Selection of the Best Wavelet Packet Nodes Based on Mutual Information for Speaker Identification

Rafael Fernández, Ana Montalvo, José R. Calvo, and Gabriel Hernández

Advanced Technologies Application Center, Havana, Cuba
{rfernandez,amontalvo,jcalvo,gsierra}@cenatav.co.cu

Abstract. The analysis of the speech signal using wavelet packet trees (WPT) is a very flexible tool, capable of effectively manipulate the frequency subbands thanks to the orthonormal bases it provides. Here, dimension reduction becomes very important since the number of subbands grows exponentially with the level of decomposition, and their discriminative relevancy is different, which leads to different resolution for each one. A method based on mutual information is proposed in order to keep as much discriminative information as possible and the less amount of redundant information.

Keywords: WPT, mutual information, feature selection, speaker identification.

1 Introduction

The task of feature extraction is a crucial step in a speaker recognition system. The performance of the later components – speaker modeling and pattern matching – is strongly determined by the quality of the features extracted in this first stage [1,2]. Many methods have been proposed like MFCC (*Mel-Frequency Cepstral Coefficients*), linear predictive cepstral coefficients (LPCC), and many others, in order to model the characteristics of the speech or the vocal tract. However, a more flexible tool, capable to discover the optimum frequency subband decomposition, is necessary. The application of the multiresolution analysis is a powerful way to deal with this problem. In this field, the wavelet theory has been widely applied in problems like noise reduction, detection of discontinuities and wave forms, and signal coding and compression. In particular, Wavelet Packet Transform has proved its effectiveness as a signal processing tool in a variety of speech processing applications [3,4], and it is an alternative to the traditional Fourier Transform based techniques for analyzing time series.

On the other hand, feature selection becomes important when applying wavelet packet decomposition for two main reasons. First, the number of subbands in a wavelet packet tree grows exponentially with the number of decomposition levels. Second, the discriminative information of each subband is not the same, and therefore, their resolution should be different taking into account certain

J. Ruiz-Shulcloper and W.G. Kropatsch (Eds.): CIARP 2008, LNCS 5197, pp. 78–85, 2008.

criterion. In this work we propose a method based on mutual information in order to determine the optimum resolution at each subband.

The remainder is organized as follows: Section 2 presents a brief description of the Wavelet Packet Transform; Section 3 shows the basis of information theory and the proposed method to select the best nodes of the Wavelet Packet Tree; Section 4 contains the experimental work and the results; last section is devoted to conclusions and future works.

2 Wavelet Packet Tree

The theory of wavelets gives a flexible framework to obtain signal representations with good resolutions in both the frequency and the time domain [5]. It also allows to deal with the problem of well frequency localized noise. The decomposition of a signal in a wavelet packet tree is based on the repeated application of a couple of filters, a low-pass and a high-pass, giving the choice to split the frequency axis into intervals of various bandwidths. In this multiresolution analysis of the signal, either the low or the high frequency band can be decomposed resulting in a binary tree structure [6] showed in Figure 1.

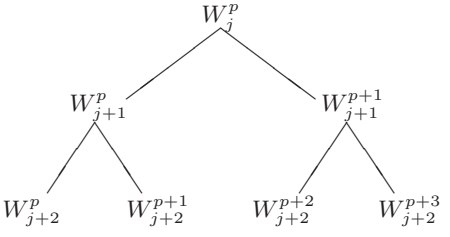

Fig. 1. Binary tree structure of the wavelet packet spaces

Each node in the tree is labeled by (j, p), where j is the depth, and p is the number of the nodes to the left of this particular node at the same depth.

The filters mentioned above characterize the orthogonal basis of $L^2(\mathbb{R})$ generated by the mother wavelet ψ, according to the multiresolution analysis, and satisfy the following conditions of orthogonality:

$$\begin{aligned} |G(\omega)|^2 + |G(\omega + \pi)|^2 &= 2, \\ G(\omega)H^*(\omega) + G(\omega + \pi)H^*(\omega + \pi) &= 0. \end{aligned} \tag{1}$$

where H and G represent the discrete time Fourier Transform of the low-pass and the high-pass filters respectively.[1] In this work we employed the polynomial spline wavelets of Battle and Lemarié [7,8], which were used in [9] for a speaker recognition system with lower EER[2] than MFCC. These wavelets are highly localized in time due to its exponential decay, and achieve similar performance

[1] A more detailed explanation of the wavelet analysis can be found in [6].
[2] Equal Error Rate.

than others with a shorter number of coefficients. Its definition in the frequency space is:

$$\hat{\psi}(\omega) = \frac{\exp(-\frac{i\omega}{2})}{\omega^{m+1}} \sqrt{\frac{S_{2m+2}(\frac{\omega}{2}+\pi)}{S_{2m+2}(\omega)S_{2m+2}(\frac{\omega}{2})}}\,, \tag{2}$$

where

$$S_n(\omega) = \sum_{k=-\infty}^{\infty} \frac{1}{(\omega + 2k\pi)^n}. \tag{3}$$

3 Feature Selection Based on Mutual Information

Reducing the dimensionality of feature vectors is usually an essential step in pattern recognition task. By removing most irrelevant and redundant features from the data, feature selection helps improve the performance of learning models by: alleviating the effect of the curse of dimensionality, enhancing generalization capability, speeding up learning process and improving model interpretability.

