Speaker Verification Using the Shape of the Glottal Excitation Function for Vowels

Michael Wagner

Human-Computer Communication Laboratory
School of Information Sciences and Engineering
University of Canberra, Australia
michael.wagner@canberra.edu.au

Abstract

This paper seeks to establish a baseline for the potential contribution of the shape of the glottal source waveform to speaker recognition. A text-dependent speaker verification experiment was performed with 4 monosyllabic words spoken repeatedly by the 16 speakers of the TI46 speech data corpus. A single fundamental period was automatically extracted from each vowel centre and inverse-filtered to obtain an approximation of one period of the glottal excitation function for the vowel. The shape of each excitation period was analysed using the magnitude and phase information obtained by means of the discrete Fourier transform. The analysis is independent of both vocal tract and fundamental frequency information and a speaker verification test yields average equal-error rates of 20% for the female speakers and 31% for the male speakers.

1. Introduction

It is well known that the shape as well as the frequency of the glottal excitation function of speech conveys information on the speaker. Particularly in speech synthesis, when a source-filter model is used, the shape of the glottal excitation function is most important for both naturalness and the speaker percept of the synthetic speech (van Santen, Sproat, Olive, & Hirschberg, 1997). Inversely, it can be argued that the shape of the glottal excitation function could be utilised for the recognition of a speaker, if it could be extracted reliably from the speech signal.

Research reporting on the usefulness of the characteristics of glottal excitation has been published occasionally over the years, e.g. (Monsen & Engebretsen, 1977; Thevenaz & Hügli, 1995; Plumpe, Quatieri & Reynolds, 1999) and more recently (Slyh, Hansen & Anderson, 2004). However, glottal source characteristics do not currently appear as standard features of contemporary speaker recognition systems.

This paper proposes an algorithm to extract a single fundamental period of speech in the centre of a vowel and to perform pitch-synchronous linear predictive inverse filtering on that period in order to obtain the equivalent period of the glottal excitation function. While previous research has mainly attempted to capture the individuality of glottal excitation through applying various glottal excitation models, e.g. (Slyh et al., 2004; Fujisaki & Ljungqvist, 1986), a method is proposed in this paper to analyse the shape of the glottal excitation period by determining the magnitude and phase spectra of a phase-normalised version of the excitation function.

Finally, speaker verification is performed using the four words “two”, “three”, “six” and “stop” from the TI46 speaker recognition corpus. The glottal excitation shape parameters were determined for a glottal period in the centre of the vowel of each word. A multivariate Gaussian model of those parameters was built for each of the 8 female and 8 male speakers from the 10 repetitions of each word in the training session. A further 16 test repetitions of each word (8 sessions x 2 tokens) from each of the 16 speakers were then used to determine the likelihoods of each test utterance given the different models.

For each of the 4 vowels and each gender group, a verification experiment was conducted, building a model of the glottal source spectrum for each of the 8 speakers of that group. For each gender group and each vowel the available data allowed 8x16=128 client trials and 8x7x16=896 non-target trials. Likelihood scores were pooled separately for the 8 male speakers and for the 8 female speakers, and were then presented in the form of DET curves and equal-error rates.

The paper will first describe the signal processing that is necessary to extract a fundamental period from the vowel centre and to determine the magnitude and phase information of the equivalent glottal function. Secondly, the paper will report the results of the speaker verification experiment.

2. Extraction of the shape of the glottal excitation function for vowels

Speaker-dependent feature vectors are determined for the 8 male and 8 female speakers of the TI46 speaker recognition data corpus. These feature vectors are based solely on the shape of the glottal excitation function as determined by inverse filtering.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

2.1. Automatic extraction of a vowel-central fundamental period

The acoustic signals containing the target vowels are first analysed by means of linear prediction analysis. A window of width 20ms was advanced by 10ms through the signal and energy and autocorrelations were determined for each frame. The voicing parameter is defined for each frame as

$$VC = \max_{k_{\min}\cdots k_{\max}} \{r(k)\}/\, r(0) \qquad (1)$$

and an expected fundamental period is defined as

$$\tau_0 = \arg\max_{k_{\min}\cdots k_{\max}} \{r(k)\}/\, f_s \qquad (2)$$

with lags k_{min} and k_{max} corresponding to the range of expected fundamental frequencies, which are 80…200Hz for male speakers and 150…350Hz for female speakers. Voiced frames are defined by the condition VC > 0.4. The centre of the vowel is then taken to be the voiced frame with the maximum energy value.

