Int. J. Appl. Math. Comput. Sci., 2014, Vol. 24, No. 2, 259–270
DOI: 10.2478/amcs-2014-0019

AUTOMATIC SPEECH SIGNAL SEGMENTATION BASED ON THE INNOVATION ADAPTIVE FILTER

RYSZARD MAKOWSKI, ROBERT HOSSA

Faculty of Electronics
Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: {ryszard.makowski,robert.hossa}@pwr.edu.pl

Speech segmentation is an essential stage in designing automatic speech recognition systems and one can find several algorithms proposed in the literature. It is a difficult problem, as speech is immensely variable. The aim of the authors' studies was to design an algorithm that could be employed at the stage of automatic speech recognition. This would make it possible to avoid some problems related to speech signal parametrization. Posing the problem in such a way requires the algorithm to be capable of working in real time. The only such algorithm was proposed by Tyagi *et al.*, (2006), and it is a modified version of Brandt's algorithm. The article presents a new algorithm for unsupervised automatic speech signal segmentation. It performs segmentation without access to information about the phonetic content of the utterances, relying exclusively on second-order statistics of a speech signal. The starting point for the proposed method is time-varying Schur coefficients of an innovation adaptive filter. The Schur algorithm is known to be fast, precise, stable and capable of rapidly tracking changes in second order signal statistics. A transfer from one phoneme to another in the speech signal always indicates a change in signal statistics caused by vocal track changes. In order to allow for the properties of human hearing, detection of inter-phoneme boundaries is performed based on statistics defined on the mel spectrum determined from the reflection coefficients. The paper presents the structure of the algorithm, defines its properties, lists parameter values, describes detection efficiency results, and compares them with those for another algorithm. The obtained segmentation results, are satisfactory.

Keywords: automatic speech segmentation, inter-phoneme boundaries, Schur adaptive filtering, detection threshold determination.

1. Introduction

Segmentation is the process of dividing a speech signal into discrete, non-overlapping fragments. This is usually a division into units of speech such as sentences, words, syllables, phonemes or even smaller phonetic units. When it comes to recordings containing numerous speakers' utterances, segmentation can consist in attributing pieces of utterances to particular speakers. The term *segmentation* is also sometimes used to refer to a division of the speech signal into frames before its parametrisation. The frames do not need to have the same length. Usages of segmentation include speech analysis and synthesis, speech quality improvement, speaker recognition or Automatic Speech Recognition (ASR). As for ASR systems, segmentation can be performed (i) at the system training stage, when segmentation is applied to the training set recordings, or (ii) at the recognition stage. In the former case, segmentation can be manual, automatic or semi-automatic, i.e., with manual correction of decisions prepared automatically. For recognition-stage segmentation, it must be performed automatically.

Regardless of the method used, speech segmentation is an important task, but it is a serious challenge to the executors, too. As for manual segmentation, it is a very laborious task, as designing high-quality speech recognition systems requires a huge amount of training data. In such a situation, it is difficult to avoid errors or inaccuracies, as manual segmentation is encumbered with subjectivity. At the same time, recognizing boundaries between phonetic units is not obvious, as changes in the geometry of the voice tract shaping the speech signal are fluid. This applies to a lesser extent to transitions to fricative and affricate phonemes. Moreover, sometimes a transition between phonemes takes place at different moments for different frequencies, which is connected with different inertness of different parts of the speech

tract. Inside fricative and affricate phonemes, signal variance may strongly fluctuate. Basically, only vowels, semivowels and nasal consonants have a stable form. Additionally, speech can contain, e.g., breaths or tongue clicks. Some phonemes may consist of spectrally different sounds or may be preceded by vocal cord activation. Moreover, speech changeability is immense. All this makes automatic segmentation a complicated task, and the literature does not provide fully satisfactory solutions.

The algorithms known from the literature could be divided into two basic groups: (i) those using information about the phonetic content (supervised segmentation) and (ii) those performing segmentation regardless of the phonetic content (unsupervised segmentation). Algorithms in the first group require signal parameterisation before they start recognition. If a phoneme sequence is known, the task consists in matching a set of observation vectors, resulting from parametrisation, to that sequence. These algorithms require at least the presence of a model composed of an acoustic speech model and a model of interphonemic transitions. Thus, these algorithms are computationally consuming. This group includes algorithms using recognition with Hidden Markov Models (HMMs) (e.g., Brugera, 1993; Tolenado, 2003; Mporas, 2008), Dynamic Time Warping (DTW) (e.g., Gomez, 2011) or Artificial Neural Networks (ANNs) (e.g., Schwarz *et al.*, 2006). The algorithms of this group are employed only in the ASR system training stage.

The other automatic segmentation group consists of algorithms which do not require any knowledge about the phonetic content and are based mostly on statistical signal analysis (e.g., Tyagi *et al.* 2006; Almanidis *et al.*, 2008, 2009; Scharenborg *et al.*, 2010; Rudoy *et al.*, 2011). Tyagi *et al.* (2006) employed speech signal modelling based on the autoregressive process. This is a variant of the Brandt algorithm (Brandt, 1983). It starts with defining prediction filter coefficients for three frames: $x_0(n) = [t - N_0, \ldots, t - 1]$, $x_1(n) = [t, \ldots, t + N_1 - 1]$ and $x(n) = [t - N_0, \ldots, t + N_1 - 1]$, so $x(n)$ is a joint of $x_0(n)$ and $x_1(n)$. For these frames, predictive filtering error variances are defined. These variances are a basis for formulating a Generalized Likelihood Ratio Test (GLRT) related to a change in the signal power spectrum density over the analysed period of time. Almanidis *et al.* (2008; 2009) presented a hybrid algorithm known as the Model Selection Criterion (MSC), using a Bayesian Information Criterion (BIC) and Kullback–Leibler information. This method requires MFCC parameterisation of the speech signal, constructing a model of speech signal segments and performing boundary detection. The whole is referred to as the DISTBIC algorithm (Delacourt *et al.* 2000). In the work by Scharenborg *et al.* (2010), the first step is also MFCC parameterisation, followed by segmentation based on the principle of finding the maximum distances between observation vectors in the selected subset, performed with the use of a method known as Maximum Margin Clustering (MMC). Finally, Rudoy *et al.* (2011) used stochastic modelling employing the standard AutoRegressive (AR) and Time-Varying AutoRegressive (TVAR) models. Detection is performed with the use of a classic GLRT test based on determining which model is more adequate for a particular segment of the signal. Except for the solutions presented by Tyagi *et al.* (2006) and Rudoy *et al.* (2011), the above-mentioned algorithms are computationally complicated.

The efficiency of defining interphonemic transitions by means of supervised algorithms, with the assumed tolerance of up to 20 ms, reaches 93%, while for unsupervised algorithms it is up to 75%, with the false acceptance error of 2% (Almanidis *et al.*, 2008; 2009). This advantage of supervised algorithms is a consequence of using information about the phonetic content.

