Study of Performance Evaluation Methods for Non-Uniform Speech Segmentation

Jakub Gałka, Bartosz Ziółko

***Abstract*—**Speech segmentation is a very difficult problem, because of continuous nature of speech. Segmenting speech into various units (phonemes, syllables, and acoustic atoms) is essential in many applications. Choosing the best method of segmentation must be preceded by evaluation of its performance. This paper is a study of various numerical measures for automatic segmentation performance.

***Keywords*—**performance evaluation, performance measures, speech segmentation, wavelet transforms.

I. Introduction

SEGMENTATION is a task of splitting continuous objects into meaningful units. For a speech input, these units are phonemes, words or sentences. The segmentation problem can be viewed as an unlabelled splitting problem where the input sequence needs to be split into subsequences. Methods for evaluating the accuracy of the segmentation problem tend to be ad-hoc with researchers using their own methods such as allowing 10% overlap between the predicted segmented and the ground truth. Since non-uniform segmentation is getting more popular some evaluation methods should be defined for common and general use.

In the vast majority of approaches to speech recognition, the speech signals need to be divided into segments before recognition can take place. The properties of the signal contained in each segment are then assumed to be constant, or in other words to be characteristic of a single part of speech. Other levels of speech segmentation are often conducted as a part of further analysis in speech recognition systems.

Speech segmentation can be used for a number of different tasks. Often it is used for word segmentation. This can be done by Viterbi and forward-backward segmentation [1]. Another applied method is based on mean and variance of spectral entropy [2]. A different problem covered by the same name is separating silence and speech from an audio recording [3]. The method uses so called TRAPS-based segmentation and Gaussian mixture based segmentation. Segmentation here means mainly removing non-speech events and additionally clustering according to speaker identities, environmental and channel conditions. Yet another possible segmentation is by phonetic features (not necessarily phonemes) [4], by applying wavelet analysis. There is also research on syllable segmentation [5]. Another meaning is segmenting due to partially correct transcriptions [6]. In this case segmentation is combined with recognition. We can also understand segmentation as the process of breaking speech into phonemes [7], [8].

In many applications, like speech recognition, the most common segmentation method is to use constant-time framing, for example into 25 ms blocks [9]. This method benefits from simplicity of implementation and the ease of comparing blocks of the same length. However, the different length of phonemes is a natural phenomenon which cannot be ignored. Moreover, boundary effects provide additional distortion. A more satisfactory approach is an attempt to find the phoneme boundaries. A number of approaches have been previously suggested for this task. Segmentation can be conducted by filter bank energy contours analysis [7]. Neural networks [10] have also been tested, but they require time consuming training. Segmentation can be applied by the segment models instead of the hidden Markov models (HMM) [11]. Partitioning is based upon the model decode. It allows boundaries to be located only on several fixed positions dependent on framing (on multiplied length of one frame). The analysis of the first derivative of power in different frequency sub-bands gives another opportunity to distinguish phoneme boundaries [8]. Many phonemes exhibit rapid changes in particular sub-bands which can determine their beginnings and endpoints. The typical approach to phoneme segmentation for creating speech corpora is to apply dynamic programming for time alignment [12]. This method is very accurate but demands transcription and hand segmentation of some utterances to start with.

II. Wavelet segmentation scheme

Presented wavelet segmentation algorithm (Fig. 1) bases on detection of huge rapid energy transitions among different wavelet sub-bands. We used 6-level dyadic decomposition scheme and discrete Meyer decomposition wavelet filters [4], [8], [13], [14]. First step of processing insists on calculating temporary power of discrete wavelet spectrum (DWT). All power signals are then smoothed using improved Tukey's

Manuscript received December 28, 2006: Revised received September 24, 2007. This work was supported by the Polish Government, Department of Science under Grant N516 00632/0684, years: 2007-2008.

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).

