Voice Source Cepstrum Processing for Speaker Identification

by

JÓN GUÐNASON

BSc., MSc.

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 Group
Department of Electrical and Electronic Engineering
Imperial College London
University of London

2007

Abstract

Voice source analysis and modelling has played a key role in important speech applications such as speech recognition, speech synthesis and speaker recognition. This work presents a robust algorithm for glottal closure detection and a novel set of voice source features for speaker recognition.

In the first part of the dissertation the DYPSA algorithm is developed for detecting glottal closure instants (GCIs). It includes a detailed study of group delay functions and their application to the linear prediction residual; glottal closure candidate generation from the group delay function; cost function design with regards to the properties of the speech signal at the point of closure; and dynamic programming algorithm used to reject unlikely glottal closure candidates. The DYPSA algorithm is evaluated on a speech database that includes simultaneous laryngograph recording to provide reference glottal closures instants. The algorithm achieves a 95.7% identification rate with 0.71 ms timing error standard deviation.

In the second part of the dissertation GCI detection allows the vocal tract transfer function to be estimated using closed-phase analysis. This is converted to cepstrum coefficients (VTCC) and subtracted from the mel-frequency cepstrum coefficients (MFCC) to derive a set of voice source cepstrum coefficients (VSCC). These are then used for speaker identification on the TIMIT database. We show that although a classifier using MFCC performs better than one using VSCC, the combination of the two gives a significant improvement in recognition rate, illustrat-

ing that additional information is present in the voice source. For example, when using test sequences of 1.0 s average duration modelled using 32 component diagonal Gaussian Mixture Model, the MFCC misclassification rate was reduced from $5.83 \pm 0.81\%$ to $2.98 \pm 0.59\%$ when combined with voice source cepstrum.

Acknowledgments

I would like to thank everyone in the Communications and Signal Processing Lab for making my time in London an enjoyable experience. I would like to extend special thanks to Glenys Benson and Mazie Paul for their helpfulness and cheerful attitude in administrating the group. For the rest of the group I would especially like to thank, David Sadler, Chin Woon Hau, Hugh Barnes, Ersi Chorti, Thushara Hewavithana, Esther Rodriguez-Villegas, Alex Wright, Nikos Mitianoudis, Dungarat Gansawat, Sarah Lee, Li Wenmin, Nikolay Gaubitch, Jesse Berent, Loic Baboulaz and Nicolas Gehrig.

I am also very grateful to Patrick Naylor for his support throughout this work. His supervision of the VERIPHON project was very educational and his help with the DYPSA algorithm was invaluable. The work done by Anastasis Kounoudes was very useful and I am indebted to him for introducing me to DYPSA. I have also had the pleasure of working with Jingjing Cui on hidden Markov models and feature extraction of radar data and I would like thank her for her collaboration.

Special thanks to my supervisor Mike Brookes whose suggestions, criticism and advice has made this work possible. I am grateful for the balance he struck between allowing me to explore different tangents and focusing on the job at hand.

Finally, I'd like to thank my family whose support proved invaluable to me. Special thanks go to Elin and Dominic for putting up with me here in London at my busiest times.

I know how important this work was to my late father, Guðni Jónsson, *so I would like to dedicate this thesis to his memory.*

Contents

List of Figures

List of Tables

Statement of Originality

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

1. *The analysis of group delay functions in Section 4.3.*

   - The definition of the energy-weighted group delay function given in Equation 4.13.
   - The time domain reformulation of the zero-frequency and energy-weighted group delay measures.

2. *Quantitative performance evaluation of GCI detection methods defined in Section 2.1.* Comparisons of methods presented in the literature has so far been hampered by the lack of performance assessments presented by the authors.

3. *Enhancement of GCI detection performance by using small analysis window size.* We present an analysis of the trade-off between false alarm and miss rate in Section 4.4 and show how this can be used to reduce miss rate when dynamic programming is used to reduce false alarm rate in Section 5.2.1.

4. *Redefinition of the DYPSA algorithm.* The basic processing steps of the DYPSA algorithm is the original work of Anastasis Kounoudes et al. [Kounoudes *et al.*, 2002b]. The algorithm has been redefined in the following way.

- We use the more effective energy-weighted group delay function for candidate generation instead of the average group delay function.
- The cost function of the dynamic programming has been changed. We have added the projected candidate cost as a cost term (Equation 5.8) and we have modified the Normalised Energy cost term (Equation 5.10) and the Idealised Slope cost term (Equation 5.12).

5. *The voice source cepstrum coefficients for speaker recognition.* Voice source analysis has been used to derive feature coefficients for speaker identification. The contribution of the vocal tract is subtracted from the derived coefficients in the cepstrum domain and therefore avoids the need for inverse filtering.

Publications

Brookes *et al.* 2006 , D. M. Brookes, P. A. Naylor, and J. Gudnason. A Quantitative Assessment of Group Delay Methods for Identifying Glottal Closures in Voiced Speech. *IEEE Transactions on Speech and Audio Processing*, 14(3), pp. 456–466, May 2006.

Naylor *et al.* 2007 , P. A. Naylor, A. Kounoudes, J. Gudnason, and D.M. Brookes. Estimation of Glottal Closure Instants in Voiced Speech using the DYPSA Algorithm. *IEEE Transactions on Speech and Audio Processing*, 15(1), pp. 34–43, January 2007.

Gudnason and Brookes 2007 , D. M. Brookes and J. Gudnason. Voice Source Cepstrum Coefficients for Speaker Identification. *In preparation.*.

Abbreviations

AM	Amplitude Modulation
AR	Auto-regressive
ARMA	Auto-regressive Moving-average
ARPA	Advanced Research Projects Agency
ARX	Autoregressive-exogenous
AV	Average (to estimate group-delay)
DC	Direct Current (to estimate group-delay)
DFT	Discrete Fourier Transform
DTFT	Discrete Time Fourier Transform
DP	Dynamic Programming
DYPSA	Dynamic Programming Phase Slope Algorithm
EAGLES	Expert Advisory Group on Language Engineering Standards
EIH	Ensemble Interval Histograms
EM	Expectation-Maximisation
EP	Energy-weighted Phase (to estimate group-delay)
EPF	Energy-weighted Phase, using estimated glottal flow (to estimate group-delay)
EPS	Energy-weighted Phase, using the speech signal (to estimate group-delay)
EPSRC	Engineering and Physical Science Research Council
EW	Energy-Weighted (to estimate group-delay)
FM	Frequency Modulation
FN	Frobenius Norm method [Ma *et al.*, 1994]

GCI	Glottal Closure Instant
GD	Group Delay method [Murthy and Yegnanarayana, 1999]
GMM	Gaussian Mixture Models
GOI	Glottal Opening Instant
HMM	Hidden Markov Model
HQTx	High Quality detection of Time of Excitation in laryngograph
ICASSP	International Conference on Acoustics Speech and Signal Processing
LF	Liljencrants-Fant model of voice source
LPC	Linear Prediction Coding
LPCR	Linear Prediction Coding Residual method [Wong *et al.*, 1979]
MFCC	Mel-Frequency Cepstrum Coefficient
NZC	Negative-going Zero Crossing
PSOLA	Pitch Synchronous Overlap Add Method
PLP	Perceptual Linear Prediction
PSP	Phase-Slope Prediction
RASTA	Representations of Relative Spectra
SNR	Signal to Noise Ratio
SVM	Support Vector Machine
UBM	Universal Background Model
VTCC	Vocal-Tract Cepstrum Coefficient
VSCC	Voice Source Cepstrum Coefficient

List of Symbols

n	time index (in an utterance)
m	time index (within a window)
M	window size
$w(m)$	(Hamming) window
f_s	sampling frequency
E_n	total squared error
$s(n)$,$s_n(m)$	speech signal
$u(n)$,$u_n(m)$	LPC residual
$u_L(n)$	lip volume velocity
$u_G(n)$	glottal volume velocity
$u_D(n)$	glottal flow derivative
z	complex variable in the z-domain
$R(z)$	lip radiation transfer function
$V(z)$	vocal tract transfer function
$G(z)$	glottal pulse transfer function
K_R	lip radiation gain constant
K_V	vocal tract gain constant
K_V	glottal pulse gain constant
P	number of segments in lossless tube model
p	tube index; LPC parameter index

A_p	cross sectional area of tube p
L	length of vocal tract
v_c	speed of sound
T_v	temperature in vocal tract
ρ_p	reflection coefficients
a_p	vocal tract transfer function denominator coefficient
$\hat{a}_p$	estimated LPC parameter
b_q	vocal tract transfer function nominator coefficient
q	transfer function nominator index
P_q	number of nominator coefficients
$\jmath$	lag index
$\Phi_n(i,p)$	short term covariance of $s(n)$
$\phi_n(i)$	autocorrelation function of $s_n(n)$
$z_{1,2}$	poles in the two-pole source model
$\Re$,$\Im$	real and imaginary components
$M_{1,2}$	Rosenberg and LF time parameters
$\alpha_{1,2}$	LF shape parameters
$\mathcal{A}_{1,2}$	LF amplitude parameters
ϵ	pre-emphasis, leaky integrator coefficient
$H(z)$	recording equipment high pass transfer function
$A(z)$	phase correcting filter
$\breve{n}_r$	reference glottal closure instant
$\hat{n}_r$	reference glottal opening instant
r	GCI sample index
$\mathcal{C}_n$	set of all closed phase samples in window n
$\eta(n)$	normalised total squared error
$f(n)$	frobenius norm function

M_F	number of observation in the Frobenius norm data matrix
M_f	window size for the Frobenius norm data matrix
ζ	timing error between detected and true GCI
σ	standard deviation of ζ
$x_n(m)$	windowed residual (input for group-delay measure)
$\tilde{X}_n(\omega)$	DTFT of $x_n(m)$
ω	angular frequency
$\tilde{\tau}_n(\omega)$	continuous group delay
$\breve{\tilde{X}}_n(\omega)$	DTFT of $mx_n(m)$
k	discrete frequency $\omega = \frac{2\pi k}{M}$
$\tau_n(k)$	discrete group delay
$X_n(k)$	DFT of $x_n(m)$
$\breve{X}_n(k)$	DFT of $mx_n(m)$
$\delta(m)$	impulse function
m_0	index of impulse within a window
$d_*(n)$	group delay measure with * denoting DC,AV,EW,EP,EPS,EPF
$\jmath$	time lag index
$d'_*(n)$	shifted group delay measure
β and β'	relative strength of two impulses
N_{DP}	DP look back
ψ	total cost in DP
$\bar{\psi}$	cost vector in DP
ψ_*	cost terms with * denoting A,P,J,F,S
λ_*	cost weights
ξ_{n_1,n_2}	covariance of segments centred on n_1 and n_2
Δ_P	pitch consistency measure
$\mathcal{K}$	pitch deviation constant
$F(n_r)$	Frobenius energy at n_r

$\breve{F}$	local minimum of $F(n_r)$
$M_{\breve{F}}$	window length for $\breve{F}$
M_S	window length for group delay slope
$\dot{d}$	group delay slope
$\breve{n}$	reference GCI
$\mathbf{c}$	mel-frequency cepstrum coefficient vector
$\mathbf{c_{vs}}$	voice source cepstrum coefficient vector
$\mathbf{c_{vt}}$	vocal tract cepstrum coefficient vector
$S(n,k)$	discrete Fourier transform of speech frame $s_n(m)$
K_s	number of DFT points
$Y(n,j)$	Output of mel-filterbank applied on $\|S(n,k)\|$
$Y_{vt}(n,j)$	Output of mel-filterbank applied on $\|V(n,k)\|$
$Q_j(k)$	Mel-filter number j
r	mel-filterbank index
N_r	number of mel-filters
$c(n,l)$	mel-frequency cepstrum coefficient
$c_{vt}(n,l)$	vocal tract cepstrum coefficient
$c_{vs}(n,l)$	voice source cepstrum coefficient
N_c	Number of cepstrum coefficients
l	cepstrum coefficient index
$\mathcal{E}_n\{\cdot\}$	expected value operator
κ	window size parameter for delta coefficients
ι	delta coefficient frame lag
C_i	Utterance spoken by speaker i
i	speaker index
$\hat{\imath}(C_i^t)$	estimated speaker index

γ_i	misclassification rate for speaker i
t	test utterance index
T_i	number test utterances from speaker i
N_s	number of speakers in a set
$\bar{\gamma}_i$	average misclassification rate
γ	test-set misclassification rate
T	total number of test utterances
$\bar{\gamma}_{M \text{ or } F}$	male/female average misclassification rate
$\bar{\gamma}_{GB}$	gender-balanced misclassification rate
$\mathcal{X}_\gamma$	random variable for error
σ_γ	standard error
χ_i	event that speaker i has spoken
$C = \{c_1, \ldots, c_{N_T}\}$	feature vector sequence
c_j	feature vector j
j	test utterance feature vector index
N_T	number of feature vectors in a test utterance
$Pr\{\cdot\}$	probability of event
$f_C(C)$	probability density function of C
$\ell_i(c)$	log-likelihood of C given χ_i
o	mixture component index
D	dimension of feature vector C
N_o	number of mixture components in GMM
$\beta_m^{(i)}$	GMM mixture weights for model i
$\boldsymbol{\mu}_{\mathbf{m}}^{(\mathbf{i})}$	GMM mixture means for model i
$\boldsymbol{\Sigma}_{\mathbf{m}}^{(\mathbf{i})}$	GMM mixture covariance matrices for model i
θ	combination weight

Chapter 1

Introduction

THIS work focuses on feature extraction for speaker identification using voice source analysis. The acoustic speech signal is realised by the voice production mechanism with injection of air through the vocal tract and out through the lips. In voiced speech, the voice source is effected by the lung pressure and the vibrating vocal folds and our approach represents this signal for speaker identification.

The hypothesis for this work as a whole is that speaker identification can be improved by explicitly presenting information about the voice source to the classifier and that to describe the voice source, we need an accurate representation of the vocal tract. This representation is made possible with closed phase analysis of voiced speech which relies on the identification of glottal closure instants.

State-of-the-art speaker recognisers rely on the mel-frequency cepstrum coefficients for classification. These coefficients were originally designed to distinguish between different phonemes in speech [Davis and Mermelstein, 1980] and have become fundamental to feature extraction in speech recognition. It is not obvious why these coefficients should be used for speaker recognition since their design was originally intended to distinguish between different phonemes and not between speakers.

The reason why we concentrate on the voice source signal is because it has

been shown to contain speaker characteristics useful for speaker recognition [Karlsson, 1988]. Most of the recent work done on voice source related features has concentrated on prosody where larger scale features such as pitch and intensity are extracted for recognition [Shriberg *et al.*, 2005]. We will focus on smaller scale spectral features of the voice source that have thus far received less attention [Plumpe *et al.*, 1999].

In this chapter we give an overview of background material that is relevant to this work. We present a brief discussion on speech and voice in Section 1.1 and review previous work on linear speech modelling in Section 1.2. We present an overview of speaker recognition in Section 1.3 and conclude the chapter by outlining the research presented in this dissertation.

1.1 Speech and Voice

1.1.1 Speech physiology

Spoken language is produced by the movement of air from the lungs through the trachea, through the larynx and the vocal tract out the mouth and the nose. The vocal organs are shown in Figure 1.1. The source of energy is the abdominal and thoracic muscles drawing air into the lungs. It is expelled back by contracting the rib cage and increasing the lung pressure. The lung pressure is kept relatively steady during speech and is only modified to change the intensity of the speech [Flanagan, 1972].

The air flow passes through the larynx throat cavity which contains the vocal folds, which are sometimes confusingly called vocal cords since they are not cords but flaps of elastic tissue that can vibrate when air passes through. The vocal folds control the voicing of the sound so that when they are tight together they vibrate

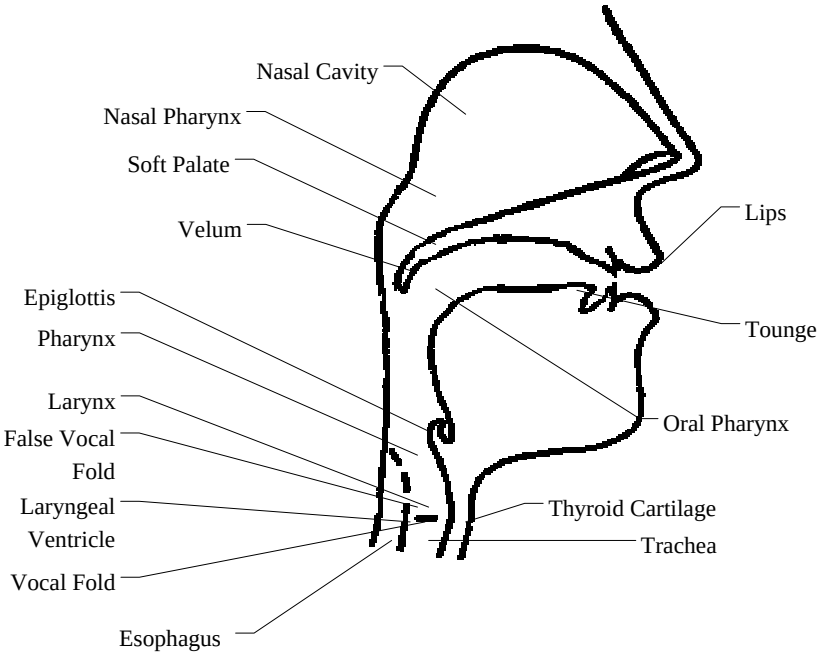

Figure 1.1: Schematic of the vocal organs.

at the fundamental frequency of the voiced speech. But when the vocal folds are relaxed the air passes through freely and unvoiced speech is formed either by a turbulent airflow past a restriction in the vocal tract or brief transient sounds can be generated when a point of total closure in the vocal tract is opened and the pressure is abruptly released [Rabiner and Juang, 1993; Jurafsky and Martin, 2000].

Voiced speech is produced by the vibrating vocal folds and the modulation of the airflow by the vocal tract. The vocal folds are kept tense and the steady flow from the lungs cause them to self oscillate. The tension in the folds determines the frequency of their vibration which in turn determines the fundamental frequency of the speech. The voluntary movements of the articulators, namely the velum, tongue, jaw and lips, change the cross-sectional area function, and thereby the acoustical response of the vocal tract. The glottal volume velocity flow passes through the vocal tract and resonances of the pressure wave effect the formant structure and the perceived sound [Flanagan, 1972].

1.1.2 Articulatory phonetics

Speech sound units are referred to as phones and their systematic description is called phonology. This is the description of every possible sound unit that can be produced by the human voice and is therefore understandably quite extensive and involved. Another description is provided by phonemics, which describe a speech utterance by a string of phonemes. A phoneme is a mental abstraction of a physical sound so two different phones can be said to represent the same phoneme if they are perceived to be the same sound. In this case it is said that the two phones are allophones of the same phoneme. The aim of speech recognition is to describe speech using phonemics. The standard alphabet to describe phones is the International Phonetic Alphabet and a subset of this is the ARPAbet which is relevant for the English language [Deller *et al.*, 1993; Jurafsky and Martin, 2000].

Phonemes are divided into two main categories: consonants and vowels. Consonants are created by restricting or blocking the airflow in some ways and may be voiced or unvoiced. Vowels are longer, louder, and voiced, and are created with less obstruction in the vocal tract. Other categories of phonemes are diphthongs, which are formed with a slide from one vowel to another, and semivowels which have some properties of both vowels and consonants [Gold and Morgan, 2000; Rabiner and Juang, 1993].

Consonants can be described by the manner and place of articulation, which is the position of the restriction in the vocal tract when they are generated. Consonants are considered to be labial, dental, aveolar, palatal, velar or glottal depending on the places of articulation. Examples are (in the same order), /b/ as in *bear*, /th/ as in *thing*, /s/ as in *sage*, /jh/ as in *sage*, /k/ of *catnip* and /h/ as in *hat* (sometimes the glottal consonants are considered as "all other" consonants). Consonants can also be described according to their manner of articulation and are then referred

to as stops, nasals, fricatives, approximants or taps [Jurafsky and Martin, 2000; Botinis *et al.*, 2001].

Vowels are normally among the phones of largest amplitude and duration and are therefore very useful in speech analysis. Their duration can be from 40 to 400 ms and they vary according to the cross-sectional area of the vocal tract. As with consonants, vowels can be classified by their place of articulation. Their manner of articulation is said to be "vowel", though linguists tend to use three parameters to describe vowels: frontness, height, and roundness.

Frontness is defined by where the tongue is raised, front or back. For example, /i/ as in *b<u>i</u>t* is a front vowel and /o/ as in *b<u>oa</u>t* is a back vowel. Frontness correlates with the second formant frequency [Gold and Morgan, 2000]. Height is determined by how much the lower jaw is dropped when forming the vowel. High vowels have the lower jaw close to the upper, e.g. /u/ in *b<u>oo</u>t* and low vowels have the jaw further away from the upper jaw, e.g. /aa/ as in *p<u>o</u>ppy*. Roundness refers to the shape of the lips. Vowels articulated with rounded lips are referred to as rounded, e.g. /uw/ as in *t<u>u</u>lip* [Rabiner and Juang, 1993].

1.1.3 Speech signal characteristics

The speech signal is considered to be slowly time-varying which means that stationarity can only be assumed over a short time interval. This can be seen in Figure 1.3 which shows the wide- and narrow band spectrograms and the speech signal of a sentence uttered by a male saying: "*She had your dark suit in greasy wash water all year.*" The two spectrograms show the log-magnitude of the signal represented in reverse grey-scale with time represented on the abscissa and frequency on the ordinate. The wideband spectrogram (upper graph) is evaluated over a shorter time window giving features with sharp changes in time more emphasis. The narrowband

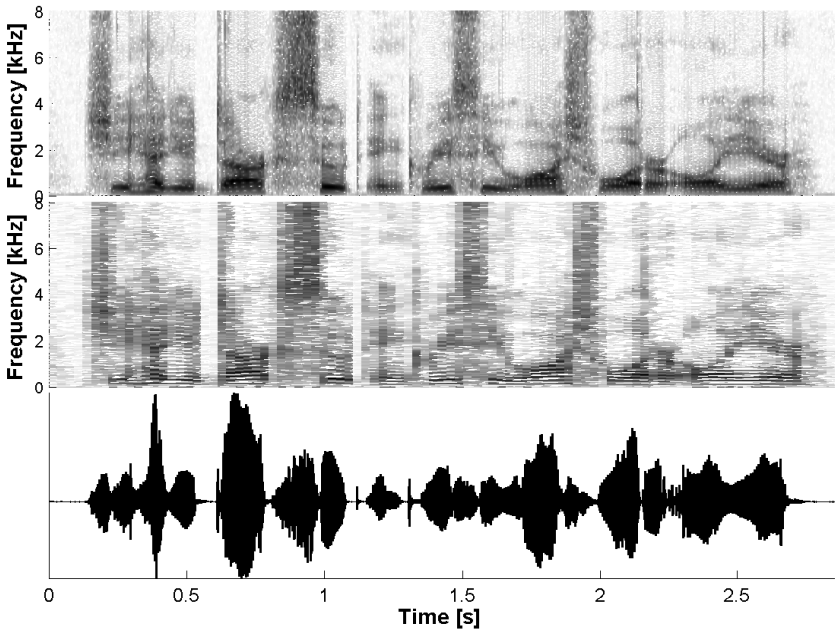

Figure 1.2: Speech characteristics represented by wideband (above) and narrowband (middle) spectrograms.

spectrogram (middle graph) is evaluated over a longer time window giving greater resolution in frequency. We can see many typical characteristics of speech in the spectrogram. Voiced vowel sounds appear as high amplitude periodic segments with the energy concentrated in low frequency, such as the /a/ sound in *dark*, around 0.65 s. The wide-band /s/ sound from *suit*, at around 0.90 s, can be seen in the spectrograms to extend over the low and high frequency part of the spectrum.

Figure 1.3 shows a voiced 35 ms segment of the sentence and two spectral representations of that segment. The middle trace shows the discrete Fourier Transform magnitude spectrum of the segment, showing energy concentrated in the low-frequency part of the spectrum, as is typical for a voiced segment like this, and two or three peaks almost visible in 2.2 and 4.0 kHz. The bottom trace shows the autoregressive (AR) magnitude spectrum derived by estimating an 18 coefficient AR filter of the segment. We will discuss this representation in more detail in next section, but we can see from the graph that a parametric representation of speech

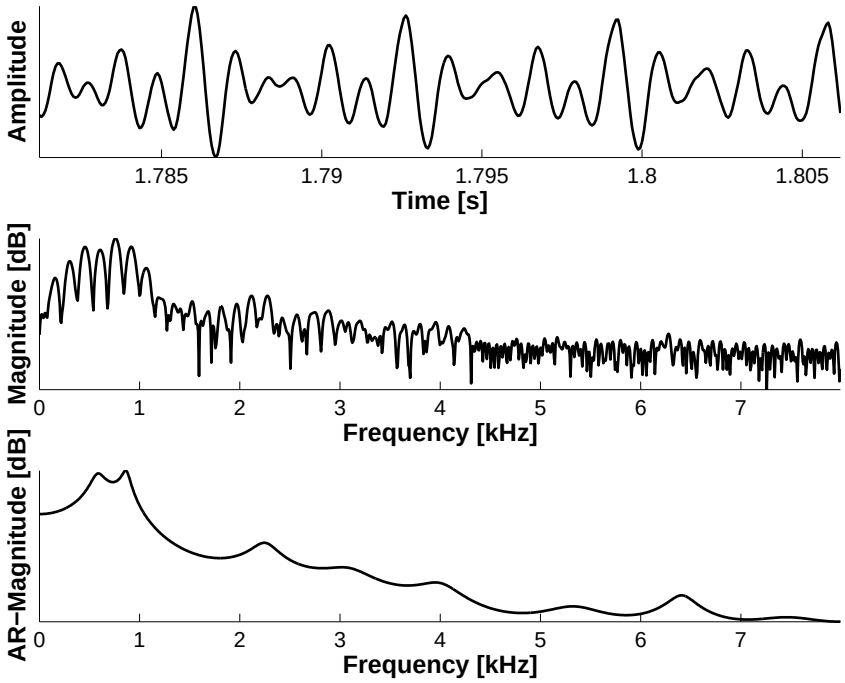

Figure 1.3: A small window of speech, its Fourier spectrum, and its AR spectrum.

can emphasise features that are latent in the full spectrum.

1.1.4 Acoustics in the Speech Production

The acoustic wave can be modelled at each stage of the voice production usually only relying on the recorded speech but the laryngograph and video recordings can also be used to aid the modelling [Krishnamurthy and Childers, 1986; Miller, 1959]. If the shape of the vocal tract is known the acoustic sound wave in the vocal tract can be described as a function of the glottal air-volume velocity, by solving the wave equation with boundary conditions appropriate to the tube which the tract forms [Beranek, 1954; Morse and Ingard, 1968]. It is theoretically possible to solve a Navier-Stokes equation for non-planar sound waves and boundary conditions appropriate for a lossy vocal tract but this analysis needs the knowledge of the exact physical properties of the vocal tract and its shape.

The shape of the vocal tract has been studied by X-ray tracings and it has been represented by time-dependent variables specifying the positions of various articulators such as the jaw, the tongue, the lips and the velum [Mermelstein, 1973]. This kind of model can include spatial constraints that follow those of natural articulation and the motion of the articulators to imitate the transition from one articulatory state to another [Coker, 1976]. This approach has, for example, been used recently for articulatory speech synthesis [Birkholz *et al.*, 2006]. The vocal tract shape was determined for vowels [Sorokin, 1992] and fricatives [Sorokin, 1994].

The modelling of the sound wave in the vocal tract has received a lot of attention and has led to the linear source-tract model [Fant, 1960] that we will describe in detail in the following section. Non-linear models that don't rely on all the assumptions needed for the linear source-tract model have also been developed. A model for the wave propagation in a lossy vocal tract has been suggested where thermal losses and viscosity are taken into account [Sondhi, 1974]. Other non-linear methods include the "black-box" approach, where the system parameters do not correspond to any physical variables. For example, the dynamics of speech was modelled using nonlinear predictors [Tishby, 1990] and neural networks have been used for speech coding [Thyssen *et al.*, 1994]. Nonlinear dynamical analysis of speech has also been presented [Kumar and Mullick, 1996].

1.2 The Linear Source-Tract Model

The linear modelling of the speech production was motivated by the lossless tube model of the vocal tract [Dunn, 1950; Dunn, 1961]. Fant conducted a study on speech intelligibility [Fant, 1960; Fant, 1968] and Miller estimated the voice source signal by using the inverse of the first vocal resonance and the vocal fold opening area measured by video [Miller, 1959]. A linear source-tract model was proposed

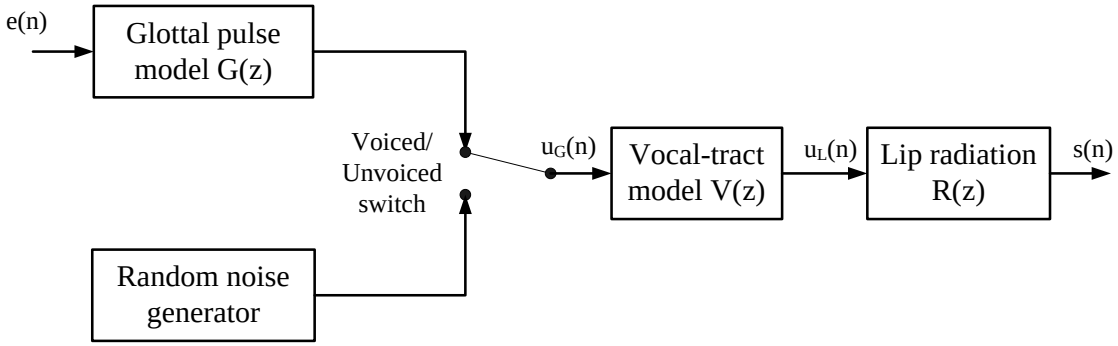

Figure 1.4: The source-filter model of voice production for voiced and unvoiced speech.

to represent the radiation impedance, vocal tract, and the glottal source as linear filters identified using covariance analysis [Strube, 1974; Markel and Gray, 1976; Wong *et al.*, 1979].

1.2.1 Linear speech production model

A general discrete-time linear speech production model, shown in Figure 1.4, describes the voiced and unvoiced modes of speech separately. The signals in the model are only superficially analogous to the waveforms present in and by the vocal organs themselves and the emphasis has been to represent the speech signal itself [Rabiner and Schafer, 1978].

The linear model for voiced speech is shown in Figure 1.5(left), where $s(n)$ is the sampled speech waveform, n is the sample number, $u_L(n)$ is the volume velocity signal at the lips and $u_G(n)$ is the glottal signal which is the input into the vocal tract. The system is assumed to be time-invariant over a time period of c.a. 30 ms so the filters can be reordered to simplify the analysis [Rabiner and Schafer, 1978]. The radiation transfer function $R(z)$ is moved back so that the glottal volume velocity signal is filtered by $R(z)$ before it is processed by the vocal tract, as shown in

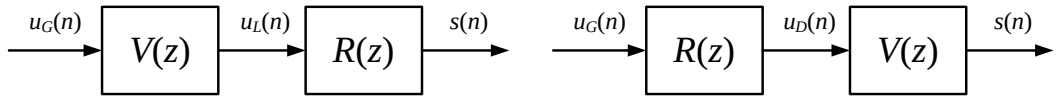

Figure 1.5: The filter model for voiced speech. The two filters represent the vocal tract and the lip radiation.

Figure 1.5(right).

The signals and the filters in the system are the following:

$u_G(n)$ is the volume velocity signal through the glottis

$u_D(n)$ is the $R(z)$-filtered glottis volume velocity signal (only in the reordered model of Figure 1.5)

$u_L(n)$ is the lip volume velocity signal through the vocal tract (only in the original model of Figure 1.5)

$s(n)$ is the speech pressure signal recorded by the microphone

$V(z)$ is the vocal tract transfer function

$R(z)$ is the lip radiation transfer function

The recorded speech, $s(n)$, is a sound pressure wave and is related to the acoustic velocity at the lips, $u_L(n)$ through the radiation transfer function, $R(z)$. Acoustic theory predicts $R(z)$ to be a first order high-pass filter with a cut-off frequency of around 5 kHz so for sampling frequency of less than 20 kHz, a good approximation for $R(z)$ is [Wong *et al.*, 1979; Deller *et al.*, 1993],

$$R(z) = K_R(1 - z^{-1}) \tag{1.1}$$

where K_R is a gain constant and the acoustic delay is ignored[1].

[1]This depends on the distance from the lips to the microphone and we assume that it does not vary fast enough to affect any of the analysis of this work.

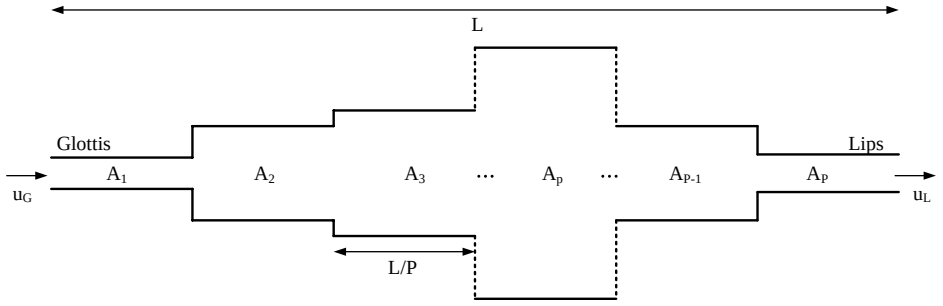

Figure 1.6: The vocal tract represented by concatenated lossless tubes of equal length but different cross-sectional area.

1.2.2 The vocal tract transfer function

The vocal tract filter $V(z)$ can be be approximated as an all-pole filter [Dunn, 1950; Fant, 1968]. The all-pole filter characterising the vocal tract can be derived from the wave-equation by making approximations and assumptions regarding the acoustic properties of the vocal tract and the behaviour of the sound wave in the process [Morse and Ingard, 1968]. The vocal tract shape is approximated as a concatenation of P rigid tubes each of constant diameter but equal length L/P, where L is the length of the vocal tract, typically $15-17$ cm in adults. This is demonstrated in Figure 1.6. The length of each segment needs to be approximately equal to the distance that sound travels in half a sample period [Rabiner and Schafer, 1978],

$$\frac{L}{P} \approx \frac{v_c}{2f_s}, \tag{1.2}$$

where $v_c \approx 20\sqrt{T_v} \approx 350$ m/s is the speed of sound in the vocal tract, $T_v \approx 305$°K is the air temperature in the vocal tract, and f_s is the sampling frequency. This means that for a speech signal with $f_s = 16$kHz sampling frequency, we need $P \approx \frac{2Lf_s}{v_c} \approx 15$ tube segments to approximate the vocal tract.

The objective of this analysis is to derive the transfer function $V(z)$ between the glottal air-flow, $u_G(n)$, and the flow at the lips, $u_L(n)$. The two main assumptions

used in this derivation is that there is no energy loss in the vocal tract and that the sound pressure waves in the vocal tract are longitudinal plane waves. The assumption of no energy loss in the vocal tract means that the walls are assumed to be rigid and that there is no turbulent flow or viscosity. The longitudinal plane wave assumption means that the sound pressure wave is independent of cross sectional coordinates.

The analysis of the lossless tube model relies on representing the acoustic wave in the vocal tract as the superposition of the acoustic velocity in the direction of the lips and the acoustic velocity in the direction of the lungs [Kelly and Lochbaum, 1962; Rabiner and Schafer, 1978]. The flow entering tube $p+1$ in the direction of the lips can then be expressed as the scaled and delayed version of the flow entering tube p depending on the tube's length L/P and the cross-sectional areas of the tubes A_p and A_{p+1}. Similarly, the flow exciting tube p in the direction of the lungs can be expressed in terms of the flow exiting tube $p+1$. A linear transfer function $V_p(z)$ for one tube-segment can be derived using these expressions and the transfer function for the concatenated tube-segments is derived by multiplying these together, yielding the vocal tract transfer function,

$$V(z) = \frac{z^{-P/2}\prod_{p=0}^{P}(1+\rho_p)}{1-\sum_{p=1}^{P}a_p z^{-p}} = \frac{z^{-P/2}K_V}{1-\sum_{p=1}^{P}a_p z^{-p}} \tag{1.3}$$

where $K_V = \prod_{p=0}^{P}(1+\rho_p)$ is a gain constant, $z^{-P/2}$ is the acoustic delay of the vocal tract, the reflection coefficients are given by,

$$\rho_p = \frac{A_{p+1}-A_p}{A_{p+1}+A_p}. \tag{1.4}$$

and a_k is obtained using Cholskey decomposition or the Levinson-Durbin algorithm [Durbin, 1959]. The estimation of a_p from the speech waveform is very efficient computationally which is a strong motivation for using the lossless tube approach

for voice modelling.

1.2.3 Lossless Tube Modelling Assumptions

The linear model of the vocal tract, $V(z)$, is justified by acoustic theory using the lossless tube model but it is in fact a gross simplification of the complicated effects that occur during voice production. Many assumptions and approximations are made in the derivation of the discrete linear filter model and when the filter parameters of $V(z)$ are estimated. Here we discuss the strengths and weaknesses of the linear filter model of the voice production.

The first step in deriving the transfer function in Equation 1.3 using the lossless tube model, is to approximate the vocal tract area function as being piecewise-constant. By carefully choosing the number of tube segments, P, as explained in Equation 1.2, any adverse effect of this approximation can be reduced. Using fewer tube segments has an adverse effect on the modelling whereas choosing the number of tube segments above the optimum number only affects the modelling if there is not enough speech data to estimate the paramters [Markel and Gray, 1976].

The acoustic wave in the vocal tract is also assumed not to suffer any energy loss. The walls of the vocal tract are not rigid so acoustic energy is in fact lost due to the vibration of the walls and viscosity and turbulence of the airflow. The mathematical analysis is computationally intractable as it requires the solution to wave equations with difficult boundary conditions and is normally not considered [Rabiner and Schafer, 1978]. Usually, this effect is either ignored or accounted for by raising and broadening the formant frequency peaks. The vibration of the walls causes the low frequency formant peaks to broaden and shift and the turbulent flow will affect the high frequency formants in the same way. The lossless tube model can therefore be adjusted after the estimation process [Atal and Hanauer, 1971;

Flanagan, 1972].

The acoustic sound wave is assumed to be constant over the cross-section of the vocal tract. This is not a bad approximation if we can assume that the voice source is decoupled from the vocal tract. This has been disputed since the sound pressure wave does not uniformly fill the vocal tract as the vocal folds open [Teager, 1980; Cranen and Boves, 1987; Cranen and Boves, 1988].

The lossless tube model includes the boundary effect between the glottis and the vocal tract and the vocal tract and the lips by including two infinitely long fictitious tubes at either end of the lossless tube model. The tube representing the larynx boundary effect is justified by the assumption that the acoustic wave going into the subglottal cavity is completely is absorbed [Deller *et al.*, 1993]. The lip boundary effect is represented by infinitely long (fictitious) tube and represents the way the flow is broken by a large flat radiation surface (which in reality is the head of the speaker) [Rabiner and Schafer, 1978].

