VOICE SOURCE PARAMETERS FOR SPEAKER VERIFICATION

Andreas Neocleous and Patrick A. Naylor
Dept. of Electrical and Electronic Engineering,
Imperial College, Exhibition Road
London SW7 2BT, UK
Tel: +44 (0)171 594 6235; fax: +44 (0)171 594 6234
e-mail: a.neocleous@ic.ac.uk, p.naylor@ic.ac.uk

ABSTRACT

In this paper we report on a study of the variability of voice source parameters in the context of speaker characterisation, and we propose a speaker verification system which incorporates these parameters. The motivation for this approach is that, whilst we have conscious control over the action of our vocal tract articulators such as the tongue and jaw, we have only limited voluntary muscle control over the vocal cords. The conjecture is, therefore, that impostors are less likely to be able to mimic vocal cord effects than vocal tract effects. The hybrid speaker verification system that is proposed incorporates two sub-systems to improve the overall performance: (i) a cepstral-based HMM with cohort normalisation and (ii) voice source parameters derived from Multi-cycle Closed-phase Glottal Inverse Filtering (MCGIF). Preliminary experimental results show that the hybrid system performs better than either of the sub-systems in terms of the equal error rate (EER). Specifically, the hybrid system improved the performance of the cepstral-based HMM system by 78% on average, resulting in a mean EER of 0.42% for the specific tests conducted.

1 INTRODUCTION

Speaker verification aims to verify the claimed identity of a speaker based on a sample of their voice. There has been an increased commercial interest in speaker verification systems in the form of security applications and access control, such as voice activated door locks, smart card security and telephone banking. In a customary speaker verification system, the decision rule for accepting or rejecting a claimed speaker is based on the score of a test utterance for the claimed speaker and a predefined threshold. Previous work on speaker verification has described the use of one or a combination of techniques as a means to classify and distinguish between speakers. Such techniques include vector quantizers [8], neural networks [13], and more commonly Hidden Markov Models (HMM) [3] and Gaussian Mixture Models (GMM) [12].

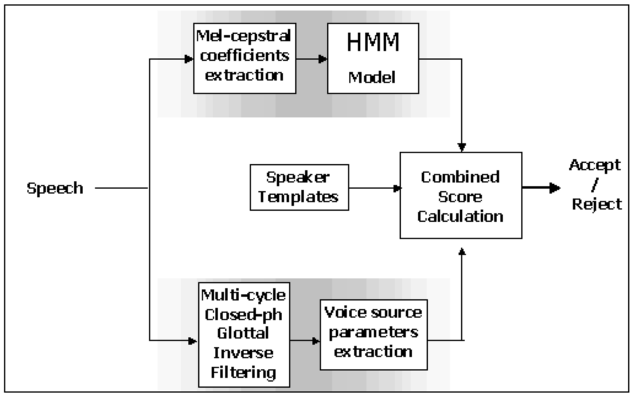
Mel-cepstral coefficients extraction
HMM Model
Speech
Speaker Templates
Combined Score Calculation
Accept / Reject
Multi-cycle Closed-ph Glottal Inverse Filtering
Voice source parameters extraction

Figure 1: The Hybrid Speaker Verification System

In this paper we propose a new approach for speaker verification (Fig. 1), which incorporates firstly an HMM using mel-cepstral coefficients and secondly a set of parameters which describe specific characteristics of the glottal flow, namely the voice source parameters. Both approaches provide cohort-normalised scores [6] for the test utterance which combine to give the final decision to accept or reject the speaker. The main aims of this work are (i) to investigate the intra- and inter-speaker variability of the voice source parameters and hence, suitability to speaker verification, and (ii) to propose a way to integrate the voice source parameters into a speaker verification system and show results of the performance of such a system. A future aim will be to compare the performance of the system with and without the use of the Electroglottograph (EGG) signal [4] to determine the instants of glottal closure and opening.

2 GLOTTAL INVERSE FILTERING

Speaker verification using voice source parameters is based on the hypothesis that there are pattern characteristics in the waveform of the glottal flow derivative which can distinguish one speaker from others. Figure 2 shows a typical cycle of the glottal flow derivative waveform and the definition of seven parameters describing the dimensional features of the waveform. MCGIF [15] is used to extract the glottal flow derivative from a speech signal utilising the EGG to locate the closed-phase period. Glottal source modelling principles derived from the Liljencrants-Fant Model [9] are used to define the specific features of the inverse filter output.

