Residual Phase Cepstrum Coefficients with Application to Cross-lingual Speaker Verification

Jianglin Wang[1], *Michael T. Johnson*[1]

Speech and Signal Processing Laboratory
Department of Electrical and Computer Engineering
Marquette University, Milwaukee, USA
{jianglin.wang, mike.johnson}@marquette.edu

Abstract

Speaker identification and verification has received a great deal of attention from the speech community, and significant gains in robustness and accuracy have been obtained over the past decade [1], [2]. However, the features used for identification are still primarily representations of overall spectral characteristics, and thus the models are primarily phonetic in nature, differentiating speakers based on overall pronunciation patterns. This creates difficulties in terms of the amount of enrollment data and complexity of the models required to cover the phonetic space, especially in tasks such as cross-lingual verification where enrollment and testing data may not have similar phonetic coverage. This paper introduces the use of a new feature for speaker verification, residual phase cepstral coefficients (RPCC), to capture speaker characteristics from their vocal excitation patterns. Results on a cross-lingual speaker verification task taken from the NIST 2004 SRE demonstrate that these RPCC features are significantly more accurate than traditional mel-frequency cepstral coefficients **(**MFCC) when the amount of enrollment data available for training is limited. Additionally, because of the significant differences in the nature of the features, combining MFCC and RPCC features shows an improvement in verification results over MFCCs alone.

Index Terms: speaker verification, glottal source excitation, residual phase cepstrum, GMM, UBM.

1. Introduction

Speaker identification and verification is an important application which has received a great deal of attention from the speech community, and significant gains in robustness and accuracy have been obtained over the past decade [1], [2]. However, the features for identification and verification, such as cepstral coefficients are still primarily representations of the overall spectral characteristics, and thus the models are primarily phonetic in nature, with systems differentiating speakers through characterization of pronunciation patterns. Little progress has been made toward identifying individually unique speech characteristics that are independent of phonetic content and language. This causes several significant limitations, including the need for models that represent a speaker's entire phonetic space and enough enrollment data to cover this model space. Additionally, there are some types of identification applications where the phonetic characteristics of the enrollment data does not necessarily match that of the test data, such as cross-lingual verification. This is important in multilingual environments, where speakers may access the system in any one of multiple languages.

This paper proposes the use of a new feature for speaker verification, residual phase cepstral coefficients (RPCC), which captures characteristics from speakers' excitation rather than vocal tract characteristics and is more compact across a wide range of phonetic conditions. The goal of this alternative feature is to rapidly capture of the characteristic physiological features of a speaker, requiring less enrollment data and less complex models and enabling better performance in cross-lingual or phonetically misaligned enrollment/test conditions.

Vocal tract related features such as mel-frequency cepstral coefficients (MFCCs) and linear predictive cepstral coefficients (LPCCs) have been the dominant features for speaker recognition for a long time. These features capture the characteristics of vocal tract and are thus very useful for speech recognition. MFCCs can also give excellent performance for speaker recognition since MFCCs capture comprehensive information about speaker spectral characteristics and pronunciation patterns.

However, MFCCs require sufficient phonetic coverage to train a phonetically dependent speaker model since MFCCs are primarily related to the phonetic content. For example, MFCCs have an excellent performance when sufficient training and test data is available. In contrast, state-of-the-art speaker recognition systems suffer performance degradation when only small amounts of data (less than 10 seconds) is available [3]. Kinnunen [4] shows that MFCCs are not effective for speaker recognition on the 10sec-10sec condition of the NIST 2006 SRE corpus. Speaker recognition systems using MFCC features also show an accuracy decrease when the speaker model is trained in one language but testing is performed in a different language [5][6]. There is still a significant need for identification of unique speaker-specific features that are more independent of phonetic content and spoken language, and whose performance is less dependent on the amount of data.

