VOCAL SOURCE FEATURES FOR BILINGUAL SPEAKER IDENTIFICATION

Jianglin Wang, Michael T. Johnson

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

ABSTRACT

This paper introduces the use of two new features for speaker identification, Residual Phase Cepstrum Coefficients (RPCC) and Glottal Flow Cepstrum Coefficients (GLFCC), to capture speaker-specific characteristics from their vocal excitation patterns. Results on a cross-lingual speaker identification task taken from the NIST 2004 SRE demonstrate that these RPCC and GLFCC features are significantly more accurate than traditional mel-frequency cepstral coefficients **(**MFCC). In particular, these two new features give better results with smaller amounts of training data, due to lower model complexity.

Index Terms **—** Speaker identification, Glottal source excitation, IAIF and GMM.

1. INTRODUCTION

The task of speaker recognition is an important application which has received a great deal of attention from the speech community, and there have been substantial gains in accuracy as well as channel and background robustness [1, 2]. However, features for speaker identification 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 identification.

This paper proposes two new features for speaker identification, Residual Phase Cepstral Coefficients (RPCC) and Glottal Flow Cepstrum Coefficients (GLFCC), which capture characteristics from speakers' excitation rather than vocal tract characteristics and is more compact across a wide range of phonetic conditions. The goal of these two alternative features is to rapidly capture of using the characteristic physiological features of a speaker, requiring less complex models and enabling better performance in cross-lingual or phonetically misaligned enrollment/test conditions.

This paper is organized as follows. Section 2 provides the details of each feature extraction method. The experiment data, classification method and results are described in Section 3 and 4. Final conclusions are given in Section 5.

2. FEATURE EXTRACTION

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 frame is divided into Mel-spaced triangular frequency bins, and then a Discrete Cosine Transform (DCT) is applied to calculate the desired number of cepstral coefficients.

2.2. RPCC

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 [3]. Listening experiments have also shown that residual provides valuable information that allows humans to distinguish between speakers [4]. 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 [4].

To address this, we proposed to use residual phase. The residual phase is the cosine of the phase function of the analytic signal [5]. The analytic signal is derived from the LP residual of a speech signal, defined as the error between the actual value $s(n)$ and the predicted value $\hat{s}(n)$, given by

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

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

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

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

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

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:

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

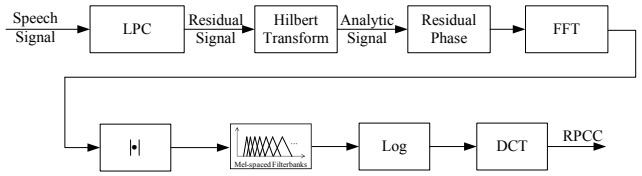

Figure 1: Block diagram for the proposed RPCC implementation.

In [5], 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.

2.3. GLFCC

The glottal flow is the airflow arising from the trachea and passing through the vocal folds. There are many reasons that the glottal flow should be speaker specific. Videos of vocal fold vibration [6] show large variations in the movement of the vocal folds from one speaker to another. For some individuals the vocal folds never close completely and in other cases vocal folds close completely and rapidly. The closing manner of vocal fold vibrations is also speaker dependent. The closure of vocal folds for some individuals shows a zipper-like pattern, while others close along the length of the vocal folds about the same time. In addition, the configuration of the area of the opening shows differences for different individuals [7, 8]. The glottal opening for some individuals is approximately equal in width along the length of the glottis, such as pressed phonation. For some individuals, a more triangular shaped opening will occur according to their own anatomical structure of vocal folds. Because of this glottal flow contains speaker specific information, and features derived from glottal flow are useful for speaker identification.

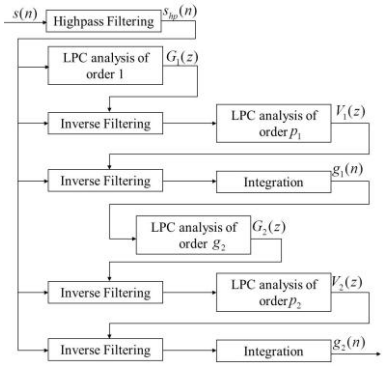

Figure 2: Structure of the IAIF Algorithm.

The accurate estimation of glottal flow has been a target of speech research for several decades. Many different methods have been developed. Among these methods, Iterative Adaptive Inverse Filtering (IAIF) [9] is a popular and has been proven to be an efficient method for estimation of the glottal flow. A flow diagram of IAIF is shown in Figure 2.

The Iterative Adaptive Inverse Filtering (IAIF) algorithm is used to estimate the glottal waveform of speech signal by filtering the original speech signal using an inverse model of the vocal tract filter, modeled as an all-pole system [9]. An example of glottal inverse filtering is shown in Figure 3. The top one is a speech signal, the middle is the LP residual and the IAIF output is on the bottom.

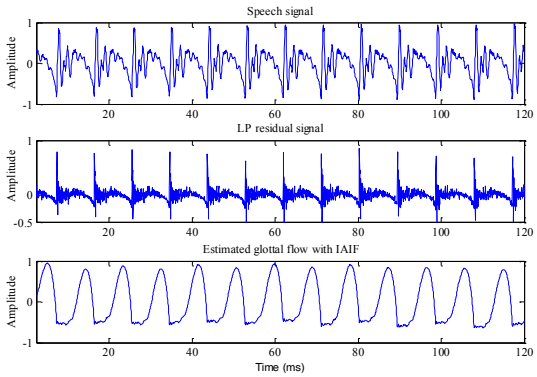

Figure 3: A diagram of glottal inverse filtering.

