Modeling Speaker Information from Source Spectrum

Debadatta Pati and S. R. M. Prasanna

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

Abstract

This work models and experimentally evaluates the speaker information present in the excitation source spectrum. The magnitude spectrum of the linear prediction (LP) residual derived from the speech signal is used as the source spectrum. The source spectrum from the voiced speech portion will have pitch and harmonics. The amplitudes at the pitch and harmonics constitute energy associated with the excitation source. The nature of harmonic structure constitutes the dynamics associated with the excitation source. Both these aspects of the source spectrum may have speaker information. Further, since they represent different aspect of source spectrum, the speaker information may also be different. The speaker recognition studies conducted using NIST-03 database indeed confirm both these aspects.

Index Terms: LP residual, source spectrum, speaker information, RMFCC, MPDSS

1. Introduction

Speech production system consists of vocal tract that generates speech when excited by an excitation source. Most of the existing speaker recognition studies exploit the speaker information from the vocal tract. The reasons may be the near complete representation in terms of resonances and their bandwidth, fueled with the availability of efficient signal processing methods to compactly represent them. Alternatively, the source component also contributes equally to the speaker information, but the performance is not at par with vocal tract information. The reason may be the dynamic nature of excitation source and hence difficulty in representing its speaker information compactly. However, from our experience about listening to normal and whispered speech, we may agree that the speaker information from the source may be relatively more robust compared to the vocal tract for any degradations like background, sensor, channel and so on. Hence the continued interest in exploring the methods for modeling speaker information from the excitation source.

The existing attempts for exploiting speaker information from the excitation source mostly use the linear prediction (LP) residual as a representation of the excitation signal [1, 2, 3]. These attempts broadly grouped into two categories, namely, time domain and frequency domain approaches. In the time domain approach, the LP residual signal is analyzed and processed at the segmental, sub-segmental and supra-segmental levels to model the speaker information. These include both neural network and other techniques like vector PCM [4, 5]. In the frequency domain, the magnitude spectrum of the LP residual is analyzed at the segmental level. These include cepstral analysis, harmonic structure analysis and wavelet octave coefficients of residues (WOCOR) [1, 2, 3, 6]. Both the approaches seem to provide nearly same performance. However, the methods based on source spectrum represent speaker information in more compact manner and computationally more efficient.

The information present in the magnitude spectrum of the LP residual may be broadly classified into two aspects, namely, the spectral amplitude values, primarily at the pitch and harmonics, and the nature of harmonic structure. The spectral amplitudes give information about the energy associated with the excitation source [2]. The nature of harmonic structure reflects about the physical structure of the excitation source and its dynamics [7]. Since they represent different aspect of source spectrum, they may contain different speaker-specific source information. These issues will be explored in this work. The energy as well as harmonic structure have earlier been studied independently for speaker recognition [3, 2]. This work carries out some refinements in the methods employed for extracting them. A detailed exploration of both these aspects to know their different nature is then made. Studies will also be made to observe how well they combine with the existing speaker recognition system based on vocal tract information.

The rest of the paper is organized as follows: Section 2 describes briefly the methods for extracting the energy and harmonic structure information from the source spectrum. Experimental results of the different speaker recognition studies based on these information and their different nature are discussed in Section 3. The summary, conclusion and future scope of the present work are given in Section 4.

2. Energy and Harmonic Information from LP residual Spectrum

2.1. LP Residual Spectrum

The speech signal $s(n)$ is processed by the p^{th} order LP analysis to extract the LP Coefficients (LPCs) a_ks, where, $k = 1, 2, \ldots, p$. The LP residual $e(n)$ is computed from the speech signal by inverse filtering [8]. For proper LP order, for instance 10-12 for speech signal sampled at 8 kHz, LPCs mostly represent the vocal tract information and the LP residual mostly represent the excitation source information [4]. In this work magnitude spectrum of 10^{th} order LP residual obtained as $E(k) = \sum_{n=0}^{N-1} e(n)e^{-j\frac{2\pi nk}{N}}$ is used as the excitation source spectrum. Examples of speech, residual and their spectra for two speakers (*Speaker-A*, *Speaker-B*) are shown in the Figure 1(a)-(d). The modulation in the speech spectra mostly represent the vocal tract information, where as, the amplitude and harmonic structure of residual spectra represent excitation source information. For instance, *Speaker-B* shows much stronger periodicity than *Speaker-A*.