The present article focuses on segmentation of speech into phonemes performed to the use in ASR systems. It proposes a segmentation algorithm whose starting point is estimation of the reflection coefficients with the use of an adaptive Schur algorithm. This algorithm is characterized by rapid adaptation to the changing signal, stability and robustness (Lee *et al.*, 1981; Lopatka *et al.*, 2005; 2006; Makowski and Zimroz, 2013). Based on Schur coefficients, a parametric time-varying spectrum is defined and subsequently converted to time-varying signal powers in subbands with the use of mel frequency scale. Based on these powers, a GLRT test is formulated. It is commonly acknowledged that a mel or bark spectrum is the signal statistic on which speech recognition performed by the human hearing organ and the brain is based. Detection of boundaries between speech sounds (phonemes and other sounds) performed by the proposed algorithm is based exclusively on the signal. Besides, this algorithm is relatively uncomplicated computationally and it can operate in real time. Thus, it can be used in ASR systems at the recognition stage.

Any definite and relevant information about automatically recognized speech signal is valuable. Even an incomplete division into phonemes would allow, e.g., avoiding problems related to estimation of the observation vector. One problem connected with such estimation is incorrect recognition of frames containing pieces of two adjacent phonemes, particularly in transitions from affricates to most other phonemes.

Section 2 of this work provides a description of the proposed algorithm, an analysis of its properties as well as a general plan of boundary processing and detection. Section 3 contains a description of the inter-phoneme boundary detection procedure allowing for speech signal properties. Section 4 discusses the values of the employed processing parameters, the defined measures of detection effectiveness, as well as the results of automatic

segmentation performed by using the proposed algorithm obtained for a recording corpus for the Polish language, also comprising manual segmentation of utterances. The obtained results are compared with those achieved by using a modified algorithm by Tyagi *et al.* (2006). Finally, Section 5 summarizes the main points of the work.

2. Description of the proposed algorithm

2.1. Innovation adaptive filter. A method of modelling the speech signal with the use of autoregressive process, i.e., describing it with a predictive filter coefficients or reflection coefficients, has been widely used, e.g., by Rabiner and Yuang (1993). In the present work, such modelling will be employed for detecting changes in the speech signal spectrum and subsequently for its segmentation. The speech signal is non-stationary, although its local stationarity is assumed. It will be analysed with the use of the innovation adaptive filter (Schur filter). For every discrete time instant t, the Schur reflection coefficients describing the autoregressive process will be estimated optimally in a mean square sense. This means that the filter is capable of following changes in the second order statistics (the autocovariance function in this case) of the analysed signal. The ladder realization of the innovation adaptive filter is shown in Fig. 1 (Lee *et al.*, 1981; Lopatka *et al.*, 2005; 2006).

The innovation filter is composed of P sections. Each section is completely described by the time-varying Schur coefficient $\rho(n,t), n = 1,\ldots,P$. The inputs of each section are the subsequent samples of the forward $e(n,t)$ and backward $r(n,t)$ prediction error signals, and for the first section these are normalized signal samples. The Schur coefficients are updated every time instant t. This updating is in fact a procedure of minimizing the mean-square prediction error. All the computations are summarized in the following three equations based on the recursive orthogonalization principle (Lee *et al.*, 1981; Lopatka *et al.*, 2005; 2006):

$$\begin{aligned}&\rho(n+1,t)\\&\quad=\rho(n+1,t-1)\sqrt{1-e^2(n,t)}\sqrt{1-r^2(n,t-1)}\\&\qquad-e(n,t)r(n,t-1),\end{aligned}$$

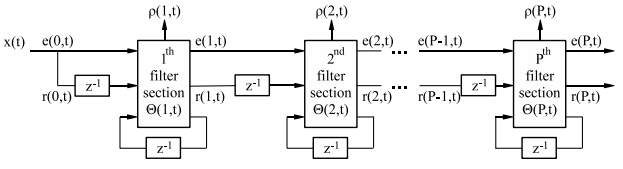


Fig. 1. Ladder-form realisation of the adaptive innovation filter.

$$e(n+1,t)=\frac{e(n,t)+\rho(n+1,t)r(n,t-1)}{\sqrt{(1-\rho^2(n+1,t))}\sqrt{1-r^2(n,t-1)}},\tag{1}$$

$$r(n+1,t)=\frac{\rho(n+1,t)e(n,t)+r(n,t-1)}{\sqrt{1-\rho^2(n+1,t)}\sqrt{1-e^2(n,t)}}.$$

The Schur filter requires an initialization. Simultaneously, the analysed signal is normalized in order to provide its good numerical properties. Specifically (Lee *et al.*, 1981; Lopatka *et al.*, 2005; 2006),

- the first sample of the registered signal $x_d(0)$ is normalized according to the relation

$$x(0)=\frac{x_d(0)}{\sqrt{c(0)}},\tag{2}$$

  where $c(0)$ is an estimate of the signal variance,

$$c(0)={x_d}^2(0)+\delta,\tag{3}$$

  while δ is a small-value constant preventing the occurrence of a division by zero;

- the successive signal samples are normalized according to the principle

$$x(t)=\frac{x_d(t)}{\sqrt{c(t)}},\tag{4}$$

  where $c(t)$ is an estimate of the signal variance at time instant t,

$$c(t)=\lambda c(t-1)+x_d^2(t),\tag{5}$$

  while λ is a forgetting factor.

The key role in the Schur algorithm is normalization expressed by the relations (4) and (5). The forgetting coefficient, $\lambda \in (0,1)$ is a significant parameter in this normalization. This coefficient balances the influence of a new signal sample both on the variance and covariance function values, and it determines the quality of estimators and the inertial properties of the algorithm.

After initialization, the two parameters of the algorithm are essential: the filter order P and the value of the forgetting factor λ. Selection of the filter order depends on the signal frequency structure and it should ensure a satisfactory model of the signal. It is worth remembering that every local maximum of the spectral power density of the process is represented by a pair of reflection coefficients. Thus, assuming that a speech signal spectrum contains 4–5 formants, it will usually be satisfactory to assume P = 10 to P = 14. On the other hand, the adopted value of the forgetting factor should depend on the rate of the signal change and should change adaptively (Makowski and Zimroz, 2013), but such an approach is difficult to realize. Therefore, a constant value

of λ is employed here. The equivalent rectangular window length T in samples, corresponding to an exponential window related to the forgetting factor λ, is provided by the relation $T = 1/(1-\lambda)$. The length T should approximately be equal to the period of local stationarity of the speech signal.

2.2. Parametric spectrogram and signal powers in mel bands. The analysis of reflection coefficient trajectories $\rho(n,t)$ is not easy for many reasons including their complicated relation to the signal spectrum. The hearing organ is a spectrum analyser. Therefore, speech signal analysis in the frequency domain is commonly applied. Having defined the Schur coefficients, we can convert them to an innovation filter in accordance with the following relation (e.g., Kay, 1988):

$$b(n,t) = a(n,P,t), \quad n = 0,\ldots,P, \tag{6}$$

where $a(i,j,t)$ are determined by iteration for successive p and n, $p = 0,\ldots,P-1$ and $n = 1,\ldots,p$,

$$a(p+1,n,t) = a(p,n,t) + \rho(p+1,t)a(p,p+1-n,t), \tag{7}$$

$$a(p+1,p+1,t) = \rho(p+1,t). \tag{8}$$

Then, by using innovation filter coefficients, a parametric signal spectrum, referred to as maximum entropy spectrum, can be defined for each time instant t:

$$S(f,t) = |1 - \sum_{n=1}^{N_f-1} b(n,t)e^{-j2\pi nf}|^{-2}, \tag{9}$$

where N_f is the length of the Fourier transform. This transformation produces a parametric time-frequency signal transform. It is worth mentioning that some difficulties with speech analysis/synthesis are observed due to the pitch periodicity (e.g., Kroon and Deprettere, 1988), particularly for high pitched part of the speech. Fortunately, we analyze the changes in the vocal tract by means of observing the spectrum defined by Eqn. (9), based on the innovation filter coefficients. This spectrum, due to the relatively low filter order, has a smoothed form with strongly suppressed influence of the pitch periodicity.

