

New Filter Structure based on Admissible Wavelet Packet Transform for Text-Independent Speaker Identification

Mangesh S. Deshpande and Raghunath S. Holambe

Department of Instrumentation Engineering, SGGS Institute of Engineering and Technology, Nanded, India
mangesh8374@yahoo.com, rsholambe @sggs.ac.in

Abstract— Identical acoustic features like Mel frequency cepstral Coefficients (MFCC) and Linear predictive cepstral coefficients (LPCC) are being widely used for different tasks like speech recognition and speaker recognition, whereas the requirement of speaker recognition is different than that of speech 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. Further MFCC uses short time Fourier transform (STFT), which has fixed time-frequency resolution. Considering above facts, in this paper we have proposed a new filter structure based on admissible wavelet packet transform for text-independent speaker identification. Multiresolution capabilities of wavelet packet transform are used to derive the new features. The performance of the proposed features is evaluated using the most commonly used Gaussian mixture model (GMM) as well as the continuous density hidden Markov model (CDHMM) classifiers. Improved speaker identification rate is obtained using the proposed features compared to the MFCC and other Wavelet transform based features. Further the results show that CDHMM works better than the GMM for small number of mixture densities. Identification accuracy of 99.76% is achieved by conducting the experiments on TIMIT database.

Index Terms— Speaker identification, AWP Transform, MFCC, GMM, HMM

I. INTRODUCTION

The speech signal consists of several levels of information. The information can be mainly divided into two categories. First, the speech signal conveys the words or message being spoken and secondly, the signal also conveys information about the identity of the talker. While the area of the speech recognition is concerned with extracting the underlying linguistic message in an utterance, the area of speaker recognition is concerned with extracting the identity of the person speaking the utterance. Speaker identification is the task of determining who is speaking from a set of known speakers. Furthermore, the speech can be constrained to be a known phrase (text-dependent) or totally unconstrained (text-independent).

Linear Prediction Cepstral Coefficient (LPCC) and Mel Frequency Cepstral Coefficient (MFCC) features are widely used as acoustic features for speech recognition.

These features are also the dominant features used in most of the state-of-the-art speaker recognition systems [1]. The LPC features can well model the vocal tract by using an all-pole model which reflects the main vocal tract resonances in acoustic spectra [2]. While MFCC features take the auditory nonlinear frequency resolution mechanism into consideration. However, the purpose of speech recognition is quite different from that of speaker recognition, the former task needs to emphasize linguistic information and suppress speaker individual information, while the later task needs more speaker individual information. This contradiction suggests that LPCC and MFCC may not meet both speech and speaker recognition requirements.

The LPC feature emphasizes the formant structure that concerns major individual differences of the speakers, while some significant details of individuals such as the nasal, piriform fossa and other side branches are ignored. 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-specific information caused by different articulatory speech organs is distributed non-uniformly in high frequency bands also [3]. The information of the glottis is mainly encoded in a low frequency band (between 100 Hz and 400 Hz), the information of the piriform fossa in a high frequency band (between 4 kHz and 5 kHz), the constriction of the consonants would be another factor in the higher frequency region around 7.5 kHz [4]. This kind of non-uniform distribution of speaker information in frequency bands was also confirmed in [3]. In contrast, most speech phonemic discriminative information, such as the first three formants, is encoded in a low and middle frequency region from 200 Hz to 3 kHz, which is very important for speech recognition [2].

Traditional feature extraction methods focus on the large spectral peaks caused by the movements of the vocal tract and emphasize the lower frequency bands. Furthermore the traditional methods use the short time Fourier transform (STFT), which has uniform resolution over the time-frequency plane and it pays more attention to the static features while neglect the speaker's dynamic characteristics. However, many experiments have shown

that the dynamic characteristics are very important features of the speaker. Wavelets provide an alternative approach to the traditional Fourier transform based techniques. The driving impetus behind wavelet analysis is their property of being localized in time (space) as well as scale (frequency). Many researchers have used wavelet and wavelet packet transform for feature extraction [5- 9].

