ARCHIVES OF ACOUSTICS
18, 1, 3–15 (1993)

# AUTOMATIC SPEECH SIGNAL SEGMENTATION WITH CHOSEN PARAMETRIZATION METHOD

Cz. BASZTURA and T. SAWCZYN

Institute of Telecommunication and Acoustics Wrocław Technical University
(50-317 Wrocław, ul. Janiszewskiego 7/9)

This paper is dedicated to the problem of automatic segmentation of a speech signal into so-called phonetic segments, i.e. speech signal segment with homogeneous physical structure which can be ascribed with adequate phonetic mean. This is the second trend in segmentation, as opposed to speech signal segmentation into short fixed segments.

A segmentation algorithm is presented. It is based on calculations of the phonetic function at speech, which makes it possible to find the boundaries of these phonetic segments.

The usability of three different parametrization methods — based on the analysis of zero-crossings, spectral analysis and linear prediction coding — was analyzed. No significant differences were observed in the efficiency of investigated parameters.

## 1. Introduction

A technological development of the civilized world has led to an increased demand for new means of man-machine communication. New methods of communication include automatic speech and speaker recognition, speech synthesis, among others. Simple ASR systems (Automatic Speech Recognition) are limited to the recognition of several to several hundred isolated words. A limited vocabulary, as well as a need for a definite articulation discipline from the operators are the shortcomings of this method [3, 4, 9]. Therefore, at the same time research has been performed on automatic recognition of continuous speech. Features disadvantageous from the point of view ASR are more clearly distinguishable during continuited language units. These defects are as follows:

— less clear pronounciation and greater differences between speakers,
— consonants pronounced in continuous speech are shorter,
— coarticulation effects are observed between words,
— variable influence of intonation and accent on the speech signal's physical structure.

Speech may be presented in the form of a between the words and weak made evident. This is an additional difficulty in the realization of ASR systems.

The classical ASR procedure requires the first necessary step to be the determination of the number and type of objects (classes), or in other words the so-called base code.

In the case of global recognition of a limited vocabulary the base code consists of all the words included in the vocabulary. When phonems are the base code (e.g. $M = 37$

phonems for the Polish language), then the recognition at the lowest level is limited to the recognition of phonems [2, 5, 6, 8]. In this case the problem of speech signal segmentation into phonetic segments, which can be assigned with phonem designations, arises. A quasistationary section of a speech signal, which can be assigned to a definite phonem or other unit of speech, is called a phonetic segment. Two parallel trends in segmentation can be distinguished [1, 7]:

— segmentation into established permanent quasistationary segments, in literature called also "implicit segementation"

— so-called "explicit segmentation", which consists in the segmentation into segments defined by phonetic transcription. In fact, it is a segmentation into defined above phonetic segments. Definite properties can be distinguished in both trends. Most important ones are presented below:

**Constant segmentation**

1. The number of segments depends on the word's length.
2. Segments are not related to the phonetic description.
3. Boundaries of segments are precisely defined.
4. There is a great number of segments.

**Segmentation into phonetic segments**

1. The number of segments is correlated with the description given from phonetic transcription.
2. Segments can be labelled in accordance to the phonetic description.
3. Boundaries of segments are approximate.
4. Limited number of segments.

Segmentation into permanent segments was applied in most cases of research on ASR for continuous speech. This results in a need of great storage capacity when segmentation of a word into 10 to 20 ms permanent segments is performed. Hence, the segmentation procedure became more complex. This led to increased interest in phonetic or quasi-phonetic (phonetic segments) segmentation. When the bulk of calculations are shifted to the segmentation procedure, then decision procedures on beyond acoustic levels can be layed-out more clearly.

Results of previous research [1, 2, 6] indicate that segments formed in the process of time decomposition of a speech signal can be divided into the following qualitative classes:

a) stationary segments,
b) transient segments,
c) short segments,
d) pause (silence).

The "overlapping" effect of physical structures of adjoining phonems is a serious problem in speech signal segmentation. For example:

— a voiced and an unvoiced phonem have overlapping formants and noise structures,
— a phonem and silence (or the other way around) appear at the beginning and end of a word or sequence.

In this case the main problems related with the application of phonetic segmentation are:

1. Selection of such a segmentation algorithm which would allow reproducible generation of a time function $P(t)$, then used as a basis for the determination of boundaries of segments, or their midpoints.

