PHYSIOLOGICALLY-MOTIVATED FEATURE EXTRACTION FOR 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 three physiologically-motivated features for speaker identification, Residual Phase Cepstrum Coefficients (RPCC), Glottal Flow Cepstrum Coefficients (GLFCC) and Teager Phase Cepstrum Coefficients (TPCC). These features capture speaker-discriminative characteristics from different aspects of glottal source excitation patterns. The proposed physiologically-driven features give better results with lower model complexities, and also provide complementary information that can improve overall system performance even for larger amounts of data. Results on speaker identification using the YOHO corpus demonstrate that these physiologically-driven features are both more accurate than and complementary to traditional mel-frequency cepstral coefficients **(**MFCC). In particular, the incorporation of the proposed glottal source features offers significant overall improvement to the robustness and accuracy of speaker identification tasks.

Index Terms— Speaker distinctive feature, Speaker identification, Glottal source excitation and GMM-UBM

1. INTRODUCTION

The task of speaker identification and verification has received a great deal of attention from the research community in the past decade, with substantial gains in accuracy as well as channel and background robustness [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. This causes several significant limitations, including the need for models that represent a speaker's entire phonetic space and higher model complexity to cover this space.

Although MFCCs have been widely applied to speech and speaker recognition, MFCC features also have limitations [3]. One of the goals of using MFCCs for speech recognition is to eliminate speaker-specific information for different speakers [4, 5], capturing the common representative acoustic information. In contrast, the goal for speaker recognition is to extract speaker-specific information while minimizing the impact of unrelated phonetic information. Since pronunciation patterns are unique, MFCCs are still effective features for speaker recognition, but this is also somewhat contradictory considering the opposite nature of these two tasks. The use of the same representation for both speech and speaker recognition is ironic, and indicates an opportunity for generating better performance by characterizing and incorporating features that are less connected to phonetic information.

The glottal source waveform contains much information about the unique physiological properties of an individual's speech production mechanism [6]. There has been some recent work in this direction [7-9], and this paper focuses on further development of effective vocal source features for speaker identification. The glottal source signal represents the musculature and tension of the vocal folds, and the associated glottal pulse parameters, including the rate of the closing phase and the degree of the glottal opening. The vibratory pattern of the vocal folds not only produces a voicing source for speech production, but also characterizes unique nonlinear flow patterns for each speaker [10]. The quasi-periodic motion of relevant vocal organs generates a pulse-like epoch shape that varies among speakers. These characteristics are unique to a given speaker's speech production system. Hence, features derived from the vocal source have capacity to provide valuable information for speaker recognition.

Feature extraction methods for capturing the vocal tract characteristics of speaker, such as MFCCs, linear predictive cepstral coefficients (LPCCs) [11], line spectral frequencies (LSFs) [11] and log area ratios (LARs) [12], have been investigated for many years. These features can accurately characterize the vocal tract configuration of a speaker, and

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

can achieve good performance in current speaker recognition systems. However, the usefulness of vocal source excitation related features is still under-investigated. Inspired by the physiological significance of the vocal source characteristics residing in speech production system, this paper investigates several speaker specific source features using novel signal processing approaches.

This paper is organized as follows. Section 2 provides the details of the proposed feature extraction methods. The baseline GMM-UBM speaker identification system is described in Section 3. Section 4 describes the experimental data, setup and results, with final conclusions in Section 5.

2. VOCAL SOURCE FEATURE EXTRACTION

2.1. Residual phase cepstrum coefficients

The Linear Predictive Coding (LPC) residual of a speaker represents the impulse-like excitation which is related to the region around the glottal closure instant within each pitch period. These regions are known to contain speaker-specific information [8, 13]. Listening experiments have also shown that residual provides valuable information that allows humans to distinguish between speakers [14]. 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. Gautherot reported that the magnitude spectrum of the LPC residual is flat, suggesting that substantial information retained in the phase [14].

