WAVELET PACKET TRANSFORM FEATURES WITH APPLICATION TO SPEAKER IDENTIFICATION

Ruhi Sarikaya, Bryan L. Pellom and John H. L. Hansen

Robust Speech Processing Laboratory
Duke University, Box 90291, Durham, NC 27708-0291
http://www.ee.duke.edu/Research/Speech ruhi@ee.duke.edu bp@ee.duke.edu jhlh@ee.duke

ABSTRACT

This study proposes a new set of feature parameters based on wavelet packet transform analysis of the speech signal. The new speech features are named subband based cepstral parameters (SBC) and wavelet packet parameters (WPP). The ability of each parameter set to capture speaker identity conveyed in the speech signal is compared to the widely used Mel-frequency cepstral coefficents (MFCC). The proposed parametrization methods are shown to achieve **48%** and **67%** reduction in relative error over MFCC for 630 and 168 speakers, respectively using the TIMIT (downsampled to 8 kHz) database.

1. INTRODUCTION

The problem of speaker identification can be divided into two major stages. The first is feature extraction and the second is classification of speakers based on the extracted features. Although these two components may appear to be independent, they are highly coupled. To be effective, the features should be capable of separating the speakers from each other in its space, whereas the classifier should be tuned to differentiate the different classes in a given feature space.

Recently, extensive research has been conducted on the problem of speaker identification (ID). The identification results can be as high as 99.5%[3] for the noise free TIMIT (sampled at 16 kHz) database. However, for the same data set transmitted over telephone channels, the identification accuracy is reduced to 60% [3]. In fact, in practice, the major applications for speaker identification such as information retrieval from large speech databases, automatic sorting of voice mail messages, telephone based financial transactions and multimedia applications usually require speech to be transmitted over a noisy channel. Hence, one of the biggest obstacles in operating speech processing systems in real environments is the presence of background and convolutional channel noise. One way to improve the performance of speech processing systems is to formulate parameters which are less sensitive to such environments.

Although the MFCC parameters achieve high ID rates over 16 kHz sampled TIMIT, the results are not as good for the telephone channel version. Furthermore, the dramatic degradation in ID rates over telephone speech can be partly attributed to the MFCCs which are not immune to noise. These reasons motivated us to formulate two new features entitled subband based cepstral parameters (SBC), and wavelet packet transform parameters (WPP), which allow embedded denoising or enhancement in the feature extraction stage rather than filtering the speech for improved speaker identification.

2. NEW FEATURE EXTRACTION

2.1 Subband Decomposition via Wavelet Packets

A detailed discussion of wavelet analysis is beyond the scope of this paper, and we therefore refer interested readers to a more complete discussion presented in [9]. In continous time, the Wavelet Transform is defined as the inner product of a signal $x(t)$ with a collection of wavelet functions $\psi_{a,b}(t)$ in which the wavelet functions are scaled (by a) and translated (by b) versions of the prototype wavelet: $\psi(t)$.

$$\psi_{a,b}(t) = \psi\left(\frac{t-b}{a}\right) \tag{1}$$

$$W_\psi x(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t-b}{a}\right) dt \tag{2}$$

Discrete time implementation of wavelets and wavelet packets are based on the iteration of two channel filterbanks which are subject to certain constraints, such as low pass and/or high pass branches on each level followed by a subsampling-by-two unit. Unlike the wavelet transform which is obtained by iterating on the low pass branch, the filterbank tree can be iterated on either branch at any level, resulting in a tree structured filterbank which we call a wavelet packet filterbank tree. The resultant transform creates a division of the frequency domain that represents the signal optimally with respect to the applied metric while allowing perfect reconstruction of the original signal. Because of the nature of the analysis in the frequency domain, it is also called subband decomposition where subbands are determined by a wavelet packet filterbank tree.

In this paper, we consider a 24 subband wavelet packet tree which approximates the Mel-scale frequency division as shown in Fig. 1. The wavelet packet tree is constructed by cascading the basic two channel filterbank into various levels.