2. Selection of such parametrization methods which would ensure an effective description of the $P(t)$ function.

3. Determination of segmentation criteria and mean of the base code of obtained segments.

This paper is aimed at the investigation of possibilities of segmentation into phonetic segments on the basis of the phonetic function of speech FFM, including chosen parametrization and verification criteria.

## 2. Methods

### *2.1. Segmentation algorithm in the recognition process*

As segmentation is to be applied in the continuous speech recognition system, it has to be performed with signal parametrization at the same time, but points on the axis, which denote boundaries of segments should be sent to the parametrization block in sequences. Figure 1 presents a simplified block diagram of an ASR input with segmentation.

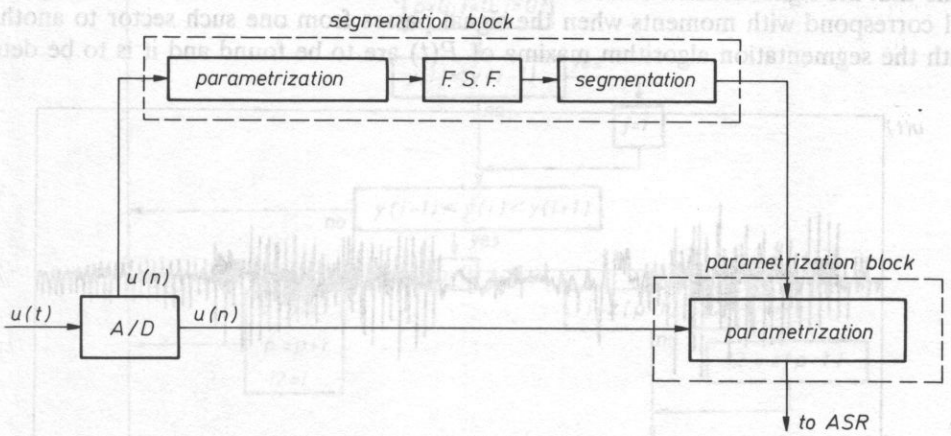


FIG. 1. Block diagram of the acoustic input system of the ARM system.

The input signal $u(t)$ is converted into digital form and then fed independently to two blocks: segmentation block and block of parametric discription. The segmentation block detects boundaries of segments in the arriving signal and transfers this information to the block of parameter extraction in the form of time marks. This is a general scheme, which has to be extended, by such elements for example as signal's buffer store or parametrization block system with a system of syntax rules in a back coupling loop, depending on the model of recognition system. It should be noted that in such a segmentation model, parametrization in the segmentation block and the block of the phonetic function parameter extraction have completely separate functions, although they can apply similar or identical techniques of parameter calculation, e.g. spectral analysis, linear prediction coding or distribution of time internals between zero-crossings of the speech signal.

### 2.2. *Segmentation procedure*

Three functional parts can be separated in the segmentation block:

1) parametrization,

2) calculation of the phonetic function of speech (F.F.M.), acting as a time function $P(t)$,

3) detection of segments boundaries on the basis of $P(t)$.

Following samples of the input signal $u(n)$ are grouped into time windows with length $t_n$ and for every time point $t$ a vector of parameters describing in a definite manner the state of the signal in the definite time window, is calculated (see part 3). Vectors of parameters are used to calculate $P(t)$ according to formula [2]

$$P(t) = \frac{1}{P} \sum_{p=1}^{p} \alpha_p * \left[ \ln \frac{\mathbf{R}(t,p)}{R[(t-\tau),p]} \right]^2 \tag{1}$$

where: $\mathbf{R}(t,p)$ — vector of parameters in time window $t_n$, $\alpha_p$ — weight of $p$-parameter, $P$ — number of parameters, $\tau$ — time shift.

