Perceptual Wavelet Decomposition for Speech Segmentation

Mariusz Ziółko[1], Jakub Gałka[1], Bartosz Ziółko[1] and Tomasz Drwięga[2]

[1] Department of Electronics, AGH University of Science and Technology, Kraków, Poland
[2]Faculty of Applied Mathematics, AGH University of Science and Technology, Kraków, Poland
{ziolko,jgalka,bziolko}@agh.edu.pl, drwiega@wms.mat.agh.edu.pl

Abstract

A non-uniform speech segmentation method based on wavelet packet transform is used for the localisation of phoneme boundaries. Eleven subbands are chosen by applying the mean best basis algorithm. Perceptual scale is used for decomposition of speech via Meyer wavelet in the wavelet packet structure. A real valued vector representing the digital speech signal is decomposed into phone-like units by placing segment borders according to the result of the multiresolution analysis. The final decision on localisation of the boundaries is made by analysis of the energy flows among the decomposition levels.

Index Terms: speech segmentation, wavelet packet transform, speech recognition

1. Introduction

Speech signals typically need to be divided into small segments before starting a recognition procedure. Analysis and classification of these frames can determine the likelihood of a particular phoneme being present within the frame.

Speech is a non-stationary signal in the sense that frequency components change continuously over time, but it is generally assumed to be a stationary process within a single frame. Naturally, this causes recognition difficulties if the frame contains the end of one phoneme and the beginning of another caused by phonetic coarticulation.

Segmentation methods currently used in speech recognition do not consider where phoneme begins and ends are. Uniform segmentation causes transient information to appear at the boundaries of phonemes. For more accurate modelling, non-uniform phoneme segmentation can be useful in speech recognition [1].

Many speech segmentation algorithms (see [2], [3]) have been used in speech processing systems, but only a few of them use the wavelet spectra [2, 4]. The discrete wavelet transform (DWT) belongs to the group of frequency transformations. Wavelet methods are known to be very useful in the time-frequency analysis of non-stationary signals $\{s(n)\}$ [5, 6]. Wavelet transform combines the best properties of classic frequency and time analysis in a common tool. DWT may be more similar than other methods to the principles of the operation of human hearing system equipped with subsystem for frequency analysis to reveal the important information for the human speech recognition ability. Dyadic frequency division makes the DWT much more compatible with the human hearing system than other methods.

2. Wavelet Decomposition

It was observed that Mayer wavelets give the separation of frequency band with a better resolution than other wavelets. Therefore, the discrete Mayer wavelet is used in the method presented below. The Meyer wavelet with the frequency band from 4 kHz to 16 kHz is defined by formula

$$\widehat{\psi}(f) = \frac{10^{-3}}{12} \begin{cases} e^{\frac{j\pi f}{12000}} \sin\left(\frac{\pi}{2}\nu\left(\frac{|f|}{125\cdot 10^{6}} - 1\right)\right) \\ \qquad \text{if } 4000 \leq |f| \leq 8000 \\ e^{\frac{j\pi f}{12000}} \cos\left(\frac{\pi}{2}\nu\left(\frac{|f|}{125\cdot 10^{6}} - 1\right)\right) \\ \qquad \text{if } 8000 \leq |f| \leq 16000 \\ 0 \qquad \text{if } |f| \notin [4000, 16000] \end{cases} \tag{1}$$

where $j^2 = -1$. The scaling function with the frequency band limited to 8 kHz is defined by formula

$$\widehat{\varphi}(f) = \frac{10^{-3}}{12} \begin{cases} 1 & \text{if } |f| < 4000 \\ \cos\left(\frac{\pi}{2}\nu\left(\frac{|f|}{25\cdot 10^{5}} - 1\right)\right) \\ \quad \text{if } 4000 \leq |f| \leq 8000 \\ 0 & \text{if } |f| > 8000 \end{cases} \tag{2}$$

where

$$\nu(x) = \begin{cases} 0 \quad \text{if } x < 0 \\ x^4(35 - 84x + 70x^2 - 20x^3) \;\text{ if } 0 \leq x \leq 1 \\ 1 \quad \text{if } x > 1 \end{cases} \tag{3}$$

and condition $\nu(x) + \nu(1-x) = 1$ is fulfilled for $x \in [0, 1]$.