Potential speaker-specific feature candidates would be those based on vocal source or excitation information, which contains much of the unique physiological properties of a speaker's speech production. These characteristics are unique to a given speaker's speech production system. The vocal source can also represent the tension of the vocal fold, which is associated with the glottal pulse parameters, such as the rate of the glottal closing, and the degree of the glottal opening. Examples of such features might be derivatives of fundamental frequency characteristics like jitter, shimmer, and harmonic amplitudes [7]. Orthogonal linear prediction coefficients have also been proposed as features for speaker identification [8] because they are more constant across utterances and thus are more independent of the linguistic context and indicative of the speaker.

Recently, it has been reported by K.S.R. Murty and B. Yegnayaranana [9] that LP residual phase extracted from vocal source also contains speaker-specific source information [10]. The changes in the phase around the glottal closure instants are different from one speaker to another. Their results demonstrate the complementary nature of the residual phase to the conventional system based on spectral features such as MFCCs. In [9], the entire residual phase of each frame (20ms) is directly applied for speaker recognition. This large feature dimension increases the complexity of the model and causes difficulty due to temporal variability.

In the study presented here, we develop a new feature extraction method based on the residual phase by performing cepstral analysis of the residual phase signal to give a lower feature dimension and de-correlate the feature vector. This de-correlation analysis allows the feature vectors to be modeled with the same Gaussian Mixture Model (GMM) approach typically used in speaker verification systems. Because of the decoupling of these features from the phonetic variation of the vocal tract and articulators, the amount of enrollment data needed for training is reduced.

This paper is organized as follows. Section 2 provides the details of the baseline system and the proposed feature extraction method. The baseline speaker verification system based on GMM-UBM-MAP is described in Section 3. Section 4 describes the experimental data, setup and results, and the final conclusions are given in Section 5.

2. Feature extraction

The LP residual signal of a speaker represents the impulse-like excitation which is related to the region around the glottal closure instant within each pitch period, corresponding to a high signal-to-noise ratio region. These regions are known to contain speaker-specific information [11]. Listening experiments have also shown that the residual carries significant speaker specific information, for it is known that residual provides valuable information that allows humans to distinguish between speakers [12]. Vocal tract excitation differs among speakers and stays stable within a given speaker. This leads to the possibility that features extracted from the residual signal may be useful in speaker recognition. Most features related to the residual are based on the magnitude spectrum of the LP residual signal, with the phase spectrum discarded. The large fluctuation of the residual causes difficulty deriving useful features from the LP residual. Gautherot reported that the magnitude spectrum of LP residual is flat, suggesting that the major information component is retained in the phase [12].

In this study, the proposed feature uses a Hilbert transformation to obtain the analytical signal of the LP residual [13], [14] and then extracts the spectrum shape of the residual phase by cepstral analysis. Compared to using the residual phase signal directly, this novel feature compactly captures perceptually meaningful source-like information from residual phase, and provides more speaker specific information about a speaker with a lower dimension.

2.1. MFCC

MFCCs are commonly used in most speech and speaker recognition systems. These approximate the perceptual model of the human auditory system by warping the linear frequency axis to match the Mel-scale cochlear frequency map. Although there are several possible methods for computation, here the filterbank approach is used, where the spectrum of each Hamming-windowed signal is divided into Mel-spaced triangular frequency bins, then a Discrete Cosine Transform (DCT) is applied to calculate the desired number of ceptral coefficients.

2.2. Residual phase cepstral coefficient (RPCC)

The definition of residual phase is the cosine of the phase function of the analytic signal [9]. The analytic signal is derived from the LP residual of a speech signal. The calculation of LP residual is equal to the error between the actual value $s(n)$ and the predicted value $\hat{s}(n)$, given by

$$r(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \qquad (1)$$

where *p* is the order of prediction and a_k are the linear prediction coefficients obtained from LPC analysis. Then, the phase of the analytic signal is calculated for the posterior feature extraction processing.