2.2 Wavelet Packet Transform Based Feature Extraction Procedure

Here, speech is assumed to be sampled at 8 kHz. A frame size of 24 msec with a 10 msec skip rate is used to derive

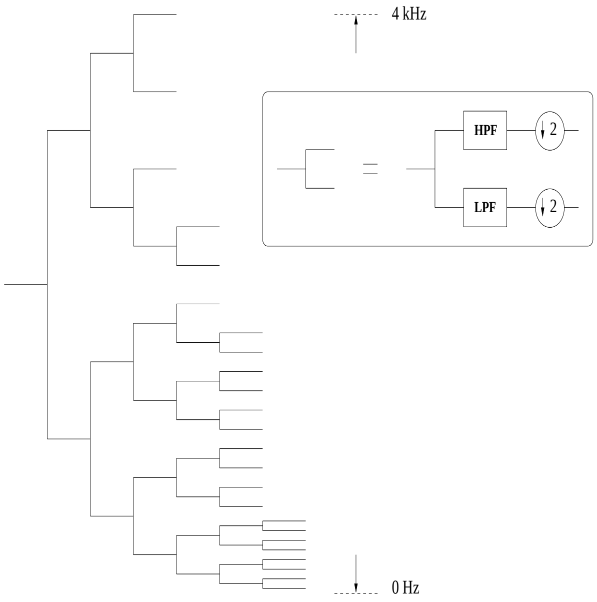

Figure 1: 24-subband wavelet packet tree.

Filters	1	2	3	4	5	6
MFCC	28	89	154	224	300	383
WPP	31	94	156	219	281	344
Filters	7	8	9	10	11	12
MFCC	472	569	674	787	910	1043
WPP	406	469	563	688	813	938
Filters	13	14	15	16	17	18
MFCC	1187	1343	1512	1694	1892	2106
WPP	1063	1188	1313	1438	1563	1688
Filters	19	20	21	22	23	24
MFCC	2338	2589	2860	3154	3472	3817
WPP	1875	2125	2375	2750	3250	3750

Table 1: Comparison of center frequency (Hz) of 24 uniformly spaced (in mel-scale) MFCC filterbanks and WPP subbands

the SBC and WPP features, whereas a 20 msec frame with the same skip rate is used to derive the MFCCs. We have used the same configuration proposed in [3] for MFCC. The reason for using a 24 msec window for WPP is due to our analysis which requires the total number of samples in the frame to be divisible by 64 while having comparable frame sizes for all parameters under consideration. Next, the speech frame is Hamming windowed and preemphesized. We found after some experimentation that the tree given in Fig. 1, gave the best overall result among a reasonable set of wavelet packet trees. The resulting subband divisions finely emphasize frequencies between 0-500Hz which normally contain large portions of the signal energy. Equal partitions are used between 500-1750Hz where each subband width is 125 Hz. The remaining frequency axis is virtually the same as a Mel-scale division. Therefore, the proposed tree assigns more subbands between low to mid frequencies while keeping roughly a log-like distribution of the subbands across frequency. In Table 1, the center frequencies of wavelet packet subbands, and the 24 filters distributed uniformly in Mel-scale are given for comparison purposes. The wavelet packet transform is computed for the given wavelet tree, which results in a sequence of subband signals or equivalently the wavelet packet transform coefficients, at the leaves of the tree. In effect, each of these subband signals contains only restricted frequency information due to inherent bandpass filtering. The complete block diagram for computation of SBC and Wavelet Packet Parameters (WPP) are given in Fig. 2. The energy of the sub-signals for each subband is computed and then scaled by the number of transform coefficients in that subband. The subband signal energies are computed for each frame as,

$$S_i = \frac{\sum_{m\epsilon i}[(W_\psi x)(i),m)]^2}{N_i} \qquad (3)$$

$W_\psi x$: wavelet packet transform of signal x,
i : subband frequency index ($i = 1, 2...L$),
N_i : number of coefficients in the i^{th} subband.

The analysis steps up to this point are common to the derivation of both SBC and WPP. Since we use the orthogonal filters corresponding to Daubechies's orthogonal wavelets [6] in the wavelet packet transform, energy is preserved in the transformation.