Then, by using such a transform, we can determine time-varying signal powers $L(k,t)$ in the subbands with index k, defined by the relation

$$L(k,t) = \sum_{f=f_{lk}}^{f_{hk}} S(f,t), \quad k = 1,\ldots,K, \tag{10}$$

where f_{lk} and f_{hk} are band cutoff frequencies taking into account the mel scale of sound perception, $f_{l0} = 0$, $f_{l,k+1} = f_{hk}$.

Averaging in frequency bands enables reducing the sizes of the analysed functions and estimator variances, while employing a mel frequency scale makes it possible to take the properties of the human hearing organ into account. $L(k,t)$ for the preset t is the mel spectrum, well-known in ASR issues though determined in a different way.

2.3. Normalized power difference and the error variance ratio. In order to simplify the criterion of signal spectrum change detection, let us introduce a normalised signal power difference for each band k

$$R(k,t) = \frac{L(k,t+d) - L(k,t)}{0.5\,[L(k,t+d) + L(k,t)]}, \tag{11}$$

where d is the distance between the mel spectra. $R(k,t)$ is then a measure of the mel spectrum change at the time interval d, determined from spectral slices. Figure 2 illustrates this with sample plots of function $R(k,t)$ defined by the formula (11) for the word *zapamjentaj*.

In functions $R(k,t)$ shown in Fig. 2, one can see local maxima and minima reflecting signal power changes

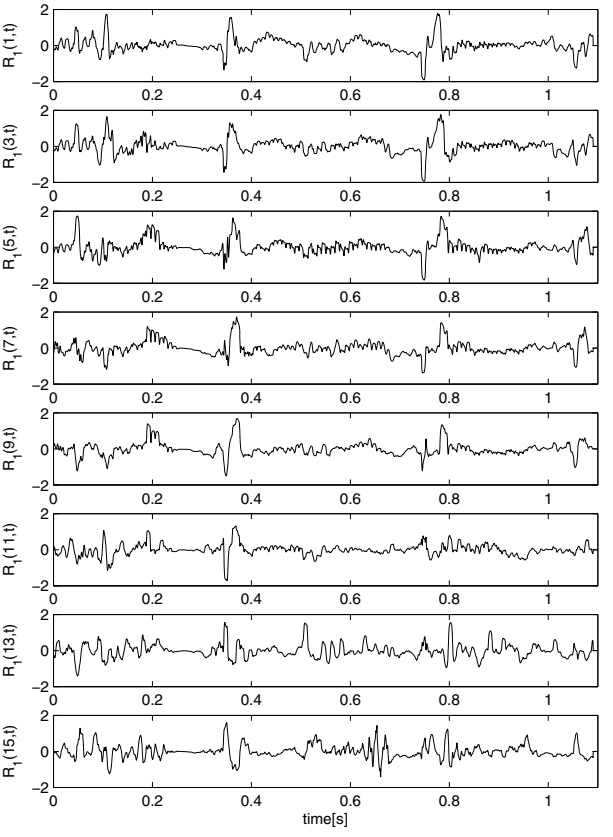


Fig. 2. Sample course of function $R(k,t)$ defined by the formula (11) for the signal *zapamjentaj* shown in Fig. 7, for P =10, $\lambda \approx 0.9917$ and d =90, for several mel bands.

in subbands within time intervals d. If the speech signal in a frequency band with index k does not change over a time interval d, then with an accuracy of the estimation error the value of $R(k,t)$ is zero. If the above condition is valid for all the k bands, this will mean that within a time interval from t to $t+d$ the signal does not change ($\mathcal{H}_0$ hypothesis). As a result, one can conclude that a given signal segment comprises the same phoneme, but it could be a different speech-related sound, e.g., a breath, a tongue click or a piece of a phoneme. To simplify the description, all these sounds will be referred to as phonemes.

Otherwise, when the function $R(k,t)$ exceeds a given threshold Θ for any k band, this will mean that the signal spectrum changes over this interval (hypothesis $\mathcal{H}_1$). Consequently, finding the local maxima of the function $R(k,t)$ which exceed the value of the detection threshold Θ enables acknowledging, with a given probability, that this is the boundary between different phonemes. A few problems remain to be solved, the most important of them being defining the values of the adaptation factor λ and detection threshold Θ.

In order to increase of the effectiveness of boundary detection, let us introduce another statistic, being the ratio of estimators of the prediction error variance for time instances spaced at g intervals, i.e.,

$$G(t) = \frac{\sigma_P^2(t+g)}{\sigma_P^2(t)}, \tag{12}$$

where $\sigma_P^2(t)$ is the estimator of the prediction error variance from the final section of the Schur filter for time t, defined by the relation

$$\sigma_P^2(t) = \frac{1}{M}\sum_{s=t-M+1}^{t} e^2(P,t). \tag{13}$$

Linear prediction errors are used in the algorithm described by Tyagi *et al.* (2006). As the prediction error is defined in the Schur algorithm, too, adding a statistic defined by the relations (12) and (13) only slightly increases the computation time.

Figure 3 shows an example evolution of function $G(t)$ for $g = M = 240$. There are distinct local extrema related chiefly to speech termination or initiation.

In the upper plot in Fig. 3 one can observe the influence of a pitch on the prediction error signal in voiced parts of the speech. Our attempts to filter out the pitch from the signal before inputting it to the Schur filter input failed due to estimation problems of pitch parameter values. Fortunately, the use of the averaging of the prediction error in the window of length M effectively suppresses the influence of pitch periodicity (see the lower graph of Fig. 3). Similarly, the pitch periodicity affects the function $R(k,t)$ (see Fig. 2), but the influence is small and decreases with an increase in λ.

2.4. Theoretical discussion.

2.4.1. Signal change detection based on the generalized maximum likelihood method. Let us consider a situation when changes in the signal $x(t)$ are to be detected by analysing signals in two moving windows with identical lengths M, shifted in time by g samples: $\mathbf{x}(t)$ and $\mathbf{x}(t+g)$. It is assumed here that the signal is modelled by means of the Gaussian AR process of order P described by the vector of coefficients

$$\mathbf{b}(t) = [b(1,t)\, b(2,t) \cdots b(P,t)].$$

As a result of such assumptions, the following hypothesis tests will be considered:

- Hypothesis $\mathcal{H}_0$: no changes in signal $x(t)$ over the distance of g samples; coefficients $\mathbf{b}(t)$ and variance $\sigma_P^2(t)$ do not change;
- Hypothesis $\mathcal{H}_1$: changes occur in signal $x(t)$ over the distance of g samples; there are changes in coefficients $\mathbf{b}(t)$ and variance $\sigma_P^2(t)$.