The analytical signal of the LP residual $r(n)$ is given by

$$r_a(n) = r(n) + jr_h(n)\,, \qquad (2)$$

where $r_h(n)$ is the Hilbert transform of $r(n)$ and is given by

$$r_h(n) = \begin{cases} IDFT\left[-jR(\omega)\right], & 0<\omega<\pi \\ -IDFT\left[-jR(\omega)\right], & -\pi<\omega<0 \\ 0 & \omega=0,\pi \end{cases} \qquad (3)$$

where $R(\omega)$ is the discrete Fourier transform of $r(n)$ and IDFT denotes the inverse discrete Fourier transform.

The cosine of the phase information is calculated by the following equation:

$$\text{ResidualPhase} = \frac{R_e(r_a(n))}{|r_a(n)|} \qquad (4)$$

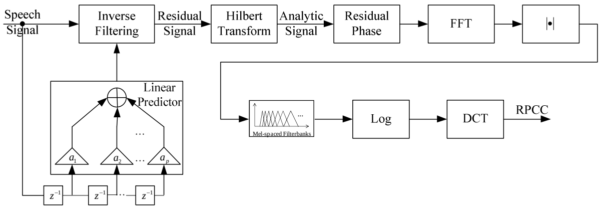

Figure 1: *Block diagram for the proposed RPCC implementation.*

In [9], the residual phase is directly implemented as a complementary feature to MFCC into their speaker recognition system. Instead, the method proposed here performs mel-spaced cepstral analysis on residual phase as shown in Figure 1. The magnitude spectrum of the residual phase is computed and warped to the Mel frequency scale followed by the usual log and DCT to obtain RPCC.

3. Method

State-of-the-art approaches for text-independent speaker verification are often based on Gaussian Mixture Model-Universal Background Model (GMM-UBM) [15]. The UBM is a speaker-independent GMM trained with speech samples from a large set of speakers to represent general speech characteristics. The hypothesized speaker model is derived from the UBM using Maximum A Posteriori (MAP) adaptation with the corresponding speech samples from a particular enrolled speaker. The strategy of adapting the target speaker model is based on the similarity between the enrollment data of target speaker and UBM, adjusting the UBM to the speaker training data. During adaptation, the distributions of the UBM which are far from the feature of target speaker remain almost unchanged. The block diagram of a speaker verification system based on GMM-UBM is showed as Figure 2.

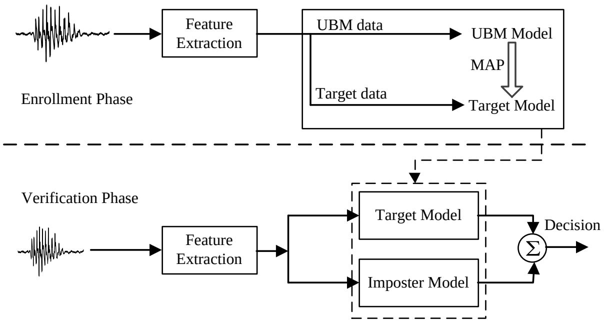

Figure 2: *Speaker verification based on GMM-UBM modeling techniques.*

4. Experiments

This section describes the experimental data, setup and results for implementing a baseline MFCC and the proposed RPCC speaker verification systems based on the UBM-GMM framework.

4.1. Data

For this particular cross-lingual speaker verification experiment, the bilingual speaker data is extracted from 2004 NIST SRE corpus. The NIST speaker corpus is a standard corpus to evaluate the performance of a speaker recognition system. Since 2004, a special effort has being made to recruit bilingual speakers who can speak Arabic, Mandarin, Russian or Spanish in addition to English. This corpus was originally collected to evaluate the effect of language, particularly differences between training and testing language, on speaker recognition systems. However, the main task of 2004 NIST SRE corpus involves speaker detection. The bilingual data of sixty-two bilingual speakers is extracted from this corpus to satisfy the data requirements of the bilingual speaker verification task. The information about the individual speakers' languages is provided by NIST.

