Speaker Information from Subband Energies of Linear Prediction Residual

Debadatta Pati and S R M Prasanna

Department of Electronics and Communication Engineering,
Indian Institute of Technology Guwahati
Email: {debadatta, prasanna }@iitg.ernet.in

***Abstract*—The objective of this work is to demonstrate the significant speaker information present in the subband energies of the Linear Prediction (LP) residual. The LP residual mostly contains the excitation source information. The subband energies extracted using the mel filterbank followed by cepstral analysis provides a compact representation. The resulting cepstral values are termed as Residual-mel Frequency Cepstral Coefficients (R-MFCC). The speaker identification studies conducted using R-MFCC as features and Gaussian mixture model (GMM) on a subset of 30 speakers from NIST-1999 provides 87% accuracy. The performance using MFCC extracted directly from speech provides 87% accuracy. Further, the combination of the two provides 90% accuracy indicating the different aspect of speaker information present in R-MFCC.**

I. Introduction

Speaker recognition is the task of recognizing speakers from their speech signal [1]. Speaker recognition can be either identification or verification. In case of identification, the most likely speaker of the test speech is identified by comparing with the stored reference models. Validating the identity claim by comparing the test speech with the claimed speaker model is the verification task. The identification task is further divided into closed set and open set based whether identification is only among the set of enrolled speakers or provision for unknown speakers also, respectively. Depending on the text, text-dependent mode will use speech for the same text and no such restrictions in case of text-independent mode. This work considers text-independent closed-set identification approach for the study. For given speech data, the performance of the speaker recognition system depends mainly on the type of feature and the modeling technique employed. State of the art speaker recognition systems use mel Frequency Cepstral Coefficients (MFCC) derived from speech as features and Gaussian Mixture Model (GMM) as the modeling technique [2]. To further improve the performance, there are several attempts to explore alternative features and models [3], [4], [5], [6]. The feature level exploration include excitation source features [4] and suprasegmental features [3]. The model level exploration include AutoAssociate Neural Network (AANN) model [5] and Support Vector Machines (SVM) [6]. Most of these studies have shown that they capture different speaker information or model speakers in different way. Due to this, the combined system is demonstrated to provide better performance [7], [8].

The focus of this work is to develop a new method for parameterizing the speaker information in the excitation source component of speech. The earliest effort in this direction include using pitch information [9]. However, pitch is only one measurement and the excitation source component has richer speaker information [9]. Motivated by this, the first attempt to parameterize the source component by cepstral analysis using Linear Prediction (LP) residual is done in [10]. There are several attempts to learn the speaker information present in the LP residual implicitly using AANN models [4], [8]. Attempts have also been made to parameterize the source component in terms of glottal wave characteristics [11]. All these studies demonstrated the potential of the excitation source information for speaker recognition. From the simplicity point of view, the initial work of using cepstral analysis over LP residual [10] is preferable over other methods. It is to be cautioned at this stage that the cepstral analysis is only used as a computation method for obtaining compact representation and not from the original aim of source and system separation [9]. This method can be improved by considering the subband energies. Subband energies in this work refers spectral subband energies. Since the LP residual has flat spectrum, we may benefit by accumulating the spectral energies over subbands and then use them as features. Further, we may benefit by transforming these subband energies into the time domain using a non-linear log operation, applying mel filterbank and then taking Inverse Discrete Fourier Transform (IDFT).

Initially, the existing method of performing cepstral analysis on the LP residual reported in [10] is described. The subband energies computed from the LP residual using uniform filterbank employing rectangular and triangular windows are directly used as features. The cepstral coefficients extracted from the subband energies of the LP residual over uniform filterbank are then used as features. Finally the cepstral coefficients extracted from the subband energies of the LP residual over mel filterbank termed as Residual-MFCC (R-MFCC) are used as feature vectors. A comparative study is made across these different cases to observe the potential of subband energies from the LP residual. In particular, whether it is better than using the LP residual as it is? Also among the different approaches for subband energies representation, which one provides better performance? Finally a comparative study will also be made with the speaker identification system using MFCC computed from speech. This study is to analyze

978–1–4244–6385–5/10/$26.00 ©2010 IEEE

the nature of speaker information present among the two MFCCs and also their potential for combination to improve the performance.

The rest of the paper is organized as follows: Section II describes briefly the method proposed in [10] and also the approaches for computing the subband energies. Section III describes the different speaker identification studies conducted using subband energies. The speaker identification studies conducted using the existing systems is described in Section IV . The combined speaker identification system using proposed R-MFCC and existing MFCC features is developed in Section V. The summary, conclusion and future scope of the present work are given in Section VI.

