VOICE SOURCE CEPSTRUM COEFFICIENTS FOR SPEAKER IDENTIFICATION

Jon Gudnason and Mike Brookes

Department of Electrical and Electronic Engineering,
Exhibition Road, Imperial College London, UK
Email: {jon.gudnason, mike.brookes} @imperial.ac.uk

ABSTRACT

We propose a novel feature set for speaker recognition that is based on the voice source signal. The feature extraction process uses closed-phase LPC analysis to estimate the vocal tract transfer function. The LPC spectrum envelope is converted to cepstrum coefficients which are used to derive the voice source features. Unlike approaches based on inverse-filtering, our procedure is robust to LPC analysis errors and low-frequency phase distortion. We have performed text-independent closed-set speaker identification experiments on the TIMIT and the YOHO databases using a standard Gaussian mixture model technique. Compared to using mel-frequency cepstrum coefficients, the misclassification rate for the TIMIT database reduced from 1.51% to 0.16% when combined with the proposed voice source features. For the YOHO database the misclassification rate decreased from 13.79% to 10.07%. The new feature vector also compares favourably to other proposed voice source feature sets.

Index Terms— Vocal Systems, Speech Analysis, Cepstral Analysis, Speaker Recognition

1. INTRODUCTION

This paper presents a procedure for speaker identification feature extraction using voice source analysis. The voice source features are compatible with mel-frequency cepstrum coefficients (MFCC) and, when combined with them, achieve superior speaker identification performance.

The preferred feature sets for speech and speaker recognition are the MFCCs and perceptual linear predictive (PLP) coefficients both of which are based on the magnitude spectrum of the speech analysis window [1, 2]. In this paper we show how to derive the magnitude spectrum of the voice source signal which describes the air flow through the glottis in voiced speech. The voice source signal is a function of the shape and the movements of the vocal folds and has been shown to give consistent variation between speaker types [3]. Using the magnitude spectrum of the voice source allows the derivation of voice source (mel-frequency) cepstrum coefficient (VSCC). This gives us two sets of coefficients, MFCCs derived from the magnitude spectrum of the speech and VSCCs derived from the magnitude spectrum of the voice source signal.

The estimation of the voice source signal is essentially that of blind system estimation. If we rely on linear prediction modelling of the speech production then we are assuming that the voice source has a flat spectrum and the source becomes encoded in the estimated vocal-tract transfer function. To avoid this, we apply closed phase LPC analysis to circumvent the problem and solve the blind nature of the estimation process.

The shape and form of the voice source signal has been the subject of many studies in the past. An early study estimated the voice source signal by using an inverse LCR circuit network analysis and the vocal fold opening area, measured by video [4]. Further developments replaced the LCR circuit network with linear filters identified using covariance analysis [5, 6]. Other methods have been suggested for voice source analysis such as the two channel analysis approach, using electroglottography [7]. Identifying the glottal closure instant (GCI) in each larynx cycle [8, 9] makes it possible to perform closed-phase analysis [10], so that the vocal tract filter is evaluated separately from the source. We have recently developed the dynamic programming projected phase-slope algorithm (DYPSA) [11], based on the group delay function [9, 12], to detect GCIs for this purpose.

This paper is organized as follows. In Section 2, the VSCC feature extraction process is presented and Section 3 defines the speaker identification task and describes how the MFCC and VSCC classifiers are combined. The experimental results are given in Section 4 and the paper is concluded in Section 5.

2. VOICE SOURCE CEPSTRUM COEFFICIENTS

The voice source is represented by cepstrum parameters so as to avoid difficulties associated with inverse-filtering including low-frequency distortion. Fig. 1 shows how the voice source cepstrum coefficients (VSCC) are extracted. The closed-phase portion of the voiced speech is identified using the DYPSA algorithm. An autoregressive (AR) model of the vocal tract is estimated using multi-cycle closed-phase analysis [10] and its spectral envelope evaluated. The envelope is then passed through a mel-filter bank, the logarithm taken, and the discrete cosine transform applied [13] to produce vocal-tract cepstrum coefficients (VTCC). This is different from LPC cepstrum since the mel-scale is applied in the frequency domain. The VSCC are then computed by subtracting the VTCC from the MFCC extracted from the same frame [1].