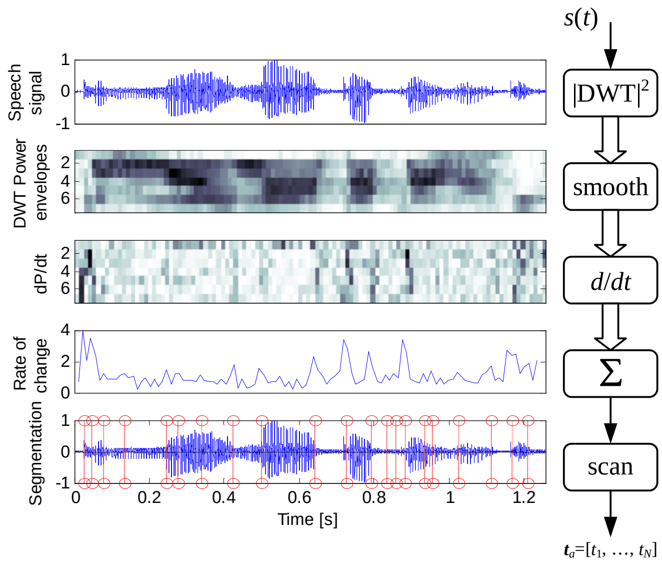

Fig. 1 Intermediate signals and wavelet segmentation of an utterance *'Tym można atakować'* (left) and algorithm simplified scheme (right).

non-linear smoothing algorithm [15]. Time-derivatives of power envelopes are calculated and aggregated by summing operation. Obtained rate-of-change function is then scanned to find the prominent peaks using algorithm similar to one presented in [16]. Two algorithm parameters (g, l) may be set to adjust global sensitivity and local restriction of peak detector. Changing them alters the segmentation accuracy and evaluation results. To illustrate the evaluation process parameters vary within specified ranges: g=[0.05, 0.1, 0.2, …, 0.9] (dots in figures) and l=[0.5, 0.1, 0.3, 0.6, …, 3] (lines in figures) in this work. Increasing g and/or decreasing l forces algorithm to find more segments.

III. Performance Evaluation Methods

Three general classes of performance measures will be presented in this paper. Each of them is suitable for different purposes. Their goal is to give rate of how precise and apposite the segmentation results are.

We present results obtained with non-uniform wavelet segmentation on the Polish speech database *Corpora*'97 [17]. Evaluation was performed on the set of male 5 speakers of 356 utterances each, what gives 1825 utterances (25906 reference phonemes in total). For comparison uniform segmentation has been performed for various frame lengths (w = [5, 10, 20, …, 100] milliseconds, depicted by crossed line plots in most figures).

A. Subjective approach

Working with human-related signals, like speech, gives an opportunity for using our senses to investigate and quote the results of work. However, this may be misleading, since resolution and capacity of aural and visual apparatus are limited [18].

Aural inspection of speech segmentation is possible, but not efficient. Human is able to distinguish two acoustic events separated at least by 20 ms time-gap. Therefore, no short acoustic units like plosives, stops, may be precisely distinguished and located in signal.

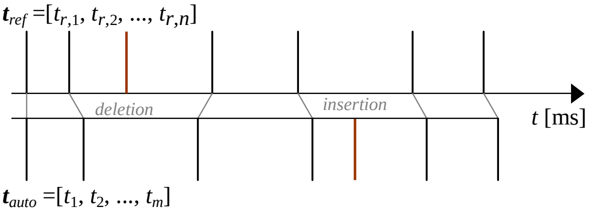
t_ref =[t_r,1, t_r,2, ..., t_r,n]
deletion
insertion
t [ms]
t_auto =[t_1, t_2, ..., t_m]

Fig. 2 Reference/automatic segmentation model. Missing border (deletion) and false alarm (insertion) cases have been illustrated.

Visual inspection of the signal gives an opportunity to time-freeze the speech. This approach utilizes time-scope of the waveform or its spectrogram. Short-time Fourier-based transforms (S-TFT, DCT), and many types of features (MFCC, LPC, PLP, Wavelet spectrum, Entropy, Zero-crossing rate) are most commonly used. Unfortunately time-spectrum (of any kind) not always explicitly represents nuances of the speech and proper training of examiner is necessary.