II. Subband Energies from LP Residual

A. LP Residual

The speech signal $s(n)$ is processed by the p^{th} order LP analysis to extract the LP Coefficients (LPC) a_k, where, $k = 1, 2, \ldots, p$ [12]. The LP residual $e(n)$ is computed from the speech signal by inverse filtering, given in the time domain as $e(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k)$. For proper LP order, the LPC mostly represent the vocal tract information and the LP residual mostly represent the excitation source information. For instance, in case of speech sampled at 8 KHz, LP order of 10-12 is found to be most suitable. In most of the speech processing applications, LP residual is therefore used as the representation of excitation source information.

B. Cepstral Analysis of LP Residual

The earliest attempt of using LP residual for speaker recognition is made by performing cepstral analysis on the LP residual [10]. The motivation for this work was to demonstrate that the LP residual has richer information compared to pitch alone. The magnitude spectrum of the LP residual is computed by the Discrete Fourier Transform (DFT) to get $E(k) \leftrightarrow e(n)$. The logarithm is applied on the magnitude spectrum and IDFT is taken to get the cepstral coefficients $c(n) = IDFT(log|E(k)|)$. The reason for the choice of cepstral analysis as explained in [10] is as follows: LP residual is a time domain feature that includes residual amplitude, phase and pitch information. The objective of the work was not to consider pitch epochs and simultaneously want to get rid of its phase contribution. To achieve this, magnitude spectrum of residual is computed which tends to be flat. This intermediate feature is meaningful only when transformed back to the time domain. Furthermore, they arbitrarily introduce a logarithmic non-linearity and then take IDFT which result in the cepstral coefficients. The cepstral coefficients termed as Residual-FFT (Fast Fourier Transform) derived cepstral coefficients (R-FFTCC, in this work), provide a compact representation for the residual magnitude spectrum information. Using these cepstral coefficients speaker verification study is performed. It is demonstrated that even though the performance is relatively poor compared to the cepstral coefficients from speech, they combine well with speech cepstral coefficients to provide improved combined performance.

C. Subband Energies

This work concentrates on some more manipulations in the frequency domain for the LP residual. The earlier work in [10] concentrated on directly applying non-linear log operation and then computing IDFT. Instead of that we thought a more compact representation is to first compute the subband spectral energies. This is done by multiplying the residual magnitude spectrum with a uniform filterbank having 24 rectangular windows and summing the residual amplitudes in each rectangular window. There is half window overlapping for the computation of all subband energies. These 24 subband energies, termed as residual rectangular subband energies (R-RSE) themselves are used as features for speaker identification. In the next step, the triangular window type was chosen to compute the subband energies termed as residual triangular subband energies (R-TSE). The motivation is to check the effect of shape of window on the performance and also for later comparison with mel-filterbank which uses triangular windows. In the next step, the log magnitude spectrum is passed through the uniform filterbank with triangular windows and IDFT is computed to get the cepstral coefficients. Since the triangular windows are uniformly placed, these cepstral coefficients are termed as Residual Uniform Frequency Cepstral Coefficients (R-UFCC). As a last attempt the log magnitude spectrum is passed through the non-uniform filterbank with triangular windows placed on the mel-frequency scale and IDFT is computed to get the cepstral coefficients. Since these triangular windows are non-uniformly placed, these cepstral coefficients are termed as residual mel-frequency cepstral coefficients (R-MFCC). All these different forms of the subband energies are used as feature vectors for speaker identification to experimentally select the representation that provides the best performance.

III. Speaker Identification using Subband Energies

A. Speaker Identification Database

The NIST-1999 database consists of speech data from 539 speakers (230 male and 309 female) [13]. The training data for each target speaker consists of two utterances of about one minute each, obtained by concatenating consecutive turns of the speaker. The test data ranges between few seconds to minutes. The data is collected over land line telephone, sampled at 8 KHz and stored with 8 bits/sample resolution with mu-law format. A detailed description of the database can be found in the NIST-1999 evaluation plan [13]. Among these speakers, 15 male and 15 female speakers having matched conditions and testing data of at least 30 sec are selected to form the subset. This subset is used as 30 speaker database for all the studies.

B. Speaker Identification using R-RSE Features

The training speech of each speaker is processed in blocks of 20 msec and 10 msec block shift using 10^{th} order LP analysis to extract the LP residual. The residual subband energies using the rectangular windows i.e., R-RSE are extracted. These R-RSE features are modeled using Vector Quantization (VQ) and Gaussian Mixture Modeling (GMM) techniques [2]. In this way all the speaker models are developed. The testing speech is also processed in the similar way and matched with the speaker models using Euclidean distance in case of VQ and log-likelihood ratio in case of GMM. The speaker of the model with least Euclidean distance and highest log-likelihood ratio is identified as the speaker in the case of VQ and GMM, respectively. The performance of speaker identification system using VQ model is tabulated in Table I and that using GMM model is tabulated in Table II. The highest performance of 57% accuracy in case of VQ and 83% accuracy for GMM indicate that these subband energies in their raw form contain speaker information.

