Impact of Voice Excitation Features on Speaker Verification

Cemal Hanilci[1], and Figen Ertas[1]

[1]Department of Electronic Engineering, Uludag University, 16059, Bursa, Turkey
chanilci@uludag.edu.tr, fertas@uludag.edu.tr

Abstract

Mel-frequency cepstrum coefficients (MFCC) have been the most popular features used in speaker recognition. It has been recently shown that residual signal estimated through linear prediction (LP) also conveys speaker-specific information, and applied to speaker identification. In this paper, we investigate on the impact of LP-residual cepstrum coefficients (LPRC) on speaker verification along with MFCC and linear predictive cepstrum coefficients (LPCC) as well, and make comparisons of their performance in verification by conducting experiments on NIST 2001 SRE corpus, including modern classifiers. It is shown that LPRC features are as useful as MFCC and LPCC features in speaker verification, and fusing the LPRC, LPCC, and MFCC features in pairs improves the verification performance.

1. Introduction

Cepstral representations of speech signal such as mel-frequency cepstrum coefficients (MFCC) and linear predictive cepstrum coefficients (LPCC) have been widely used in speaker recognition. Both MFCC and LPCC are vocal tract related acoustic features and aim to model spectral envelope or formant structure of the vocal tract [1], [2]. Cepstral features based on short-term spectral analysis are the dominant features used in speaker recognition because the ease of their computation and their effectiveness in the performance of speaker recognition. But there is no feature set that only conveys the information about speaker identity.

Voice excitation features which paremeterize the glottal flow also carry useful informaton about speaker identity. The most popular feature which is related to voice source is the fundemental frequency (F0). The most common approach when estimating the glottal flow is that speech signal is the output of a filter (vocal tract filter) and the input of the filter is the vocal source excitation signal [3]. Vocal tract filter parameters are commonly estimated via well-known linear prediction (LP) analysis. There are several works in the literature which parameterize the glottal flow and used in speaker recognition such as Zheng et.al. [4] have applied the wavelet transform to residual signal, Plumpe et.al. [5] performed inverse filtering, and residual phase [6]. Recently [7] it has been shown that vocal tract spectrum fused with MFCC features improves the speaker recognition accuracy.

Using the residual signal to estimate the glottal flow is a common approach in speech applications as used by [8]. Chetouani et al. [8] has shown that residual signal conveys speaker-specific information and when it is fused with MFCC features, the speaker identification performance improves. However, we use LPRC features in speaker verification. Our work is different from that in [8] in the following ways: i) a detailed analysis of LPRC features (not only the base features but also delta and double delta) on speaker verification performance, ii) analyzing the fusion of classifier scores not only with MFCC but also with LPCC features iii) we use a well-known and modern classifiers, e.g., Gaussian mixture model with universal bacground model (GMM-UBM). Furthermore the experiments of [8] were conducted on NTIMIT and GAUDI databases and NTIMIT database is a telephone speech database but it was recorded in laboratory environment with limited length of training and test utterances for each speaker (there are total of 10 sentences for each speaker where each sentence is approximately 3 seconds length), and GAUDI database which is more realistic compared to NTIMIT but it has a small group of speakers (49 speakers) with different conditions such as language, interval sessions and microphone and as in NTIMIT it consists of studio recordings. We analyze the effect of LP-residual features on a more challenging NIST 2001 corpus [9].

2. Voice Excitation Features

The flowchart of feature extraction procedure is shown in Fig. 1. Three different type of features, MFCC, LPCC, and LP-Residual cepstrum (LPRC), are extracted from speech signal. The detailed description of extraction of each feature set is given in the following subsections.