4.2. Experimental setup

In this experiment, the UBM is trained using data from all sixty-two non-English speakers in the NIST corpus, representing 17 Arabic speakers, 19 Mandarin speakers, 16 Russian speakers, and 10 Spanish speakers. The total number of samples for initial UBM training is 552, while there are an additional 564 samples from the target speakers used for verification purposes. The total duration of training and testing data sets is 147 minutes and 149 minutes, respectively. There were an average of 9 speech samples per speaker, with an average length of about two minutes. Each target speaker's model is adapted from the global UBM using the individual English language speech samples , and the verification is performed using their alternative language speech samples.

For comparison, MFCCs are used as the baseline feature. The analysis window size is 12.5ms with a overlap of 6.25ms. Twenty two MFCCs are calculated and an LPC order of 22 is used to calculate the residual phase. The LPC residual is used to calculate RPCC features as described in the previous section, with a matching RPCC dimension of twenty-two.

Three comparison experiments have been implemented. The first experiment focuses on evaluating the performance of the system as a function of the number of mixtures. The second investigates the impact of amount of enrollment data on system performance. Each individual speaker model is trained using an increasing amount of training data. The third investigates the degree of complementary information between RPCC and MFCC features by combining the two in a single system.

4.3. EER with the increasing number of mixtures

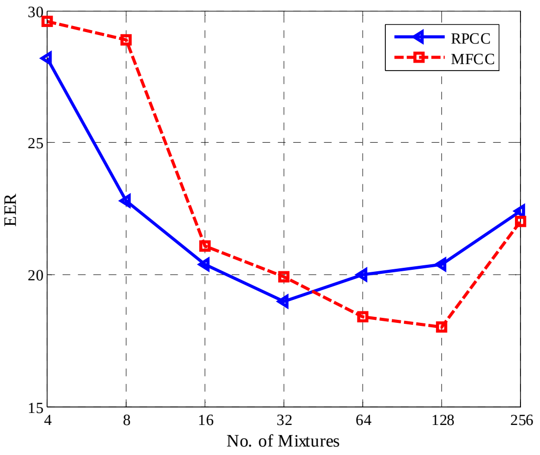

Figure 3: *EER versus increasing number of mixtures.*

Figure 3 shows the EER as a function of increasing number of mixtures using all available training data. This result supports the idea that RPCC features are more compact, needing a smaller number of model parameters to represent the information for each speaker. The RPCC features give lower error for all small model sizes up to 32 mixtures. Above that, the MFCC features give better performance, suggesting that once there is sufficient model complexity and training data the amount of total information in the MFCC features relevant to speaker identification is higher than that of RPCC.

4.4. EER with the increasing amount of training time

Figure 4 shows the EER of both MFCC and RPCC features across an increasing amount of training time with 256 mixtures. Results also support the idea that RPCC features are more compact with less dependence on phonetic content, showing lower EER in the 1, 5, and 10 second conditions. MFCC features show better performance with larger amount of data, indicating that the additional spectral and phonetic information contained in the MFCC gives better overall discriminability once enough information is available to train the models.

