WAVELET PARAMETERIZATION FOR SPEECH RECOGNITION

Jakub Gałka[1], Mariusz Ziółko[1],

[1] Department of Electronics, AGH University of Science and Technology,
Al. Mickiewicza 30,
30-059 Kraków, Poland
{jgalka, ziolko}@agh.edu.pl

Abstract. Typical parameterization schemes utilize linear prediction or mel-scaled filter-banks, which are classic windowed DFT based methods. In this paper a new optimized adaptive wavelet parameterization scheme is presented. A novel extension of the Best Basis algorithm is used on wavelet-packet cosine transform (WPCT) instead of typical filter bank. Obtained features are tested using Polish language HMM phone-classifier.

Keywords: wavelet transform, best basis, speech parameterization, speech recognition

1 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 and the most often used among other methods. These methods are based on algorithms developed from windowed discrete Fourier transform (DFT). Its main disadvantage is caused by an equal window size applied to each of various analyzed frequencies. The same time-resolution is used to measure different frequencies (too high or too low). It is inadvisable and may lead to noticeable border effect propagation for some frequencies, followed by time-resolution loss for others, also when psychoacoustic mel-scale had been applied.

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, however they 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 filters have been warped or 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. Other works mention use of this algorithm in a parameterization of plosive consonants [15].

Unfortunately, the well known Best Basis and Joint Best Basis (JBB) algorithms can not be used for sets of variable length data. In this paper, a new method of best wavelet basis selection is presented. It is applicable to sets of a non-uniform data, like various-length phoneme samples.

2 MEAN BEST BASIS DECOMPOSITION

2.1 Wavelet Packet Cosine Transform (WPCT)

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

$$\boldsymbol{d}_{m,j} = \left[d_{m,j} \right] \in \Re^K \ , \tag{1}$$

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

$$\boldsymbol{W}^{WPT} = \left\{ W_{m,j}^{WPT} \right\}_{m,j} : \quad W_{m,j}^{WPT} \leftrightarrow \boldsymbol{d}_{m,j} \ , \tag{2}$$

with a sample of speech signal $\boldsymbol{d}_{0,0}$ (single frame of speech) related to its root, and wavelet coefficients $\boldsymbol{d}_{m,j}$ 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

$$\widehat{d}_{m,j}(k) = \sum_{n=1}^{N_m} d_{m,j}(n) \cdot \cos\left(2\pi \frac{nk}{N_m} \right) , \tag{3}$$

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

2.2 Best Basis Algorithm

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

$$\boldsymbol{W}^{opt} = \underset{\boldsymbol{W}^{*}}{\operatorname{argmin}} \sum_{\varkappa_{m,j} \leftrightarrow \boldsymbol{W}^{*}} \varkappa_{m,j} \ , \tag{4}$$

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

$$\varkappa\left(\hat{\boldsymbol{d}}_{m,j}\right) = -\sum_{n=1}^{N_d} \left(\frac{\hat{d}_{m,j}{}^{2}(n)}{\left\|\hat{\boldsymbol{d}}_{m,j}\right\|^{2}} \log\left(\frac{\hat{d}_{m,j}{}^{2}(n)}{\left\|\hat{\boldsymbol{d}}_{m,j}\right\|^{2}} \right) \right) , \tag{5}$$

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

$$\boldsymbol{W}^{\sigma,opt} = \underset{\boldsymbol{W}^{\sigma}}{\operatorname{argmin}} \sum_{\varkappa^{\sigma}_{m,j} \leftrightarrow \boldsymbol{W}^{\sigma}} \varkappa^{\sigma}_{m,j} \ , \tag{6}$$

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.

2.3 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 this 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

$$\bar{\boldsymbol{W}}^{opt} = \underset{\bar{\boldsymbol{W}}^{*}}{\operatorname{argmin}} \sum_{\bar{\varkappa}_{m,j} \leftrightarrow \bar{\boldsymbol{W}}^{*}} \bar{\varkappa}_{m,j} \ , \tag{7}$$

of a full binary tree $\bar{\boldsymbol{W}}^{\varkappa}$ of nodes' entropy mean values $\{\bar{\varkappa}\}$ 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 $\{\boldsymbol{s}\}_i$ of signals calculate full WPCT tree

$$\boldsymbol{W}^{WPCT} = \{W_{m,j}^{WPCT}\}: \quad W_{m,j}^{WPCT} \leftrightarrow \hat{\boldsymbol{d}}_{m,j} \ . \tag{8}$$

- Find entropy value

$$\varkappa_{m,j}^{i} = \varkappa\left(\hat{\boldsymbol{d}}_{m,j}^{i}\right), \tag{9}$$

  for each node of all previously calculated WPCT trees.

