International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:4, No:1, 2010

Speaker Identification Using Admissible Wavelet Packet Based Decomposition

Mangesh S. Deshpande and Raghunath S. Holambe

Abstract—Mel Frequency Cepstral Coefficient (MFCC) features are widely used as acoustic features for speech recognition as well as speaker recognition. In MFCC feature representation, the Mel frequency scale is used to get a high resolution in low frequency region, and a low resolution in high frequency region. This kind of processing is good for obtaining stable phonetic information, but not suitable for speaker features that are located in high frequency regions. The speaker individual information, which is non-uniformly distributed in the high frequencies, is equally important for speaker recognition. Based on this fact we proposed an admissible wavelet packet based filter structure for speaker identification. Multiresolution capabilities of wavelet packet transform are used to derive the new features. The proposed scheme differs from previous wavelet based works, mainly in designing the filter structure. Unlike others, the proposed filter structure does not follow Mel scale. The closed-set speaker identification experiments performed on the TIMIT database shows improved identification performance compared to other commonly used Mel scale based filter structures using wavelets.

Keywords—Speaker identification, Wavelet transform, Feature extraction, MFCC, GMM.

I. Introduction

IN state-of-the-art methods for speaker identification, the feature extraction is one of the most important aspects along with an appropriate classifier to model the speaker. The cepstral coefficients such as linear predictive cepstral coefficients (LPCC) and Mel frequency cepstral coefficients (MFCC) are the dominant features used in most of the speaker recognition systems. These traditional methods use the short time Fourier transform (STFT), which has uniform resolution over the time-frequency plane.

With conventional frequency analysis technique such as STFT, high frequency localization results in poor time resolution and high time resolution results in poor frequency localization. Speech sounds (phonemes) encompass a wide variety of characteristics, in both time and frequency domains. As an example, vowels are typically lower in frequency for longer time duration, whereas fricatives have high frequency contents for short time duration. It leads to the fact that to analyze the non-stationary signals like speech, both time and frequency resolutions are important. Therefore while extracting features; it would be useful to analyze the signal from multiresolution perspective.

Wavelets provide an alternative approach to the traditional STFT based techniques. The driving impetus behind wavelet analysis is their property of being localized in time (space) as well as scale (frequency). In recent years, multi-resolution analysis based on wavelet theory was applied in many recognition tasks [1]-[10]. In the aspect of speaker identification, many studies had developed the Mel filter-like structure to integrate the concept of Mel scale and multiresolution capabilities [4], [6], and [10]. The advantage of wavelet packet (WP) parameters presented in Mel scale is that the model of extracted features will approach humans' auditory system; moreover, the number of parameters will be decreased.

More particularly, wavelets have been used two fold. In the first approach wavelet transformation is used instead of discrete cosine transform (DCT) in the feature extraction stage [1]. Whereas in the second approach, wavelet transform is applied directly on the speech signal and either wavelet coefficients with high energy are taken as features [2], [3] or sub-band energies are used instead of Mel filter-bank sub-band energies [4], [6]. As MFCC features are the most widely used, wavelet packet bases used in [4], [5] and [6] are close approximations of the Mel-frequency division using Daubechies' orthogonal filters. In these cases, frequency warping is prescribed by the way the human auditory system functions. It follows the rule that generally the frequency resolution is fine in the lower frequency bands while it gets considerably coarser in the higher frequency bands. This information about the human auditory system has been used extensively for speech recognition. However, the needs of speaker recognition might be somewhat different, and this information may not be properly fitted for eliciting any speaker relevant characteristics.

In this paper we propose a filter structure using the admissible wavelet packet (AWP) tree that best represents the speech signal without taking into consideration any underlying knowledge of the human auditory system. The AWP tree gives the freedom to partition the low frequency band or high frequency band [4]. The admissible wavelet packet transform is combined with the Gaussian mixture model (GMM) to accomplish the closed-set speaker identification.

The remainder of the paper is organized as follows. Section II gives a brief introduction to wavelet transform. Section III explains the proposed filter bank structure for feature extraction. Matching algorithm used is discussed in Section IV. Section V explains the experimental set-up and the results obtained. Section VI draws conclusion from the results obtained.