Then boundaries of segments are determined on the basis of function $P(t)$. If we assume that the signal consists of short quasistationary sectors then boundaries of segments will correspond with moments when the signal passes from one such sector to another. With the segmentation algorithm maxima of $P(t)$ are to be found and it is to be deter-

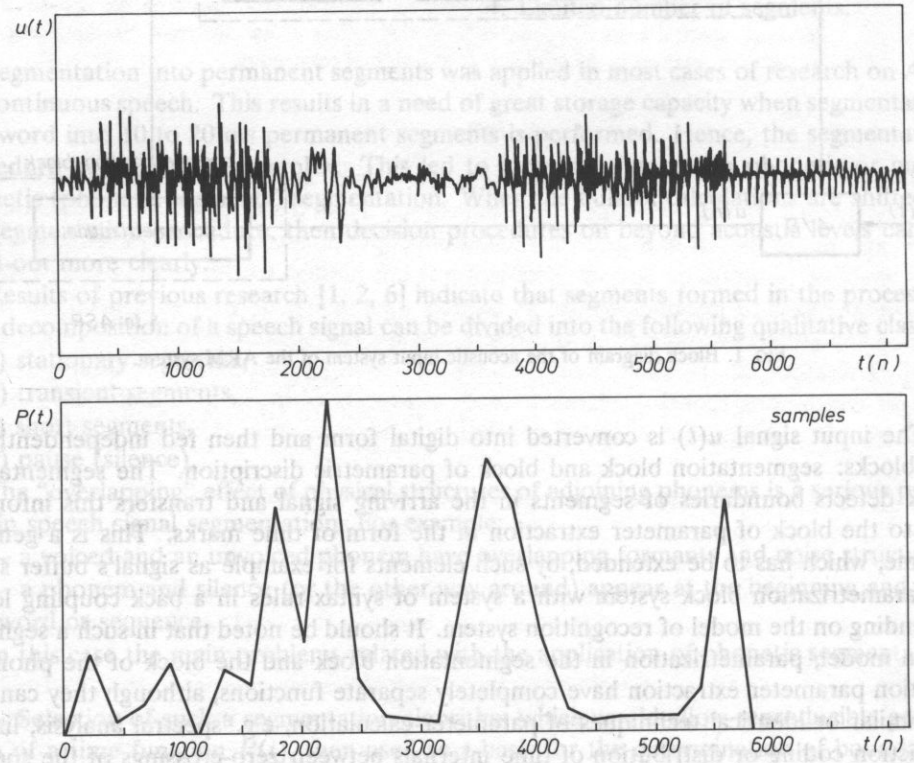


FIG. 2. Speech signal $u(t)$ and its phonetical speech function $P(t)$.

mined on the basis of certain definite additional criteria (discussed further in the paper after the algorithm itself is described) whether the instance indicated by the maximum can be a boundary between two successive segments. For example, Fig. 2 presents the signal of the word /oɕem/ („eight") and the phonetic function of speech for this signal achieved with the described above method with the use of parameters from FFT analysis.

The time function $P(t)$ has several local maxima which denote the boundaries between phonetic segments. They appear as an effect of even small changes in the signals spectral structure, specially within noise segment of speech phoneme, /ɕ/ and in segments accepted