Time consumption of examining the segmented recordings is the important problem. It is impossible to evaluate the segmentation performance over a whole speech corpus (like TIMIT or AURORA) by a single man, thus obtained evaluation results are never precise and satisfactory. This kind of measuring of the performance is useful in introductory and system-designing works only.

B. Purpose-oriented measures

Speech segmentation is usually a part of a complex speech recognition, annotation, synthesis, compression or processing system. Result of the system activity can be measured with usually well defined criterions like:

- Word recognition ratio
- Word error rate (WERR)
- Phone recognition ratio
- Phone error rate (PERR)
- Signal-to-noise ratio (SNR)
- Compression ratio
- and others…

This approach is widely used because results can be easily compared and are in fact most important criterions of efficiency, when segmentation is component of the system. Results obtained in this way are comparable and strongly advisable in researcher's practice.

C. Objective, numerical measures

In this chapter most important objective and numerical criterions will be presented. Most of them rely on the reference segmentation which is usually difficult to obtain. Because of continuous nature of speech production, no true segmentation may be defined at all for some phoneme boundaries (diphthongs). This is the main problem in creating reference segmentation (by hand and/or by using optimization algorithms) and using it for evaluation of automatic segmentation algorithms. However, proposed methods still have many important advantages. They are computationally efficient and easily applicable. Their major advantage is ability of processing big amounts of data for obtaining comparable results, what is not possible with supervision. Described methods provide also possibility of verifying robustness of segmentation algorithm in wide range of signal conditions and when no purpose-oriented approach may be used. Objective measures may provide strong, useful and reliable evaluation tool, when accompanied by other methods, what is described in section III.

Most of presented methods are statistical, what is caused by stochastic nature of the measured object and usually big variable amounts of data. In most cases typical apparatus of statistical mathematics is being used (mean value, deviation, histograms or probability density functions). Each result may be presented in absolute (usually milliseconds) or relative (percentage) form.

Reference speech segmentation of signal $s(t)$ is defined by vector

$$\boldsymbol{t}_{ref} = [t_{r,1}, t_{r,2}, ..., t_{r,N}] \qquad (1)$$

of N segment border's time-marks t_{rn}.

Automatic segmentation is defined in similar way by vector

$$\boldsymbol{t}_{auto} = [t_1, t_2, ..., t_M] \qquad (2)$$

of M automatically derived segment borders. Both vectors are usually of different lengths and finding a corresponding borders is necessary at the beginning.

Border accuracy

Most intuitive approaches base on measuring what is the difference in the localization of the detected and reference borders. This is defined by

$$\delta_m = t_{r,n} - t_m \qquad (5)$$

and should be taken into account only when proper correspondence of the borders is set (i. e. within the defined neighborhood). This value may be normalized with the reference segment length to obtain *relative accuracy*

$$\delta'_m = \frac{\delta_m}{t_{r,n+1} - t_{r,n}}, \qquad (6)$$

however, results presented in absolute form (milliseconds) are very intuitive and well comparable.

Mean accuracy value

$$\bar{\delta} = \frac{\sum_{m=1}^{M} \delta_m}{M}, \qquad (7)$$

where M stands for the number of examined borders, may be computed as well. This measure is reliable as far as the reference segmentation is. For accuracy value less then few milliseconds it loses its meaning because reference data are not usually prepared so precisely.

Interval accuracy (*refinement*) is strongly related to the accuracy measure. It is defined as a percentage (or absolute number) of boundaries which accuracy varies within specified ranges. Histogram of accuracy values (Fig. 4) provides lot of information on the segmentation algorithm properties and characteristics and was widely used in many works [20]. It provides reliable and easily comparable results. This is one of the predominatingly used and important approaches to segmentation evaluation.

