ON SEPARATING GLOTTAL SOURCE AND VOCAL TRACT INFORMATION IN TELEPHONY SPEAKER VERIFICATION

Tomi Kinnunen

Dept. of Computer Science and Statistics
Univ. of Joensuu, FINLAND

Paavo Alku

Dept. of Signal Processing and Acoustics
Helsinki Univ. of Technology, FINLAND

ABSTRACT

The popular mel-frequency cepstral coefficients (MFCCs) capture a mixture of speaker-related, phonemic and channel information. Speaker-related information could be further broken down according to articulatory criteria. How these underlying components are exactly mixed in the features is not well understood. To this end, in this paper we aim at separating the spectra of glottal source and vocal tract using glottal inverse filtering, with an application to speaker recognition over telephone lines. Our experiments on the 10sec-10sec condition of the NIST 2006 SRE corpus suggest that the mel-frequency cepstrum of the voice source is not too useful for recognizing speakers. On the contrary, fusing the vocal tract spectrum with conventional MFCCs improves accuracy, suggesting that vocal tract information should be enhanced.

Index Terms— Glottal inverse filtering, speaker recognition, source-filter model, mel-frequency cepstrum

1. INTRODUCTION

Cepstral features [1] such as mel-frequency cepstral coefficients (MFCCs) and linear predictive cepstral coefficients (LPCCs) have been the dominant features for a long time in speaker recognition. Cepstral features mainly capture properties of vocal tract and they are standard features in speech recognition since the vocal tract contains the major "message-bearing" articulators. Since MFCCs capture a mixture of phonemic and speaker-related information their use has resulted in excellent performance in speaker recognition. The question of how these two factors are exactly mixed in the feature coefficients is largely unanswered.

Since MFCCs depend on the spoken text, their use requires sufficient phonetic coverage in the training and test data for reliable recognition. For instance, state-of-the-art speaker recognition systems utilizing MFCCs achieve respectable performance (error rates $\sim$ 2 %) when *minutes* or *tens of minutes* of training and test data is available [2]. In contrast, the results deteriorate by an order of magnitude (error rates $\sim$ 20 %) when only 10 seconds of data is available. There is a clear need for features that are truly text-independent and whose performance is less dependent on the amount of data.

A potential candidate for text-independent features would be *voice source* features, that is, features that characterize the source of voiced sounds known as the *glottal volume velocity waveform* or simply the *glottal flow*. The most popular voice source feature is the rate of vibration of the vocal folds, referred to as the *fundamental frequency* (F0). Other parameters describe the shape of the glottal pulse, such as the duration of the closing phase, and the corresponding frequency domain effects. These contribute to voice quality which can be described for example, as modal, breathy, creaky or pressed [3]. It can be hypothesized that these parameters also carry useful speaker-specific information.

When estimating the glottal flow, a common assumption is to consider the glottal source and the vocal tract filter to be linearly related and independent from each other. These simplifications allow one to first estimate the vocal tract filter parameters using, for example, the well-known linear prediction (LP) model (e.g. [1]). Once the vocal tract filter is known, or reasonably-well estimated, an estimate of the glottal flow can be obtained by filtering the original signal using the inverse model of the tract. The process is referred to as *glottal inverse filtering* and illustrated in Fig. 1. The glottal flow can be parameterized, for instance, by fitting a physical glottal flow model to the inverse filtered signal [4]. Other approaches include wavelet transforms [5], residual phase [6], cepstral coefficients [7, 8] and higher-order statistics [8] to mention a few.

Often the residual signal obtained from autocorrelation LP analysis (e.g. [5, 6, 8, 9]) is used as a crude estimate of the glottal flow (derivative) waveform. An alternative approach uses *closed-phase covariance analysis* during the portions when the vocal folds are closed [4, 7, 10]. This leads to improved estimates of the vocal tract and glottal flow. On the other hand, accurate detection of closed phase is required and can be difficult in the presence of noise or in soft phonation.