Meyer wavelets are frequency band-limited functions whose Fourier transforms (1) and (2) are smooth. The scale function spectrum defined by (2) and (3) is presented in Fig.1. It is worth to notice that spectrum has compact support which leads to infinite support of scale function in the time domain. By applying the inverse Fourier transform to spectrum (2) we obtain

$$\begin{aligned} \varphi(t) = 1200^{-1} \left[\pi^{-1}t^{-1}\sin(8000\pi t) + \right. \\ \int\limits_{-8000}^{-4000} \cos\left(0.5\pi\nu\left(\tfrac{|f|}{4000} - 1\right)\right)\cos(2\pi f t)df + \\ \left. \int\limits_{4000}^{8000} \cos\left(0.5\pi\nu\left(\tfrac{f}{4000} - 1\right)\right)\cos(2\pi f t)df \right], \end{aligned} \tag{4}$$

where integrals must be computed numerically. The supporting area in time domain is not limited, however the time decay of this wavelet is high. A compact support approximation of (2) is used as discrete Mayer wavelet.

In order to obtain the DWT, the coefficient $c_{m+1,i}$ of series

$$s(n) = \sum_i c_{m,i}\varphi_{m,i}(n) \tag{5}$$

are computed for the m-th resolution level, where

$$\varphi_{m,i}(n) = 2^{\frac{m}{2}}\varphi\left(2^m n\Delta t - i\right) \tag{6}$$

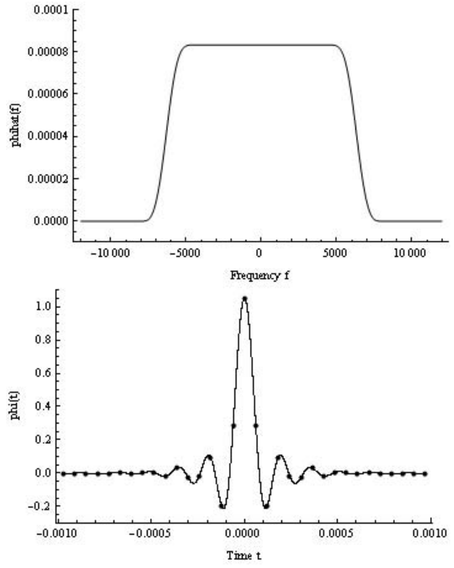

Figure 1: Spectrum (2) (upper figure) of Meyer scale function (4) with $N = 33$ samples (lower figure)

is the ith wavelet function and Δt is the sampling density. An example of wavelet function $\varphi(t)$ and its spectrum is presented in Fig. 1. Due to the orthogonality of wavelet functions $\{\varphi_{m+1,i}\}_i$ we obtain

$$c_{m,i} = 2^{\frac{m}{2}} \int\limits_{-\infty}^{+\infty} s_a(t)\, \varphi(2^m t - i)\, dt =$$

$$2^{\frac{m}{2}} \sum_{n=-\infty}^{+\infty} s(n) \int\limits_{-\infty}^{+\infty} \varphi(2^m t - i) \frac{\sin(\pi(t - n\Delta t)/\Delta t)}{\pi(t - n\Delta t)/\Delta t} dt, \tag{7}$$

where $s_a(t)$ is an analog signal and its samples create the discrete signal $s(n)$, *i.e.*

$$s_a(n\Delta t) = s(n).$$

Formula (7) has two computational disadvantages. First, it is difficult to compute integrals numerically when wavelet supports are unlimited. Secondly, the numerical computations of integrals are time-consuming, because the high quality standard needs 16 000 elements of series (5) for each second of the recorded speech signal. Approximation

$$c_{m,i} \approx \sum_{n \in D_i} s(n)\, \varphi_{m,i}(n), \tag{8}$$

is used instead of formula (7) to avoid these difficulties, where D_i are compact supports of $\varphi_{m,i}$.

The support of scale function $\varphi(t)$ must be compact to provide the fast calculations in the real time. For the opposite case it is a common feature of the scale functions that $\varphi(t) \longrightarrow 0$ very fast as $|t| \longrightarrow +\infty$. The support can be limited to the segment $[-T, T]$ where

$$T = \max\{t \in \mathbb{R} : |\varphi(t)| \geq h\}. \tag{9}$$

The threshold h should depend on the extreme value of the scale function.

