Combining Vocal Source and MFCC Features for Enhanced Speaker Recognition Performance Using GMMs

Danoush Hosseinzadeh and Sridhar Krishnan
Department of Electrical and Computer Engineering
Ryerson University, Toronto, ON - M5B 2K3 Canada
Email: (danoushh@hotmail.com) (krishnan@ee.ryerson.ca)

***Abstract*— This work presents seven novel spectral features for speaker recognition. These features are the spectral centroid (SC), spectral bandwidth (SBW), spectral band energy (SBE), spectral crest factor (SCF), spectral flatness measure (SFM), Shannon entropy (SE) and Renyi entropy (RE). The proposed spectral features can quantify some of the characteristics of the vocal source or the excitation component of speech. This is useful for speaker recognition since vocal source information is known to be complementary to the vocal tract transfer function, which is usually obtained using the Mel frequency cepstral coefficients (MFCC) or linear predication cepstral coefficients (LPCC). To evaluate the performance of the spectral features, experiments were performed using a text-independent cohort Gaussian mixture model (GMM) speaker identification system. Based on 623 users from the TIMIT database, the spectral features achieved an identification accuracy of 99.33% when combined with the MFCC based features and when using undistorted speech. This represents a 4.03% improvement over the baseline system trained with only MFCC and ΔMFCC features.**

I. Introduction

Speaker recognition has many potential applications as a biometric tool for resources that can be accessed via the telephone or internet. In these applications, the identity of users cannot be verified because there is no direct contact between the user and the service provider. Hence, speaker recognition is a cost effective and practical technology that can be used for enhanced security.

Often in literature, the entire speech system is modeled with a time-varying excitation and a short-time-varying filter [1]. Using this model, the source and filter are assumed independent and hence the speech signal ($s(t)$) is modeled by the linear convolution of:

$$s(t) = x(t) * h(t) \tag{1}$$

where, $x(t)$ is a periodic excitation (for voiced speech) or white noise (for unvoiced speech) and $h(t)$ is a time-varying filter which constantly changes to produce different sounds. Although $h(t)$ is time varying, it can be considered stable over a period of a few milliseconds (ms); typically around 10-30ms is commonly used in literature [1]. This convenient short-time stationary behavior is exploited by many speaker recognition systems in order to characterize the vocal tract configuration given by $h(t)$, which is known to be a unique speaker-dependent characteristic for a given sound. While assuming a linear model, this information can be easily extracted from speech signals using well established deconvolution techniques such as homomorphic filtering or linear prediction methods.

To date, the most effective features for speaker recognition have been the Mel frequency cepstral coefficient (MFCC) and the linear prediction cepstral coefficients (LPCC) [2][1][3]. These features can accurately characterize the vocal tract configuration of a speaker and can achieve good performance. Part of the success of these features is that they provide a compact representation of the vocal tract which can be modeled effectively. The first several MFCCs can characterize the speaker's vocal tract configuration and LPCCs generally define lower order polynomials [1]. Additionally, the first derivative of the MFCC feature (ΔMFCC) is largely uncorrelated with the MFCC feature and has been shown to enhance recognition performance.

Although the MFCC and LPCC based features have proven to be effective for speaker recognition, they do not provide a complete description of the speaker's speech system. Hence, vocal source information can complement these traditional features by quantifying some speaker-dependent characteristics such as pitch, harmonic structure and spectral energy distribution [4][5].

This work proposes seven novel spectral features for speaker recognition that can quantify the vocal source. These features are the spectral centroid (SC), spectral bandwidth (SBW), spectral band energy (SBE), spectral crest factor (SCF), spectral flatness measure (SFM), Shannon entropy (SE), and Renyi entropy (RE). These spectral features can be used to complement the MFCC or LPCC features since they can quantify characteristics of the vocal source.

It is also known that there is some degree of coupling between the vocal source and vocal tract [6][4] - i.e. the linear model assumed when calculating MFCC and LPCC is not entirely accurate. Therefore, the vocal source signal is to some extent predictable for a given vocal tract configuration. Given these factors, features that characterize the vocal source can be expected to improve the performance of existing speaker recognition systems. In this work, the seven proposed spectral features are extracted from the speech spectrum and used to enhance the performance of MFCC-based features in order to illustrate their effectiveness.

Others have attempted to use the vocal source for improving performance of speaker recognition systems. Attempts have been made to develop features from the LPCC residual [7][8] with some success. In these cases, the authors have noted improved performance by complementing vocal tract features with vocal source information.

The paper is organized as follows. Section II defines the baseline system used for testing and presents the spectral features. Section III presents the results as well as the experimental conditions and Section IV concludes the paper.