An evaluation of mutual information as a criterion to select the best WPT for speaker recognition is presented in this work.

In probability theory and information theory, the mutual information of two random variables is a quantity that measures their mutual dependence [10]. With this method, low information redundancy is achieved and, in contrast to others like PCA[3] – which project the features along directions of high variance – dimensionality is reduced trying to keep as much speaker discriminative information as possible.

Let $\mathcal{S}$ and X be the variables for the speaker class and the speech feature vector respectively. The mutual information between $\mathcal{S}$ and X is given by:

$$I(\mathcal{S}, X) = H(\mathcal{S}) + H(X) - H(\mathcal{S}, X) = H(\mathcal{S}) - H(\mathcal{S}|X), \tag{4}$$

where $H(\cdot)$ is the entropy function, which is a measure of the uncertainty of the variable. For a discrete-valued random variable Y, it is defined as:

$$H(Y) = -\sum_i p(Y = y_i) \log_2 p(Y = y_i), \tag{5}$$

where the y_i are the possible values of Y. From (4), mutual information measures the uncertainty reduction of $\mathcal{S}$ knowing the feature values. Those features with low speaker information have low values of mutual information with the speaker class. The best K features – following this criterion – are those $\{y_{i_1}, \ldots, y_{i_K}\} \subset \{y_1, \ldots, y_N\}$ which maximize the mutual information with the speaker class:

$$\{y_{i_1}, \ldots, y_{i_K}\} = \arg \max_{\{y_{j_1}, \ldots, y_{j_K}\}} I(\{y_{j_1}, \ldots, y_{j_K}\}, \mathcal{S}), \tag{6}$$

where N is the total number of features.

[3] Principal Component Analysis.

If the features were statistically independent, the search in (6) would be reduced to find those features iteratively. If we know the first $n-1$ features, the n-th is obtained as follows:

$$y_{i_n} = \arg \max_{y_k \notin \{y_{i_1}, \ldots, y_{i_{n-1}}\}} I(y_k, \mathcal{S}), \qquad n = 1, \ldots, K. \tag{7}$$

However, the latter is not always true. Then, the problem of finding out the best subset – see Eq. (6) – becomes a search for all the $\binom{N}{k}$ combinations.

In order to select the best nodes of the WPT, a sub-optimal method [11,12] was applied. If we have the first $n-1$ features, the n-th is selected according to:

$$y_{i_n} = \arg \max_{y_k \notin \{y i_1, \ldots, y_{i_{n-1}}\}} \left[I(y_k, \mathcal{S}) - \frac{1}{n-1} \sum_{m=1}^{n-1} I(y_k, y_{i_m}) \right]. \tag{8}$$

The idea is to look for those features with high mutual information with the speaker class and low average mutual information with the features previously selected. Last term in (8) can be thought of as a way to reduce the redundant information. Here, mutual information between two variables is the only estimation needed, which avoids the problem of estimating the probability densities of high dimension vectors. We used histogram method to calculate the probability densities.

3.1 The Algorithm

Based on the stated in the previous section, we developed a method to prune the wavelet packet tree, originally with 128 leaves ($D = 7$ levels). First, the random variables y_k were defined as the coefficients corresponding to each node's children. Then we start, from the last level, pruning the two children with less information. Once these children are pruned, their father becomes a terminal node, which is included in the list of possible nodes to be pruned, only if his brother – in another iteration – becomes a terminal node too. This cycle is repeated until the wished number of K features is reached:

Algorithm 1. Proposed method

$nFeat := 2^D$;
$memory := \{\}$;
$SearchList := \{2^{D-1}, 2^{D-1}+1, \ldots, 2^D - 1\}$;
while $nFeat > K$ **do**
  Find node $n \in SearchList$ whose children carry less information;
  Prune children of node n ;
  $SearchList := SearchList \setminus \{n\}$;
  if $\lfloor \frac{n}{2} \rfloor \in memory$ **then**
    $SearchList := SearchList \bigcup \{\lfloor \frac{n}{2} \rfloor\}$;
  end
  $memory := memory \bigcup \{\lfloor \frac{n}{2} \rfloor\}$;
  $nFeat := nFeat - 1$;
end

As first approach, we only used the mutual information of each couple of children with the speaker class – according to Eq. (7)– as a measure of information. Here, the variables are considered as though they were independent. We call this *Individual Pruning*. In the second approach, we also consider the influence of the mutual information between the couples – according to Eq. (8). We call this *Collective Pruning*.