In this paper, our main goal is to decompose the magnitude spectral features of speech into the underlying processes corresponding to the vocal tract filter and the glottal voice source. This would lead to a better understanding of

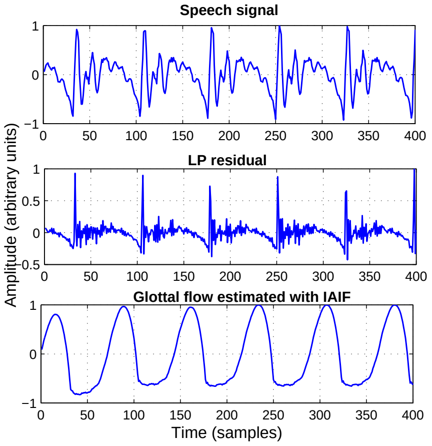

Fig. 1. An example of glottal inverse filtering. A speech signal (top), LP residual (middle) and IAIF output (bottom).

the speaker discrimination power of the individual components. Interestingly, an approach close to ours was utilized in a recent study [7], in which the closed-phase covariance analysis was used for extracting the vocal tract information. Both the voice source and the filter were parameterized using mel-cepstrum. The vocal tract cepstrum was computed from the spectral envelope obtained from the closed-phase analysis, whereas the source cepstrum was computed as the difference between the MFCCs and the vocal tract cepstrum.

Our work differs from [7] in that we use a more straightforward *iterative adaptive inverse filtering* (IAIF) method [11] rather than a closed phase covariance analysis in the source-tract separation. In this way, we avoid the most critical part of the closed phase analysis, the glottal closure detection, and we obtain both the vocal tract and the source parameters simultaneously with a smaller computational cost. Furthermore, the experiments of [7] were reported on TIMIT and YOHO databases consisting of high-quality studio recordings. We report our results on the more challenging NIST 2006 corpus with telephone transmission and channel variability.

2. SOURCE AND FILTER SEPARATION

A flowchart of the feature extraction process is shown in Fig. 2. The IAIF method helps in separating the source- and filter-related features. We extract three synchronous feature streams that aim at capturing complementary properties of the speech signal. In the following subsections we detail each of the components.

2.1. Glottal Flow Estimation Using IAIF

The IAIF method [11] is a straightforward glottal inverse filtering method that estimates the glottal flow. IAIF estimates the contribution of the glottal source to the speech spectrum with low-order (order q) all-pole modeling during several fundamental periods. By canceling the estimated effect of the source, an all-pole model of order p for the vocal tract is computed either with LP [11] or with *discrete all-pole* (DAP) model [12]. This computation is executed in a repetitive manner consisting of two phases. The detailed description of the IAIF algorithm including its flow diagram can be found in [11, 12]. In this paper, we have selected to use the LP as a vocal tract model due to its better computational efficiency. Typical orders of the vocal tract and glottal source filters are $8 \leq p \leq 14$ and $2 \leq q \leq 4$, respectively.

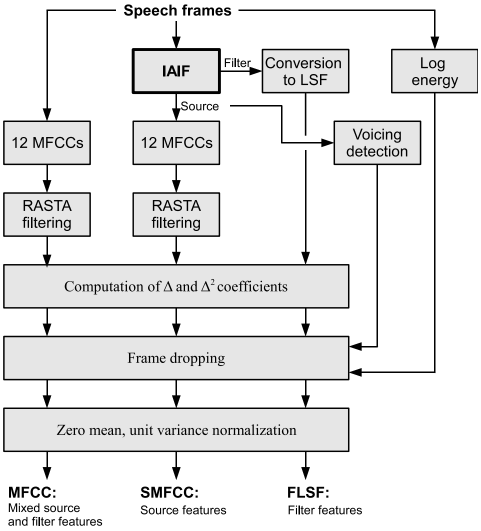

Fig. 2. Using IAIF to separate the source (ie. the glottal flow) and the filter (ie. the vocal tract) components.