II. Proposed Testing Method

GMM based speaker recognition systems have become the most popular method to date. This is because GMMs can capture the acoustic phenomena or acoustic classes that are present in speech [2]. In fact, some of the GMM clusters have been found to be highly correlated with particular phonemes [9]. As a result, good recognition performance can be achieved with GMM based systems. The performance of the proposed spectral features will be compared to the baseline system, which is a cohort text-independent GMM classifier trained with 14-dimensional MFCC vectors and 14-dimensional ΔMFCC vectors extracted from 30ms speech frames. The log-likelihood function is used to find the user model that best matches a given utterance.

TABLE I
SUBBAND ALLOCATION USED TO CALCULATED SPECTRAL FEATURES.

Subband	Lower Edge (Hz)	Upper Edge (Hz)
1	300	627
2	628	1060
3	1061	1633
4	1634	2393
5	2394	3400

A. Training and GMM Estimation

The expectation maximization (EM) algorithm was used to estimate the parameters of the GMM models. In the past, model orders of 8-32 have been commonly used in literature however, good results have been obtained with cohort GMM systems using as little as 16 components [2][10]. A model order of 24 was in this work to account for the additional features being used in the system and also, preliminary experimental results indicate that this model order was the optimal order for the proposed feature set given models of order 16, 20, 24, 28 and 32. The k-means algorithm was used to obtain the initial estimate for each cluster since it has been shown that the initial grouping of data does not significantly affect the performance of GMM based recognition systems [2].

A diagonal covariance matrix was used to estimate the variances of each cluster in the models since it is well known that diagonal covariance matrices are much more computationally efficient than full covariance matrices. Furthermore, diagonal covariance matrices can provide the same level of performance as full covariance matrices because they can capture the correlation between the features if a larger model order is used [11]. For these reasons, diagonal covariance matrices have almost been exclusively used in previous speaker recognition works. Each element of these matrices is limited to a minimum value of 0.01 during the EM estimation process to prevent singularities in the matrix, as recommended by [2].

B. Spectral Features

The proposed spectral features can be expected to improve the performance of MFCC or LPCC features because they can capture complementary information related to the vocal source such as pitch, harmonic structure, energy distribution, bandwidth of the speech spectrum and even voiced or unvoiced excitation. To illustrate the effectiveness of these features, they are extracted from the speech spectrum and used to enhance the performance of MFCC and ΔMFCC features.

Spectral features should be extracted from multiple subbands, as shown in Table I. This extraction method will provide better discrimination between different speakers because the trend for a given feature can be captured from the spectrum. This is better than obtaining one global value from the spectrum, which is not likely to show speaker-dependent characteristics.

The proposed subbands are linearly spaced on the Mel scale and spans the range of a practical telephone channel (300Hz-3.4kHz). This allocation scheme reflects the fact that most of the energy of the speech signal is located in the lower frequency regions and therefore, narrowly defined subbands are used in the lower frequency regions in order to capture more detail. This is also consistent with the non-linearities of human auditory perception, which shows more sensitivity to lower frequencies than higher frequencies. This non-linearity has been shown to be important for cepstral based features such as the MFCC feature [3].

Spectral features are extracted from 30ms speech frames as follows. Let $s_i[n]$ for $n \in [0, N]$, represents the i^{th} speech frame and $S_i[f]$ represents the spectrum of this frame. Then, $S_i[f]$ can be divided into M non-overlapping subbands where, each subband (b) is defined by a lower frequency edge (l_b) and a upper frequency edge (u_b). Now, each of the seven spectral features can be calculated from $S_i[f]$ as shown below.

Spectral Centroid (SC) - SC as given below is the weighted average frequency for a given subband, where the weights are the normalized energy of each frequency component in that subband. Since this measure captures the center of gravity of each subband it can locate large peaks in subbands. These peaks correspond to the approximate location of formants [12] or pitch frequencies.

$$SC_{i,b} = \frac{\sum_{f=l_b}^{u_b} f|S_i[f]|^2}{\sum_{f=l_b}^{u_b} |S_i[f]|^2} \tag{2}$$

Spectral Bandwidth (SBW) - SBW as given below is the weighted average distance from each frequency component in a subband to the spectral centroid of that subband. Here, the weights are the normalized energy of each frequency component in that subband. This measure quantifies the relative spread of each subband for a given sound and therefore, it might characterize some speaker-dependent information.

$$SBW_{i,b} = \frac{\sum_{f=l_b}^{u_b} (f - SC_{i,b})^2 \, |S_i[f]|^2}{\sum_{f=l_b}^{u_b} |S_i[f]|^2} \tag{3}$$