When nasal consonants are produced the velum is lowered so that the nasal tract is coupled to the pharynx and the oral tract is simultaneously closed (see Figure 1.1). For nasal vowels the velum is also lowered but the oral tract is kept open as for other vowels. An appropriate modelling of such a side cavity is to introduce zeros to the all-pole formulation of the vocal tract [Rabiner and Schafer, 1978]. The vocal tract model is therefore represented as an ARMA model,

$$V(z) = \frac{\sum_{q=0}^{P_q} b_q z^{-q}}{1 - \sum_{p=1}^{P} a_p z^{-p}} \tag{1.5}$$

where the acoustic delay of the vocal tract has been omitted. The estimation of the parameters b_q and a_p can be done using standard linear-least squares methods [Kay, 1988]. The zeros are difficult to estimate accurately and including them does not, for example, improve the representation of the speech signal synthesised by the

model [Kleijn and Paliwal, 1995]. It is a common practice to exclude the zeros and make up for any modelling deficiency with more poles instead [Krishnamurthy and Childers, 1986].

The only known quantity in the linear voice source model, illustrated in Figure 1.5, is the output signal $s(n)$. We can also approximate $R(z)$ up to a fixed scale and time delay, and we have determined the form of $V(z)$ given the approximations and assumptions mentioned earlier. The parameters of $V(z)$ are a_p and depend on the location of the articulators, such as the tongue and lips, which are moved to produce different sounds. Therefore the parameters need to be estimated over successive speech frames where the articulators can be assumed to be still. We have presented the source-tract model by assuming that the place of the source is the larynx, whereas for many consonants, the place of articulation can be viewed as a place of the source for a shorter tube model of the vocal tract [Rabiner and Schafer, 1978].

For voiced speech the voice source signal $u_G(n)$ does not have a flat spectrum. During voiced speech the vocal folds vibrate at the pitch frequency but the voice source pulse shape is not an impulse. The vocal folds stays closed over a certain period of the larynx cycle as can be observed using video [Miller, 1959] or the laryngograph [Abberton *et al.*, 1989]. The voice source is analysed in more detail in Chapter 3.

1.3 Speaker Recognition

Speaker recognition is a part of the diverse field of speech processing and has had a relatively fast development in the last few years [Doddington, 1985; Naik, 1990; Gish and Schmidt, 1994; Campbell, 1997; Furui, 1997; Bourlard *et al.*, 1998; Reynolds, 2002; Bimbot *et al.*, 2004]. Speaker recognition itself encompasses all applications

involving the determination of a speaker's identity, whether it be a claimed identity, as in *speaker verification* [Naik, 1990; Bimbot *et al.*, 2004] or an identity from a set of known speakers as in *speaker identification* [Doddington, 1985; Reynolds, 1995; Campbell, 1997]. Closed-set speaker identification means that the system is forced to choose a speaker from the set whereas in open-set speaker identification it can reject the speech utterance deciding that it does not belong to any of the speakers in the test. Other applications include for example speaker detection and tracking where a new speaker is enroled and recognised on-the-fly [McLaughlin and Reynolds, 2002; Reynolds, 2002; Klusacek *et al.*, 2003]. Speaker recognition can be text-dependent [Che *et al.*, 1996] which requires the speaker to utter a specific word or a sentence or it can be text-independent [Reynolds, 1995; Bimbot *et al.*, 2004] where the context of the speaker's sentence is unknown to the recognition. Here we give a description of speaker identification and verification before describing the most common approaches to speaker recognition.

1.3.1 Speaker identification

Speaker identification assigns an identity to a test utterance from a set of known speakers. Normally the utterance is assumed to be from a known set of speakers [Reynolds, 2002] and is therefore called closed-set speaker identification but open-set speaker identification can also reject a speaker as "none of the above" if the best speaker score does not exceed a certain threshold.

Speaker identification is normally carried out by training a speaker model for every speaker in the set and the test is implemented by pattern matching, as shown in Figure 1.7. The preprocessing and feature extraction process treats the speech signal waveform and represents it as feature vectors. These feature vectors are modelled in the training part or the system, for example by using Gaussian mixture models[2].

[2]Neural networks, support vector machines and hidden Markov models could also be used.

The identification process performs pattern matching of the derived feature vectors with each speaker model to give a likelihood and the decision module selects the most likely speaker.

1.3.2 Speaker verification

Speaker verification accepts or rejects an identity claim based on a speech utterance. If the utterance matches the claimed speaker model, the claim is accepted but rejected otherwise [Campbell, 1997]. As indicated in Figure 1.7, the comparison between the utterance and the speaker model results in a likelihood which is compared against a threshold and the accept/reject decision follows. The setting of the threshold is crucial to the performance of the system and is often based on a world-model or a cohort-model [Reynolds and Rose, 1992; Reynolds *et al.*, 2000]. Research on speaker verification has centred on this issue in recent years with the speaker likelihood and world model being normalised with respect to things such as handsets/microphones, communication channels, impostor distribution or gender [Bimbot *et al.*, 2004]. Most verification systems assume casual impostors, i.e. the causes of errors are due to difficulties with distinguishing between speakers that talk normally. Studies have also been done on vulnerability to voice mimicking, where actors imitated utterances from genuine speakers [Pellom and Hansen, 1999; Lau *et al.*, 2004].

1.3.3 Speaker recognition Methods

There are many open research questions in speaker recognition. They range from questions on how to treat speech input of varying quality; how to represent the input speech as feature vectors; what classifier to use; how to choose the correct subsets for cohort modelling; how to implement text-independent and text-dependent classifiers;

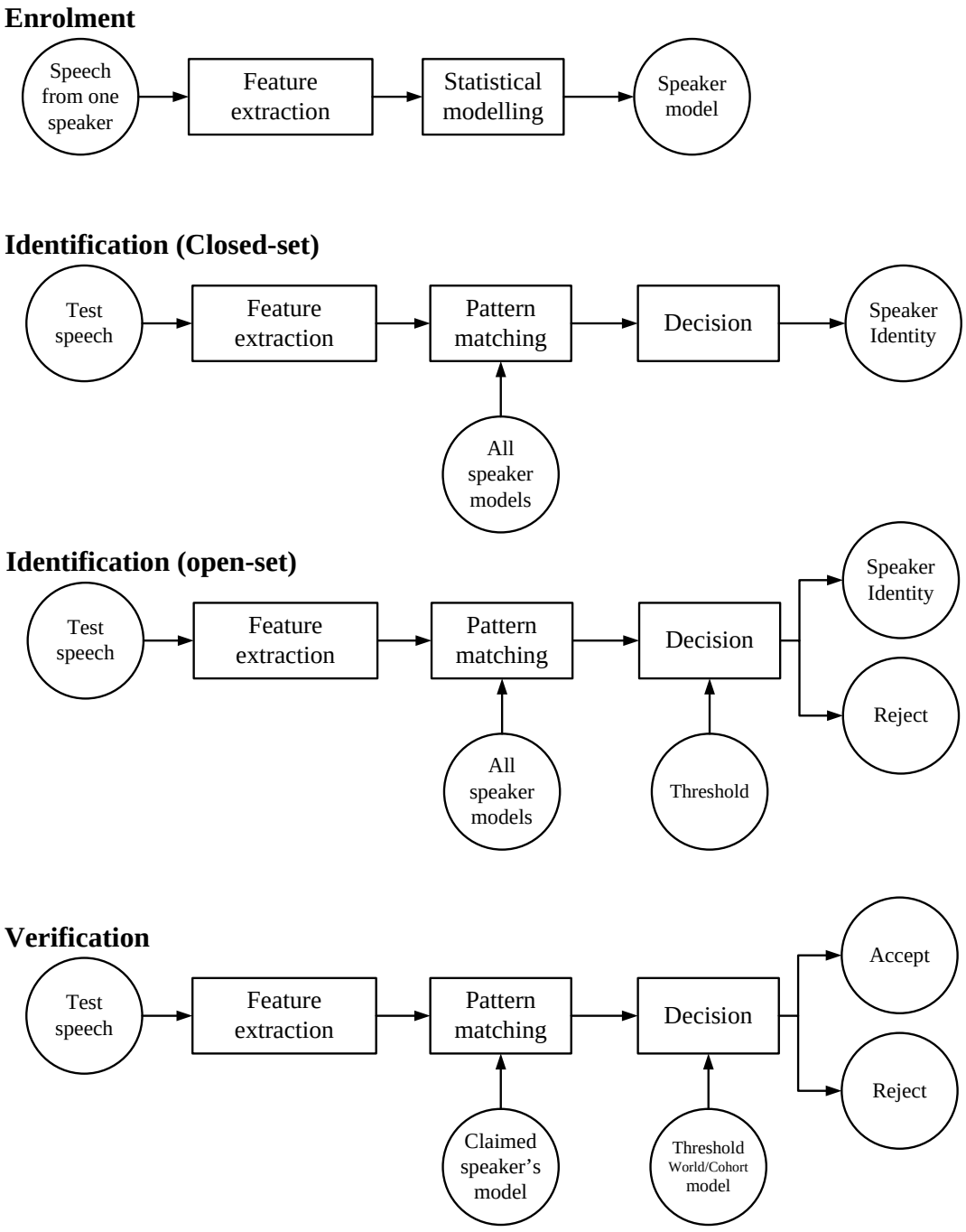

Figure 1.7: Enrolment, identification and verification of speakers.

and how to deal with uncooperative speakers.

Text-dependent speaker recognition compares the sequence of feature vectors with a model based on a known string of phonemes. One method represents a speaker with a hidden Markov model (HMM). When an unknown speaker is prompted with a sentence, the appropriate state sequence is assembled from the HMM model of the speaker being tested against. The likelihood used for recognition is then given by the comparison between the test utterance and the state sequence [Che *et al.*, 1996]. Another study applied dynamic time warping, vector quantisation and HMMs to text-dependent speaker recognition [Yu *et al.*, 1995]. Pitch and duration have been used in text-dependent speaker verification using auto-associative neural networks [Yegnanarayana *et al.*, 2005] and speaker verification has been implemented with dynamic programming using multiple templates for a given password [Ramasubramanian *et al.*, 2006]. Multi-stream hidden Markov models have been used to fuse video and speech for text-dependent speaker recognition [Lucey *et al.*, 2005].

Text-independent speaker recognition does not rely on a known string of phonemes and does therefore not rely on temporal dependency between vectors in the sequence in the way text-dependent recognisers do. A popular approach is to represent each speaker with a Gaussian mixture model (GMM) probability density function [Reynolds, 1994; Reynolds, 1995; Reynolds and Rose, 1995; Wildermoth and Paliwal, 2003]. When performing identification, the test utterance is evaluated against every speaker's GMM and the highest likelihood determines the identity of the speaker [Reynolds and Rose, 1995]. The reason why the GMM has become so popular is its versatility in approximating the probability density function of speech features.

For speaker verification, a likelihood ratio test was implemented with the likelihood of claimed speaker GMM compared against likelihood of a world- or a cohort GMM derived from a set of speakers [Reynolds, 1995]. Cohort modelling was

proposed to represent a set of speakers close to the claimed speaker [Higgins *et al.*, 1991; Matsui and Furui, 1993] but world model normalisation was proven to be just as effective but reducing the amount of computation needed in cohort modelling [Carey *et al.*, 1991; Heck and Weintraub, 1997].

The current state-of-the-art Gaussian mixture modelling approach uses a universal background model (UBM - or a world model) for score normalisation and Bayesian adaption [Braverman, 1962; Bernardo and Smith, 1996] for the speaker modelling. The UBM-GMM is estimated using speech data from all the speakers in the set and it can be chosen to contain many more mixture components than a GMM evaluated for a single speaker because there is much more data available from the entire set of speakers. The Bayesian adaption uses the UBM-GMM as the a-priori model and fits a GMM to the speech available from a single speaker [Gauvain and Lee, 1994; Reynolds *et al.*, 2000].

The problem of speaker verification can be cast into finding out whether the mismatch between a test utterance and an utterance from a known speaker is due to inter-session variability (so in this case the test utterance could be from that speaker), or if the mismatch is because of inter-speaker variability (which means that the test utterance is not from the speaker). To address this, score normalisation has been proposed and can, for example, be either an overall distribution normalisation of the likelihood sequence (Z-Norm) [Li and Porter, 1988], normalisation with respect to different handsets (H-Norm) [Reynolds, 1996], or with respect to impostor distributions (T-Norm) [Ben *et al.*, 2002; Bimbot *et al.*, 2004]. Channel compensation can be done by front-end processing with feature mapping [Reynolds, 2003] or by adapting the speaker- and background models using for example eigenchannels or factor analysis [Kenny and Dumouchel, 2004; Kenny *et al.*, 2006a; Kenny *et al.*, 2006b]

An alternative approach has emerged recently that uses support vector ma-

chines (SVM) instead of Gaussian mixture models [Campbell, 2002; Campbell, 2003; Wan and Renals, 2003; Wan and Renals, 2005]. Nonlinear mapping is performed from the input space to an SVM expansion space where linear classification techniques can be applied [Vapnik, 1999]. The training of SVM is discriminative which means that background models do not need to be estimated and they are computationally efficient for identification. But this approach has only been proved to achieve similar results to the state-of-the-art Bayesian adaptive UBM GMM approaches [Campbell *et al.*, 2006].

1.4 Outline of Thesis

The dissertation is organised as follows:

Chapter 2 Recorded speech data and evaluation methods are presented in this chapter. The APLAWD and SAM0 speech corpora contain contemporaneous laryngograph recordings from which glottal closure instants can be extracted with some certainty. We use these to evaluate glottal closure instant detection methods studied and proposed in this work. We also use the TIMIT speech corpus to evaluate the speaker identification proposed in the work.

Chapter 3 The voice source is considered and how best to separate it from the vocal tract. The most simple approach is to model the voice source with two real poles. More effective methods include applying ARX (autoregressive-exogenous) modelling on the speech signal and closed phase analysis. Identification of the closed phases of speech allows the vocal tract transfer function is evaluated during time periods where the excitation can be assumed to be zero.

Chapter 4 Glottal closure instant detection is necessary if closed phase analysis

is to be used to extract the voice source. Here we review proposed methods for GCI detection and study the group delay function and its ability to detect impulses in synthetic signals and speech. We show that the energy weighted group delay function is extremely good in detecting glottal closure instants but the false alarm rate is also high.

Chapter 5 To reduce the number of false alarms, we present the DYPSA algorithm to detect glottal closure instants in voiced speech. We base the DYPSA algorithm on candidate generation by the energy weighted group delay function and evaluate its performance.

Chapter 6 Voice source cepstrum processing is based on closed phase analysis of the speech signal which relies on a reliable detection of the glottal closures. Voice source- and vocal tract cepstrum coefficients are extracted and used in speaker identification experiments on the TIMIT database.

Chapter 7 We present a discussion on the results obtained in Chapter 4, 5 and 6. We conclude that the DYPSA algorithm can reliably detect the glottal closure instants and that information contained in the voice source is useful for speaker recognition.

Chapter 2

Speech Corpora and Evaluation Methods

THERE are two sets of experiments presented in this work. One is on detecting glottal closure instants in voiced speech and the other on identify a speaker in a set of known speakers given an unknown utterance. Here we describe the performance assessment and the data we use for these experiments.

2.1 Detecting Glottal Closure Instants

The experiments we use to evaluate the performance of glottal closure instant detectors rely on the laryngograph. We extract glottal closure instants from the laryngograph and use as a reference "truth." We describe how the laryngograph signal is recorded, what features can be seen in it and how the HQTx algorithm derives the instants of closure from it. We define the performance measures we use for GCI detection, based on the HQTx derived instants. Finally the the two speech corpora used for evaluating GCI detection performance are described. Both these corpora contain contemporaneous laryngographic recordings.

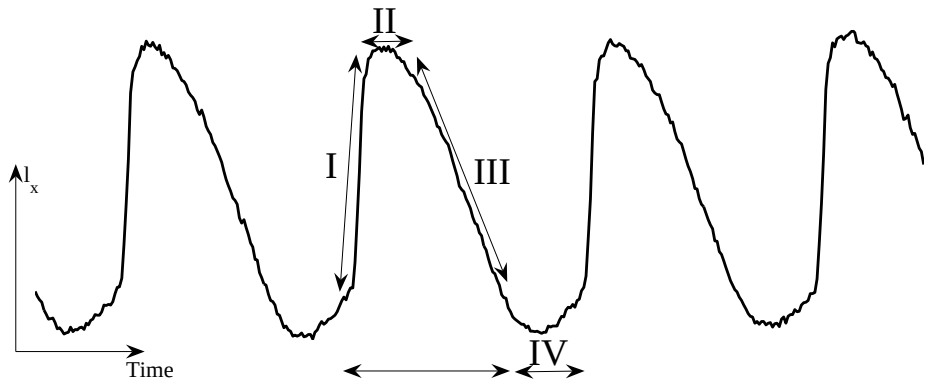

Figure 2.1: The laryngograph, l_x for voiced speech. Four features in the signal can be identified as the (I) closing phase, (II) phase of maximum contact, (III) opening phase, and (IV) open phase. The phases (I-III) are referred to as the closed phase.

2.1.1 The Laryngograph

The laryngograph is a device that measures the electrical conductance across the larynx [Abberton *et al.*, 1989]. Each electrode is placed superficially on either side of the neck next to the larynx and held in place by an elastic band. A constant AC voltage is applied across the larynx and the varying conductance is measured during speech. We refer to the demodulated conductance signal, l_x, as the laryngograph. During voiced speech, the vocal folds vibrate and the laryngograph becomes quasi-periodic according to the vocal fold frequency.

Few periods of the laryngograph are shown in Figure 2.1. Four distinct features in the signal correspond to the closing phase, maximum contact, opening phase and open phase of the vocal fold vibration. This has been verified using high-speed larynx photography and x-ray flashing imaging [Abberton *et al.*, 1989].

We use glottal closure instants determined from the laryngograph as the reference "truth" when evaluating the performance of methods that rely only on the speech signal itself. Glottal closure and opening can be reliably observed in the laryngograph signal [Scherer *et al.*, 1995]. The HQTx algorithm identifies GCIs from

Measure	No. of GCIs
Miss	0
Identification	1
Detection	1 or 2
False alarm	> 1

the laryngograph signal using the following definitions which we adopt in this work [Huckvale, 2000]. The starting points of glottal closure and opening are defined respectively as positive-going and negative-going zero crossings in the smoothed laryngograph time-derivative. The interval between the start of closure and the start of opening is defined as a glottal pulse if its duration and the amplitude of the laryngograph within the interval are within defined limits. A GCI is defined to occur at the maximum of the smoothed laryngograph time-derivative during a glottal pulse [Hess and Indefrey, 1984].

2.1.2 Glottal Closure Instants Detection Evaluation

We define the *larynx cycle* as the range of samples n such that, given a reference GCI at sample $\breve{n}_r$, with preceding and following reference GCIs at samples $\breve{n}_{r-1}$ and $\breve{n}_{r+1}$ respectively, $\frac{1}{2}(\breve{n}_{r-1} + \breve{n}_r) \leq n < \frac{1}{2}(\breve{n}_r + \breve{n}_{r+1})$. Table 2.1.2 defines the performance measures we use to assess GCI detection algorithms, with reference to Figure 2.2. They are *miss rate*, which is the ratio of larynx cycles where no GCI is identified, *identification rate*, which is the ratio of larynx cycles where only one GCI is identified, *detection rate*, which is the ratio of larynx cycles where either one or two GCIs are identified and *false alarm rate*, which is the ratio of larynx cycles where more than one GCI is identified.

Furthermore, we define the *identification error*, ζ, as the timing error between the reference GCI and the identified GCI in larynx cycles for which exactly one GCI has been identified and similarly the *detection error* as the timing error between

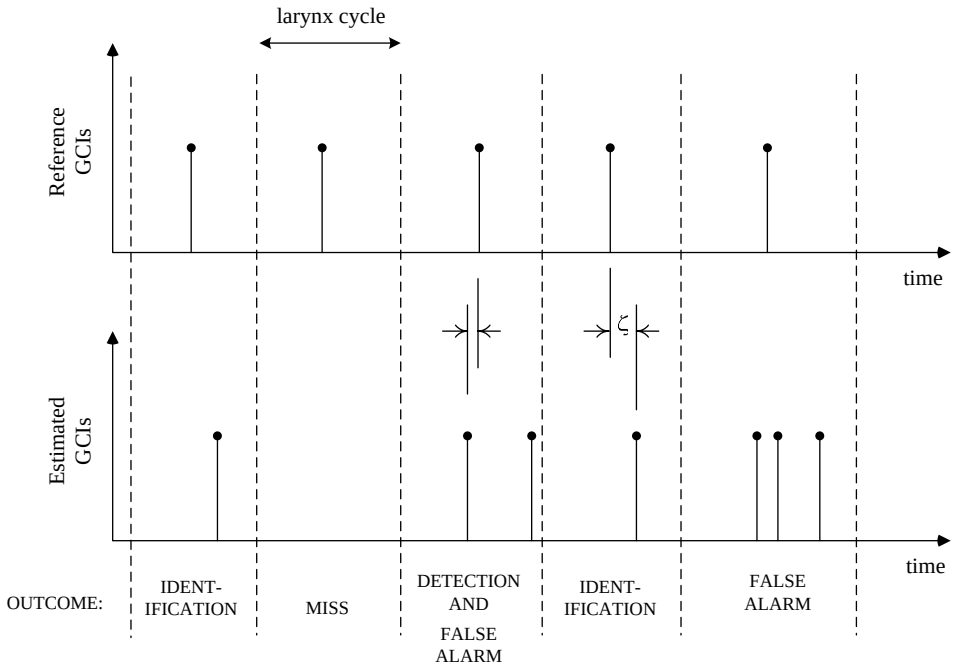

Figure 2.2: Characterisation of GCI estimates showing four larynx cycles with examples of each possible outcome from GCI estimation. Identification accuracy is measured by ζ.

the reference GCI and the detected GCI in larynx cycles for which there are either one or two GCIs detected. In those cycles for which exactly two GCIs are detected, the smaller timing error is used. The identification- and detection *accuracy* are the standard deviations of the respective errors produced in each experiment.

2.1.3 Speech Corpora

We used two spoken language corpora to assess glottal instant detection methods. They include a laryngograph channel which provides a direct measurement of glottal activity [Krishnamurthy and Childers, 1986; Abberton *et al.*, 1989] so the instants of glottal closure can be determined using the HQTx program from the Speech Filing System software suite [Huckvale, 2000]. Both corpora were recorded anechoically at a sample rate of 20 kHz. A headrest was used to keep the lip-to-microphone as constant as possible. This allowed us to use a constant time-alignment of the

laryngograph with the speech signal. Requiring a contemporaneous laryngographic recordings with the speech restricted the choice of speech databases. The corpora were chosen because they were recorded under near ideal conditions and therefore the design and test procedure of the GCI detectors are based on speech signal characteristics rather than on artifacts introduced by the recording process.

APLAWD

The APLAWD corpus [Lindsey *et al.*, 1987] includes ten repetitions from each of ten British English speakers (five male, five female) of the following sentences

S1: "George made the girl measure a good blue vase"

S2: "Why are you early you owl?"

S3: "Cathy hears a voice amongst SPAR's data"

S4: "Be sure to fetch a file and send their's off to Hove"

S5: "Six plus three equals nine"

for a total of 500 utterances. Ten of the utterances contained recording errors and, after excluding voiced segments with fewer than five cycles, the remaining 490 utterances contained 129537 glottal closures whose time were delayed by 1 ms to provide a first order correction for the larynx-to-microphone delay. The APPLAWD database contains reference square-wave recordings. These can be used to reduce the effect of phase distortion introduced in the recording process and we discuss this in Section 3.7.

The DYPSA algorithm, which is presented in Chapter 5 required an optimisation of cost function weights. We found that the performance was not sensitive to the precise values of these weights but to derive the optimum weights we designated the initial utterance of the first two sentences from every speaker as a training

utterance. We did not use these utterances for assessing the performance of the algorithms.

SAM

The Speech Assessment Methods (SAM) is a preparation phase of the European Union information technology programme, Esprit [Chan *et al.*, 1995]. As a part of this project, five language spoken corpus was designed consisting of digits and passages. Laryngograph signals were also recorded simultaneously. We use, in this work, the English subset of this initial corpus, which has now been expanded to include more languages and more speakers for each language. Our subset was recorded at 16 kHz sampling rate and contains 4 British English speakers (two female and two male) each reading a continuous passage of approximately 2 minutes. After excluding voiced segments with fewer than five cycles, the number of glottal closure instants in the corpus was 42005.

2.2 Speaker Identification Evaluation

2.2.1 Performance measures

The measure of performance we use for a speaker identifier is obtained by counting the number of misclassifications made by the system. Statistical experimental design allows us to estimate the probability of misclassification from the error count of the sample of speakers [Papoulis, 1991]. The quality of the assessment depends on the size of the sample and the number of errors. In assessing the speaker identification performance, we followed the recommendations of Expert Advisory Group on Language Engineering Standards (EAGLES), published in the *Handbook of Standards and Resources for Spoken Language Systems* [Gibbon *et al.*, 1998]. In closed-set

identification, an estimated speaker index $\hat{\imath}$, is assigned to an utterance C_i spoken by speaker i. Since all utterances belong to one of the registered speakers, a *misclassification error* occurs for the t-th test utterance produced by speaker i when

$$\hat{\imath}(C_i^t) \neq i \qquad \text{or if} \qquad \delta(\hat{\imath}(C_i^t) - i) = 0 \tag{2.1}$$

where $\delta(\cdot)$ is the delta function. *Misclassification rate* for speaker i is defined as

$$\gamma_i = 1 - \frac{1}{T_i} \sum_{t=1}^{T_i} \delta(\hat{\imath}(C_i^t) - i) \tag{2.2}$$

where T_i is the number of utterances from speaker i and γ_i is an estimate of the probability of the classifier choosing a different speaker from i. The *average misclassification rate* is defined as

$$\bar{\gamma}_i = \frac{1}{N_s} \sum_{i=1}^{N_s} \gamma_i \tag{2.3}$$

where N_s is the number of speakers in the set, and we assume that there is a test utterance for every speaker in the set. The *test set misclassification rate* is defined as

$$\gamma = 1 - \frac{1}{T} \sum_{i=1}^{N_s} \sum_{k=1}^{T_i} \delta(\hat{\imath}^k - 1) \tag{2.4}$$

where $T = \sum_i T_i$ is the total number of test utterances in the entire set. We note that $\bar{\gamma} = \gamma$ if $T_{i_1} = T_{i_2} \ \ \forall i_1, i_2$ which is the case in our experiments and we therefore only report γ.

A *gender-balanced misclassification rate* is derived by calculating the average misclassification rate separately for males and females and averaging

$$\bar{\gamma}_{GB} = \frac{1}{2}(\bar{\gamma}_M + \bar{\gamma}_F). \tag{2.5}$$

Speaker recognition performance varies with speaker's gender and we therefore report this measure as is recommended [Gibbon *et al.*, 1998].

The standard error is a measure of how well the test-set misclassification rate, γ, is estimated and quantifies how far the estimate is likely to be from its expected value. If we let $\mathcal{X}_\gamma$ be the random variable in question, then the standard error is,

$$\begin{aligned} \sigma_\gamma &= \sqrt{\mathcal{E}\{(\mathcal{X}_\gamma - \mathcal{E}\{\mathcal{X}_\gamma\})^2\}} \\ &\simeq \frac{\sqrt{K\gamma(1-\gamma)}}{K}. \end{aligned} \tag{2.6}$$

In Section 6.4, we present all results in the form of $\gamma \pm \sigma_\gamma$ to indicate the degree of uncertainty of the error assessment.

2.2.2 What performance to expect

Various factors dictate the performance of speaker identification systems. These include the amount of training data, the length of the test speech utterance, the mismatch between the training and the test data and the number of speakers in the set. Text-dependent speaker recognition experiments with clean speech from the YOHO database of 138 speakers have achieved between 0.20 and 0.65% gender balanced misclassification rate using 1-3 mixture hidden Markov model [Che *et al.*, 1996]. The same experiments using Gaussian mixture models for text-independent speaker identification achieved between 3.55 and 16.1% misclassification rate depending on how much training data and how many feature coefficients were used [Wildermoth and Paliwal, 2003]. The Switchboard database contains noisy conversational telephone speech from 113 speakers. Text-independent speaker recognition experiments on the Switchboard database using Gaussian mixture models were reported to be 17.2% [Reynolds, 1995]. The performance of text-independent GMM speaker identifier using the 630 speaker TIMIT database has been reported to be

0.97% misclassification rate and the equivalent experiment on the noisy NTIMIT database achieved 55.6% misclassification rate [Wildermoth and Paliwal, 2003].

For verification the equal error rate is when the false rejections rate equals the false acceptance rate. For the state-of-the-art recognisers the equal error rate can be expected to be around 0.1% for text-dependent clean speech verification, 2% for text-dependent verification of 10-digit string telephone speech, 10% for text-independent verification of conversational telephone speech and 25% for text-independent verification of noisy, military radio speech all tested in multiple-session experiments [Reynolds, 2002].

2.2.3 TIMIT

The TIMIT database contains utterances from 630 speakers [Fisher *et al.*, 1986]. There are 10 utterances from each speaker of an average duration of 2.5s. The use of the TIMIT data base makes it possible to test the speaker identifier under near ideal conditions. The speech from each speaker is recorded in one session at 16 kHz sampling rate, with negligible noise, distortion nor session variability. This means that any occurred error is due to overlapping speaker distributions which is a function of the applied feature set and modelling technique. A common choice for speaker identification evaluations is to use the first eight sentences of each speaker for training and the last two for testing [Reynolds, 1995]. This gives 5040 training and 1260 test utterances.

The TIMIT database was originally chosen because of its phonetic labelling. To begin with, we used the phonetic labelling to aid the feature extraction and prove the concept but later improved the process so that the labelling was no longer needed. It was decided to continue using the TIMIT database because it was one of few databases that was already available and there were other speaker identification

results reported on this database which were deemed to be very relevant to this work [Reynolds, 1995; Plumpe *et al.*, 1999]. The disadvantage of this experimental setup is that it produces very few errors and therefore a poor estimation of misclassification. We therefore made the task harder and divided each test utterance into eight equal parts which we tested on separately. This can be seen by the standard error of the estimation of the probability of misclassification which we report. Furthermore, dividing the test utterances makes us able to evaluate the performance as a function of test utterance duration.

Chapter 3

Voice Source Analysis

THE estimation of the voice source signal is addressed in this chapter. We will give an overview of the background material relevant to the study of the voice source signal and in Chapter 6 we will use voice source features for speaker identification. We will make the assumption that the voice source is linearly separable from the vocal tract and the interaction between the sub- and supraglottal systems is not modelled. More complex models of the voice source assume such interaction [Cummings and Clements, 1995] and have become popular where the objective is to synthesise the speech time waveform accurately but this can be computationally expensive [Ananthapadmanabha and Fant, 1982; Titze, 1984].

For voiced speech, the voice source is often modelled as the output of a filter with two real poles close to unity and an impulse train as an input [Rabiner and Schafer, 1978]. More accurate models have been suggested which approximate the glottal pulse with piecewise continuous functions and the estimation of the vocal tract can take this shape into account [Rosenberg, 1971; Fant *et al.*, 1985]. If the vocal tract model is known then the voice source signal can be obtained by inverse filtering. The vocal tract estimation can be achieved without knowing the exact

shape of the voice source pulse when applying closed phase analysis, e.g. [Wong *et al.*, 1979; Chan and Brookes, 1989]. The parameters of the vocal tract model are derived when the vocal folds are closed and the voice source signal is assumed to be zero [Atal and Hanauer, 1971].

This chapter is organised as follows. We give a brief review of the vocal folds physiology in Section 3.1 before we formulate inverse filtering in Section 3.2 and describe the two-pole model of the glottal pulse in Section 3.3. The Rosenberg- and the Liljencrants-Fant models of the glottal pulse are given Section 3.4 and methods for simultaneous estimation of the vocal tract and the glottal pulse are reviewed in Section 3.5. Finally we describe closed phase analysis in Section 3.6 and conclude the chapter with discussion in Section 3.7.

3.1 Vocal Folds Physiology

Voiced speech is generated by a normally steady flow of air from the lungs, through the trachea, and into the glottis. It is interrupted by the vocal folds which vibrate in a self-oscillating mode producing quasi-periodic pulses of air that excite the vocal tract. The period of oscillation is therefore dictated by the flow of air from the lungs and the physiological properties of the vocal folds and not by any explicit neurological excitation signal [Cummings and Clements, 1995]. The major excitation of the vocal tract is associated with the rapid closure of the vocal folds and the slope of closure increases with the vocal effort [Miller, 1959].

Figure 3.1 shows a segment of voiced speech, its corresponding laryngograph, its estimated glottal volume velocity and its glottal flow derivative signal. The laryngograph has delayed by 1 ms so that it is time-aligned with the speech waveform. Also, the time delay in the inverse filtering has been ignored so that the voice source signals are time aligned with the speech signal. We can identify the larynx cycle

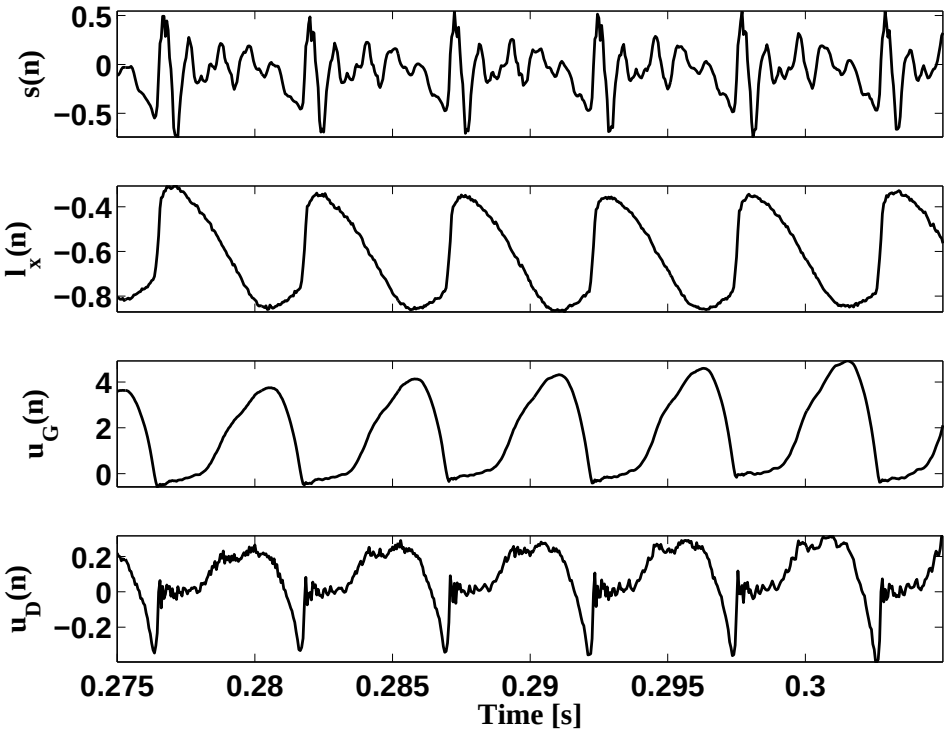

Figure 3.1: The first strip shows a small portion of a speech signal from the APLAWD database. Speaker 4 is saying: "Why are you early you owl?" and the portion is from the phone /ay/ in "Why". The second strip shows the laryngograph, the third shows the glottal volume velocity and the fourth shows the glottal flow derivative.

in the speech signal beginning with an excitation followed by exponentially damped sinusoids. The excitation can also be seen in the laryngograph as a sudden rise to the maximum value and in the graph of the glottal flow derivative $u_D(n)$ it is shown as a sudden return to zero. The graphs of the glottal volume velocity $u_G(n)$ and the glottal flow derivative $u_D(n)$ show the closed phase of each larynx cycle as the signals being close to zero. The open phase is shown as a low value in the laryngograph and as an increase in glottal volume velocity and glottal flow derivative. The graphs also demonstrate the difference between the abrupt closure and more gentle opening instants of the vocal folds.

Studies using video recordings of the vocal folds during voiced phonation also show that they open slowly until they close rapidly at the glottal closure instant

(GCI) and then they remain closed during the closed phase at the end of which they start opening again [Miller, 1959]. Parameters of the vocal fold movements such as the glottal width and area have been measured using photoglottography [Childers *et al.*, 1980] and modelled together with with lung pressure [Ananthapadmanabha and Fant, 1982]. Electroglottography has also been used to study the vocal folds [Abberton *et al.*, 1989; Scherer *et al.*, 1995; Childers, 1992]. The laryngograph in Figure 3.1 shows how the conductance changes over a larynx cycle as we explained in Chapter 2. Strong relationship between the laryngograph, the glottal area function and the glottal volume velocity function was observed [Scherer *et al.*, 1988]. In particular, the study supports the the notion of a linear relationship between the value of the laryngograph and the vocal fold contact area.

3.2 Inverse Filtering

If the vocal tract filter $V(z)$ has been estimated as $\hat{V}(z)$, the glottal volume velocity signal $U_G(z)$, can be estimated as $\hat{U}_G(z)$ from the speech signal using inverse filtering [Miller, 1959; Fant, 1960; Markel, 1973]. If $\hat{V}(z)$ is an all-pole filter, the inverse filter $\hat{V}^{-1}(z)$ is an FIR filter and we get,

$$\hat{U}_G(z)R(z) = \frac{S(z)}{\hat{V}(z)} \tag{3.1}$$

where the acoustic delay of the vocal tract, $z^{-P/2}$, and the acoustic delay from the lips to the microphone has been ignored, so that $s(m)$ and $\hat{u}_G(m)$ are time-aligned. If this is not desired, the delay from the voice source signal to the speech signal has to be determined and corrected for. $\hat{U}_G(z)$ is an estimate of the glottal volume velocity waveform represented as $U_G(z)$ in the z-domain.

The inverse of the lip radiation transfer function, $R(z)$, also needs to be applied to obtain the glottal volume velocity signal. In practice, we can not apply a

true integrator $R^{-1}(z)$ since the initial conditions for the integrator are not known, so a leaky integrator with the pole placed near, but not at, zero frequency has to be applied [Rothenberg, 1973]. We therefore use,

$$\frac{1}{\hat{R}(z)} = \frac{K_R^{-1}}{1-(1-\epsilon)z^{-1}} \tag{3.2}$$

where ϵ is small. The glottal volume velocity can therefore be estimated as

$$\hat{U}_G(z) = \frac{S(z)}{\hat{R}(z)\hat{V}(z)}. \tag{3.3}$$

In the time domain the glottal flow derivative is derived by,

$$\hat{u}_D(m) = -\frac{1}{K_V}\sum_{p=0}^{P} a_p s(m-p) \tag{3.4}$$

and

$$\hat{u}_G(m) = \frac{1}{K_R}\hat{u}_D(m) + (1-\epsilon)\hat{u}_G(m-1). \tag{3.5}$$

Figure 3.2 shows a voiced speech segment $s(n)$, the estimated glottal volume velocity $\hat{u}_G(n)$ and the estimated glottal flow derivative $\hat{u}_D(n)$ determined using inverse filtering. It also shows $\hat{u}(n)$ which is the preemphasised $\hat{u}_D(n)$, resembling a periodic impulse train. The glottal volume velocity was estimated by deriving vocal tract filter parameters over a 30 ms frames repeating every 10 ms. Inverse filtering was applied on each frame using a linear interpolation of the filter parameters across frames. The closed phase segments can be seen in both the voice source signal and its derivative as the constant, close-to-zero portions of the larynx cycle. The slow rise in the glottal volume velocity signal during closure is due to the leaky integrator used when applying the inverse of the lip transfer function $\hat{R}(z)$. The abrupt closure is represented as a fast return to zero in the volume velocity signal and a negative peak in its derivative.