Autocorrelations were determined by a variant of the ordinary autocorrelation function, which ensures an equal number of summands for all examined lags

$$r(k) = \sum_{i=1}^{N-k_{\max}} x(i)x(i+k) \qquad (3)$$

with signal values x(i) and frame size N.

A single fundamental period is then extracted from the vowel-central frame by first finding the signal maximum in the frame at sample x(imax1). This sample is the start sample of the target fundamental period.

The precise duration of that period is found by finding the maximum autocorrelation of a sequence of samples surrounding x(imax1) for lags between $0.95\tau_0$ and $1.05\tau_0$

$$\tau_1 = \arg\max_{0.95\tau_0 \ldots 1.05\tau_0} \{r(k)\}/\, f_s \qquad (4)$$

with

$$r(k) = \sum_{i=i\max 1-\tau_0/8}^{i\max 1+\tau_0/8} x(i)x(i+k) \qquad (5)$$

The extraction process is illustrated in Figure 1, showing the signal, the energy, voicing and F0 contours in the first 4 panes. The maximum-energy voiced frame is marked in panes 2-4 and is shown magnified in the 5th pane, where the samples x(imax) and x(imax+τ_1) are marked. The last pane shows the extracted fundamental period. Note that the beginning of this period obviously does not correspond to the time of glottal opening, but we found that the maximum signal value in the frame provides the most reliable point for the determination of the precise duration of the glottal period.

2.2. Inverse filtering

In order to obtain the glottal excitation waveform, the extracted signal values are subjected to linear prediction analysis and inverse filtering (Markel & Gray, 1975). Throughout the analysis, the extracted signal period is treated as a single period of a periodical signal, such that difference values and autocorrelation functions are evaluated on indices modulo the length of the period. The signal is preemphasised with α=15/16 and a predictor of order 15 is determined by means of the autocorrelation method in order to model the expected 6-7 poles within the signal bandwidth of 6.25kHz. The unpreemphasised signal is then inverse-filtered with the resulting 15-pole LP filter, yielding the required approximation of the glottal excitation waveform.