In order to solve the detection problem thus formulated, a log-likelihood ratio function between two hypothesis will be created (Barkat, 1991; Puig, 2010; Jamouli *et al.*, 2012):

$$\log(U(\mathbf{x}(t),g)) = \log\left(\frac{p(\mathbf{x}(t+g)|\mathbf{b}(t+g),\sigma_P^2(t+g))}{p(\mathbf{x}(t)|\mathbf{b}(t),\sigma_P^2(t))}\right), \tag{14}$$

where

$$p(\mathbf{x}(a)|\mathbf{b}(a),\sigma_P^2(a)) = \frac{1}{(2\pi\sigma_P^2(a))^{M/2}}\exp\left(\frac{-1}{2\sigma_P^2(a)}\sum_{i=0}^{M-1} e_P^2(a-i)\right) \tag{15}$$

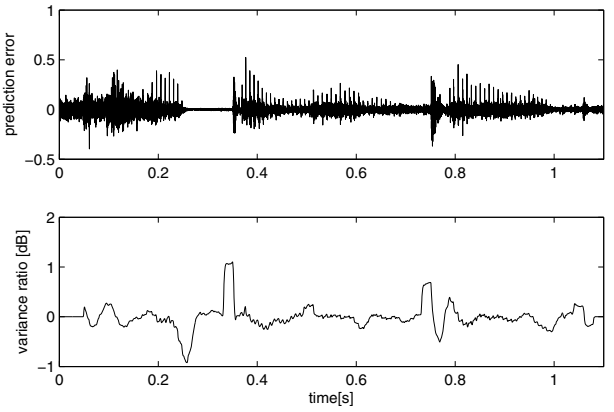


Fig. 3. Sample evolution of the function $G(t)$ defined by the formula (12) for the signal *zapamjentaj* presented in Fig. 7, for $g = M = 240$ (on a logarithmic scale).

and

$$\sigma_P^2(a) = \frac{1}{M}\sum_{i=0}^{M-1} e_P^2(a-i), \tag{16}$$

where

$$e_P(a) = x(a) - \sum_{i=1}^{P} b(i,a)x(a-i). \tag{17}$$

In the next step, we will use a generalized GLRT rule based on independent maximisation of the numerator and denominator of the likelihood function in relation to unknown parameters (Kay, 1998):

$$\begin{aligned}
&\log(U(\mathbf{x}(t), g) \\
&= \log\left(\frac{\max_{\mathbf{b}(t+g)} p(\mathbf{x}(t+g)|\mathbf{b}(t+g), \sigma_P^2(t+g))}{\max_{\mathbf{b}(t)} p(\mathbf{x}(t)|\mathbf{b}(t), \sigma_P^2(t))}\right) \\
&= \frac{M}{2}\log\left(\frac{\sigma_P^2(t+g)}{\sigma_P^2(t)}\right),
\end{aligned} \tag{18}$$

which leads to the formula

$$\log(U(\mathbf{x}(t), g) = \frac{M}{2}\log\left(\frac{\sigma_P^2(t+g)}{\sigma_P^2(t)}\right). \tag{19}$$

Since the prediction errors have normal distributions, i.e., $e_P(t) \sim \mathcal{N}(0, \bar{\sigma}_P^2(t))$ and $e_P(t+g) \sim \mathcal{N}(0, \bar{\sigma}_P^2(t+g))$, the statistic

$$\eta(a) = \frac{\sigma_P^2(a)}{\bar{\sigma}_P^2(a)} = \frac{1}{M}\sum_{i=0}^{M-1} e_p^2(a-i)/\bar{\sigma}_P^2(a) \tag{20}$$

has a chi-squared distribution with M degrees of freedom, i.e., $\eta(a) \sim \chi^2(M)$. In the relation (20), $\bar{\sigma}_P^2(a)$ denotes the precise variance value. In the next step, the following statistic will be defined:

$$\begin{aligned}
g(t) &= \frac{\eta(t+g)/M}{\eta(t)/M} = \frac{\sigma_P^2(t+g)}{\sigma_P^2(t)}\frac{\bar{\sigma}_P^2(t)}{\bar{\sigma}_P^2(t+g)} \\
&= G(t)\frac{\bar{\sigma}_P^2(t)}{\bar{\sigma}_P^2(t+g)} \sim F(M,M)
\end{aligned} \tag{21}$$

where $F(M,M)$ is the F-Snedecor distribution with (M,M) degrees of freedom. As a result, at the preset confidence level $1-\alpha$ for hypothesis $\mathcal{H}_0$, F-Snedecor distribution tables can be used to find critical values Θ_a and Θ_b for a two-tailed test:

$$\Theta_a \le g(t) = G(t)\frac{\bar{\sigma}_P^2(t)}{\bar{\sigma}_P^2(t+g)} \le \Theta_b, \tag{22}$$

or, in the equivalent form, for statistic $G(t)$,

$$\Theta_a\frac{\bar{\sigma}_P^2(t+g)}{\bar{\sigma}_P^2(t)} \le G(t) \le \Theta_b\frac{\bar{\sigma}_P^2(t+g)}{\bar{\sigma}_P^2(t)}. \tag{23}$$

If we assume that the maximum value defining the variance change over the distance of g samples is

$$\Gamma_{\max} = \max_t \frac{\bar{\sigma}_P^2(t+g)}{\bar{\sigma}_P^2(t)}, \tag{24}$$

then, in order to ensure the assumed confidence level, the following inequalities should be satisfied:

$$\begin{aligned}
\frac{\Theta_a}{\Gamma_{\max}} &\le \Theta_a\frac{\bar{\sigma}_P^2(t+g)}{\bar{\sigma}_P^2(t)} \le G(t) \\
&\le \Theta_b\frac{\bar{\sigma}_P^2(t+g)}{\bar{\sigma}_P^2(t)} \le \Theta_b\Gamma_{\max}
\end{aligned} \tag{25}$$

For instance, if the confidence level for statistic $G(t)$ is preset at 98%, then the values obtained for $M = 240$ will be $\Theta_a = 0.74$ and $\Theta_b = 1.35$. Observation of prediction error variances of the analysed speech signals demonstrates that the variance over distance $d = 240$ does not exceed 4.2 times, i.e., $\Gamma_{\max} \approx 4.2$. Due to signal normalization, before being fed to the Schur filter, $\Gamma_{\max}$ does not depend on a signal scaling or a local intensity of the signal $x_d(t)$. Bearing in mind the above estimation of $\Gamma_{\max}$, the following decision rule for testing hypothesis $\mathcal{H}_0$ is obtained:

$$0.176 \le G(t) \le 5.670, \tag{26}$$

or, on a logarithmic scale,

$$-0.754 \le \log G(t) \le 0.753. \tag{27}$$

2.4.2. Signal change detection in the spectral domain. Let us accept the assumption (Rabiner and Gold, 1975) that normalized spectral power density components $S_F(f,t)/\sigma_S^2(f,t)$ of the analysed speech signal have a chi-squared distribution with two degrees of freedom, i.e., $S_F(f,t)/\sigma_S^2(f,t) \sim \chi^2(2)$, where $\sigma_S^2(f,t)$ is a variance of spectral power density for preset f and t. In such a case, time-varying powers $L(k,t)$ defined by the relation (10) and normalized in relation to the variance $\sigma_L^2(k,t)$ also have a chi-squared distribution with J_k degrees of freedom:

$$\frac{L(k,t)}{\sigma_L^2(k,t)} \sim \chi^2(2J_k), \tag{28}$$

where J_k is the number of spectral lines in subband k and $\sigma_L^2(k,t) = \sigma_S^2(f,t)$ within the analysed subband.

