

# Measures On Wavelet Segmentation of Speech

Michał Dyrek, Jakub Gałka, Bartosz Ziółko

**Abstract**—Speech segmentation is widely used in many speech applications. We propose a new wavelet-based extension of the typical spectrum-based non-uniform speech segmentation methods. The use of wavelets improves computation performance and provides easy and flexible adjusting of algorithm parameters. Segmentation accuracy measures are introduced and applied for evaluation as well.

**Keywords**—speech analysis, speech recognition, speech segmentation, wavelet transforms.

## I. INTRODUCTION

TWO general types of segmentation algorithms are distinguishable [1]. First one is model based approach, which bases usually on some kind of optimization and dynamic programming like Viterbi Decoding or Dynamic Time Warping. The second approach is a signal feature scanning [2].

First one finds optimal, from the model's point of view, placements of speech unit borders [3]. Speech unit is usually an allophone or phoneme. This method can be used only when the phonetic/acoustic model is previously prepared, what is usually quite complicated and time consuming process. Model-based segmentation may be time-consuming itself. This is a problem in real-time processing applications but not when annotating recorded speech for use in speech databases. In this case annotation is performed usually once, during database creation and is most important part of this process.

Segmentation method presented in this paper is of the second type. It bases on tracking of specific changes in temporary discrete wavelet spectrum of speech [4]-[6]. No training or acoustic knowledge is needed for segmentation.

Lack of fast and accurate methods of speech segmentation caused domination of uniform segmentation in speech applications [7].

The use of non-uniform segmentation reduces total number of segments to be processed by higher-level parts of ASR systems (usually HMM modeling). The effect is a radical decrease in Viterbi decoding search-space and computational cost. Furthermore no complicated HMM state duration

modeling is needed. Ten seconds of uniformly segmented speech signal typically produces about 750 overlapping frames of duration 20 ms each. Non-uniform segmentation reduces this number to approximately 100 non-uniform frames (average duration of segment in investigated speech corpus is about 100 ms). This is a significant difference in complex decoding process. Non-uniform segmentation may also cause some degradation in features quality, because of increased in-frame diversity of signal, thus decision of using non-uniform segmentation should be considered individually by a system designer.

## II. WAVELET SEGMENTATION ALGORITHM

Presented algorithm is based on assumption that phonemes are characterized by the quasi-stationary spectral properties. Boundaries between them should be marked by the rapid energy flows between frequency sub-bands. At these points large changes in the signal spectra shall occur. This is not always fulfilled. In case of plosive sounds or diphthongs spectral changes occur in the middle of phoneme and it will be split into two or more parts, which should be taken into account at the classification stage. The algorithm consists of following steps:

Wavelet decomposition of speech signal  $S(t)$ , using six-level, dyadic decomposition tree and discrete Meyer wavelet filters [4], [5], [8], [9]. Decomposition process produces seven vectors of wavelet coefficients (sub-bands):

$$\hat{S} = \{B_1, B_2, B_3, B_4, B_5, B_6, B_7\} \quad (1)$$

where  $B_1$  represents lowest frequency sub-band and  $B_7$  – the highest one. Wavelet decomposition causes that vectors have various lengths:  $B_7$  consist of  $L/2$  elements,  $B_6 - L/4, \dots, B_2$  and  $B_1 - L/64$  elements each ( $L$  – length of the input signal).

Calculating

$$P_i(t) = B_i^2(t) \quad (2)$$

power function  $P_i(t)$  for each sub-band  $i$ .

Reduction of the number of samples in various bands. Desired length is obtained by summing  $N$  adjacent samples. We decided to equalize all bands to the shortest bands' length ( $B_1$  and  $B_2$ ). Therefore length reduction factor  $N_i$  for  $i$ -th sub-band is given by

$$N_i = \begin{cases} 2^{i-2} & , i > 1 \\ 1 & , i = 1 \end{cases} \quad (3)$$

Calculating power envelopes  $\hat{P}_i(t)$  by smoothing  $P_i(t)$  with combination of running median and FIR filtering presented in Fig. 2 and originally proposed by Tukey [10]. This smoother

Manuscript received December 28, 2006; Revised received September 25, 2007. This work was supported by the Polish Government, Department of Science under Grant N516 00632/0684, years: 2007-2008.