In this paper, we propose a filter bank implemented using admissible wavelet packet (AWP) transform, which gives the freedom to partition the low frequency band or high frequency band [10], [11]. While proposing the filter bank, we have not considered the speech perception mechanism. Rather we have analyzed the signal in different frequency bands using wavelet transform. Finally the performance of proposed features is evaluated using Gaussian mixture model (GMM) as well as hidden Markov model (HMM). It is found that the proposed filter structure based on AWP tree improves the speaker identification performance.

Remainder of the paper is organized as follows. Section II describes the multi-resolution filter bank approach. Section III describes the new feature extraction method. Section IV provides a brief description of GMM and HMM. The experiments performed are discussed in section V. Finally, section VI outlines the conclusions.

II. MULTIREOLUTION ANALYSIS

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

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 may require some features from high frequency sub-bands. 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. 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 [11]. 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

The MFCC has been the most widely used features for speaker identification. The Mel scale is a mapping between the real frequency scale (Hz) and the perceived frequency scale (Mel). This mapping is virtually linear below 1 kHz and logarithmic above. The drawback of MFCC is that it uses STFT, which has fixed time-frequency resolution. In addition to this it assumes that the signal is stationary during the window period, actually which may not be strictly stationary. Further, MFCC does not take into consideration the contribution of the vocal tract elements like piriform fossa, which results into high frequency components. In order to overcome these limitations of MFCC, we proposed a new filter structure using AWP tree.

In the experiments performed, the frequency analysis of each 1 kHz frequency band of the speech signal up to 8 kHz bandwidth was done using DWT. Then energy of each decomposed level was calculated. The decomposition of each 1 kHz frequency band was carried out still noticeable energy in the last decomposed level was obtained. The detailed process of extracting the features is discussed below.

After huge experimentation it was found that the tree given in Fig.1, gives the best overall result among a reasonable set of AWP trees. Initially, a three level wavelet packet decomposition was carried out. This partitioned the frequency axis into eight bands each of 1 kHz (the speech in the TIMIT database is sampled at 16 kHz, giving an 8 kHz bandwidth signal). Then the first four frequency bands, 0-1 kHz, 1-2 kHz, 2-3 kHz, and 3-4 kHz were further decomposed up to 4 levels. Frequency bands 4-5 kHz and 5-6 kHz were decomposed up to 3 levels and the last two bands, 6-7 kHz and 7-8 kHz were further decomposed into two bands each and obtained a total of 32 frequency bands.

After performing this decomposition of a 32 ms speech frame, energy in each of the frequency bands was calculated. The energy was normalized by the number of

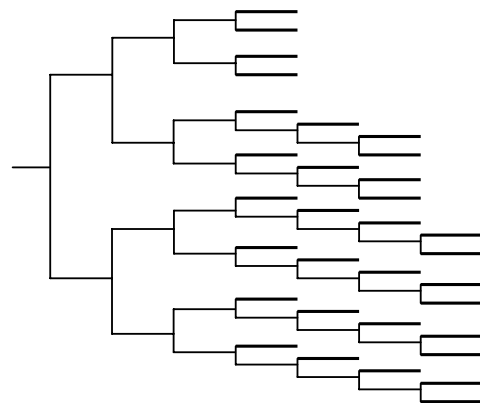


Figure 1. Proposed filter structure achieved using AWP tree.

wavelet coefficients in the corresponding band, thereby giving an average energy per frame in each band as,

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

where $W_j^P f(i)$ is the i^{th} 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 was performed and a DCT was applied on the logarithmic sub-band energies to reduce dimensionality:

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

where r is the number of feature parameters. Only the first 24 coefficients were considered to construct a feature vector.

IV. SPEAKER MODELLING USING GMM AND HMM

Gaussian mixture model is the classifier commonly used in speaker identification/ verification [1], [14] and [15]. This classifier is able to approximate the distribution of the acoustic classes representing broad phonetic events occurring in speech production (e.g., during the production of vowels, nasals, fricatives etc.).