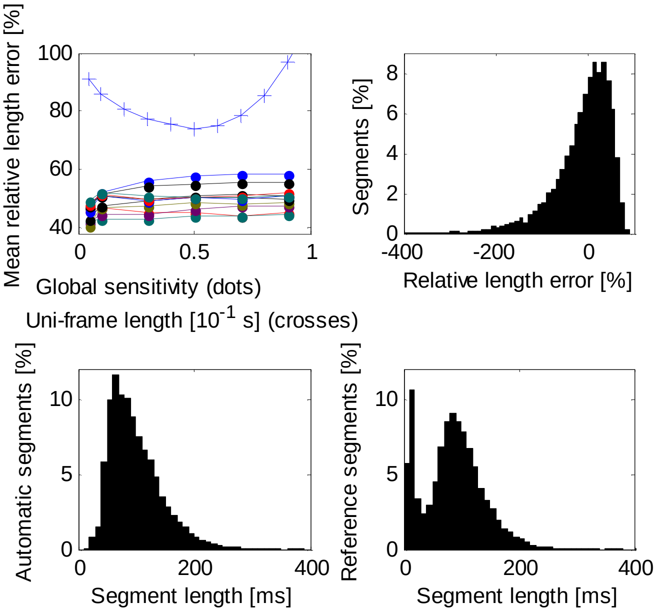

Fig. 3 Mean relative length error for different (*g*,*l*) values of segmentation algorithm ('*g*' – dots, '*l*' - lines) and various uniform segmentation duration (crosses). Histograms of relative error and segment durations for *g*=1, *l*=0.4 prepared in 10% and 10 ms bins respectively. Local minimum of the uniform segmentation length error corresponds with the half of the average length of reference segmentation (50 m).

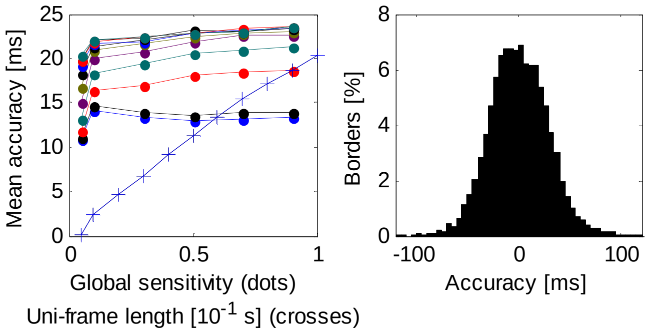

Fig. 4 Mean accuracy of automatic (dots) and uniform (crosses) borders (left) and histogram of border accuracy for *g*=1, *l*=0.4 in 5 ms bins (right).

Fuzzy Recall and Precision

Fuzzy logic is a tool for embedding structured human knowledge into workable algorithms. In a narrow sense, fuzzy logic is considered a logical system aimed at providing a model for modes of human reasoning that are approximate rather than exact. In a wider sense, it is treated as a fuzzy set theory of classes with unsharp boundaries [21]. Fuzzy logic found many applications in artificial intelligence due to the

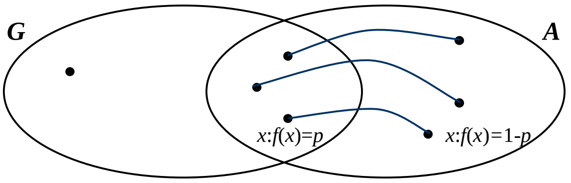

Fig. 5 The general scheme of sets: ***G*** with reference boundaries and ***A*** with detected ones. Elements of ***A*** have a grade *f*(*x*) standing for probability of being a correct boundary. In a set ***G*** there can be elements which were not detected (in the very left part of the plot).

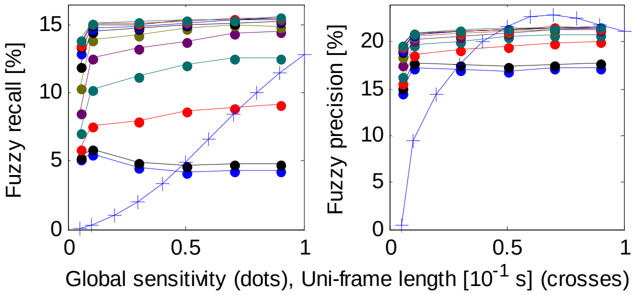

Fig. 6 Fuzzy recall and precision of the automatic (dots) and uniform (crosses) segmentation.

introduction of the opportunity of numerical and symbolic processing of a humanlike knowledge. This kind of processing is needed in proper evaluating of many types of segmentation. In our case we are interested in phoneme boundary location. Detected boundaries may be shifted more or less with respect to a reference segmentation. This ’more or less’ makes a crucial difference and cannot be mathematically described in a boolean logic.