II. Wavelet Transform

To overcome the problem of fixed resolution of STFT, the wavelet transform uses an adaptive window size, which

Mangesh S. Deshpande is with the Electronics and Telecommunication Engineering Department, SRES College of Engineering, Kopargaon, Maharashtra, India (e-mail: mangesh8374@ yahoo.com).

Raghunath S. Holambe is with the Instrumentation Engineering Department, SGGS Institute of Engineering and Technology, Nanded, Maharashtra, India (e-mail: rsholambe@sggs.ac.in).

allocates more time to the lower frequencies and less time for the higher frequencies. The decomposition of the signal into 'approximation' and 'detail' space is called the multiresolution approximation, which can be realized using a pair of low pass and high pass filters. These filters form one stage of decomposition. Wavelets are families of functions $\psi_{j,k}(t)$ generated from a single base wavelet, called the 'mother wavelet', by dilation and translation, i.e.

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k), j, k \in Z \quad (1)$$

where Z is the set of all integers, j is the dilation (scale) parameter and k is the translation parameter. Discrete wavelet transform (DWT) results in a binary tree like structure which is left recursive. It performs the recursive decomposition of the lower frequency bands in dyadic fashion thereby giving more features derived from the lower frequency bands. However speaker discrimination also requires some features from high frequency sub-bands [11], [12]. It can be achieved by wavelet packet transform (WPT). In WPT, lower as well as higher frequency bands are decomposed thereby giving a balanced binary tree structure. Each node, W_j^P in the tree is indexed by its depth j and number of subspaces p below it. The two wavelet packet orthogonal bases generated from a parent node, W_j^P are defined as,

$$\psi_{j+1}^{2p}(k) = \sum_{n=-\infty}^{\infty} h[n]\psi_j^p(k - 2^j n) \quad (2)$$

$$\psi_{j+1}^{2p}(k) = \sum_{n=-\infty}^{\infty} g[n]\psi_j^p(k - 2^j n) \quad (3)$$

where $h[n]$ is the low pass filter and $g[n]$ is the high pass filter. For a full j level wavelet packet decomposition there will be more than $2^{2^{j-1}}$ orthogonal bases in which all of them are not useful as features for recognition. Therefore the best basis selection criterion needs to be derived. However application of a best basis algorithm to the pattern recognition problem is difficult, as they are not translation invariant [13]. To overcome the above problem, AWP decomposition can be used. The AWP tree, which is in between DWT and WPT, gives the freedom to partition the low frequency band or high frequency band. By using AWP, more sub-bands in the frequency region carrying more discriminatory information can be obtained.

III. Feature Extraction

We performed a systematic and application oriented search through a reasonable set of AWP trees exploiting the tree structure as shown in fig.1, which yields the best results. The numbers in fig. 1 indicate the pass band of each filter in Hz.

The speech in the TIMIT database is sampled at 16 kHz (8 kHz bandwidth). After preprocessing (pre-emphasis, framing and windowing) the signal, wavelet packet decomposition is carried out up to three levels. This partitions the frequency axis into eight bands each of 1 kHz bandwidth. Then the first four frequency bands, 0-1 kHz, 1-2 kHz, 2-3 kHz, and 3-4 kHz are further decomposed up to 4 levels. Frequency bands 4-5 kHz and 5-6 kHz are decomposed up to 3 levels and the last two bands, 6-7 kHz and 7-8 kHz are further

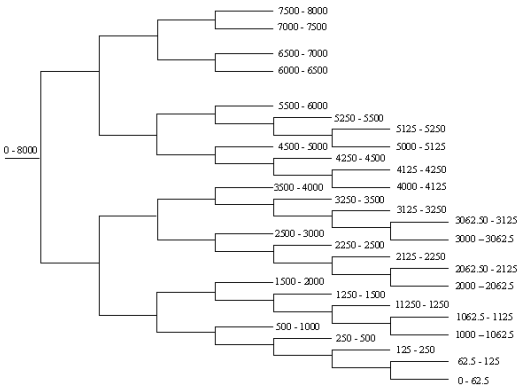

Fig. 1. Proposed filter bank structure achieved by admissible wavelet packet decomposition.