At the next stage of the discussion, let us introduce the statistic

$$r(k,t) = \frac{L(k,t+d)\,/\,2J_k\sigma_L^2(k,t)}{L(k,t)\,/\,2J_k\sigma_L^2(k,t)}, \tag{29}$$

which has the F-Snedecor distribution with (J_k, J_k) degrees of freedom, and it can be assumed that the power variance $\sigma_L^2(k,t)$ for subband k with a shift by d samples

does not change its value, so $\sigma_L^2(k,t) = \sigma_L^2(k,t+d)$. For the preset confidence level $1-\alpha_k$, let us define the critical values for hypothesis $\mathcal{H}_0$:

$$\beta_{a,k} \le r(k,t) \le \beta_{b,k}. \tag{30}$$

Statistic $R(k,t)$ introduced by the relation (11) could be expressed with $r(k,t)$ in the form

$$\begin{aligned} |R(k,t)| &= \frac{|L(k,t+d) - L(k,t)|}{0.5\,[L(k,t+d) + L(k,t)]} \\ &= \frac{|r(k,t)-1|}{0.5\,[r(k,t)+1]}. \end{aligned} \tag{31}$$

If decision thresholds $\beta_{a,k}$ and $\beta_{b,k}$ are adopted in statistic (31), the following condition for testing hypothesis $\mathcal{H}_0$ will be obtained:

$$\begin{aligned} &0 \le |R(k,t| \\ &\quad \le \max\left(\frac{|\beta_{a,k}-1|}{0.5(1+\beta_{a,k})}, \frac{|\beta_{b,k}-1|}{0.5(1+\beta_{b,k})}\right) = \Theta_k. \end{aligned} \tag{32}$$

When an additional requirement for the simultaneous absence of a power change in all the subbands (hypothesis $\mathcal{H}_0$) is introduced, the probability of such an event, with the assumed independence of events in particular subbands, can be described with the formula

$$P(\mathcal{H}_0) = \prod_{k=1}^{K} P_k(\mathcal{H}_0) = \prod_{k=1}^{K} P(|R(k,t)| \le \Theta_k). \tag{33}$$

Subsequently, one can observe that probability $P(\mathcal{H}_0)$ defined in (33) fulfils the inequality

$$[P_{k,\min}(\mathcal{H}_0)]^K \le P(\mathcal{H}_0) \le [P_{k,\max}(\mathcal{H}_0)]^K, \tag{34}$$

which enables estimating the confidence interval for hypothesis $\mathcal{H}_0$.

For example, if we assume that Θ_k =1.76, and Θ_k is identical for all the subbands, then the corresponding threshold values are $\beta_{a,k}$ =0.0638 and $\beta_{b,k}$ =15.66, respectively. Moreover, for a case where $J_k = 5$, tables for the F-Snedecor distribution with (10,10) degrees of freedom can be used to find the corresponding critical values $P_{a,k} = P_{b,k}$=0.00008, which demonstrates that $1 - \alpha_k = 0.00016$ while $P_k(\mathcal{H}_0) = P_{k,\min}(\mathcal{H}_0) = 1-\alpha_k$ =0.99984. Finally, by adopting K =16, we obtain

$$P(\mathcal{H}_0) \ge (P_{k,\min}(\mathcal{H}_0))^K = 0.99984^{16} = 0.997, \tag{35}$$

which demonstrates that α for hypothesis $\mathcal{H}_0$ does not exceed the value of 0.02.

2.5. General processing and detection plan. Bearing in mind the fact that phoneme lengths vary a lot and range from a few to a few hundred milliseconds, the authors propose segmentation based on two sets of functions: $R_1(k,t)$ and $R_2(k,t)$, defined for various values of λ_1 and λ_2, and consequently for various d_1 and d_2 as well as Θ_1 and Θ_2. The diagram of the proposed automatic segmentation algorithm is shown in Fig. 4.

The sampling frequency f_s of the analysed signals is 12 kHz. At first, boundary detection, based on functions $R_1(k,t)$, is performed for $\lambda_1 \approx 0.9917$ ($T_1 = 120$), $P_1 = 10$ and $d_1 = 90$. For such parameters, the function $R_1(k,t)$ enables detection of mostly fast changes in the mel spectrum of the signal. In the next step, detection using $R_2(k,t)$ is performed for $\lambda_2 \approx 0.9979$ ($T_2 = 480$), $P_2 = 14$ and $d_2 = 360$. A larger distance between spectral slices enables detecting slower changes occurring in the signal. At the same time, a higher estimation quality, related to an increase in T and in the order of the Schur filter, enables a more accurate analysis in the frequency domain. In detection based on $R_2(k,t)$, local extrema are sought, allowing for the boundaries defined earlier based on $R_1(k,t)$, while maintaining the minimum distance between particular boundaries (cf. Section 3). Finally, in the third step, detection based on the function $G(t)$ is performed, also taking into account the earlier detection results. The summing operation shown by the graph in Fig. 4 indicates that the set of inter-phonemic boundaries detected in *Detection 1* is supplemented with boundaries detected in *Detections 2* and *3*.

3. Algorithm for signal spectrum change detection

The proposed spectral change detection algorithm has restrictions which make it possible to take into account specific characteristics of the speech signal. Firstly, the speech signal contains pauses, so defining boundaries between phonemes in the periods where the speaker is inactive is pointless. Therefore, a Voice Activity Detection (VAD) algorithm is applied first. It is based on the power of successive 20 ms signal frames shifted with a step of 5 ms. Inter-phoneme boundary detection is performed exclusively for the activity fragments. Secondly, inertion of the speech organ results in the fact that transitions from one phoneme to another take time.

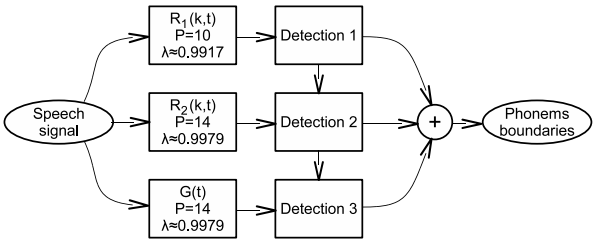


Fig. 4. Diagram of the proposed automatic segmentation algorithm.

This yields restrictions concerning the minimum distance between such boundaries. A flow chart of such a detection for $R(k,t)$ is shown in Fig. 5.

Detection starts with sample t_p, which is the onset of a given activity period or with a sample delayed from the previously defined boundary by the assumed minimum distance. Detection finishes with sample t_k, which is the termination of the speaker's activity. The decision blocks (1) and (2) determine respectively the local maximum and minimum. The positions of these extrema are written in table M_b and their number reaches m_b. Owing to the cyclical nature of pitch excitation (for the voiced segments of speech), there are often a few local extrema of the function $R(k,t)$ around the boundary of interphonemic transitions (cf., e.g., Fig. 2). Therefore, the detected extrema are first recorded in an auxiliary table M_b, and it is only after meeting the conditions defined in the decision block (3), the most important of which being that the distance between the extrema must be larger than d_m, that the ultimate choice is made. When $m_b = 1$, the boundary position is copied to table M. When $m_b > 1$, the arithmetic mean $R_s(t)$ of the absolute values of all the functions $R(k,t)$ in the range of the occurrence of the boundary group M_b is calculated and the position of the defined boundary is established at the point of the maximum of the function $R_s(t)$.