2.3 Subband based Cepstral Parameters (SBC)

As in MFCCs, the derivation of parameters is performed in two stages. The first stage is the computation filterbank energies and the second stage would be the decorrelation of the log filterbank energies with a DCT to obtain the MFCC. The derivation of the SBC parameters follows the same process except that the filterbank energies are derived using the wavelet packet transform rather than the short-time Fourier transform. It will be shown that these features outperform MFCCs. We attribute this to the computation of subband signals with smooth filters. In the computation of the MFCC, the spectrum of the signal is filtered with either triangular or raised cosine type filters to obtain filterbank energies. However, the filterbank between 0-1 kHz are uniformly arranged with a 100 Hz bandwidth. For a frame size of 20 msec, the frequency resolution is 31 Hz. For MFCC, we partition the spectrum between 0-1 kHz to obtain 10 uniformly spaced frequency bands. Energy estimates within each of these bands are obtained by multiplying the spectrum with a window of 3 sample points only. This filter is very coarse regardless of whether it is triangular or a raised cosine. On the other hand, the effect of filtering as a result of tracing through the low-pass/high-pass branches of the wavelet packet tree, is much smoother due to the balance in time-frequency representation. We believe that this will contribute to improved speech/speaker characterization over MFCC. These parameters have been shown to be effective for speech recognition in car noise[10] and for classification of stressed speech [1]. SBC parameters are derived from subband energies by applying the Discrete Cosine Transformation transformation:

$$SBC(n) = \sum_{i=1}^{L} \log S_i \cos\left(\frac{n(i-0.5)}{L}\pi\right), \quad n = 1, ...n' \qquad (4)$$

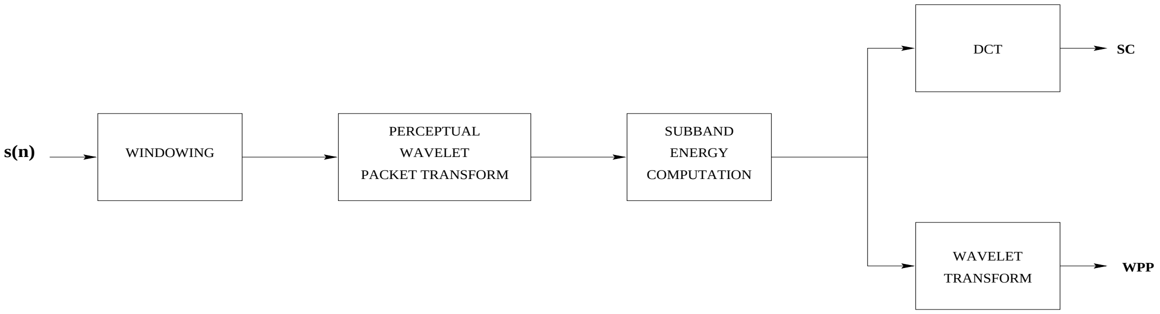
s(n)
WINDOWING
PERCEPTUAL WAVELET PACKET TRANSFORM
SUBBAND ENERGY COMPUTATION
DCT
SC
WAVELET TRANSFORM
WPP

Figure 2: Block diagram for Wavelet Packet Transform based feature extraction procudure.

where $n^{'}$ is the number of SBC parameters and L is the total number of frequency bands. Because of the similarity to root-cepstral [8] analysis, they are termed as subband based cepstral parameters.

2.4 Wavelet Packet Parameters (WPP)

Essentially, the DCT step in the calculation of the MFCC features, decorrelates the filterbank energies. It has been shown [7] however that the wavelet transform is a better decorrelator in coding applications. We know that the Gaussian mixture densities typically used to model speakers for ID, have diagonal covariances assuming that the components of the feature vector are independent of each other. The degree to which this assumption is satisfied partly depends on the transform which makes the decorrelation. We hypothesize that using a wavelet transform instead of the DCT may satisfy the assumption better, which in turn should lead to improved speaker identification. In order to verify this hypothesis, we consider the following scenario which is based on the idea that if features are well decorrelated, the energy contributed by the off-diagonal terms in the autocovariance matrix should be smaller. We have derived the same number of features from a fixed set of sentences each of which is approximately 3 seconds in duration. The same frame size and skip rate are used to align both features for a fair comparison. Furthermore, the feature energy for each frame is normalized to 1.0 to eliminate the differences resulting from a different scaling of the parameters. The autocovariance matrix $R(n,m)$ is formed over time for each frame and the absolute values of the off-diagonal terms are summed to obtain an accumulated measure. The larger the off-diagonal elements the higher the correlation between the components of the feature vector. Here, we define the term η based on the autocovariance of the feature vectors, which denotes the total correlation for both wavelet and discrete cosine transforms.