Spectral Band Energy (SBE) - SBE as given below is the energy of each subband normalized with the combined energy of the spectrum. The SBE gives the trend of energy distribution for a given sound and therefore, it contains some speaker-dependent information.

$$SBE_{i,b} = \frac{\sum_{f=l_b}^{u_b} |S_i[f]|^2}{\sum_{f,b} |S[f]|^2} \tag{4}$$

Spectral Flatness Measure (SFM) - SFM as given below is a measure of the flatness of the spectrum, where white noise has a perfectly flat spectrum. This measure is useful for discriminating between voiced and un-voiced components of speech [13].

$$SFM_{i,b} = \frac{\left[\prod_{f=l_b}^{u_b} |S_i[f]|^2\right]^{\frac{1}{u_b - l_b + 1}}}{\frac{1}{u_b - l_b + 1} \sum_{f=l_b}^{u_b} |S_i[f]|^2} \tag{5}$$

Spectral Crest Factor (SCF) - SCF as given below provides a measure for quantifying the tonality of the signal. This measure is useful for discriminating between wideband and narrowband signals by indicating the relative peak of a subband. These peaks correspond to the most dominant pitch frequency in each subband.

$$SCF_{i,b} = \frac{max(|S_i[f]|^2)}{\frac{1}{u_b - l_b + 1} \sum_{f=l_b}^{u_b} |S_i[f]|^2} \tag{6}$$

Renyi Entropy (RE) - RE as given below is an information theoretic measure that quantifies the randomness of the subband. Here, the normalized energy of the subband can be treated as a probability distribution for calculating entropy and α is set to 3, as commonly found in literature [14]. This RE trend is useful for detecting the voiced and unvoiced components of speech.

$$RE_{i,b} = \frac{1}{1 - \alpha} \log_2 \left(\sum_{f=l_b}^{u_b} \left| \frac{S_i[f]}{\sum_{f=l_b}^{u_b} S_i[f]} \right|^{\alpha} \right) \tag{7}$$

Shannon Entropy (SE) - SE as given below is also an information theoretic measure that quantifies the randomness of the subband. Here, the normalized energy of the subband can be treated as a

probability distribution for calculating entropy. Similar to the RE trend, the SE trend is also useful for detecting the voiced and unvoiced components of speech.

$$SE_{i,b} = -\sum_{f=l_b}^{u_b} \left| \frac{S_i[f]}{\sum_{f=l_b}^{u_b} S_i[f]} \right| \cdot \log_2 \left| \frac{S_i[f]}{\sum_{f=l_b}^{u_b} S_i[f]} \right| \quad (8)$$

To the best of our knowledge, these features are being used for the first time in speaker recognition although they have previously been used in other areas [15]. These spectral features along with the MFCC and ΔMFCC features will be extracted from each speech frame and appended together to form a combined feature matrix for the speech signal. These vectors can then be modeled and used for speaker recognition. Equation 9 shows the feature matrix that can be extracted based on only one spectral feature, say the SC feature, from i frames; where the bracketed number is the length of the feature. It should be noted that any other spectral feature can be substituted in for the SC feature in the feature matrix.

$$\overrightarrow{\mathcal{F}} = \begin{bmatrix} MFCC_1(14) & \Delta MFCC_1(14) & SC_1(5) \\ \vdots & \vdots & \vdots \\ MFCC_i(14) & \Delta MFCC_i(14) & SC_i(5) \end{bmatrix} \quad (9)$$

The spectral features are expected to be largely uncorrelated with the MFCC based features because the spectral features can capture some information about the vocal source, whereas the MFCC features tend to capture information about the vocal tract. Among the spectral features, there may be some correlation between the SC and the SCF features because they both quantify information about the peaks (locations of energy concentration) of each subband. The difference is that the SCF feature describes the normalized strength of the largest peak in each subband while the SC feature describes the center of gravity of each subband. Therefore, these features will perform well if the largest peak in a given subband is much larger than all other peaks in that subband. The RE and SE features are also correlated since they are both entropy measures. However, the RE feature is much more sensitive to small changes in the spectrum because of the exponent term α. Therefore, although these features quantify the same type of information, their performance may be different for speech signals.

III. Experimental Results

All speech samples used in these experiments were obtained from 623 speakers of the TIMIT speech corpus. Since the TIMIT database has a sampling frequency of 16kHz, the signals were down sampled to 8kHz which is well suited for telephone applications. Features were extracted from 30ms long frames with 15ms of overlap with the previous frames and a Hamming window was applied to each frame to ensure a smooth frequency transition between frames. Twenty seconds of undistorted speech from each speaker was used to train the system and the remaining samples were used for testing. Although the tests were performed with undistorted audio, it is expected that some of these features will remain robust to different linear and non-linear distortions [15].