- For each of the obtained trees $\boldsymbol{W}_i^{\varkappa}$ normalize entropy values within the whole tree according to its root entropy value

$$\forall_i \forall_{m,j} \quad \varkappa_{m,j}^{i} = \frac{\varkappa_{m,j}^{i}}{\varkappa_{0,1}^{i}} \ . \tag{10}$$

  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

$$\bar{\boldsymbol{W}}^{\varkappa} = \{\bar{W}_{m,j}^{\varkappa} \leftrightarrow \bar{\varkappa}_{m,j}\}: \ \bar{\varkappa}_{m,j} = \frac{1}{\left|\{\boldsymbol{s}\}\right|} \sum_{\varkappa_{m,j}^{i} \leftrightarrow \boldsymbol{W}_i^{\varkappa}} \varkappa_{m,j}^{i} \ , \tag{11}$$

  the general tree of mean entropy values over all signals with all entropy values normalized.

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

The wavelet decomposition scheme obtained 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 Daubechie's 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.

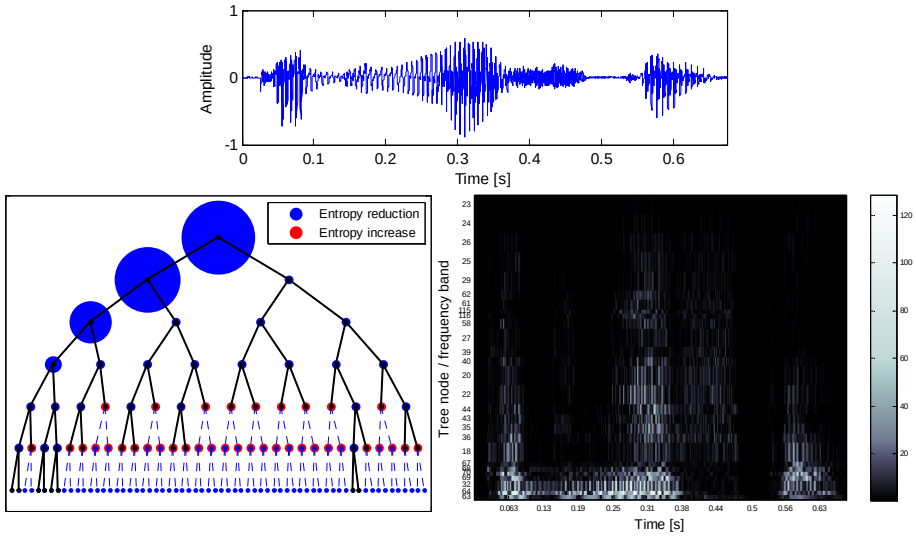

Fig. 1. Optimized MBB wavelet decomposition tree for polish speech database Corpora, using Daubechie wavelet and Shannon entropy (*solid lines, left plot*). Utterance *"Agn'jeSka"* (SAMPA notation, *top*) and its MBB optimized spectrum (*right*).

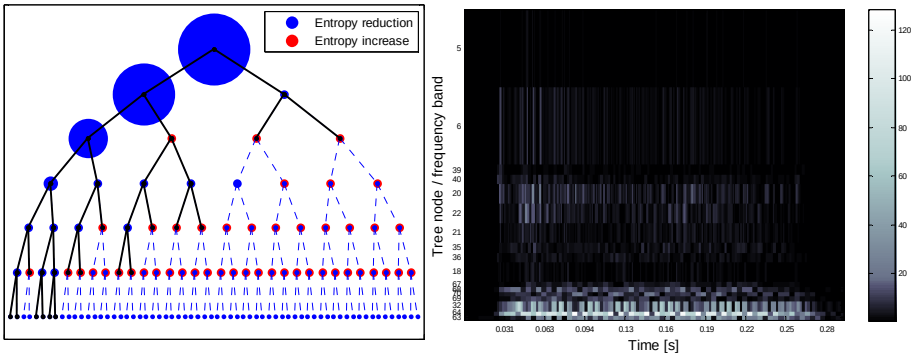

Fig. 2. Optimized MBB wavelet decomposition tree for Polish vowels, using Daubechies wavelet and Shannon entropy (*left*). MBB vowels-optimized wavelet spectrum of the phoneme /*e*/ (*right*).

2.4 Feature Extraction

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

$$x(k) = \sum_{\boldsymbol{d}_{m_k, j_k} \leftrightarrow \boldsymbol{W}^{opt}} \left\| \boldsymbol{d}_{m_k, j_k} \right\|^2 , \tag{12}$$

of wavelet coefficient in each leaf was computed. Obtained values form a vector $\boldsymbol{x}$ of a length equal to the optimized tree's leaf quantity. Normalization and DCT decorelation of the vector is then applied to use it with an HMM phone recognizer.

3 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 Hidden Markov Model tri-phone classifier with no higher-level language context knowledge [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.