$$R = E[XX^T] \tag{5}$$

$$\eta_{wt} = \sum_{t=1}^{t=T} \sum_{m=1}^{N} \sum_{\substack{n=1 \\ n \neq m}}^{N} |R_{wt}(n,m)| \tag{6}$$

where η_{wt} denotes the the total correlation between components of the feature vector over the entire speech data. $R(n,m)$ is the autocovariance matrix of the feature vector X, N is the feature length and T is total number of frames.

The total correlation term for the DCT used for SBC, η_{dct}, is calculated in a similar manner. For all speech evaluated, we consistently observed that $\eta_{wt} < \eta_{dct}$. This result confirms our hypothesis that the wavelet transform decorrelates the subband log energies better than a DCT.

WPPs are derived by taking the wavelet transform of the log-subband energies. The wavelet parameters a and b in Eq. 4 are continuous. For discrete implementation one often samples $\psi_{ab}(t)$ using $a = {a_0}^m$ and $b = nb_0{a_0}^m$,

$$\psi_{mn}(t) = {a_0}^{-m/2}\psi({a_0}^{-m}t - nb_0) \tag{7}$$

where $m, n \in Z$. Therefore, the WPP which are the wavelet coefficients of the subband energies are obtained by,

$$WPP(m,n) = \frac{1}{a_0}^{m/2} \int \log S_i \psi({a_0}^{-m}t - nb_0)dt. \tag{8}$$

It has been shown [6, 9] that multirate filters arranged in a dyadic tree can be used to compute the coefficents $WPP(m,n)$. In our case, we used Daubechies' 4 tap filters to compute a 3 level wavelet transform.

2.5 Mel-Frequency Cepstral Parameters (MFCC)

A *Mel* is a unit of measure of *perceived pitch* or *frequency* of a tone. The Mel-scale is therefore a mapping between the real frequency scale (Hz) and the perceived frequency scale (mels). The mapping is virtually linear below 1 kHz and logarithmic above. Extensive research on MFCCs indicate that they are less sensitive to noise compared to other currently used parameters and provide better recognition/identification performance than other parametrization schemes [5, 2]. Although the triangular filterbank is used in this study, other windows such as Hamming or Hanning type could be used. After windowing the incoming speech signal, the Discrete Fourier Transform of the the frame of speech is taken. A magnitude spectrum is computed and frequency warped in order to transform the spectrum into Mel frequency in which the filterbank is uniformly spaced. The filters multiplied with the magnitude spectra of the frame and log energies are computed. Next, the discrete time cosine transform of the filterbank log energies are taken to find the MFCCs. In this study, 20 filterbanks and 19 MFCCs are used for the simulations.

3. THE SPEAKER IDENTIFICATION SYSTEM

2.1 The Gaussian Mixture Model

In this study, a Gaussian Mixture Model approach proposed in [2] is used where speakers are modeled as a mixture of Gaussian densities. The use of this model is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes and the capability of Gaussian mixtures to model arbitrary densities.

The Gausssian Mixture Model is a linear combination of M Gaussian mixture densities, and given by the equation,

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{9}$$

where $\vec{x}$ is a D-dimensional random vector, $b_i(\vec{x})$, $i = 1, .., M$ are the component densities and p_i, $i = 1, .., M$ are the mixture weights. Each component density is a D-dimensional Gaussian function of the form

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} \mid \sum_i \mid^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)^T \sum\nolimits_i^{-1} (\vec{x} - \vec{\mu}_i)\} \tag{10}$$

where $\vec{\mu}$ denotes the mean vector and $\sum_i$ denotes the covariance matrix. The mixture weights satisfy the law of total probability, $\sum_{i=1}^{M} p_i = 1$. The major advantage of this representation of speaker models is the mathematical tractibility where the complete Gaussian mixture density is represented by only the mean vectors, covariance matrices and mixture weights from all component densities.