Glottal Flow Cepstral Coefficients are computed using mel-spaced cepstral analysis on glottal flow as shown in Figure 4. The IAIF method here helps in separating the source and filtering related information. The magnitude spectrum of the glottal flow, similar to the process of RPCC feature extraction, is computed and warped to the Mel frequency scale followed by the usual log and DCT to obtain GLFCC.

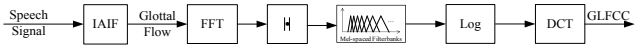

Figure 4: The feature extraction of glottal flow cepstrum coefficients.

3. METHOD

For these experiments, the Gaussian Mixture Model-Universal Background Model (GMM-UBM) [10] is applied for speaker identification. The UBM is a speaker-independent GMM trained with speech samples from a large set of speakers to represent general speech characteristics. The individual speaker model is derived from the UBM using Maximum A Posteriori (MAP) adaptation with the corresponding speech samples from a particular enrolled speaker. The UBM technique is incorporate into the GMM speaker identification system to reduce the time requirement for recognition significantly. 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.

4. EXPERIMENT RESULTS

4.1. Data

For this particular cross-lingual speaker identification experiment, 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 twenty-four bilingual speakers is extracted from this corpus to satisfy the data requirements of the bilingual speaker identification 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 twenty-four non-English speakers in the NIST corpus, representing 8 Arabic speakers, 7 Mandarin speakers, 6 Russian speakers, and 3 Spanish speakers. The total number of samples for initial UBM training is 262, while there are an additional 260 samples from the target speakers used for identification. There is an average of 8 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 identification 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 an overlap of 6.25ms. Twenty 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 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 is to evaluate their individual performance with the same dimension size.

4.3. Accuracy with the increasing number of mixtures

The accuracy versus increasing number of mixtures for MFCC and GLFCC features is shown in Figure 5. The proposed GLFCC feature shows much better performance with an increasing number of mixtures than the baseline feature MFCC. GLFCC, in particular, has a good accuracy even with the lower complexity of speaker model (small amount of mixtures). This result supports the idea that GLFCC features are more compact, needing a smaller number of model parameters to represent the information for each speaker.

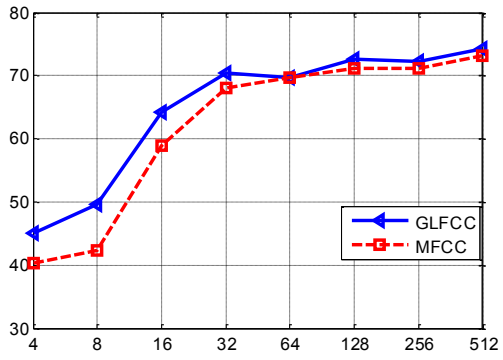

Figure 5: Accuracy versus increasing number of mixtures (MFCC&GLFCC).

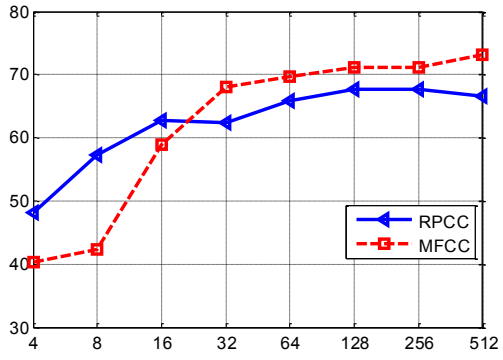

Figure 6: Accuracy versus increasing number of mixtures (MFCC&RPCC)

Figure 6 shows the accuracy versus increasing number of mixtures for MFCC and RPCC features. RPCC gives a better performance with a small number of mixtures than MFCC. MFCC features show better performance with a large number of mixtures, but results support the idea that RPCC features are more compact with less dependence on phonetic content, showing higher accuracy in the 4, 8, and 16 mixtures.

According to the above results, the feature of GLFCC and RPCC are clearly the strongest individual component within the excitation-related measures.

4.4. The accuracy of individual feature

Table 1: The classification accuracy of individual features

Individual Feature	Accuracy (%)
MFCC	71.2
GLFCC	72.3
RPCC	67.7

Table 1 shows the results of each individual feature with the same dimension 20. The number of mixtures for all three features is 256. Individually, the best overall feature is the GLFCC feature followed by MFCCS. Although RPCC has the lowest accuracy comparing to other two features, RPCC gives highest classification with the small number of mixtures as shown in Figure 5 and Figure 6.

5. CONCLUSIONS

This paper has introduced two speaker-specific features for speaker identification based on GMM-UBM system. The experimental results show that the proposed features provide information about speaker characteristics that is significantly different in nature from the phonetically-focused information present in traditional speaker identification features such as MFCCs. These two new features give better results with lower model complexities. The fact that these new features are less dependent on the phonetic content of the speaker makes it useful for tasks with language or other mismatch conditions between training and testing data, such as cross-lingual speaker identification or verification.

REFERENCES

[1] J. P. Campbell, "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, pp. 357-366, 1980.

[2] N. Zheng, T. Lee, and P. C. Ching, "Integration of complementary acoustic features for speaker recognition," *IEEE Signal Proc. Letters*, vol. 14, 2006.

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

[4] O. Gautherot, "LPC residual phase investigation," *in Proc. of EuroSpeech*, 1989.

[5] K. S. R. Murthy and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13, pp. 52-56, 2006.

[6] B. T. Labs, "High speed motion pictures of the human vocal cords," *Bureau of Publication*, 1937.

[7] H. Hanson, "Glottal characteristics of femal speakers," *Ph.D. dissertation*, vol. Harvard University, 1995.

[8] 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 and Audio Proc.*, 1999.

[9] P. Alku, "Glottal Wave Analysis With Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Communication*, pp. 109-118, 1992.

[10] D. A. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.