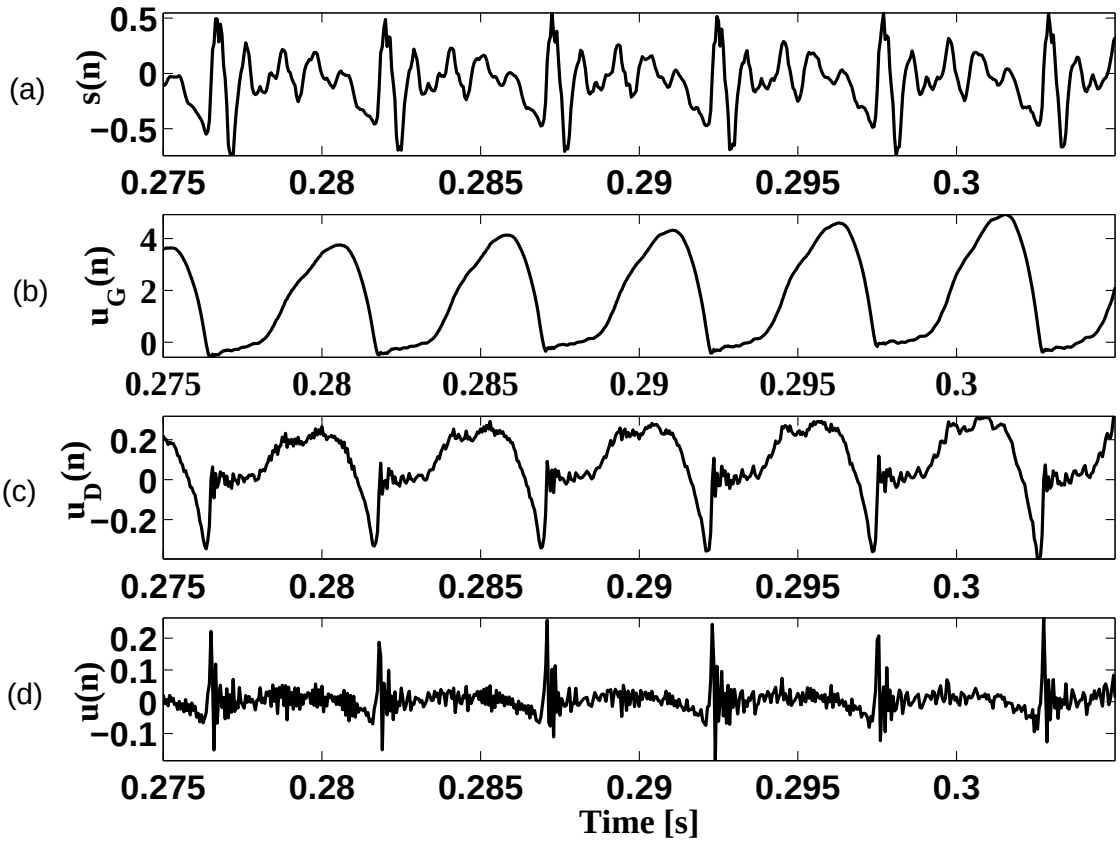

Figure 3.2: (a) Voiced segment of speech $s(n)$, (b) the glottal volume velocity $\hat{u}_G(n)$, (c) the filtered flow velocity $\hat{u}_D(n)$ and (d) the preemphasised flow velocity $\hat{u}(n)$, resembling an impulse train.

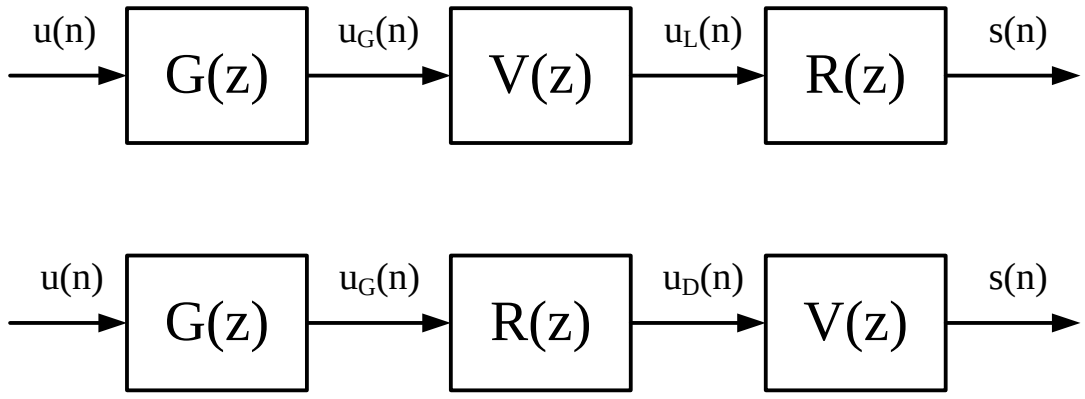

Figure 3.3: The vocal tract and glottal pulse model for voiced speech.

3.3 Two-Pole Model of Glottal Flow

Instead of deriving the voice source signal through inverse filtering, the voice source can be approximated by a glottal pulse model $G(z)$. This filter is driven by an impulse train with periods that correspond to the pitch frequency and the pulse is formed as the output of $G(z)$. The linear voice production model for voiced speech is therefore changed to that presented in Figure 3.3. The transfer function of the glottal pulse model is often assumed to have only two positive real poles close to unity and no zeros [Atal and Hanauer, 1971; Markel and Gray, 1976],

$$G(z) = \frac{K_G}{(1 - z_1 z^{-1})(1 - z_2 z^{-1})} \tag{3.6}$$

with K_G as a gain related to the amplitude of the glottal flow and z_1 and z_2 are the poles. An example of a pulse generated by such filter can be seen in Figure 3.4(a), where the poles are at $z_1 = 0.98$ and $z_2 = 0.95$ (Figure 3.4(c)). The joint transfer function of the lip radiation and the glottal pulse can thus be approximated by a single pole,

$$G(z)R(z) = \frac{K_G K_R (1 - z^{-1})}{(1 - z_1 z^{-1})(1 - z_2 z^{-1})} \approx \frac{K_G K_R}{(1 - z_2 z^{-1})}. \tag{3.7}$$

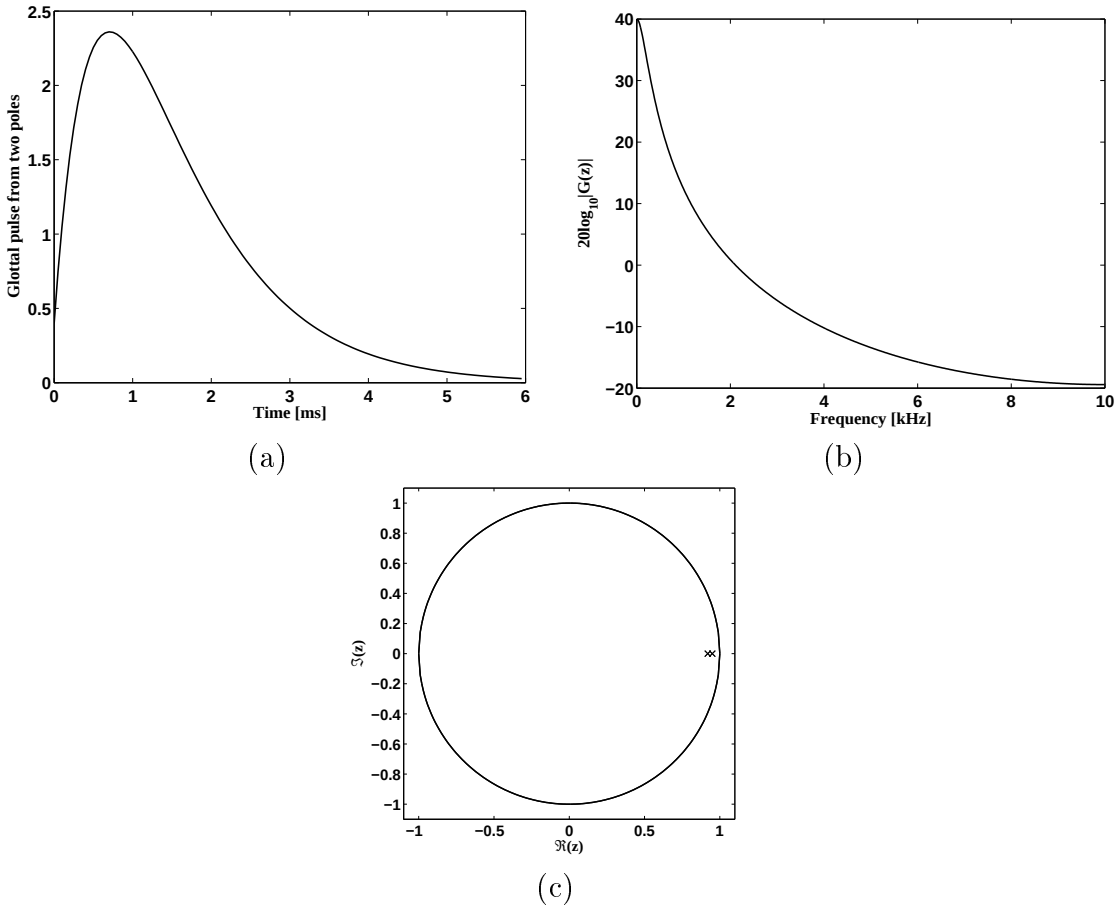

Figure 3.4: (a) The impulse response of the a glottal pulse filter with poles at $z_1 = 0.98$ and $z_2 = 0.95$, (b) its frequency response and (c) the pole plot of the filter $G(z)$.

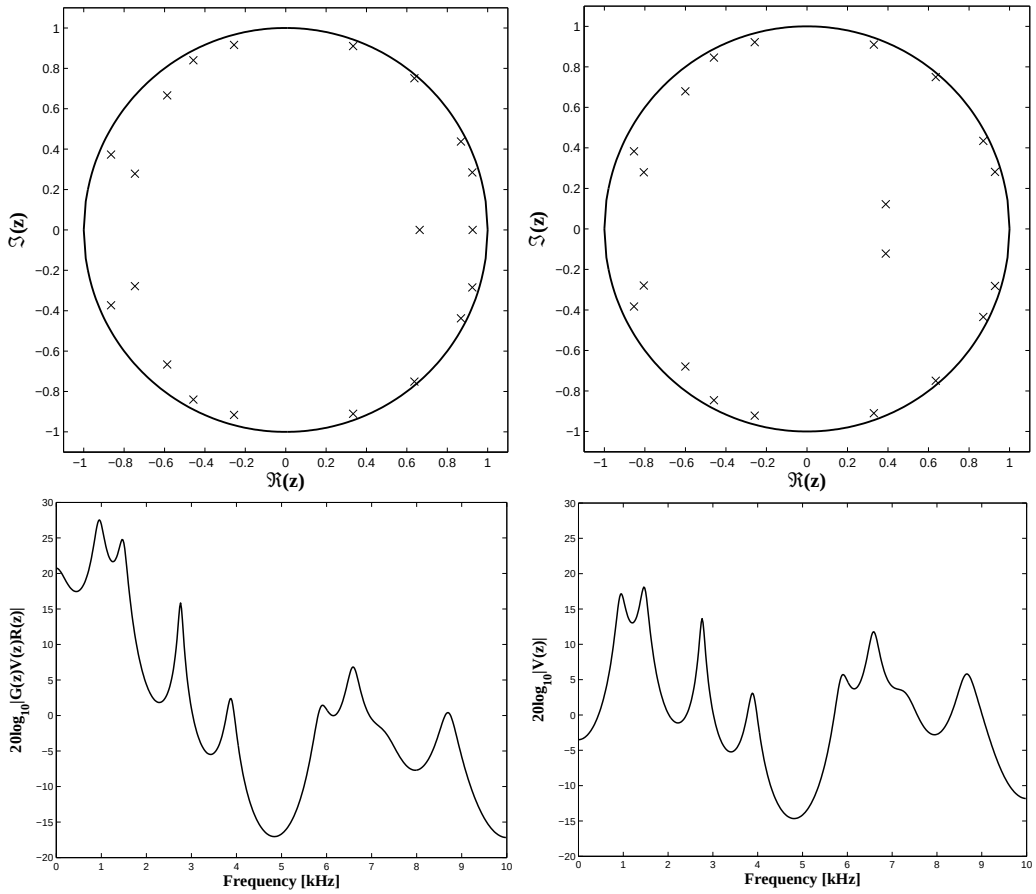

Figure 3.5: The poles and the frequency response of the linear prediction coefficients derived directly on the speech waveform (left) and on the pre-emphasised speech (right).

The remaining pole can be removed using preemphasis, which is a finite impulse response (FIR) filtering of the speech with a single zero corresponding to the pole. Alternatively, linear prediction can be applied to estimate the joint transfer function of the glottal pulse, the vocal tract and the lip radiation, $G(z)V(z)R(z)$ using an extra parameter to represent the extra pole. Figure 3.5 shows the pole-plots and the frequency responses for either case. We can see the poles on the positive real axis for $G(z)V(z)R(z)$ and the corresponding spectral tilt in the frequency response. There are no real poles for the pre-emphasised speech and spectral tilt has been removed. If we compare the estimated glottal volume velocity in Figure 3.2 with the impulse response of the two pole model of $G(z)$ shown in Figure 3.4 we can see that the

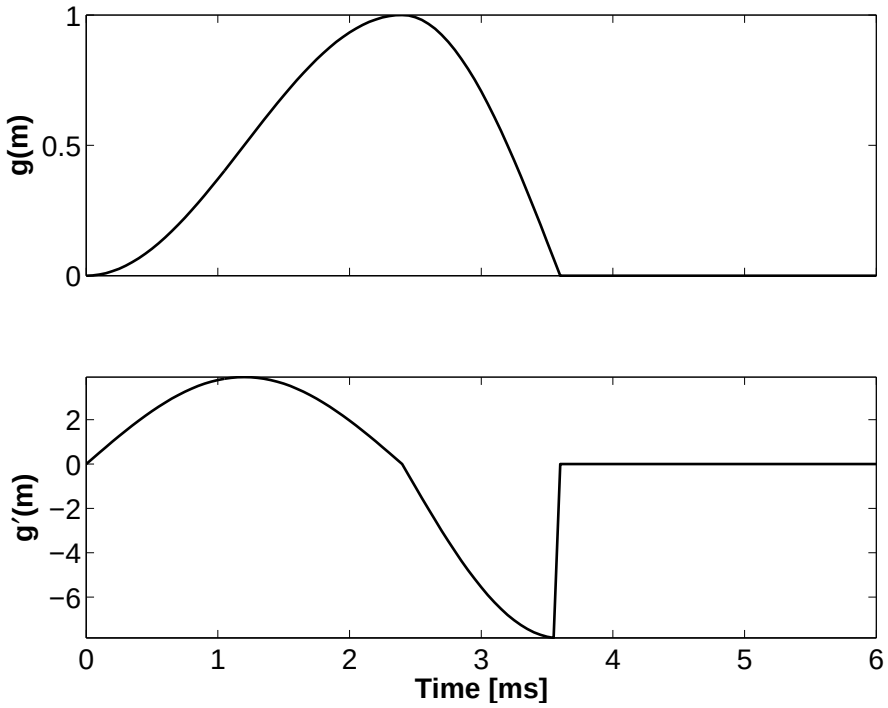

Figure 3.6: The glottal volume velocity $g(m)$ and its derivative for the trigonometric Rosenberg model $g\prime(m)$.

two-pole model does not resemble the glottal volume velocity pulse shape very well. This has led to the voice source being represented by all-pole filters of higher order [Akande and Murphy, 2005].

3.4 Parametric Modelling of Glottal Flow

Another way of characterising the glottal flow is to parameterise it so that within one larynx cycle it is represented by a piece-wise continuous functions [Cummings and Clements, 1995]. An early contribution to this approach was made where the quality of synthetic speech was tested using parametric approximations to the glottal flow. Rosenberg presented six pulse shape approximations varying from a triangle to a piecewise linear trapezoidal pulse with the other models comprised of either polynomials or trigonometric functions [Rosenberg and Sambur, 1975]. The three-

segment trigonometric model is often referred to as the Rosenberg model,

$$g(m) = \begin{cases} \frac{1}{2}(1 - \cos(\pi m / M_1)), & \text{for } 0 \leq m \leq M_1, \\ \cos(\frac{\pi}{2}(m - M_1)/M_2), & \text{for } M_1 \leq m \leq M_1 + M_2, \\ 0, & M_1 + M_2 < m \leq M - 1. \end{cases} \tag{3.8}$$

Where the parameters M_1 and M_2 determine the model shape. The glottal flow of the Rosenberg model and its derivative are shown in Figure 3.6.

Another commonly used model is the two-segment Liljencrants-Fant (LF-) model of the glottal flow derivative $g\prime(m)$. This model was developed in a similar fashion to that of Rosenberg's by piecewise fitting of trigonometric functions [Fant, 1979; Fant and Liljencrants, 1979] and was later refined to add an exponential recovery phase at the instant of closure [Fant *et al.*, 1985],

$$g\prime(m) = \begin{cases} \mathcal{A}_1 e^{\alpha_1 m} \sin(\pi m / M_1) & \text{for } m \leq M_2 - 1, \\ \mathcal{A}_2 \left[1 - e^{-\alpha_2 (m - M)}\right] & \text{for } M_2 \leq m \leq M - 1. \end{cases} \tag{3.9}$$

The seven parameters, $\mathcal{A}_1$, $\mathcal{A}_2$, α_1, α_2, M_1, M_2, and M determine the model and are constrained by requiring $g\prime(t)$ to be continuous at M_2 and sums to zero on the interval zero to $M - 1$ to ensure that $u_G(0) = u_G(M - 1)$. The glottal volume velocity and its derivative for the LF model are shown in Figure 3.7.

Many other models of the glottal flow rely on more parameters [Fujisaki and Ljungqvist, 1987; Klatt and Klatt, 1990; Brookes and Chan, 1994; Alku and Backstrom, 2002; Backstrom *et al.*, 2002]. The parameters of these models can be determined using least-square fits to the inverse filtered speech [Strik *et al.*, 1993; Brookes and Chan, 1994].

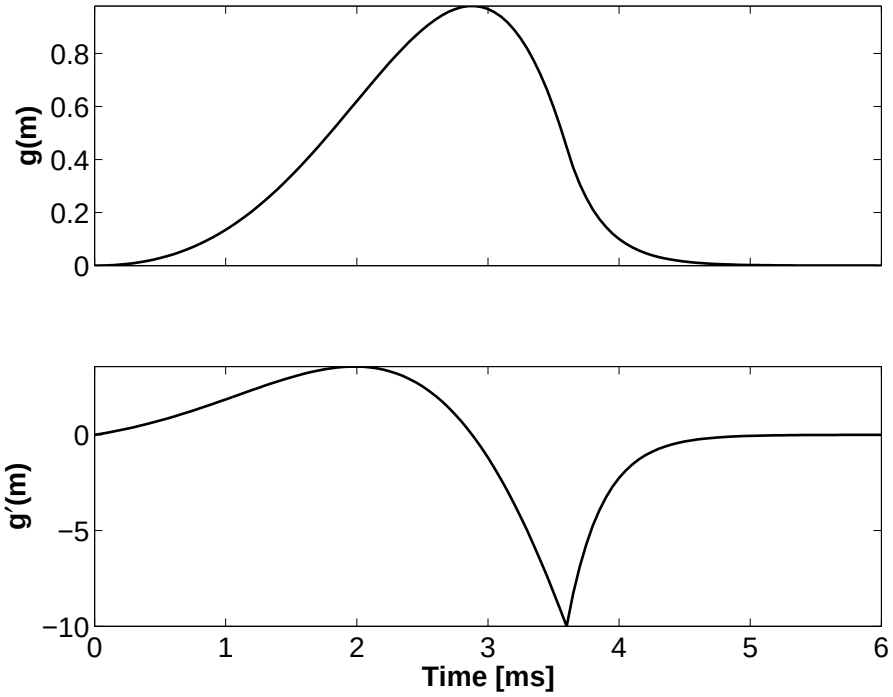

Figure 3.7: The glottal flow velocity $g(m)$ and its derivative for the LF model $g\prime(m)$.

3.5 Simultaneous Estimation of Tract and Source

We assume that speech is generated by the linear vocal tract filter by the excitation of the voice source signal. To determine the system we need to perform a blind estimation, since neither the voice source signal nor the vocal tract filter is known beforehand. But since we can assume a lot about the vocal tract filter as we did in Chapter 1 and the signal shape, as we did in last section, the task becomes tractable. A simple approach is to characterise the source as a filter of two real poles as we have already discussed but here we give a brief overview of methods which rely on parametric models of the glottal flow.

One formulation of simultaneous estimation of the vocal tract and the voice requires the minimisation of two error signals with respect to the parameters that describe the vocal tract and the voice source [Milenkovic, 1986]. The first is the

difference between the parameterised voice source, $G(z)$, and estimated voice source $\hat{U}_G(z)$, both modulated by $R(z)$, i.e.,

$$E_1(z) = G(z)R(z) - \hat{U}_G(z)R(z) = G(z)R(z) - \frac{S(z)}{\hat{V}(z)}. \tag{3.10}$$

The second is the difference between the speech signal and the synthesised speech signal,

$$E_2(z) = S(z) - \hat{S}(z) = S(z) - G(z)\hat{V}(z)\hat{R}(z). \tag{3.11}$$

The parameters in question are those of the the vocal tract all-pole filter $\hat{V}(z)$ and the parameters of the glottal pulse model $G(z)$. The glottal pulse was represented as a superposition of two polynomials, each representing the excitation generated by the closure and opening of the vocal folds [Milenkovic, 1986]. The voice source parameters are the coefficients of each polynomial and the two weights of the superposition of the two waveforms, but the laryngograph was used to determine the instant of closure and opening so that the polynomials could be shifted correctly. The speech $S(z)$ was lowpass filtered before it was applied in 3.10 and 3.11 to approximate the smoother synthesised speech signal, $\hat{S}(z)$.

Another approach is to formulate an ARMA model for the speech production [Fujisaki and Ljungqvist, 1987],

$$A(z)S(z) = B(z)G(z) \tag{3.12}$$

which leads to the error

$$E(z) = S(z) - \frac{B(z)}{A(z)}U(z) \tag{3.13}$$

where $B(z)$ is the nominator and $A(z)$ is the denominator of the ARMA model and $U(z)$ denotes the input waveform[1]. This error function leads to a nonlinear

[1]This does not conform to the notation we use in this work but $A(z)$ corresponds to $V(z)$ whereas $B(z)$ and $U(z)$ correspond to a combination of $R(z)$, $G(z)$ and $U_G(z)$.

minimisation which is avoided by minimising $A(z)E(z)$ instead and the coefficients of $A(z)$ and $B(z)$ are estimated. A piece-wise continuous functions for the voice source derivative, called the Fujisaki-Ljungqvist model, was proposed for $R(z)G(z)$ and a joint optimisation of its parameters with the coefficients of $A(z)$ and $B(z)$ was developed [Fujisaki and Ljungqvist, 1987].

Similarly, an ARMA filter and the glottal flow derivative were estimated recursively using simulated annealing and using the Fujisaki-Ljungqvist model as an objective waveform [Lobo, 2001]. The glottal closure instants were re-estimated from the wavelet-smoothed excitation signal derived using weighted recursive least squares with variable forgetting factor [Childers *et al.*, 1995].

The Rosenberg-Klatt model [Klatt and Klatt, 1990] has also been used to model the voice source where the estimation of the vocal tract and the voice source parameters was done using autoregressive exogenous input (ARX) modelling [Ding and Kasuya, 1996; Zhu and Kasuya, 1996]. The optimisation of the parameters were based on simulated annealing and the adaptive Kalman filter algorithm. Alternatively, convex optimisation of the ARX parameters has been used [Lu and Smith, 1999; Fu and Murphy, 2006].

3.6 Closed Phase Analysis

The inverse filtering of speech relies on the estimate of the vocal tract filter $V(z)$. Early researchers determined the formant locations and bandwidths interactively by varying the coefficients of an inverse-filter [Miller, 1959; Lindqvist-Gauffin, 1970; Rothenberg, 1973; Hunt, 1978]. For example, two formants have been inverse filtered by using an LCR-circuit with adjustable components whose values are determined from the spectrogram of the speech and the knowledge that the glottal volume velocity is zero in the closed phase [Miller, 1959]. Automatic estimation methods

of the vocal tract filter have been developed, relying on some form of nonstationary linear prediction analysis during the closed glottis interval [Strube, 1974; Markel and Gray, 1976; Steiglitz and Dickinson, 1977; Wong *et al.*, 1979; Hedelin, 1984; Larar *et al.*, 1985; Veeneman and BeMent, 1985; Krishnamurthy and Childers, 1986; Chan and Brookes, 1989; Brookes and Loke, 1999; Akande and Murphy, 2005]. Closed phase analysis is of this nature, relying on an undriven segment of the voiced speech signal to estimate the vocal tract parameters.

The input signal is assumed to be zero during the closed phase of the speech. The laryngograph in Figure 3.1 shows the closed phase with high conductivity across the glottis. This part of the speech signal is freely oscillating and the parameters of $V(z)$ can be estimated with linear prediction analysis [Atal and Hanauer, 1971]. Vocal tract parameters have been extracted for each of the closed phases in the speech by applying the minimisation over that period, LPC is restricted to the closed phase by [Veeneman and BeMent, 1985],

$$s(m) = \sum_{p=1}^{P} a_p s(m-p) \qquad \text{for } \breve{n} \leq m < \hat{n} \tag{3.14}$$

where $\breve{n}$ and $\hat{n}$ are first samples of consecutive closed and open periods of the larynx cycle. A further development of this was needed to increase the number of speech samples used in the parameter estimation. A multi-glottal closed phase analysis was presented to include adjacent glottal cycles [Chan and Brookes, 1989]. With fixed frame period and analysis frame size, the vocal tract parameters can be derived using a slightly altered version of Equation 3.14,

$$s_n(m) = \sum_{p=1}^{P} a_p s_n(m-p) \qquad \text{for } m \in \mathcal{C}_n \tag{3.15}$$

where $\mathcal{C}_n$ is the set of all samples in a closed phase within the window excluding the

first $P-1$ samples of each phase. The closed phase covariance matrix is then,

$$\Phi_n(i,p) = \sum_{m \in \mathcal{C}_n} s_n(m) s_n(m+i-p). \tag{3.16}$$

The difficulty with this analysis is to determine the closed phases $\mathcal{C}_n$. There are various ways of estimating these time periods, for example using video recordings [Miller, 1959], the laryngograph [Veeneman and BeMent, 1985; Krishnamurthy and Childers, 1986; Chan and Brookes, 1989] or automatic estimation directly from the speech waveform [Strube, 1974; Wong *et al.*, 1979]. We will focus on using the speech signal directly for closed phase detection in Chapters 4 and 5.

3.7 Discussion

There are many practical difficulties in implementing a robust inverse filtering scheme. Some of the difficulties arise from the fact that the time-domain voice source waveform is very sensitive to any small estimation errors in the inverse-filter transfer function. For example, the voice source signal can not be used for the detection of any abnormality of the vocal folds vibrations, if there are any undetected resonances in the vocal tract transfer function. This can happen for example if the LPC closed phase analysis interval erroneously contains excitations and demonstrates the importance of the accuracy of the glottal closure instants detection.

Another problem with detecting the time-domain voice source waveform is low-frequency phase distortion. Speech recordings normally contain phase distortion due to a high-pass filter effect in the recording process. The high-pass filter normally suppresses amplitudes at very low frequencies but still distorts the phase at higher frequency. While this is not a problem for perception, the phase distortion can extend well into frequencies that are of interest to speech analysts, especially if the

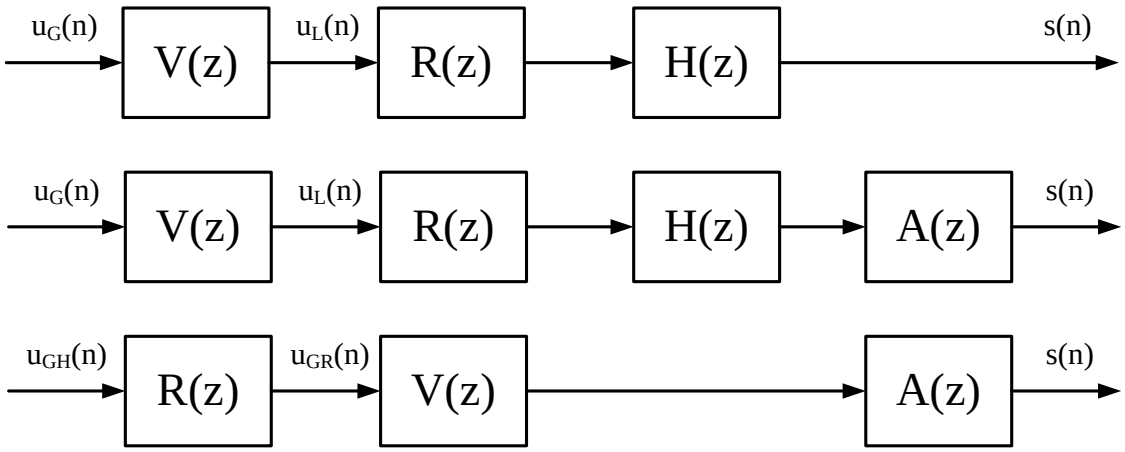

Figure 3.8: The low-pass distortion in speech is represented by $H(z)$ whose phase is corrected with the all-pass filter $A(z)$. The filter model (top) represents the model as it happens. We acknowledge the phase-distortion by denoting the glottal volume velocity as u_{GH} (middle) and the reordering of the vocal tract and the lip radiation filters are shown (bottom).

shape of the time-domain waveform is to be preserved. The phase distortion is manifested in the speech as larger delays of low frequency components compared to high frequency components. This is why such analyses either require the recording to be carefully carried out so that the phase distortion becomes negligible or the phase distortion to be corrected for.

The phase distortion has been corrected by using a second order all-pass filter, estimated from a reference square-wave recording [Holmes, 1975; Hunt, 1978]. This procedure does not affect the amplitude but shifts the low frequency phase. To represent the distortion we modify the model of the speech production by adding a high-pass filter $H(z)$ with a low cutoff frequency and an all-pass filter $A(z)$ to correct the phase distortion. This is represented in the upper system of Figure 3.8. Phase distortion presents considerable difficulties where there are no reference square-wave recordings and the all-pass filter can not be estimated.

The voice source can be modelled using a filter with only two poles, placed close to unity. The effect of one of these poles is negated by the zero in the radiation

impedance and the effect from the other pole is normally removed in LPC analysis using pre-emphasis. More elaborate approximations of the voice source use piecewise continuous functions to represent the pulse shape. An ARX vocal tract model parameter estimation can be implemented using such voice source approximations as an input signal. Alternatively, an AR vocal tract parameter estimation can be done using only the time periods when the voice source signal is assumed to be zero. We adopt this approach and concentrate on determining the glottal closure instants from the speech signal so that the closed phase analysis of speech can be implemented.

Chapter 4

Glottal Closure Instants Detection

IN voiced speech, the primary acoustic excitation normally occurs at the instant of vocal-fold closure. This marks the start of the closed-phase interval during which there is little or no airflow through the glottis. As we saw in Chapter 3, accurate identification of the closed phases allows the blind inverse filtering of the vocal tract with the use of closed phase analysis. The resultant characterisation of the glottal source gives benefits to speaker identification as we shall see in Chapter 6 but in this chapter we concentrate on the accurate identification of glottal closure instants (GCI). The identification of GCIs has also proven to be necessary in preserving coherence across segment boundaries in PSOLA-based concatenative synthesis and voice-morphing techniques [Hamon *et al.*, 1989; Stylianou, 1999].

4.1 Overview of Methods

Several algorithms have been proposed for estimating glottal closure instants from a speech waveform without the use of the laryngograph signal. For convenience we categorise these into algorithms relying on 1) short-term energy in the speech signal, 2) the predictability of an all-pole linear predictor, and 3) the negative going

zero crossings of a group delay measure of the speech or derived signals. We note however that methods that we place in one category could also belong to another given another interpretation of the method.

4.1.1 GCI from speech energy

Glottal closure instants can be detected from energy peaks in waveforms derived directly from the speech signal [Ma *et al.*, 1994; Jankowski Jr. *et al.*, 1995] or from features in its time-frequency representation [Tuan and d'Alessandro, 1999; Navarro-Mesa *et al.*, 2001]. The Frobenius norm offers a short-term energy estimate of the speech signal and, using a sliding window, this estimate gives an energy value for every speech sample. The peaks in this waveform indicate glottal closure instants. We assess this method in more detail in Section 4.2 as the FN method. The energy peaks can also be detected in a time-frequency representation of the speech signal. The wavelet transform has been used to represent the speech and detect glottal closure instants [Tuan and d'Alessandro, 1999]. Lines of amplitude maxima in the time-frequency plane were identified and the GCIs were determined to correspond to the line carrying the maximum accumulated amplitude within each pitch period. Alternatively, a Cohen's class time-frequency representation of speech was constructed and used to detect GCIs [Navarro-Mesa *et al.*, 2001]. They were detected as peaks in a spectral density correlator derived from the time-frequency representation.

4.1.2 GCI from linear prediction

Many approaches detect discontinuities in a linear model of the speech production. An early approach used a predictability measure to detect GCIs by finding the minimum of the Gram determinant of the auto-covariance matrix of the speech

signal [Strube, 1974]. This method, however, does not work well for some vowel sounds, particularly when many pulses occur in the prediction residual around the instant of closure. Furthermore it is quite computationally expensive.

GCIs have been detected using discontinuities in the derivative of the glottal air flow [Ananthapadmanabha and Yegnanarayana, 1975; Ananthapadmanabha and Yegnanarayana, 1979]. The drawback of this method is that noise can cause similar discontinuities to those caused by excitations of the voice production and this confuses the detection. Similarly, work on energy flow in the lossless-tube model has suggested that the signal representing acoustic input power at the glottis could be used to determine the instants of glottal closure and opening [Brookes and Loke, 1999].

GCIs were detected at the minimum of the total energy in the LPC residual derived over a sliding window [Wong *et al.*, 1979]. We assess this method further in Section 4.2 as the LPCR method. The formant modulation has been shown to be slower in the closed phase than in the open phase [Ananthapadmanabha and Fant, 1982] and this was used to enhance the LPCR method [Plumpe *et al.*, 1999].

The difficulty with using the LPC residual is that it often contains resonances as the derived filter does not fully predict the lower formant frequencies. The excitation energy peaks become less prominent in the residual and it becomes harder to detect GCIs. A maximum likelihood estimate of GCIs was proposed using the Hilbert transformed LPC residual signal [Cheng and O'Shaughnessy, 1989]. The maximum likelihood takes periodicity into account and the Hilbert transform filters out harmonic components caused by formants still present in the residual.

Kalman filtering has been applied to detect closed phases in voiced speech [McKenna, 2001]. The boundary of the closed phase, i.e. the GCI and GOI are detected using the log determinant of the error covariance matrix of the Kalman

filter. This measure assesses the predictability of the speech signal and is able to detect GCIs well but the timing accuracy is not good.

4.1.3 GCI from group delay measures

A group delay function can be evaluated for either the speech signal or the LPC residual to detect GCIs [Smits and Yegnanarayana, 1995; Yegnanarayana and Smits, 1995; Murthy and Yegnanarayana, 1999]. We assess this approach in more detail in the next section. Furthermore, we give a detailed analysis of group delay measures in Section 4.3 and base our contribution to GCI detection on the group delay function in Chapter 5.

4.2 Selected Methods

We have implemented three GCI detection methods to evaluate their accuracy and see which one was the most suitable for closed phase analysis. We chose to implement the LPC residual method (LPCR) [Wong *et al.*, 1979], the Frobenius Norm method (FN) [Ma *et al.*, 1994] and the Group Delay method (GD) [Smits and Yegnanarayana, 1995] which we describe here. Figure 4.1 shows the speech, the laryngograph, the LPC residual and the signals from each of the methods used to detect the closure instants. The detected instants are shown as circles in the graphs, with the circles in the speech and laryngograph graphs representing the closures detected by the laryngograph.

4.2.1 LPCR

The LPCR method calculates the normalised total squared error which is the energy in the LPC residual divided by the speech signal energy over a sliding window of

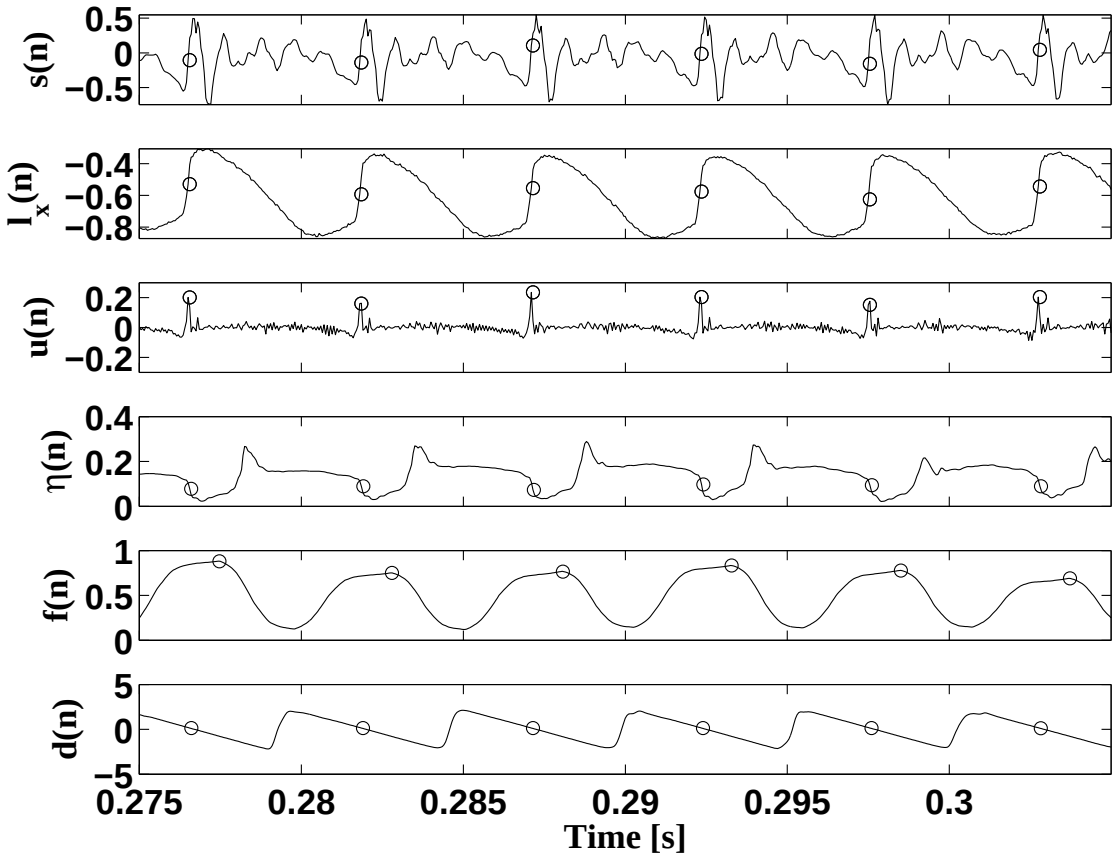

Figure 4.1: Glottal closure instants detected using the laryngograph, $l_x(n)$ with the HQTx algorithm or using the speech signal $s(n)$ directly using the normalised total squared error $\eta(n)$ which is derived from the speech and the LPC residual $u(n)$, the Frobenius norm $f(n)$, and the average group delay function $d(n)$.