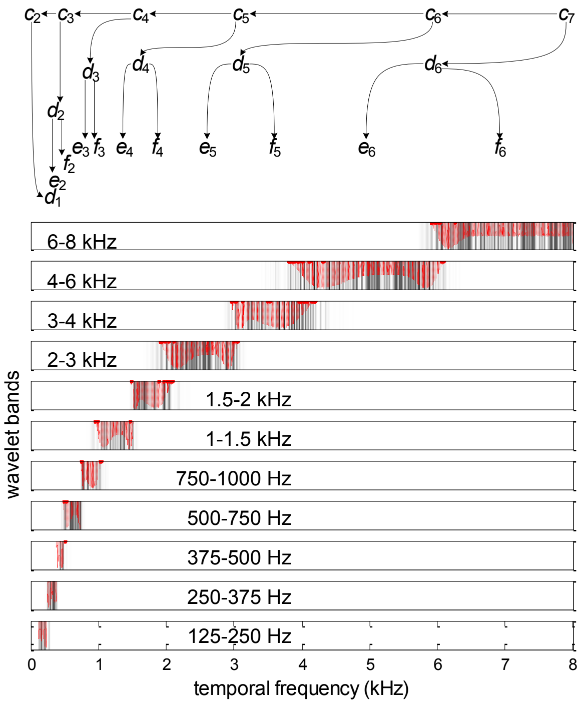

Figure 2: Perceptual speech feature extraction analysis, based on wavelet decomposition (top) and temporal frequency-sweep response of the decomposition (bottom)

The coefficients of the lower level are calculated by applying the well known [5, 7] formulae

$$c_{m-1,n} = \sum_i h_{i-2n} c_{m,i} \tag{10}$$

$$d_{m-1,n} = \sum_i g_{i-2n} c_{m,i} \tag{11}$$

where $\{h_i\}$ and $\{g_i\}$ are the coefficients which depend on the assumed scale function φ and wavelet ψ. In other words, the speech spectrum is decomposed by digital filters and downsampling operations defined by (10) and (11). It means that given the wavelet coefficients $c_{m,i}$ of the mth resolution level, (10) and (11) are applied to compute the coefficients of the $(m-1)$th resolution level. Coefficients $\{c_{m-1,n}\}_n$ are known as the coarse approximation while $\{d_{m-1,n}\}_n$ are called details coefficients. The coefficients of next resolution levels are calculated recursively by applying formulae (10) and (11). The multiresolution analysis gives a hierarchical and fast scheme for the computation of the wavelet spectrum for a given signal s.

The elements of the DWT for a mth level may be collected into a vector $\mathbf{d}_m = (d_{m,1}, d_{m,2}, \ldots)^T$. In this way the values of DWT for $M+1$ levels can be obtained. It means that dyadic discrete wavelet spectrum

$$\mathrm{DWT}(s) = \{\mathbf{c}_1, \mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_M\} \tag{12}$$

is created.

The undertaken experiments show that the speech signal decomposition into $M = 6$ levels is sufficient to cover the frequency band of voice. The energy of the speech signal above 8 kHz and below 125 Hz is very low and can be neglected.

The above presented theory is based on the wavelet analysis and leads to the dyadic decomposition. The classical DWT results in a logarithmic frequency resolution [8]. The low frequencies have narrow bandwidths and the high frequencies have wide bandwidths. Therefore, the low frequencies are investigated with finer resolution, while the wide bandwidths at high frequencies result in a poor resolution. In case of the perceptual scale [9] the number of subbands must be increased. The wavelet packet system is a generalisation of wavelet transform, in which at all stages both the low-pass and high-pass bands are split, as illustrated in Fig. 2. Therefore, it allows a finer resolution at high frequencies. It also gives a rich structure that enables adaptation the time-frequency analysis to particular signal properties. Vectors $\mathbf{d}_m$ (where $2 \le m \le 6$), which constitute a part of spectrum (12), should be split into two vectors $\mathbf{e}_n$ and $\mathbf{f}_m$ to represent the additional frequency bands according to the decomposition tree presented in Fig. 2. The elements of these vectors are computed by applying formulae

$$f_{m,n} = \sum_i h_{i-2n} d_{m,i} \qquad (13)$$

$$e_{m,n} = \sum_i g_{i-2n} d_{m,i}, \qquad (14)$$

where $2 \le m \le 6$.

From acoustic point of view [9], eleven subbands $\{\mathbf{d}_1, \mathbf{e}_2, \mathbf{f}_2, \mathbf{e}_3, \mathbf{f}_3, \ldots, \mathbf{e}_6, \mathbf{f}_6\}$ seems to be the best frequency representation of the speech properties in terms of speech analysis for non-uniform segmentation. They corresponds to human hearing system properties, and this is why the approach can be called perceptual.

3. Segmentation Scheme

The role of the segmentation algorithm is to detect the significant transitions of the energy among the wavelet sub-bands. It is marked and scored as a spectral-phonetic event. It is assumed that events occur when the energy transition changes the order of the power-sorted wavelet bands.