With a slight delay, another detection, based on $R_2(k,t)$, is performed. The procedure is similar to that for $R_1(k,t)$, the only differences occurring in parameter values. Finally, after these two detections, the third one, based on $G(t)$, is carried out in a manner similar to the former two.

If the distance between the current time t and the activity onset or the last determined boundary is greater than d_a, the threshold value is lowered linearly to a certain minimum value Θ_m. This enables detecting parts of boundaries in these speech segments where interphonemic transitions run smoothly, so the values of the functions $R_1(k,t)$, $R_2(k,t)$ or $G(t)$ are lower. Figure 6 illustrates the lowering of the detection threshold value.

Finally, for some phonemes, chiefly fricative and affricate ones, significant local spectral changes occur within phonemes, which is related to the turbulent nature of the air flow. Such incorrectly defined boundaries can be eliminated based on the function

$$U(t) = \frac{\sum_{f=N_f/4+1}^{N_f/2} S(f,t)}{\sum_{f=0}^{N_f/4} S(f,t)}, \tag{36}$$

where N_f is the length of the Fourier transform. The function $U(t)$ expresses the ratio of the power over $f_s/4$ to the power below $f_s/4$. For fricative and affricate phonemes, this ratio reaches relatively high values. The criterion of eliminating such an incorrectly defined boundary for time t has the following form:

$$U(t - r/2) > \Omega \quad \wedge \quad U(t + r/2) > \Omega, \tag{37}$$

where r is the distance between the verified points $U(t)$ and Ω is the elimination threshold.

4. Algorithm quality evaluation

The values of algorithm and detection parameters have a crucial effect on the results of automatic segmentation. At the same time, applications of the proposed algorithm in the speech domain can be multiple, including

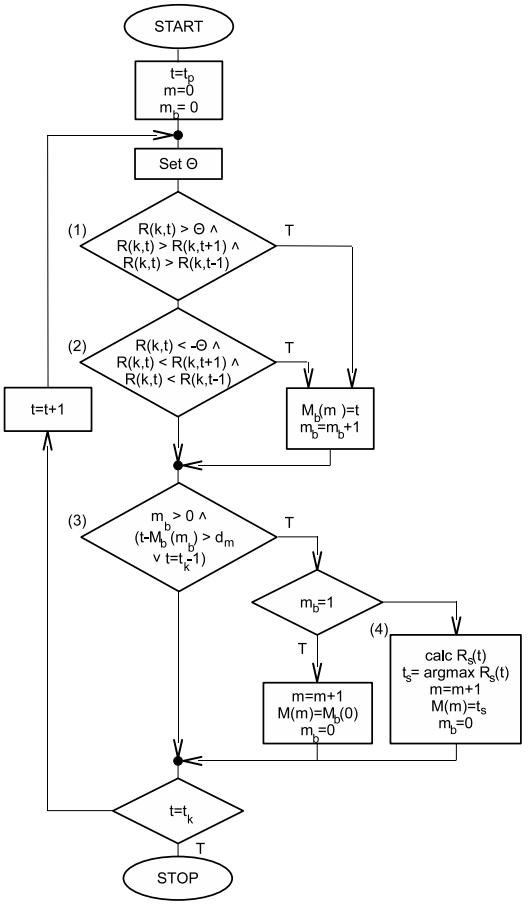


Fig. 5. Flowchart of the detection algorithm.

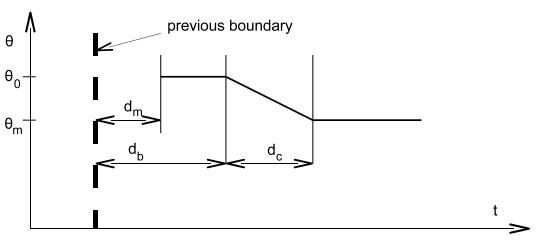


Fig. 6. Illustration of a detection threshold change.

- manually-assisted speech signal segmentation: in such a strategy it will be important to detect all the boundaries between speech sounds while accepting the fact that some of the boundaries will be redundant;
- segmentation using information about the number of phonemes in a given utterance; this is a widely used segmentation strategy at the stage of ASR system training;
- segmentation at the speech recognition stage; in this strategy it will be important to avoid wrong decisions (redundant boundaries) while accepting the fact that some boundaries between speech sounds will not be detected.

This section, presenting experiment results, will focus on the latter strategy.

4.1. Segmentation quality ratings. Let S denote the number of boundaries identified by manual segmentation, J—the number of boundaries determined automatically, J_G—the number of boundaries whose distance from the manual segmentation boundary is not greater than 10 ms, J_B—the number of boundaries whose distance from the manual segmentation boundary is greater than 10 ms but smaller than 20 ms, J_R—the number of redundant boundaries, i.e., those whose distance from the nearest boundary is greater than 20 ms. In order to obtain a global statistical evaluation of the results of automatic segmentation, let us introduce four quality measures:

- rating P_G of accurately defined boundaries,

$$P_G = \frac{J_G}{J}, \tag{38}$$

- rating P_B of inaccurately defined boundaries,

$$P_B = \frac{J_B}{J}, \tag{39}$$

- rating P_R of redundant boundaries,

$$P_R = \frac{J_R}{J}, \tag{40}$$

- rating P_U undetected boundaries,

$$P_U = \frac{S - J}{S}. \tag{41}$$

From the point of view of detection theory, P_R represents a probability of false acceptance, i.e., $P(\mathcal{H}_1|\mathcal{H}_0)$, and P_U—a probability of false rejection, i.e., $P(\mathcal{H}_0|\mathcal{H}_1)$.

Table 1. Values of detection parameters.

Parameter	Detect.1	Detect.2	Detect.3
$d_m[s]$	0.014	0.035	0.035
$d_b[s]$	0.054	0.075	0.060
$d_c[s]$	0.040	0.060	0.050
Θ_0	1.76	1.82	0.75
Θ_m	1.68	1.60	0.72

4.2. Processing and detection parameter values. The values of the employed adaptation factors λ_1 and λ_2, the numbers of sections P_1 and P_2, as well as the distances between mel spectra slices d_1 and d_2, were discussed in Section 2.5. In order to reduce estimator variances, simple smoothing of reflection coefficient trajectories was applied. Parametric spectra obtained from the relation (9) were defined with a step of 5 samples. The value of M in the formula (13) is 240 and $g = M$. Boundary frequencies f_{lk} from the relation (10) in the range of up to 1 kHz are distributed linearly (8 bands), and higher up they form a geometric series with a multiplier of 1.223. The total number of mel bands is 16. The values of the remaining parameters are given in Table 1.