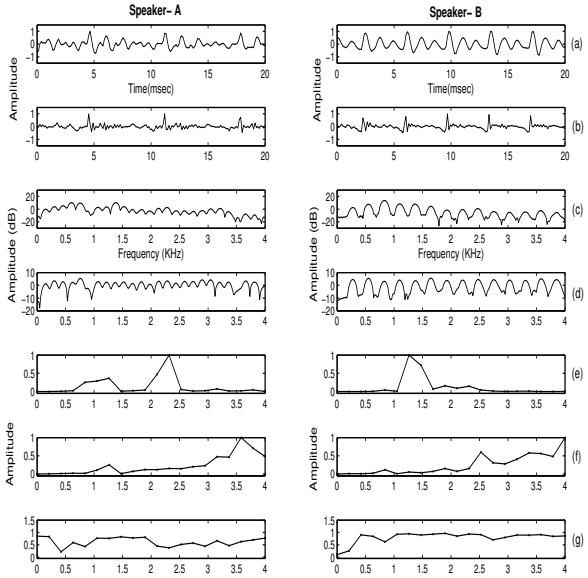

Figure 1: Vocal tract and excitation source information of two different speakers. (a) Speech signals. (b) LP residuals. (c) Magnitude spectra of speech signals. (d) Magnitude spectra of LP residuals. (e) MSE contours of speech signals. (f) MSE contours of LP residuals. (g) MPDSS contours of LP residuals.

2.2. Energy Information from LP residual Spectrum

The magnitude spectrum of the LP residual is nearly flat. The energy information extracted directly from the flat spectrum may not have useful discriminating information. In [3], it was shown that the distribution of energies on the mel scale called as mel subband energies (MSE) provide enhanced discrimination about the speaker. These energies are computed from the mel warped spectrum of the LP residual using Equation 1.

$$MSE(m) = \sum_{k=l_m}^{h_m} E_m(k), m = 1, 2, \ldots, M \tag{1}$$

where, $E_m(k) = |E(k)|H_m(k)$ is the mel warped magnitude spectrum of LP residual and l_m, h_m are the lower and upper limit of frequencies, respectively, in the m^{th} mel filter $H_m(k)$. M is the number of filters in the mel filterbank.

One can observe from Figure 1 (e) and (f) that the distribution of energies for speech and residual signals are significantly different. Further, they are also different across speakers. A more compact representation of these energies is proposed in [3]. The information related to subband energies are derived using the residual mel frequency cepstral coefficients (RMFCC). The method to compute RMFCC are similar to the conventional mel frequency cepstral coefficients (MFCC) using Equation 2, except that the input is LP residual spectrum. These coefficients essentially represent the mel scale distribution of the subband energies in time domain.

$$RMFCC(c) = \sum_{m=1}^{M} X_m cos[c(m - \frac{1}{2})\frac{\pi}{M}] \tag{2}$$

where, $m = 1, 2, \ldots, M$ is the number of filters in the mel filterbank. $c = 1, 2, \ldots, C$ is the number of cepstral coefficients computed (usually C<M) and $X_m = log10(\sum_{k=0}^{N-1} E_m(k))$, represents the log energy output of the m^{th} mel filter.