Fuzzy recall and *precision* segmentation evaluation method [22] is based on the well-known recall and precision evaluation method [23]. However, in [22] approach, calculated boundary locations are elements of a fuzzy set and a binary operation T-norm describes their memberships. T-norm is defined as a function T : [0; 1] × [0; 1] → [0; 1] which satisfies commutativity, monotonicity, associativity and for which 1 acts as an identity element. As usual in recall and precision, one set contains relevant elements. The other is the set of retrieved boundaries. An evaluation grade using the number of elements in each of them and in their intersection was calculated. The comparison of the number of relevant boundaries and a number of elements in intersection gives precision. In a boolean version of the evaluation method it is information about how many correct boundaries were found. In [22] it is evaluated not only how many boundaries were detected but how accurately they were detected by incorporating fuzzy logic. The comparison of the number of retrieved elements and intersection gives recall, which is a grade of wrong detections. In this case fuzzy logic allows to evaluate not only a number of wrong detections but also their incorrectness. Each retrieved boundary has a probability factor which represents being correct information.

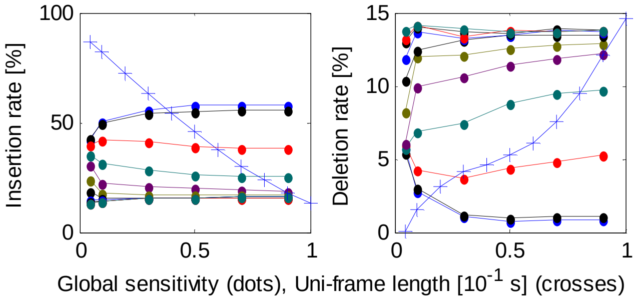

Fig. 7 Segment insertion and deletion rate of the automatic (dots) and uniform (crosses) segmentation.

Insertions and deletions

All previous methods concern automatically obtained segments for which reference could be found. When no corresponding reference segments can be found, segmentation errors of different nature take place. This situation usually occurs when to many borders have been automatically detected. This is over-segmentation which is measured by segments’ *insertion rate* (*false border rate, false alarm rate*).

While processing big amounts of data, absolute number of insertions may be difficult for interpretation, thus *relative insertion rate* defined as

$$\lambda = \frac{number\ \ of\ \ insertions}{number\ \ of\ \ automatically\ obtained\ \ segments} \qquad (7)$$

is advisable. This value gives information on what percentage of automatically obtained segments is a false alarm (Fig. 7). Increasing algorithm sensitivity usually increases insertion rate.

Deletions occur when no automatically derived border can be assigned with a reference border. This is sub-segmentation measured with *deletion rate* (*missed border rate*). This is very important criterion because missing a segment causes huge changes in spectral properties of segment and usually a big decrease in general system performance (recognition, compression). Insertions errors are much easier to handle with.

Relative deletion rate is a good way of presenting the result. It is defined as

$$\mu = \frac{number\ \ of\ \ deletions}{number\ \ of\ \ reference\ segments} \qquad (8)$$

and reports what percentage of reference borders had not been detected in automatic segmentation process (Fig. 7).

Correct detection rate

$$\gamma = \frac{number\ of\ \ correct\ detections}{number\ of\ \ automatically\ obtained\ \ segments} \qquad (9)$$

can be defined in similar way.

These methods give information on sensitivity of the segmentation algorithm and are very important in applications and therefore for general performance evaluation. Relative rates are useful when comparing the algorithm evaluated on different sets of data or referencing to other’s works. Besides *interval accuracy*, methods presented in this subsection are most widely used [16], [24], [25].