The value of the parameter $d_{m1} = 14$ ms corresponds to the minimum assumed pitch frequency $f_0 \approx 71$Hz. This is, at the same time, the minimum distance between definable boundaries. In the section, devoted to elimination of redundant boundaries from noise fragments of speech (relation (13)), the adopted $r = 30$ ms and $\Omega = 1.2$. The values of thresholds and other segmentation parameters were defined through analysis (cf. Section 2.4) and based on speech signal properties, and then verified through a number of numerical experiments. The optimization criterion was established so that the value of P_R rating is 2% and the value of P_U be as low as possible.

4.3. Research material and an example segmentation result. In order to test the effectiveness of the proposed algorithm, automatic segmentation was performed for words from the *vdITA* recording corpus. This is a corpus of 36 male voices, each uttering 33 words. These utterances had earlier undergone manual phonemic segmentation where the list of phonemes comprised 37 entries plus one undescribed speech sound. The total number of manually defined boundaries was 8353. Before automatic segmentation was applied, the utterances were disturbed by noise to the level of SNR ≈ 40 dB. Figure 7 shows an example segmentation result for the word *zapamjentaj*.

The thick lines at the bottom of Fig. 7 mark the manually designated phoneme boundaries. The thick lines on top of the drawing indicate the boundaries of voice activity periods, while the dotted lines represent phoneme

Table 2. Values of automatic segmentation quality ratings in functions Θ_{01} and Θ_{m1}.

Θ_{01}	1.60	1.68	1.76	1.84
Θ_{m1}	1.52	1.60	1.68	1.76
$P_G[\%]$	85.2	86.8	86.9	86.0
$P_B[\%]$	10.7	10.7	11.1	12.0
$P_R[\%]$	4.1	2.6	2.0	2.1
$P_U[\%]$	44.3	46.7	48.8	50.5

Table 3. Values of automatic segmentation quality ratings in functions Θ_{02} and Θ_{m2}.

Θ_{02}	1.66	1.74	1.82	1.90
Θ_{m2}	1.44	1.52	1.60	1.68
$P_G[\%]$	83.5	85.6	86.9	88.2
$P_B[\%]$	12.2	11.3	11.1	10.4
$P_R[\%]$	4.3	3.1	2.0	1.4
$P_U[\%]$	43.7	46.2	48.8	51.8

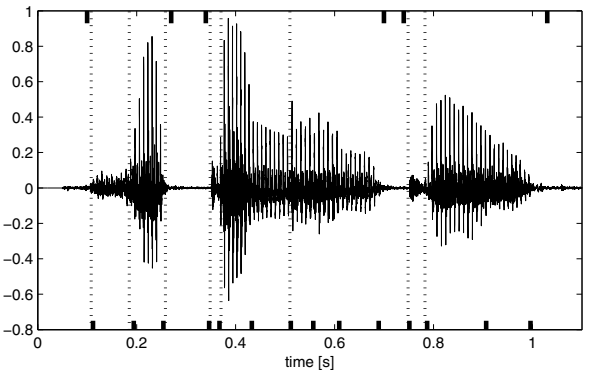
Table 4. Values of automatic segmentation quality ratings in functions Θ_{03} and Θ_{m3}.

Θ_{03}	0.67	0.71	0.75	0.79
Θ_{m3}	0.64	0.68	0.72	0.76
$P_G[\%]$	85.5	86.2	86.9	87.5
$P_B[\%]$	11.7	11.5	11.1	10.5
$P_R[\%]$	2.9	2.4	2.0	1.9
$P_U[\%]$	46.0	47.5	48.8	50.4

boundaries determined automatically. Out of the fourteen manual segmentation boundaries, the proposed algorithm detected eight. Their accuracy is very satisfactory. Out of the six undefined boundaries, two are signal decay boundaries at the ends of two activity fragments. This is a regularity observed for almost all the analysed utterances. However, it does not cause any difficulty, as this problem is dealt with by the VAD algorithm, although with a lower accuracy. Four of the undefined boundaries are related to smooth transition between phonemes. These boundaries are accompanied by local extrema of functions $R_1(k,t)$, $R_2(k,t)$ or $G(t)$ (cf. Figs. 5 and 6), but as thresholds had been adjusted in such a way that the number of false boundaries is minimised, they were not detected.

4.4. Global segmentation results. Tables 2, 3 and 4 present the values of automatic segmentation quality indices for several parameter values: Θ_{01} and Θ_{m1}, Θ_{02} and Θ_{m2}, as well as Θ_{03} and Θ_{m3}, respectively, with the values of the remaining parameters like in Section 4.2.

For the parameter values given in Section 4.2, the value of P_R is 2%, and that of P_U rating is 48.8%. A detailed analysis proves that a lot of redundant boundaries are found in those speech fragments which comprise sounds other than phonemes or in places where the pronunciation of a phoneme changes noticeably. Such false detections can hardly be considered an algorithm error. A large number of boundaries not defined precisely enough (rating P_B) are boundaries in the places of smooth transitions between phonemes. In such situations, there is quite a good chance that manually designated boundaries are not accurate. Such inaccuracies do not occur for plosive phonemes. In all the analysed recording material, a large number of unidentified boundaries is a consequence of the employed strategy of minimising the number of false detections, i.e., minimising the probability of false acceptance. Some of these boundaries will be defined with a lesser accuracy by the VAD algorithm.



vspace*-3mm

Fig. 7. Diagram of a sample utterance of the word *zapamjentaj* with manually defined phoneme boundaries (rectangles at the bottom of the picture), identified activity ranges (rectangles at the top of the picture) and automatically determined phoneme boundaries (dotted lines).

The obtained results of automatic segmentation were compared with those achieved with the algorithm described by Tyagi *et al.* (2006). Among other well-known algorithms, this one has a similar complexity and it is an unsupervised algorithm. It was slightly modified, as it had originally been designed for defining boundaries of speech signal parametrisation frames whose maximum length was set at 60 ms. Therefore, in this form they are unsuitable for segmentation. The function on which boundary detection was based has the form (cf. Section 1)

$$C(t) = \frac{1}{2}\Big[(N_0 + N_1) \log \sigma_e^2 - \big(N_0 \log \sigma_{e_0}^2 + N_1 \log \sigma_{e_1}^2\big) \Big], \qquad (42)$$

where e_0, e_1 and e are linear prediction error signals for signal x_0 being a fragment preceding the analysis point t, the signal x_1 being a fragment following the analysis point t, and the joint of these fragments, respectively. The adopted lengths of N_0 and N_1 were 20 ms. To make the

Table 5. Values of automatic segmentation quality ratings in function Θ_T.

Θ_T	33	38	43	48
$P_G[\%]$	87.4	88.6	89.9	90.6
$P_B[\%]$	9.8	8.9	8.1	8.1
$P_R[\%]$	2.8	2.4	2.0	1.2
$P_U[\%]$	57.8	60.4	69.1	73.2

results comparable, the analysis point was shifted by every 5 samples. Table 5 contains the values of segmentation quality ratings for this algorithm as a function of the threshold value Θ_T.

The presented results obtained by both the algorithms demonstrate that, with the assumed probability of false acceptance $P(\mathcal{H}_1|\mathcal{H}_0)$ at 2%, the efficiency of the proposed algorithm is over 20% higher.