It is also worth emphasizing that even though IAIF uses LP as a computational tool, it is, in principle, greatly different from the approach in which the LP residual is treated as a crude estimate of the glottal flow. In the IAIF computation, namely, the spectral decay of the estimated glottal excitation is allowed to vary, thus enabling mimicking the spectral behaviour of the true glottal flow generated by the vocal fold fluctuation. In turn, the LP residual is always spectrally flat due to use of a single LP analysis and as a result, it does not reflect properly the frequency domain features of the true glottal excitation.

2.2. Feature Extraction

We consider three feature sets in this study as illustrated in Fig. 2. Our baseline features, shown on the leftmost signal path are the conventional MFCCs. The processing steps shown in the figure are fairly standard in current speaker recognition front-ends for telephone quality speech. Firstly, 12 MFCCs are extracted from 30 millisecond Hamming-windowed frames with 50 % overlapping. The MFCCs are computed using a 27-channel filterbank, followed by logarithmic compression and discrete cosine transform (DCT). The filterbank consists of triangular bandpass filters spaced uniformly on the mel-frequency scale defined as $f_{\mathrm{Mel}} = (1000/\log_{10} 2) \cdot \log_{10}(1 + f_{\mathrm{Hz}}/1000)$. Here, 1000 Hz is chosen to correspond to 1000 Mels. To improve robustness against convolutive distortions, the MFCC streams are processed with *RelAtive SpecTrAl* (RASTA) filter. Next, the delta (Δ) and double-delta (Δ^2) coefficients are computed using a differentiator and appended to the feature vectors. Finally, only voiced frames are retained, followed by normalization of each feature to have zero mean and unit variance over the entire utterance. This last normalization further reduces channel effects and equalizes the numeric ranges of the features.

The second feature set, which we dub as the *source mel-frequency cepstral coefficients* (SMFCCs), aims at capturing the frequency-domain characteristics of the voice source. The computation follows exactly the same steps as MFCC processing, except that the input is the estimated glottal flow rather than the original speech frame. We had two motivations to study such a feature in the beginning of this work. Firstly, the method is straightforward and readily integrable into existing systems. Secondly, comparing accuracies of the MFCC and SMFCC features would lead to better understanding of the intrinsic speaker discrimination power of the source and filter features.

The third and last feature set aims at capturing the frequency-domain characteristics of the vocal tract filter. These features are derived by converting the filter coefficients obtained from IAIF into an equivalent but more robust representation of the *line spectral frequencies* (LSFs) (e.g. [1]). In order to make explicit distinction from typical computation of LSFs with a single LP analysis on a speech frame, we refer to the new feature set as the *filter line spectral frequencies* (FLSFs). Note that, unlike MFCCs and SMFCCs, the FLSFs neither undergo psychoacoustically motivated transformations nor processed by RASTA filtering. The RASTA filter is designed for cepstral and log-spectral features to attenuate the low and high modulation frequencies. It is not obvious whether such procedure applies to LSFs.

2.3. Frame Dropping

Frame dropping uses a combination of energy and periodicity information to select voiced frames. Both the energy detector and the periodicity detector make binary decision on each frame, which are then combined by logical "AND" operator to give the final frame labels. Short-term energy is measured from the original frame and normalized by the maximum energy over all frames, followed by thresholding (here, we set the energy difference to 30 dB). For the voicing detector, we first compute the autocorrelation sequence of the glottal flow $g[n]$ as $R_g[l] = \sum_{n=0}^{N-l-1} g[n]g[n-l]$ for $l_{\min} \leq l \leq l_{\max}$ and detect the lag corresponding to the maximum value. By denoting this lag as l^*, the frame is marked as voiced if the energy-normalized autocorrelation value $R_g[l^*]/R[0]$ exceeds a pre-set threshold value (here, 0.65). The lag range $[l_{\min}, l_{\max}]$ is set according to expected fundamental periods (here we consider F0 values from 80 Hz to 500 Hz).

3. EXPERIMENTAL SETUP

We report our results on the 10sec-10sec condition of the 2006 NIST speaker recognition evaluation (SRE) corpus. Each training and test utterance contains approximately 10 seconds of speech. The condition consists of 33555 verification trials (3064 genuine speakers and 30491 impostors) from 732 target speakers (415 females and 316 males).