decomposed into two bands each. This gives a total of 32 frequency bands. After performing this decomposition of a 32 ms speech frame, energy in each of the frequency bands is calculated. The energy is normalized by the number of wavelet coefficients in the corresponding band. More specifically, the sub-band signal energies are computed for each frame as,

$$E_j = \frac{\sum_{i=1}^{N_j}[W_j^p f(i)]^2}{N_j}, j = 1, ...B. \quad (4)$$

where $W_j^p f(i)$ is the ith coefficient of the wavelet packet transform of a signal f at node W_j^p of the wavelet packet, B is the total number of nodes used, and N_j is the total number of coefficients consisting node j.

Finally, a logarithmic compression is performed and a DCT is applied on the logarithmic sub-band energies to reduce dimensionality:

$$F(i) = \sum_{n=1}^{B} \log_{10} E_n cos\left[\frac{i(n-1/2)}{B}\right], i = 1, ...r. \quad (5)$$

where r is the number of feature parameters. We compute only the first 24 coefficients, since we found that they represent 99.99% of the energy of the complete set of parameters. Fig. 2 shows the block schematic of proposed feature extraction procedure.

IV. Matching Algorithm

As a typical, Gaussian mixture model (GMM) [14] has been used to characterize speakers' voice in the form of probabilistic model. A GMM can be viewed as a parametric, multivariate

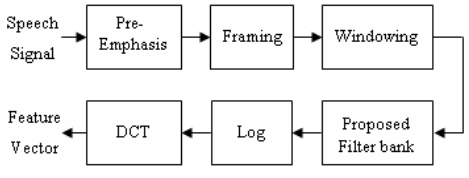

Fig. 2. Block schematic of feature extraction method.

probability distribution model that is capable of modeling arbitrary distributions and is currently the dominant method of modeling speakers in speaker recognition systems. A Gaussian mixture density is a weighted sum of M component densities and is given by the equation,

$$p(\bar{x}/\lambda) = \sum_{i=1}^{M} c_i p_i(\bar{x}), \tag{6}$$

where $\bar{x}$ is a D dimensional feature vector, $c_i, i = 1, ..., M$ are the mixture weights and $p_i(\bar{x}), i = 1, ..., M$, are the component densities of the form,

$$\frac{1}{(2\pi)^{D/2}\left|\sum_i\right|^{1/2}} exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu_i})' \left(\sum_i\right)^{-1} (\vec{x}-\vec{\mu_i})\right\} \tag{7}$$

with mean vector $\vec{\mu_i}$ and covariance matrix $\sum_i$. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} c_i = 1$. The complete Gaussian mixture density is represented by the notation,

$$\lambda = \left\{p_i, \vec{\mu_i}, \sum_i\right\}, i = 1, ..., M. \tag{8}$$

Given training utterance of a speaker, the goal of speaker model training is to estimate the parameters of the GMM, λ. The well-established method for estimating GMM parameters is the maximum likelihood (ML) estimation. For speaker identification, a group of s speakers $S = \{1, 2, 3, ..., s\}$ is represented by GMM's $\{\lambda_1, \lambda_2, \lambda_3, ..., \lambda_s\}$. The objective is to find the speaker model which has the maximum a posteriori probability for a given observation sequence $X = \{\bar{x_1}, \bar{x_2}, ..., \bar{x_T}\}$ for an utterance with T frames. The maximum a posteriori probability can be obtained by,

$$\hat{S} = argmax \sum_{t=1}^{T} logp(\vec{x_t}/\lambda_k), 1 \leq k \leq s \tag{9}$$

in which $p(\vec{x_t}/\lambda_k)$ is given in Eq. (6).