5. Conclusions

The paper proposes a real-time algorithm for automatic segmentation of a speech signal. The discussed solution has a few valuable features such as the independence of signal scaling related to normalization performed in innovation filtering, fast adaptation to time-varying signal statistics, numerical stability, the possibility of performing detection with different temporal resolutions and adopting a mel scale of sound perception. For statistics of variance and power change in subbands, when employed in the detection process, a formal analysis was performed. It consisted of determining the threshold values that guarantee the required confidence level. The threshold values obtained through theoretical deliberations were applied in research conducted on real speech signals which confirmed the correctness of the performed analyses and the effectiveness of the designed method. What is also worth emphasizing is the low computational complexity of the discussed method, which makes it applicable for real time work at the speech recognition stage. Moreover, the properties of the algorithm are so effective that it copes with the problem of automatic segmentation better than the referential algorithm.

The level of incorrect acceptances is satisfactory and it should not pose any problems. A vast majority of undetected inter-phoneme boundaries are attributable to smooth transitions between phonemes. Also, the probability of not detecting a boundary between spectrally different boundaries is very small, and such transitions pose the biggest problems for speech signal parametrisation at recognition stage.

Acknowledgment

Preparation of this manuscript was supported by the Ministry of Science and Higher Education of Poland (grant no. S-20119).

References

Almpanidis, G. and Kotropoulos, C. (2007). Phonetic segmentation using the generalized Gamma distribution and small sample Bayesian information criterion, *Speech Communication* **50**(1): 38–55.

Almpanidis, G., Kotti, M. and Kotropoulos, C. (2009). Robust detection of phone boundaries using model selection criteria with few observations, *IEEE Transactions on Audio, Speech, and Signal Processing* **17**(2): 287–298.

Barkat, M. (1991). *Signal Detection and Estimation*, Artech House, Boston, MA.

Brandt, A.V. (1983). Detecting and estimating the parameters jumps using ladder algorithms and likelihood ratio test, *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Boston, MA, USA*, pp. 1017–1020.

Brugnara, F., Falavinga, D. and Omolongo, M. (1993). Automatic segmentation and labeling of speech based on hidden Markov models, *Speech Communication* **12**(4): 357–370.

Delacourt, P. and Wellekens, C.J. (2000). DISTBIC: A speaker-based segmentation for audio data indexing, *Speech Communication* **32**(1–2): 111–126.

Gomez, J.A. and Calvo, M. (2011). Improvements on automatic speech segmentation at the phonetic level, *in* C. San Martin and S.-W. Kim (Eds.), *CIARP 2011*, Lecture Notes in Computer Science, Vol. 7042, Springer-Verlag, Berlin/Heidelberg, pp. 557–564.

Haykin, S. (1996). *Adaptive Filter Theory*, Prentice-Hall, Englewood Cliffs, NJ.

Jamouli, H., Al Hail, M.A. and Sauter, D. (2012). A mixed active and passive GLR test for a fault tolerant control system, *International Journal of Applied Mathematics and Computer Science* **22**(1): 9–23, DOI: 10.2478/v10006-012-0001-1.

Kay, S.M. (1988). *Modern Spectral Estimation*, Prentice-Hall, Englewood Cliffs, NJ.

Kay, S.M. (1998). *Fundamentals of Statistical Signal Processing, Vol. II: Detection Theory*, Prentice-Hall, Englewood Clifft, NJ.

Kroon, P. and Deprettere, E.F. (1988). A class of analysis-by-synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 kbits/s, *IEEE Journal on Selected Areas in Communications* **6**(2): 353–363.

Lee, D.T.L., Morf, M. and Friedlander, B. (1981). Recursive least squares ladder estimation algorithms, *IEEE Transactions on Circuits and Systems* **28**(6): 627–641.

Lopatka, M., Adam, O., Laplanche, C., Zarzycki, J. and Motsch, J-F. (2005). Effective analysis of non-stationary short-time signals based on the adaptive Schur filter, *IEEE/SP 13th Workshop on Statistical Signal Processing, Bordeaux, France*, pp. 251–256.

Lopatka, M., Adam, O., Laplanche, C., Motsch, J-F. and Zarzycki, J. (2006). Sperm whale click analysis using a recursive time-variant lattice filter, *Applied Acoustics* **67**(11–12): 1118–1133.

Makowski, R. and Zimroz, R. (2013). A procedure for weighted summation of the derivatives of reflection coefficients in adaptive Schur filter with application to fault detection in rolling element bearings, *Mechanical Systems and Signal Processing* **38**(1): 65–77.

Mporas, I., Ganchev, T. and Fakotakis, N. (2008). Phonetic segmentation using multiple speech features, *International Journal of Speech Technology* **11**(1): 73–85.

Park, S.S. and Kim, N.S. (2007). On using multiple models for automatic speech segmentation, *IEEE Transactions on Audio, Speech, and Language Processing* **15**(8): 2202–2212.

Prasad, V.K., Nagarajan, T. and Murthy, H.A. (2004). Automatic segmentation of continuous speech using minimum phase delay functions, *Speech Communication* **42**(3–4): 429–446.

Puig, V. (2010). Fault diagnosis and fault tolerant control using set-membership approaches: Application to real case studies, *International Journal of Applied Mathematics and Computer Science* **20**(4): 619–635, DOI: 10.2478/v10006-010-0046-y.

Rabiner, L. and Gold, B. (1975). *Theory and Application of Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ.

Rabiner, L. and Juang, B-H. (1993). *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ.

Rudoy, D., Quatieri, T.F. and Wolfe, P.J. (2011). Time-varying autoregressions in speech: Detection theory and applications, *IEEE Transaction on Audio, Speech, and Language Processing* **19**(4): 977–989.

Scharenborg, O., Wan, V. and Ernestus, M. (2010). Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries, *Journal of Acoustical Society of America* **127**(2): 1084–1095.

Schwarz, P., Matejka, P. and Cernocky, J. (2006). Hierarchical structures of neural networks for phoneme recognition, *IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse, France,* Vol. 1, pp. 325–328.

Sharma, M. and Mammone, R. (1996). Blind speech segmentation: Automatic segmentation of speech without linguistic knowledge, *Proceedings of the International Conference on Spoken Language Processing, Philadelphia, PA, USA,* pp. 1237–1240.

Toledano, D.T., Hernandez Gomez, L.A. and Villarrubia Grande, L. (2003) Automatic phonetic segmentation, *IEEE Transactions on Speech and Audio Processing* **11**(6): 617–625.

Tyagi, V., Bourlard, H. and Wellekens, C. (2006). On variable-scale piecewise stationary analysis of speech signals for ASR, *Speech Communication* **48**(9): 1182–1191.

Ryszard Makowski obtained his Ph.D. in acoustics from the Faculty of Electronics at the Wrocław University of Technology in 1982. Since 1996 he has been an associate professor at the same faculty. He is the author of about 70 scientific papers primarily in the field of digital signal processing in areas such as acoustics, mechanics, mining seismology and telecommunications. Currently his main scientific interests focus on automatic speech recognition, speech enhancement and machinery condition monitoring.

Robert Hossa obtained his Ph.D. in control and robotics in 1997. Since 1998, he has been an assistant professor in the Signal Theory Section, Faculty of Electronics, Wrocław University of Technology. His scientific interests focus on the field of statistical methods of signal processing, detection algorithms, optimal adaptive filtering, array processing and sensor data fusion.

Received: 21 January 2013
Revised: 16 June 2013