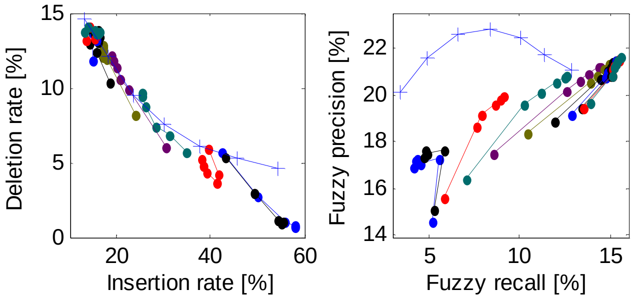

Fig. 8 Plots of deletions vs insertions (left) and Fuzzy precision vs recall (right).

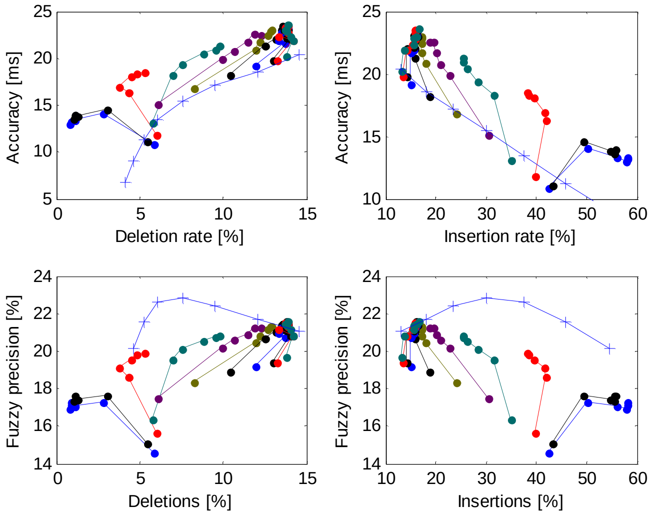

Fig. 9 Hybrid plots of different evaluation measures provide insight how uniform segmentation (crosses) differs from automatic non-uniform segmentation (dots). Properties of different evaluation methods may be compared.

Other segmentation measures

Many another approaches can be used for evaluation, if different, specific properties of an algorithm have to be examined.

Use of a various acoustic distance measures (Itakura-Saito, Log-Likelihood) can lead to interesting results, especially when only non-reliable reference segmentation is known. It is usually used for total inter-segment distance calculation. The greater inter-segment distance, the better is the segmentation result.

Statistical significance test like: χ^2, ANOVA or *t-Student*'s have been used with success in some works [26], [27].

Image segmentation has been intensively studied for many years, and several 2D evaluation methods are known. Properly modified can be used for evaluation of speech segmentation. Some suitable methods are: *scalable discrepancy*, *distortion rates*, *Baddeley distance*, and others [28]-[30].

IV. Conclusion

Methods presented in section III may be amended by simple operations. Plotting one measure result versus another provides opportunity to judge some general features of segmentation algorithm. Because of importance of the *insertion* and *deletion* rates combined plot (Fig. 8, left) is very important way of quoting the results. Plots in Fig. 9 not only gives information on performance of segmentation but also allow comparison of different evaluation methods. One can compare results of accuracy and fuzzy precision obtained with the same data. Significant difference is noticeable for uniform segmentation case within each column in Fig. 9. Various performance rates can be combined to obtain overall performance rate. Arithmetic product of chosen rates is the most straightforward solution.

Among all presented approaches few are the most important in general use. Purpose oriented measures (WERR, PERR, WRR, PRR, SNR) should be always used if possible. When no such information is available simultaneous use of insertion, deletion and accuracy or fuzzy rates is the best option [25].

Acknowledgment

This work was supported by the Polish Government, Department of Science under Grant N516 00632/0684, years: 2007-2008.

References

[1] Demuynck, K. and T. Laureys, 2002. A comparison of different approaches to automatic speech segmentation. Proceedings of the 5th International Conference on Text, Speech and Dialogue:277–284.

[2] Subramanya, A., J. Bilmes, and C. P. Chen, 2005. Focused word segmentation for ASR. Proceedings of Interspeech 2005, pp. 393–396.

[3] Zheng, C. and Y. Yan, 2004. Fusion based speech segmentation in DARPA SPINE2 task. Proceedings of ICASSP 2004:I–885–888.