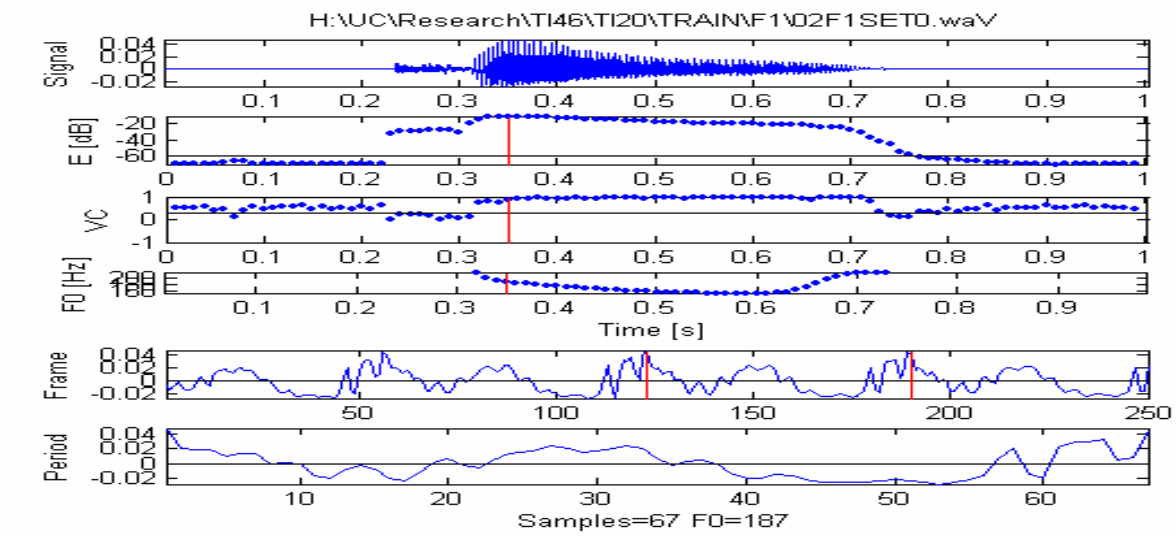

Figure 1: Extraction of a single fundamental period from the vowel-centre frame in the word "two".

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

The final step of the determination of the glottal excitation function is the normalization of the phase of the function, which is done by defining the sample with the minimum value as the start of the glottal excitation period, e.g. sample 53 in pane 6 of Figure 1. In many of the signal files, this value is close to the actual opening of the glottis, but this is by no means guaranteed. The importance of the rule is that is was found to reliably yield an equivalent position in most of the signal files examined.

2.3. Glottal excitation shape parameters

The shape of the glottal excitation function is described by the complex coefficients of its discrete Fourier transform (DFT)

$$X(k) = \sum_{i=0}^{\tau_1 - 1} x(i) \exp\{j 2\pi i k / \tau_1\} \qquad (6)$$

for $1 \le k \le 15$, where $\tau 1$ is the duration of the fundamental period determined according to (4) and (5).

Equation (6) implies that the frequencies k are normalized with respect to the duration of the fundamental period and are therefore not related to the actual fundamental frequency of the signal. The set of glottal excitation shape parameters for the subsequent speaker verification experiments is formed by various subsets of the magnitudes and phases of the 15 spectral coefficients determined according to Equation (6).

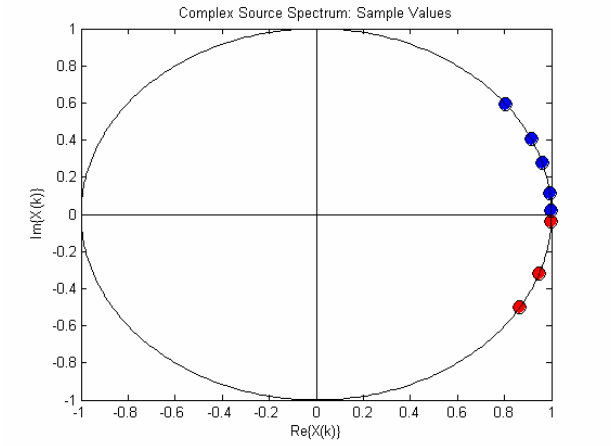

Figure 2: Determination of mean and variance for glottal source phase values: the red dots indicate large positive phase values, which need to be converted to small negative values before calculating the phase statistics

2.4. Speaker models

The training set of the TI46 corpus comprises 10 repetitions of each word spoken by each speaker during a single training session. For each of the 16 speakers and each of the 4 vowels, a Gaussian mixture model (GMM) is built from the respective set of glottal source parameters. All GMMs are constrained to have diagonal covariance matrices. For some of the experiments, the GMMs are trained solely with the spectral magnitude values, in which case the determination of the means and covariances is straightforward. However, GMMs that are trained with both magnitude and phase values, require an algorithm that takes into account the arithmetic modulo 2π of the phase values as is illustrated in Figure 2. If the mean phase of the 8 samples in Figure 2 is to be determined, the 4 samples with phase values just less than 2π (red dots) need to be adjusted by 2π in order to give the expected mean value. Assuming a set of phase values $\{p_i\}$, the mean m and variance s^2 are determined by the following algorithm:

```
Sort {pᵢ} in ascending order
m := mean{pᵢ}
s² := variance{pᵢ}
i := 1
while i≤size{pᵢ} and pᵢ<π
   pᵢ := pᵢ + 2π
   if variance{pᵢ} < s²
      m := mean{pᵢ}
      s² := variance{pᵢ}
   end if
   i := i + 1
end while
```

For the different experiments, 2-, 4-, 8- and 15-dimensional speaker models are built from the first 2, 4, 8 and 15 magnitude coefficients. To also include the phase information, additional 4-, 8-, 16- and 30-dimensional speaker models are built from the first 2, 4, 8 and 15 magnitude and phase coefficients.