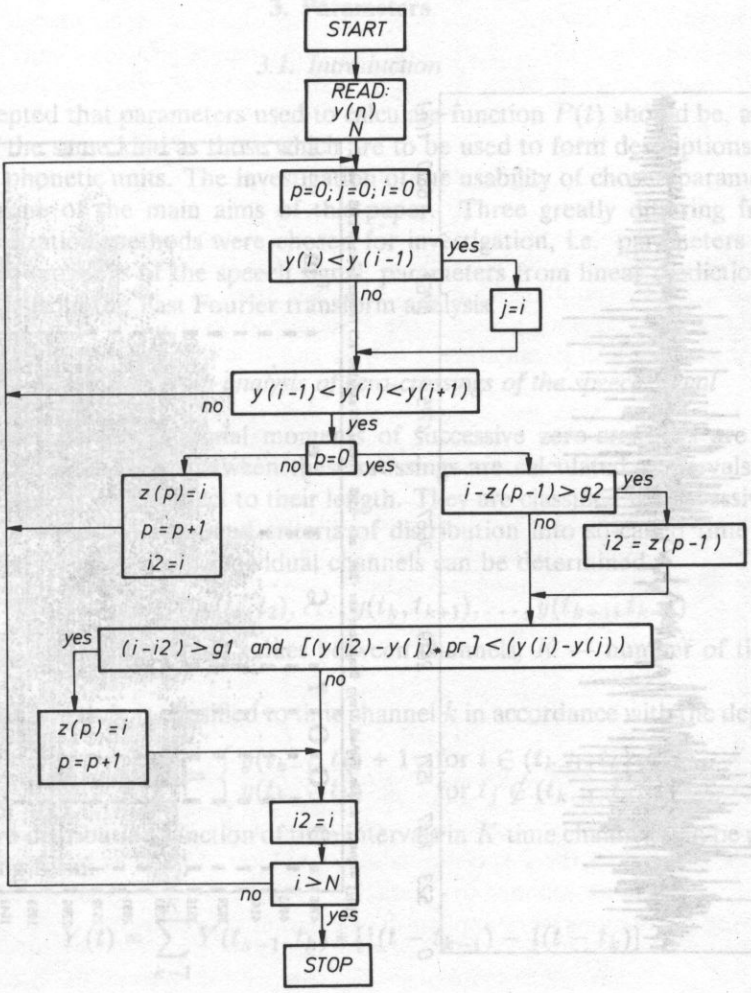


FIG. 3. Flow chart for segment detection algorithm.

as quasistationary (e.g. vowels) for individual or emotional changes in the articulation. The elimination of points denoting local maxima from the set of all hypothetical boundaries of segments is the greatest problem in the realization of the segmentation algorithm. There are two possible approaches here:

1. Application of adequate smoothing filters,
2. expert algorithm created on the basis of heuristic knowledge.

We have accepted the second approach in this paper. The block diagram of the algorithm is shown in Fig. 3.

Succeeding local maxima of function $P(t)$ are found in the algorithm. Every maximum is accepted as a prototype for determination of the boundary between phonetic segments. However, it is accepted that a maximum corresponds with the boundary between segments

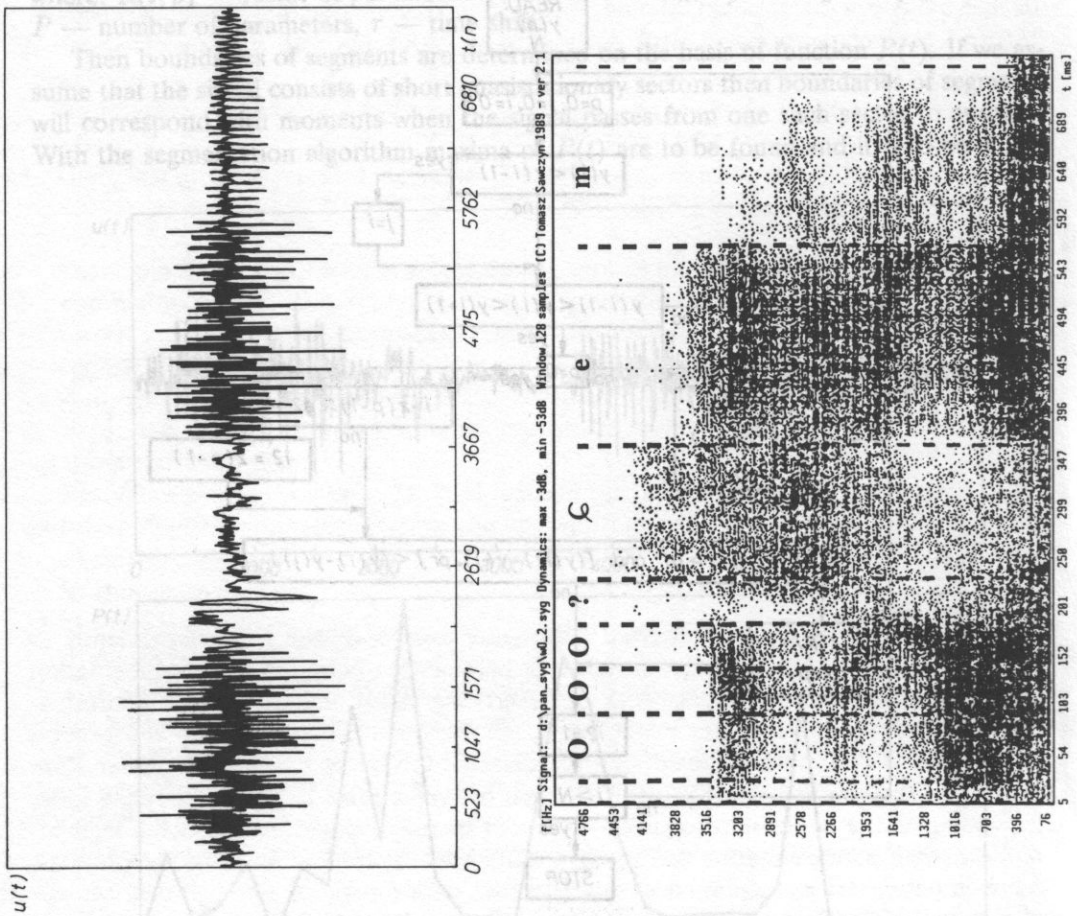


FIG. 4. Speech signal and its digital spectrogram with marked segments and phonetic description.