2.1. Vocal Tract Filter Estimation

According to speech production model speech signal can be modelled as a source filter system where the source excitation signal is filtered by a vocal tract filter. The vocal tract filter can be considered as a p^{th} order filter with transfer function

$$\frac{1}{A(z)} = \frac{1}{1 + \alpha_1 z^{-1} + ... + \alpha_p z^{-p}} \quad (1)$$

where $\alpha_i, i = 1, 2..., p$ are the vocal tract filter parameters. The filter parameters are commonly estimated using LP analysis due to its better computational efficiency. The LP analysis estimates the filter coefficients by minimizing the prediction error. The predicted signal is related to p past samples of signal in the form

$$\hat{s}(k) = \sum_{i=1}^{p} s(k-i) \quad (2)$$

The prediction error (a.k.a. residual signal) is the difference of actual signal and predicted signal, namely

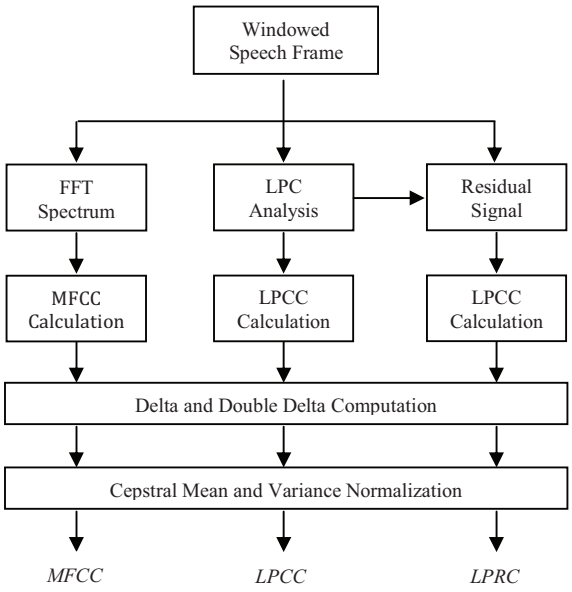

Fig. 1. Flowchart of feature extraction procedure

$$r(k) = s(k) - \hat{s}(k) \tag{3}$$

An example for original speech signal, predicted speech signal by a $p=12^{th}$ order LP analysis and residual signal is given in Figure 2.

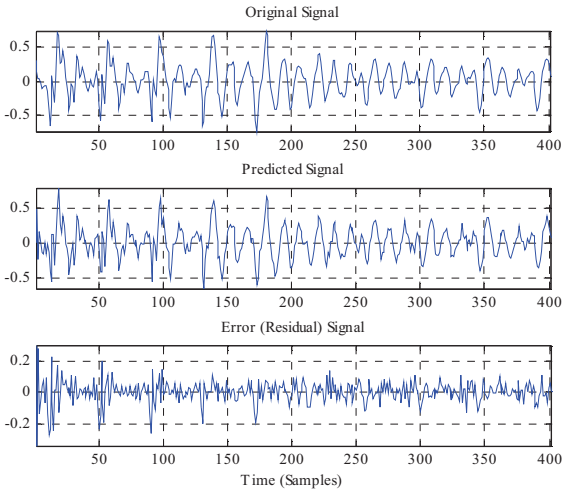

Fig. 2. Prediction of a speech frame via LP analysis and residual signal

The spectral envelope which is the most important part of speech spectrum in speaker recognition that conveys the speaker specific information about the resonance properties of the vocal tract can be estimated using vocal tract filter coefficients. The Fast Fourier Transform (FFT) spectrum (absolute value of Fourier transform) envelope and LP spectrum envelope for a speech frame is given in Figure 3. In Fig. 3 the vocal tract resonances appear as peaks in LP spectrum.