The testing set of the TI46 corpus comprises 16 further repetitions of each word by each speaker, 2 each recorded in 8 separate sessions. Each of these tokens is evaluated against each of the 16 speaker models by computing the log likelihood of the token given the model. The resulting likelihoods are pooled among same-sex speakers and then used to derive detection error trade-off curves as well as equal-error rates for each speaker and each vowel. The results are shown in the next section.

3. Results

A DET curve and corresponding equal error rate were determined for different parameter sets and for each of the 4 vowels under examination. The DET curves in Figures 3 to 6 show the detection error trade-off for each vowel for the female speakers.

The 4 solid lines represent the detection error for using only the magnitudes of the source spectra, and the 4 dotted lines represent the detection error for using both the magnitudes and phases of the source spectra. The black curves represent the first 2 spectral magnitudes (and phases), the red curves represent the first 4 spectral magnitudes (and phases), the blue curves represent the first 8 spectral magnitudes (and phases), and the green curves represent the first 15 spectral magnitudes (and phases).

The equal-error rates for the different conditions for the female speakers are shown in Table 1. The equal-error rates range between 18.0% (green) and 33.0% (orange) for the different vowels and experimental conditions and it can be seen that the addition of the source spectrum phase information diminishes the EER for females in all conditions.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

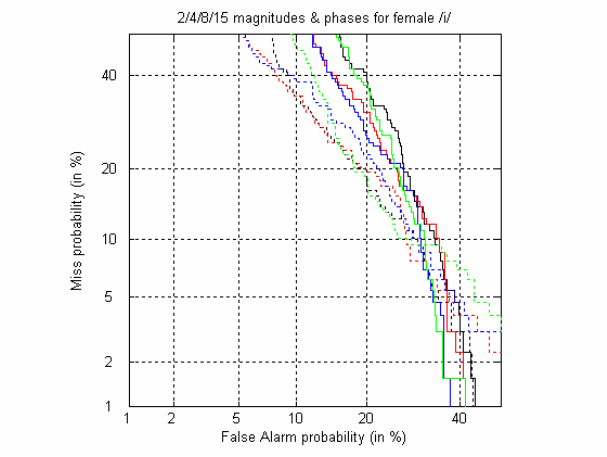

Figure 3: DET curves for /i/ by female speakers, using source spectrum magnitudes and phases with different spectral vector sizes: N=2 (black), N=4 (red), N=8 (blue), N=15 (green. Solid curves for spectral magnitude only, dotted curves for magnitude & phase.

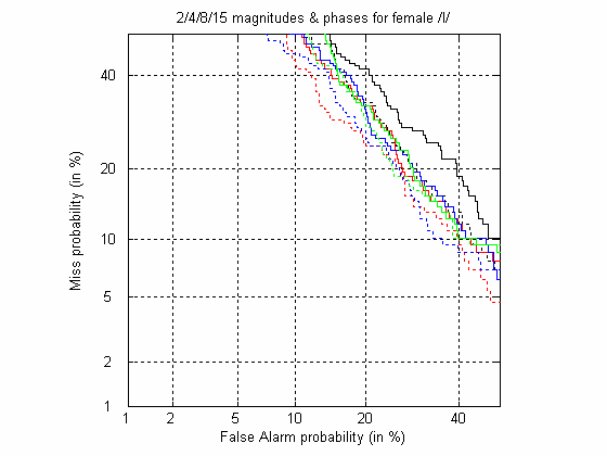

Figure 5: DET curves for /ɪ/ by female speakers, using source spectrum magnitudes and phases as in Fig.3.