M. Dyrek is with the Department of Computer Science, AGH University of Science and Technology, Kraków, Poland (e-mail: mdyrek@agh.edu.pl).

J. Gałka is with the Department of Electronics, AGH University of Science and Technology, Kraków, Poland (phone: +48-12-6173639; fax: +48-12-6332398; e-mail: jgalka@agh.edu.pl).

B. Ziółko is with the Department of Computer Science, University of York, York, United Kingdom (e-mail: bziolko@cs.york.ac.uk).

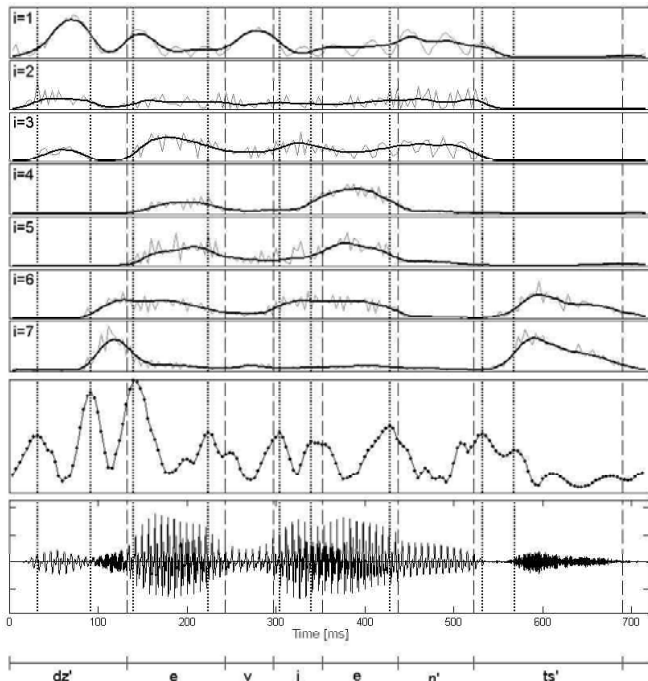


Fig. 1. An example of segmentation of polish digit „9 (dz'evjen'ts')". Dashed lines are reference segmentation boundaries; dotted lines are automatic segmentation boundaries. Upper plots depict power functions  $P_i(t)$  and envelopes  $P_i(t)$  for  $i=1, \dots, 7$ . Two lower plots present *rate-of-change* function  $D(t)$  and SAMPA annotated input signal  $s(t)$ .

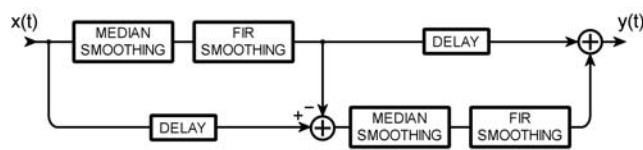


Fig. 2. Nonlinear smoothing system [10].

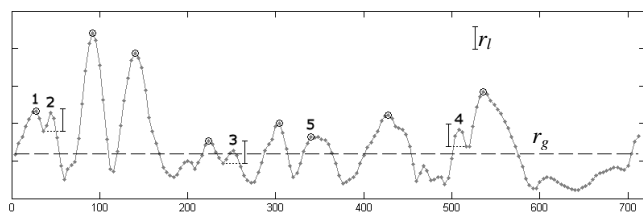


Fig. 3. *Rate-of-change* function  $D(t)$ . Maxima indicating phoneme boundaries are marked with circles. Peak (1) has no minimum in its left neighborhood and right min-max span is greater than  $r_l$ . Peaks (2), (3), and (4) are not fulfilling min-max span condition.

combines advantages of FIR filtering which well separates high (noise-like) frequency components and running medians, which preserves fast slopes – this is especially important in consideration of creating *rate-of-change* function.

Calculating *rate-of-change* function  $D(t)$  by summing absolute values of time derivatives of envelopes  $P_i(t)$  in every time point  $t$ . Function  $D(t)$  is the measure of wavelet-spectral variability of the signal. According to initial assumptions we expect it to be large at the boundaries of phonemes so

its peaks should indicate them. Because the sampling frequency of input signals is 16 kHz and all bands have now length 64 times smaller than input signal each sample of  $P_i(t)$  corresponds to 4 ms.