A Gaussian mixture density is a weighted sum of M component densities and is given by the equation,

$$p(\vec{x}/\lambda) = \sum_{i=1}^M c_i p_i(\vec{x}), \quad (3)$$

where \vec{x} is a D dimensional feature vector, $c_i, i=1, \dots, M$ are the mixture weights and $p_i(\vec{x}), i=1, \dots, M$, are the component densities. The complete Gaussian mixture density is represented by the notation,

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\}, i=1, \dots, M. \quad (4)$$

In recent studies, HMM has become the most popular statistical tool for speech as well as text-dependent speaker identification. Two main variations of HMM have been widely used: discrete HMM (DHMM) and continuous density HMM (CDHMM). The main problem of DHMM is the loss of information about the input signal during the vector quantization process. CDHMM avoids this problem by using probability density functions. Thus, CDHMM modeling seems to be a more flexible and complete tool for speaker modeling [12].

In text-dependent speaker recognition, a left-to-right HMM can be used because the phoneme sequence of the input speech is predetermined. However in text-

independent speaker recognition, it is difficult to know the phoneme sequence beforehand. It is also true that a Markov chain should be able to revisit the earlier states, because the states of the HMM reflect the vocal configuration of a speaker and the variations of vocal configuration may repeat in pronunciation. Therefore an ergodic model which allows transitions to any other states has been assumed to be effective for text-independent speaker identification.

A parameter set of HMM is given by $\lambda_{HMM} = (A, B, \pi)$, where A, B and π denote a set of state transition probability, a set of output probability density functions, and a set of initial state probabilities, respectively. For an ergodic HMM, every state can be reached from every other state. The probability density function of certain observations o being in state j has the following general form:

$$b_j(o) = c_{jk} \aleph(o, \mu_{jk}, U_{jk}), 1 \leq j \leq Q \quad (5)$$

where $\aleph(o, \mu_{jk}, U_{jk})$ and c_{jk} respectively represent the multi-dimensional Gaussian PDF and the weight for the k^{th} mixture component of state j [13].

To evaluate the performance using GMM, it is assumed that the feature vectors are independent of each others, whereas such assumption is not required in case of HMM. Truly speaking the feature vectors can not be independent of each others.

V. EXPERIMENTS AND RESULTS

A. Database Description

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). For speaker identification task it is much easier to use the entire database. In the following experiments, all 630 speakers (438 males and 192 females) from the TIMIT database are used. 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 seconds each) for testing and average identification results are noted.

B. Performance Evaluation

The speech signal was pre-emphasized using a pre-emphasis filter, $H(z) = 1 - 0.97z^{-1}$ then windowed using 32ms Hamming window with a 16 ms frame shift. Proposed features were obtained using 6th order Daubechies' orthogonal filter. To compare the performance of the proposed features with others, the well known MFCC features and AWP based energy

features proposed by Farooq and Datta (F-D) in [10] were estimated. To obtain these features the same pre-processing was used, i.e. the signal was pre-emphasized then windowed using 32ms Hamming window with a 16 ms frame shift. The features proposed by Farooq and Datta are considered for comparison (even though the primary aim in [10] was the phoneme recognition and not speaker identification) because the frequency band spacing in F-D features is similar to Mel scale (i.e. high resolution at low frequencies and low resolution at high frequencies) and they have also used the 6th order Daubechies' orthogonal filter.

Both GMM as well as HMM classifiers were used to evaluate the performance. Diagonal covariance matrices were considered to model the speakers with 32 mixtures GMMs. It was shown in [16] that the identification rates using an ergodic CDHMM are strongly correlated with the number of states, number of mixtures per state and the amount of data used for training. A 2 state, 4 mixture ergodic CDHMM were considered in the following experiments. Diagonal covariance matrices were considered to train the CDHMM using Baum-Welch algorithm.

Generally, as the speaker population increases the identification rate goes down. To evaluate the effect of population size on the proposed features, experiments were conducted using three different population sizes as 200, 400 and 630 speakers.