2.3. Harmonic Information from LP residual Spectrum

Rate of vocal folds vibration and manner in which the vocal folds open and close show variations across speakers [7]. In some cases the vocal folds close rapidly and completely corresponding to hard voice. In this case the flow is discontinuous and the excitation is more impulse-like in nature and the residual magnitude spectrum is more flat. For soft speaking speakers the folds never close completely and there is smooth air flow, so the residual spectrum is comparatively less flat. Such excitation characteristics are reflected in the harmonic structure of the residual spectra. To determine quantitatively, power difference of spectrum in subband (PDSS) measure is proposed in [2]. These differences are computed from spectral flatness measure using uniform filter bank consisting of rectangular windows. We prefer to use mel filters [9]. The reason for using mel filters is the property of the mel filter bank that provides less spectral samples to lower bands and more to higher bands (beyond 1 kHz). In case of LP residual most of the energies are concentrated in the higher range. Since flatness is a statistical measure, with increase in number of samples in the higher bands, flatness may be more accurately measured. The PDSS values are computed from the 20 subband spectra of LP residual using the Equation 3. These PDSS values are called as mel PDSS (MPDSS) and shown in Figure 1 (g) for two speakers. One can observe that, lower the difference in peak and dip, higher is the MPDSS value and vice versa. Further these values are different across speakers. These differences show that, the harmonic structure of residual spectra depends on speaker and expected to be an effective speaker individual information.

$$MPDSS(m) = 1 - \frac{\left[\prod_{k=l_m}^{h_m} E_m(k)\right]^{\frac{1}{N_m}}}{\frac{1}{N_m}\sum_{k=l_m}^{h_m} E_m(k)}, \tag{3}$$

where $N_m = h_m - l_m + 1$ is the sample number of frequency points in the m^{th} mel warped subband spectrum.

2.4. Analysis of Energy and Harmonic Information

Although RMFCC and MPDSS feature are derived from the source spectrum, but they reflect different aspect of speaker information. RMFCC represent the amplitude of the excitation where as the MPDSS represent the modulation of the excitation in the subbands spectra. One can also observe from Equation 2 and 3 that, their computational procedure are also different. Each RMFCC is computed from the whole source spectrum. So RMFCC represent the gross information about the source spectra. Where as, MPDSS are computed from individual subband spectra and thus represent information about the local variation in each subband spectra. Therefore these two information reflect different aspect of speaker information. This can also be observed from the Figure 1 (f) and (g). They are entirely different from each other. This is further confirmed from the recognition experiments in Section 3.4. They give different decisions in many instances.

3. Speaker Recognition Study

3.1. Experimental Setup

Speaker recognition is the task recognizing speakers and can be either identification or verification. Identification mode finds the most likely speaker of the test data, and verification mode

Table 1: Identification performance (%) using GMM model.

Feature	Set-1	Set-2
RMFCC	63	35
MPDSS	53	31
RMFCC+MPDSS	70	43
MFCC	83	50
MFCC+RMFCC+MPPDSS	87	57

validates the identity claim. In this paper, Gaussian mixture model (GMM) technique using 128 Gaussian functions is used to build the speaker models [10]. Decision is taken based on the log-likelihood ratio (LLR). For identification, the speaker of the model having highest (LLR) is the identified speaker. Performance is expressed in terms of average accuracy. In case of verification, for every trail LLR is assigned to each claimant. The performance is given by detection error tradeoff (DET) curve based on genuine and imposter LLR [11]. From the DET curve equal error rate (EER) is found based on threshold θ such that false acceptance rate (FAR) is equal to false rejection rate (FRR). We conduct both identification and verification experiments. In both tasks feature extraction and modeling techniques remain same.

The speaker recognition experiments presented in this paper are conducted using the NIST-2003 database [12]. There are 149 male speakers and 207 female speakers. The duration of training data for each speaker is around 2 min and the test data ranges between 15-45 sec. For preliminary identification study (set-1) consist of 15 male and 15 female speakers having matched conditions and testing data of at least 30 sec are selected. The results are further verified on a large database (set-2) consisting of 146 male and 203 female target speakers. For verification experiments there are 3428 test utterances. Each test utterance has 11 claimants, where the genuine speaker may or may not be present. All speech signals were sampled at 8 kHz.

3.2. Speaker Recognition using RMFCC Features

The training speech of each speaker is processed using LP analysis in blocks of 20 $msec$ with a shift of 10 $msec$ to extract the LP residual. The RMFCC features are computed from residual spectrum using 24 overlapping mel filters as described in Section 2.2. The first 13 coefficients excluding the first coefficient is used as feature vector. These features are used for building the speaker models. The testing speech is also processed in a similar way. The identification performance for both the sets are given in the first row of the Table 3.2. The performance achieved for set-1 is 63% and for set-2 is 35%. In case of verification, the DET curve is shown in Figure 2 and an EER of 36% is achieved. The performance achieved for both identification and verification tasks indicate that these RMFCC features contain good amount of speaker information.