We use a standard Gaussian mixture classifier with diagonal covariance matrices [13]. Two gender-dependent universal background models (UBMs) are trained with the expectation maximization (EM) algorithm from the 1-conversation training files of the NIST 2004 SRE corpus, including 246 males and 370 females. A relevance factor of 16 is used to adapt the target models. Only the mean vectors are adapted. In the recognition phase, 10 top-scoring Gaussians from the UBM are used in the fast log-likelihood ratio computation [13].

In the experiments, the speech signal is high-pass filtered with a 50-tap linear-phase FIR filter with a cut-off frequency of 60 Hz to remove possible low-frequency noise components that could adversely affect inverse filtering. We fixed the source and filter LP analysis orders in the IAIF method to $p = 12$ and $q = 4$. The value $p = 12$ was mainly chosen so as to have an equal dimensionality for all the three feature sets considered.

In evaluating accuracy, we use two well-known metrics. The first one, *equal error rate* (EER), corresponds to the decision threshold that gives equal false acceptance rate (FAR) and false rejection rate (FRR). The second one, *minimum detection cost function* (MinDCF), punishes heavily false acceptances. It is used in the NIST SRE evaluations and defined as the minimum value of the function $0.1 \times \mathrm{FRR} + 0.99 \times \mathrm{FAR}$.

4. RESULTS

We first study different combinations of the base coefficients and delta features for each feature set as shown in Table 1. The conventional MFCC features (EER 25$\sim$30 %) clearly outperform the source-related SMFCC features (EER $\approx$ 40 %) whereas the accuracy of the filter-related FLSF features (EER $\approx$ 30 %) is only slightly worse than that of the MFCCs. Appending the first order deltas with base coefficients always improves accuracy, whereas adding the second order deltas does not make much difference. The detection error trade-off (DET) curves, not shown here due to page limitations, confirmed this last observation.

Table 1. Accuracy on the 10sec-10sec condition of the NIST 2006 SRE corpus, equal error rates (EER %).

Feature	Model order (# Gaussians)		
	$M = 32$	$M = 64$	$M = 256$
Baseline features			
MFCC	28.3	27.7	**27.2**
MFCC + Δ	27.8	26.4	**25.8**
MFCC + Δ + Δ^2	27.7	27.0	**26.0**
Source features			
SMFCC	**40.3**	40.5	40.9
SMFCC + Δ	39.7	39.5	**39.1**
SMFCC + Δ + Δ^2	39.1	39.3	**38.8**
Filter features			
FLSF	31.5	30.7	**30.3**
FLSF + Δ	30.9	30.2	**29.1**
FLSF + Δ + Δ^2	31.6	30.6	**29.9**

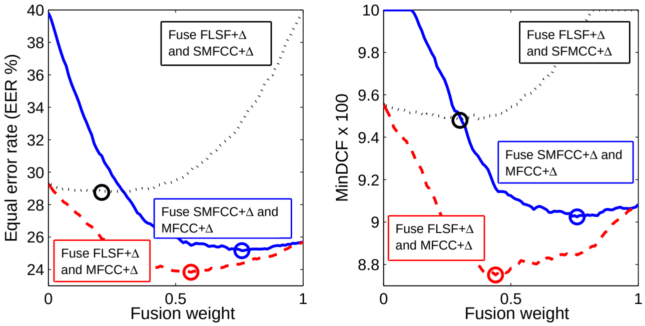

Fig. 3. Fusing different features by linear match score weighting. Optimum points are indicated by circles.

Next, we study fusion of the different feature set pairs by considering the base coefficients with first order deltas and having GMMs with 256 Gaussians. We normalize the scores of each classifier to have zero mean and unit variance, followed by linear score combination of the form $w\text{LLR}_1 + (1-w)\text{LLR}_2$. Here, $\text{LLR}_{\{1,2\}}$ are the log-likelihood ratios of the individual classifiers and $0 \leq w \leq 1$ is the combination weight. Both the EER and MinDCF values are displayed in Fig. 3 as a function of the combination weight. The SMFCC features do not fuse with the MFCC or the FLSF features, whereas fusing MFCC and FLSF features slightly improves accuracy. It is interesting to notice that the optimum fusion weight in this case is close to $w = 0.5$ for both error criteria, suggesting that both features are equally important. Regarding relative accuracies of individual features, the order of the source- and filter-related features is consistent with [7] for entirely different corpora. Enhancing the filter rather than source contribution in the magnitude spectrum seems therefore helpful.