Picking the prominent peaks of *rate-of-change* function. As can be seen in Fig. 3  $D(t)$  function is quite rough so detection of peaks that indicates phonemes' boundaries is not straightforward. Algorithm based on observed features of  $D(t)$  function picks local maxima fulfilling three conditions [2]:

Local maximum value is greater than threshold defined as

$$r_g = \text{mean}(D) - \text{std}(D) \cdot g \quad (4)$$

where  $g$  is an adjustable parameter called *global sensitivity*.

No local minimum is present within left 12 ms neighborhood or otherwise there is neighboring maximum within next 12ms which value is smaller than considered maximum or local min-max span is greater than threshold  $r_l$  defined as

$$r_l = \text{std}(D) \cdot l \quad (5)$$

where  $l$  is also adjustable parameter called *local restriction*. Its increasing causes overall sensitivity decreasing.

No local minimum is present within right 12 ms neighborhood or there is neighboring maximum within next 12 ms which value is smaller than considered maximum or local min-max span is greater than threshold  $r_l$ .

Proposed wavelet segmentation algorithm is flexible because of using mean and standard deviation (*std*) values instead of fixed thresholds. Its sensitivity can be easily regulated with  $g$  and  $l$  in full range. Setting  $g = -\infty$ , causes that algorithm finds no boundaries. Values  $g = \infty$ ,  $l=0$  causes that all maxima of  $D(t)$  will be accepted.

Time resolution of the algorithm depends on 4-th step. In described case ( $f_s=16$  kHz, 6-level decomposition) it is 8 ms.

### III. PERFORMANCE EVALUATION

Defining an ideal and objective criterion for evaluating accuracy of speech segmentation is not possible because speech production is a continuous process and no phoneme borders can be pointed unambiguously at all. When segmentation is a part of the specific system (like speech recognition) the best measure of segmentation efficiency would be general performance of the system (word error/accuracy rate) [11]. Otherwise, it is hard to define numerical, not purpose-oriented segmentation quality measure. Usually some kind of reference segmentation has to be known, what is not easy to prepare, because of necessity of hand-annotating of big amounts of recordings.

*Average inaccuracy* measures mean percentage displacement of automatically detected borders against reference segmentation. Time distance from detected border to the reference border location is normalized by reference segment length. Mean value is then calculated over all detected borders. It is the most typical indicator of segmentation performance.

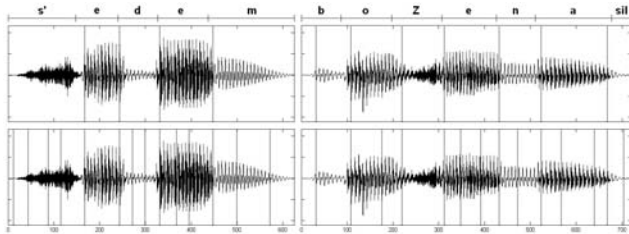


Fig. 4. Examples of segmentation of polish words: digit „7 (s`edem)” (left plots) and female name “Bożena (boZena)” (right plots) with medium ( $g=0$ ,  $l=1$ ; upper plots) and maximum ( $g \gg 1$ ,  $l=0$ ; lower plots) algorithm sensitivity.

When no border is detected in neighborhood of reference border then sub-segmentation error occurs (segment deletion). *Missed border rate* measures the relative number of such errors according to reference segments quantity [2].

Opposite situation occurs when too many changing points are mistakenly detected in over-segmentation process (segment insertion). In this case, relative *false border rate* measures number of inserted segments according to the total number of detected segments [2].

Tuning of algorithm parameters and evaluation of its accuracy was performed with a biggest Polish speech database “Corpora’97” [12]. Male speech of 28 speakers was used. Each speaker representation consists of 365 different recordings (names, short phrases and utterances). It has to be said, that only little part of recordings was segmented and annotated manually. Dynamic optimization algorithm was used for segmenting and annotating the rest [12].