A. Results and Discussions

MFCC based features are very effective for characterizing the vocal tract configuration. Although this is the main reason for the success of the MFCC based features, they do not provide a complete description of the speaker's speech system. The proposed spectral features are expected to increase identification accuracy of MFCC based systems because they provide some information about the vocal source.

TABLE II
Experimental results using 7s test utterances (298 tests)

Feature	Accuracy(%)
MFCC & ΔMFCC (Baseline system)	95.30
MFCC & ΔMFCC & SC	97.32
MFCC & ΔMFCC & SBE	97.32
MFCC & ΔMFCC & SBW	96.98
MFCC & ΔMFCC & SCF	96.31
MFCC & ΔMFCC & SFM	81.55
MFCC & ΔMFCC & SE	90.27
MFCC & ΔMFCC & RE	98.32
MFCC & ΔMFCC & SBE & SC	96.98
MFCC & ΔMFCC & SBE & RE	96.98
MFCC & ΔMFCC & SC & RE	**99.33**

Table II demonstrates the identification accuracy of the system when using spectral features in addition to the MFCC based features with undistorted speech. The table also shows several combinations of the best performing features. The accuracy rate represents the percentage of test samples that were correctly identified by the system, as shown below.

$$\text{Accuracy} = \frac{\text{Samples Correctly Identified}}{\text{Total Number of Samples}} \quad (10)$$

It is evident from these results that there is some speaker-dependent information captured by most of the proposed features since they improved identification rates when combined with the standard MFCC based features. In fact, when two of the best performing spectral features (SC and RE) were simultaneously combined with the MFCC based features, an identification accuracy of 99.33% was achieved, which represents a 4.03% improvement over the baseline system. These results suggest that the proposed spectral features provide complementary and discriminatory information about the speaker's vocal source and system, which leads to enhanced identification accuracies.

The best performing feature was the RE feature. This feature is very effective at quantifying voiced speech which is quasi-periodic (relatively low entropy) and un-voiced which is often represented by AWGN (relatively high entropy). However, we suspect that the RE feature may also be characterizing another phenomena other than voice and unvoiced speech. This is likely since the SE feature did not show any performance benefits and it too is an entropy measure capable of discriminating between voiced and unvoiced speech. One possibility is that the exponential term α in the RE definition is contributing to this performance improvement. Since the spectrum is a normalized to the range of [0,1] before calculating these features, the exponent term α has the effect of significantly reducing the contributions of the low energy components relative to the high energy components. Therefore, the RE feature is likely to produce a more reliable measure since it heavily relies on the high energy components of each subband. However, the entropy features in general are susceptible to random noise and will not perform well in all conditions.

Figure 1(a) shows that the SC feature can capture the center of gravity of each subband. Since the subband's center of gravity is related to the spectral shape of the speech signal, it implies that the SC feature can also detect changes in pitch and harmonic structure since they fundamentally affect the spectrum. Pitch and harmonic structure convey some speaker-dependent information and are complementary

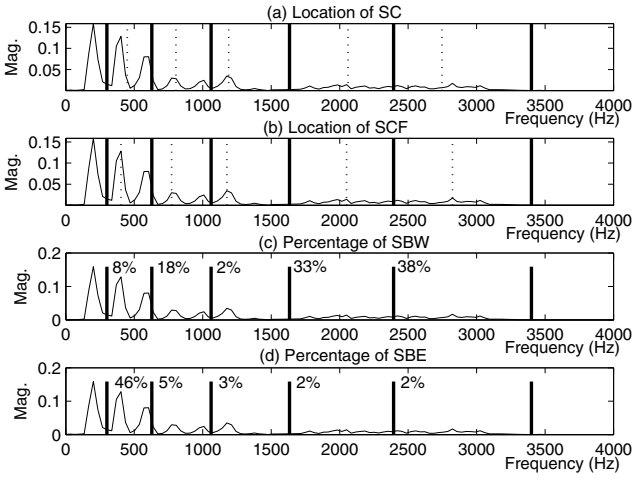

Fig. 1. Plot of the spectral features. Subband boundaries are indicated with dark solid lines and feature location is indicated with dashed lines. (a) Location of the SC (b) Location of the SCF (c) SBW as a percentage of the five subbands. (d) SBE as a percentage of the of the whole spectrum.

to the vocal tract transfer function for speaker recognition. In addition, the SC feature can also locate the approximate location of the dominant formant in each of the subbands since formants will tend towards the subband's center of gravity. These properties of the SC feature provide complementary information and led to the improved performance of the MFCC based classifier.