C. Speaker Identification using R-TSE Features

In this case the experimental study remains same as in the case of R-RSE features, except for using the R-TSE features. The performance of speaker identification system by modeling R-TSE features using VQ and GMM are given in Table I and II, respectively. In the VQ case, R-TSE features shows highest performance of 50%. In the GMM case, R-TSE features provides highest performance of 83%. The performance of the systems using subband energies from the rectangular and triangular windows are nearly same. Most of the existing filterbank approaches in case of speech, employ triangular filters. Hence in all our studies the subband energies are extracted using triangular windows.

D. Speaker Identification using R-UFCC Features

The residual subband energies using uniformly placed triangular windows i.e., R-TSE are computed for the log magnitude spectrum. The IDFT of these subband energies are taken to get R-UFCC features. The performance of speaker identification systems by modeling R-UFCC features using VQ and GMM are given in Table I and II, respectively. In both VQ and GMM modeling cases, the performance using R-UFCC is significantly better compared to using only R-TSE features. This result reinforces the advantage achieved by transforming the subband energies into time domain as reported in [10].

E. Speaker Identification using R-MFCC Features

This study is same as in the case of R-UFCC, except for the fact of computing R-MFCC by using non-uniformly placed triangular windows based on the mel scale. The performance of the speaker identification system by modeling R-MFCC features using VQ and GMM are given in Table I and II, respectively. The performance of R-MFCC based speaker identification systems is better compared to R-UFCC counterparts. This infers that we gain by using non-uniformly placed triangular windows compared to the uniformly placed ones. From this result we can say that even though the LP residual has flat spectrum, gain may be achieved while computing subband energies by using non-uniformly placed windows. We therefore propose that whenever subband energies from the residual are to be used, R-MFCC can be used as their representation to achieve maximum gain.

TABLE I
IDENTIFICATION PERFORMANCE (%) USING VQ MODEL.

Codebook Feature	16	32	64	128	256
R-RSE	43	37	50	57	53
R-TSE	37	50	43	43	50
R-UFCC	83	77	80	80	80
R-MFCC	73	80	87	87	83
R-FFTCC	70	70	77	73	77
MFCC	93	90	90	93	90

TABLE II
IDENTIFICATION PERFORMANCE (%) USING GMM MODEL.

No.Gaussian Feature	16	32	64	128	256
R-RSE	83	83	73	80	70
R-TSE	73	77	83	83	80
R-UFCC	83	83	83	83	87
R-MFCC	83	87	93	93	87
R-FFTCC	80	80	83	77	73
MFCC	90	90	90	87	87

IV. SPEAKER IDENTIFICATION USING EXISTING SYSTEMS

A. Speaker Identification using R-FFTCC

The first and foremost comparison we need to make the present work is with the system proposed in [10]. For this the LP analysis and residual computation is made as in the earlier studies. The R-FFTCC are used as features for modeling and testing. The performance of speaker identification systems using VQ and GMM modeling techniques are given in Table I and II, respectively. The speaker identification system using VQ modeling shows significantly better performance for the proposed R-MFCC compared to the R-FFTCC features. The speaker identification system using GMM modeling shows almost same performance for both R-MFCC and R-FFTCC for lower number of Gaussian mixtures, but better performance to R-MFCC for higher number of Gaussian mixtures. From these observations we propose the R-MFCC as features for representing speaker information in the residual amplitude values compared to the existing R-FFTCC.

B. Speaker Identification using MFCC

The last comparison for the present work is with the state art feature, that is, MFCC derived from speech signal. For this, speech is processed in blocks of 20 msec and shift of 10 msec and processed to extract the MFCC features. The MFCC features are modeled using VQ and GMM modeling techniques. The models are tested using the MFCC from the test speech signals. The performance of the speaker identification systems using MFCC features for VQ and GMM modeling cases are given in Table I and Table II, respectively. It should be noted that MFCC provides near complete representation for the vocal tract information, where as, R-MFCC models only the residual magnitude information. Thus the proposed R-MFCC features may be augmented with other excitation source features like pitch and phase information to further improve the performance.

TABLE III

IDENTIFICATION PERFORMANCE (%) OF COMBINED SYSTEM USING VQ (32) AND GMM (256) MODELS. THE FIRST THREE ROWS REFER TO VQ BASED SYSTEMS AND THE NEXT THREE ROWS REFER TO GMM BASED SYSTEMS.