3.75 ms (as recommended in [Wong *et al.*, 1979]). The windowed residual at sample n is denoted as $u_n(m)$ and the windowed speech as $s_n(m)$, so the normalised total squared error is

$$\eta(n) = \frac{\sum_m u_n^2(m)}{\sum_m s_n^2(m)} \tag{4.1}$$

calculated for every time sample n and m is the window index whose last sample corresponds to n. A low value of $\eta(n)$ indicates that the speech fits the LPC model.

The fourth trace of Figure 4.1 shows $\eta(n)$ during five larynx cycle of voiced speech. The glottal closure instant is identified in the larynx cycle as the beginning

of the period over which $\eta(n)$ stays small. The closed phase follows and the opening instant can be detected as the end of this period.

4.2.2 FN

The FN method detects glottal closure instants using the Frobenius norm of the speech data matrix defined as,

$$S = \begin{bmatrix} s(M_f+1) & s(M_f) & s(M_f-1) & \cdots & s(1) \\ s(M_f+2) & s(M_f+1) & s(M_f) & \cdots & s(2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s(M_f+M_F) & s(M_f+M_F-1) & s(M_f+M_F-2) & \cdots & s(M_F) \end{bmatrix}. \quad (4.2)$$

We simplify the derivation of the Frobenius norm method by noting that the energy of the Frobenius norm of the data matrix, formed by an advancing rectangular window can be computed, within a scaling factor, as the output of a trapezoidal filter with the speech as an input,

$$f(n) = \sum_{m=-M_F}^{M_F} h_t(m) s^2(n-m) \quad (4.3)$$

where $h_t(m)$ of a trapezoidal shape,

$$h_t(m) = \min(M_f, M_F - |m|). \quad (4.4)$$

M_f is the window size of the data matrix in samples corresponding and was recommended to correspond to 1 ms, and M_F is the number of observations in the data matrix recommended to correspond to 2 ms [Ma *et al.*, 1994]. The Frobenius norm filter impulse response is shown in Figure 4.2.

The Frobenius norm function $f(n)$ is depicted in Figure 4.1 as the fifth trace.

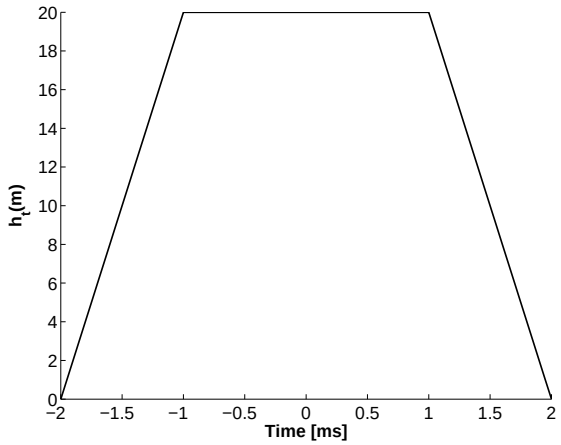

Figure 4.2: The Frobenius Norm filter is trapezoidal and where the length corresponds to the number of rows in the data matrix and the length of the constant interval to the size of the "advancing rectangular window" in [Ma *et al.*, 1994].

We can see that the waveform peaks around the instant of excitation but the maxima are poorly localised for this segment of speech. The maxima give a biased estimate of the glottal closure instants so that the average *identification error* is high but the *identification accuracy* is still low since since the variance of the error turns out to be adequately small (as can be seen in subsequent performance evaluation).

4.2.3 GD

The group delay method was proposed in [Smits and Yegnanarayana, 1995] and developed in [Murthy and Yegnanarayana, 1999]. We give a detailed definition of group delay functions in Section 4.3 where we analyse these methods in full. The group delay function plotted in sixth trace of Figure 4.1 is derived from the average group delay of the LPC residual $u(n)$

$$d(n) = \frac{1}{M} \sum_{k=0}^{M-1} \tau_n(k) \tag{4.5}$$

Table 4.1: Performance comparison for GCI detection methods on the APPLAWD database.

	Identification Rate (%)	Miss-Rate (%)	False Alarm Rate (%)	Identification Accuracy, σ (ms)
LPCR	39.9	53.5	6.61	1.38
FN	59.3	3.07	40.4	0.62
GD	81.7	2.35	16.0	0.52

where $\tau_n(k)$ is the discrete group delay of a windowed residual $u_n(m) = w(m)u(m+n)$ and $w(m)$ is in this case a Hamming window. We can see that the effect of the vocal tract has been removed from this clearly voiced, periodic interval of speech and the group delay function of the LPC residual detects the glottal closure instants as the negative going zero crossings.

4.2.4 Performance

We use the laryngograph as described in Chapter 2 to evaluate the performance of the three methods described above. For the LPCR method [Wong *et al.*, 1979], the identification rate was 39.9%, where only one GCI is determined in a larynx cycle, and the timing accuracy of those was 1.38 ms. The Frobenious Norm (FN) method [Ma *et al.*, 1994] achieved 59.3% identification rate and an accuracy of 0.62 ms. The Group Delay (GD) method [Smits and Yegnanarayana, 1995] achieved the best result of 81.7% identification rate and 0.52 ms accuracy. The results are summarised in Table 4.1 where the miss rate is the percentage of larynx cycles where no GCI was identified and the false alarm rate is the percentage of larynx cycles where more than one GCI was identified.

Identifying the closed phases also requires the determination of the glottal opening instants. GOI excitations are normally very small and so the reliable identification of GOIs remains a very challenging task with, as yet, little reported work in

the literature. The only recent study we are aware of uses the multiscale product of the wavelet transform of the speech signal to detect openings [Bouzid and Ellouze, 2004] but GOI identification will remain outside the scope of this work since errors arising from inaccurate identification of openings are less severe than errors arising from inaccurate detection of GCIs.

After detecting the glottal closure instants, the closed phase can be determined as, for example, the first 30% of the larynx cycle and multi-glottal closed phase analysis can be performed [Abberton *et al.*, 1989]. The identification accuracy is important because it determines how many milliseconds we need to discard in the beginning of each closed phase to be confident that it doesn't contain the excitation itself. The trade-off between misses and false alarms is also important. Missing a GCI causes the vocal tract not to be estimated for that particular larynx cycle, whereas a false alarm causes the vocal tract to be estimated outside the period of closure.

Another consideration is the postprocessing of the detected GCIs. There is a trade-off between missing GCI and false alarms which we address in Chapter 5 where dynamic programming is used to eliminate false alarms from a set of GCI candidates.

4.3 Group-Delay Functions

We use a group-delay function in this work to locate an energy peak in a signal. A glottal closure instant excites the vocal tract and introduces a burst of energy in the speech signal. This is much more prominent in the inverse-filtered speech, where the effect of the vocal tract has been removed. For the purpose of identifying this energy burst, we can use the LPC residual derived by inverse filtering relying on the

two-pole model of the voice source as we described in Chapter 3 [1].

In this section we describe four group-delay functions, three of which have been previously published by other authors [Smits and Yegnanarayana, 1995; Stylianou, 1999] and one developed as part of this work [Brookes *et al.*, 2006]. The difference between the methods depends on what measure to use to characterise the group-delay of the sliding window. The first method picks the zero frequency value of the group-delay [Stylianou, 1999]; the second takes the average over all frequencies [Smits and Yegnanarayana, 1995]; the third is the one we propose and characterises the group-delay of the sliding window as an energy-weighted average of the group-delay [Brookes *et al.*, 2006]; and the fourth uses energy-weighted phase [Stylianou, 1999]. We will evaluate these methods with respect to window size, signal-to-noise ratio, spurious impulses and real speech.

4.3.1 Group-delay measures

Given the linear prediction residual signal, $u(n)$, and applying a sliding M-sample Hamming window $w(m)$, we obtain a signal segment

$$x_n(m) = \begin{cases} w(m)u(m+n) & \text{for } m = 0, 1, \ldots, M-1 \\ 0 & \text{otherwise.} \end{cases} \tag{4.6}$$

The discrete-time Fourier transform (DTFT) of $x_n(m)$ is

$$\tilde{X}_n(\omega) = \sum_{m=-\infty}^{\infty} x_n(m) e^{-j\omega m} \tag{4.7}$$

[1] In practice this means that we perform AR analysis on the pre-emphasised speech.

where we let $\tilde{}$ denote a continuous function of ω. The group-delay, defined as the negative of the derivative of the phase, can be expressed as,

$$\begin{aligned}
\tilde{\tau}_n(\omega) &= -\frac{d\arg\big(\tilde{X}_n(\omega)\big)}{d\omega} \\
&= -\Im\Big(\frac{d\ln(\tilde{X}_n(\omega))}{d\omega}\Big) = -\Im\Big(\frac{1}{\tilde{X}_n(\omega)}\frac{d\tilde{X}_n(\omega)}{d\omega}\Big) \\
&= -\Im\Big(\frac{\sum_{m=1}^{M-1} -jmx_n(m)e^{-jm\omega}}{\tilde{X}_n(\omega)}\Big) \\
&= \Re\Big(\frac{\breve{\tilde{X}}_n(\omega)}{\tilde{X}_n(\omega)}\Big) \qquad (4.8)
\end{aligned}$$

where $\breve{\tilde{X}}_n(\omega)$ is the Fourier transform of $mx_n(m)$ and $\Re$ and $\Im$ indicate the real and the imaginary part respectively. We can convert this to discrete frequency by sampling

$$\begin{aligned}
\tau_n(k) &= \tilde{\tau}_n(2\pi k/M) \\
X_n(k) &= \tilde{X}_n(2\pi k/M) \qquad (4.9) \\
\breve{X}_n(k) &= \breve{\tilde{X}}_n(2\pi k/M)
\end{aligned}$$

for $k = 0, 1, \ldots, M-1$ and the discrete group-delay can be expressed as,

$$\tau_n(k) = \Re\Big(\frac{\breve{X}_n(k)}{X_n(k)}\Big). \qquad (4.10)$$

The motivation for using the group-delay is that it allows us to identify the position of an impulse within the analysis window $x_n(m)$. If $x_n(m) = \delta(m-m_0)$ is a noise-free impulse at $m = m_0$ then it follows directly from Equation 4.8 that $\tau_n(k) = m_0 \forall k$. If, however, $x_n(m)$ contains noise, then $\tau_n(k)$ will no longer be constant and we need to form some sort of an average over k.

Here we describe four different measures that estimate a single value for the delay from the start of the window to the impulse.

Zero frequency group-delay

The group-delay at $k = 0$ was proposed as a way of estimating the instant of excitation and is given by [Stylianou, 1999]

$$d_{DC}(n) = \frac{\breve{X}_n(0)}{X_n(0)} = \frac{\sum_{m=0}^{M-1} m x_n(m)}{\sum_{m=0}^{M-1} x_n(m)}. \tag{4.11}$$

This expression may be interpreted as the "centre of gravity" of $x_n(m)$. This measure is easy to calculate but it is, as we shall see, sensitive to noise and its value is unbounded if the mean value of $x(n)$ is close to zero.

Average group-delay

The frequency-averaged group-delay is given by [Smits and Yegnanarayana, 1995; Yegnanarayana and Smits, 1995; Murthy and Yegnanarayana, 1999]

$$d_{AV}(n) = \frac{1}{M} \sum_{k=0}^{M-1} \frac{\breve{X}_n(k)}{X_n(k)} \tag{4.12}$$

where the conjugate symmetry of $X(k)$ and $\breve{X}(k)$ ensures that d_{AV} is real. Direct evaluation of Equation 4.12 requires two Fourier transforms per output sample but the computation may be reduced by recursive formulae [Brookes *et al.*, 2006]. A disadvantage of this measure is that if, for some k, $X_n(k)$ is near zero, then the resultant quotient will dominate the summation in Equation 4.12 and result in a very large value for d_{AV}. To avoid such extreme values, it is recommended that a 3-term median filter be applied to $\breve{X}_n(k)/X_n(k)$ along the n-axis before performing the summation [Murthy and Yegnanarayana, 1999].

Energy-weighted group-delay

The problem of unbounded terms in Equation 4.12 may be circumvented by weighting each term by $|X_n(k)|^2$, the energy at frequency index k. We therefore propose the energy-weighted group-delay

$$\begin{aligned} d_{EW}(n) &= \frac{1}{\sum_{k=0}^{M-1}|X_n(k)|^2}\sum_{k=0}^{M-1}|X_n(k)|^2\frac{\breve{X}_n(k)}{X_n(k)} \\ &= \frac{\sum_{k=0}^{M-1}\breve{X}_n(k)X_n^*(k)}{M\sum_{m=0}^{M-1}x_n^2(m)}. \end{aligned} \tag{4.13}$$

This expression may be simplified further by noting that

$$\begin{aligned} \sum_{k=1}^{M-1}\breve{X}_n(k)X_n^*(k) &= \sum_{k,\jmath,m} m x_n(m)x_n(\jmath)e^{-\jmath 2\pi(m-\jmath)k/M} \\ &= M\sum_{m,\jmath} m x_n(m)x_n(\jmath)\delta(m-\jmath) \\ &= M\sum_{m=0}^{M-1} m x^2(m). \end{aligned} \tag{4.14}$$

Substituting this into Equation 4.13 gives,

$$d_{EW}(n) = \frac{\sum_{m=0}^{M-1} m x^2(m)}{\sum_{m=0}^{M-1} x_n^2(m)} \tag{4.15}$$

which may be viewed as the "centre of energy". Unlike the previous measures, d_{EW} is bounded and, provided that $x_n(m)$ is not identically zero, lies in the range 0 to $M-1$.

Energy-weighted phase

Equation 4.15 may be viewed as a weighted average of m using $x_n^2(m)$ as the weighting function. An alternative way of averaging m is to associate the sample positions within the window with M complex numbers of the form $e^{\jmath\pi(2m+1)/M}$ evenly spaced

around the unit circle on the complex plane. Using $x_n^2(m)$ as the weights, we then take the argument of a weighted average of these complex numbers and multiply by $M/(2\pi)$ to convert back to a delay. This forms the energy-weighted phase measure

$$d_{EP}(n) = \frac{M}{2\pi} \arg\Big(\sum_{m=0}^{M-1} x_n^2(m) e^{j\pi(2m+1)/M}\Big) - \frac{1}{2} \tag{4.16}$$

where $0 \leq \arg(\cdot) < 2\pi$. The discontinuity of $\arg(\cdot)$ has been chosen to lie midway between the complex numbers associated with $m = M-1$ and $m = 0$. It is clear from Equation 4.16 that $d_{EP}(n)$ is restricted to the range $-\frac{1}{2}$ to $M-\frac{1}{2}$. This quantity is essentially the same as one of those proposed for aligning waveform segments in a speech synthesis system [Stylianou, 1999]. This measure can also be interpreted as the phase of the fundamental term of the Fourier transform of $x_n^2(m)$.

4.3.2 Properties of group-delay measures

Here we will study the properties of the group-delay measures by examining their behaviour with synthetic signals that consist of impulses with additive white Gaussian noise. These are consistent with those already reported [Smits and Yegnanarayana, 1995; Murthy and Yegnanarayana, 1999] but we extend these studies to include an analysis of multiple impulses and a quantitative comparison between different measures.

The effect of window size

The aim is to identify excitation instants in the LPC residual and for that purpose, we will examine the group-delay functions of an impulse train, using the measures described in the previous section. The first trace in Figure 4.3 shows an idealised version of the impulse train with additive white Gaussian noise at 10 dB signal-to-noise ratio (SNR). The dominant pulse period is 100 samples with an additional

Figure 4.3: The upper trace shows an impulse train $u(n)$ with a dominant period of 100 samples and an SNR of 10 dB. The remaining traces show the waveform of $d'_{EP}(n)$ for different window size M. The circles mark the negative-going zero crossings (NZCs).

pulse in the fourth period and with the third pulse being of half amplitude.

The group delay function is shifted so that an impulse at the centre of the analysis window corresponds to zero group delay (this can be done by shifting the window $w(n)$) so that

$$d'_*(n) = d_*(n - \frac{M-1}{2}) - \frac{M-1}{2} \tag{4.17}$$

where $*$ is one of DC, AV, EW, or EP. The lower traces in Figure 4.3 show the waveform of $d'_{EP}(n)$ for four different window sizes. The window is in all cases a symmetrical Hamming window of period M. The effect of varying the window size is similar for all the measures, so we discuss it in detail only for $d'_{EP}(n)$.

For a noise-free impulse $x_n(m) = \delta(m-m_0)$ all four measures give the correct result $d_*(n) = m_0$. All measures are shift-invariant so that if $w(m) = 1$ and $u(n) =$

$u(M+n) = 0$ then

$$d_*(n+1) = d_*(n) - 1. \tag{4.18}$$

With white additive noise this is also close to being true when the impulse is close to the centre of the window and the window size less than or equal to the impulse period, i.e. $M \leq 100$. For these cases therefore, we see in Figure 4.3 that $d'_{EP}(n)$ has a negative-going zero crossing (NZC) with a gradient of -1 whenever an impulse is present in $u(n)$. Each NZC is marked with a circle.

When the window size equals the period ($M = 100$) there is a clearly defined NZC for each impulse and no spurious NZCs are introduced. However when the window size is much less than the period ($M = 50$) there are intervals between each impulse where the window contains only noise. In these intervals $d'_{EP}(n)$ is almost flat and numerous spurious NZCs are introduced. The local gradient at these spurious NZCs is close to 0 rather than -1 and this provides a possible way of distinguishing them.

If the window size is increased then it becomes likely that two or more impulses will lie within the window and individual impulses may no longer be resolved. Thus when $M = 150$, we see that the two impulses that are closest together (40 sample separation) have generated a single NZC approximately midway between them. As the window size is increased further ($M = 300$), each impulse now represents only a small fraction of the energy in the window. This means that the amplitude of the $d'_{EP}(n)$ waveform is low and the timing accuracy with which impulse locations can be identified degrades. In this example, the low level third impulse contains so little energy compared to other nearby pulses that it fails to generate an NZC at all.

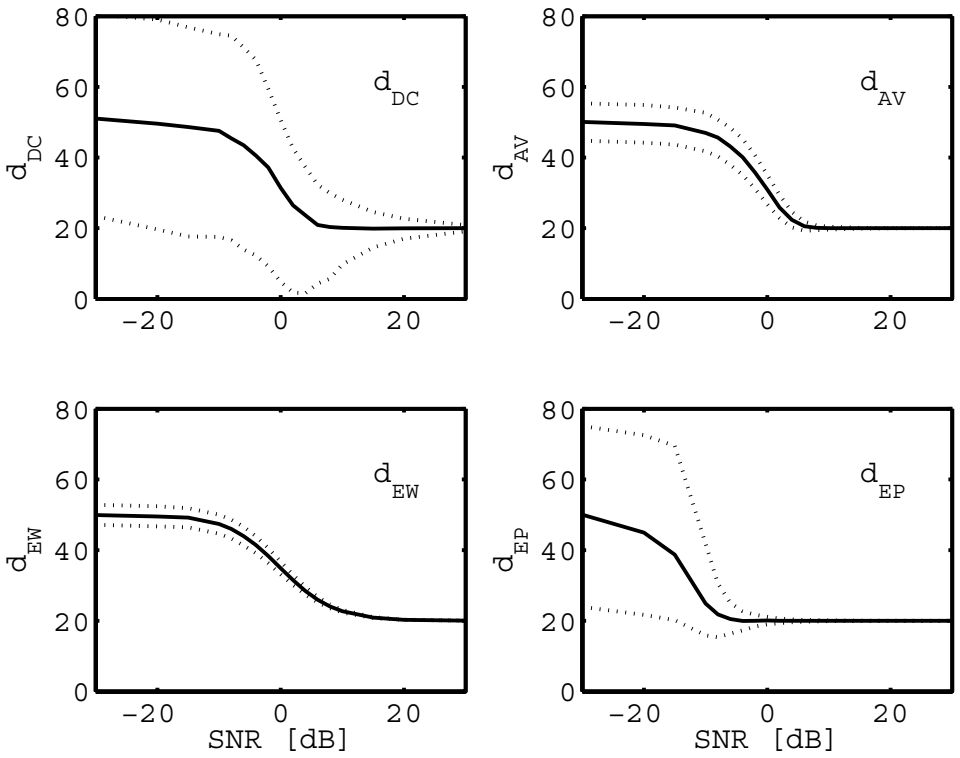

Figure 4.4: The variation of d'_{DC}, d'_{AV}, d'_{EW} and d'_{EP} as the signal-to-noise ratio (SNR) varies from -30 to +30 dB for an input consisting of a single impulse at $m_0 = 20$ with additive white Gaussian noise in a window size of $M = 101$. For each measure, the graph shows the median value of d'_* and the upper and lower quartiles.

Robustness to noise

To assess the effect of noise on the group-delay measures, we have applied them to a signal $x(m)$ consisting of a single impulse with additive white Gaussian noise. Figure 4.4 shows the behaviour of each measure as the SNR is varied from -30 to +30 dB for an impulse at sample $m_0 = 20$ within a rectangular window of size $M = 101$. For each measure, the corresponding graph shows the median value of d_* and the upper and lower quartiles. We use the median rather than the mean because of the unbounded values sometimes generated by d_{DC} and d_{AV}. At an SNR of +30 dB all measures correctly give $d_* = m_0$ with a very small inter-quartile range. As the SNR is reduced, all measures show an increasing spread and a progressive bias with the median values tending to the middle of the window, $m = 51$. The most robust measure is d_{EP} whose median value is barely affected by noise until the SNR falls below -6 dB. For this measure, the effect of noise is to add onto the summation in Equation 4.16 a random complex number of arbitrary phase. It follows that the noise will not affect the median value of d_{EP} unless the noise amplitude is large enough to cause the value of the summation to cross the positive real axis where there is a discontinuity in the arg(·) function. For impulses near the centre of the window, the summation in Equation 4.16 lies on or near the negative real axis and so for positive SNR values, the noise has little effect on the median of d_{EP}.

The measure whose median is most sensitive to noise is d_{EW} for which the effects are noticeable in Figure 4.4 for SNR as high as 14 dB. Since this measure calculates the centre of energy of the windowed signal, the bias introduced depends directly on the SNR and at an SNR of 0 dB, for example, d_{EW} will be halfway between n_0 and the window centre. The median curves for d_{DC} and d_{AV} are almost identical to each other and lie between those of the other two measures with significant bias only for SNR worse than 5 dB. Although low levels of noise have little effect on the median value of d_{DC}, they have a substantial effect on its inter-quartile

range which is considerably larger than that of the other measures.

When noise is added to an impulse train like that in Figure 4.3(a) the NZCs are affected in two ways. Firstly, the bias towards the window centre means that $d'_*(n)$ is pulled towards zero either side of the NZC and so its gradient will be less steep. It is possible, therefore, to use the gradient of $d'_*(n)$ at an NZC to estimate the SNR of the signal. The second effect is that the combination of the bias and the increased variance will add uncertainty to the position of the NZC. Figure 4.5 shows, as a function of SNR, how far an impulse must be from the centre for $d'_*(n)$ to have the probability of 0.75 of having the correct sign. We can view this as a measure of how accurately the position of the impulse will be located and how this accuracy degrades with noise. The algorithms attain a precision of 5 samples (5% of the window size) with 0.75 probability at SNR levels of 11.9, -0.5, -2.4 and -6.6 dB for the d'_{DC}, d'_{AV}, d'_{EW} and d'_{EP} measures respectively. This indicates that the timing of the NZCs is least affected by noise when using d'_{EP} and is most affected when using d'_{DC}.

Response to multiple impulses

It is possible for the analysis window to contain multiple impulses either because the window is longer than the pulse period or because, as is often the case with the LPC residual, the signal includes additional pulses or other impulsive features. We consider here the behaviour of the measures when the window contains two impulses. From the shift invariance property, expressed in Equation 4.18, we may, without loss of generality take the impulses to be at positions $m = \{0, m_0\}$ giving

$$x_n(m) = (1 - \beta)\delta(m) + \beta\delta(m - m_0) \tag{4.19}$$

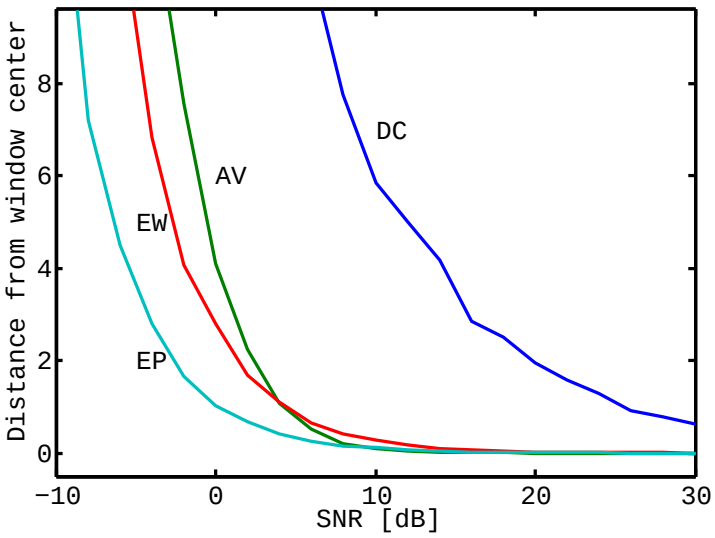

Figure 4.5: The graph shows, as a function of SNR, how far an impulse must be from the centre of 101 sample window to ensure that d'_{DC}, d'_{AV}, d'_{EW} and d'_{EP} have the correct sign with a probability of 0.75.

where the factor β lies in the range 0 to 1 and determines the relative amplitude of the two impulses. We can evaluate the four measures analytically [Brookes *et al.*, 2006] to obtain the following exact results. It is convenient to express them in terms of $\beta' = 1 - \frac{1}{\beta}$ which ranges from 0 to $-\infty$ and is the negative of the ratio of the impulse magnitudes

$$\begin{aligned} d_{DC} &= \frac{m_0}{1-\beta'} \\ d_{EW} &= \frac{m_0}{1-\beta'^2} \\ d_{AV} &= \frac{m_0}{1-\beta'^{M/\gcd(m_0,M)}} \\ d_{EP} &= \frac{M}{2\pi}\arg(\beta'^2 + e^{j2\pi m_0/M}) \mod (M) \end{aligned} \tag{4.20}$$

where $\gcd(\cdot,\cdot)$ denotes the greatest common divisor and the equation for d_{EP} should be regarded as modulo M with $-\frac{1}{2} \leq d_{EP} < M - \frac{1}{2}$.

Figure 4.6 plots the expressions from Equation 4.20 versus β for the particular

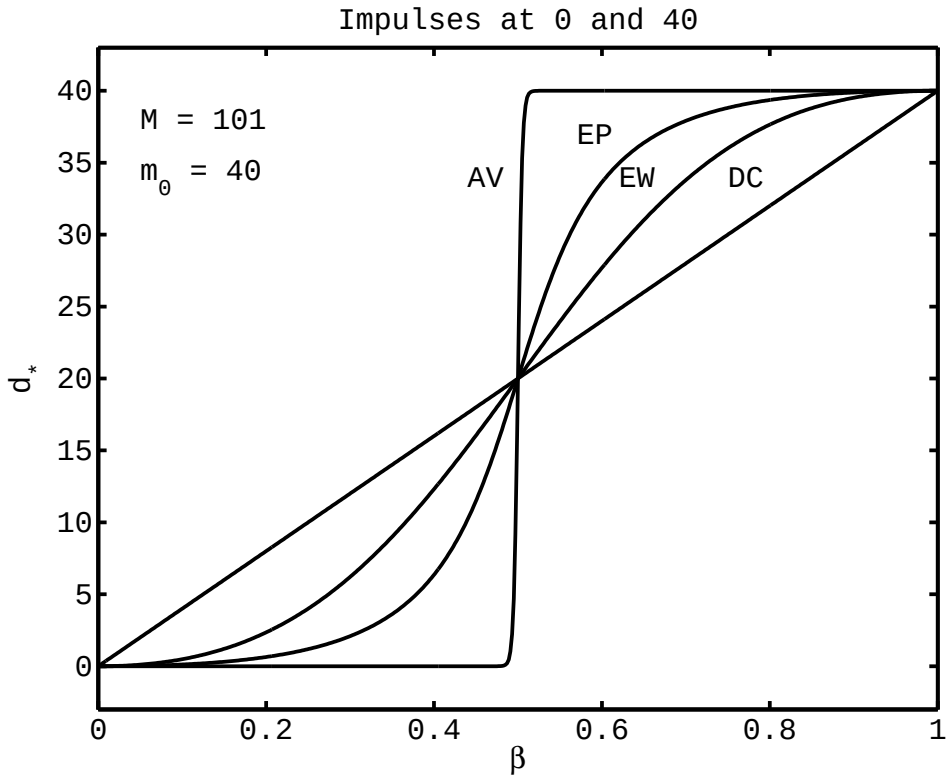

Figure 4.6: The values of d'_{DC}, d'_{AV}, d'_{EW} and d'_{EP} for a signal containing impulses at samples 0 and 40 of amplitudes $1-\beta$ and β respectively. The window size is 101 and β varies between 0 and 1.

case of $M = 101$ and $m_0 = 40$. As β varies from 0 to 1 all the measures change from $d_* = 0$ to $d_* = m_0 = 40$. Measure d_{DC} equals the centre of gravity of the pair of impulses and it therefore changes linearly with β. Measure d_{EW} on the other hand, which equals the centre of gravity of the squared input signal, is biased towards the position of larger impulse giving rise to the S-shaped curve shown. In the expression for d_{AV}, the exponent of β' depends on $\gcd(m_0, M)$ and is, for this case, equal to 101. Because this is so high, d_{AV} makes an extremely abrupt transition at $\beta = 0.5$ and this measure essentially locates the position of the highest peak in the window. It is possible to obtain a similar behaviour for d_{EW} or d_{EP} by increasing the exponent of $x_n(m)$ in Equations 4.15 and 4.16 but we have found that this does not improve their performance with real speech and so we do not discuss the resultant measure in detail. The behaviour of d_{EP} varies according to the separation of the two impulses. When they are close to each other it is almost the same as d_{EW} but as their separation increases to half the window size its graph approaches that of d_{AV}. For separation greater than $M/2$ the graph changes completely and as β increases from 0, d_{EP} decreases towards -0.5, wrapping around abruptly to $M - 0.5$ then continuing down to m_0.

4.4 Evaluation with Speech Signals

The four measures defined in Section 4.3.1 have been evaluated using the sentence subset of the APLAWD database [Lindsey *et al.*, 1987]. Figure 4.7 shows the histograms of larynx period, obtained from HQTx, for all the male and the female speakers of the APLAWD's sentence subset.

We will test the group-delay measures using identification rate and accuracy defined in Chapter 2 but we will be interested in the *detection rate* which we define as the fraction of larynx cycles that contain one or two NZC. We consider the

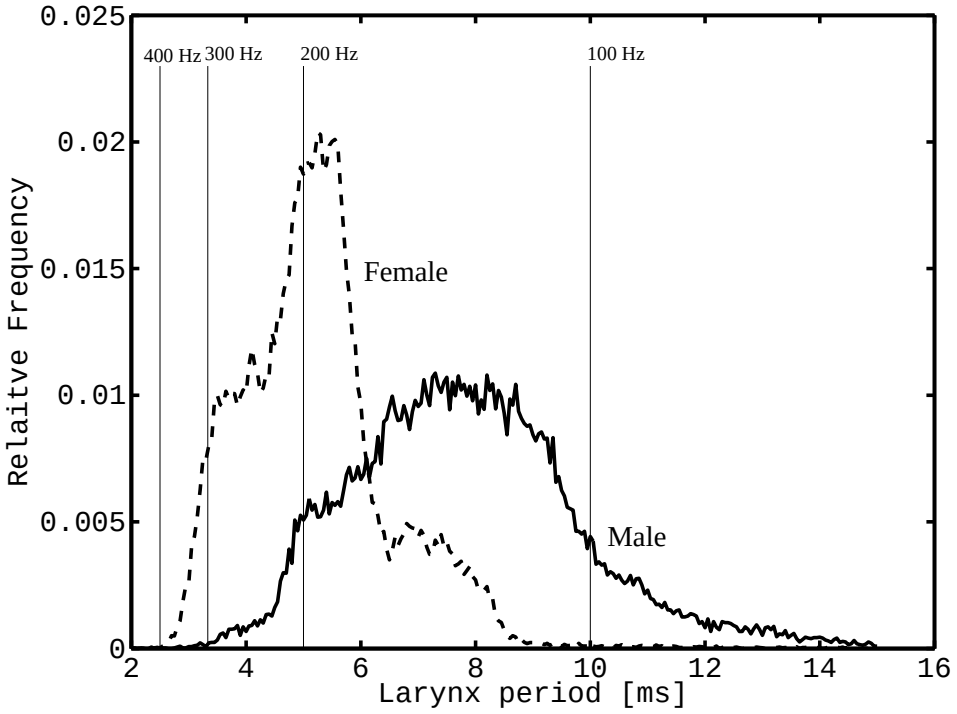

Figure 4.7: Histogram of larynx cycle periods for male and female speakers.

detection rate to be a good performance assessment of the potential of the group-delay measure to locate the GCIs provided that other techniques can be used to reject spurious NZCs. The *detection accuracy* is the standard deviation of the timing error between the GCI (identified using the HQTx algorithm) and the closest NZC for cycles containing either one or two NZCs.

4.4.1 Waveform processing

Figure 4.8 shows (a) a segment of speech with (b) the laryngograph waveform, (c) the LPC residual, $u(n)$, and (d) the waveform of $d'_{EP}(n)$ with its negative going zero-crossings (NZCs) marked by circles. The boundaries of the larynx cycles as defined in Chapter 2 are shown as vertical dashed lines. The speech is first passed through a 1^{st} order pre-emphasis filter with a 50 Hz cut-off frequency and then processed using autocorrelation LPC of order 22 with 20 ms Hamming windows overlapped

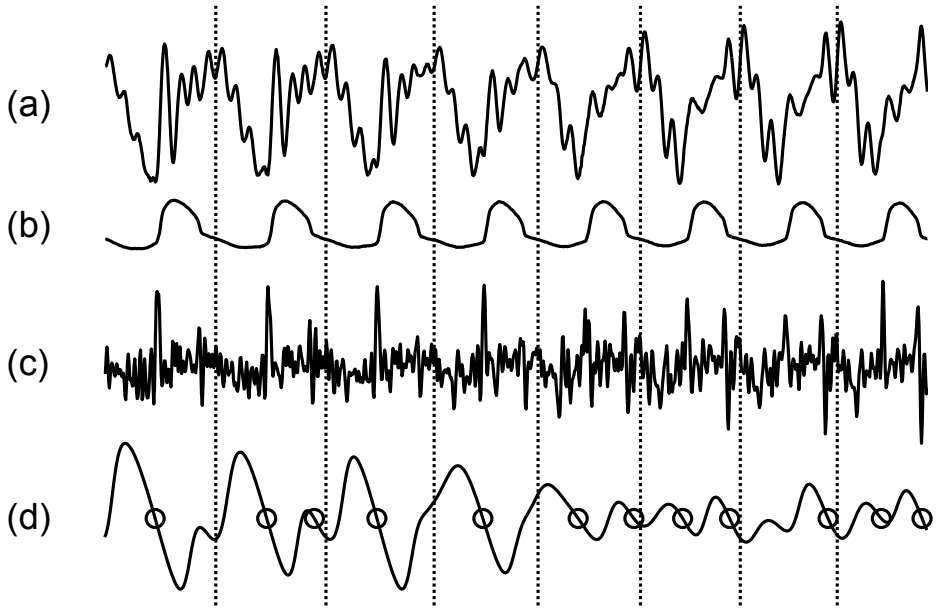

Figure 4.8: (a) Segment of male speech from diphthong /aɪ/ with (b) the laryngograph waveform, (c) the LPC residual and (d) the waveform of $d'_{EP}(n)$ with NZCs identified by circles. The vertical dashed lines indicate the larynx cycle boundaries.

by 50%. The pre-emphasised speech is inverse filtered with linear interpolation of the LPC coefficients for 2.5 ms either side of the frame boundary. Finally, in order to remove high frequency noise, the residual is low-pass filtered at 4 kHz using a 2^{nd} order Butterworth filter to obtain the signal $u(n)$. A sliding Hamming window is applied to $u(n)$ and the delay measures from Section 4.3.1 are calculated. The energy weighting, median filter and 1.5 kHz low pass filter are applied to the d'_{AV} measure and a 3-point median filter is also applied to d'_{DC} in order to remove the extreme values that are sometimes generated [Murthy and Yegnanarayana, 1999].

The speech segment of Figure 4.8 has been chosen to illustrate some of the difficulties that arise in detecting the GCIs. Identifying the GCIs has proved more difficult for this particular male speaker than for any of the other speakers in our database. His speech sometimes contains an unusually strong excitation at glottal opening which, as can be seen from the last four cycles of the LPC residual waveform

in Figure 4.8(c), may be comparable in strength to the excitation at glottal closure. In each of the first four larynx cycles a strong excitation is visible in the LPC residual at glottal closure and this results in a well-defined NZC in $d'_{EP}(n)$ at or near the centre of the cycle. In the second four larynx cycles, the poor signal-to-noise ratio of the LPC residual results in a low amplitude $d'_{EP}(n)$ waveform. In these cycles, the secondary excitation at glottal opening gives rise to an additional NZC and, in the penultimate cycle, the excitation at glottal closure is so weak that no NZC results although a ripple in $d'_{EP}(n)$ is visible. In Section 5.1.2, we will use a projection technique to determine NZC-equivalent time instants from the turning points of such ripples [Kounoudes *et al.*, 2002b; Kounoudes *et al.*, 2002a; Naylor *et al.*, 2007].