Segmentation performance was examined using various values of  $g$  and  $l$  parameters in ranges chosen to cover most reasonable results. In Fig. 6 (left plot) dependence of segmentation inaccuracy on  $g$  and  $l$  parameters is presented. Impact of sensitivity is not significant therefore inaccuracy varies within 23% - 31% range only.

It is acceptable since speech corpus was automatically segmented and no guarantee is given for reference segmentation accuracy. No ideal reference segmentation may be defined. Such values of borders dislocation have no significant impact on segment quality when proper windowing method is then applied to each segment. *Missed border* and *false border* rates seem to be much more important indicators of segmentation efficiency [2]. Right plot in Fig. 6 visualizes dependency of these values. When general application (i.e. speech recognition) of segmentation is known, best desired ( $g$ ,  $l$ ) combination can be chosen. Results (Fig. 5, Fig. 6) show, that changing of sensitivity parameter values  $g$  does not affect border placement accuracy but rather *false* and *missed border* rates.

Similar results were obtained in other works as well, but methods used there were different and usually model-based [1], [2], [11], [13], [14].

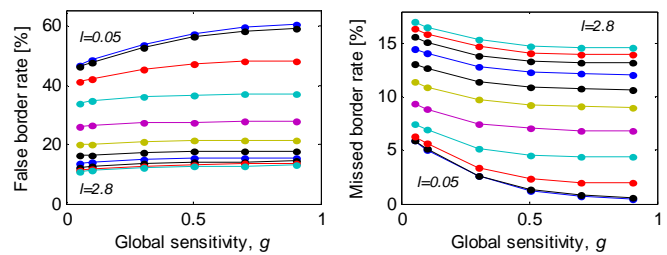


Fig. 5. *False* (left) and *missed* (right) border rate for various *local restriction* values  $l = (0.05, 0.1, 0.4, 0.7, 1, 1.3, 1.6, 1.9, 2.2, 2.5, 2.8)$ .

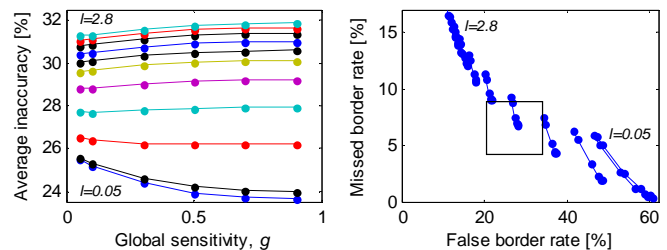


Fig. 6. Average inaccuracy (left) and missed vs false border rate dependency (right) for *local restriction* values  $l = (0.05, 0.1, 0.4, 0.7, 1, 1.3, 1.6, 1.9, 2.2, 2.5, 2.8)$  (curves) and *global sensitivity* values  $g = (0.05, 0.1, 0.3, 0.5, 0.7, 0.9)$  (dots). Best *local restriction* values ( $l=0.7$ ,  $l=1$ ) for non-uniform speech recognition were marked with square.

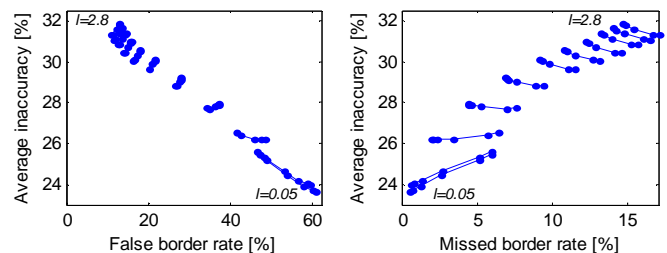


Fig. 7. Average inaccuracy dependency from *false* and *missed* border rates for *local restriction* values  $l = (0.05, 0.1, 0.4, 0.7, 1, 1.3, 1.6, 1.9, 2.2, 2.5, 2.8)$  (curves) and *global sensitivity* values  $g = (0.05, 0.1, 0.3, 0.5, 0.7, 0.9)$  (dots).

#### IV. CONCLUSION

The use of a big amount of speech data for evaluation could degrade the results. However, general performance is very good, comparing to other works. In consequence one can say that proposed algorithm is speaker independent and robust, what increases its universality.

Further works will concern impact of wavelet decomposition scheme on segmentation quality and use of this algorithm in non-uniform speech recognition system build with HTK. Robustness for noise and distortion of signal will be examined as well.