[4] Tan, B. T., R. Lang, H. Schroder, A. Spray, and P. Dermody, 1994. Applying wavelet analysis to speech segmentation and classification. H. H. Szu, editor, Wavelet Applications, volume Proc. SPIE 2242:750–761.

[5] Villing, R., J. Timoney, T. Ward, and J. Costello, 2004. Automatic blind syllable segmentation for continuous speech. Proceedings of ISSC 2004, Belfast.

[6] Cardinal, P., G. Boulianne, and M. Comeau. Segmentation of recordings based on partial transcriptions. Proceedings of Interspeech 2005:3345–3348.

[7] Grayden, D. B. and M. S. Scordilis, 1994. Phonemic segmentation of fluent speech. Proceedings of ICASSP:73– 76.

[8] Ziółko, B., S. Manandhar, R. C. Wilson, and M. Zi´ołko, 2006. Wavelet method of speech segmentation. Proceedings of 14th European Signal Processing Conference EUSIPCO.

[9] Young, S., 1996. Large vocabulary continuous speech recognition: a review. IEEE Signal Processing Magazine, 13(5):45–57.

[10] Suh, Y. and Y. Lee, 1996. Phoneme segmentation of continuous speech using multi-layer perceptron. In Proceedings of ICSLP.

[11] Ostendorf, M., V. V. Digalakis, and O. A. Kimball, 1996. From HMM's to segment models: A unified view of stochastic modeling for speech recognition. IEEE Transactions on Speech and Audio Processing, 4:360–378.

[12] Holmes, J. N., 2001. Speech Synthesis and Recognition. London: Taylor and Francis.

[13] 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.

[14] 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.

[15] J. W. Tukey, “Nonlinear (Nonsuperposable) Methods for Smoothing Data”, Proceedings of EASCON’74, pp. 673, 1974.

[16] 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

[17] S. Grocholewski, “First Database for Spoken Polish”, Proceedings of International Conference on Language Resources and Evaluation, pp. 1059-1062, Grenada, 1998.

[18] Rabiner, L., and Juang, B., Fundamentals of Speech Recognition. Prentice-Hall Inc., 1993.

[19] Moe Pwint, Student Member, IEEE and Farook Sattar, Member, IEEE, A SEGMENTATION METHOD FOR NOISY SPEECH USING GENETIC ALGORITHM

[20] Seung Seop Park and Nam Soo Kim, Automatic Speech Segmentation Based on Boundary-Type Candidate Selection, IEEE SIGNAL PROCESSING LETTERS, VOL. 13, NO. 10, OCTOBER 2006 pp 640-643

[21] Kecman, V., 2001. Learning and Soft Computing. US: Massachusetts Institute of Technology.

[22] Bartosz Ziółko , Suresh Manandhar , Richard C. Wilson, Fuzzy Recall and Precision for Speech Segmentation Evaluation, LTC 07

[23] van Rijsbergen, C. J., 1979. Inforamtion Retrieval. London: Butterworths.

[24] Shih-sian Cheng and Hsin-min Wang, METRIC-SEQDAC: A Hybrid Approach for Audio Segmentation

[25] 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.

[26] James M. McQueen, Segmentation of Continuous Speech Using Phonotactics, JOURNAL OF MEMORY AND LANGUAGE 39, 21–46 (1998).

[27] Morten H. Christiansen & Joseph Allen, Coping with Variation in Speech Segmentation

[28] Y. J. Zhang and J. J. Gerbrands, Objective and quantitative seg,mentation evaluation and comparison, Signal Processing, vol. 39, pp. 43-54, 1994

[29] S. Philipp-Foliguet, Evaluation de la segmentation, Rapport Technique, 2001.

[30] Baddeley, D.L. Wilson and R.A. Owens, A New Metric for grey Scale Image Comparison, International Journal of computer Vision, vol. 24, 1995, pp. 5-17.

[31] Brian Ekman and William Goldenthal, Time-Based Clustering for Phonetic Segmentation pp.1225-1228

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 Ziółko 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.