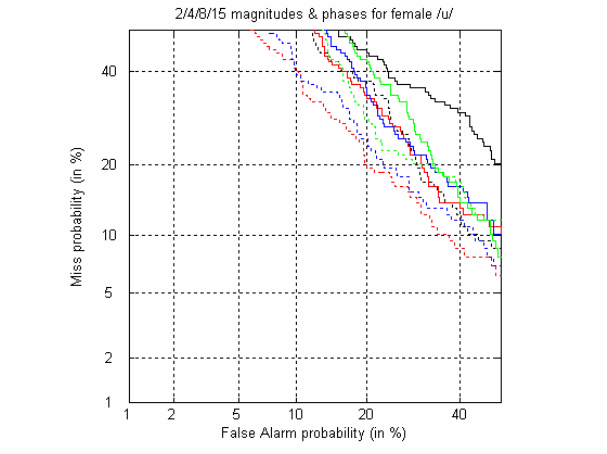

Figure 4: DET curves for /u/ by female speakers, using source spectrum magnitudes and phases as in Fig.3.

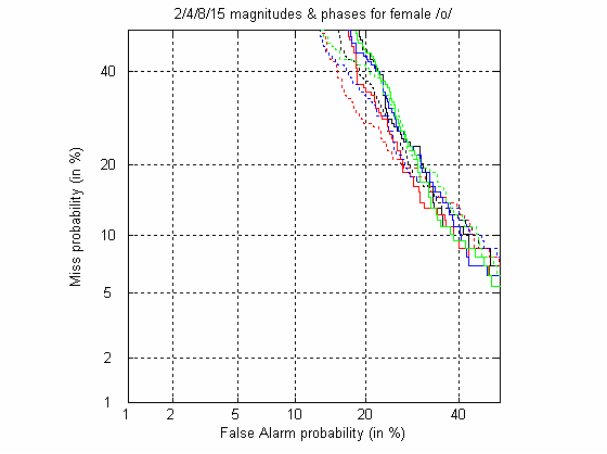

Figure 6: DET curves for /ɔ/ by female speakers, using source spectrum magnitudes and phases as in Fig.3.

Table 1: Equal-error rates for female speakers, 4 vowels and 8 experimental conditions of source spectrum parameters.

Order	/i/		/u/		/ɪ/		/ɔ/	
	Mag	M&Ph	Mag	M&Ph	Mag	M&Ph	Mag	M&Ph
2	25.8%	19.0%	33.0%	25.9%	27.0%	24.9%	26.0%	24.9%
4	22.7%	19.5%	26.5%	19.5%	25.0%	22.7%	25.0%	23.4%
8	22.5%	21.1%	25.4%	21.1%	24.0%	23.4%	26.3%	25.0%
15	24.2%	18.0%	28.2%	22.7%	24.8%	22.7%	26.6%	26.6%

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

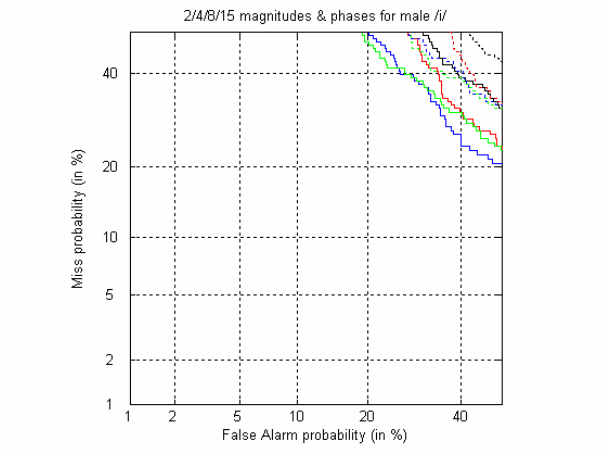

Figure 7: DET curves for /i/ by male speakers, using source spectrum magnitudes and phases as in Fig.3.