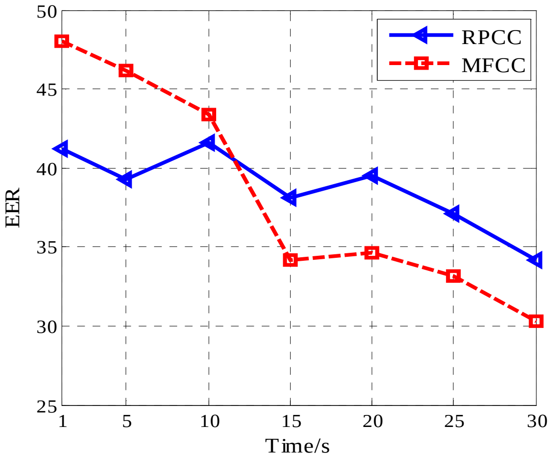

Figure 4: *EER versus duration of enrollment data.*

4.5. DET curves

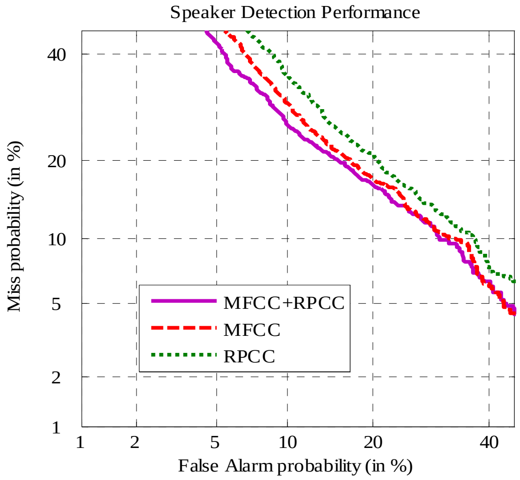

Figure 5: *DET curves comparing MFCC, RPCC and combined features.*

Figure 5 shows DET curves for MFCC features, RPCC features, and a combined feature set with 128 mixtures, using all available training data. The EER for MFCC reduced by the combined system illustrates that the information contained in these two feature sets is complementary, and that RPCC features can have benefit even with a more complex model and full length training data. The primary benefit, however, as illustrated in Figure 3 and Figure 4, is for circumstances where the amount of data and corresponding model complexity is low.

5. Conclusions

The experimental results confirm that the proposed feature provides information about speaker characteristics that is significantly different in nature from the phonetically-focused information present in traditional speaker identification features such as MFCCs. This new feature gives better results with smaller amounts of enrollment data and lower model complexities, and also provides complementary information that can improve overall system performance even for larger amounts of data. The fact that this new feature is less dependent on the phonetic content of the speaker makes it useful for low enrollment data tasks and also for tasks with language or other mismatch conditions between training and testing data, such as cross-lingual speaker identification or verification.

6. References

[1] J.P. Campbell, “speaker recognition: a tutorial,” in Proc. IEEE, vol. 85, no. 9, pp. 357-366, Aug. 1980.

[2] N. Zheng et al, “Integration of complementary acoustic features for speaker recognition,” IEEE Signal Proc. Letters, 2006.

[3] “ The 2008 NIST SRE evaluation results,” 2008. [Online]: http://www.nist.gov/speech/tests/sre/2008/official_results/index.html

[4] T. Kinnunen and P. Alku, “On separating glottal source and vocal tract information in telephony speaker verification”, in Proc. of ICASSP, Taipei, Taiwan, 2009.

[5] M. Akbacak and J. H. L. Hansen, “Language normalization for bilingual speaker recognition systems,” in Proc. of ICASSP, Hawai'i, 2007.

[6] D. Durou, “Multilingual text-independent speaker identification,” in Proc. of MIST, Leusden, The Netherlands, 1999.

[7] X. Li and M.T. Johnson, etc., “Stree and emotion classification using jitter and shimmer features,” in Proc. of ICASSP, Hawai’I, USA, 2007.

[8] R.S. Marvin, “Speaker recognition using orthogonal linear prediction,” IEEE Trans. Acoust., Speech and Signal Processing, 1976.

[9] K.S.R. Murthy and B. Yegnanarayana, “Combining evidence from residual phase and MFCC features for speaker recognition,” IEEE Signal Process. Lett., 2006.

[10] C. S. Gupta, “Significance of source features for speaker recognition,” M.S. thesis, Dept. Comput. Sci. Eng., Indian Inst.Technol., Chennai, India, 2003.

[11] T. C. Feustel, G. A. Velius, and R. J. Logan, “Human and machine performance on speaker identity verification,” Speech Tech, pp. 169-170, 1989.

[12] O. Gautherot et al. “LPC residual phase investigation,” in Proc. of EuroSpeech, 1989.

[13] S.L. Hahn, “Hilbert transforms in signal processing,” Artech House, Inc., Boston, 1996.

[14] S.R. Long et al., “The Hilbert Techniques: An Alternate Approach for Non-Steady Time Series Analysis,” IEEE Geoscience and Remote Sensing Society Newsletter, 1995.

[15] D.A. Reynolds, “Speaker Verification using Adapted Gaussian Mixture Models,” Digital Signal Processing, 2000.