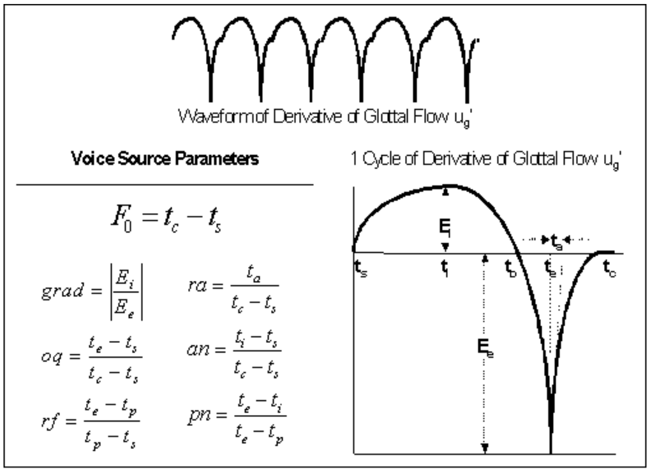
Waveform of Derivative of Glottal Flow u_g'
Voice Source Parameters
1 Cycle of Derivative of Glottal Flow u_g'
$F_0 = t_c - t_s$
$grad = \left|\frac{E_i}{E_e}\right|$
$ra = \frac{t_a}{t_c - t_s}$
$oq = \frac{t_e - t_s}{t_c - t_s}$
$an = \frac{t_i - t_s}{t_c - t_s}$
$rf = \frac{t_e - t_p}{t_p - t_s}$
$pn = \frac{t_e - t_i}{t_e - t_p}$
E_i
t_a
t_s
t_i
t_p
t_e
t_c
E_e

Figure 2: Derivation of Voice Source Parameters

Glottal Inverse Filtering (GIF) is the process by which a speech pressure wave is de-convolved into the vocal-tract filtering process and the driving source function to the vocal tract. Closed-phase GIF (CGIF) assumes the structure of a vocal-tract filter determines the transfer function of the filter via an objective spectral estimator (LP analysis is used in this case) and computes the parameters of the filter during the period when the glottis is closed. This technique is accurate because the formant structure of the speech signal is better described during the closed phase. Under certain conditions, the formulation of the closed phase analysis is linear and can be solved efficiently; the main disadvantage is the lack of information in the unvoiced parts of speech. CGIF's capabilities are restricted from the fact that the closed phase period (mostly in female speech) can sometimes be too short for accurate measurement of low frequency formants [14]. This restriction is diminished by using a technique termed Multi-cycle Closed-phase Glottal Inverse Filtering (MCGIF) which uses speech data from adjacent closed-phase periods to make up for the lack of data in the closed-phase period in question [2]. In MCGIF, we consider the closed phases of k consecutive larynx cycles $cp1, cp2, ..., cpk$ of a speech pressure wave (where $s(n)$ is the nth sample). The total prediction error can then be expressed as [2],

$$E_{total} = E_{cp_1} + E_{cp_2} + ... + E_{cp_k} \qquad (1)$$

where

$$E_K = \sum_K e^2(n) = \sum_K \left(s(n) + \sum_{i=1}^{p} a_i s(n-i) - G_K \right)^2 \qquad (2)$$

$$for\ K = cp_1, cp_2, ..., cp_k$$

and a_i is the i_{th} LPC coefficient and G_k is the k_{th} magnitude of constant excitation The unknowns a_i and G_K can then be determined by setting,

$$\frac{\partial E_{total}}{\partial a_j} = 0\ for\ j = 1, 2, ..., p-1, p \qquad (3)$$

$$\frac{\partial E_{total}}{\partial G_K} = 0\ for\ K = cp_1, cp_2, ..., cp_k \qquad (4)$$

which results to the following set of simultaneous equations:

$$\begin{aligned} \sum_{i=1}^{p} a_i \sum_{cp_1,\ldots,cp_k} s(n-i)s(n-j) - G_{cp_1} \sum_{cp_1} s(n-j) & \\ - G_{cp_2} \sum_{cp_2} s(n-j) & \\ \vdots \quad & \\ - G_{cp_k} \sum_{cp_k} s(n-j) & \\ = - \sum_{cp_1,\ldots,cp_k} s(n)s(n-j) & \qquad (5) \\ for\ j = 1, 2, ..., p & \end{aligned}$$

$$\sum_{i=1}^{p} a_i \sum_K s(n-i) - \sum_K G_K = - \sum_K s(n) \qquad (6)$$

$$for\ K = cp_1, cp_2, ..., cp_k$$

These can be solved efficiently using the Cholesky decomposition method [14], yielding the required LPC coefficients, and hence, the voice source parameters.