only after three conditions are checked:

1. the distance in time between the current checked maximum and preceding maximum can not be less than the minimal distance, denoted by $g1$, accepted from experiment;

2. if the distance between the current local maximum $mb$ and local maximum $mp$, found directly before it, exceeds the complex maximal distance $g2$, then the maximum qualified as the boundary between segments is accepted as the last local maximum $mp$;

3. the relative value of the maximum has to exceed the assumed fraction of the relative value of the last maximum $(pr)$

Figure 4 presents the same signal as in Fig. 2 and its digital spectrogram word /osiem/ („eight") with marked boundries of segments obtained from the described algorithm.

## 3. Parameters

### *3.1. Introduction*

It was accepted that parameters used to calculate function $P(t)$ should be, as far as it is possible, of the same kind as those which are to be used to form descriptions (images) of recognized phonetic units. The investigation of the usability of chosen parametrization methods was one of the main aims of this paper. Three greatly differing from each other parametrization methods were chosen for investigation, i.e. parameters from the analysis of zero-crossings of the speech signal, parameters from linear crediction coding, and parameters from the Fast Fourier transform analysis.

### *3.2. Parameters from analysis of zero-crossings of the speech signal*

For the analysed speech signal moments of successive zero-crossings are detected and lengths of time intervals between these crossings are calculated. Intervals are then grouped and ordered with respect to their length. They are classified to successive groups according to previously determined criteria of distribution into so-called time channels [3]. The number of intervals in individual channels can be determined

$$y(t_d) = (t_0, t_1), y(t_1, t_2), \ldots, y(t_k, t_{k+1}), \ldots, y(t_{k-1}, t_{k=g}) \tag{2}$$

where: $t_d, t_1, \ldots, t_g$ — threshold values between channels, $K$ — number of time channels.

An interval with length $t_j$ is classified to time channel $k$ in accordance with the dependence

$$y(t_{k-1}, t_k) = \begin{cases} y(t_{k-1}, t_k) + 1 & \text{for } t \in (t_{k-1}, t_k) \\ y(t_{k-1}, t_k) & \text{for } t_j \notin (t_{k-1}, t_k) \end{cases} \tag{3}$$

The cumulative distribution function of time intervals in $K$ time channels can be presented in the following form:

$$Y(t) = \sum_{k-1}^{K} Y(t_{k-1}, t_k) * [1(t - t_{k-1}) - 1(t - t_k)] \tag{4}$$

where:

$$1(t) = \begin{cases} 0 & \text{for } t \ll 0 \\ 1 & \text{for } t > 0 \end{cases}$$

### 3.3. *Spectral parameters* (FFT)

FFT calculation algorithms given by G.D. BERGLAND and M.T. DOLAN [10] were used to calculate spectral parameters. Applied parameters satisfy equation

$$X(k) = \sum_{n=0}^{N-1} x(n) * e^{-j\frac{2\pi}{N}-nk} \tag{5}$$

where $x(n)$ — real input sequence, $X(k)$ — complex transform coefficients, $k, 0, 1, \ldots$ $\ldots, N/2$.

The signal was described with parameters, which were the signal's energies in 16 one-third octave frequency channels or 6 octave channels. The energy in individual frequency bands was calculated in accordance with the following formula:

$$Y_i = \frac{1}{K_i - P_i} * \sum_{j=P_i}^{K_i} [\mathrm{re}\, X(j)]^2 + [\mathrm{im}\, X(j)]^2 \tag{6}$$

where

$$P_i = f_i * N/f_{pr}\,,$$
$$K_i = f_{i+1} * N/f_{pr}\,,$$

$f_i$ — boundary frequency between the $(i-1)$ and $i$ frequency band, $N$ — number of FFT samples, $f_{pr}$ — sampling frequency of signal $x(n)$.

