Mean Best Basis Algorithm for Wavelet Speech Parameterization

Jakub Galka, Mariusz Ziółko
Department of Electronics
AGH University of Science and Technology
Kraków, Poland
{jgalka,zio1ko}@agh.edu.pl

Abstract—In this paper a new optimized adaptive wavelet parameterization scheme for speech recognition is presented. A novel extension of the Best Basis algorithm is used on wavelet-packet cosine transform (WPCT) instead of typical Mel-scaled filter bank. Obtained features are tested using Polish language HMM phone-classifier.

Keywords—speech recognition; wavelets; best basis

I. INTRODUCTION

Almost all speech recognition systems transform acoustic waveforms into vectors that represent important features of the speech signal. This process is called the feature extraction or parameterization, and has been studied for a long time. Its aim is to reduce redundancy of the representation of a signal without losing its content.

Mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) are the most popular among all other methods. These methods are based on algorithms developed from windowed discrete Fourier transform (DFT). Its main disadvantage is caused by an equal-length windowing applied to each of various analyzed frequencies. The same time-resolution (often too high or too low) is used to measure different frequencies. It is inadvisable and may lead to a noticeable border effect propagation for some frequencies, followed by time-resolution loss for others.

Wavelet transform performs analysis of various frequencies (related to wavelet scales) using various and adequate windows lengths, therefore above-mentioned disadvantages can be reduced. Classic discrete decomposition schemes: dyadic (DWT), and packet wavelet (WP), do not fulfill all essential conditions required for direct use in parameterization. DWT do not provide sufficient number of frequency bands for effective speech analysis; however it is a good approximation of the perceptual frequency division [1], [2]. Wavelet packets do provide enough frequency bands, but do not respect the non-linear frequency perception phenomena [3], [4], [5].

Various decomposition schemes for an efficient speech parameterization had been presented [6]. Most of works present approximation of perceptual frequency division with an arbitrary or empirically chosen decomposition subtree [7], [8], [9], [10], [11], [12]. These papers do not provide description of the subtree selection method. In some works wavelet a-scale has been properly chosen to obtain mel-frequency scale in a wavelet transform [13].

Wickerhouser’s best wavelet basis selection (BB), entropy-based algorithm [14] has been used by Datta and Long [6] to obtain the best decomposition schemes of single phonemes. Others mention use of this algorithm in a parameterization of consonants [15]. Unfortunately, the well known Best Basis and Joint Best Basis (JBB) algorithms can not be used for sets of variable-length data. According to Wickerhouser’s works [14], the best basis selection requires the computation of the variance of the data in the training set. It can be computed only when all data vectors in the set are of the same length. In this paper, a new method of the best wavelet basis selection, which overcomes this limitation, is presented.

II. MEAN BEST BASIS ALGORITHM

A. Wavelet Packet Cosine Transform

Multi-level wavelet packets produce $2^M$ wavelet coefficient vectors, where $M$ stands for the number of decomposition levels. Wavelet coefficient vectors

$$d_{m,j} = [d_{m,j}] \in \mathbb{R}^N,$$

represent uniformly distributed frequency banks. Decomposition process may be represented by a full binary tree

$$W_{WP} = \{W_{WP}^{m,j}\}_{m,j} : W_{WP}^{m,j} \rightarrow d_{m,j},$$

with a sample of speech signal $d_{0,0}$ (single frame of speech) related to its root, and wavelet coefficients $d_{m,j}$ related to its nodes and leafs (when $m=M$) [16], [17].

For a better spectral entropy extraction from the speech signal we applied the discrete cosine transform

$$\tilde{d}_{m,j}(k) = \sum_{n=1}^{N} d_{m,j}(n) \cdot \cos\left(2\pi \frac{nk}{N_m}\right),$$

to each of the WP tree nodes to obtain the Wavelet Packet Cosine Transform (WPCT). It eliminates the problem of a time shifting in the entropy measure and takes account of more important spectral content. This is a very important step since the speech is a time-spectral phenomenon [3], [4].
B. Best Basis Algorithm

The best wavelet basis subtree $W_{\text{opt}}$ may be defined as a set $W$ of tree nodes

$$W_{\text{opt}} = \arg\min_{W} \sum_{x_{m,j} \in W} x_{m,j},$$

which minimizes its total entropy and generates an orthogonal decomposition base [14], where the node split cost function

$$x(d_{m,j}) = -\sum_{n=1}^{N} d_{m,j}^{2}(n) \log \left( \frac{d_{m,j}^{2}(n)}{\|d_{m,j}\|^{2}} \right),$$

is the Shannon entropy of the Wavelet-Packet Cosine Transform coefficients (WPCT).

Best Basis algorithm may be applied to a single signal when it is needed. However, finding the best decomposition scheme for a set of signals can not be done using this method. When a set of signals is given, Joint Best Basis algorithm may be used [18], [19]. It utilizes a tree of signal variances

$$W_{\sigma,\text{opt}} = \arg\min_{W} \sum_{x_{m,j} \in W} x_{\sigma,m,j},$$

to select an optimized subtree. Unfortunately, computation of variance requires each signal to be of equal length and normalized in terms of energy and amplitude, what is even more important, when energy dependent cost function is used [14]. This is a serious limitation, since in practice signals may be of various lengths. Next section presents the solution of this problem by calculation of mean entropy values instead of signals’ variances.