4 Experiments and Results

4.1 Database

Recordings of 98 male speakers extracted from Ahumada speech database [13] were used for all the experiments. They consist of approximately one minute of spontaneous speech for the training and similar recordings but in a different session for the test.

4.2 Front End Processing

The WPT extraction scheme is as follows: the speech signal – previously sampled at 8 kHz – is filtered by a fifth order Butterworth filter with pass-band from 80 to 3800 Hz. Pre-emphasis filter $H(z) = 1 - 0.97z^{-1}$ and framing (32 milliseconds of frame size and overlap of 16 milliseconds) are applied. No Hamming or other complex window was required. Silence was removed employing a voice activity detector based on energy and zero-crossing. Next, wavelet packet decomposition is applied at a maximum depth of $D = 7$, corresponding to a frequency resolution of 31.25 Hz. The wavelet packet tree constructed provides a total of $2^D = 128$ subbands. The normalized energy at each node is computed as:

$$E_j^p = \frac{1}{N_j} \sum_{i=1}^{N_j} \left[W_j^p f(i) \right]^2, \tag{9}$$

where $W_j^p f(i)$ is the i-th coefficient of the wavelet packet transform of the signal f at node W_j^p, and N_j is the number of coefficients at that node. The final structure of the tree is defined through different criteria. One structure considered – and taken as reference – is based on the concept of critical bandwidth introduced by Fletcher [14], and applied in [9]. In that work, M. Siafarikas et al. proposed a 66-subband tree, whose structure is obtained taking into account the EER calculated for each subband. We call this approach *Critical bandwidth*. The other two structures are based on the approaches defined in the end of the previous section. Here we choose $K = 66$ in order to establish a comparison between the three schemes.

Once the wavelet packet tree is determined, logarithmic compression and DCT (*Discrete Cosine Transform*) are applied to the set of subband energies:

$$c_i = \sum_{n=1}^{K} \log E_n \cos\left[\frac{\pi}{2K} (2n-1)(i-1) \right], \qquad i = 1, \ldots, K. \tag{10}$$

Table 1. Evaluation for 35-dimension feature vectors

Tree Structure	% Id.	EER	DCF
Critical bandwidth	85.7	**13.3**	**6.2**
Individual Pruning	**88.8**	16.8	6.9
Collective Pruning	84.7	14.3	6.3

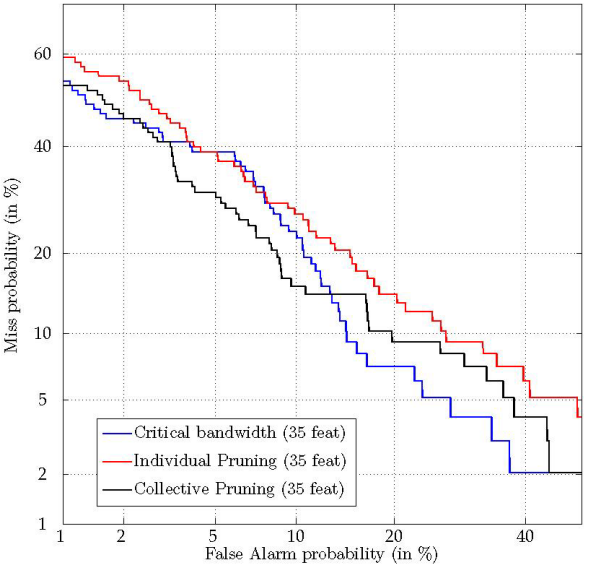

Fig. 2. Detection error trade-off curves for 35-dimension feature vectors

Table 2. Evaluation for 21-dimension feature vectors and its dynamics inclusion

Tree Structure	% Id.	EER	DCF
Individual Pruning(21 features)	**82.7**	**14.2**	**5.6**
Collective Pruning(21 features)	78.6	15.3	7.1
Individual Pruning(21 feat. + Δ)	79.6	14.6	7.0
Collective Pruning(21 feat. + Δ)	74.5	22.4	8.4

Only the first 35 coefficients are calculated, since they represent more than 99.9% of the energy of the complete set of 66. The results[4] of the experiments with these three structures are shown in Table 1 and Fig. 2.

In order to find a more compact speech signal representation – really necessary in real-time application or when processing huge amounts of information – we analyzed 21-subband trees using both Individual and Collective Pruning. Here,

[4] Every approach is evaluated by its % identification (% Id.), EER, and Detection Cost Function (DCF).