Speakers Feature	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	Identification Accuracy (%)
R-MFCC	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	1	0	80
MFCC	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	90
Combined system	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	93
R-MFCC	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	87
MFCC	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	87
Combined system	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	90

V. SPEAKER IDENTIFICATION USING R-MFCC AND MFCC

The proposed R-MFCC features are derived from the LP residual where as MFCC features are derived from the speech signal. Thus it looks like the two features may have different information about the speaker, one mostly excitation source and the other mostly vocal tract. Since both the features give nearly same performance, it is difficult to observe the different information present in them. However, in case of VQ for 32 size codebook and in case of GMM for 256 number of mixtures, the different information captured by R-MFCC and MFCC can be observed from the detailed identification performance given in Table III. In this table, 1 indicates correct identification and 0 indicate missing of the speaker. The patterns of $1's$ and $0's$ in both the cases are different indicating that both have different speaker information. This can be further confirmed by combining the two systems by adding the scores of the two systems and then identifying the speakers. This is also given in the Table III. The performance of the combined system is better than the individual systems using R-MFCC and MFCC. This infers that we gain by combining R-MFCC and MFCC evidences for speaker identification.

VI. CONCLUSION

The objective of this work was to experimentally evaluate the speaker information present in the residual subband energies and obtain the best possible compact representation for the same. In this direction we explored different variants of residual subband energies which include R-RSE, R-TSE, R-UFCC and R-MFCC. Among all these R-MFCC provided the best performance. Further R-MFCC provides better performance than earlier R-FFTCC, but lower than MFCC. R-MFCC is demonstrated to represent different aspect of speaker information compared to MFCC with the improved performance of the combined system.

The discriminating ability of the proposed R-MFCC feature needs to be verified on a larger database. It is also observed that performance of R-MFCC is still inferior compared to MFCC. Thus future work should focus on using other excitation source related features along with R-MFCC to improve the performance.

ACKNOWLEDGMENT

This work is supported by the UK-INDIA Education and Research Initiative (UKIERI) project titled ”study of source features for speech synthesis and speaker recognition” between IIT Guwahati, IIIT Hyderabad and CSTR, University of Edinburgh, UK.

REFERENCES

[1] J. P. Campbell, “Speaker recognition: a tutorial,” *Proc. IEEE*, vol. 85, pp. 1436–1442, 1997.

[2] D. A. Reynolds and R. C. Rose, “Robust text -independent speaker identification using gaussian mixture speaker models,” *IEEE Trans.Speech and Audio Proc.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[3] B. Yegnanarayana, S. R. M. Prasanna, J. M. Zachariah, and C. S. Gupta, “Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system,” *IEEE Transactions on speech and audio processing*, vol. 13, no. 4, pp. 575–582, July 2005.

[4] S. R. M. Prasanna, C. S. Gupta, and B. Yegnenarayana, “Extraction of speaker-specific excitation information from linear prediction residual of speech,” *Speech Communication*, vol. 48, pp. 1243–1261, 2006.

[5] B. Yagnenarayana and S. P. Kishore, “AANN: An alternative to GMM for pattern recognition,” *Neural Networks, Vol. 15, No. 3 (2002) 459469*, vol. 15, no. 3 (2002), pp. 459–469, 2002.

[6] H. Fenglei and W. Bingxi, “Text-independent speaker recognition using support vector machine,” *Proc. IEEE 2001 International conference on Info-tech and Info-net ICII , Beijing*, vol. 3, pp. 402–407, 2001.

[7] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308– 311, May 2006.

[8] K. S. R. Murthy and B. Yegnanarayana, “Combining evidence from residual phase and mfcc features for speaker recognition,” *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, 2006.

[9] B. S. Atal, “Automatic speaker recogntion based on pitch contours,” *J.Acoust.Soc.Amer.*, vol. 52, no. 6, pp. 1687–1697, 1972.

[10] P. Thevenaz and H. Hugli, “Usefulness of the LPC-residue in text-independent speaker verification,” *Speech Communication*, vol. 17, pp. 145–157, 1995.

[11] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, “Modelling of glottal flow derivative waveform with application to speaker identification,” *IEEE Trans.Speech and Audio Proc.*, vol. 7, no. 5, pp. 569–586, Sep. 1999.

[12] J. Makhoul, “Linear prediction: A tutorial review,” *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[13] M. Przybocky and A. Martin, “The NIST-1999 speaker recognition evaluation-An overview,” in *Digital Signal Processing*, vol. 10, 2000, pp. 1–18.