4.4.2 Timing error histograms

In most larynx cycles the measures will generate a single NZC at or near the instant of glottal closure. If, for example, a window size of 8 ms is used, then about 88% of larynx cycles give exactly one NZC in $d'_{EP}(n)$. Figure 4.9(left) shows a histogram of the deviation of the NZC from the true larynx closure as determined using HQTx applied to the laryngograph signal. The mean value is close to zero which confirms the value of 1 ms used for the larynx-to-microphone delay compensation. The standard deviation is 0.55 ms, but the underlying accuracy of the GCI estimation is somewhat better than this because variations in the larynx-to-microphone acoustic delay due to head movement can add as much as 0.1 ms onto this figure. Of the remaining 12% of larynx cycles, over three quarters contain exactly two NZCs; in most cases these occur at glottal opening and closure respectively giving rise to the histogram shown in Figure 4.9(right). The standard deviation of this tri-modal distribution is not a useful measure. Instead, we consider in our statistics only the NZC in each larynx cycle that is closest to the GCI and make the assumption that

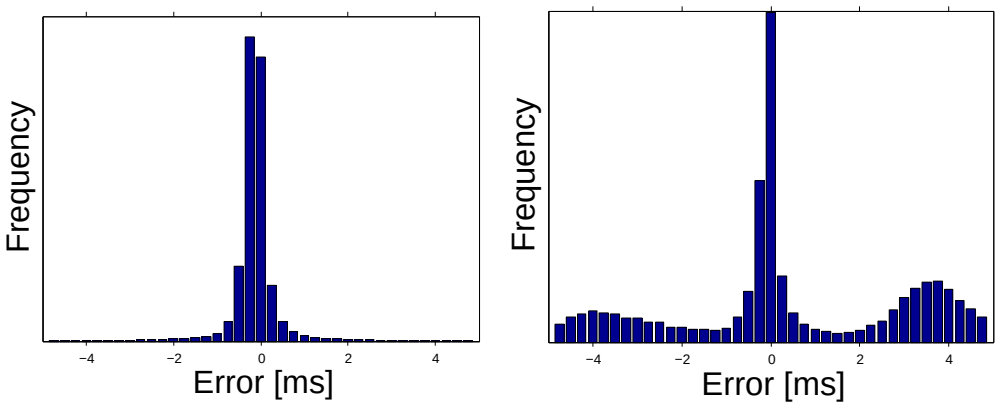

Figure 4.9: Histograms of the deviation between the instant of glottal closure and the negative zero crossings of $d'_{EP}(n)$. The histograms are for larynx cycles containing exactly one (left) and exactly two (right) NZCs respectively.

the other NZC can be rejected using other techniques [Kounoudes *et al.*, 2002b; Naylor *et al.*, 2007] which we address in Section 5.1. For this example, the standard deviation, of these "closest" NZCs is 0.97 ms and if we combine these with the single-NZC cycles, we can detect the GCI in over 97% of larynx cycles with a standard deviation of 0.6 ms. The remaining 3% of cycles either contain more than two NZCs or else contain none at all and we assume, pessimistically, that the glottal closure instant cannot be identified for any of these cycles.

4.4.3 Accuracy and detection rate

In Figure 4.10 we plot the identification rate against the identification accuracy for each of the four algorithms for window size varying between 4 ms and 13 ms in steps of 1 ms. Each curve is labelled with its algorithm abbreviation and in all cases the leftmost point corresponds to the shortest window (4 ms). The curves labelled "EPF" and "EPS" use alternative input signals and are discussed in Section 4.4.5. To take a specific example, the $d'_{EP}(n)$ measure is identified by circles and we see from the first point on the graph that for a 4 ms window, its identification accuracy is 0.34 ms but its identification rate is only 36%. This low rate arises because with a

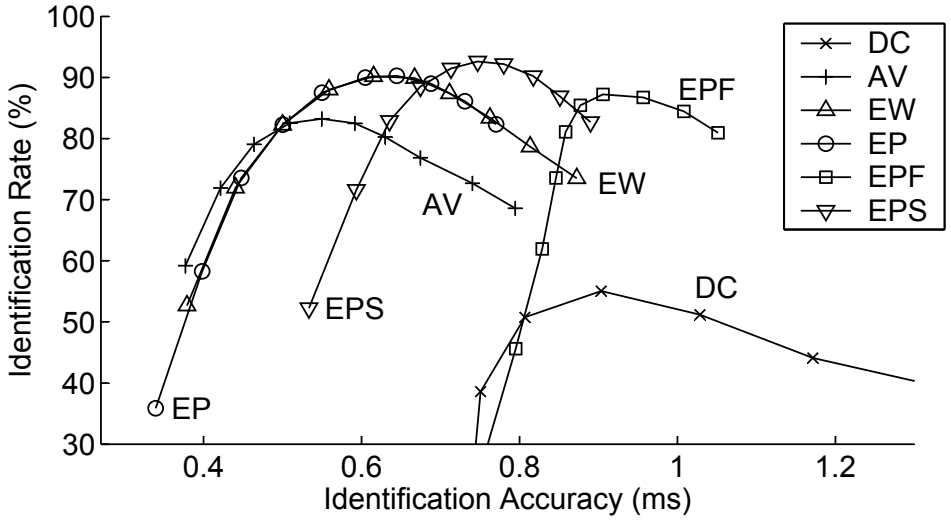

Figure 4.10: Identification rate and identification accuracy for cycles containing exactly one NZC. For each measure the window size varies from 4 ms (left most point) to 13 ms in steps of 1 ms.

window as short as this, most larynx cycles will contain more than one NZC. As the window size is increased the accuracy steadily worsens but the identification rate improves and reaches a peak of over 90.0% at a window size of 10 ms. Beyond this point, the identification rate falls again as an increasing number of cycles contain no NZC at all. The performance of the $d'_{EW}(n)$ measure is almost identical to that of $d'_{EP}(n)$ measure but reaches its peak at the shorter window size of 8 ms. The $d'_{AV}(n)$ measure has a somewhat worse performance and only achieves a peak of 83.2% while the $d'_{DC}(n)$ measure is by far the worst with a peak identification rate of only 55.0%.

In Figure 4.11 we show the same curves but this time for the detection rate and detection accuracy that are based on the larynx cycles that contain either one or two NZCs. The $d'_{EP}(n)$ and $d'_{EW}(n)$ measures again show the best performance and reach a detection rate of 97.1% for window size of 8 ms and 7 ms respectively. The $d'_{AV}(n)$ measure is slightly worse with a peak detection rate of 94.6% and although the $d'_{DC}(n)$ measure reaches a peak of 90.0%, its detection accuracy is off the graph

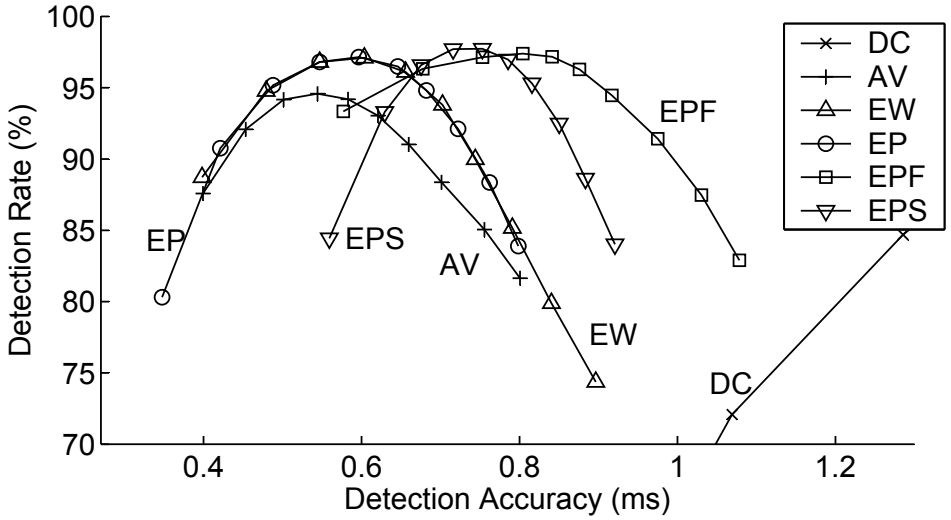

Figure 4.11: Detection rate and detection accuracy for cycles containing either one or two NZCs. For each algorithm the window size varies from 4 ms (left most point) to 13 ms in steps of 1 ms.

at 1.4 ms. In general, as the window size is decreased, the number of NZCs rises and accuracies improve. It is not surprising, therefore, that for all measures the peak detection rate has a better accuracy than the peak identification rate and occurs with a window size that is between 1 ms and 2 ms shorter.

4.4.4 Gender and linguistic content differences

In Figure 4.12 the detection rate is shown for each of the ten speakers as a function of the window size using the $d'_{EP}(n)$ measure. It can be seen that the female speakers (solid lines) are closely bunched and the peak detection rate is achieved with a window size of between 6 and 7 ms. The male speakers are less tightly bunched and have slightly worse detection rates than the female speakers with peak performance occurring at window size between 7 and 10 ms. The male speaker used in the example of Figure 4.8 shows the poorest detection rate. His speech is notable for the high proportion of cycles that include a strong excitation at glottal opening and consequently his speech also shows the worst identification rate. If a single window

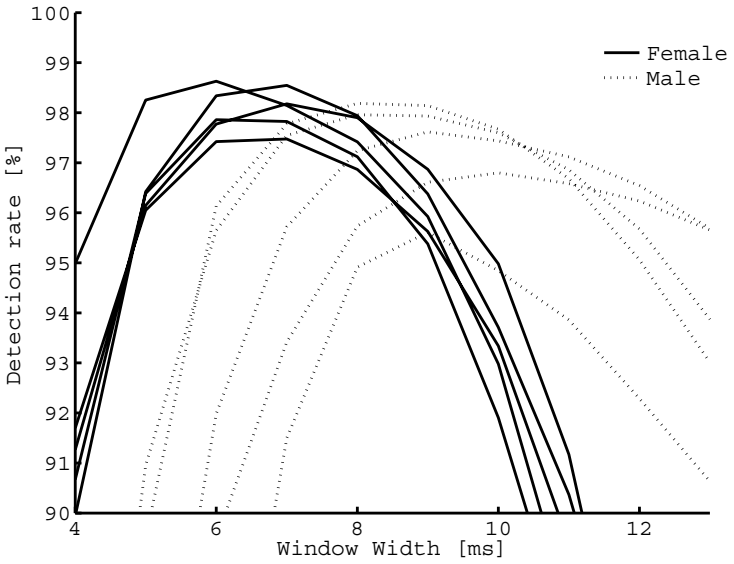

Figure 4.12: Detection rate for $d'_{EP}(n)$ as a function of window size. A separate curve is shown for each female (solid) and male (dotted) speaker.

is used for all speakers, then the optimum compromise is a window size of 8 ms. If the best window size is used for each speaker the detection rate for $d'_{EP}(n)$ measure rises from 97.1% to 97.8% with the identification rate remaining at 87.4%. It is therefore likely that the use of an auxiliary pitch estimator and an adaptive window size would give a modest improvement in performance.

Evaluating the performance of $d'_{EP}(n)$ measure on individual sentences revealed only one significant difference. The fully voiced sentence, S2, gave a slightly higher detection rate (97.8%) with much better accuracy (0.45 ms) than the other sentences which all gave similar results of 97% and 0.62 ms. We have not analysed the reasons for this in detail but we suggest that the lack of frication in sentence S2 may be a contributory factor.

4.4.5 Alternative input signals

The group-delay measures may be applied to any signal containing an energy peak at the time of glottal closure. We include in Figures 4.10 and 4.11 the results of applying

$d'_{EP}(n)$ measure to the preemphasised speech (EPS) and to the estimated glottal energy flow (EPF). The use of the preemphasised speech energy to detect glottal closures has been proposed [Ma *et al.*, 1994] and the estimation of the glottal energy flow has been developed in the literature [Brookes and Loke, 1999]. We see that applying the $d'_{EP}(n)$ measure to these signals gives good results and that the peak identification and detection rates were respectively 92.6% and 97.7% for EPS and 87.2% and 97.4% for EPF. The identification rate for EPS and the detection rates for both EPF and EPS are higher than those obtained when the $d'_{EP}(n)$ measure is applied to the LPC residual but this improvement comes at the cost of poorer accuracy. It can also be seen that as the window size is decreased below 8 ms, the EPF identification rate decreases very rapidly while its detection rate remains well above 90% even for windows as short as 4 ms. This behaviour means that the EPF measure is detecting exactly two acoustic excitations in a large fraction of cycles and indicates that it could potentially be effective in identifying the closed phase intervals. We have also evaluated the $d'_{EP}(n)$ measure on speech that had not been preemphasised but, with peak identification and detection rates of 85% and 96% respectively; this did not perform as well as EPS.

4.5 Comments

After an overview of glottal closure instant detection approaches and quantitative performance assessments of three selected methods, we developed the group delay method further. We presented four group-delay measures and evaluated the effect of analysis window size, robustness to noise, response to multiple impulses and performance for GCI detection in speech signals. It turned out that when evaluated on synthetic signals the energy-weighted phase measure performed best, but on GCI detection the energy-weighted group-delay measure showed similar perfor-

mance. Their peak detection rate of 97.1% was achieved using 8 ms and 7 ms window size respectively with 0.6 ms detection accuracy.

Despite the good performance obtained from these measures, they do not provide a complete solution to the problem of detecting GCIs. The problem is the trade-off between false alarms and misses which is adjusted by the analysis window size. We showed by using the detection rate assessment that, by reducing the window size, the GCI in 97% of larynx cycles could be detected, at the expense of increasing false alarm rate. To eliminate these false alarms it is necessary to combine them with a selection procedure which we describe in the next chapter.

Chapter 5

Detecting Epochs in Speech with DYPSA

WE outlined the speech production process in Chapter 1 and showed how the identification of closed and open phases, and more specifically identification of glottal closures is important to voice modelling in Chapter 3. Here the Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) for identification of glottal closure instants (GCIs) in voiced speech is presented and evaluated. Apart from speaker identification, detecting glottal closure instants in speech has many other practical applications in speech processing. It is for example, important in PSOLA-based concatenative synthesis and voice-morphing techniques so that coherence across segment boundaries is preserved [Hamon *et al.*, 1989; Stylianou, 1999].

The DYPSA algorithm was originally designed and implemented as part of Anastasis Kounoudes doctoral work [Kounoudes, 2001] and as part of an EPSRC project[1] called Veriphon. In that phase of the project the GCI candidates were derived from the average group delay function and the dynamic programming can-

[1]Engineering and Physical Sciences Research Council UK. Grant number GR/N01569.

didate selection was implemented. The group delay function projection technique for extra candidate generation was also designed. Needless to say, the research of Dr. Anastasis Kounoudes has been very valuable to this work.

In this work, however, we used the energy weighted group delay function as the basis for candidate generation. The further analysis and definitions which we presented in Chapter 4 were done at a later stage which explains why we used the energy weighted group delay measure instead of the energy weighted phase measure. We did not feel that the slight improvement of this measure warranted a revision of DYPSA. The contribution of this work is clarified in the Statement of Originality.

We have designed DYPSA using dynamic programming to select the best subset of candidate GCIs generated from using the group-delay function [Kleijn and Paliwal, 1995, pp. 495–518]. The dynamic programming is based on a cost function consisting of terms derived from pitch deviation, Frobenius norm, slope at zero crossings and speech similarity. The slope projections are used to generate additional GCI candidates where the group-delay function fails to change sign between a maximum and minimum.

We found that the DYPSA algorithm achieved the highest identification rate and timing accuracy of the four algorithms tested. In voiced regions, the DYPSA algorithm achieved 95.7% identification rate and 0.71 ms standard deviation of timing for the APLAWD sentences. The second best was the Group-Delay algorithm which achieved 81.7% identification rate and 0.52 ms standard deviation timing accuracy. The accuracy of the DYPSA algorithm is more than sufficient to use to segment speech for closed-phase voice analysis.

The chapter is organised as follows. In Section 5.1, we describe the slope projection and dynamic programming part of the DYPSA algorithm. We present the evaluation results in Section 5.2, where reference GCIs from subsets of the APLAWD

and SAM0 corpora were compared with the GCI outputs of the four algorithms. A summary of the work is presented in Section 5.3.

5.1 GCI Detection with DYPSA

The Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) was developed to extract glottal closure instants from voiced speech and has proven to be reliable and accurate. Preliminary results were presented at the ICASSP conference [Kounoudes *et al.*, 2002b] and the fully developed algorithm has been published recently [Naylor *et al.*, 2007].

The main idea of the DYPSA algorithm is to generate GCI candidates using the NZC of a group-delay function and a projection technique, and to select an optimum set of instants using dynamic programming based on an optimisation function that is defined using our knowledge about voiced speech.

In previously published work of our group [Kounoudes *et al.*, 2002b; Kounoudes *et al.*, 2002a; Naylor *et al.*, 2007], the the negative of the group-delay function is referred to as the phase-slope function. In this thesis, we refrain from this terminology to avoid confusion and only use the term group-delay function. Phase slope projection is referred to as group-delay projection or simply projection where appropriate.

5.1.1 Overview of the algorithm

The two main steps in the DYPSA algorithm are to generate a set of GCI candidates and to select a subset of these as the GCIs. A diagram of the main functions in the algorithm is shown in Figure 5.1. The preemphasised speech is the input and the GCIs are the output. The energy-weighted group delay function, $d'_{EW}(n)$ is

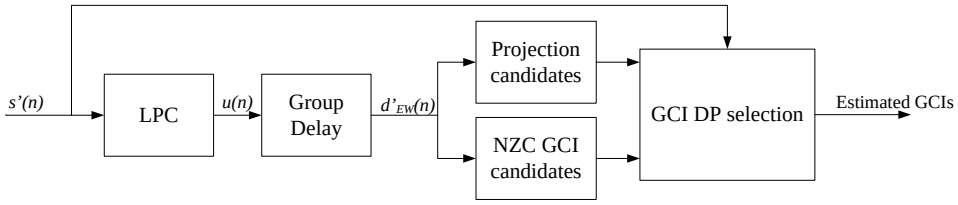

Figure 5.1: DYPSA processes speech and returns glottal closure instants (GCI). The process involves the extraction of candidate GCIs using the negative-going zero crossings (NZC) and projections of the group-delay function. The optimum set of GCIs is selected using Dynamic Programming (DP) based on costs determined by the speech signal.

computed from the LPC residual $u(n)$ and the negative-going zero crossing and the projected candidates are produced. The cost function in the dynamic programming selection is based on the speech signal which is fed into that process. The cost terms will be presented in Section 5.1.3.

The set of projected candidates is likely to pick out missed zero crossings of the group-delay function and we added these to the negative-going zero crossing (NZC) candidates. Because the selection procedure allows us to reject bad candidates, it is important to add candidates that have the potential to correspond to real closures and the projections allows us to do this.

5.1.2 Group-delay projection

In Section 4.4 we saw that GCI events can go undetected because the group-delay measure fails to cross zero appropriately, even though the turning-points and general shape of the waveform are consistent with the presence of an impulsive event indicating a GCI. An example can be seen in Figure 5.2 in which (a) shows an example segment of speech, (b) shows the laryngograph with reference GCIs extracted from the laryngograph using HQTx, (c) shows the LPC residual signal and (d) shows the group-delay function with zero-crossings indicating GCIs (marked as circles). Figure 5.2(e) shows the detail near 1593 ms and includes an example, marked by

'x', where the group-delay function fails to cross the zero axis. A GCI candidate at this instant is indicated by successive turning points but would be undetected by methods relying only on zero-crossings. To recover such otherwise undetected GCI candidates, we introduce the group-delay projection technique as illustrated in Figure 5.2(e). In this method, whenever a local maximum is followed by a local minimum without an intervening negative going zero-crossing, the midpoint between the two turning points is identified and its position projected with unit negative slope onto the time axis. This technique draws on the assumption that if the signal is a single impulse, in the absence of noise, the slope of the group-delay at a zero-crossing is unity. The number of detection misses is reduced by more than half, or from 3.6% to 1.6%, by defining the set of GCI candidates to be the union of all negative going zero-crossings and projected zero-crossings as will be shown in Section 5.2.

Most often, one pulse in the prediction residual can be expected at the instant of glottal closure. However, for some talkers, LPC analysis can give a prediction residual containing additional strong pulses, possibly for example at the start of glottal opening such as at 1586 ms in Figure 5.2(c), or an absence of any significantly strong pulses such as near 1593 ms. This latter case explains the missed zero-crossing observed in the group-delay function. The group-delay projection technique recovers such missed GCI candidates as shown by the result of group-delay projection on the 'missing' zero-crossing at 1593 ms indicated by a cross.

5.1.3 Dynamic programming cost function

Given a set of candidate GCIs determined as described above, we now wish to choose from those candidates a subset corresponding to the true GCIs. The selection of GCIs from a set of candidates is performed by minimising a cost function using N-best dynamic programming (DP) [Chen and Soong, 1994]. Procedures employing N-best DP maintain information about the N_{DP} most likely hypotheses at each step.

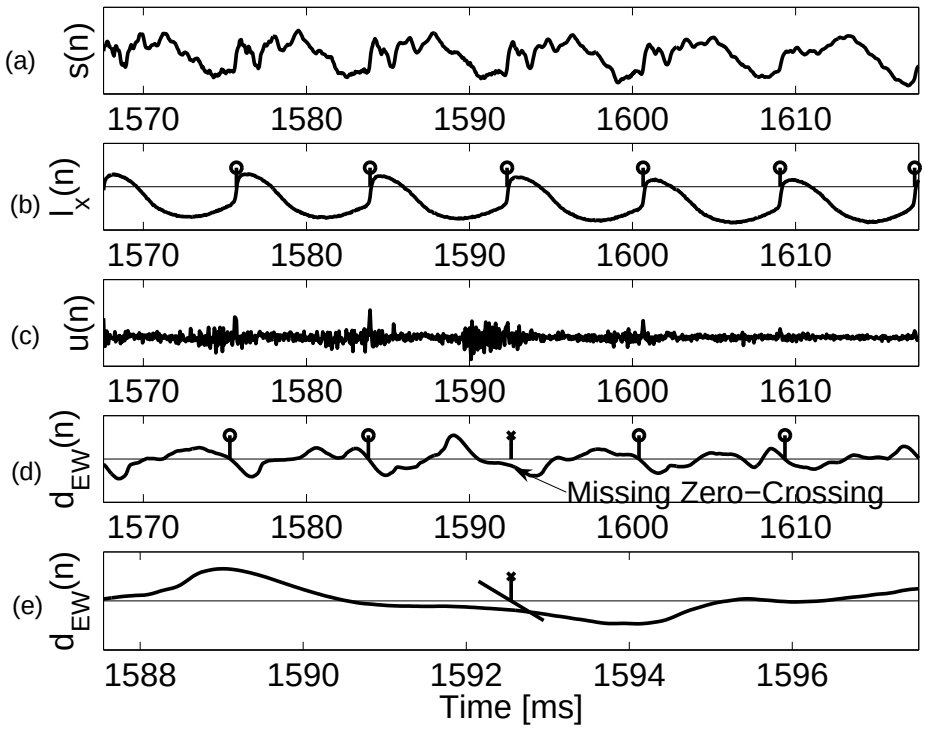

Figure 5.2: Group-delay projection: (a) Voiced speech signal, (b) laryngograph with reference GCIs extracted from the laryngograph using HQTX, (c) LPC prediction residual, (c) group-delay function with zero-crossings indicating GCIs (circles) and a missed GCI recovered using group-delay projection (cross), (e) detail showing the projection of a "missed" zero crossing onto the horizontal axis.

The value of $N_{DP} = 5$ has been chosen in this work as discussed in Section 5.2.

The terms used in the construction of the cost function are based on the attributes of the GD and FN methods and known characteristics of voiced speech including spectral quasi-stationarity and the periodic behaviour of the vocal folds [Talkin, 1995]. DYPSA employs DP to select GCIs from the set of all GCI candidates so as to minimise the cost function

$$\psi = \sum_{r} \lambda^T \cdot \bar{\psi}(r) \tag{5.1}$$

where r is the index of a GCI candidate occurring at sample n_r and $[\cdot]^T$ represents the transpose operation. The weights were determined to be

$$\lambda = [\lambda_A, \lambda_P, \lambda_J, \lambda_F, \lambda_S]^T = [0.8, 0.5, 0.4, 0.3, 0.1]^T \tag{5.2}$$

by optimisation procedure which exhaustively searched each parameter over the range $\{0, 0.1, \ldots, 1\}$ using a training subset of the APLAWD database.

The elements of the cost vector evaluated for the r^{th} GCI are

$$\bar{\psi}(r) = [\psi_A(r, r-1), \psi_P(r, r-1, r-2), \psi_J(r), \psi_F(r), \psi_S(r)]^T \tag{5.3}$$

all lie in the range $[-0.5, 0, 5]$ and are defined as the following.

Speech similarity cost

The speech waveform similarity cost uses the normalised cross-correlation estimator calculated from the speech signal as

$$\psi_A(r, r-1) = -\frac{1}{2} \frac{\xi_{r-1,r}}{\sqrt{\xi_{r-1,r-1}\ \xi_{r,r}}} \tag{5.4}$$

where $\xi_{r-1,r}$ is the covariance of 10 ms speech segments centred at samples n_{r-1} and n_r, and $\xi_{r-1,r-1}$ and $\xi_{r,r}$ are similarly computed auto-covariances. The size of the segments are chosen to be 10 ms to ensure that they include at least one larynx cycle [Talkin, 1995]. During voicing, it is common that the speech wavefrom near an instant of excitation is well correlated to the waveform at the previous excitation. The cost allows us to include the amplitude consistency at the candidates. A high cost is therefore applied to any candidate that occurs where the speech signal is not well correlated with the previous candidate. This serves effectively to penalise candidates that occur, for example, part way through a larynx cycle. Additionally, ψ_A is insensitive to the stationary amplitude and phase distortion that can be introduced by speech input devices or during transmission since it is concerned only with relative variations between consecutive larynx cycles.

Pitch deviation cost

The pitch deviation cost is a function of the current and previous two GCI candidates under consideration by the DP and is defined as

$$\psi_P(r, r-1, r-2) = 0.5 - \exp(-\left((\Delta_P - 1)N_{DP}\right)^2) \tag{5.5}$$

where the pitch consistency measure is

$$\Delta_P = \frac{\min\left((n_r - n_{r-1}), (n_{r-1} - n_{r-2})\right)}{\max\left((n_r - n_{r-1}), (n_{r-1} - n_{r-2})\right)} \tag{5.6}$$

and n_r, n_{r-1} and n_{r-2} are the sample indices of GCI candidates r, $r-1$ and $r-2$ respectively. Constant pitch is achieved with $\Delta_P = 1$.

The relationship between the pitch consistency measure Δ_P and the pitch deviation cost ψ_P is shown in Figure 5.3. The cost increases nonlinearly with Δ_P from -0.5 to $+0.5$, applying relatively small penalties for minor pitch changes based

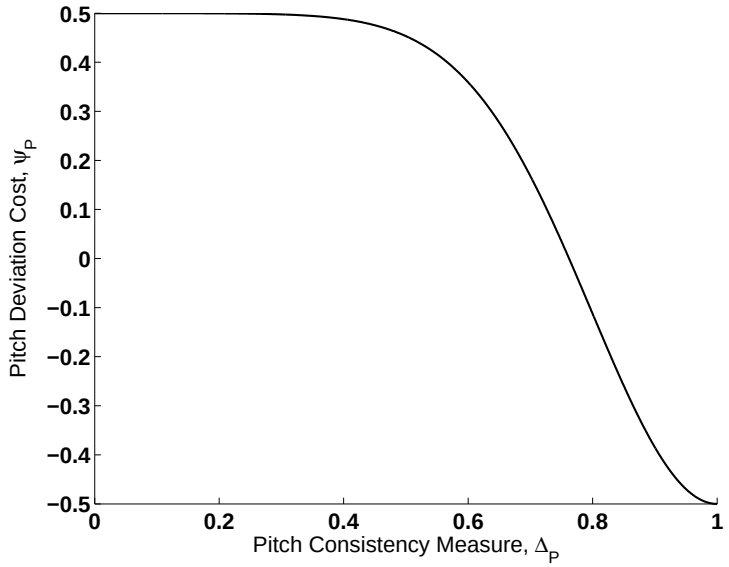

Figure 5.3: The relationship between the pitch consistency measure Δ_P and the pitch deviation cost ψ_P.

on an assumption of smooth variation in pitch over short segments of voiced speech. The rate of increase of cost with pitch deviation is controlled by ψ and zero cost is obtained at

$$\Delta_{P_0} = 1 + \frac{1}{\varsigma}\sqrt{-\ln\left(\tfrac{1}{2}\right)}. \tag{5.7}$$

In our experiments, $\varsigma = 3.3$ has been employed so as to obtain zero cost at pitch consistency of 25%. The DYPSA algorithm does not require a supplemental pitch estimator.

Projected candidate cost

The projected candidate cost penalises a GCI candidate that arises from a projection of the group-delay function onto the time-axis as described in Section 5.1.2 such that

$$\psi_J(r) = \begin{cases} 0.0, & \text{candidates from group-delay zero crossings;} \\ 0.5, & \text{candidates from group-delay projection.} \end{cases} \tag{5.8}$$

This cost function term is included because, as well as recovering GCIs that are not detectable as zero-crossings, group-delay projection can generate spurious GCIs due to noise in the LPC residual.

Normalised energy cost

The normalised energy cost is formulated as

$$\psi_F(r) = 0.5 - \frac{F(n_r)}{\breve{F}(n_r)} \tag{5.9}$$

where $F(n_r)$ is the energy of the speech signal $s(n)$ in the vicinity of GCI candidate r. This is computed using (as discussed in Section 4.2.2)

$$F(n_r) = \sum_{m=-M_F}^{M_F} \min\left(M_f, M_F - |m|\right) s^2(n_r - m) \tag{5.10}$$

where we take M_f and M_F to be the number of samples in 1 ms and 2 ms times the sampling frequency respectively [Ma *et al.*, 1994]. The term $F(n_r)$ differs only by a scale factor from the Frobenius norm measure [Ma *et al.*, 1994] but is computed more efficiently. The normalisation term $\breve{F}(n_r)$ is an estimate of the local maximum of $F(n)$ in the vicinity of GCI candidate r calculated using a sliding window of size $M_{\breve{F}}$

$$\breve{F}(n_r) = \max_m \left(F(n_r - m)\right), \;\; 0 \leq m < M_{\breve{F}}. \tag{5.11}$$

The choice of $M_{\breve{F}}$ should be large enough to ensure that the window contains at least one excitation event in voiced speech and a duration corresponding to 16 ms has therefore been chosen.

The cost ψ_F is smallest when the GCI candidate occurs at a local maximum in the short-term signal energy. This measure is used to penalise candidates that do not correspond to high energy in the speech signal such as candidates that arise

due to opening of the glottis or noise events.

Ideal group-delay slope cost

In the absence of noise, an impulsive event at the input of the group-delay function that DYPSA employs for candidate generation gives rise to a zero crossing with unit gradient at its output. Since the group-delay function is applied to the LPC residual signal containing noise, the events are not normally true impulses and therefore the gradient at the zero-crossing will deviate from minus one [McKenna, 2001]. The ideal group-delay function deviation cost is used to provide a measure of confidence in the LPC residual and the candidates obtained from it. Candidates arising from zero-crossings with gradients close to negative unity are favoured. This cost is set to zero for candidates arising from group-delay projections. We define

$$\psi_S(r) = \max(0.5 + \dot{d}(n_r), -0.5) \tag{5.12}$$

where $\dot{d}(n_r)$ is the mean value of the slope of the group-delay calculated over a short window centred on candidate r such that

$$\dot{d}(n_r) = \frac{1}{M_S}\left(d(n_r + \tfrac{M_S}{2}) - d(n_r - \tfrac{M_S}{2})\right) \tag{5.13}$$

where M_S is the even window size in samples. From our tests we have found 0.3 ms to be a satisfactory choice for the window duration and have observed that overall performance of DYPSA is insensitive to the choice of window duration over the range of 0.3 ms to 1.0 ms.

5.2 Evaluation of DYPSA

An initial evaluation of existing techniques LPCR [Wong *et al.*, 1979], FN [Ma *et al.*, 1994] and GD [Murthy and Yegnanarayana, 1999] was carried out in Chapter 3 in order to determine which of the variously proposed methods in the literature is most effective at generating GCI candidates. The window-size used in the GD method was chosen as 7.5 ms so as to be in the range of approximately one to two times the average pitch period [Murthy and Yegnanarayana, 1999]. Subsequent experiments were performed to test the effectiveness and to qualify the overall performance of DYPSA in comparison to the existing techniques.

5.2.1 Window size

We plot, in Figure 5.4, the identification, false alarm and miss rate against the identification accuracy of DYPSA obtained by varying the analysis window size of the group-delay function (denoted M in Equation 4.15). The identification rate is plotted separately in the figure on the left and and the false alarm and miss rate plotted together on the right. Together they all sum up to 100%. The window size was altered from 1 to 2 ms in increments of 0.2 ms and from 2 to 10 ms in increments of 1 ms. The points corresponding to 1, 2 and 10ms are labelled accordingly.

The plot of the identification rate can be considered in conjunction with Figure 4.10 where each group-delay function measure is compared. The difference is that the peak is reached for a much smaller window size when using the DYPSA algorithm due to the selection process performed by the dynamic programming. When detecting glottal closures using only the group-delay function, it is important to choose the right window size since a small one results in extra zero crossings and false alarms. These false alarms can be eliminated because of the selection process in DYPSA. The abrupt change in the miss rate appears when the window size is

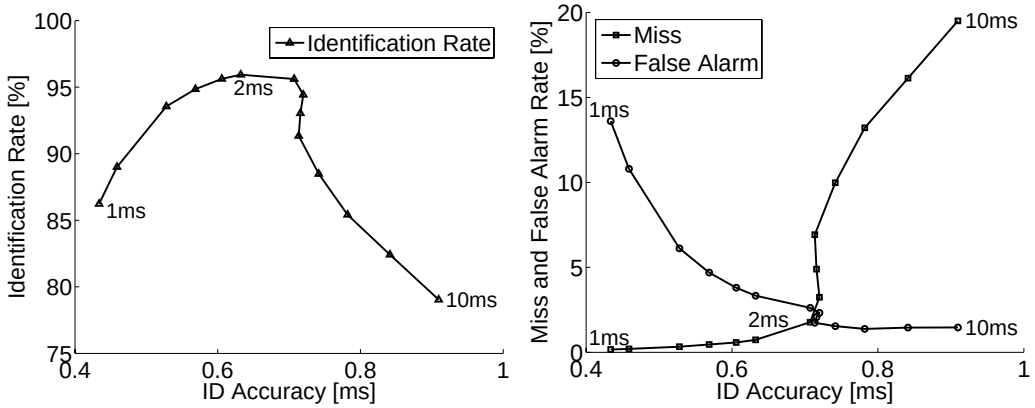

Figure 5.4: Identification rate (left) and miss- and false alarm rate (right) plotted against identification accuracy for different window size. The window was increased from 1 to 2 ms in increments of 0.2 ms and from 2 to 10 ms in increments of 1 ms.

between 3 and 7 ms. This range coincides with a window size which is increasingly unlikely to contain more than one glottal closure instant, but still likely to contain an opening. The algorithm is therefore more likely to detect the presence of a closure but the accuracy is still reduced because of the presence of an opening. Optimum window size for DYPSA appears to be 3 ms when the tradeoff between false alarms and misses is at minimum. Degradation in performance of DYPSA when the window size is too large is the same as when using a group-delay function, i.e. in increasing number of misses. Lower performance when the window size is decreased is caused by the inability of the dynamic programming to choose between too many spurious glottal closure candidates.

In Figure 5.5 the identification rate is shown for each of the ten speakers in the APLAWD database. We can see that for most speakers the identification rate reaches a broad peak between 2 and 4 ms and that identification rate for female speakers (marked with crosses) taper off more quickly as the window size increases. We use 3 ms window size in the subsequent experiments.

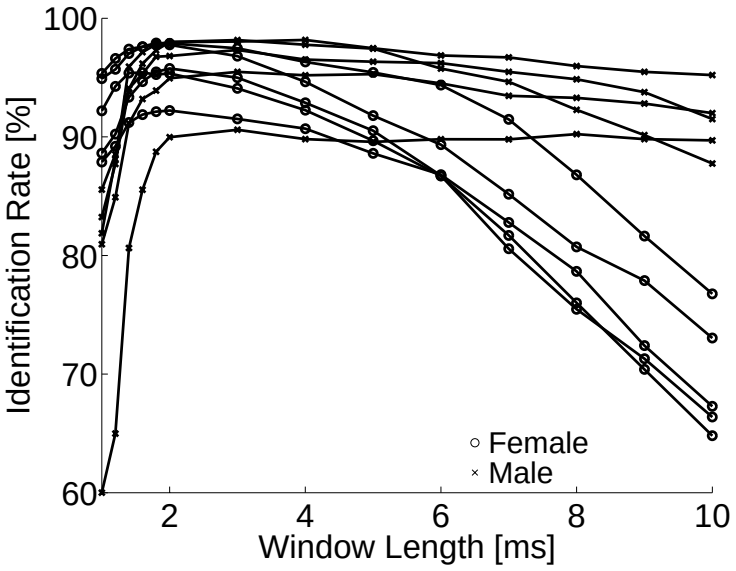

Figure 5.5: Identification rate for DYPSA as a function of window size. A separate curve is shown for each female (circles) and male (crosses) speaker in the APLAWD database.

5.2.2 Performance comparison

We performed the experiments on the APLAWD and SAM databases and the results are presented in Tables 5.1 and 5.2. The columns of both tables indicate the identification-, miss-, and false alarm rates and the identification accuracy as defined previously, for the LPCR, FN, GD and DYPSA methods. We also include results on the APLAWD database using the DYPSA algorithm without group-delay projections.

We can see from the tables that in our tests the GD method performs best out of the previously published algorithms. This motivated our choice of the group-delay function, as used in the GD method, as the principal GCI candidate generator for use within DYPSA. The identification rate for DYPSA on the APLAWD and SAM databases were 95.7% and 93.1% respectively. This significant improvement over the GD method can be accounted for by considering the capability of the DP within DYPSA to reject GCI candidates generated from the group-delay function

Table 5.1: Performance comparison for GCI detection methods on the APLAWD database. Results for DYPSA without group-delay projection are indicated by 'w/o PSP'.

	Identification Rate (%)	Miss-Rate (%)	False Alarm Rate (%)	Identification Accuracy, σ (ms)
LPCR	40.2	53.1	6.7	1.38
FN	59.5	0.3	40.2	0.62
GD	81.7	2.3	16.0	0.52
DYPSA	95.7	1.6	2.7	0.71
DYPSA w/o PSP	94.0	3.61	2.4	0.74

Table 5.2: Performance comparison for GCI detection methods on the SAM database.

	Identification Rate (%)	Miss-Rate (%)	False Alarm Rate (%)	Identification Accuracy, σ (ms)
LPCR	42.3	50.5	7.19	1.46
FN	58.7	1.36	39.9	0.59
GD	82.6	4.79	12.6	0.55
DYPSA	93.1	3.97	2.96	0.65

for which the DP cost is high. This reduces the false alarm rate typically in larynx cycles for which more than one candidate has been generated. Although DYPSA has no explicit knowledge of the range of each larynx cycle, and does not attempt to estimate it, the DP cost function can be seen effectively to penalise GCI candidates so as to reject all but one candidate per larynx cycle in most cases. The low value of σ indicates that the remaining GCI candidate per larynx cycle is close in time to the reference GCI. A further factor towards the improved performance comes from the use of group-delay projections that recover GCI candidates that would otherwise be missed. The last row of Table 5.1 shows the performance of DYPSA without group-delay projection and indicates that the group-delay projection technique identifies, with good identification accuracy, GCIs that would otherwise be missed, resulting in a rise of identification rate from 94.0% to 95.7%.