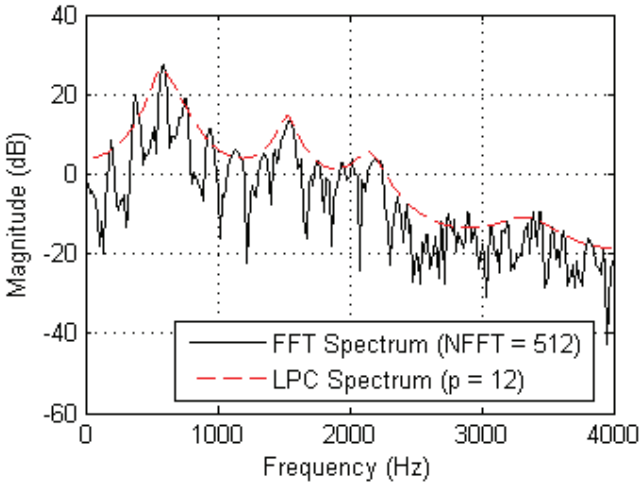

Fig. 3. Spectrum estimation with FFT (solid line) and spectrum envelope estimation with LP (dashed line).

2.2. Extraction of LPCC Features

Once the vocal tract filter coefficients, $\alpha_i, i = 1, 2..., p$ are estimated by LP analysis for each frame of speech the filter coefficients are converted into cepstral coefficients using the following equation:

$$c_m = \alpha_m + \sum_{i=1}^{m-1} \left(\frac{i}{m}\right) c_i \alpha_{m-i}, 1 < m \le p \tag{4}$$

where c_m are the m^{th} cepstral coefficient and p is the order of LP analysis or number of LP coefficients. In the experiments we extract $m=12$ LPCC features from 30 milliseconds Hamming-windowed frames with 10 milliseconds overlap.

2.3. Extraction of LPRC Features

LPRC features are computed using residual signal *r(k)* which is defined in Eq (3). First the LP coefficients are calculated using residual signal and then they converted into cepstral coefficients as described in Eq (4). The only difference between LPCC and LPRC is that LPCC uses the original speech signal whereas LPRC uses the error (residual) signal. We used the same frame and overlap length as in LPCC during the experiments.

2.4. Extraction of MFCC Features

The MFCC features which are shown on the leftmost part of Fig.1 are the baseline features we use in the experiments. 12 MFCCs are extracted from 30 milliseconds Hamming-windowed frames with 10 milliseconds overlap. We compute the MFCCs using a 27-channel triangular filterbank. Logarithmic filterbank outputs are converted into cepstral coefficients by Discrete Cosine Transform (DCT). For more details about MFCC extraction refer to [5].

For the three different feature set we analyze the effect of appending the first and second order derivatives (delta and double delta) of features on speaker recognition performance.

The delta features are computed by convolving the features with the kernel $h = [1\ 0 - 1]$. Double delta features are computed by applying the same kernel to the delta features.

3. Experimental Setup

In the experiments we use NIST 2001 speaker recognition evaluation (SRE) corpus which consists of conversational telephone speech in English. NIST 2001 SRE contains 174 speakers (74 males and 100 females) with total number of 22,418 trials (2,038 genuine and 20,380 impostor trials). The training data for each speaker is with the amount of 2 minutes and the length of test segments vary from a few seconds up to 1 minute.

We use a well-known and modern classifier, Gaussian mixture model with universal background model (GMM-UBM) which is introduced by Reynolds, Quatieri, and Dunn (2000). For GMM-UBM we use diagonal covariance matrices and only mean vectors are adapted with a relevance factor of 16. Two gender-dependent background models are trained from the development set of database. For the LP analysis we fixed the predictor order *p = 12*.

In the experiments, we used Equal Error Rate (EER) as the evaluation metric. EER corresponds to threshold which gives equal false acceptance rate (FAR) and false rejection rate (FRR). As the second evaluation metric we used *minimum detection cost function* (MinDCF) which is used in the NIST speaker recognition evaluations and defined as the minimum value of the function $0.1 \times FRR + 0.99 \times FAR$. At the final stage of experiments we used score fusion of feature set pairs. We used FOCAL toolkit for score fusion.