3.3. Speaker Recognition using MPDSS Features

The procedure to compute the LP residual spectrum remains same as in the case of RMFCC. Then MPDSS values are computed using 20 overlapping mel filters as described in Section 2.3. Each MPDSS feature is represented by 20 MPDSS values. These features are used to build the speaker models. The identification performance for both the sets are given in second row of the Table 3.2. The performance achieved for set-1 is 53% and for set-2 is 31%. In case of verification, from the DET curve shown in Figure 2 an EER of 33% is achieved. The performance achieved for both identification and verification tasks indicate the presence of speaker information.

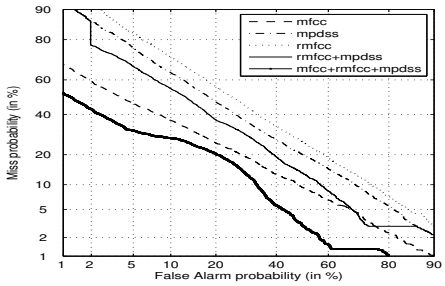

Figure 2: Performances of speaker verification systems with RMFCC, MPDSS, MFCC, RMFCC+MPDSS and MFCC+RMFCC+MPDSS.

3.4. Combination of RMFCC and MPDSS Features

The different aspect of speaker information present in RMFCC and MPDSS can also be observed from the detailed recognition performance. In case of identification task for set-1, the detailed identification performance is given in first two rows of the Table 3.2. In this table, 1 indicates correct identification and 0 indicate missing of the speaker. The patterns of 1's and 0's are different indicating that both have different speaker information. Further, RMFCC that provides better performance compared to MPDSS, in certain cases, for example, speakers 7 and 23, gives wrong decision, but MPDSS gives correct decision. By combining both the evidences identification performance may be improved. Simple linear combination with predefined weights may not necessarily provide the best result [6]. Because fusion of scores may result in a wrong decision. Since we have the ground truth, we use logical *OR* combination. In this combination if any one system is giving correct decision, we consider it as a correct decision. The performance of the combined system shown in Table 3.2 is improved by an average of 20% for both the data sets. Similarly in case of verification, these two approaches give different decisions. Figure 3 shows the distribution of the LLR scores attained by RMFCC and MPDSS for genuine and imposters trails. Let θ_{RMFCC} and θ_{MPDSS} denote the LLR thresholds for RMFCC and MPDSS based systems. The regions where the systems are giving different decision are labeled as region I and II. For region I, RMFCC proposes rejection, but MPDSS suggests acceptance. Further RMFCC rejects some genuine speakers, which are actually accepted by MPDSS system. Similarly some of the imposters that are falsely accepted by MPDSS are rejected by RMFCC system. Similar observation can be made on region II. By combining both the evidences, the verification performance may be improved. Since we have the key reference about the speakers, we propose to modify the scores in the boundary region of the best system based on the information of the other. The boundary region is the spreading of the scores distribution. In case of genuine scores, if a genuine from the left side of the region of MPDSS is in right side of the region of RMFCC, then the genuine score is replaced by the mean plus standard deviation of the genuine scores given by MPDSS. Similarly for imposter scores, if an imposter from the right side of the region of MPDSS is in left side of the region of RMFCC, then the imposter score is replaced by the mean minus standard deviation of the imposter scores given by MPDSS. The redistribution of the scores reject some imposters and accept some genuine speakers based on

Table 2: Detail identification performance of RMFCC, MPDSS and MFCC features using GMM modeling technique.