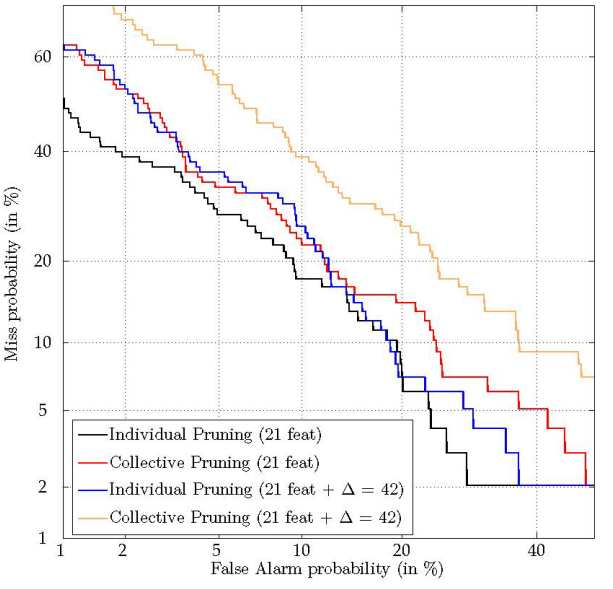

Fig. 3. Detection error trade-off curves for 21-dimension feature vectors and its dynamics inclusion

the 21 coefficients resulting from DCT were employed. The performance of these features with the addition of their dynamical information – in a feature vector of 42 dimensions – was also studied. In Table 2 and Fig. 3 the results of these experiments are shown.

5 Conclusions and Future Work

A mutual information criterion has been applied to build a wavelet packet tree for speaker identification. The proposed method – for 35-dimension feature vectors – shows better identification rate than the critical bandwidth criterion although the latter achieves better EER and DCF. The inclusion of the mutual information between the features using the model stated in (8) improved the results only for the EER and DCF in the experiments with 35-dimension features vectors. Thus, better models must be studied in order to take advantage of this information effectively. The method showed its usefulness for dimension reduction – see Table 2 – obtaining in one case the best DCF of all the experiments. Another interesting result was the performance decrease when dynamics information was included. Two factors could explain this behavior: first, the course of dimensionality; second, the feature selection method employed, which leads to features for whom its dynamics does not add considerably new information. Further studies must be done to find not only the optimal configuration, but also the optimal dimension for the features. Other ways to find the most informative time-spectral regions will be analyzed in the future.

References

1. Campbell Jr., J.P.: Speaker recognition: a tutorial. Proceedings of the IEEE 85(9), 1437–1462 (1997)
2. Kinnunen, T.: Spectral Features for Automatic Text-Independent Speaker Recognition. Licentiate's thesis, University of Joensuu, Department of Computer Science, Joensuu, Finland (2004)
3. Sarikaya, R., Hansen, H.L.: High resolution speech feature parametrization for monophone-based stressed speech recognition. IEEE Signal Processing Letters 7(7), 182–185 (2000)
4. Farooq, O., Datta, S.: Mel-scaled wavelet filter based features for noisy unvoiced phoneme recognition. In: International Conference on Spoken Language Processing ICSLP, pp. 1017–1020 (2002)
5. Goswami, J.C., Chan, A.K.: Fundamentals of Wavelets: Theory, Algorithms, and Applications. John Wiley & Sons, Chichester (1999)
6. Mallat, S.: A wavelet tour of signal processing. Academic Press, San Diego (1998)
7. Battle, G.: A block spin construction of ondelettes. Part I: Lemarié functions. Comm. Math. Phys. 110, 601–615 (1987)
8. Lemarié, P.G.: Ondelettes à localisation exponentielle. J. Math. Pures Appl. 67, 227–236 (1988)
9. Siafarikas, M., Ganchev, T., Fakotakis, N.: Wavelet Packet Based Speaker Verification. In: Ortega-Garcia, J., et al. (eds.) The Speaker and Language Recognition Workshop ODYSSEY (2004)
10. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience, Chichester (1991)
11. Peng, H.C., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of Maxdependency, Max-relevance and Min-redundancy. IEEE Trans. On Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)
12. Lu, X., Dang, J.: Dimension reduction for speaker identification based on mutual information. In: Interspeech, pp. 2021–2024 (2007)
13. Ortega-Garcia, J., Gonzalez-Rodriguez, J., Marrero-Aguiar, V.: AHUMADA: A large speech corpus in Spanish for speaker characterization and identification. Speech Comm. 31, 255–264 (2000)
14. Fletcher, H.: Auditory patterns. Reviews of Modern Physics 12, 47–65 (1940)