Fig. 2 shows the speaker identification performance as a function of population size for all 3 different types of features using GMM as a classifier. It shows that the performance degrades as the population increases from 200 to 630 speakers. Further the proposed AWP features show better performance than MFCC as well as F-D features for population size as 200 and 400. For 630 speakers' population the performance of AWP features is same as MFCC features and better than F-D features. Fig. 3 shows the similar plots using CDHMM as a classifier. It shows that the performance of AWP features is same as that of MFCC features for the population of 200 and 400 speakers and it is slightly better than MFCC features for 630 speakers population. The identification rate is 99.76 % at the population of 630 speakers. Fig. 2 and 3 shows that the AWP feature performance is comparable to that of well known and most widely used MFCC features. It also emphasizes the need to obtain a different type of filter structure (other than perceptual based) for identifying speakers. Using both the classifiers, the proposed features work better than the F-D features and equally good as that of the MFCC features. The plot in Fig. 4 compares the speaker identification performance obtained using GMM and CDHMM for different types of features. It shows that for all 3 types of features CDHMM performance is slightly better than GMM.

Further the performance is evaluated by reducing the feature vector dimensions and using CDHMM as a classifier. Fig. 5 shows the speaker identification performance as a function of feature vector dimensions for three different types of features. With 16 dimension feature vectors, the identification rate achieved is

99.45%, which is slightly greater than the identification rate achieved with MFCC features (99.12%) and the features proposed by Farooq and Datta (98.34%). It also shows that the identification performance increases with increase in the proposed AWP feature vector dimensions. With 24 dimension feature vectors, the identification rate achieved is 99.76%, where as with 16 dimensions it is

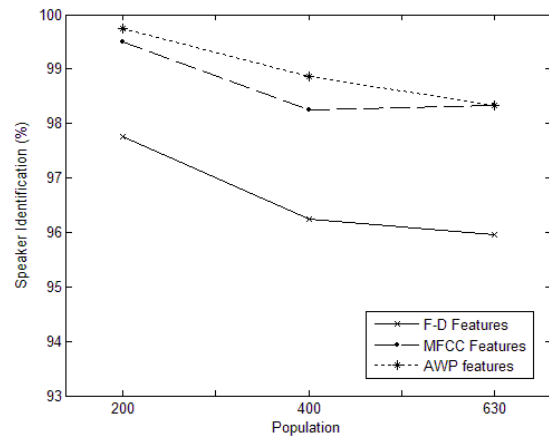


Figure 2. Speaker identification rate in percent as a function of population size for F-D features, MFCC features and AWP features using GMM as a classifier.

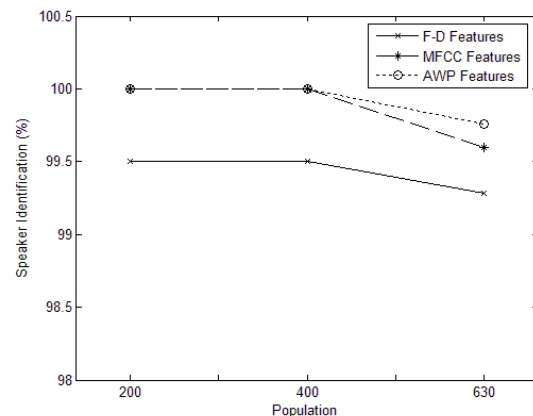


Figure 3. Speaker identification rate in percent as a function of population size for F-D features, MFCC features and AWP features using CDHMM as a classifier.

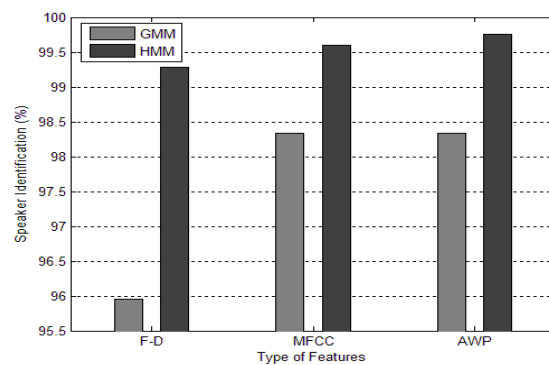


Figure 4. Speaker identification performance obtained using GMM and HMM for different types of features.