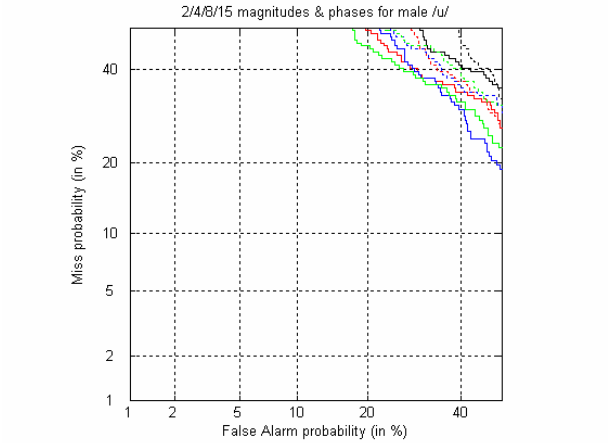

Figure 8: DET curves for /u/ by male speakers, using source spectrum magnitudes and phases as in Fig.3.

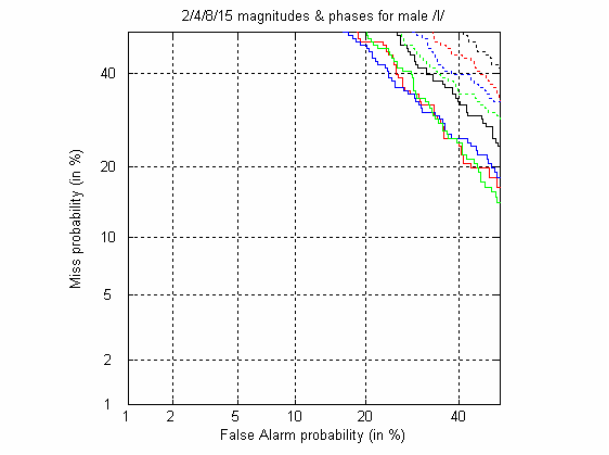

Figure 9: DET curves for /ɪ/ by male speakers, using source spectrum magnitudes and phases as in Fig.3.

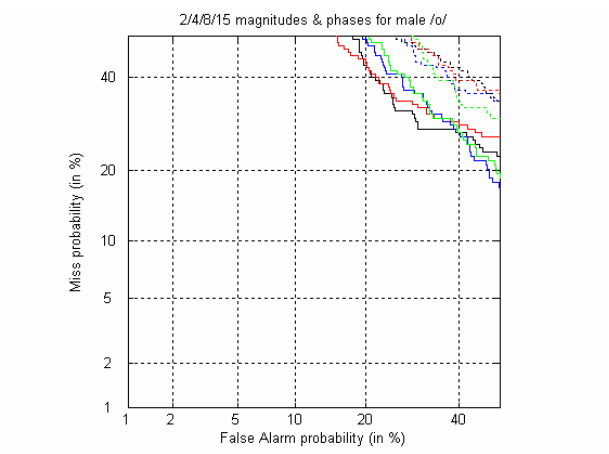

Figure 10: DET curves for /ɔ/ by male speakers, using source spectrum magnitudes and phases as in Fig.3.

Table 2: Equal-error rates for male speakers, 4 vowels and 8 experimental conditions of source spectrum parameters.

Order	/i/		/u/		/ɪ/		/ɔ/	
	Mag	M&Ph	Mag	M&Ph	Mag	M&Ph	Mag	M&Ph
2	38.9%	45.2%	40.2%	42.3%	36.5%	44.6%	29.7%	40.5%
4	34.9%	41.0%	35.5%	37.8%	31.1%	42.7%	32.0%	39.1%
8	32.9%	39.7%	34.6%	36.7%	31.0%	38.2%	32.6%	37.8%
15	33.9%	36.5%	35.3%	39.4%	32.5%	37.4%	32.8%	38.3%

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

The DET curves in Figures 7 to 10 show the detection error trade-off for each vowel and the different source-spectrum conditions for the male speakers. The colour coding of the lines is as described above for the female speakers. It is very clear that for the same conditions of determining the vowel-central fundamental period and the source-spectrum parameters, the detection-error rates are significantly and consistently higher more the male speakers.