The Residual Phase (RP) is defined as the cosine of the phase function of the analytic signal [15, 16]. The analysis is based on the residual of a speech signal, which is defined as the error between the actual value $s(n)$ and the LPC predicted value $\hat{s}(n)$. Rather than using the residual directly or the residual magnitude spectrum, the analytic signal is calculated via a Hilbert transform and the phase component is then extracted as cosine of the analytic signal $r_a(n)$ [16]:

$$\text{ResidualPhase} = \frac{\text{Re}\,al(r_a(n))}{|r_a(n)|} \quad (1)$$

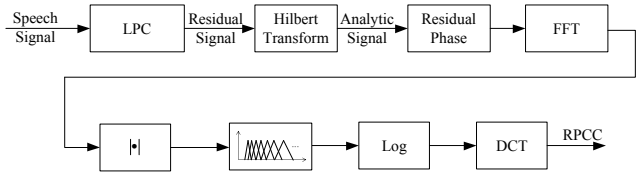

Fig. 1. Residual Phase Cepstrum Coefficients (RPCC)

In contrast to [16], where the residual phase is directly implemented as a complementary feature to MFCCs for speaker recognition, the method proposed here performs mel-spaced cepstral analysis on the spectrum of the residual phase signal, followed by log and DCT operations, as shown in Fig. 1. The resulting Residual Phase Cepstral Coefficients (RPCC) compactly represent the phase information of the underlying excitation waveform.

2.2. Glottal flow cepstrum coefficients

The glottal flow is the airflow arising from the trachea and passing through the vocal folds. There is significant supporting evidence demonstrating that the glottal flow is speaker specific [17]. Videos of vocal fold vibration [18] 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 duration of vocal fold opening and closing, the glottal closing instants (GCIs) and opening instants (GOIs), and the shape of the glottal flow vary significantly across speakers. These variations correspond to the variations in the glottis, and then are reflected in the glottal flow. Therefore, the glottal flow contains speaker distinctive information and features derived from glottal flow are expected to be useful for speaker identification.

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, Pitch Synchronous Iterative Adaptive Inverse Filtering (PSIAIF) [19] is popular and has been proven to be an efficient method for estimation of the glottal flow. The PSIAIF 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. In this work the magnitude spectrum of the PSIAIF-estimated flow waveform is used to represent the glottal flow characteristics. The FFT magnitude spectrum is warped to the Mel frequency scale followed by log and DCT operators to obtain Glottal Flow Cepstral Coefficients (GLFCC). An overview of this process is shown in Fig. 2. The GLFCC features thus represent the spectral magnitude characteristics of a speaker's glottal excitation pattern.

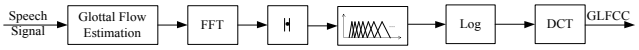

Fig. 2. Glottal Flow Cepstrum Coefficients (GLFCC)

2.3. Teager phase cepstrum coefficients

Most speech processing models are based on the traditional linear source-filter speech production model, assuming that the airflow propagates in the vocal tract as a plane wave. It is well understood that the true airflow propagation is much more complex [17]. Research by Teager [20] models the air flow through a series of separate and simultaneous vortices

distributed throughout the vocal tract. The resulting instantaneous Teager-Kaiser energy operator (TEO) provides an advantage over Fourier analysis methods in capturing the characteristics of nonlinear systems [21], measuring the underlying energy required for production rather than the energy of the resulting waveform. The TEO is applicable for analysis and estimation of the nonlinear characteristics of the existing amplitude and frequency modulation patterns in a vocal excitation signal. Based on this approach, we have used the TEO to characterize the vibration characteristics yielded by the vocal folds for potential speaker-specific feature extraction. Features derived from the TEO are used to reflect properties of the speech production process that are not covered by features derived from the linear model of speech production.

The TEO operator in the discrete-time form is

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1) \quad (2)$$

where $x(n)$ is the sampled speech signal and $\Psi[\cdot]$ is the TEO operator. This TEO is typically applied to a band-pass filtered speech signal since its purpose is to reflect the energy of this nonlinear flow within the vocal tract for a single resonant frequency. The corresponding TEO profile can be used to decompose a speech signal into its amplitude modulation (AM) and frequency modulation (FM) components within a certain frequency band [17]:

$$f(n) \approx \frac{1}{2\pi T}\arccos\left(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]}\right) \quad (3)$$

$$|a(n)| \approx \sqrt{\frac{\Psi[x(n)]}{[1-(1-\frac{\Psi[y(n)]+\Psi[y(n+1)]}{4\Psi[x(n)]})^2]}}, \quad (4)$$

where $y(n) = x(n) - x(n-1)$ is the time domain difference signal, $f(n)$ is the FM component at sample n, and $a(n)$ is the AM component at sample n.

For the task of speaker recognition being investigated here, we have again used the phase of the signal rather than the magnitude spectrum to represent speaker-specific characteristics, following the computational structure used for the RPCC.