Following Fig. 1, detailed description of the feature extraction is as follows. We detect if the frame is unvoiced or voiced. For unvoiced frames (middle path), the AR spectral envelope covariance LPC coefficients of the frame is extracted. For voiced frames (left path), the closed phases in the frame are identified by using the DYPSA algorithm [11]. The DYPSA algorithm provides an accurate detection of glottal closure instants. The closed phase LPC coefficients can be estimated in the interval following the glottal closure instants. We avoid including the glottal closure instants in this analysis since the detection provided by of the DYPSA algorithm is sufficiently accurate. We do not attempt to locate the glottal opening instants, since they are less pronounced in the voice source signal, but we assume the closed phase to be the first 33% of the larynx cycle [14]. We have found that the LPC analysis is not sensitive to

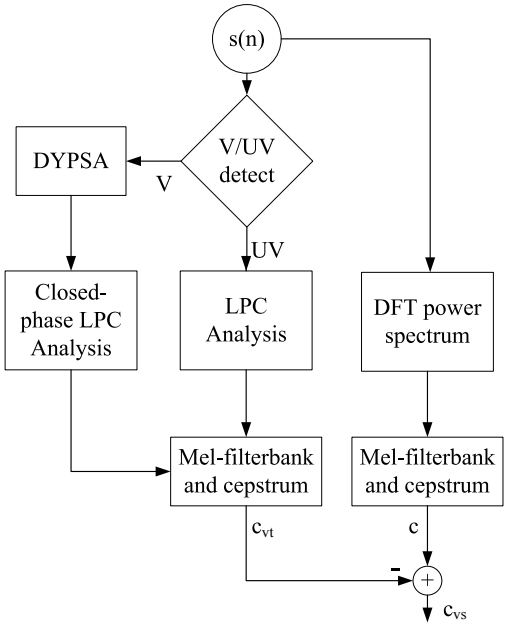

Fig. 1. Computation of voice source cepstrum coefficients using closed-phase analysis.

this choice. Excluding a part of the closed phase only means that we have less data to estimate the parameters and since the opening is gradual (compared to the closure) the effect of including a small portion of the start of the open phase is rarely serious.

The closed-phase covariance analysis is performed on a fixed 32 ms frame with a frame increment of 10 ms. The LPC parameters a_p are evaluated only for the closed-phase portion of the speech where the voice source is assumed to be zero [10, 15]. For unvoiced frames, we perform covariance analysis on the entire frame since this is consistent with the modelling assumptions. For speech sampled at 16 kHz we use a prediction order $P = 16$ [16].

The spectral envelope of the speech signal is evaluated as,

$$S(k) = \frac{\sigma_u}{\sum_{p=0}^{P} a_p e^{-j2\pi kp/N_s}}, \tag{1}$$

where σ_u is the magnitude of the closed-phase LPC residual and N_s determines the frequency resolution of the envelope. We apply a mel-filter bank to attain the filter outputs,

$$Y(r) = \sum_{k=0}^{N_a-1} S(k) M_r(k), \tag{2}$$

where $M_r(k)$ is the r-th mel-filter and $r \in \{1, \ldots, 26\}$ [13]. The vocal tract cepstrum coefficients are then computed as the cosine transform of the logarithm of the filterbank output,

$$c_{vt}(m) = \sum_{r=1}^{N_r-1} \log(Y(r)) \cos\Big(\frac{(2r+1)m\pi}{2N_r}\Big), \tag{3}$$

where $m = \{1 \ldots N_c\}$ and $N_c = 12$ and we discard $c_{vt}(0)$. We consider the VTCC to represent the vocal tract, since they are derived during a period when there is no input from the voice source.