Figures 5.6 and 5.7 show the histograms of timing errors ζ for the APLAWD database and SAM database, respectively. The tri-modal shape of the histogram of the LPCR method indicates that glottal opening instants are being detected. The other methods do not exhibit such a shape. The last graphs in Figures 5.6 and 5.7 show the corresponding distribution of timing errors, ζ for DYPSA. These show that the timing errors are closely distributed around the instant of closure and that there are no other modes in the distributions.

5.2.3 Complexity tradeoff

In these experiments, a reasonable trade-off between complexity and performance of DP for the choice of number of best paths has been found when $N_{DP} = 5$. This choice is supported by Figure 5.8 which is plotted from results of an experiment in which N_{DP} was varied and shows the frequency of selection finally made by the DP at each GCI. The results indicate that the choice $N_{DP} = 3$ is adequate in 96.6% of cases, but small improvements in overall performance are obtained by increasing

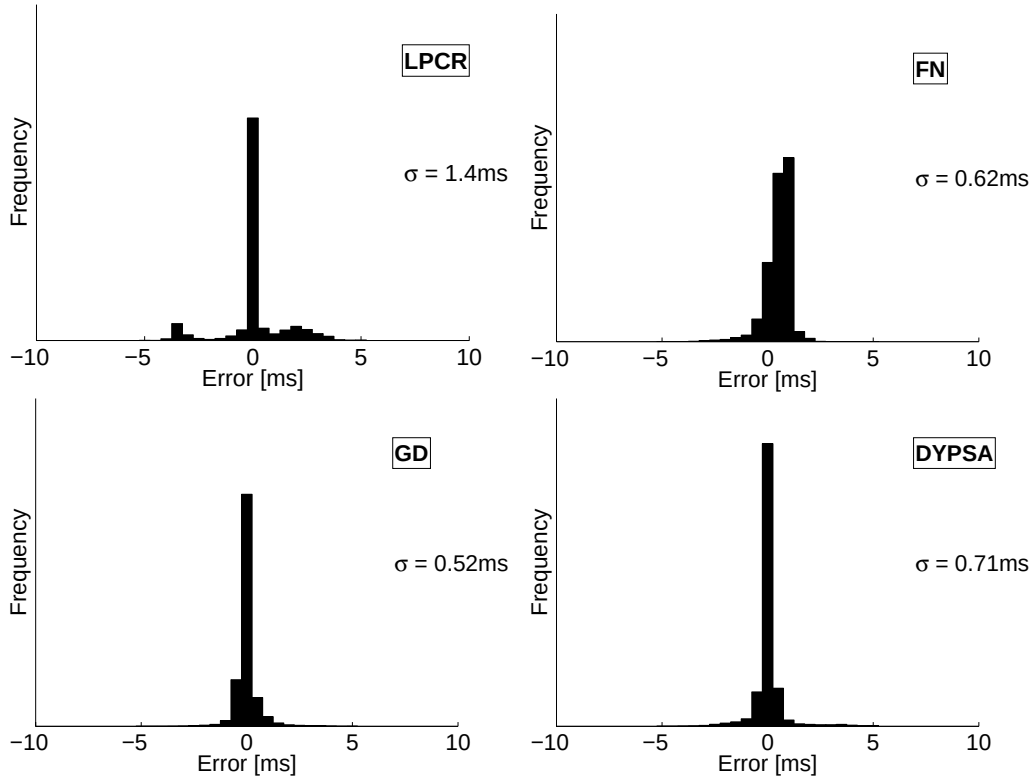

Figure 5.6: GCI timing errors, ζ, for LPCR, FN, GD and DYPSA on APLAWD

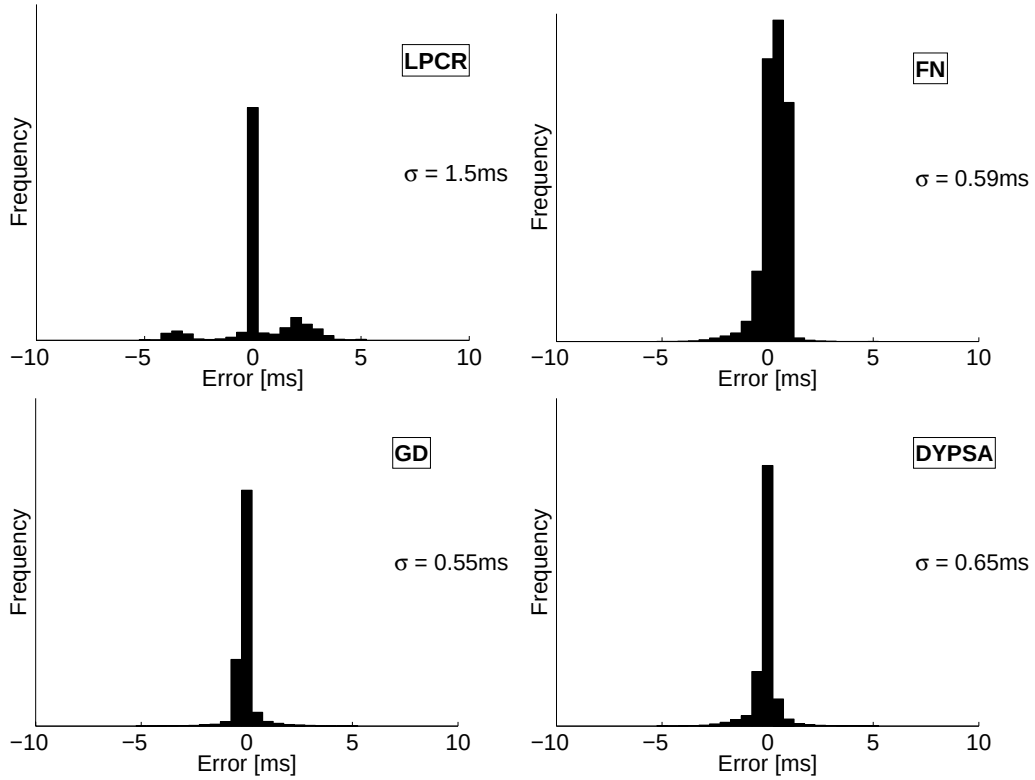

Figure 5.7: GCI timing errors, ζ, for LPCR, FN, GD and DYPSA on SAM

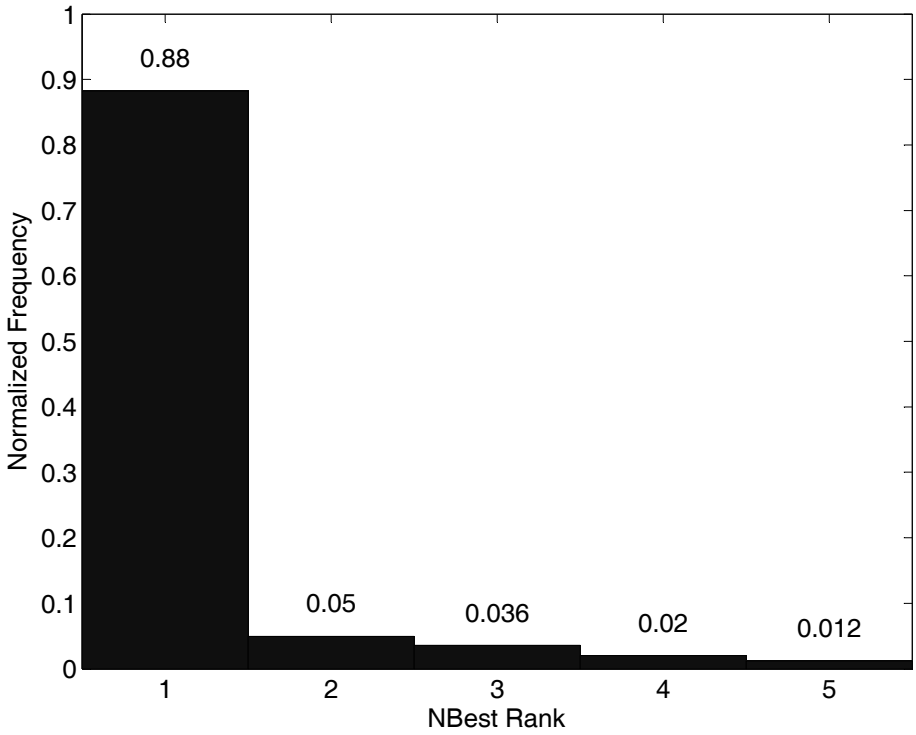

Figure 5.8: Frequency of selection from the N-best paths at each GCI.

N_{DP} at the cost of increased computation.

5.2.4 DYPSA in operation

Figure 5.9 shows an example of DYPSA's operation. For the utterance shown in Figure 5.9(a) and the detail of the same data shown in Figure 5.9(b), the lower and upper traces of ticks indicate respectively the reference GCIs obtained from the laryngograph using HQTx and GCIs obtained from DYPSA. This example has been chosen to illustrate two different types of missed GCIs. It can be seen that DYPSA's GCIs match well with the HQTx-derived GCIs except near the onset and offset of voiced regions where DYPSA misses GCIs due to the use of consistency measures in the cost function. We also see that the HQTx algorithm misses closure instants in the voiced offset close to 0.8 s which demonstrates that in these regions of speech,

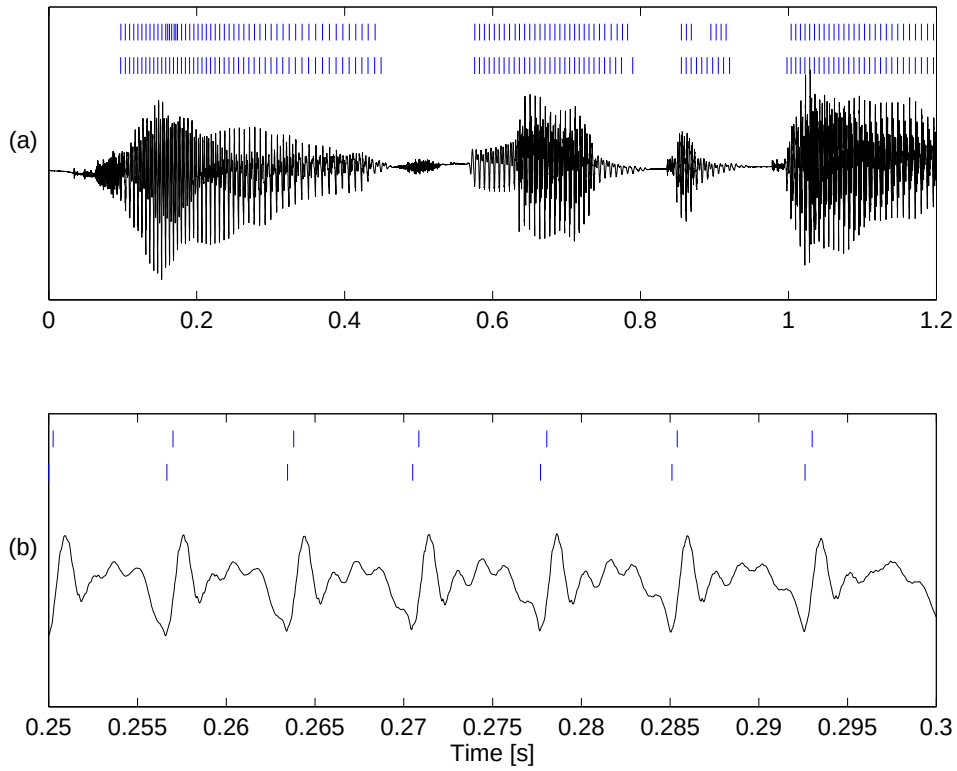

Figure 5.9: GCI identification using DYPSA, (a) speech signal and (b) detail at 0.25s. The lower row of ticks are reference GCIs determined from the laryngograph. The upper row of ticks are obtained from DYPSA. Unvoiced regions are excluded by DYPSA.

i.e. at onset and offset of voicing, the closure instants are not as well defined as in the middle of sustained period of voicing. We claim, therefore, that our performance assessment must be pessimistically skewed since they include such regions of speech.

Our implementation of DYPSA does not include a voiced/unvoiced detector. DYPSA can therefore produce unwanted "GCIs" outside periods of voicing. So far we have not counted those as false alarms since we have assumed that they can be eliminated using a voiced/unvoiced detector. We opted, in this work, to use the voiced/unvoiced/silence detector proposed in [Atal and Rabiner, 1976]. This method derived five measurements from the speech signals which are the signal energy, the zero crossing rate, the autocorrelation coefficient at unit sample delay, the first predictor coefficient and the prediction residual energy. A mean vector and a covariance matrix for each class: voiced, unvoiced, silence, was estimated from training data and the detection used a normalised Mahalanobis distance[2], between an unknown measurement vector and the mean and the covariance matrix of each class.

DYPSA also misses GCIs occasionally within a voiced segment such as that illustrated in this example near 0.9 s. Figure 5.9(b), showing a detail from the waveform, illustrates that the GCIs obtained from DYPSA are aligned with a consistent offset to the reference GCIs. Such an offset may, for example, arise from imperfect time-alignment between the speech and laryngograph in the test data and is not included in our assessment of accuracy.

Figure 5.10 presents an illustrative example of the components of the DYPSA cost function. A segment of voiced speech is shown in Figure 5.10(a) in which the upper ticks represent the candidate GCIs and the vertical lines indicate the GCIs selected by the DP. Figure 5.10(b) shows the time-variation of four components of

[2]This could be improved by using the class likelihood of the measurements. Furthermore, a more elaborate probability density function for each class could be estimated, such as a Gaussian mixture model.

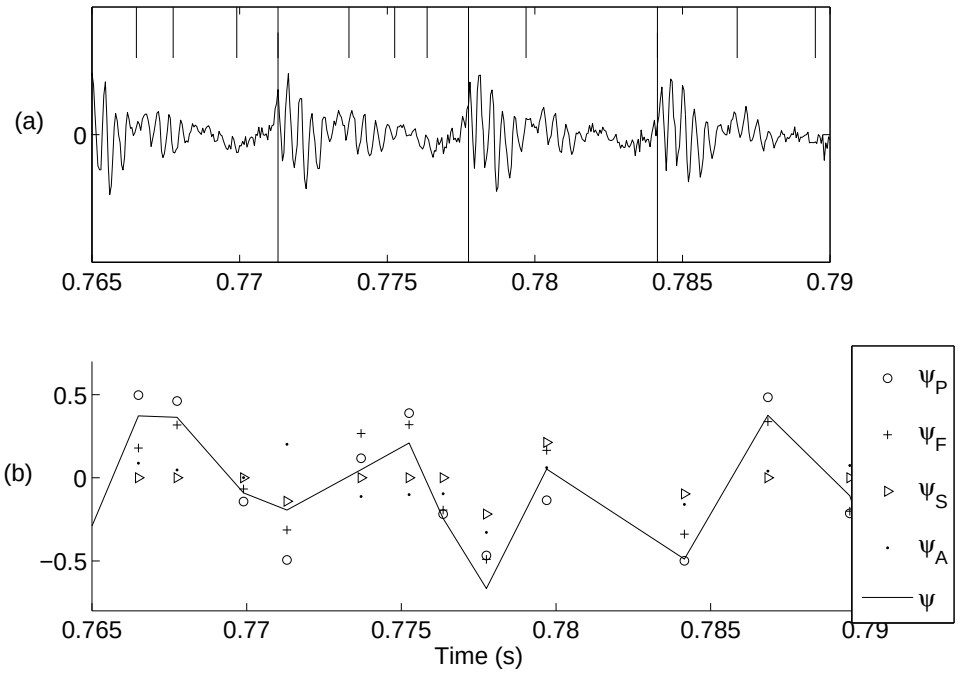

Figure 5.10: Components of the DYPSA Cost Function, (a) voiced speech segment with GCI candidates (upper ticks) and selected GCIs determined by DYPSA (vertical lines), (b) components of the cost function and total cost c.

the cost function and their weighted sum, $\lambda^T \cdot \bar{\psi}(r)$, for each of the candidates. For a given candidate r, the cost function components $\psi_F(r)$ and $\psi_S(r)$ can be determined independently of any other GCI selections. However, the other cost components are dependent on the particular selection of GCIs from candidates made by the DP. Therefore, in this example, the cost of selecting a given candidate r to be the j^{th} GCI is found using DP as the optimal cost across all possible selections for which candidate r is selected to be GCI j. It can be seen that, as expected, the overall cost is higher for the rejected candidates than for the selected GCIs. The pitch deviation cost, ψ_P, can be seen to discriminate well in most cases and this is consistent with the high weighting of this cost component $\lambda_P = 0.5$. Near 0.78 s, however, its cost of zero indicates uncertainty and the successful rejection of the candidate is achieved by the other cost function components in the DP. The component with the highest weighting is the speech similarity cost (amplitude consistency), ψ_A, with $\lambda_A = 0.8$.

It can be seen that during the second half of this example ψ_A discriminates the GCIs correctly but that during the first half it incorrectly penalises a GCI. Nevertheless, the contributions of the other cost function components are sufficient to lead the DP to select the GCI correctly.

5.3 Concluding Remarks

5.3.1 Summary

Candidate GCIs are obtained in DYPSA as negative going zero-crossings of the group-delay function. The choice of the analysis window size M for calculation of the group-delay function in Equation 4.13 is important. If it is too long relative to the pitch period then it is likely to span more than one excitation event giving rise to missed GCI candidate zero-crossings as discussed in the literature [Smits and Yegnanarayana, 1995]. The likelihood of the analysis window spanning more than one excitation event for a chosen value of M increases in speech with unusually high pitch or when there is a strong excitation at opening as well as closure. Unusually high pitch can occur for some females as well as for talkers speaking under stress. Alternatively, if the analysis window is too short relative to the pitch period then many spurious GCI candidate zero-crossings will be generated. Noise can be expected to give rise to a similar effect, although detailed study of the effects of noise on DYPSA are outside the scope of the current study. The use of DP within DYPSA makes the algorithm robust to spurious candidates since they are penalised in the cost function. In contrast, a missed zero-crossing represents an error which the DP cannot recover. We therefore incorporated two important features into DYPSA's candidate generation technique. Firstly, because of the introduction of DP, we could employ a shorter window than previously proposed [Smits and Yegnanarayana, 1995]. Secondly, we introduced the group-delay projection technique. These techniques ensure

the inclusion of valid GCI candidates that would otherwise be missed and result in improved robustness to the choice of analysis window size M and its relation to the pitch.

Chapter 6

Voice Source Cepstrum for Speaker Identification

CLOSED phase analysis is used to derive vocal tract and voice source parameters for speaker identification. Studies of speaker recognition using features related to the voice source have shown good promise in recent years, e.g. [Sönmez *et al.*, 1997; Plumpe *et al.*, 1999; Shriberg *et al.*, 2005]. Many of them rely on large scale features such as pitch and intensity contours whereas little has been reported on spectral features of the voice source. The motivation for using such features stems from studies done on voice source analysis for characterising different speaker traits [Karlsson, 1985; Karlsson, 1988]. It was shown that voice source features such as open quotient, peak flow and DC flow of the voice source signal can be used to categorise speakers into different speaker types. Similarly, the voice source signal was later extracted and processed for speaker recognition feature extraction [Plumpe *et al.*, 1999]. Our approach is comparable to this but, although this method shows some promising results, we choose to avoid the inverse filtering operation in the time domain. The reason for circumventing time-domain processing in the feature extraction is to avoid introducing the phase distortion that is present

in the recorded speech, which we discussed in Chapter 3. We estimate the vocal tract spectrum using closed phase analysis and convert this into cepstrum parameters [Davis and Mermelstein, 1980]. Voice source cepstrum coefficients are obtained as the difference between mel-frequency and vocal-tract cepstrum coefficients.

We choose to demonstrate the voice source cepstrum coefficients with text-independent closed-set speaker identification with a statistical pattern classification approach using Gaussian mixture models to approximate the probability density of the coefficients. Each speaker has a Gaussian mixture probability density function that overlaps with those of other speakers. With this classification setup, any misclassifications are due to these overlaps and are not corrected for or affected in any other way by things such as likelihood normalisation, thresholds, or temporal dependency of the feature vector sequence.

The chapter is organised as follows. We give an overview feature extraction for speaker recognition in Section 6.1 before proposing the voice source feature extraction in Section 6.2. Speaker classification is explained in Section 6.3 and experimental results are presented in Section 6.4. The chapter is concluded with short discussion in Section 6.5.

6.1 Speaker Recognition Feature Extraction

The plethora of feature extraction techniques for speech and speaker recognition developed over the past 40 years have converged to use a cepstral representation derived from a filterbank designed according to a model of the auditory system. The MFCC was shown to outperform feature sets such as the reflection coefficients and linear predication cepstrum parameters on monosyllabic word recognition task [Davis and Mermelstein, 1980] and have been popular ever since [O'Shaughnessy, 2003]. MFCCs have also become the most popular choice for feature vectors in

speaker recognition [Reynolds, 1995]. The basic processing steps are shown in Figure 6.1 and we describe these steps in detail in Section 6.2 as part of the proposed feature set developed in this work. Alternative approaches are reviewed here.

6.1.1 Auditory approaches

The development of front-end processing techniques for speech has benefitted from research of the auditory process. The mel-scale is an example of a widely used front-end processing component developed from the study of how the ear warps the frequency scale and how we perceive frequencies accordingly. Mel-frequency cepstrum coefficients are based on these principles [Davis and Mermelstein, 1980]. Perceptual linear prediction (PLP) coefficients have also been proposed and have become increasingly common [Hermansky, 1990; Hermansky and Morgan, 1994]. The processing steps of PLP and MFCC are similar but they differ only in the way the non-linear compression is done, where PLP uses the cube-root instead of the logarithm, and the way cepstral smoothing is performed by autoregressive modelling in PLP instead of the high quefrency cepstral coefficients being discarded [Hermansky, 1990; Gold and Morgan, 2000].

Other techniques inspired by auditory research have been suggested, but none have achieved such widespread use as MFCC or PLP. The ensemble interval histogram (EIH) simulates the outer part of the auditory periphery and represents the speech in terms of the auditory nerve firing rate and synchrony [Ghitza, 1994]. The spectral processing is performed in the time-domain using cochlear filters and interval histograms of level crossings computed in each frequency band. For phone classification, EIH was shown to perform better then mel-cepstrum on telephone speech although mel-cepstrum still performed better on clean speech [Sandhu and Ghitza, 1995]. This line of research has led to sub-band spectral processing and the success of these methods has mainly been to suppress noise and channel affects for

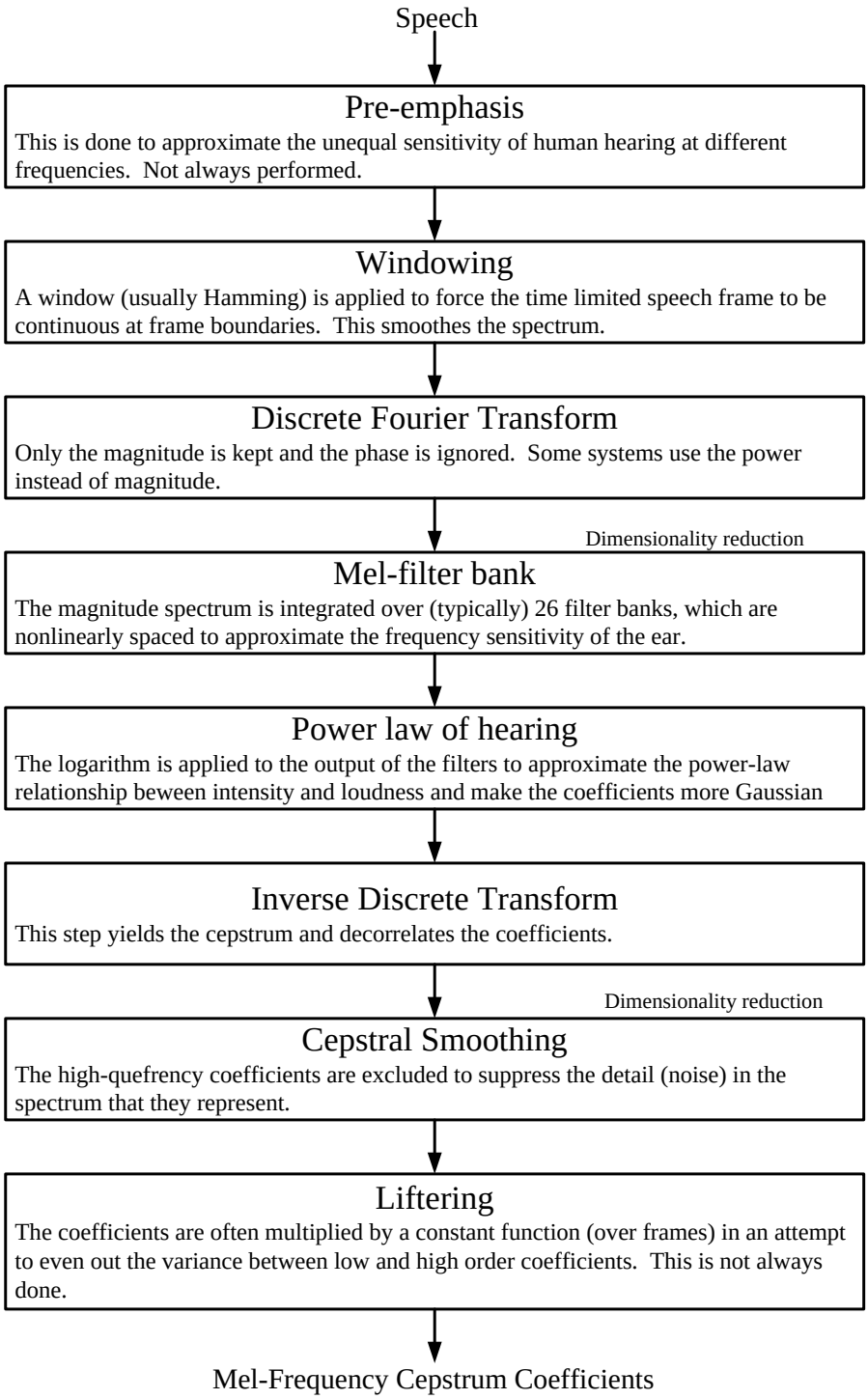

Figure 6.1: The processing steps involved computing mel-frequency cepstrum coefficients.

improved recognition [Gajic and Paliwal, 2006].

6.1.2 Other small scale features

Detection of amplitude and frequency (AM-FM) modulations in the speech signal was used to extract features for speech and speaker recognition [Dimitriadis *et al.*, 2005]. Such features can also be based on the Teager energy operator [Teager, 1980] that can track the instantaneous amplitude and frequency modulations within one pitch period [Jabloun and Enis Cetin, 1999].

There are not many reported approaches for small scale voice source feature extraction. The time-domain voice source signal was extracted and analysed for speaker identification [Plumpe *et al.*, 1999]. The method used twelve voice source parameters and fourteen LPC cepstral coefficients. Seven of the twelve voice source parameters were derived from the coarse structure of the voice source signal and five from the fine structure. We compare the speaker identification performance of this method with our proposed features in Section 6.5. A recent study proposes voice source features based on the LPC residual of the speech for speaker verification [Murty and Yegnanarayana, 2006]. The residual phase is computed from the residual using the Hilbert transform. The recognition is based on capturing nonlinear relations in the residual phase around the GCIs by using auto-associative neural networks. The speaker verification experiment is also performed using mel-frequency cepstrum coefficients and the combination of the two classifiers. Using the NIST-2003 database [NIS, 2003], the voice source features achieve 22% equal error rate, whereas the MFCC features achieve 14%, but combined the equal error rate is reduced to 10.5%. We will discuss this further in Chapter 7.

6.1.3 Prosodic features

Prosody is a term used to describe suprasegmental patterns in speech derived from pitch (or intonation), timing (or rhythm) and intensity (or stress) [Botinis *et al.*, 2001]. They can convey structural, semantic and functional linguistic information but they can also contain information about speaker characteristics such as identity. Features derived from such patterns are extracted over long period of time compared to the time interval over which spectral features such as MFCC are computed. Larger databases have recently enabled more accurate modelling of such patterns [Shriberg *et al.*, 2005] and prosodic features have been applied to many speech tasks such as sentence and topic segmentation [Shriberg *et al.*, 2000], evaluation of nativeness for language learners [Teixeira *et al.*, 2000], and recently, promising approaches have been proposed in speech recognition that combine pitch features with spectral features in a hidden Markov model [Morgan *et al.*, 2005].

In a study using pitch for speaker recognition, pitch contours were extracted from speech utterances of from 2 s duration. The data consisted of 10 speakers each uttering six sentences and the reported misclassification rate was 3% [Atal, 1972]. Another early study of speaker verification relied on pitch, intensity and the three lowest formant frequencies [Lummis, 1973]. Good results were achieved on a larger database by computing the mean, variance, skew and kurtosis over a 10 ms frame and used this as a feature vector [Carey *et al.*, 1996]. A piecewise-linear pitch contour was estimated over voiced regions and the distribution of the segment median, slope, duration, voiced segment duration, and pause duration were compared with the true speaker distribution for verification [Sönmez *et al.*, 1997; Sönmez *et al.*, 1998]. Pitch frequency features were combined with mel-frequency cepstrum by estimating the joint probability distribution of the two sets [Ezzaidi *et al.*, 2001b; Ezzaidi *et al.*, 2001a]. The quality of the combination is shown to depend on the test-segment's duration, with increasing performance as the segments get

longer. Lexical information has also been used in conjunction with prosodic features for speaker identification [Weber *et al.*, 2002]. In addition to the prosodic pitch- and duration-related features, they added word usage and conversational style features including relative frequency of disfluency classes, such as pause fillers.

6.1.4 Combining feature sets

The aim of the *SuperSID* project was to exploit high-level speech information for speaker recognition [Reynolds *et al.*, 2003]. The information applied to speaker verification tests included prosodic features, such as pitch trajectories and duration statistics [Adami and Hermansky, 2003]; phone features [Adami *et al.*, 2003], using n-grams or binary trees to determine phone sequence likelihoods [Navratil *et al.*, 2003]; and conversational features, based on summary statistic pitch and phone occurrence in each turn of the conversation [Peskin *et al.*, 2003] and pronunciation features [Klusacek *et al.*, 2003]. Lexical information, such as word usage and frequency, were also mentioned as a possible source for speaker recognition features [Reynolds *et al.*, 2002].

The problem with assessing the performance of higher-level feature sets such as prosodic features is the challenging benchmark set by acoustic features sets such as MFCCs. The obvious approach is to combine features derived at different levels of scale. However this is not straightforward since the data rate is different at each level and the feature vector sequences are not synchronised. This is a more serious problem in text-dependent speaker recognition and speech recognition where temporal information is crucial. The classifier needs to be adapted to cope with two or more feature sequences at different data rates [Morgan *et al.*, 2005; Jin *et al.*, 2003]. Classification fusion techniques can be applied for text-independent speaker recognition [Campbell *et al.*, 2003].

6.2 Voice Source Feature Extraction

The proposed method for voice source feature extraction depends on closed phase analysis. We refrain from fitting parametric models to the voice source time waveform since this is severely affected by phase distortion, as discussed in Chapter 3. Instead of deriving mel-frequency cepstrum coefficients directly from the voice source signal we circumvent inverse filtering by estimating the AR spectrum of the vocal tract by using the vocal tract filter $V(z)$ derived with closed phase analysis. The AR spectrum is processed in the same way as the DFT spectrum is treated in mel-frequency cepstrum processing as shown in Figure 6.2. The voice source cepstrum coefficients, $\mathbf{c_{vs}}$, are then derived as the difference between the mel-frequency cepstrum coefficients $\mathbf{c_{mf}}$ and the vocal tract cepstrum coefficients $\mathbf{c_{vt}}$. This processing is described and analysed in this section.

6.2.1 Mel-frequency cepstrum coefficients

The MFCCs are derived as part of our proposed method and here we review this processing. The speech signal is processed on a frame-by-frame basis. Each frame is multiplied by a Hamming window to suppress discontinuities in the periodic extension required for the subsequent discrete Fourier transform (DFT). The windowed speech is denoted as,

$$s_n(m) = w(m)s(m+n) \qquad \text{for } m = 0, 1, \ldots, M-1 \tag{6.1}$$

and the window size is determined by setting $M/f_s \approx 32$ ms where f_s is the speech sampling frequency. We take n to be the first sample of each frame, repeating every 10 ms.

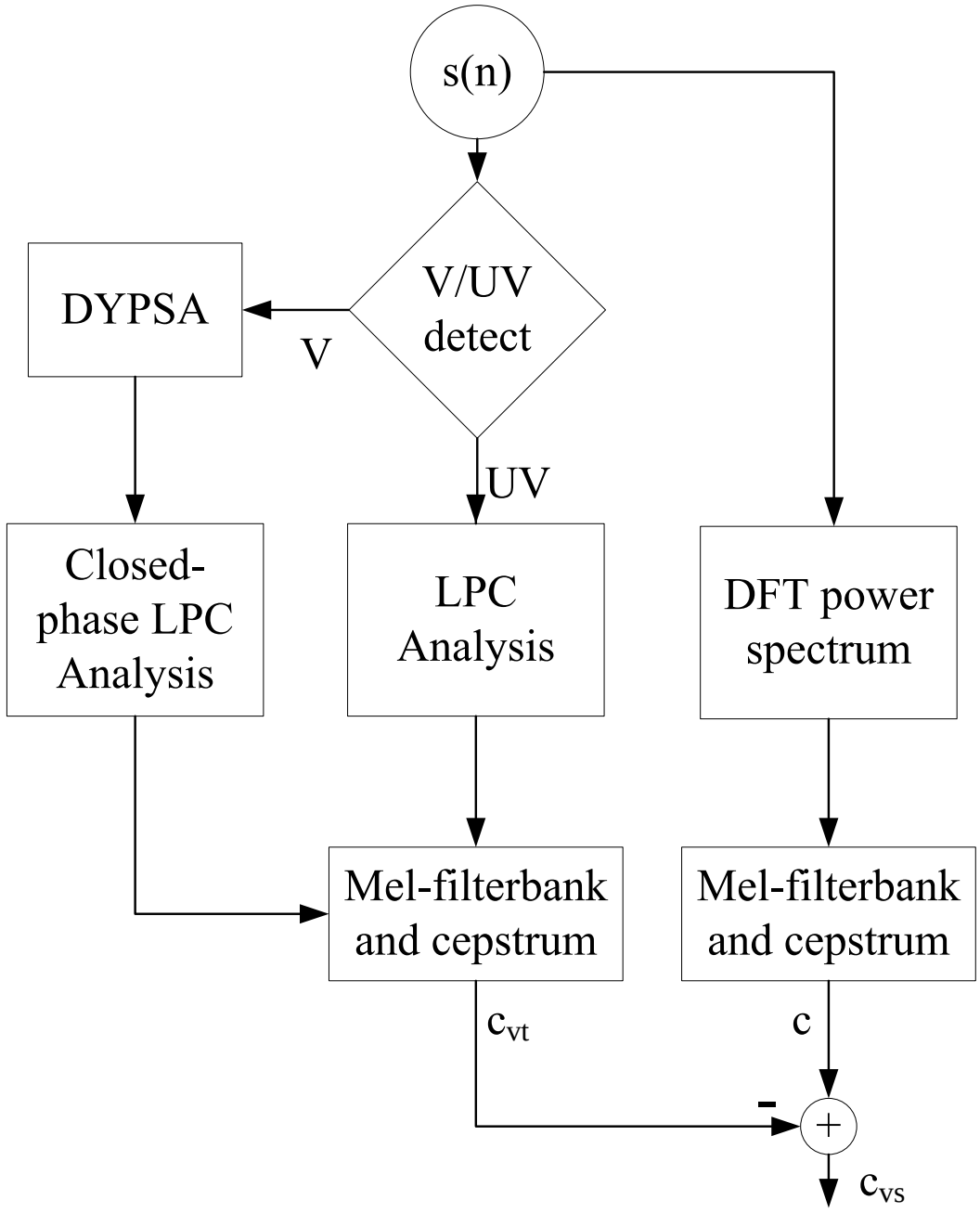

Figure 6.2: The processing steps involved computing voice-source cepstrum coefficients.

The spectrum is estimated with the discrete Fourier transform,

$$S(n,k) = \sum_{m=0}^{M-1} s_n(m) e^{-j2\pi \frac{km}{K_s}} \tag{6.2}$$

where $k \in \{0, 1, \ldots, K_s\}$ is the discrete frequency index and $K_s \geq M$ is the number of DFT points. The energy spectrum $|S(n,k)|^2$ [Davis and Mermelstein, 1980] or the magnitude spectrum $|S(n,k)|$ [Young *et al.*, 2002] is then used but the phase is ignored. The speech magnitude spectrum, $|S(n,k)|$ is shown for a typical voiced speech frame in Figure 6.3(left) together with the closed phase AR spectral envelope $|V(n,k)|$, described below. The plot on the right also shows the voice source spectrum $\frac{|S(n,k)|}{|V(n,k)|}$.

A filterbank is applied to the spectrum to determine the strength of the signal in each frequency band. The filter spacing approximately follows the frequency sensitivity of the ear. Figure 6.4 shows the triangular mel-filters used in the process[1]. The output of the filterbank is,

$$Y(n,j) = \sum_{k=0}^{K_s-1} |S(n,k)| Q_j(k), \tag{6.3}$$

where $Q_j(k)$ is the j-th mel-filter and $j \in \{1, \ldots, N_j\}$, and typically, $N_j = 26$.

The mel-frequency cepstrum coefficients are computed as

$$c(n,l) = \sum_{j=1}^{N_j} \log_{10}(Y(n,j)) \cos\big(\frac{(2j+1)l\pi}{2N_j}\big), \tag{6.4}$$

where $l = \{0, \ldots, N_c\}$ with typically $N_c = 12$ and the zeroth coefficient, $c(n,0)$ is ignored since it is the sum of the logarithm of the filter outputs and does therefore represent the intensity of the speech frame. This is normally not considered useful for

[1] An alternative to the mel-scale is the Bark-scale and Hamming-shaped filters have also been proposed instead of triangular shaped filters [Rabiner and Juang, 1993; Gold and Morgan, 2000; Young *et al.*, 2002].

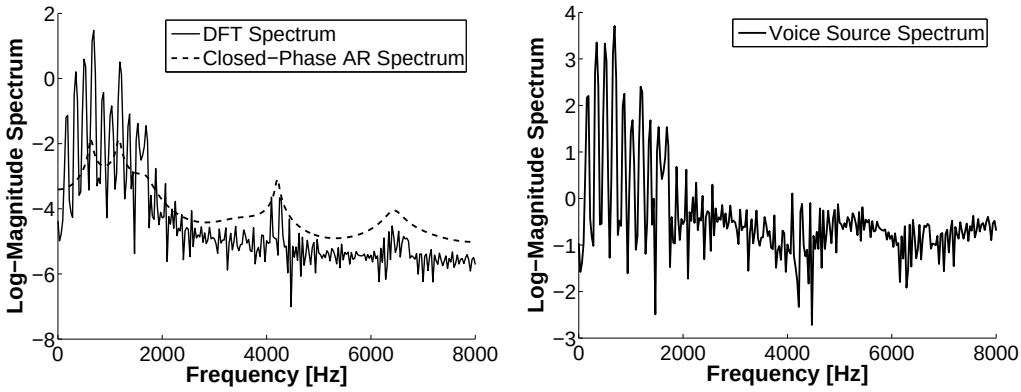

Figure 6.3: The DFT magnitude spectrum $|S(n,k)|$ of a voiced frame of speech and the closed phase autoregressive spectral envelope $|V(n,k)|$ (left). The voice source spectrum $\frac{|S(n,k)|}{|V(n,k)|}$ (right)

recognition since the intensity is dependent on the arbitrary volume of the speech so this coefficient is either omitted or replaced by the log-energy of the speech frame. However the change in intensity can prove useful and so the zeroth coefficient is included when dynamic features are added, as we describe below. Figure 6.4 shows the twelve mel-frequency cepstral coefficients corresponding to the spectra shown in Figure 6.3.