The SCF feature shown in Figure 1(b) quantifies the normalized strength of the dominant peak in each subband. Given that the dominant peak in each subband corresponds to a particular pitch frequency harmonic, it shows that the SCF feature is pitch dependent and therefore, it is also speaker-dependent for a given sound. This dependence on pitch frequency is useful when the vocal tract configuration (i.e. MFCC) is known as seen by the enhanced performance. Moreover, the SCF feature is a normalized measure and should not be significantly affected by the intensity of speech from different sessions.

The SBE feature, shown in Figure 1(d), also performed well in the experiments. This feature provides the distribution of energy in each subband as a percentage of the entire spectrum, which is another measure that can quantify the harmonic structure of the signal. The SBE feature is also a normalized energy measure and should not be significantly affected by the intensity (or relative loudness) of speech from different sessions. The results in Table II suggests that for a given vocal tract configuration the SBE trend is predictable and complementary for speaker recognition.

The SBW feature is largely dependent on the SC feature and the energy distribution of each subband therefore, it has also performed well for the reasons mentioned above. Figure 1(c) shows the SBW for each subband as a percentage of all subbands.

The SFM feature did not perform well because it quantifies characteristics that are not well defined in speech signals. For example, the SFM feature measures the tonality of the subband, a characteristic that is difficult to define in the speech spectrum since its energy is distributed across many frequencies.

IV. Conclusion

Features such as the SC, SCF and SBE provide vocal source information as it relates to harmonic structure, pitch frequency and spectral energy distribution, while the entropy features quantify the spectrum in terms of voiced and unvoiced speech. The proposed features were shown to be complementary in nature and enhanced performance when used with the vocal tract transfer function (i.e. MFCC). This is mainly because the vocal tract transfer function is the most discriminating feature for speaker recognition and it greatly influences the spectral shape and harmonic structures of speech.

Experimental results show that the proposed spectral features improve the performance of MFCC based features. Based on 623 users from the TIMIT database, the combined feature set of MFCC, ΔMFCC, SC and RE achieved an identification accuracy of 99.33% (for clean speech) by incorporating information about the vocal source. This represents a 4.03% improvement over the baseline system, which only used the MFCC based features.

The good performance of spectral features for speaker recognition in this speaker identification system is very promising. These features should also produce good results if used with more sophisticated speaker recognition techniques, such as universal background model (UBM) based approaches.

References

[1] J. P. Campbell, "Speaker recognition: A tutorial." *Proc. of the IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

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

[3] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[4] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification." *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, Sept. 1999.

[5] J. M. Naik, "Speaker verification: A tutorial." *IEEE Communications Magazine*, vol. 28, no. 1, pp. 42–48, Jan. 1990.

[6] C. Che and Q. Lin, "Speaker recognition using HMM with experiments on the yoho database." in *Proc. Eurospeech*, Sept. 1995, pp. 625–628.

[7] W. Chan, T. Lee, N. Zheng, and H. Ouyang, "Use of vocal source features in speaker segmentation." in *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, May 2006, pp. 14–19.

[8] N. Zheng and P. Ching, "Using haar transformed vocal source information for automatic speaker recognition." in *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, May 2004, pp. 77–80.

[9] R. Auckenthaler, E. S. Parris, and M. J. Carey, "Improving a GMM speaker verification system by phonetic weighting." in *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 1999, pp. 313–316.

[10] J. Gonzalez-Rodriguez, S. Gruz-Llanas, and J. Ortega-Garcia, "Biometric identification through speaker verification over telephone lines." in *Proc. IEEE Int'l Carnahan Conf. on Security Technology*, Oct. 1999, pp. 238–242.

[11] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixtures models." *Digital Signal Processing*, vol. 10, pp. 19–41, Jan. 2000.

[12] K. K. Paliwal, "Spectral subband centroid features for speech recognition." in *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, vol. 2, May 1998, pp. 617–620.

[13] R. E. Yantorno, K. R. Krishnamachari, and J. M. Lovekin, "The spectral autocorrelation peak valley ratio (SAPVR) — a usable speech measure employed as a co-channel detection system." in *Proc. IEEE Int'l Workshop on Intelligent Signal Processing (WISP)*, May 2001.

[14] P. Flandrin, R. G. Baraniuk, and O. Michel, "Time-frequency complexity and information." in *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, Apr. 1994, pp. 329–332.

[15] A. Ramalingam and S. Krishnan, "Gaussian mixture modeling of short-time fourier transform features for audio fingerprinting." *IEEE Trans. on Information Forensics and Security*, vol. 1, no. 4, pp. 457–463, 2006.