The non-uniform segmentation algorithm consists of the following steps:

1. Decompose signal s into spectrum $\mathbf{W} = \{\mathbf{d}_1, \mathbf{e}_2, \mathbf{f}_2, \ldots, \mathbf{e}_6, \mathbf{f}_6\}_l$ which consists of eleven levels.
2. Calculate the sum of power samples in all frequency subbands l according to rule

$$B_{l,k} = \sum_{n=(k-1)\cdot 2^{6-m}+1}^{k\cdot 2^{6-m}} \mathbf{w}_{n,l}^2, \qquad (15)$$

where k is a new discrete time index with the sampling period 4 ms, due to energy aggregation over the summation range and $\mathbf{w}_{n,l}$ are elements of $\mathbf{W}$.

3. Calculate the power envelopes as running mean values

$$B_{l,n}^{env} = \frac{1}{K} \sum_{k=n-\frac{K}{2}}^{n+\frac{K}{2}} B_{l,k}, \qquad (16)$$

where $K = 2^{-M} \Delta t_\mu f_s$ for expected mean duration Δt_μ of the speech segments. For the given $\Delta t_\mu = 0.1$ s, f_s=16 000 Hz and $M = 6$ we obtain $K = 25$ samples.

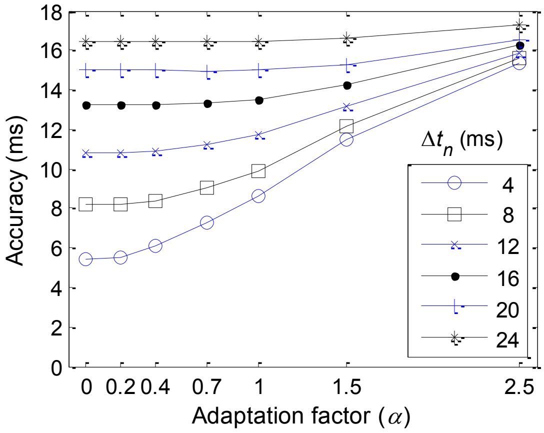

Figure 3: Accuracy of the non-uniform segment borders' localisation for various algorithm settings (α, Δt_n)

4. Generate importance matrix $\mathbf{L} = [L_{i,k}] \in \mathbb{R}^{11 \times L_s}$ of frequency bands by sorting the envelopes in each time k position *i.e.*

$$\mathbf{L} = \left\{\{l_i\}_{i=1}^{11} : B_{l_1,n}^{env} \ge \cdots \ge B_{l_{11},n}^{env}\right\}_n \qquad (17)$$

where L_s depends on the length of the speech signal utterance.

5. Compute event-function

$$f(n) = \sum_{i=1}^{11} \frac{|L_{i,n+1} - L_{i,n}|}{i}. \qquad (18)$$

6. Segment border's locations can now be extracted from $f(n)$ by choosing its local maxima, which are greater than specified threshold f_{tr} and where each of them is the highest within the neighbourhood of Δt_n milliseconds.

Time-range condition rejects multiple changes related to the same border and segments shorter than Δt_n. Threshold adjusts sensitivity of the segmentation. By increasing its value we reduce the number of chosen events. It is reasonable to set its value on-line, according to the varying values of detection function

$$f_{tr}(n) = \frac{\alpha \cdot \sum_{k=-P}^{P} f(n-k)}{2P}, \qquad (19)$$

where P is an adaptation range corresponding to 100 milliseconds, and $\alpha \ge 0$ is a sensitivity factor.

4. Results

The presented speech segmentation algorithm was tested on almost six hours of the hand-annotated and labelled Polish speech recordings extracted from *Corpora*'97 database. Selected utterances covered most of the possible Polish diphones spoken by 26 different male speakers (365 utterances each). Assigned hand-made labelling of the speech was used as an ideal reference segmentation in each evaluation test.

Accuracy of the segmentation is defined as mean value of the differences between detected and reference borders' position. Best (lowest) values have been obtained for short Δt_n

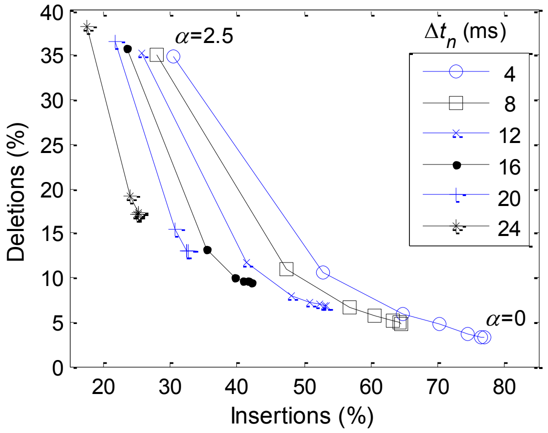

Figure 4: False border detection (insertions) and missing border (deletions) error rates of the presented system