A problem with this feature set is that the variance of the coefficients is approximately inversely proportional to the square of the coefficient index. This does not present any theoretical difficulties in the subsequent modelling but to avoid numerical problems, such as variance floors set for the model parameters, the coefficients are scaled using cepstral liftering using the raised sine function of the form,

$$c'(n,l) = \left(1 + \frac{N_l}{2}\sin(\frac{\pi n}{N_l})\right) c(n,l) \tag{6.5}$$

where a typical value of N_l is 22.

Cepstral coefficients have many nice properties which has made them popular in various speech applications. Specifically, if the speech input is corrupted by a

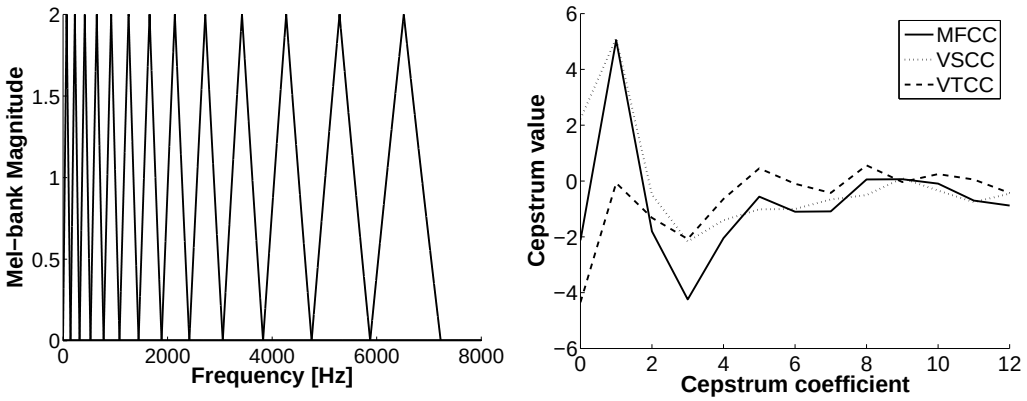

Figure 6.4: Filters from the mel-filterbank used in the mel-frequency cepstrum processing (left). Only every second filter is plotted for clarity. An example of mel-frequency, vocal tract and voice-source cepstrum for a single frame of voiced speech (right).

convolutional transmission channel, the effect is to multiply the speech spectrum by the transmission channel's transfer function. This is equivalent to addition in the log domain and if we assume that the mean of the clean speech is zero we can compensate for the channel effect by subtracting the mean from the coefficients [Mammone *et al.*, 1996],

$$\hat{c}(n,l) = c(n,l) - \mathcal{E}_n\{c(n,l)\} \tag{6.6}$$

where $\mathcal{E}_n$ denotes expectation taken over the frames n. This line of research has been extended to include a more elaborate filtering over the frames to suppress noise and channel effects. An example of this is RASTA processing which applies bandpass filters to the coefficients to emphasise critical bands of the speech [Hermansky and Morgan, 1994]. We did not employ these methods in this work since the data we work with contains little noise or channel effects.

Dynamic features are derived from the cepstral parameters by estimating their average change. The delta features proposed by Furui are now widely used in speech and speaker recognition and are derived using the regression formula [Furui,

1981]

$$\Delta c(n,l) = \frac{\sum_{i=-\kappa}^{\kappa} ic(n+i,l)}{\sum_{i=-\kappa}^{\kappa} i^2}. \tag{6.7}$$

These are appended to the cepstrum vector, so that the dimensionality is increased by a factor of two. Also typically, $\Delta c(n,0)$ is also included even though $c(n,0)$ is not and the acceleration coefficients, $\Delta\Delta c(n,l)$ are computed by applying the same regression formula, on $\Delta c(n,l)$ and the number of coefficients is increased to a total of 38 from 12.

6.2.2 Voice source cepstrum coefficients

The voice source cepstrum is derived by analysing the vocal tract response. The speech frame is represented by AR coefficients $\{a_{n,p}\}_{p=1}^{P}$, derived using multi-glottal closed phase analysis when the frame is from a voiced part of speech. We used the DYPSA algorithm, from Chapter 5, to identify the glottal closure instants and we assumed the closed phase to be 30% of the larynx-cycle. Figure 6.5 shows a voiced speech frame with the closed phases being identified. The AR modelling is performed on the entire frame when the speech is unvoiced since then the source can be modelled as white noise[2].

Figure 6.2 shows the steps involved in computing the voice source cepstrum. We can see that the mel-frequency cepstrum is derived as part of the process (on the right), while the vocal tract cepstrum is derived from the closed phase AR analysis (on the left). Since the assumption is that the vocal-tract cepstrum does not contain any contribution from the voice source we subtract it from the mel-frequency cepstrum which contains contribution from both the vocal tract and the voice source. This is equivalent to deconvolution in the time domain but bypasses problems such

[2]This is a first approximation for unvoiced speech. The emphasis of this work is on the voiced portion of speech so we have left the model for unvoiced speech unmodified. We discuss this further in Section 6.5.

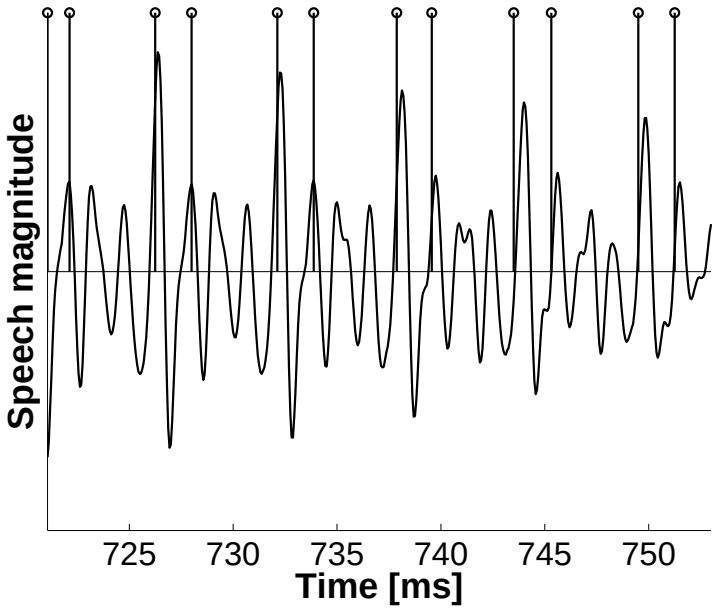

Figure 6.5: Frame of speech indicating the identified closed phases.

as phase-distortion. Furthermore, the information has been compacted into the cepstrum vectors through the filter integration and cepstral smoothing.

The vocal tract cepstrum is computed from the autoregressive spectrum envelope,

$$V(n,k) = \frac{\sigma_e}{1 - \sum_{p=1}^{P} a_{n,p} e^{-j2\pi kp/K_s}}. \tag{6.8}$$

The magnitude $|V(n,k)|$ is shown in Figure 6.3(left) together with the discrete Fourier magnitude spectrum $|S(n,k)|$.

We derive the vocal-tract cepstrum from $|V(n,k)|$ in the same manner we derive the mel-frequency cepstrum from $|S(n,k)|$. We apply a mel-filter bank to obtain the filter outputs,

$$Y_{vt}(n,j) = \sum_{k=0}^{K_s-1} |V(n,k)| Q_j(k) \tag{6.9}$$

and the vocal tract cepstrum coefficients are then computed as the cosine transform

$$c_{vt}(n,l) = \sum_{r=1}^{N_j-1} \log_{10}(Y_{vt}(n,j)) \cos\Big(\frac{(2r+1)l\pi}{2N_j}\Big) \tag{6.10}$$

and $l = \{0, \ldots, N_c\}$ with $N_c = 12$ as with the mel-frequency cepstrum. Liftering, cepstral mean subtraction and delta regression can be done in parallel with the mel-frequency cepstrum.

The voice source cepstrum coefficients are derived by subtracting the vocal tract cepstrum coefficients from the mel-frequency cepstrum coefficients,

$$c_{vs}(n,l) = c(n,l) - c_{vt}(n,l). \tag{6.11}$$

This is equivalent to deconvolution in the time domain and division in the spectrum domain. Figure 6.4 shows an example of the three coefficient vectors derived from a voiced frame of speech.

6.3 Speaker Classification

The purpose of speaker modelling is to determine a probability density function so that a likelihood[3] can be produced when an utterance from an unknown speaker is tested. If χ_i is the event that speaker i has spoken and an unknown utterance represented by the feature vector test sequence is $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_T\}$, then the aim is to calculate the conditional probability $Pr\{\chi_i|C\}$ so that we can determine the speaker of the unknown utterance by finding the maximum over i. This probability can be evaluated using Bayes's rule

$$Pr\{\chi_i|C\} = \frac{f_C(C|\chi_i)Pr\{\chi_i\}}{f_C(C)} \tag{6.12}$$

[3]Alternatively some similarity- or distance measure could be the objective of speaker modelling but here we concentrate on statistical pattern matching using likelihoods.

where $f_C(C)$ is the probability density function of all feature vector sequences. But since $f_C(C)$ is constant for all speakers and we assume that the prior probabilities $Pr\{\chi_i\}$ are equal, we can make the decision based on the conditional probability density function $f_C(C|\chi_i)$. We assume that the elements of C are independent observations so

$$f_C(C|\chi_i) = \prod_{j=1}^{T} f_C(\mathbf{c}_j|\chi_i) \tag{6.13}$$

but usually the log-likelihood,

$$\ell_i(C) = \log(f_C(C|\chi_i)) = \sum_{j=1}^{T} \log(f_C(\mathbf{c}_j|\chi_i)) \tag{6.14}$$

is preferred since it is computationally faster to process.

The probability density function we use is the Gaussian mixture model [Duda *et al.*, 2001],

$$f_C(\mathbf{c}|\chi_i) = \sum_{o=1}^{N_o} \frac{\nu_o^{(i)}}{\sqrt{(2\pi)^D |\mathbf{\Sigma}_o^{(i)}|}} e^{-\frac{1}{2}(\mathbf{c}-\mu_o^{(i)})^T (\mathbf{\Sigma}_o^{(i)})^{-1} (\mathbf{c}-\mu_o^{(i)})} \tag{6.15}$$

where $\nu_o^{(i)}$, $\mu_o^{(i)}$ and $\mathbf{\Sigma}_o^{(i)}$ are the weight (scalar), mean vector and covariance matrix for the o-th mixture component and i-th speaker and D is the dimension of the feature vector $\mathbf{c}$. In our case $D = N_c = 12$, the number of cepstrum coefficients.

We tested each classifier varying the number of mixture components using $N_o = \{2, 4, 8, 16, 32, 64\}$. We used diagonal covariance matrices and trained each model using the EM algorithm with 15 iterations [Dempster *et al.*, 1977; Moon, 1996]. Because the EM algorithm does local optimisation on the model parameters we retrained every model 5 times and chose the one achieving the highest likelihood on the training data . Each mixture model was initialised by setting all weights equal and the mean vectors were drawn randomly from a single Gaussian distribution of the training data. The covariance matrices were set to be equal to the covariance of

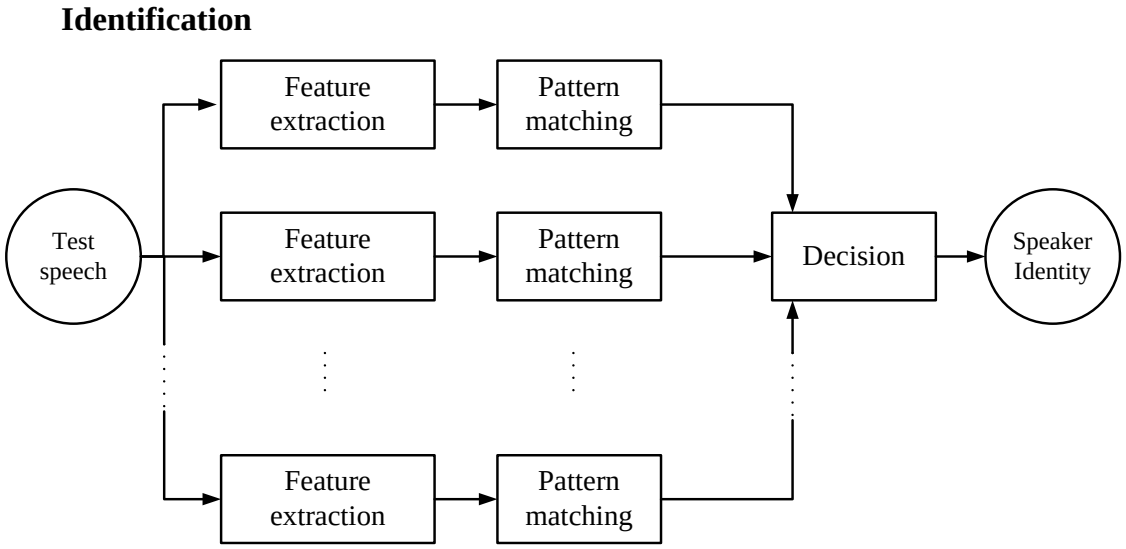

Figure 6.6: Many classifier combined by a more elaborate decision process.

the single Gaussian, divided by the model order, N_o.

6.3.1 Decision process

The decision process for closed-set speaker identification follows,

$$\hat{\imath} = \underset{i \in \{1,\ldots,N_S\}}{\arg\max} \log(f_C(C|\chi_i)) \tag{6.16}$$

where $\hat{\imath}$ is the index of the identified speaker in the set of N_S speakers. An open-set speaker identification setup would reject the utterance (represented by C) if $\log(f_c(C|\chi_i))$ would fall below a set threshold as indicated in Figure 1.7. We present speaker identification results for mel-frequency, vocal-tract and voice-source cepstrum coefficients, in Section 6.4.

We also present results for classification using two feature sets in combination [Kittler *et al.*, 1998]. We implemented this with decision fusion as indicated in Figure 6.6. Alteratively, the feature sets can either be modelled jointly or transformed into one feature set with linear discriminant approaches. We demonstrate the

effectiveness of the voice source by combining two classifiers, one based on the traditional mel-frequency cepstrum coefficients and the other on voice source cepstrum coefficients. We will base the decision on a weighted sum of the two log-likelihoods, $\ell_i(C_{VS})$ and $\ell_i(C_{MF})$ so Equation 6.16 becomes,

$$\hat{\imath} = \underset{i\in\{1,\ldots,N_s\}}{\arg\max} \; \theta\ell_i(C_{VS}) + (1-\theta)\ell_i(C_{MF}) \tag{6.17}$$

where $\theta \in [0,1]$ is a weight constant which is either undetermined, or derived using the training set or a validation set that is separate from the eventual test set [Duda *et al.*, 2001; Jang *et al.*, 1997; Bishop, 1995].

6.3.2 Baseline classifier

The baseline classifier was implemented using the mel-frequency cepstrum front-end processing using the same Gaussian mixture model classifier as the vocal-tract and voice source cepstrum classifiers used. The best results achieved by the baseline classifier on the TIMIT database using the first eight sentences for training and the remaining two for testing gave a test set misclassification rate of $\gamma = 1.51 \pm 0.34\%$. We consider this misclassification rate to be the main reference point in assessing performance of the new suggested classifiers.

6.4 Speaker Identification Results

Speaker identification using GMMs is presented in this section. Experimental results with respect to model order, test utterance duration and classifier combination are presented. We show that using only one feature set, the mel-frequency cepstrum coefficients perform better than voice source cepstrum coefficients. We then analyse the correlation between decisions made by the classifiers and show that, when applied

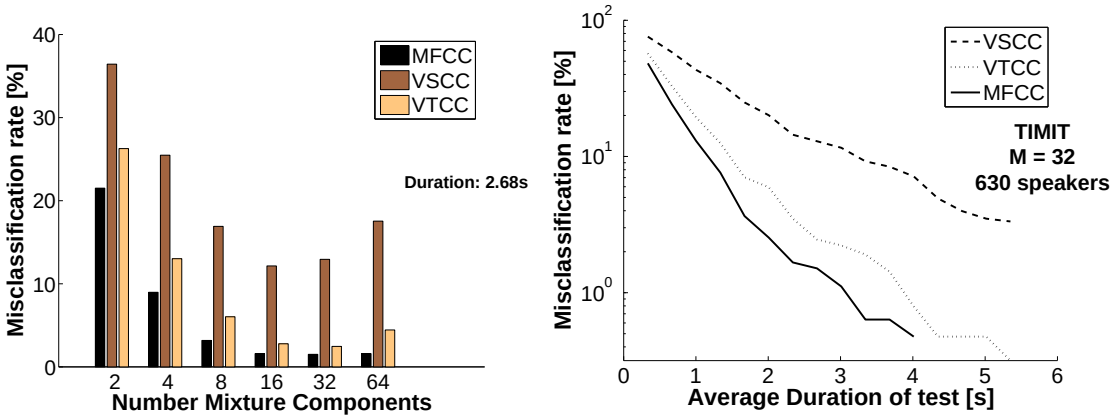

Figure 6.7: Test set misclassification rates for the three classifiers using different number of mixture components (left) and different test utterance duration (right). The misclassification for the MFCC classifier goes to zero for test utterances of duration greater than 4 s and is not shown on the logarithmic scale.

together, the combined result outperforms that of using only a single feature set.

6.4.1 Model order

Limited training data puts a cap on the model complexity we can employ. In the case of Gaussian mixture models this is represented by the number of mixture components and the shape of the covariance matrix. As the number of mixture components is increased more model parameters need to be estimated and overfitting is likely to occur. We implemented six speaker identification experiments using 2, 4, 8, 16, 32, and 64 mixture components. The results are shown in the bar-chart of Figure 6.7(left). The bars show the test set misclassification rates for each of the three classifiers, the baseline classifier using MFCC features, the VSCC features and the VTCC features. We found that the performance did not increase when the number of mixture components were increased beyond 32 but on the contrary in many of our experiments the performance decreased somewhat. This chart also displays the difference in performance between the three feature sets. We can see that the baseline MFCC features outperformed the VSCC and the VTCC

features. For the 32 mixture component case, the test set misclassification rate was $1.51 \pm 0.34\%$, $12.9 \pm 0.95\%$ and $2.46 \pm 0.44\%$ for the MFCC, VSCC and VTCC feature sets respectively.

6.4.2 Utterance duration

We varied the test utterance duration by splitting the two test sentences of each speaker up to eight parts. The test sentences were of an average duration of 2.68s so the shortest test duration was 0.334s and the longest was 5.35s on average. The plot in Figure 6.7(right) shows the test set misclassification rate of the MFCC, VSCC and VTCC classifiers using 32 mixture components. We see how the test set misclassification rate drops as the test duration is increased. When only 0.334s segments are used for testings the misclassification rates are $48.3 \pm 0.50\%$, $75.8 \pm 0.43\%$ and $57.1 \pm 0.49\%$ for the MFCC, VSCC and VTCC feature sets respectively, but for 5.35s , no errors were recorded for the MFCC feature set, the VSCC feature sets produced $3.33 \pm 0.72\%$ errors and the VTCC features misclassified $0.32 \pm 0.22\%$ of the speakers. The tests using longer utterances are less useful since so few misclassifications are produced and hence difficult to estimate the probability of error (or impossible in case of zero misclassification).

6.4.3 Comparing classifier decisions

The errors made by each classifier are compared in Table 6.1 for 32 mixtures and 2.68 s test utterances. The three tables compare the classification results of the three pairs of classifiers, the MFCC vs. VSCC, the MFCC vs. VTCC and the VTCC vs. VSCC classifier. We can see from the table on the left that the MFCC and VSCC classifiers agree to make a correct decision 86.3% of the time and also agree to make a wrong decision 0.8% of the time. The rest of the decisions are

Table 6.1: Cross tabulation of classifier decisions showing the higher correspondence between the MFCC and VSCC classifiers than that of the MFCC and VTCC classifiers.

(%) MFCC/	VSCC: correct	error
correct	86.3	12.1
error	0.7	0.8

(%) MFCC/	VTCC: correct	error
correct	96.9	1.6
error	0.6	0.9

(%) VTCC/	VSCC: correct	error
correct	85.5	12.1
error	1.6	0.9

disagreed upon with the VSCC classifier making a correct decision 0.7% of the time when the MFCC makes an error whereas it makes an error 12.1% of the time when the MFCC classifier makes a correct decision.

The tables demonstrate the correspondence between the decisions made by the classifiers and indicate what improvement may be possible by combining them. The interesting result is is that the decisions made by the MFCC and VSCC classifiers are less correlated than the decisions made by the MFCC and VTCC classifiers. This indicates that a combination between MFCC and VSCC features has more scope for improvement. Such results depend on the implementation details of the decision fusion.

6.4.4 Combination of classifiers

We have implemented a weighted likelihood sum decision fusion defined in Equation 6.17. There are various alternatives for combining classifier decisions [Kittler *et al.*, 1998], but since it is not the objective to study fusion techniques specially we have only relied on one fusion technique to draw conclusions on the quality of the extracted feature sets.

Figure 6.8 shows how the VSCC and MFCC feature sets (left) and VTCC and MFCC feature sets (right) are combined as a function of weight (θ in Equation 6.17). Each trace shows the test set misclassification rate when using a test token of given duration. The combination of VSCC with MFCC features works better than the

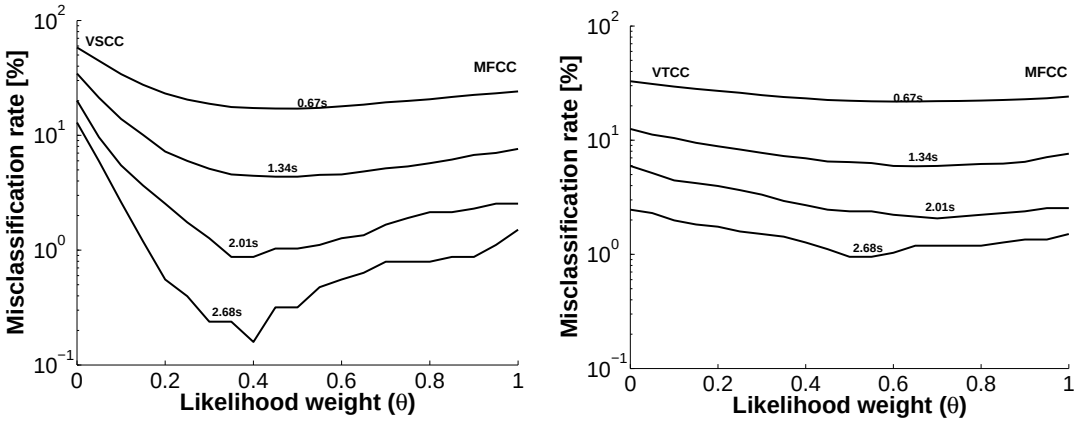

Figure 6.8: Combinations of the VSCC and MFCC classifiers (left) and VTCC and MFCC classifiers (right) for different test utterance duration.

Table 6.2: Misclassification rate using the three feature sets and combined classifiers. Each classifier uses 32 mixture components and the test utterance duration is 2.68s.

	Misclassification rate [%]	
Classifier	Test set γ	Gender bal. $\bar{\gamma}_{GB}$
MFCC	1.51 ± 0.34	1.52 ± 1.08
VSCC	12.94 ± 0.95	13.40 ± 0.03
VTCC	2.46 ± 0.44	2.43 ± 1.64
MFCC+VSCC	0.16 ± 0.11	0.11 ± 0.08
MFCC+VTCC	0.95 ± 0.27	0.98 ± 0.69
VTCC+VSCC	0.48 ± 0.19	0.49 ± 0.35

combination of VTCC and MFCC features with test set misclassification rate of $0.16 \pm 0.11\%$ reached when combining VSCC with MFCC but $0.95 \pm 0.27\%$ when combining VTCC with MFCC using 2.68s test tokens.

We have presented the three classifiers using MFCC, VSCC and VTCC feature sets and combined them using weighted likelihood sum. The summary of the results are presented in Table 6.2 where the test set and gender balanced misclassification rates are given for each classifier. The combination results are given using the best possible combination weight, but it must be noted that the values of these weights have not been optimised using a specific training or a validation set.

Table 6.3: Contingency table for comparing the MFCC and the combined MFCC and VSCC classifiers.

Classifiers	Correct	Error	Total
MFCC only	1241	19	1260
MFCC+VSCC	1258	2	1260
Total	2499	21	2520

We see that, of the three non-combined classifiers, best performance is achieved by the base classifier using the MFCC feature set but very significant improvements are attained by combining it with the VSCC feature set. The significance of the improvement between the MFCC classifier ($1.51 \pm 0.34\% = 19$ errors in 1260 tests) and the combined MFCC and VSCC classifier ($0.16 \pm 0.11\% = 2$ errors in 1260 tests) can be estimated using χ^2-test [Papoulis, 1991]. The contingency table is shown in Table 6.3. The value of the χ^2 distribution is 13.88 which is significant at the 0.1% level. Alternatively, the Fisher's Exact Test can be applied, since it is more appropriate to imbalanced tables and small values. The p-value for the onesided Fisher's Exact Test is $1.04 \cdot 10^{-4}$ which is therefore also significant at the 0.1% level [Conover, 1999].

6.4.5 Test utterance duration

We show the combination between all permutations of the MFCC, VSCC and VTCC classifiers in the plots of Figures 6.9 and 6.10 for test durations of 0.67s, 1.34s, 2.01s, and 2.68s. It can be seen from the plots that the combination of the MFCC and VTCC classifiers does not improve the misclassification rate of the MFCC classifier, whereas combining the VSCC classifier with the MFCC classifier results in a lower misclassification than that of the MFCC classifier. The lowest misclassification rate was $0.16 \pm 0.11\%$, achieved by combining the MFCC and VSCC classifiers weighting the VSCC likelihood with $\theta = 0.4$. The lowest misclassification rate achieved by the combination of VSCC and VTCC was $0.48 \pm 0.19\%$ also weighting the VSCC

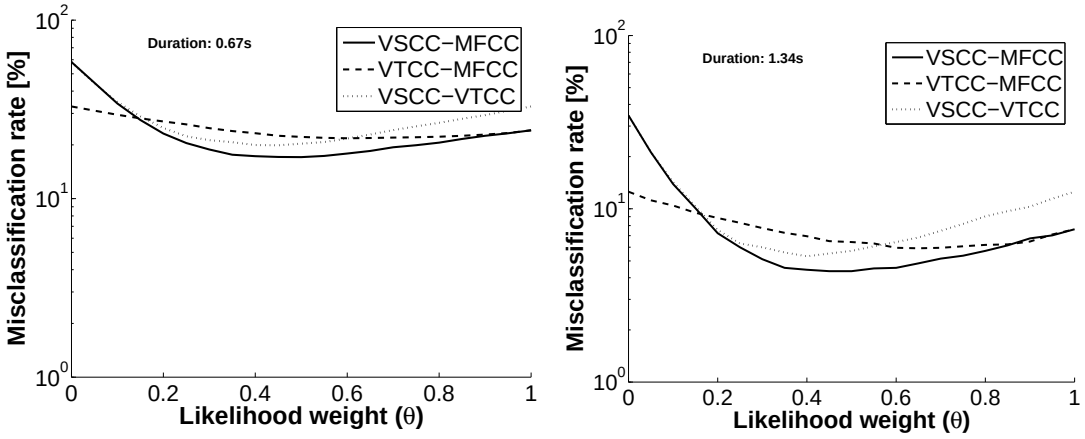

Figure 6.9: Comparison of the three combinations of classifiers using test utterances of 0.67s **(left) and** 1.34s **(right) duration.** $\theta = 0$ **corresponds to the feature set mentioned first in the legend box.**

likelihood with $\theta = 0.4$.

Figure 6.11 shows the misclassification rate of three classifiers, the VSCC, MFCC and the combined VSCC and MFCC classifier with the decision weight set to $\theta = 0.4$. We can see how the combination of classifiers is consistently better than the MFCC classifier for test utterances of short and long duration and that it improves faster than the MFCC classifier as the test utterance duration increases.

6.5 Concluding Remarks

We presented the voice source cepstrum coefficients and applied them to a closed-set speaker identification task with good results. We used the segmentation provided by the techniques developed in Chapter 5 and applied closed-set AR modelling, described in Chapter 3, to represent the vocal tract. By subtracting the cepstrum representation of the spectral envelope from the mel-frequency cepstrum we characterised the voice source with cepstrum features which we used for speaker identification. The results were very positive. They show that there is discriminative power in the voice source and misclassification rate was improved when combined

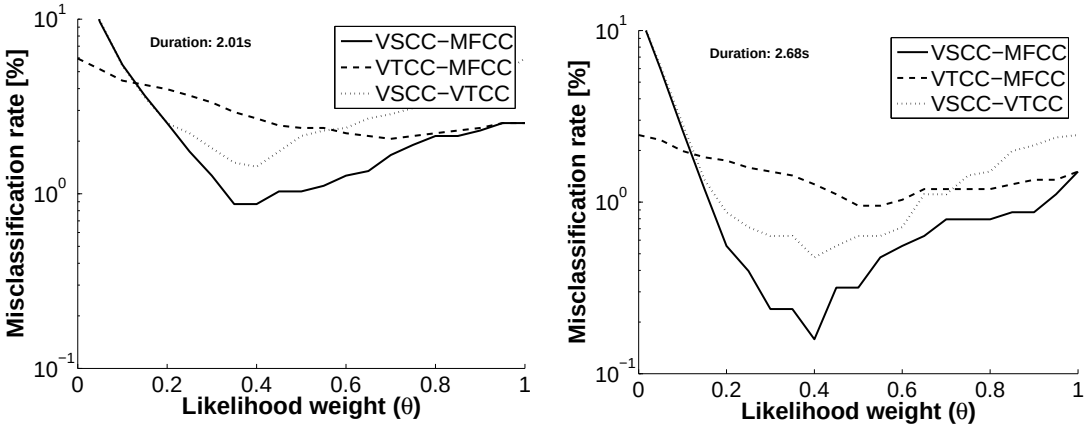

Figure 6.10: Comparison of the three combination of classifiers using test utterances of 2.01s (left) and 2.68s (right) duration. $\theta = 0$ corresponds to the feature set mentioned first in the legend box.

with the mel-frequency cepstrum representation of the speech.

We have not attempted to compare speaker identification performance of voice source coefficients relying on different configuration of DYPSA. The experiments in this chapter were presented to demonstrate the feature extraction design and voice source processing for speaker identification. The methodology and the database are not suited to distinguish between different configurations of DYPSA because the low misclassification rate for such tests would not be useful.

Voice source representation for speaker recognition has been used before combined with LPC cepstrum on a subset of TIMIT [Plumpe *et al.*, 1999]. The time-domain voice source signal is derived using inverse-filtering and the voice source features are derived as a composite of coarse and fine features. The coarse features are the equivalent to the LF parameters we described in Chapter 3 and the fine features are based on the difference between the parametric LF model and the estimated voice source signal. They achieved gender balanced misclassification rate of $\bar{\gamma}_{GB} = 28.64\%$[4] using only a 168 speaker subset of TIMIT whereas the gender balanced misclassification rate of the VSCC features, using the same subset, achieved

[4]Deduced from the third line of Table IV ("Source") in [Plumpe *et al.*, 1999].

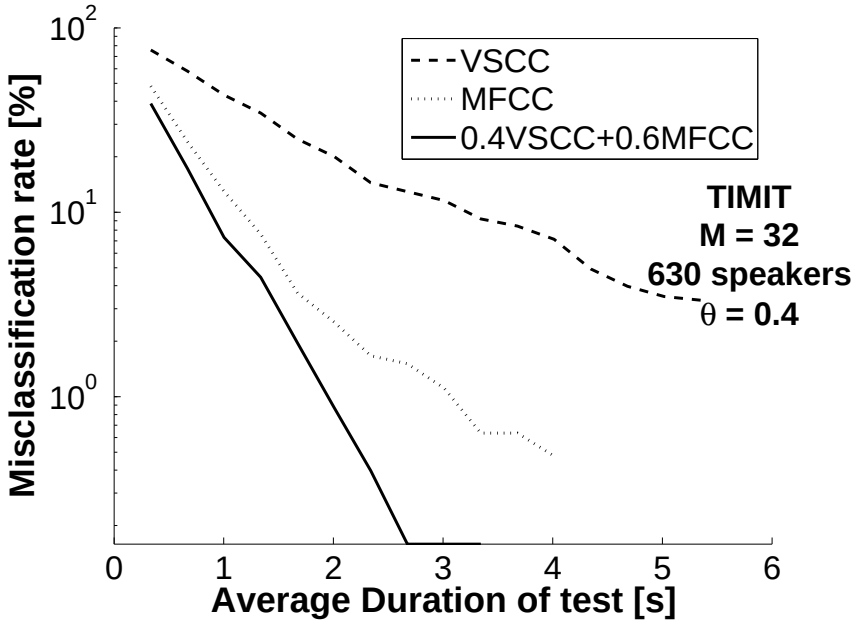

Figure 6.11: Misclassification rate as a function of test utterance duration for the VSCC, MFCC and the combined VSCC and MFCC classifiers with the decision weight $\theta = 0.4$.

$\bar{\gamma}_{GB} = 5.1 \pm 1.2\%$[5] This superior performance is due to the fact that the VSCC processing avoids inverse filtering and parametric modelling of the glottal flow. The coefficients are also less correlated because of the discrete cosine transform performed in their processing and their distribution is more Gaussian because of the logarithm taken in the frequency domain. Plumpe et. al. also presented results combining their voice source features with LPC cepstrum classifier but comparison of these with our combination results is unfair since LPC cepstrum is considered inferior to MFCCs.

[5]The results presented in this work was $\bar{\gamma}_{GB} = 12.87\%$ for the entire 630 speaker set.

Chapter 7

Discussion

THIS work has concentrated on front-end processing for speech or speaker recognition applications with the aim of forming feature parameters for speaker identification. Here we give a brief summary of the work, highlight the major conclusions that can be drawn from each chapter and suggest ideas for further work.

7.1 Summary

The dissertation covers the topic of voice source processing using a cepstrum approach to speaker identification. We gave an overview of background material in Chapter 1, where LPC analysis and speaker recognition were discussed. In Chapter 2, we described the speech corpora we used to evaluate the algorithms developed in this study and defined our performance assessment measures. The review in Chapter 3 focused on the voice source analysis, derived from the lossless tube model and made practicable by the autocorrelation and covariance LPC. Multi-glottal closed phase LPC analysis was presented and the extraction of the voice source signal was discussed. We discussed how the voice source signal is affected by small modelling

errors in the LPC analysis and, more severely, by low frequency phase distortion from the recording equipment.

Closed phase analysis depends on segmenting voiced speech into closed and open phases and glottal closure instant detection is crucial for this purpose. We tested GCI detection algorithms presented in the literature and discovered that better segmentation was needed. We investigated the properties of four types of group delay function in Chapter 4. There are many different ways of representing the group delay at any given point in time since the group delay is also a function of frequency. We studied the zero-frequency, average, energy-weighted, and energy-weighted-phase group delay functions all derived from different measures of the group delay. We based the DYPSA algorithm, presented in Chapter 5, on the energy-weighted group delay function and developed the dynamic programming cost function. The glottal closure instants detection performance was evaluated against three other published methods. We developed voice source cepstrum coefficients for speaker identification in Chapter 6. The parameters were based on closed-phase LPC analysis facilitated by the accurate GCI detection of DYPSA. The vocal tract cepstrum coefficients were extracted from the closed phase AR spectrum in parallel with mel-frequency cepstrum coefficients. Speaker identification experiments were performed using the three parameter sets, with the mel-frequency cepstrum coefficient classifier for the baseline. Furthermore the classifiers based on different parameter sets were combined to take advantage of their decision disagreement.

7.2 Conclusions

7.2.1 Voice production and glottal closures

The loss-less tube model of the voice production mechanism leads to a linear time-invariant model of speech. We have used closed-phase analysis to evaluate the vocal tract transfer function which facilitates a more sophisticated model of the voice source signal. Voice production analysis still remains an active research topic, but the contribution of this work has shown how linear models can be extended and refined using appropriate assumptions about the voice source. The DYPSA algorithm has been shown to produce an accurate detection of glottal closure instants compared to other algorithms. It relies on the group delay function and phase slope projection to generate GCI candidates for the dynamic programming that selects the best sequence of closure instants based on criteria that describe the physical properties of voiced speech. The implementation of an accurate GCI detector has lead to a more precise voiced speech production model.

7.2.2 Speaker identification

The speaker identification experiments in Chapter 6 demonstrated how much voice source features contribute to the classification performance. In extracting voice source cepstrum coefficients we had to derive vocal-tract cepstrum coefficients based on the AR spectrum of the vocal-tract filter. This allowed us to circumvent inverse filtering in the time domain and avoid the low-frequency phase distortion which is normally present in the time domain voice source signal. The voice source has been modelled using only few parameters compared to the parameters needed in modelling the vocal tract in LPC analysis [Rosenberg, 1971; Fant *et al.*, 1985]. It still contributes to speaker identification as we have shown in

this work. The errors produced by the voice source cepstrum classifier were different from the errors produced by the mel-frequency cepstrum classifier and combining their results improved recognition performance. For test utterances of average duration of 2.68 s, the combination of the voice source cepstrum coefficients with the mel-frequency cepstrum coefficient classifier reduced the misclassification rate from $1.51 \pm 0.34\%$ to $0.16 \pm 0.11\%$, which is more than 75% reduction in misclassification.

K.S.R. Murty and B. Yegnayaranana made the same arguments for using voice source features in speaker recognition as has been presented in this work [Murty and Yegnanarayana, 2006]. They claim that the residual phase contains speaker specific information which is complimentary to that of MFCCs. They also show how speaker verification can be improved by combining their voice source classifier with an MFCC classifier. Their results are not directly comparable with the results presented in this work, since their classifier design is based on neural networks, whereas here, GMM classifiers are used. Furthermore, they use the NIST-2003 corpus [NIS, 2003] but we use the TIMIT corpus.

M.D. Plumpe presented speaker identification experiments using voice source features which are comparable with the results presented in this work [Plumpe *et al.*, 1999]. On a 168 speaker subset of the TIMIT database, the VSCC features achieved a gender-balanced misclassification rate of $\bar{\gamma}_{GB} = 5.1 \pm 1.2\%$ whereas the previous published comparable results were $\bar{\gamma}_{GB} = 28.64\%$. The superior performance of the VSCC features is explained by the classification qualities of cepstrum features and that time-domain processing is avoided in our work.