V. Experimental Set-Up and Results

The TIMIT database consists of 630 speakers, 70% male and 30% female from 8 different dialect regions in America. The speech was recorded using a high quality microphone at a sampling frequency of 16 kHz. The speech is designed to have rich phonetic contents. It consists of 2 dialect sentences (SA), 450 phonetically compact sentences (SX) and 1890 phonetically diverse sentences (SI). The database is divided into train and test sets by its creators, but this division is useful for speech recognition task. For speaker identification task it is much easier to use the whole database.

The evaluation of the proposed feature extraction method was carried out by closed-set speaker identification experiments on complete 630 speakers of TIMIT database. We have also evaluated the performance for different population sizes as 200, 400 and 630. The speaker models were trained using eight sentences, five SX and three SI (approximately 24 seconds). The two SA sentences per speaker were used separately (a total of 1260 tests of 3 s each) for testing and average identification results were noted. The speech signal was pre-emphasized using a pre-emphasis filter with impulse response $h[n] = \{1, -0.97\}$. The pre-emphasized signal was divided into frames of 32 ms with 50% overlap and Hamming window was applied on each frame. Then proposed features were obtained using 6th order Daubechies' orthogonal filter as discussed in section III. Finally diagonal covariance matrices were used to model the speakers with 32 mixtures GMMs.

General speaker identification performance by various researchers on TIMIT database (for all 630 speakers) with 32 mixtures GMM is depicted in Table I. The researchers in [6], [9], and [13] obtained various dimension feature vectors and evaluated the performance with same training and testing conditions, i.e. eight sentences for training purpose and two sentences for testing. In [6], Ruhi Sarikaya considered a 24 sub-band wavelet packet tree that approximates the Mel scale frequency division, sub-band based cepstral parameters as well as wavelet packet parameter based features. In [9], S.-Y. Lung used the wavelet packet feature selection based on neuro-fuzzy evaluation index for speaker identification. In [15], K. Markov and S. Nakagava obtained cepstrum as well as delta cepstrum coefficients. Due to a better representation of the speaker specific variations in speech signal, the proposed features demonstrated a superior performance than other wavelet packet based speech features and Mel frequency cepstral coefficients.

TABLE I
Comparison of speaker identification performance by various researchers on TIMIT database

System	Type of features	Dimensions	Speaker identification rate (%)
R Sarikaya [6]	MFCC	19	94.8
R Sarikaya [6]	SBC[a]	24	96
R Sarikaya [6]	WPP[b]	24	97.3
S. Y. Lung [9]	Wavelet packet	16	69
K. Markov [15]	Cepstrum + delta cepstrum	20(10+10)	94.3
K. Markov [15]	Cepstrum	10	94.3
Proposed	AWP	24	98

[a] subband based cepstral parameters, [b] wavelet packet parameters.

To compare the performance of the proposed features with the most commonly used MFCC features, we have implemented a 32 triangular shape linearly spaced filter bank using Mel scale warping and 24 dimension MFCC feature vectors were considered. Further we have implemented a Mel filter like structure using AWP tree as proposed by Farooq and Datta (F-D) in [4]. Even though the primary aim in [4] was the phoneme recognition and not speaker identification, the frequency band spacing is similar to Mel scale. Table II shows the speaker identification performance of proposed features, MFCC features and F-D features for different population sizes. It shows that F-D feature performance degrades as population size increases. The proposed features performance is superior compared to F-D wavelet based features and it is equally good as that of the most widely used MFCC features even for large population.

TABLE II
SPEAKER IDENTIFICATION PERFORMANCE.

Population	24 Dimensions		
	F-D Features (%)	MFCC(%)	Proposed Features(%)
200	97.75	97.75	99.75
400	96.25	98.25	98.87
630	95.95	98.33	98

VI. CONCLUSION

The state-of-the-art filter bank structures for feature extraction are based on frequency warping by the way the human auditory system functions (i.e. Mel scale). In Mel scale, generally the frequency resolution is fine in the lower frequency bands (approximately up to 1 kHz) and gets considerably coarser in the higher frequency bands. This structure has worked very well for speech recognition but the need of speaker recognition might be somewhat different. Considering this we have proposed a filter structure that best represent the speech signal without taking into consideration any underlying knowledge of the human auditory system. We have experimentally shown that the proposed filter structure which is different than the Mel scale gives better results compared to the widely used MFCC as well as Mel filter bank implementation using wavelet packet transform. The reason for getting better performance is that the proposed filter structure is fine tuned to some of the frequency bands which are more important for speaker discrimination. This study shows that the need of filter structure to extract speaker specific features is somewhat different than the commonly used filter structure based on Mel scale warping.