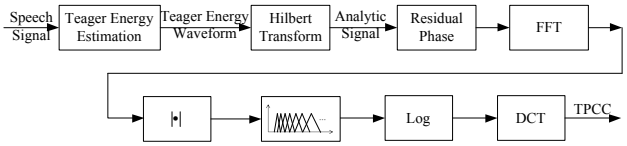

Fig. 3. Teager Phase Cepstrum Coefficients (TPCC)

Fig. 3 shows a block diagram of the proposed Teager Phase Cepstrum Coefficients (TPCC). The excitation energy contour is initially calculated through the Teager energy operator. Then the fine energy structure is obtained by a Hilbert transformation, and the cepstrum of the fine energy structure is computed and warped to the Mel frequency scale followed by a log and DCT operation, to obtain TPCC. The TPCC features computed in this way represent the phase characteristics of the Teager nonlinear energy model of the speech production process.

3. METHODS

The baseline classification framework in this paper is based on a Gaussian Mixture Model and Universal Background Model (GMM-UBM) approach [22], commonly used in the speech processing community to perform speaker recognition. This approach has advantages in its flexibility and robustness to duration and temporal alignment differences between training and testing examples. 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.

4. EXPERIMENTAL RESULTS

4.1. Data corpus

The proposed new features were evaluated on the YOHO speaker identification task. The YOHO corpus was collected by ITT under a US government contract and was designed for speaker recognition systems in office environments with limited vocabulary [23]. This database was recorded using a telephone handset in a real office environment and sampled at an 8 kHz sampling frequency with 16 bits per sample. This corpus consists of 138 speakers each with 24 training utterances and 40 test utterances recorded in different sessions. The vocabulary consists of 56 two-digit numbers, ranging from 21 to 97 spoken continuously in sets of three (e.g., 32-56-68) in each utterance. In this work, all the utterances identified as enrollment data were used to train the model, and utterances in the verification data set were used for testing. Each enrollment session consists of 24 phrases and each testing utterance is single example. There are about 6 minutes of speech used for training each speaker, and 2.4 seconds of speech for testing.

4.2. Experimental setup

As introduced in Section 3, the GMM-UBM speaker identification framework is used for evaluation. Initially, a universal background model is trained using the training utterances from all 138 speakers. Following this each speaker's model is adapted from the corresponding training utterance using the MAP adaption approach. The identification experiments were conducted on the YOHO

database as the number of mixtures is increased. This experimental configuration is designed to evaluate if the proposed features can rapidly build an accurate model with lower model complexity.

The speech utterances were analyzed using a 32ms frame with a 50% frame overlap, and twelve coefficients of each feature (MFCC, RPCC, GLFCC and TPCC) are derived from each frame. A baseline system using MFCC features was evaluated. The first experiment uses individual features alone to assess their individual performance respectively, and then proposed features are appended to the baseline MFCC in order to evaluate their complimentary characteristics to the baseline feature.

Since the motivation behind the proposed features is to identify information that corresponds more to physiological structure and less to phonetic characteristics and pronunciation patterns, the accuracy of the system as a function of model complexity, represented by number of mixtures, is used to evaluate whether the proposed features are able to better model speaker differences in a model space with reduced parameters and representative power.

4.3. Accuracy of individual features

The accuracy versus increasing number of mixtures for the proposed features is shown in Fig. 4. At low model complexities, RPCC and GLFCC features show better performance than the baseline MFCC features, although the TPCC features do not. Of great interest is that the GLFCC features outperform the traditional MFCC features across all model configurations. This is particularly meaningful because the GLFCC features are based entirely on the glottal flow waveform, with spectral information removed.

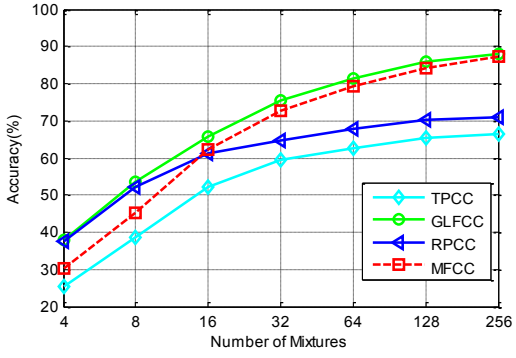

Fig. 4. SID performance on YOHO with an increasing number of Gaussian mixture components

4.4. Accuracy of combined features

Fig. 5 shows classification accuracy using MFCC features combined with the proposed source features against increasing model complexity. The primary observation is that, as hypothesized, the spectral information of the MFCCs and the vocal excitation information of the proposed features is clearly complementary, with each combination outperforming the baseline by a substantial margin. Combining the MFCCs with all proposed features gives noticeable additional improvement.