7.3 Further Research

7.3.1 Improved DYPSA

Although the DYPSA algorithm has proven to be a very reliable way of detecting glottal closure instants in clean speech, it needs to be developed for noisy and/or reverberant speech; the effects of voice pathologies, such as whisper and creakiness, have to be examined; and its robustness in voiced/unvoiced/silence transitions needs to be improved. The DYPSA algorithm can be improved, for instance, by adapting its cost function to these scenarios. A voiced/unvoiced/silence estimator should be implemented and added to DYPSA, either explicitly as an additional method of pruning candidates, or implicitly, for example as a term in the dynamic programming cost function.

We can see in Table 5.1 that the identification rate is increased when using the group-delay projection technique for GCI candidate generation but the identification accuracy is decreased. This suggests that the accuracy of the projected candidates chosen by the dynamic programming could be improved. The first step would be to identify whether the projection technique gives a different offset to that of the non-projected GCIs. This could be corrected for by shifting all GCI candidates generated from projection by the estimated offset. The projection technique could also be altered so that the point of inflection between the maxima and minima would be chosen instead of the midpoint to project from and the gradient of the projection line could be altered as well.

7.3.2 Closed phase analysis and feature extraction

The solution for closed phase analysis presented in this work only extends to that of finding the beginning of the closed phases in voiced speech. The DYPSA algo-

rithm identifies GCIs at higher rate than any of the other tested algorithms with good accuracy. How this accuracy and the obtained identification rate affects the closed phase analysis has yet to be evaluated. Furthermore, the identification of glottal opening instants remains challenging. The waveforms of Figure 4.8, appear to indicate the possibility of using $d'_{EP}(n)$ to detect glottal GOIs in addition to the GCIs. However, in many other speakers, the GOI excitations are very small and so the reliable identification of GOIs remains a very challenging task with, as yet, little reported work in the literature.

The knowledge of the exact timing of glottal closures opens up many possibilities for speech feature extraction for speech or speaker recognition. We have employed this technique for closed-phase LPC analysis but these could be used for other approaches such as subband analysis feature extraction. For example the ensemble interval histogram [Ghitza, 1994] and subband spectral centroids histogram [Gajic and Paliwal, 2006] feature extraction approaches could be improved with the knowledge of GCIs. Glottal closure instants can also be used to characterise pitch and prosody features [Shriberg *et al.*, 2005].

We have made the somewhat coarse assumption that the voice source is white for unvoiced speech. Our voice source parameters have therefore become the difference between mel-frequency cepstrum coefficients based on a direct spectrum estimation of the speech frame and the vocal tract model spectrum. This assumption is good for some unvoiced sound but inappropriate for others. Epoch analysis of the unvoiced portion of the speech signal, similar to that of DYPSA, would provide cues that enabled the distinction between the source and the tract when constructing feature vectors.

7.3.3 Classifier considerations

In combining two feature sets we used a weighted sum of the log-likelihoods obtained from the two feature set classifiers. Systematic assessment of classifier combination strategies [Kittler *et al.*, 1998] has not been performed in this work but could become the focus of a future study. Another approach for combining different feature set is through linear discriminant analysis or more recently developed heteroscedastic linear discriminant analysis [Burget, 2004].

In this work, the classifier of choice was the Gaussian mixture model classifier. The EM algorithm was used to derive a speaker's GMM using training data from that speaker and the likelihood of the test utterance was derived for all the speaker models. Other classifier approaches such as the adapted GMM classifier [Reynolds *et al.*, 2000] and the support vector machine, e.g. [Campbell *et al.*, 2006], should be tested on the proposed feature sets.

London, March 2007.

Bibliography

[Abberton *et al.*, 1989] E. R. M. Abberton, D. M. Howard, and A. J. Fourcin. Laryngographic Assessment of Normal Voice: A Tutorial. *Clinical Linguistics and Phonetics*, 3(3):263–296, 1989.

[Adami and Hermansky, 2003] A. G. Adami and H. Hermansky. Segmentation of Speech for Speaker and Language Recognition. *Eurospeech, Geneva*, pages 841–844, 2003.

[Adami *et al.*, 2003] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey. Modelling Prosodic Dynamics for Speaker Recognition. *International Conference on Acoustics, Speech, and Signal Processing*, 4:788–791, 2003.

[Akande and Murphy, 2005] O. O. Akande and P. J. Murphy. Estimation of the Vocal Tract Transfer Function with Application to Glottal Wave Analysis. *Speech Communication*, 46(1):15–36, May 2005.

[Alku and Backstrom, 2002] P. Alku and T. Backstrom. Normalized Apmlitude Quotient for Parametrization of the Glottal Flow. *Journal of the Acoustical Society of America*, 112(2):701–710, August 2002.

[Ananthapadmanabha and Fant, 1982] T. V. Ananthapadmanabha and G. Fant. Calculations of True Glottal Volume-Velocity and its Components. *Speech Communication*, 1:167–184, 1982.

[Ananthapadmanabha and Yegnanarayana, 1975] T. V. Ananthapadmanabha and B. Yegnanarayana. Epoch Extraction of Voiced Speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(6):562–570, December 1975.

[Ananthapadmanabha and Yegnanarayana, 1979] T. V. Ananthapadmanabha and B. Yegnanarayana. Epoch Extraction from Linear Prediction Residual for Identification of Closed Glottis Interval. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-27(4):309–320, August 1979.

[Atal and Hanauer, 1971] B.S. Atal and S.L. Hanauer. Speech Analysis and Synthesis by Linear Prediction of the Speech Wave. *Journal of the Acoustical Society of America*, 50(2B):637–655, August 1971.

[Atal and Rabiner, 1976] B.S. Atal and L.R. Rabiner. A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Application to Speech Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-24(3):261–272, June 1976.

[Atal, 1972] B.S. Atal. Automatic Speaker Recognition Based on Pitch Countours. *Journal of the Acoustical Society of America*, 52(6):1687–1697, July 1972.

[Backstrom *et al.*, 2002] T. Backstrom, P. Alku, and E. Vilkman. Time-Domain Parameterization of the Closing Phase of Glottal Airflow Waveform from Voices Over a Large Intensity Range. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 10(3):186–194, March 2002.

[Ben *et al.*, 2002] M. Ben, R. Blouet, and F. Bimbot. A Monte Carlo method for score normalization in automatic speaker verification using Kullback-Leibler distances. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 689–692, 2002.

[Beranek, 1954] L. L. Beranek. *Acoustics*. McGraw-Hill, New York, 1954.

[Bernardo and Smith, 1996] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, 1996.

[Bimbot *et al.*, 2004] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, and D. A. Reynolds. A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Applied Signal Processing*, 4:430–451, 2004.

[Birkholz *et al.*, 2006] P. Birkholz, D. Jackel, and B. J. Kroger. Construction and Control of a Three-Dimensional Vocal Tract Model. In *International Conference on Acoustics, Speech, and Signal Processing*, pages I–873–I876, May 2006.

[Bishop, 1995] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[Botinis *et al.*, 2001] A. Botinis, B. Granström, and B. Möbius. Developments and Paradigms in Intonation Research. *Speech Communication*, 33:263–296, 2001.

[Bourlard *et al.*, 1998] H. Bourlard, , B. Gold, and N. Morgan. Speaker Verification – A Quick Overview. 1998. A chapter in Gold and Morgan pages 521–530, 2000.

[Bouzid and Ellouze, 2004] A. Bouzid and N. Ellouze. Glottal Opening Instant Detection from Speech Signal. In *European Signal Processing Conference EUSIPCO*, pages 729–732, Vienna, September 2004.

[Braverman, 1962] D. Braverman. Learning Filters for Optimum Pattern Recognition. *IEEE Transactions on Information Theory*, 8:280–285, 1962.

[Brookes and Chan, 1994] D. M. Brookes and D. S. F. Chan. Speaker Characteristics from a Glottal Airflow Model Using Robust Inverse Filtering. *Proceedings of the Institute of Acoustics*, 16(5):501–508, 1994.

[Brookes and Loke, 1999] D. M. Brookes and L. P. Loke. Modelling Energy Flow in the Vocal Tract with Applications to Glottal Closure and Opening Detection. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 213–216, March 1999.

[Brookes *et al.*, 2006] D. M. Brookes, P. A. Naylor, and J. Gudnason. A Quantatative Assessment of Group Delay Methods for Identifying Glottal Closures in Voiced Speech. *IEEE Transaction on Speech and Audio Processing*, 14(3):456–466, May 2006.

[Burget, 2004] L. Burget. Combination of Speech Features Using Smoothed Heteroscedastic Linear Discriminant Analysis. In *Proc. 8th International Conference on Spoken Language Processing*, pages 2549–2552, 2004.

[Campbell *et al.*, 2003] J. P. Campbell, D. A. Reynolds, and R. B. Dunn. Fusing High- and Low-Level Features for Speaker Recognition. In *Eurospeech*, pages 2665–2668, Geneva, Switzerland, 2003.

[Campbell *et al.*, 2006] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff. SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 97–100, May 2006.

[Campbell, 1997] J. P. Campbell, Jr. Speaker Recognition: A Tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, September 1997.

[Campbell, 2002] W. M. Campbell. Generalized Linear Discriminant Sequence Kernels for Speaker Recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 161–164, 2002.

[Campbell, 2003] W. M. Campbell. A SVM/HMM System for Speaker Recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 209–212, Apr 2003.

[Carey *et al.*, 1991] M. J. Carey, E. S. Parris, and J. S. Bridle. A speaker Verification System using Alpha-Nets. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 397–400, Apr 1991.

[Carey *et al.*, 1996] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, and S. Bennet. Robost Prosodic Features for Speaker Identification. *International Conference on Spoken Language, Proceedings*, 3:1800–1803, October 1996.

[Chan and Brookes, 1989] D. S. F. Chan and D. M. Brookes. Variability of Excitation Parameters Derived from Robust Closed Phase Glottal Inverse Filtering. *European Conf. on Speech Communication and Technology*, 33(1), September 1989.

[Chan *et al.*, 1995] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouronopoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeilieger. EUROM - A Spoken Language Resource for the EU. In *European Conference on Speech Communication and Speech Technology*, pages 867–870, September 1995.

[Che *et al.*, 1996] C. Che, Q. Lin, and D. Yuk. An HMM Approach to Text-Prompted Speaker Verification. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 673–676, May 1996.

[Chen and Soong, 1994] J. K. Chen and F. K. Soong. An N-best Candidates-Based Discriminative Training for Speech Recognition Applications. *IEEE Transaction on Speech and Audio Processing*, 2:206–216, January 1994.

[Cheng and O'Shaughnessy, 1989] Y. M. Cheng and D. O'Shaughnessy. Automatic and Reliable Estimation of Glottal Closure Instant and Period. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(12):1805 – 1815, December 1989.

[Childers *et al.*, 1980] D. Childers, J. Mott, and G. Moore. Automatic Parameterization of Vocal Cord Motion from Ultra High Sspeed Films. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 65–68, 1980.

[Childers *et al.*, 1995] D.G. Childers, J.C. Principe, and Y.T. Ting. Adaptive WRLS-VFF for Speech Analysis. *IEEE Transaction on Speech and Audio Processing*, 3:209–213, May 1995.

[Childers, 1992] D. G. Childers. Detection of Laryngeal Function Using Speech and Electroglottographic Data. *IEEE Transactions on Biomedical Engineering*, 39(1):19–25, January 1992.

[Coker, 1976] C. H. Coker. A Model of Articulatory Dynamics and Control. *Proceedings of the IEEE*, 64(4):452–460, 1976.

[Conover, 1999] W. J. Conover. *Practical Nonparametric Statistics*. Wiley Series in Probability & Mathematical Statistics. John Wiley and Sons, 3 edition, 1999.

[Cranen and Boves, 1987] B. Cranen and L. Boves. On Subglottal Formant Analysis. *Journal of the Acoustical Society of America*, 81:734–746, 1987.

[Cranen and Boves, 1988] B. Cranen and L. Boves. On the Measurement of Glottal Flow. *Journal of the Acoustical Society of America*, 84(3):888–900, September 1988.

[Cummings and Clements, 1995] K. E. Cummings and M. A. Clements. Glottal Models for Digital Speech Processing - A Historical Survey and New Results. *Digital Signal Processing*, 5(1):21–42, 1995.

[Davis and Mermelstein, 1980] S. B. Davis and P. Mermelstein. Comparison of Parametric Repreesentations for Monosyllabic Word Recognition in Continously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, August 1980.

[Deller *et al.*, 1993] J.R. Deller, J.G. Proakis, and J.H.L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan, 1993.

[Dempster *et al.*, 1977] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[Dimitriadis *et al.*, 2005] D. Dimitriadis, P. Maragos, and A. Potamianos. Robust AM-FM Features for Speech Recognition. *IEEE Signal Processing Letters*, 12(9):621–624, September 2005.

[Ding and Kasuya, 1996] W. Ding and H. Kasuya. A Novel Approach to the Estimation of Voice Source and Vocal Tract Parameters from Speech Signals. In *Fourth Intl Conf on Spoken Language ICSLP 96. Proceedings.*, volume 2, pages 1257–1260, 1996.

[Doddington, 1985] G. R. Doddington. Speaker Recognition - Identifying People by their Voices. *Proceedings of the IEEE*, 73(11):1651–1664, November 1985.

[Duda *et al.*, 2001] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, 2 edition, 2001.

[Dunn, 1950] H. K. Dunn. The Calculation of Vowel Resonances, and an Electrical Vocal Tract. *Journal of the Acoustical Society of America*, 22(12):740–753, 1950.

[Dunn, 1961] H. K. Dunn. Methods of Measuring Vowel Formant Bandwidths. *Journal of the Acoustical Society of America*, 33(12):1737–1746, 1961.

[Durbin, 1959] J. Durbin. Efficient Estimation of Parameters in Moving-Average Models. *Biometrika*, 46(3-4):306–316, 1959.

[Ezzaidi *et al.*, 2001a] H. Ezzaidi, J. Rouat, and D. O'Shaughnessy. Combining Pitch and MFCC for Speaker Recognition Systems. In *A Speaker Odyssey, the Speaker Recognition Workshop, an ISCA Tutorial and Research Workshop (ITRW) on Speaker Recognition*, pages 1036–1041, June 2001.

[Ezzaidi *et al.*, 2001b] H. Ezzaidi, J. Rouat, and D. O'Shaughnessy. Towards Combining Pitch and MFCC for Speaker Identification Systems. In *Eurospeech*, pages 2825–2828, September 2001.

[Fant and Liljencrants, 1979] G. Fant and J. Liljencrants. Perception of Vowels with Truncated Intraperiod Decay Envelopes. In *STL-QPSR 1*, pages 79–84. Department of Speech, Music and Hearing, KTH, http://www.speech.kth.se, 1979.

[Fant *et al.*, 1985] G. Fant, J. Liljencrants, and Q. Lin. A Four-Parameter Model of Glottal Flow. In *STL-QPSR*, pages 1–13. Department of Speech, Music and Hearing, KTH, http://www.speech.kth.se, 1985.

[Fant, 1960] G. Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, The Netherlands, 1960.

[Fant, 1968] G. Fant. Analysis and Synthesis of Speech Processes. *Manual of Phonetics*, Chapter 8:173–276, 1968.

[Fant, 1979] G. Fant. Vocal Source Analysis, a Progress Report. In *STL-QPSR 3-4*, pages 31–53. Department of Speech, Music and Hearing, KTH, http://www.speech.kth.se, 1979.

[Fisher *et al.*, 1986] W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall. The DARPA Speech Recognition Research Database: Specifications and Status. *Proc. DARPA Workshop on Speech Recognition*, pages 93–99, February 1986.

[Flanagan, 1972] J. L. Flanagan. *Speech Analysis, Synthesis, and Prception.* Springer-Verlag, New York, 2 edition, 1972.

[Fu and Murphy, 2006] Qiang Fu and P. Murphy. Robust Glottal Source Estimation Based on Joint Source-Filter Model Optimization. *IEEE Transaction on Speech and Audio Processing*, 14:492–501, Mar 2006.

[Fujisaki and Ljungqvist, 1987] 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 *International Conference on Acoustics, Speech, and Signal Processing*, volume 12, pages 637–640, 1987.

[Furui, 1981] S. Furui. Comparison of Speaker Recognition Methods using Statistical Features and Dynamic Features. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29:342–350, 1981.

[Furui, 1997] S. Furui. Recent Advances in Speaker Recognition. *Pattern Recognition Letters*, 18:859–872, 1997.

[Gajic and Paliwal, 2006] B. Gajic and K. K. Paliwal. Robust Speech Recognition in Noisy Environments Based on Subband Spectral Centroid Histograms. *IEEE Transaction on Speech and Audio Processing*, 14(2):600–608, 2006 2006.

[Gauvain and Lee, 1994] J.-L. Gauvain and Chin-Hui Lee. Maximum a Posteriori Rstimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transaction on Speech and Audio Processing*, 2:291–298, Apr 1994.

[Ghitza, 1994] O. Ghitza. Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Transaction on Speech and Audio Processing*, 2:115–132, Jan 1994.

[Gibbon *et al.*, 1998] D. Gibbon, R. Moore, and R. Winski. *Handbook of Standard and Resources for Spoken Language Systems*, volume 3: Spoken Language System Assessment. Mouton de Gruyter, Germany, 1998.

[Gish and Schmidt, 1994] H. Gish and M. Schmidt. Text-Independent Speaker Identification. *IEEE Signal Processing Magazine*, 11:18–32, 1994.

[Gold and Morgan, 2000] B. Gold and N. Morgan. *Speech and Audio Signal Processing.* John Wiley and Sons, Inc., New York, 2000.

[Hamon *et al.*, 1989] C. Hamon, E. Moulines, and F. Charentier. A Diphone Synthesis System Based on Time-Domain Prosodic Modifications of Speech. *International Conference on Acoustics, Speech, and Signal Processing*, pages 238–241, 1989.

[Heck and Weintraub, 1997] L. P. Heck and M. Weintraub. Handset-Dependent Background Models for Robust Text-Independent Speaker Recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1071–1074, Apr 1997.

[Hedelin, 1984] P. Hedelin. A Glottal LPC-Vocoder. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 21–24, 1984.

[Hermansky and Morgan, 1994] H. Hermansky and N. Morgan. RASTA Processing of Speech. *IEEE Transaction on Speech and Audio Processing*, 2(4):578–589, October 1994.

[Hermansky, 1990] H. Hermansky. Perceptual Linear Predictive (PLP) Analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[Hess and Indefrey, 1984] W. Hess and H. Indefrey. Accurate Pitch Determination of Speech Signals by Means of a Laryngograph. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 73–76, 1984.

[Higgins *et al.*, 1991] A. Higgins, L. Bahler, and J. Porter. Speaker Verification Using Randomized Phrase Prompting. *Digital Signal Processing*, 1:89–106, 1991.

[Holmes, 1975] J. N. Holmes. Low-Frequency Phase Distortion of Speech Recordings. *Journal of the Acoustical Society of America*, 58(3):747–749, September 1975.

[Huckvale, 2000] M. Huckvale. Speech Filing System: Tools for Speech Research. Online, 2000. http://www.phon.ucl.ac.uk/resource/sfs/.

[Hunt, 1978] M. J. Hunt. Automatic Correction of Low-Frequency Phase Distortion in Analogue Magnetic Recordings. *Acoustics Letters*, 2:6–10, 1978.

[Jabloun and Enis Cetin, 1999] F. Jabloun and A. Enis Cetin. The Teager Energy Based Feature Parameters for Robust Speech Recognition in Car Noise. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 273–276, Mar 1999.

[Jang *et al.*, 1997] J.-S. R. Jang, C.-T. Sun, and E. Mizutani. *Neuro-Fuzzy and Soft Computing*. Prentice Hall, 1997.

[Jankowski Jr. *et al.*, 1995] C. R. Jankowski Jr., T.F. Quatieri, and D.A. Reynolds. Measuring Fine Structure in Speech: Application to Speaker Identification. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 325–328, May 1995.

[Jin *et al.*, 2003] Q. Jin, J. Navratil, D.A. Reynolds, J.P. Campbell, W.D. Andrews, and J.S. Abramson. Combining Cross-Stream and Time Dimensions in Phonetic Speaker Recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 800–803, Apr 2003.

[Jurafsky and Martin, 2000] D. S. Jurafsky and J. H. Martin. *Speech and Language Processing*. Prentice-Hall, New Jersey, 2000.

[Karlsson, 1985] I. Karlsson. Glottal Wave Forms for Normal Female Speakers. *STL-QPSR*, 26(1):31–36, 1985.

[Karlsson, 1988] I. Karlsson. Glottal Waveform Parameters for Different Speaker Types. *STL-QPSR*, 29(2-3):61–67, 1988.

[Kay, 1988] S. Kay. *Modern Spectral Estimation*. Prentice Hall, 1 edition, 1988.

[Kelly and Lochbaum, 1962] J. L. Kelly and C. C. Lochbaum. Speech Synthesis. In *Proceedings of the Fourth International Congress on Acoustics*, volume G42, pages 1–4, 1962.

[Kenny and Dumouchel, 2004] P. Kenny and P. Dumouchel. Experiments in Speaker Verification using Factor Analysis Likelihood Ratios. In *Proceedings of Odyssey04 - Speaker and Language Recognition Workshop*, Toledo Spain, 2004.

[Kenny *et al.*, 2006a] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Improvements in Factor Analysis Based Speaker Verification. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, May 2006.

[Kenny *et al.*, 2006b] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Joint Factor Analysis versus Eigenchannels in Speaker Recognition. Submitted to IEEE Transaction on Speech and Audio Processing, 2006.

[Kittler *et al.*, 1998] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:226–239, Mar 1998.

[Klatt and Klatt, 1990] D. H. Klatt and L. C. Klatt. Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Talkers. *The Journal of the Acoustical Society of America*, 87(2):820–857, February 1990.

[Kleijn and Paliwal, 1995] W.B. Kleijn and K.K. Paliwal, editors. *Speech Coding and Synthesis*. Elsevier, Amsterdam, 1995.

[Klusacek *et al.*, 2003] D. Klusacek, J. Navratil, D.A. Reynolds, and J.P. Campbell. Conditional Pronunciation Modeling in Speaker Detection. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 804–807, Apr 2003.

[Kounoudes *et al.*, 2002a] A. Kounoudes, P. A. Naylor, and M. Brookes. Automatic Epoch Extraction for Closed Phase Analysis of Speech. In *Digital Signal Processing, 14th International Conference on*, volume 2, pages 979 – 983, July 2002.

[Kounoudes *et al.*, 2002b] A. Kounoudes, P. A. Naylor, and M. Brookes. The DYPSA Algorithm for Estimation of Glottal Closure Instants in Voiced Speech. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I349–I352, May 2002.

[Kounoudes, 2001] A. Kounoudes. *Epoch Estimation for Closed Phase Analysis of Speech*. PhD thesis, University of London Imperial College, 2001.

[Krishnamurthy and Childers, 1986] A. K. Krishnamurthy and D. G. Childers. Two-Channel Speech Analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):730 – 743, August 1986.

[Kumar and Mullick, 1996] A. Kumar and S.K. Mullick. Nonlinear dynamical analysis of speech. *Journal of the Acoustical Society of America*, 100(1):615–629, July 1996.

[Larar *et al.*, 1985] J.N. Larar, Y.A. Alsaka, and D.G. Childers. Variability in Closed Phase Analysis of Speech. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 29, pages 1089–1093, 1985.

[Lau *et al.*, 2004] Y. W. Lau, M. Wagner, and D. Tran. Vulnerability of Speaker Verification to Voice Mimicking. In *Proceedings of International Symposium on Intelligent Multimedia, Video and Speech Processing*, pages 145–148, October 2004.

[Li and Porter, 1988] K. P. Li and J. E. Porter. Normalizations and selection of speech segments for speaker recognition scoring. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 595–598, Apr 1988.

[Lindqvist-Gauffin, 1970] J. Lindqvist-Gauffin. The voice source studied by means of inverse filtering. *STL-QPSR*, 11(1):3–9, 1970.

[Lindsey *et al.*, 1987] G. Lindsey, A. Breen, and S. Nevard. SPAR'S Archivable Actual-Word Databases. Technical report, University College London, June 1987.

[Lobo, 2001] A.P. Lobo. Glottal Flow Derivative Modeling with the Wavelet Smoothed Excitation. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 861–864, May 2001.

[Lu and Smith, 1999] Hui-Ling Lu and J.O. Smith, III. Joint Estimation of Vocal Tract Filter and Glottal Source Waveform via Convex Optimization. In *1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 79–82, October 1999.

[Lucey *et al.*, 2005] S. Lucey, T. Chen, S. Sridharan, and V. Chandran. Integration Strategies for Audio-Visual Speech Processing: Applied to Text-Dependent Speaker Recognition. *IEEE Transactions on Multimedia*, 7:495–506, Jun 2005.

[Lummis, 1973] R. C. Lummis. Speaker Verification by Computer using Speech Intensity for Temporal Registration. *IEEE Transactions on Audio and Electroacoustics*, 21:80–89, 1973.

[Ma *et al.*, 1994] C. Ma, Y. Kamp, and L. F. Willems. A Frobenius Norm Approach to Glottal Closure Detection from the Speech Signal. *IEEE Transaction on Speech and Audio Processing*, 2(2):258–264, April 1994.

[Mammone *et al.*, 1996] R. J. Mammone, X. Zhang, and R. P. Ramachandran. Robust Speaker Recognition: A Feature-based Approach. *IEEE Signal Processing Magazine*, 13(5):58–71, September 1996.

[Markel and Gray, 1976] J. D. Markel and A. H. Gray, Jr. *Linear Prediction of Speech.* Springer-Verlag, 1976.

[Markel, 1973] J. D. Markel. Application of a Digital Inverse Filter for Automatic Formant and F_0, Analysis. *IEEE Transactions on Audio and Electroacoustics*, AU-21(3):154–160, June 1973.

[Matsui and Furui, 1993] T. Matsui and S. Furui. Concatenated Phoneme Models for Text-Variable Speaker Recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 391–394, Apr 1993.

[McKenna, 2001] J. G. McKenna. Automatic Glottal Closed-Phase Location and Analyis by Kalman Filtering. In *4th ISCA Tutorial and Research Workshop on Speech Synthesis*, August 2001.

[McLaughlin and Reynolds, 2002] J. McLaughlin and D.A. Reynolds. Speaker Detection and Tracking for Telephone Transactions. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 129–132, 2002.

[Mermelstein, 1973] P. Mermelstein. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53(4):1070–1082, 1973.

[Milenkovic, 1986] P. Milenkovic. Glottal Inverse Filtering by Joint Estimation of an AR System with a Linear Input Model. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34:28–42, Feb 1986.

[Miller, 1959] R. L. Miller. Nature of the Vocal Chord Wave. *Journal of the Acoustical Society of America*, 31:667–677, 1959.

[Moon, 1996] T.K. Moon. The Expectation-Maximization Algorithm. *IEEE Signal Processing Magazine*, 13:47–60, Nov 1996.

[Morgan *et al.*, 2005] N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cetin, H. Bourlard, and M. Athineos. Pushing the Envelope - Aside. *Signal Processing Magazine*, 22(5):81–88, September 2005.

[Morse and Ingard, 1968] P. M. Morse and K. U. Ingard. *Theoretical Acoustics*. International Series in Pure and Pllied Physics. McGraw Hill, New York, 1968.

[Murthy and Yegnanarayana, 1999] P.S. Murthy and B. Yegnanarayana. Robustness of Group-Delay-Based Method for Extraction of Significant Instants of Excitation from Speech Signals. *IEEE Transaction on Speech and Audio Processing*, 7(6):609–619, November 1999.

[Murty and Yegnanarayana, 2006] K.S.R. Murty and B. Yegnanarayana. Combining Evidence from Residual Phase and MFCC Features for Speaker Recognition. *IEEE Signal Processing Letters*, 13:52–55, Jan 2006.

[Naik, 1990] J. M. Naik. Speaker Verification: A Tutorial. *IEEE Communications Magazine*, 28(1):42–48, January 1990.

[Navarro-Mesa *et al.*, 2001] J.L. Navarro-Mesa, E. Lleida-Solano, and A. Moreno-Bilbao. A New Method for Epoch Detection Based on the Cohen's Class of Time Frequency Representations. *IEEE Signal Processing Letters*, 8(8):225–227, August 2001.

[Navratil *et al.*, 2003] J. Navratil, Q. Jin, W.D. Andrews, and J.P. Campbell. Phonetic Speaker Recognition using Maximum-Likelihood Binary-Decision Tree Models. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 796–799, Apr 2003.

[Naylor *et al.*, 2007] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes. Estimation of Glottal Closure Instants in Voiced Speech using the DYPSA Algorithm. *IEEE Transaction on Speech and Audio Processing*, 15(1):34–43, January 2007.

[NIS, 2003] Nist speaker recognition evaluation plan. Proc. NIST speaker Recognition Workshop, 2003.

[O'Shaughnessy, 2003] D. O'Shaughnessy. Interacting With Computers by Voice: Automatic Speech Recognition and Synthesis. *Proceedings of the IEEE*, 91(9):1272–1305, September 2003.

[Papoulis, 1991] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Inc., 3 edition, 1991.

[Pellom and Hansen, 1999] B.L. Pellom and J.H.L Hansen. An Experimental Study of Speaker Verification Sensitivity to Computer Voice-Altered Imposters. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 837–840, March 1999.

[Peskin *et al.*, 2003] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D.A. Reynolds, and Bing Xiang. Using Prosodic and Conversational Features for High-Performance Speaker Recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 792–795, Apr 2003.

[Plumpe *et al.*, 1999] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds. Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identifica-

tion. *IEEE Transaction on Speech and Audio Processing*, 7(5):569–586, September 1999.

[Rabiner and Juang, 1993] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Signal Processing Series. Prentice Hall, New Jersey, 1993.

[Rabiner and Schafer, 1978] L. R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals*. Prentice Hall, New Jersey, 1978.

[Ramasubramanian *et al.*, 2006] V. Ramasubramanian, A. Das, and V.P. Kumar. Text-Dependent Speaker-Recognition Using One-Pass Dynamic Programming Algorithm. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 901–904, May 2006.

[Reynolds and Rose, 1992] D.A. Reynolds and R.C. Rose. An Integrated Speech-Background Model for Robust Speaker Identification. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 185–188, Mar 1992.

[Reynolds and Rose, 1995] D.A. Reynolds and R.C. Rose. Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models. *IEEE Transaction on Speech and Audio Processing*, 3:72–83, January 1995.

[Reynolds *et al.*, 2000] D. A. Reynolds, T.F. Quatieri, and R. Dunn. Speaker Verification using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[Reynolds *et al.*, 2002] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, Barbara Peskin, A. Adami, Q. Jin, David Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang. SuperSID Project Final Report. Technical report, John Hopkins University, 2002.

[Reynolds *et al.*, 2003] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Qin Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey,

D. Jones, and Bing Xiang. The SuperSID Project: Exploiting High-Level Information for High-Accuracy Speaker Recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 784–787, Apr 2003.

[Reynolds, 1994] D.A. Reynolds. Experimental Evaluation of Features for Robust Speaker Identification. *IEEE Transaction on Speech and Audio Processing*, 2:639–643, 1994.

[Reynolds, 1995] D. A. Reynolds. Speaker Identification and Verification using Gaussian Mixture Speaker Models. *Speech Communication*, 17:91–108, 1995.

[Reynolds, 1996] D.A. Reynolds. The Effects of Handset Variability on Speaker Recognition Performance: Experiments on the Switchboard corpus. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 113–116, May 1996.

[Reynolds, 2002] D.A. Reynolds. An Overview of Automatic Speaker Recognition Technology. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 4, May 2002.

[Reynolds, 2003] D. Reynolds. Channel Robust Speaker Verification via Feature Mapping. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 6–10, Apr 2003.

[Rosenberg and Sambur, 1975] A. Rosenberg and M. Sambur. New Techniques for Automatic Speaker Verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(2):169–176, 1975.

[Rosenberg, 1971] A. E. Rosenberg. Effect of Glottal Pulse Shape on the Quality of Natural Vowels. *Journal of the Acoustical Society of America*, 49(2B):583–590, February 1971.

[Rothenberg, 1973] M. Rothenberg. A New Inverse Filtering Technique for Deriving the Glottal Airflow Waveform During Voicing. *Journal of the Acoustical Society of America*, 53:1632–1645, 1973.

[Sandhu and Ghitza, 1995] S. Sandhu and O. Ghitza. A Comparative Study of Mel Cepstra and EIH for Phone Classification under Adverse Conditions. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 409–412, May 1995.

[Scherer *et al.*, 1988] R. C. Scherer, D. G. Druker, and I. R. Titze. *Vocal Physiology: Voice Production Mechanisms and Functions*, chapter Electroglottography and Direct Measurement of Vocal Fold Contact Area, pages 279–291. Raven Press Ltd, New York, 1988.

[Scherer *et al.*, 1995] R.C. Scherer, V.J. Vail, and B. Rockwell. *Producing Speech: Contemporary Issues*, chapter Examination of the Laryngeal Adduction Measure EGGW, pages 269–290. American Institue of Physics, 1995.

[Shriberg *et al.*, 2000] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür. Prosody-Based Automatic Segmentation of Speech into Sentences and Topics. *Speech Communication*, 32(1-2):127–154, September 2000.

[Shriberg *et al.*, 2005] E. Shriberg, L. Ferrera, S. Kajarekara, A. Venkataramana, and A. Stolckea. Modeling Prosodic Feature Sequences for Speaker Recognition. *Speech Communication*, 46(3-4):455–472, July 2005.

[Smits and Yegnanarayana, 1995] R. Smits and B. Yegnanarayana. Determination of Instants of Significant Exitation in Speech using Group Delay Function. *IEEE Transaction on Speech and Audio Processing*, 5(3):325–333, September 1995.

[Sondhi, 1974] M. M. Sondhi. Model for Wave Propagation in a Lossy Vocal Tract. *Journal of the Acoustical Society of America*, 55:1070–1075, May 1974.

[Sönmez *et al.*, 1997] K. Sönmez, L. Heck, M. Weintraub, and E. Shriberg. A Log-normal Tied Mixture Model of Pitch for Prosody-Based Speaker Recognition. *Proc. EUROSPEECH*, 3:1391–1394., September 1997. Rhodes, Greece.

[Sönmez *et al.*, 1998] K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub. Modeling Dynamic Prosodic Variation for Speaker Verification. *In Proceedings of International Conference on Spoken Language*, 7:3189–3192, November 1998.

[Sorokin, 1992] V. N. Sorokin. Determination of Vocal Tract Shape for Vowels. *Speech Communication*, 11:71–85, 1992.

[Sorokin, 1994] V. N. Sorokin. Inverse Problems for Fricatives. *Speech Communication*, 14:249–262, 1994.

[Steiglitz and Dickinson, 1977] K. Steiglitz and B. Dickinson. The Use of Time-Domain Selection for Improved Linear Prediction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25:34–39, Feb 1977.

[Strik *et al.*, 1993] H. Strik, B. Cranen, and L. Boves. Fitting LF-model to Inverse Filtered Signals. In *Eurospeech*, volume 1, pages 103–106, Berlin, 1993.

[Strube, 1974] H. W. Strube. Determination of the Instant of Glottal Closure from the Speech Wave. *Journal of the Acoustical Society of America*, 56(5):1625–1629, 1974.

[Stylianou, 1999] Y. Stylianou. Synchronization of Speech Frames Based on Phase Data with Qpplication to Concatenative Speech Synthesis. In *6th European Conference on Speech Communication and Technology*, volume 5, pages 23343–2346, Budapest, September 1999.

[Talkin, 1995] D. Talkin. *Speech Coding and Synthesis*, chapter 14. A Robust Algorithm for Pitch Tracking (RAPT), pages 495–518. In Kleijn and Paliwal [1995], 1995.

[Teager, 1980] H. Teager. Some Observations on Oral Air Flow During Phonation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28:599–601, 1980.

[Teixeira *et al.*, 2000] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and K. Sönmez. Prosodic Features for Automatic Text-Independent Evaluation of Degree of Nativeness for Language Learners. In *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, October 2000.

[Thyssen *et al.*, 1994] J. Thyssen, H. Nielsen, and S.D. Hansen. Non-Linear Short-Term Prediction in Speech Coding. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 185–188, Apr 1994.

[Tishby, 1990] N. Tishby. A Dynamical Systems Approach to Speech Processing. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 365–368, Apr 1990.

[Titze, 1984] I.R. Titze. Parameterization of the Glottal Area, Glottal Flow, and Vocal Fold Contact Area. *Journal of the Acoustical Society of America*, 75(2):570–580, February 1984.

[Tuan and d'Alessandro, 1999] V.N. Tuan and C. d'Alessandro. Robust Glottal Closure Detection Using the Wavelet Transform. In *Eurospeech*, pages 2805–2808, Budapest, September 1999.

[Vapnik, 1999] V.N. Vapnik. An Overview of Statistical Learning Theory. *IEEE Transactions on Neural Networks*, 10:988–999, September 1999.

[Veeneman and BeMent, 1985] D.E. Veeneman and S.L. BeMent. Automatic Glottal Inverse Filtering from Speech and Electroglottographic Signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33:369–377, April 1985.

[Wan and Renals, 2003] V. Wan and S. Renals. SVMSVM: Support Vector Machine Speaker Verification Methodology. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 221–224, April 2003.

[Wan and Renals, 2005] V. Wan and S. Renals. Speaker Verification using Sequence Discriminant Support Vector Machines. *IEEE Transaction on Speech and Audio Processing*, 13:203–210, March 2005.

[Weber *et al.*, 2002] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg. Using Prosodic and Lexical Information for Speaker Identification. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 141–144, 2002.

[Wildermoth and Paliwal, 2003] B. R. Wildermoth and K. K. Paliwal. GMM Based Speaker Recognition on Readily Available Databases. In *Proc Microelectronic Engineering Research Conf*, Brisbane, Nov 2003.

[Wong *et al.*, 1979] D. Y. Wong, J. D. Markel, and A. H. Gray, Jr. Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(4):350–355, August 1979.

[Yegnanarayana and Smits, 1995] B. Yegnanarayana and R. Smits. A Robust Method for Determining Instants of Major Excitations in Voiced Speech. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 776–779, 1995.

[Yegnanarayana *et al.*, 2005] B. Yegnanarayana, S. R. M. Prasanna, J. M. Zachariah, and Cheedella S. Gupta. Combining Evidence From Source, Suprasegmental and Spectral Features for a Fixed-Text Speaker Verification System. *IEEE Transaction on Speech and Audio Processing*, 13(4):575–582, July 2005.

[Young *et al.*, 2002] S. Young, G. Evermann, T. Hain, Kershaw D., G. More, J. O. D. Odell, D. V. V. Provey, and P. Woodland. *The HTK book*. Cambridge University Engineering Dept, 2 edition, 2002.

[Yu *et al.*, 1995] K. Yu, J. Mason, and J. Oglesby. Speaker Recognition using Hidden Markov Models, Dynamic Time Warping and Vector Quantisation. *IEE Proceedings on Vision, Image and Signal Processing*, 142:313–318, 1995.

[Zhu and Kasuya, 1996] Weizhong Zhu and Hideki Kasuya. A New Speech Synthesis System Based on the ARX Speech Production Model. In *International Conference on Spoken Language Processing*, pages 1413–1416, 1996.