The equal-error rates for the different conditions for the male speakers are shown in Table 2. The equal-error rates range between 29.7% (green) and 45.2% (orange) for the different vowels and experimental conditions. For the male speakers it can be seen that the addition of the source spectrum phase information actually increases the EER in all conditions. These results for the male speakers are most likely due to a different number of formants within the signal range of 0-6.25kHz, compared with the number of formants for the female speakers.

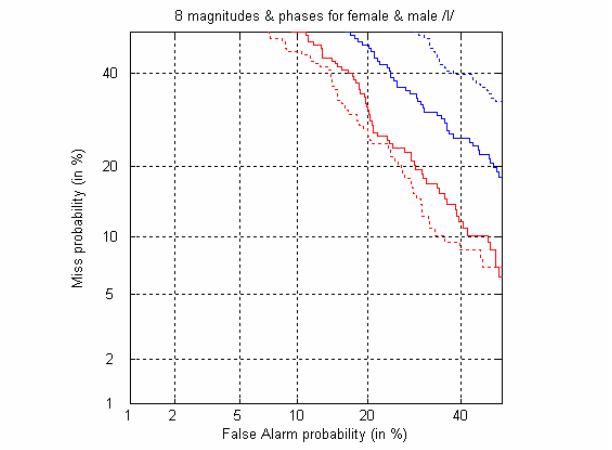

Figure 11: DET curves for /ɪ/ by female (red) and male speakers (blue), using 8th-order source spectrum magnitudes (solid) and 8th-order magnitudes & phases (dotted).

Figure 11 shows a comparison between the male and female speakers for two specific experimental conditions, namely an 8th-order magnitude spectrum and a combined 8th-order magnitude-and- spectrum phase spectrum of the glottal source signal for the vowel /ɪ/. The Figure exemplifies the findings that the error rates are smaller for the female speakers than for the male speakers if the magnitude spectrum only is used, and that the error rate is diminished further for the female speakers if glottal phase information is added while it is increased for the male speakers if glottal phase information is added.

4. Conclusions

It has been found that the shape of the glottal excitation waveform contains speaker-related information that allows speaker verification equal-error rates of between 18.0% and 33.0% for female speakers and between 29.7% and 45.2% for male speakers. Between 4 and 8 spectral magnitude coefficients are sufficient when determined from an automatically selected and normalized single fundamental period in the centre of vowels. For the female data it was found that the addition of spectral phase information diminished the error rate while this was not the case for the male data. The method promises to be valuable where vowel signals in well-defined contexts are available for speaker verification. Since the measured parameters are largely independent of both segmental information and fundamental frequency, they lend themselves very well as complementary parameters to standard cepstral and other segmental parameters used in speaker recognition.

5. References

Fujisaki, H. & Ljungqvist, M. (1986). Proposal and evaluation of models for the glottal source waveform, *Proc. ICASSP-1986*, 1605-1608.

Markel, J. & Gray, A. (1975). *Linear Prediction of Speech Signals*, Springer Verlag, Berlin.

Monsen, R.B. & Engebretsen, A.M. (1977). Study of variations in the male and female glottal wave, *J. Acoust Soc Am*, 62(4), 981-993.

Plumpe, M., Quatieri, T. & Reynolds, D. (1999). Modeling of the glottal flow derivative waveform with application to speaker identification, *IEEE Trans Speech Audio Processing*, 7(5), 569-586.

Slyh, R.E., Hansen, E.G. & Anderson, T.R. (2004). Glottal modeling and closed-phase analysis for speaker recognition, *Proc. Odyssey-2004 Speaker and Language Recognition Workshop*, 315-322.

Thevenaz, P. & Hügli, H. (1995). Usefulness of the LPC residue in text-independent speaker verification, *Speech Communication*, 17, 145-157.

van Santen, J., Sproat, R., Olive, J. & Hirschberg, J. (1997). *Progress in Speech Synthesis*, Springer Verlag, Berlin.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review