In this study we used a relatively simple GMM-UBM system, yielding accuracy EER$\sim$25 % for MFCCs. More advanced systems such as *joint factor analysis* (JFA) compensated GMMs using MFCCs achieve EER$\sim$17 % on the same data [14]. In future it would be interesting to study the proposed features with JFA. The results should be also validated under broader test conditions by varying training/test data durations, channel conditions and level of text mismatch.

5. CONCLUSIONS

We have studied separation of the voice source and the vocal tract in speaker recognition. Preliminary results suggest that the cepstrum of the voice source is not a useful feature for telephony speech. In contrary, vocal tract features were competitive with MFCCs, and fusing these two features improved accuracy, indicating importance of the vocal tract spectrum. Our future plan includes studying alternative parameterizations for the filter (cepstrum, psychoacoustically motivated transformations) and exploring time-domain parameters of the source. Comparisons with other popular features like PLPs, LFCCs and LPCCs would also be interesting.

6. REFERENCES

[1] T.F. Quatieri, *Discrete-Time Speech Signal Processing - Principles and Practice*, Prentice-Hall, 2002.

[2] "NIST 2008 SRE results page," September 2008, `http://www.nist.gov/speech/tests/sre/2008/official_results/index.html`.

[3] C.Y. Espy-Wilson, S. Manocha, and S. Vishnubhotla, "A new set of features for text-independent speaker identification," in *Interspeech 2006*, Pittsburgh, Sept. 2006, pp. 1475–1478.

[4] M.D. Plumpe, T.F. Quatieri, and D.A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech & Audio Proc.*, vol. 7, no. 5, pp. 569–586, September 1999.

[5] N. Zheng, T. Lee, and P.C. Ching, "Integration of complementary acoustic features for speaker recognition," *IEEE Sign. Proc. Lett.*, vol. 14, no. 3, pp. 181–184, March 2007.

[6] K.S.R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Sign. Proc. Lett.*, vol. 13, no. 1, pp. 52–55, Jan. 2006.

[7] J. Gudnason and M. Brookes, "Voice source cepstrum coefficients for speaker identification," in *ICASSP 2008*, Las Vegas, March-April 2008, pp. 4821–4824.

[8] M. Chetouani, M. Faundez-Zanuy, B. Gas, and J.L. Zarader, "Investigation on LP-residual presentations for speaker identification," *Pattern Recognition*, vol. 42, pp. 487–494, 2009.

[9] S.R.M. Prasanna, C.S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Comm.*, vol. 48, pp. 1243–1261, 2006.

[10] R.E. Slyh, E.G. Hansen, and T.R. Anderson, "Glottal modeling and closed-phase analysis for speaker recognition," in *Proc. Speaker Odyssey 2004*, Toledo, May 2004, pp. 315–322.

[11] P. Alku, H. Tiitinen, and R. Näätänen, "A method for generating natural-sounding speech stimuli for cognitive brain research," *Clinical Neurophysiology*, vol. 110, no. 8, pp. 1329–1333, 1999.

[12] P. Alku, B. Story, and M. Airas, "Estimation of the voice source from speech pressure signals: Evaluation of an inverse filtering technique using physical modelling of voice production," *Folia Phoniatrica et Logopaedica*, vol. 58, no. 2, pp. 102–113, 2006.

[13] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digit. Sign. Proc.*, vol. 10, no. 1, pp. 19–41, January 2000.

[14] P. Kenny, N. Dehak, P. Ouellet, P. Gupta, and P. Dumouchel, "Development of the primary CRIM system for the NIST 2008 speaker recognition evaluation," in *Proc. Interspeech 2008*, Brisbane, Sept. 2008.