3 VARIABILITY OF VS PARAMETERS

The significance of the voice source parameters and their potential in distinguishing between speakers was investigated using two experiments based on neural network structures. The experimentation was carried out by using the Archivable Priority List Actual-Word Database (APLAWD) [10] which is a 10 talkers (5 male and 5 female) with 10 repetition recordings of several one- or two- word items and sentences which also contain the corresponding EGG signals. In the first experiment, a competitive learning network [11] was used to determine whether the parameter values coming from different repetitions from the same speaker exhibit clustering properties. The network searched for clusters in the space of each parameter, without any knowledge of the identity of each repetition's speaker, and identified correctly around 50-60% of the repetitions using only one parameter at a time. The main conclusion from the competitive learning test is that the voice source parameters are exhibiting grouping properties among speakers that can be used to distinguish one speaker from another. In the second experiment, a two-layer perceptron network [11] was used to attempt to recognise speakers based solely on the voice source parameters. The two-layer perceptron can often classify non-linearly separable input vectors, with its first layer acting as a non-linear pre-processor for the second layer, which is trained as usual. Training the network on five of the repetitions, and testing on the other five, a recognition rate of 92% was achieved. However, more importantly, the neural network tests proved that for the tests performed, the voice source parameters could be used as a means to classify speakers, and hence showed that a speaker verification system could be based on a combination of voice source parameters with other more standard approaches.

4 TESTS AND RESULTS

Since the speech characteristics used in the two approaches in question describe different parts of speech (namely, vocal tract and glottal flow), it would be fair to assume that the errors made by the individual approaches are uncorrelated. The two approaches have been combined into one hybrid system in order to improve the overall performance. On this basis, a hybrid speaker verification system has been designed and implemented that employs a sub-system based on the voice source parameters, in parallel with the well-proven cepstral-based HMM system [3]. The experiments were performed using the APLAWD database, again using the first five repetitions for training and the other five for testing. The experimental evaluations were conducted using at first single vowels and finally single digit utterances. The vowel-tests showed that the combination of voice source parameters derived from glottal inverse filtering with HMM methods, resulted in a 10.3% improvement in terms of mean EER, in comparison to the HMM-system on its own. It should be realised that the speech data used in these tests are much shorter than probable user-selectable passwords in a real system. This limits the data available for constructing speaker models and for testing the system. It also means that the system does not currently take advantage of the full potential of the HMM. Once longer speech samples (such as digits) were used, the performance improved, and a 78% improvement in terms of mean EER was observed. The verification performance of each sub-system on its own and that of the hybrid system for both the phoneme- and the digit-tests are compared in Table 1, and analytic EER results for each APLAWD speaker for the digit tests are shown in Figure 3.

Method	Vowel EER(%)	Digit EER(%)
VS Parameters	29.2	17.6
HMM	2.79	1.91
Hybrid	2.50	0.42

Table 1: Equal Error Rate Comparison

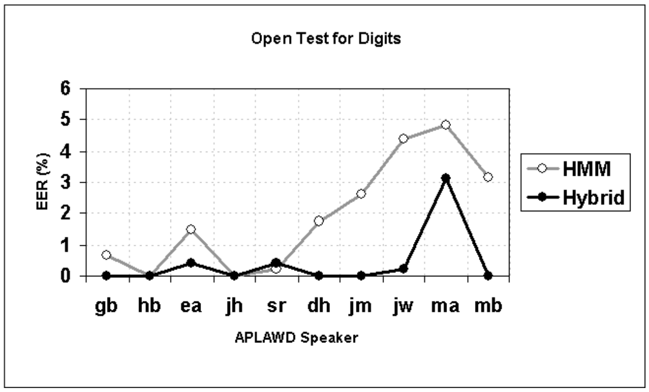

Figure 3: EER Results for the Digit Tests

Although the EER gives a good indication of the verification performance, an ROC plot is necessary for a more complete evaluation. Figure 3 shows ROC plots for the three methods. In addition to the ROC plots, the systems' performance is presented (Fig. 4) in the form of a DET Plot [7], which is a means of representing performance on detection tasks that involve a tradeoff of error types. It can be seen from both figures that the Hybrid system performs clearly better than either of the two sub-systems alone.