4. EVALUATIONS

The simulations are conducted on the TIMIT database. TIMIT contains 6300 sentences spoken by 630 speakers sampling the regional accents of U.S. Although TIMIT is sampled at 16 kHz, we downsampled to 8 kHz to simulate more realistic environments. In the speaker ID system, we used 32 mixtures with diagonal covariance matrices to model each speaker. The models are trained by using the Expectation Maximization algorithm (EM) [3]. The goal of the EM algorithm is to start with an initial model, λ, and iteratively estimate a new model λ', such that $p(X \mid \lambda') \geq p(X \mid \lambda)$. For training, 8 unique sentences (approximately 24 seconds) are used, whereas, for testing 2 unique 3 second sentences are used separately. Although the complete database consists of 630 speakers, we also evaluate the parameters on a TIMIT test speaker set consisting of 168 speakers. As a result we have 1260 and 336 test sentences for 630 and 168 speakers, respectively. Wavelet packet transform is implemented by using a 32^{nd} order Daubechies' orthogonal filters. The simulation results with 3 seconds of testing data on 630 speakers for MFCC, SBC and WPP parameters are **94.8**%, **96.0**% and **97.3**%, respectively. WPP and SBC have achieved **98.8**% and **98.5**% respectively, for 168 speakers. Although both WPP and SBC performed equally well on a limited test set (168 speakers generating 336 testing tokens), WPP outperformed SBC on the full test set. MFCC achieved **96.4**% for 168 speakers, which is better than the **95.2**% reported in [4]. WPP achieved a **48**% and **67**% reduction in relative error over MFCC for 630 and 168 speakers, respectively. Table 2 summarizes the final results for all three parameters. We have observed that these features are better suited to speaker identification than MFCC over the TIMIT database.

SPEAKER ID SCORES (%)			
TESTING	MFCC	SBC	WPP
168 Speakers	96.4	98.5	98.8
# of miss	12	5	4
630 Speakers	94.8	96.0	97.3
# of miss	66	50	34

Table 2: Speaker ID Scores of Each Feature

5. SUMMARY

A new feature set based on the wavelet packet transform of the speech signal is proposed with application to speaker identification. Two features entitled Subband Based Cepstral (SBC) parameters and Wavelet Packet Parameters (WPP) were derived. We have shown that the accumulated correlation measure for energies using wavelet transform is less than that for the discrete cosine transform. The simulation results indicate that the new parameters are able to outperform MFCC over an 8 kHz version of the TIMIT database. We are presently exploring the performance of these features for other databases such as NIMIT (telephone channel) and YOHO (session to session variability) by utilizing the denoising techniques and other signal enhancement techniques on the features itself in the wavelet domain.

References

[1] R. Sarikaya and J. N. Gowdy, "Subband Based Classification of Speech Under Stress," *ICASSP-98,* vol. 1, pp. 569-572, 1998.

[2] D. A. Reynolds, "Experimental Evaluation of Features for Robust Speaker Identification," *IEEE Transactions on SAP,* vol. 2, pp. 639-643, 1994.

[3] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Spekaer Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on SAP,* vol. 3, pp. 72-83, 1995.

[4] D. A. Reynolds, M. A. Zissman, T. F. Quatieri, G. C. O'Leary, B. A. Carlson , "The Effect of Telephone Transmission Degradations on Speaker Recognition Performance," *ICASSP-95,* pp. 329-332, 1995.

[5] S. B. Davis and P Melmelstein, "Comparison of Parametric Representationa for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on ASSP,* pp. 357-366, 1980.

[6] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Comm. Pure Appl. Math.,* pp. 909-996, 1988.

[7] M. Antonini, M. Barlaud , P. Mathieu and I.Daubechies, "Image Coding Using Wavelet Transform," *IEEE Transactions on Image Proces.,* vol. 2, pp. 205-220, 1992.

[8] P. Alexandre and P. Lockwood,"Root cepstral analysis: A unified view: Application to speech processing in car noise environments,"*Speech Communication,* v.12, pp. 277-288, 1993.

[9] O. Rioul and M. Vetterli, "Wavelets and Signal Processing,"*IEEE Signal Proc. Magazine,* vol. 8(4), pp. 11-38, 1991.

[10] E. Erzin, A. E. Cetin and Y. Yardimci, "Subband analysis for speech recognition in the presence of car noise," *ICASSP-95,* vol. 1, pp. 417-420, 1995.