If we make the loss-less tube assumption then the $c_{vt}(m)$-coefficients represent the vocal tract for both voiced and unvoiced frames. Parallel to the above processing we extract mel-frequency cepstrum coefficients [13] (rightmost path in Fig. 1) from the same frame of speech and denote as $c(m)$. The inverse-filtering that is normally used to extract the voice source signal is equivalent to subtraction in the cepstrum domain so the voice source signal is represented as

$$c_{vs}(m) = c(m) - c_{vt}(m). \tag{4}$$

3. SPEAKER IDENTIFICATION

Each speaker is represented by a Gaussian Mixture Model (GMM), ξ, formed from the training utterances using the EM-algorithm [17]. Each component in the GMM is represented by a weight, a mean vector and a diagonal covariance matrix. We vary the number of mixture components from 2 to 64 to assess how many components are needed for each set of coefficients.

A score for each test-utterance, indexed as i, is evaluated as the summed log-likelihood of the sequence of feature vectors from that utterance. An identity is assigned to each utterance according to the highest log-likelihood score.

$$\hat{\imath} = \arg\max_{\xi} L(i, \xi). \tag{5}$$

where L is the summed log-likelihood.

We present results for the following three feature sets.

1. MFCC, Mel-Frequency Cepstrum Coefficients, $c(m)$;
2. VSCC, Voice Source Cepstrum Coefficients, $c_{vs}(m)$;
3. VTCC, Vocal Tract Cepstrum Coefficients, $c_{vt}(m)$.

Results based on classifier combination of two coefficient sets are also presented. Two classifiers are combined using a weighted sum of likelihoods [18]. So, for example, instead of basing the decision on the likelihood of the MFCC or VSCC classifier only, the two likelihoods are added together,

$$L_w(i, \xi) = \theta L_{mfcc}(i, \xi) + (1 - \theta) L_{vscc}(i, \xi). \tag{6}$$

where θ is an undetermined weight factor, i is the test utterance index and ξ is the speaker model.

4. EXPERIMENTS

The evaluation of the proposed feature extraction method was done by text-independent closed-set speaker identification experiments on the TIMIT [19] and YOHO [20] databases. The TIMIT database contains speech from 630 speakers. We defined the first 8 utterances for each speaker as the training set and the remaining 2 sx utterances as the test set. The 10 utterances were all recorded during the same session at 16 kHz sampling frequency. The YOHO database consists of 138 speakers each with 24 training utterances and 40 test utterances recorded in different sessions. The data was recorded in a normal office environment at 8 kHz sampling frequency with 16 bit per sample. We divided each utterance into 32 ms frames with 10 ms frame increment and derived 12 coefficients for each method (MFCC, VSCC and VTCC) for each frame excluding the 0th coefficient. No attempt was made to pre-recognize voice activity or to enhance the speech before the feature extraction.

4.1. Model size

Fig. 2 shows the misclassification rate experiments conducted on the TIMIT database using 2, 4, 8, 16, 32, and 64 mixtures each classifier depending only on the MFCC, VSCC, and VTCC feature sets. The bars show the test set misclassification rates for each of the three classifiers, the baseline classifier using the MFCC, VSCC and 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. The results for the 32 mixture component cases is shown in the upper half of Table 1 for the TIMIT and YOHO databases. We can see that the MFCC feature set outperforms the VSCC and the VTCC feature sets when used on their own.

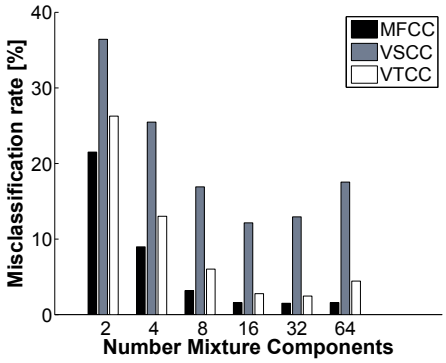

Fig. 2. Speaker identification experiments for 630-speaker TIMIT subset using MFCC, VSCC and VTCC feature sets and different number of Gaussian mixture components.