4. Results

Different combinations of features have been studied in the experiments first. Table 1 shows the EERs and MinDCF values for each feature set with delta and double delta features. The number of Gaussians (Model Order) for GMM-UBM is fixed to 64 and 256. As it seen from the table LPCC features yields slightly better recognition accuracy than the MFFC features and LPRC base features (without appending delta and double delta) (EER 17~18%) outperform the MFCC (EER 17~19%) and LPCC (EER 18~19%) base features in terms of EER. Appending the delta and double delta improves the recognition accuracy for all feature sets. Clearly the recognition rates for all feature sets are very close to each other and the most important and interesting thing is that LPRC features yields recognition rate almost as good as MFCC features. The feature sets show the same behaviour in terms of MinDCF values as in EERs. Again LPCC yields slightly better MinDCF value than MFCCs and LPRC yields slightly worse value than MFCCs and appending first and second order derivatives improves the system performance. The detection error trade-off (DET) curves for MFCC+Δ+ΔΔ, LPCC+Δ+ΔΔ and LPRC+Δ+ΔΔ with 256 Gaussians GMMs are given in Fig. 4.

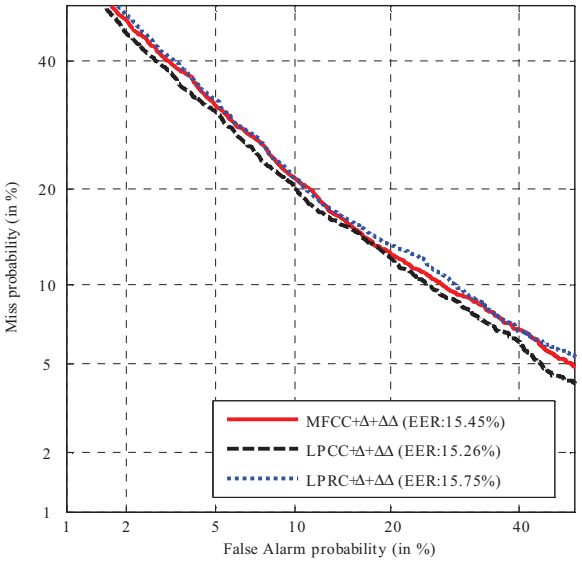

Fig. 4. DET curves of different type of features

Next we consider the fusion of scores of different feature set pairs by linear score weighting. We have used FoCal toolkit to optimize fusion weights. We fused the scores of base features with appended delta and double delta features. The MinDCF values for fused feature set pairs and the DET curves for conventional MFCC and fused feature set pairs (MFCC-LPCC, MFCC-LPRC and LPCC-LPRC) are given in Table 2 and Figure 5, respectively.

Table 1. MinDCF values for fused feature set pairs for M=64 and M=256 mixtures of GMM-UBM

Feature Set	# Gaussians	
	M=64	M=256
MFCC-LPCC	6.65	**6.22**
MFCC-LPRC	6.66	6.27
LPCC-LPRC	6.82	6.38

It can be seen from Table 2 and Figure 5 that fusing the LPCC and LPRC features with the MFCC features slightly improves the recognition accuracy both in terms of MinDCF values and EERs. The results confirm that glottal flow features (LPRC) achieves the recognition accuracy as good as or slightly worse than MFCC and LPCC. Chetouani et al. (2009) showed that LPRC features did not yield good identification rate and suggested to be fused with MFCC. Our results show that LPRC features can be used as the baseline features for speaker verification, achieving recognition rate as well as that of MFCC. All fused features outperform the conventional MFCC features.

Table 2. EERs and MinDCF values of different feature sets for *M=64* and *M=256* mixtures of GMM-UBM

	EER (in %)		**MinDCF x 100**	
	# Gaussians		# Gaussians	
Feature Set	**M = 64**	**M = 256**	**M = 64**	**M = 256**
MFCC	18.06	17.39	7.79	7.22
MFCC+Δ	16.78	15.67	7.47	6.81
MFCC+Δ+ΔΔ	16.48	**15.45**	7.28	**6.65**
LPCC	18.49	18.15	7.48	7.13
LPCC+Δ	16.87	15.75	7.01	6.63
LPCC+Δ+ΔΔ	16.13	**15.26**	6.82	**6.47**
LPRC	17.9	17.3	7.61	6.98
LPRC+Δ	17.02	16.06	7.31	6.7
LPRC+Δ+ΔΔ	17.02	**15.75**	7.24	**6.76**