#### V. ACKNOWLEDGMENT

We would like to thank Stefan Grochowski and Institute of Computer Science, Poznań University of Technology for providing a corpus of spoken Polish - CORPORA’97.

## REFERENCES

- [1] K. Demuyne, and T. Laureys, "A comparison of different approaches to automatic speech segmentation", *Proceedings of the 5th International Conference on Text, Speech and Dialogue*, pp. 277-284, 2002.
- [2] S. Cheng, and H. Wang, "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, Geneva, 2003.
- [3] N. Beringer, and F. Schiel, "Independent automatic segmentation of speech by pronunciation modeling", *Proceedings of ICPhS*, San Francisco, 1999.
- [4] B. Ziólko, S. Manandhar, R. C. Wilson, and M. Ziólko, "Wavelet method of speech segmentation", *Proceedings of 14th European Signal Processing Conference - EUSIPCO*, Florence, 2006.
- [5] B. T. Tan, R. Lang, H. Schroder, A. Spray, and P. Dermody, "Applying wavelet analysis to speech segmentation and classification", *Proceedings of SPIE*, vol. 2242, pp. 750-761, 1994.
- [6] L. Janer, J. Marti, C. Nadeu, and E. Lleida-Solano, "Wavelet transforms for non-uniform speech recognition systems", *Proceedings of Fourth International Conference on Spoken Language, ICSLP 96*, vol. 4, pp. 2348-2351, Philadelphia, 1996.
- [7] Rabiner, L., and Juang, B., *Fundamentals of Speech Recognition*. Prentice-Hall Inc., 1993.
- [8] O. Farooq, and S. Datta, "Wavelet based robust subband features for phoneme recognition", *IEE Proceedings: Vision, Image and Signal Processing*, vol. 151(3), pp. 187-193, 2004.
- [9] J. N. Gowdy, and Z. Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '00*, vol. 3, pp. 1351-1354, Istanbul, 2000.
- [10] J. W. Tukey, "Nonlinear (Nonsuperposable) Methods for Smoothing Data", *Proceedings of EASCON'74*, pp. 673, 1974.
- [11] Z. Chengyi, and Y. Yonghong, "Fusion based speech segmentation in DARPA SPINE2 task", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP'04*, vol. 1, pp. I-885-8, 2004.
- [12] S. Grochowski, "First Database for Spoken Polish", *Proceedings of International Conference on Language Resources and Evaluation*, pp. 1059-1062, Grenada 1998.
- [13] G. Adami, and H. Hermansky, "Segmentation of speech for speaker and language recognition", *Proceedings of 8th European Conference on Speech Communication and Technology - EUROSPEECH*, pp. 841-844, Geneva, 2003.
- [14] J. A. Gómez, and M. J. Castro, "Automatic Segmentation of Speech at the Phonetic Level", *Lecture Notes In Computer Science*, vol. 2396, Springer-Verlag, pp. 672-680, London, 2002.



**Michał Dyrek** received the Master of Science and Engineering degrees in 2005 from Department of Electronics at AGH University of Science and Technology in Kraków, Poland where he continues PhD studies started in the same year. Currently he works for Motorola Software Group Poland as a software engineer. His scientific interests concentrate in the areas of digital signal processing, speech recognition, neural networks and machine learning.



**Jakub Gałka** was born in Kraków, Poland in 1979. He obtained his Master of Science and Engineering degrees from Department of Telecommunications at AGH University of Technology in Kraków, where he continued his doctoral studies since 2003. He is a scientific assistant and academic teacher in Department of Electronics at AGH University of Technology. He contributed to various research programs related to digital signal processing, speech processing and recognition.



**Bartosz Ziólko** received the MSc and MEng in Electronics and Telecommunications from AGH University of Science and Technology, Kraków, Poland in 2004. He worked at Cambridge Broadband Ltd which designs modern wireless access systems in 2002. In 2003 he studied at the Tampere University of Technology as an exchange student. From 2004 to 2005 he worked as a Research Associate at the AGH University of Science and Technology. In 2005 he started PhD studies at University of York. He has published some 20 papers in journals and refereed conferences. His research interests are in speech recognition.