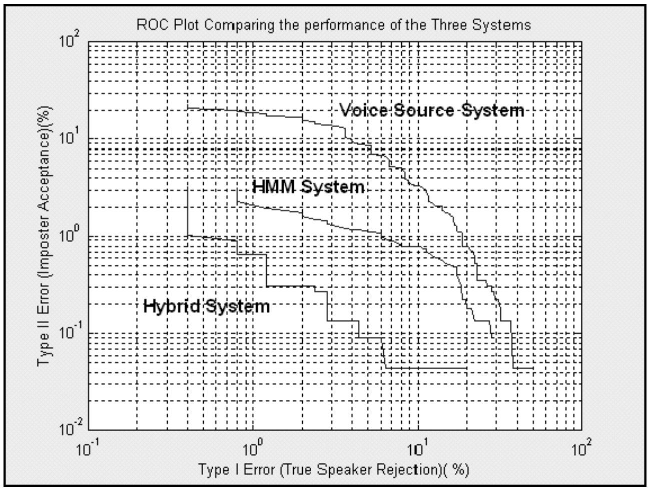

Figure 4: ROC Plots for the 3 systems

5 DISCUSSION AND CONCLUSIONS

In this paper, a study of the intra- and inter-speaker variability of voice source parameters is performed and a speaker verification system that incorporates these parameters is proposed and tested. The experimental evaluations show that combining voice source parameters derived from inverse filtering with HMM's in the manner described resulted in a 78% improvement in speaker verification tests on digit utterances. It must be noted however, that although this respresents a significant improvement in the performance of a speaker verification system, the applicability of VS paramaters in a practical system is still restricted due to the requirement for the EGG signal. Nevertheless, the task of extracting the VS parameters from the speech signal alone is currently being researched - applying the work in [4],[5],[1] as well as novel methods for determining the glottal closure instants. This will not only give way to the possible practical implementation of a hybrid system based partly on VS parameters, but it will also allow for the use of larger databases for tests which will give more solid results on the performance and capabilities of such a system.

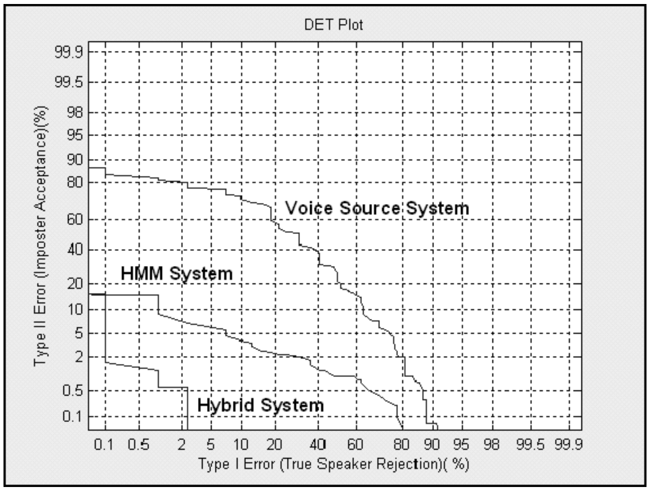

Figure 5: DET Plots for the 3 systems

References

[1] T. V. Ananthapadmanabha and B. Yegnanarayana. Epoch extraction of voice speech. *IEEE Trans. Acoust. Speech, Signal Processing*, 23(6):562–570, December 1975.

[2] M. Brookes and D. S. F. Chan. Speaker characteristics from a glottal airflow model using robust inverse filtering. *Proc. IOA*, 16(5):501–508, 1994.

[3] Q. G. Lin C. W. Che and D. S. Yuk. An hmm approach to text-prompted speaker verification. *CAIP Center, Rutgers University, USA*, 1996.

[4] Y. M. Cheng and D. O'Shaughnessy. Automatic and reliable estimation of glottal closure instant and period. *IEEE Trans. Acoust. Speech, Signal Processing,*, 37(12):1805–1815, December 1989.

[5] J. D. Markel D. Y. Wong and A. H. Gray Jr. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Trans. Acoust. Speech Signal Processing*, 27(4):350–355, August 1979.

[6] A. E. Rosenberg et al. The use of cohort normalised scores for speaker verification. *Proc. Int. Conf. On Spoken Language Processing*, 1:599–602, 1992.

[7] A. Martin et al. The det curve in assessment of detection task performance. *Proceedings EuroSpeech*, 1997.

[8] F. Soong et al. A vector quantisation approach to speaker recognition. *ATT Tech. J.*, 66:14–26.

[9] J. Liljencrants G. Fant and Q. Lin. A four parameter model of glottal flow. *STL - QPSR*, 4:1–13, 1985.

[10] A. Breen G. Lindsey and S. Nevard. *SPAR's archivable actual-word databases.* Dept. of Phonetics and Linguistics, University College, London, June 1987.

[11] S. Haykin. *Neural networks: A comprehensive foundation.* MacMillan College, 1994.

[12] E. S. Parris M. J. Carey and J. S. Bridle. A speaker verification system using alphanets. *Proceedings of the IEEE-ICASSP*, pages 397–400, 1991.

[13] J. Oglesby and J.Mason. Optimisation of neural models for speaker identification. *Proceedings of the IEEE-ICASSP*, pages 261–264, 1990.

[14] L. R. Rabiner and R. W. Schafer. *Digital processing of speech signals.* Prentice Hall, NJ, 1978.

[15] D. V. Veeneman and S. L. Bement. Automatic glottal inverse filtering from speech and electroglottographic signals. *IEEE Trans. Acoust. Speech, Signal Processing,*, 33(2):369–376, April 1985.