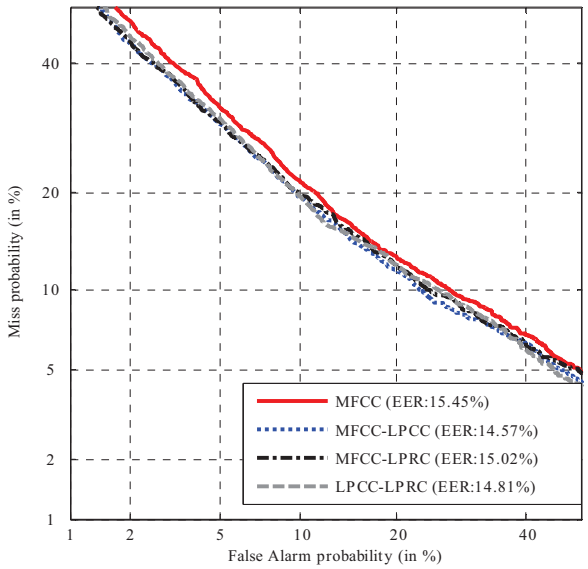

Fig. 5. DET curves of fused feature sets

5. Conclusions

We have studied the effect of glottal flow features (LPRC) on speaker verification performance. The results reveal that LPRC features convey useful speaker-specific information as much as those of MFCC and LPCC, and fusing these features with LPRC improve recognition accuracy. This indicates the usefulness of glottal flow features in speaker recognition. For future work, comparisons of MFCC, LPCC, and LPRC features with other popular feature sets such as PLPs and LFCCs would be interesting.

6. References

[1] M. Chetouani, M. Faundez-Zanuy, B. Gas, J. L. Zarader, "Investigation on LP-residual representations for speaker identification", *Pattern Recognition*, vol. *42,* no. *3,* pp.*487-494,* March, 2009.

[2] S. B. Davis, P. Mermelstein,"Comparison of parametric representation for Monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoustics Speech and Signal Proc.*, vol. *28*, no. *4*, pp. *357-366*, Aug. 1980

[3] S. Furui, "Cepstral analysis technique for automatic speaker verification", *IEEE Trans. Acoustics, Speech and Signal Proc.*, vol. *29*, no. *2*, pp. *254-272*, Apr. 1981.

[4] T. Kinnunen, P. Alku, "On separating glottal source and vocal tract information in telephony speaker verification", in *Int. Conf. on Acoustics, Speech and Signal Processing*, Taipei, 2009, pp. *4545-4548*.

[5] T. Kinnunen, H. Li, "An overview of text-independent speaker recognition: from features to supervectors", *Speech Communication*, vol. *52*, no. *1*, pp. *12-40*, Jan. 2010.

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

[7] NIST2001 Speaker Recognition Evaluation, Available: http://www.itl.nist.gov/iad/mig/tests/sre/2001/index.html (URL)

[8] M. D. Plumpe, T. F. Quatieri, D. A. Reynols, "Modelling of the glottal flow derivative waveform with application to speaker recognition", *IEEE Trans. Acoustics, Speech and Signal Proc.*, vol. *7*, no. *5*, pp. *569-586*, Sep. 1999.

[9] L. R. Rabiner, R. W. Schafer, "Digital processing of speech signals", Prentice-Hall, New Jersey, USA, 1978.

[10] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, vol. *10*, no. *1*, pp. *19-41*, Jan. 2000.

[11] N. Zheng, T. Lee, P.C. Ching, "Integration of complementary acoustic features for speaker recogition", *IEEE Signal Proc. Letters*, vol. *14*, no. *3*, pp. *181-184*, March 2007.

[12] Niko Brummer's Focal Toolkit, available: sites.google.com/site/nikobrummer/**focal (URL)**