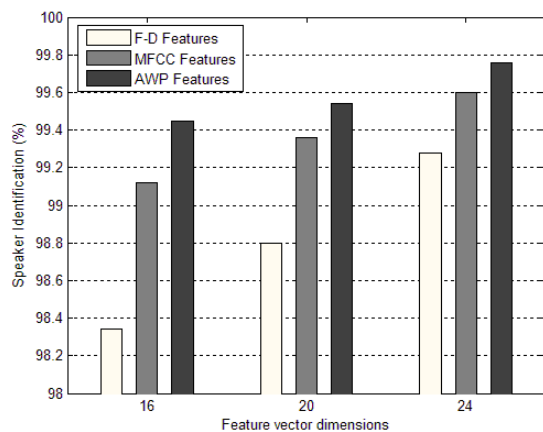


Figure 5. Speaker identification performance as a function of feature vector dimensions.

99.45%. This small improvement in identification rate is important for large population of speakers.

VI. CONCLUSION

New, AWP based features appropriate for speaker identification has been proposed. The proposed filter structure is fine tuned to emphasize some of the spectral bands important for speaker identification. A comparative experimental evaluation of the proposed features, performed on a well-known TIMIT database, proved the practical significance of the approach. 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. Mel scale is better for speech recognition where majority of the information is concentrated in the low frequency bands. For identifying speakers high frequency bands are equally important. It is further required to investigate how the information from the high frequency bands should be combined with low frequency bands to use it effectively for speaker identification. Further, to evaluate the performance of the proposed features we have used both the GMM and HMM classifiers. It is also shown that the continuous density ergodic HMM can perform better than the GMM

REFERENCES

- [1] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91-108, 1995.
- [2] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall of India, 1993.
- [3] S. Hayakawa and F. Itakura, "Text-dependent speaker recognition using the information in the higher frequency band" in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1994*, Adelaide, Australia, 1994, pp. 137-140.
- [4] Xugang Lu and Jianwu Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech Communication*, vol. 50, pp. 312-322, 2008.
- [5] Y. D. Zhang and F. Y. Sun, "A methodology based on wavelet packet for speaker transform recognition" in *Proceedings of International conference on Wavelet Analysis and Pattern Recognition*, Beijing, China, Vol. 2, pp. 767-771, Nov. 2007.
- [6] P. T. Nghia, P. V. Binh, N. H. Thai, N. T. Ha and P. Kumsawat, "A Robust Wavelet-Based Text-Independent Speaker Identification" in *International conference on Computational Intelligence and Multimedia Applications*, Vol. 2, Dec. 2007, pp. 219-223.
- [7] Y. L. Shung, "Wavelet feature selection based neural networks with application to the text independent speaker recognition," *Pattern recognition*, vol. 39, pp. 1518-1521, 2006.
- [8] C. T. Hsieh, E. Lai and Y. C. Wang, "Robust speech features based on wavelet transform with application to speaker identification," *IEE Proc. Image signal process.* April 2002, 149, (2), pp. 108-114.
- [9] H. M. Torres and H. L. Rufiner, "Automatic speaker identification by means of Mel cepstrum, wavelets and wavelets packets" in *Proceedings of EMBS, International Conference*, Chicago IL, July 2000, pp. 978-981.
- [10] O. Farooq and S. Datta, "Mel filter like admissible wavelet packet structure for speech recognition," *IEEE Signal Process. Letters*, vol. 8, no. 7, pp. 196-199, July 2001.
- [11] S. Mallat, *A wavelet tour of signal processing*, second edition, Academic Press, 1998.
- [12] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/ continuous HMMs," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 3, pp. 456-459, July 1994.
- [13] L. R. Rabiner, "A Tutorial on Hidden Markov Models and selected applications in speech recognition," *IEEE*, vol. 77, no. 2, pp. 257-286, Feb. 1989.
- [14] H. A. Murthy, F. Beaufays, L. P. Heck, M. Weintraub, "Robust text-independent speaker identification over telephone channels," *IEEE Trans. Speech and Audio Processing*, Vol. 7, no. 2, pp. 554-568, 1999.
- [15] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech and Audio Processing*, Vol. 2, no. 4, pp. 639-643, 1994.
- [16] M. S. Deshpande, R. S. Holambe, "Text-independent speaker identification using hidden Markov models," in *Proceedings of First International conference on Emerging trends in Engineering and technology, IEEE Computer Society*, Nagpur, India, July 2008, pp. 641-644.