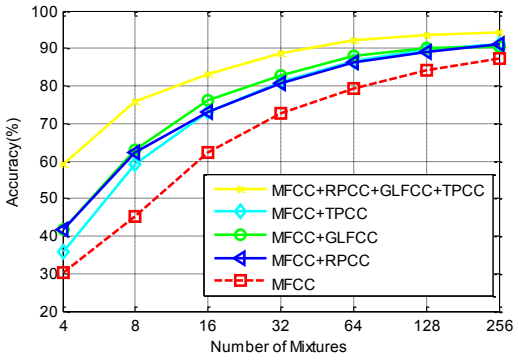

Fig. 5. SID performance of combined features with an increasing number of Gaussian mixture components

The performance across all model complexities clearly show the robustness of combining the baseline features with the proposed vocal source features. At low model complexities the improvement is quite large, e.g. 60% accuracy for the combined features vs. the 30% baseline MFCC at 4 mixtures. Table 1 shows the final system performance, with a final performance of 94.3% compared to the original 87.3% baseline, cutting error by more than half.

Table 1. Accuracy of the final system with 256 mixtures

Feature Combinations	Accuracy
MFCC + Proposed features	94.3
MFCC	87.3

5. CONCLUSIONS

This paper has introduced three speaker-distinctive features for speaker identification based on vocal source characteristics, including residual phase, glottal flow, and phase of Teager energy. These features represent unique and individually distinct aspects of the underlying vocal source excitation. 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 spectral features such as MFCCs. The incorporation of the proposed glottal source features offers significant overall improvement to the robustness and accuracy of speaker identification tasks. The fact that the proposed features have better performance at low model complexities suggests that these new features are less dependent on the underlying phonetic content of the speech and may be useful for a wide variety of speaker identification and verification applications.

6. 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] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Signal Processing*, pp. 357-366, 1980.

[4] R. D. Zilca, J. Navratil, and G. N. Ramaswamy, "Depitch and the role of fundamental frequency in speaker recognition," *ICASSP*, 2003.

[5] K. Chen and A. Salman, "Extracting speaker-specific information with a regularized Siamese deep network," *Advances in Neural Information Processing Systems*, 2011.

[6] G. Fant, "Glottal source and excitation analysis," *Speech transmission laboratory, Royal Institute of Technology, Quarterly Progress and Status Report*, vol. 20, pp. 85-107, 1979.

[7] I. Hernaez, I. Saratrxaga, J. Sanchez, E. Navas, and I. Luengo, "Use of the harmonic phase in speaker recognition," *Proc. Interspeech*, pp. 2757-2760, 2011.

[8] L. Wang, S. Ohtsuka, and S. Nakagawa, "High improvement of speaker identification and verification by combining MFCC and phase information," *Proc. ICASSP*, pp. 4529-4532, 2009.

[9] T. Drugman and T. Dutoit, "On the potential of glottal signatures for speaker recognition," *Proc. Interspeech*, 2010.

[10] B. H. Hildebrand, "Vibratory patterns of the human vocal cords during variations in frequency and intensity," in *Doctoral Diss.*: Univ. of Florida, 1976.

[11] X. Huang and A. Acero, "Spoken Language Processing," *Prentice Hall, Upper Saddle River, New Jersey*, 2001.

[12] L. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition," *Prentice Hall Press*, 1993.

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

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

[15] J. Wang, "Physiologically-motivated feature extraction methods for speaker recognition," in *Doctoral Dissertation*: Marquette University, 2013.

[16] 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.

[17] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*: Prentice Hall, 2002.

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

[19] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, pp. 109-118, 1992.

[20] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 599-601, 1980.

[21] S. Das and J. H. L. Hansen, "Detection of voice onset time (VOT) for unvoiced stops using the Teager Energy Operator (TEO) for automatic detection of accented English," *Proceedings of the 6th Nordic Signal Processing Symposium*, 2004.

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

[23] J. Campbell and H. Alan, *YOHO Speaker Verification (LDC94S16)*, 1994.