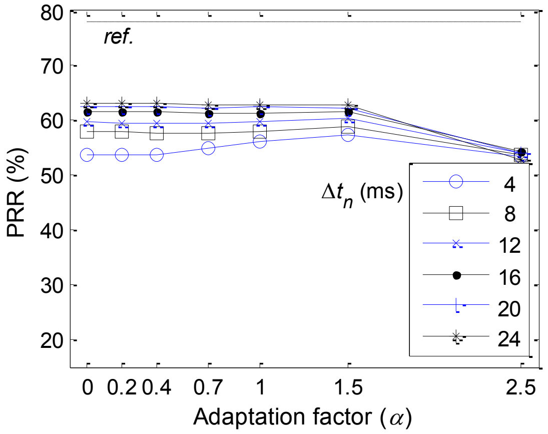

Figure 5: Phone recognition rate of the non-uniform segmentation scheme for different algorithm settings. Reference recognition rate (horizontal line) is presented for ideal, manual phonesegmentation

ranges, and correspond to the best possible resolution of the event-function (18), reduced to 4 milliseconds ($\Delta t_K = 2^6/f_s$) in step (15). Detection threshold does not impact the accuracy of the segmentation in case of longer Δt_n periods (see Fig. 3).

The most important feature of the segmentation algorithm is to detect phone transitions properly. A number of false borders' detection (so called insertions) and a number of missing borders (deletions) for various algorithm settings were computed to measure this capability. Lower values, indicates better segmentation (see Fig. 4). For speech recognition purposes, a good "insertions vs deletions" compromise should be chosen because those two factors have a major impact on the recognition performance. In this case $\Delta t_n = 20$ milliseconds and $\alpha \leq 1.5$ are suggested.

Phone classification was also performed to measure the possible impact of the segmentation settings on recognition performance. Recognition was based on k-Nearest Neighbour (k-NN, $k = 3$) with modified Itakura-Saito spectral distortion measure and a feature vector composed of eleven wavelet energy fractions within detected segment extracted from (15). The best possible phone recognition rate (PRR) was measured using reference hand-made segmentation. The phoneme classification was performed on databases of each of five speakers independently (18 088 phoneme templates) in a "Round Robin" fashion, to obtain the maximum number (18 088) of tests. No language modelling nor context data were used for classification to prevent from altering the results with a non-acoustical knowledge.

The best recognition (PRR $= 63\%$) was obtained for $\Delta t_n = 24$ milliseconds and $\alpha = 1.5$. Recognition performance for the reference segmentation was PRR$_{ref} = 78\%$. That means that presented method gained PRR$_{ref}$/PRR $= 81\%$ of the possible performance of this parameterisation/classification front-end with manual segmentation and phone labeling.

The results were very similar for just one speaker, which strongly suggests that the method is speaker independent.

5. Conclusions

Wavelet packet analysis was used as a powerful method for the detection of phoneme boundaries. The use of wavelet analysis turns out to be an effective tool in finding the boundaries between two phonemes. The use of non-uniform segmentation reduces the total number of segments to be processed by higher-level parts of ASR systems. The proposed method has several advantages over short-time Fourier analysis. The speech signal in nature is non-stationary, therefore wavelet packet transform can provide better frequency analysis. The effect is an important decrease in any word-level decoder search space and therefore significant computational cost reduction of the entire processing system. The speech recognition can also be more accurate thanks to phoneme segmentation. The proposed method was successfully verified through laboratory tests.

6. Acknowledgments

This work was supported by MNISW grant OR00001905.

7. References

[1] J. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, pp. 137–152, 2003.

[2] A. Alani and M. Deriche, "A novel approach to speech segmentation using the wavelet transform," *Proceedings of The Fifth International Symposium on Signal Processing and its Applications*, p. 127130, 1999.

[3] H.-M. W. S.-S. Cheng, "A sequential metric-based audio segmentation method via the bayesian information criterion," *Proceedings of 8th European Conference on Speech Communication and Technology, Eurospeech*, pp. 945–948, 2003.

[4] O. Farooq and S. Datta, "Wavelet based robust subband features for phoneme recognition," *IEE Proceedings: Vision, Image and Signal Processing*, vol. 151, no. 3, pp. 187–193, 2004.

[5] I. Daubechies, *Ten lectures on Wavelets*. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics, 1992.

[6] Y. Meyer, *Wavelets and applications*. Masson, 1991.

[7] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Magazine*, vol. 8, pp. 11–38, 1991.

[8] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 41, pp. 909–996, 1988.

[9] J. Gałka and M. Ziółko, "Mean best basis algorithm for wavelet speech parameterization," *Proceedings of Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Kyoto*, 2009.