4.2. Combination of classifiers

We combined the MFCC, VSCC and VTCC classifiers in pairs and show the results in Fig. 3 and 4 for the TIMIT and YOHO databases respectively. It can be seen from the traces that the combination of the MFCC and VTCC classifiers does not improve the misclassification rate of the MFCC classifier significantly, whereas combining the VSCC classifier with the MFCC classifier results in a much lower misclassification than that of the MFCC classifier.

The test-set misclassification rate γ is the ratio of all incorrectly identified speakers to the total number of test utterances [21]. Error rates are presented in the form $\gamma \pm e$ where e is an estimate of the standard error.

For the TIMIT database 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 likelihood with $\theta = 0.4$.

The YOHO database presents a more realistic challenge to speaker identification than the TIMIT database. The speech is recorded in a real office environment and the data is recorded for each speaker in different sessions increasing the intra-speaker variability. The lowest misclassification rate was for the YOHO test was $10.07 \pm 0.41\%$, also achieved by combining the MFCC and VSCC classifiers with the VSCC likelihood weight $\theta = 0.4$. The lowest misclassification rate achieved by the combination of VSCC and VTCC was $10.45 \pm 0.41\%$.

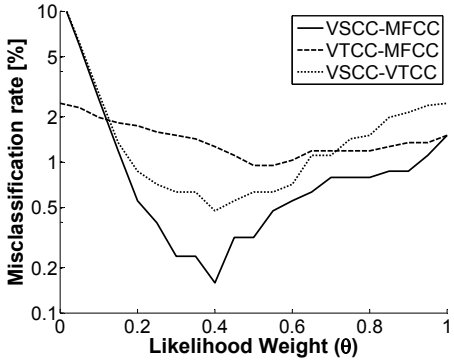

Fig. 3. Combinations of the VSCC, VTCC and MFCC classifiers for the TIMIT database.

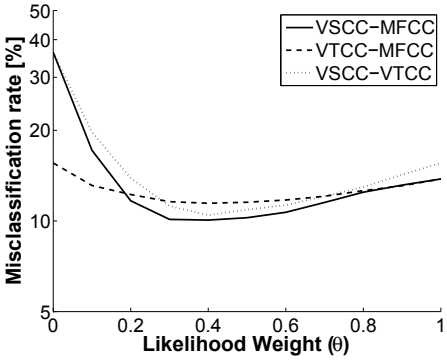

Fig. 4. Combinations of the VSCC, VTCC and MFCC classifiers for the YOHO database.

A summary of the results is presented in Table 1 where the test set misclassification rate is given for each classifier. The combination results are given using the best possible combination weight, but it should be noted that the values of these weights have not been optimized using a specific training or a validation set.

Table 1. Misclassification rate using the three feature sets and combined classifiers. Each classifier uses 32 mixture components applied to the TIMIT and YOHO databases

	Misclassification rate, $\gamma \pm e$ [%]	
Classifier	TIMIT	YOHO
MFCC	1.51 ± 0.34	13.79 ± 0.46
VSCC	12.94 ± 0.95	36.30 ± 0.65
VTCC	2.46 ± 0.44	15.58 ± 0.49
MFCC+VSCC	0.16 ± 0.11	10.07 ± 0.41
MFCC+VTCC	0.95 ± 0.27	11.45 ± 0.43
VTCC+VSCC	0.48 ± 0.19	10.45 ± 0.41

4.3. Comparison to other work

A time-domain voice source feature set was developed by M.D. Plumpe et.al. [22]. The method uses twelve voice source coefficients. Seven coefficients are derived from the coarse structure of the voice source signal found using larynx synchronous, piecewise continuous function fitting [23]. The remaining five coefficients are based on the fine structure of voice source signal derived from the

error between the fitted model and the measured signal. The experiments were done using a 168 speaker subset of the TIMIT database and cross gender tests were not performed. The average male and female misclassification rate was reported as 28.64%. When combined with 14 LPC cepstrum parameters the method achieved 6.85% misclassification rate. The VSCC features achieve 5.06% misclassification rate on the same 168 speaker TIMIT subset. On the more challenging full 630 speaker TIMIT set the result is 12.95% for the VSCC classifier and combined with MFCC features the misclassification rate is 0.16%. Another promising approach to voice source feature extraction for speaker recognition has been presented by K.S.R. Murty and B. Yegnayaranana [24] but their results are based on a different database so direct comparison is impossible.