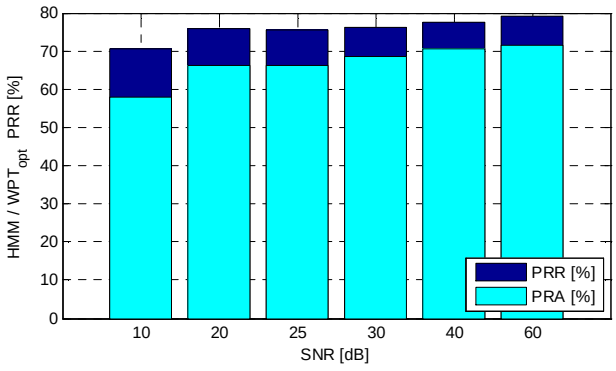

Fig. 3. Phoneme recognition results for the MBB-optimized parameterization scheme.

4 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.

Acknowledgments. 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

1. Datta, S., Farooq, O.: Phoneme Recognition Using Wavelet Based Features. An International Journal on Information Sciences. 150, Elsevier (2003)
2. Tan, B. T., Fu, M., Spray, A., Dermody, Ph.: The Use of Wavelet Transforms in Phoneme Recognition. In: Proceedings of ICSLP, (1996)
3. Gałka, J., Kepiński, M., Ziółko, M.: Speech Signals in Wavelet-Fourier Domain. In: Proceedings of The Fiftieth Open Seminar on Acoustics - Speech Analysis, Synthesis And Recognition In Technology, Linguistics And Medicine. Archives of Acoustics. vol. 28, no. 3
4. Gałka, J., Kepiński, M.: WFT context-sensitive speech signal representation. In: Advances in Soft Computing: Proceedings of the IIPWM. Springer, Heidelberg (2006)
5. Ganchev, T., Siafarikas, M., Fakotakis, N.: Speaker Verification Based on Wavelet Packets. In: LNCS, Springer, Heidelberg (2004)
6. Datta, S., Long, C. J.: Wavelet Based Feature Extraction for Phoneme Recognition, In: Proceedings of ICSLP 1996, (1996)
7. Datta, S., Farooq, O.: Mel Filter-Like Admissible Wavelet Packet Structure for Speech Recognition. IEEE Signal Processing Letters, vol. 8, no. 7, (2001)
8. Datta, S., Farooq, O.: Wavelet Based Robust Sub-band Features For Phoneme Recognition. IEE Proceedings: Vision, Image and Signal Processing, vol. 151(3), (2004)
9. Datta, S., Farooq, O.: Mel-Scaled Wavelet Filter Based Features For Noisy Unvoiced Phoneme Recognition. In: Proceedings of ICSLP, (2002)
10. Gowdy, J. N., Tufekci, Z.: Mel-Scaled Discrete Wavelet Coefficients for Speech Recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP'00, vol. 3. Istanbul (2000)

11. Sarikaya, R., Hansen, J. H. L.: High Resolution Speech Feature Parameterization for Monophone – Based Stressed Speech Recognition. IEEE Signal Processing Letters, vol. 7, no. 7, (2000)
12. Sarikaya, R., Gowdy, J. N.: Subband Based Classification of Speech Under Stress. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Seattle (1998)
13. Evangelista, G., Cavaliere, S.: Discrete Frequency Warped Wavelets: Theory and Applications. IEEE Transactions On Signal Processing, vol. 46, no. 4, (1998)
14. Wickerhauser, M. V., Coifman R. R.: Entropy-Based Algorithms for Best Basis Selection, IEEE Transactions On Information Theory, vol. 38, no. 2, (1992)
15. Łukasik, E.: Classification Of Voiceless Plosives Using Wavelet Packet Based Approaches. In: Proceedings of EUSIPCO, (2000)
16. Daubechies, I.: Ten Lectures on Wavelets. SIAM (1992)
17. Vetterli, M., Ramchandran, K., Herley, C.: Wavelets, Subband Coding, and Best Bases. Proceedings of the IEEE, vol. 84, no. 4, (1996)
18. Wickerhauser, M. V.: Designing a Custom Wavelet Packet Image Compression Scheme with Applications to Fingerprints And Seismic Data. In: Proceedings of Perspectives in Mathematical Physics: Conference in Honor of Alex Grossmann. CRC Press (1998)
19. Wickerhauser, M. V., Odgaard, P. F., Stoustrup, J.: Wavelet Packet Based Detection Of Surface Faults On Compact Discs, In: Proceedings of 6^{th} IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, (2006)
20. Vetterli, M., Ramchandran, K.: Best Wavelet Packet Bases Using Rate-Distortion Criteria. In: Proceedings of IEEE International Symposium on Circuits and Systems – ISCAS, (1992)
21. Young, S., et al.: HTK Book. Cambridge University Engineering Department, Cambridge (2005)