REFERENCES

[1] Z. Tufekci and J.N. Gowdy, "Feature extraction using discrete wavelet transform for speech recognition," in Proc. IEEE Southeastcon, USA, 2000, pp. 116-123.
[2] O. Farooq and S. Datta, "Phoneme recognition using wavelet based features," Information Sciences, vol. 150, no.1-2, Mar. 2003, pp. 5-15.
[3] O. Farooq and S. Datta, "Wavelet based robust sub-band features for phoneme recognition," in IEE Proc. Image signal process., vol. 151, no. 3, June 2004, pp. 187-192.
[4] O. Farooq and S. Datta, "Mel filter like admissible wavelet packet structure for speech recognition," IEEE Signal Process. Lett., vol. 8, no. 7, pp. 196-199, July 2001.
[5] R. Sarikaya and H. L. Hansen, "High resolution speech feature parameterization for monophone-based stressed speech recognition," IEEE Signal Process. Lett., vol. 7, no. 7, pp. 182-185, July 2000.
[6] R. Sarikaya, B. L. Pellom and H. L. Hansen, "Wavelet packet transform features with application to speaker identification," in Proc. IEEE Nordic Signal processing Symposium, Visgo, Denmark, 1998, pp. 81-84.
[7] C. T. Hsieh, E. Lai and Y. C. Wang, "Robust speech features based on wavelet transform with application to speaker identification," in IEE Proc. Image signal process., vol. 149, no. 2, April 2002, pp. 108-114.
[8] S.-Y. Lung, "Further reduced form of wavelet feature for text independent speaker recognition," Pattern recognition, vol. 37, 2004, pp. 1565-1566.
[9] S.-Y. Lung, "Wavelet feature selection based neural networks with application to the text independent speaker recognition," Pattern recognition, vol. 39, 2006, pp. 1518-1521.
[10] H. M. Torres and H. L. Rufiner, "Automatic speaker identification by means of Mel cepstrum, wavelets and wavelets packets," in Proc. IEEE international conference, EMBS, Chicago, IL, 2002, pp. 978-981.
[11] Xugang Lu and Jianwu Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," Speech communication, vol. 50, 2008, pp. 312-322.
[12] S. Hayakawa and F. Itakura, "Text-dependent speaker recognition using the information in the higher frequency band," in Proc. IEEE international conference on Acoustic Speech and signal Processing, ICASSP, Adelaide, Australia, 1994, pp. 137-140.
[13] S. Mallat, A wavelet tour of signal processing. Second ed., Academic Press, 1998.
[14] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech and Audio Processing, vol. 3, no. 1, pp. 72-83, Jan. 1995.
[15] K. Markov and S. Nakagawa, "Frame level likelihood normalization for text-independent speaker identification using Gaussian mixture models," in Proc. IEEE ICSLP, 1996, pp. 1764-1767.

Mangesh S. Deshpande received the B. E. degree from North Maharashtra University, Jalgaon, India in 1996 and Masters Degree from Shivaji University Kolhapur in 2002. Presently he is a research scholar in the Department of Instrumentation Engineering, SGGS Institute of Engineering and Technology, Nanded (India). His main research interests are in Digital Signal Processing, Speech processing and embedded systems.

Raghunath S. Holambe received the Ph. D. degree from Indian Institute of Technology, Kharagpur in India and presently he is a professor in the Department of Instrumentation Engineering, SGGS Institute of Engineering and Technology, Nanded (India). The areas of his research interest are Digital Signal Processing, Image Processing, Applications of Wavelet Transform, Biometrics, and real time signal processing using DSP Processors.