5. CONCLUSIONS

This study has developed a novel feature set for speaker identification and shown how a standard speaker identification system can be significantly improved. The results are also better than other attempts suggested in the literature. The voice source cepstrum coefficients were presented and applied to a closed-set speaker identification task. We used the segmentation provided by the techniques developed in [11] and applied closed-set AR modelling to identify the vocal tract response. We characterized the voice-source with cepstrum features for speaker identification, by subtracting the cepstrum representation of the AR spectral envelope from the mel-frequency cepstrum of the speech frame. The results show that there is discriminative power in the voice source and misclassification rate was improved when combined with the over all mel-frequency cepstrum representation of the speech.

6. REFERENCES

[1] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continously Spoken Sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[2] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[3] I. Karlsson, "Glottal Waveform Parameters for Different Speaker Types," *STL-QPSR*, vol. 29, no. 2-3, pp. 61–67, 1988.

[4] R. L. Miller, "Nature of the Vocal Chord Wave," *J. Acoust. Soc. Am.*, vol. 31, pp. 667–677, 1959.

[5] H. W. Strube, "Determination of the Instant of Glottal Closure from the Speech Wave," *J. Acoust. Soc. Am.*, vol. 56, no. 5, pp. 1625–1629, 1974.

[6] D. Y. Wong, J. D. Markel, and A. H. Gray, Jr., "Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 4, pp. 350–355, Aug 1979.

[7] A. K. Krishnamurthy and D. G. Childers, "Two-Channel Speech Analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 4, pp. 730 – 743, Aug 1986.

[8] Y. M. Cheng and D. O'Shaughnessy, "Automatic and Reliable Estimation of Glottal Closure Instant and Period," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 12, pp. 1805 – 1815, Dec 1989.

[9] P.S. Murthy and B. Yegnanarayana, "Robustness of Group-Delay-Based Method for Extraction of Significant Instants of Excitation from Speech Signals," *IEEE Trans. on Speech and Audio Proc.*, vol. 7, no. 6, pp. 609–619, Nov 1999.

[10] 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*, vol. 33, no. 1, September 1989.

[11] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of Glottal Closure Instants in Voiced Speech using the DYPSA Algorithm," *IEEE Trans. on Speech and Audio Proc.*, vol. 15, no. 1, pp. 34–43, Jan 2007.

[12] D. M. Brookes, P. A. Naylor, and J. Gudnason, "A Quantatative Assessment of Group Delay Methods for Identifying Glottal Closures in Voiced Speech," *IEEE Trans. on Speech and Audio Proc.*, vol. 14, no. 3, pp. 456–466, May 2006.

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

[14] E. R. M. Abberton, D. M. Howard, and A. J. Fourcin, "Laryngographic Assessment of Normal Voice: A Tutorial," *Clinical Linguistics and Phonetics*, vol. 3, no. 3, pp. 263–296, 1989.

[15] A. Kounoudes, P. A. Naylor, and M. Brookes, "Automatic Epoch Extraction for Closed Phase Analysis of Speech," in *Int. Conf. Digital Signal Processing*, July 2002, vol. 2, pp. 979 – 983.

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

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

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

[19] 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*, pp. 93–99, Feb 1986.

[20] J. P. Campbell, Jr., "Testing with the YOHO CD-ROM Voice Verification Corpus," *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, pp. 341–344, May 1995.

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

[22] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification," *IEEE Trans. on Speech and Audio Proc.*, vol. 7, no. 5, pp. 569–586, Sep 1999.

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

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