C. Mean Best Basis Algorithm

The set of speech signals used in this work consists of phoneme samples extracted from Polish speech database Corpora. Phonemes are of various lengths, depending on the phoneme class and case. Each pattern is actually unique. Under these conditions the use of variance-based JBB algorithm is impossible. The tree of variances cannot be fairly computed when signals are of various lengths and energies [18].

The above-mentioned problem may be solved when a new definition of the optimal tree for a set of different signals is introduced. The best decomposition tree in such case is a subtree

$$\bar{W}_{\text{opt}} = \arg\min_{\bar{W}} \sum_{x_{m,j} \in \bar{W}} \bar{x}_{m,j},$$

of a full binary tree $\bar{W}$ of nodes’ entropy mean values $\{\bar{x}\}$ over all signals in the set, for which its entire value is minimal. Having a tree of mean entropy values, one can find an optimal Mean Best Basis (MBB) subtree using the Best Basis algorithm over mean entropy tree. The algorithm consists of the following steps:

- For each element of set $\{s\}$, of signals calculate full WPCT tree

$$W_{\text{WPCT}} = \{W_{m,j}\} : W_{m,j} \leftrightarrow \hat{d}_{m,j}.\quad (8)$$

- Find entropy value

$$x_{m,j}^{i} = x\left(\hat{d}_{m,j}\right),\quad (9)$$

for each node of all calculated WPCT trees.

- For each of the obtained trees $W_{i}$ normalize entropy values within the whole tree according to its root entropy value

$$\forall i, \forall m,j x_{m,j}^{i} = \frac{x_{m,j}}{x_{0,1}}.\quad (10)$$

It makes the cost-function (entropy) independent of different signal energy values. After this step, every signal from the set will be equally important in the basis selection process.

- Calculate the general tree of mean entropy values over all signals with all entropy values normalized. This step can be performed because cost function values have been normalized in previous step, and therefore are independent from length of the signal.

- Find the best subtree using the Wickerhouser’s Best Basis algorithm with a mean-entropy tree $\bar{W}$.

The obtained wavelet decomposition scheme depends on the entropy and spectral properties of all signals used in the computations. Frequency bands containing more spectral variations among all signals in the set are represented in the optimized wavelet spectrum with a higher spectral resolution. In Fig. 1 a wavelet decomposition tree, obtained for all of the phones of Polish language with a Daubechies’ 6th order wavelet and Mean Best Basis algorithm is presented. The order of tree branches is not frequency-based because of the disordering effect of multilevel decimation / filtering present in the decomposition process [20].

In Fig. 1 one can also notice a higher resolution of the spectrum in the frequency ranges related to the 1st and the 2nd formant. The spectrum has been generated using the tree presented in the left plot. Bands in the spectrum plot are frequency-ordered.
Figure 1. Optimized MBB wavelet decomposition tree for Polish speech, using Daubechie’s wavelet (solid lines, left plot). Utterance “AgnjeSka” (SAMPA notation, top) and its MBB optimized spectrum (right).

Figure 2. Optimized MBB wavelet decomposition tree for Polish vowels, using Daubechie’s wavelet (left). MBB vowels-optimized wavelet spectrum of the phoneme /e/ (right).

D. Feature Extraction

When the optimized decomposition tree $\tilde{W}_{opt}$ is known, it may be used for an efficient spectral analysis and feature extraction [8]. In presented experiment, energy

$$x(k) = \sum_{d_{mk}} \left\| d_{mk} \right\|^{2}, \quad (12)$$

of wavelet coefficient in each leaf was computed. Obtained values form a vector $x$ of a length equal to the optimized tree’s leaf quantity. Normalization and DCT decorrelation of the vector is then applied to use it with HMM.

III. PHONEME RECOGNITION

New decomposition schemes were tested using Polish speech database Corpora. Phone recognition task had been performed using 3617 patterns. All phoneme patterns were used in the mean best basis selection. Obtained decomposition subtree had been used for speech feature extraction. In this case 27 tree leafs produced 27 features.

Its efficacy was measured with typical 3-state, left-to-right Hidden Markov Model tri-phone classifier with no higher-level language context knowledge. Observation probability was modeled with single Gaussian per state.
Standard training and testing algorithms presented in HTK Book were used for comparison purposes [21]. Various noise conditions (AWGN) had been applied to measure the robustness of the features.

Results of this task are presented in Fig. 3. For the given feature quantity (27), phone recognition and phone accuracy rates are reaching 80% and 72% respectively on clean speech. Introduction of 10dB SNR noise results in the recognition decrease by only 10% points which proves robustness of such composed wavelet parameterization scheme.

Similar recognition task run on the vowels set with only 17 feature components resulted in 90% phone recognition accuracy for clean conditions with similar HMM setup.

IV. CONCLUSIONS

A new method of choosing the best wavelet decomposition scheme for a set of signals has been presented. It is based on the well known Wickerhouser’s Best Basis algorithm, but extends it with the possibility of selecting the decomposition tree for differentiated multi-length data. The use of a WPCT - Wavelet Packet Cosine Transform, provides high robustness of the entropy value to a time-shift and focuses on the spectral properties of the signal. Decomposition schemes obtained for the real speech data and phone recognition results confirm the method’s efficacy. Presented algorithm may be used with other types of signals, e.g. image data.

Future works will focus on finding the better, aim-oriented cost function (in place of entropy) used in a tree selection process.

ACKNOWLEDGMENT

This work was supported by MNiSW grant OR00001905.

We would like to thank Stefan Grocholewski from Poznań University of Technology for providing a corpus of spoken Polish - CORPORA’97.

REFERENCES