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 (%)
RMFCC	1	1	1	0	0	1	0	1	1	1	0	0	1	1	0	1	1	1	1	0	1	1	0	1	0	0	1	0	1	1	63
MPDSS	0	0	1	0	0	0	1	1	1	1	0	0	1	1	0	1	0	1	1	0	1	1	1	1	0	0	0	0	1	1	53
RMFCC+ MPDSS	1	1	1	0	0	1	1	1	1	1	0	0	1	1	0	1	1	1	1	0	1	1	1	1	0	0	1	0	1	1	70
MFCC	1	1	1	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	83
RMFCC+ MPDSS+ MFCC	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	87

RMFCC information. As a result, the EER is reduced to 29% as shown in Figure 2. The improvement in the recognition performance by combining the evidences indeed indicate the different aspect of speaker information in RMFCC and MPDSS.

3.5. Speaker recognition using MFCC Features

The last comparison is with the standard MFCC features derived from the speech signal. Speech signal in blocks of 20 msec and with a shift of 10 msec are processed to extract the MFCC features. Similar to RMFCC features, 13 dimensional MFCC features are computed using 24 overlapping mel filters. The identification performance for both the data sets are given in the fourth row of the Table 3.2. The performance achieved for set-1 is 83% and for set-2 is 50%, which are better than corresponding RMFCC and MPDSS performances. Similarly in case of verification, the EER achieved as shown in Figure 2 is 23%, which is also better than RMFCC and MFCC. The possible reason may be MFCC provides near complete representation for the vocal tract information, where as, RMFCC and MPDSS model only the excitation energy and the voicing quality information. Since these two information are different, they may be combined to further improve the recognition performance. We combine the evidences from MFCC, and RMFCC and PDSS features as described in section 3.4. The identification performances for both the data sets given in the last row of the Table 3.2 are improved to 87% and 57%, respectively. Similarly in case of verification the EER is improved to 20% as shown in Figure 2.

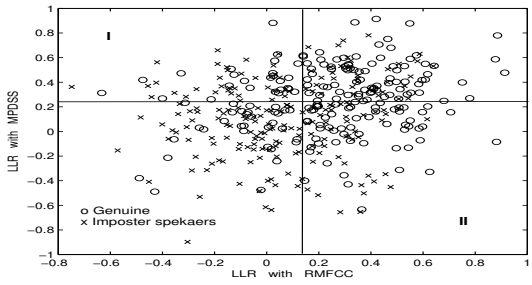

Figure 3: Distribution of 2-D LLR scores for genuine and imposter trails. $\theta_{RMFCC} = 0.15$ and $\theta_{MPDSS} = 0.23$ are chosen such that FAR=FRR.

4. Conclusion

The objective of this work was to experimentally analyze the speaker information present in source spectrum. We have shown that energy information represented by RMFCC and MPDSS indeed represent effective speaker information. Further we verified from recognition experiments that they reflect different aspect of speaker information. We also concluded that the combined use of this information is more effective. Performance of combined RMFCC and MPDSS is still inferior compared to MFCC. Thus future work should focus on using other approaches to exploit the source spectral information. Further the combination scheme used in this work is based on the available ground truth. For real time application suitable combination technique needs to be explored.

5. 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.

6. References

[1] P. Thevenaz and H. Hugli, "Usefulness of the LPC-residue in text-independent speaker verification," *Speech Commun.*, vol. 17, pp. 145–157, Aug. 1995.

[2] S. Hayakawa, K. Takeda and F. Itakura, "Speaker identification using harmonic structure of lp-residual spectrum," *Biometric personal Aunthentification, Lecture notes, Springer, Berlin*, vol. 1206, pp. 253–260, 1997.

[3] D. Pati and S. R. M. Prasanna, "Speaker information from subband energies of linear prediction residual," in *Proc. NCC 2010*, pp. 1–4.

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

[5] D. Pati and S. R. M. Prasanna, "Non-parametric vector quantization of excitation source information for speaker recognition," in *Proc. TENCON*, 2008, pp. 1–4.

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

[7] 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.

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

[9] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech and Signal Processing*, vol. 28, no. 28, pp. 357–366, Aug. 1980.

[10] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Proc.*, vol. 3, no. 1, pp. 4–17, Jan. 1995.

[11] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," vol. 4. in Proc. Eur. Conf. on Speech Communication Technology, Rhodes, Greece , 1997, pp. 1895–1898.

[12] "NIST speaker recognition evaluation plan," *in Proc. NIST speaker recognition workshop, college park, MD*, 2003.