Before FFT was calculated, signals were standarized to the energy equal to unity. A Hamming's window with parameters $\alpha = 0.54$ and $\beta = 0.46$ was superimposed onto the signal.

### 3.4. *Linear prediction coding* (LPC) *parameters*

Algorithms given by J.D. MARKEL and A.H. GRAY Jr. [10] were applied to calculate these parameters. Linear prediction methods model the signals spectrum through a "only poles" type filter with transmittance function:

$$H(z) = G/A(z) \tag{7}$$

where

$$A(z) = 1 + \sum_{k=1}^{P} a_k z^{-1}$$

is the reverse filter function, $G$ — amplification factor, $a_k$ — prediction coefficient, and $p$ is the number of poles or prediction coefficients of the model. If $H(z)$ is stable, then $A(z)$ can be realized in the form of a ladder-type filter. Coefficients of reflection which are related in a definite manner to prediction coefficients, are used to describe signals. The following expressions were applied in calculation procedures:

$$\begin{aligned} a_m^{(m)} &= k_m \\ a_j^{(m)} &= a_k^{(m-1)} + k_m a_{m-1}^{(m-1)}, \quad 1 \le j \le m-1 \end{aligned} \tag{8}$$

where $a_j^{(m)}$, $1 \le j \le m$ $m$ — prediction coefficients order.

## 4. Experimental and results

### *4.1. Introduction*

The basic questions, to which our research was to provide answers belong to two problem ranges. Within the first one, the dependence between the effectiveness of segmentation and type of parameters applied in calculations of the phonetic function of speech was to be investigated. While within the second range the effectiveness of automatic segmentation, i.e. carried out with the described algorithm, was compared with manual segmentation, i.e. conducted by an experienced phonetician.

The first one of these experiments was to give a confirmation of the thesis saying that it is possible to apply in FFM analysis parameters achieved from a simple and fast time analysis. This would be of great importance in the realization of systems working in real time. Results of the second experiment were also expected to confirm the given above thesis. They were also supposed to create a better understanding of the problem of segmentation into phonetic segments by comparing automatic and manual segmentation.

### *4.2. Analysis of segmentation parameters*

Three types of parameters applied in FFM calculations were defined. A corpus of 10 words each repeated ten times by 2 speakers was analysed. Values of constants — $g1$, $g2$ and $pr$ — were changed three times for every type of parameters. These values were dependent on the length of applied windows. Table 1 contains values of these parameters for all examined cases.

**Table 1.** Values of parameters of segmentation with three methods. $P$ — number of parameters, $N$ — length of window, $t$ — shift of window, $g1, g2, g3$ — constants of segmentation algorithm, par 1, par 2, par 3 — sets of parameters.

| Method | RICZ | | | FFT | | | LPC | | |
|---|---|---|---|---|---|---|---|---|---|
| par. | par 1 | par 2 | par 3 | par 1 | par 2 | par 3 | par 1 | par 2 | par 3 |
| $P$ | 8 | 8 | 8 | 16 | 16 | 16 | 12 | 12 | 12 |
| $N$[pr] | 150 | 150 | 300 | 128 | 128 | 256 | 150 | 150 | 300 |
| [pr] | 75 | 75 | 150 | 64 | 64 | 128 | 75 | 75 | 150 |
| $g1$ | 1 | 5 | 3 | 2 | 5 | 3 | 1 | 5 | 2 |
| $g2$ | 15 | 11 | 8 | 13 | 13 | 7 | 15 | 13 | 13 |
| $pr$ | 0.1 | 0.1 | 0.2 | 0.1 | 0.2 | 0.2 | 0.1 | 0.2 | 0.2 |

In order to perform a comparative analysis a notion of a "coefficient of segmentation conformity" $wzs$, was introduced. Its definition is as follows:

$$wzs = \frac{1}{N} \sum_{i=1}^{N} \frac{lpw}{ls} \tag{9}$$

where $lpw$ — number of overlapping segments in all repetitions of a word, $ls$ — number of all segments in all repetitions of a word, $N$ — number of words.
The $wzs$ coefficient should be interpreted as a measure of segmentation reliability within one class. The greated the segmentation repeatibility, the better the definition of class standards, and hence increased recognition reliability.

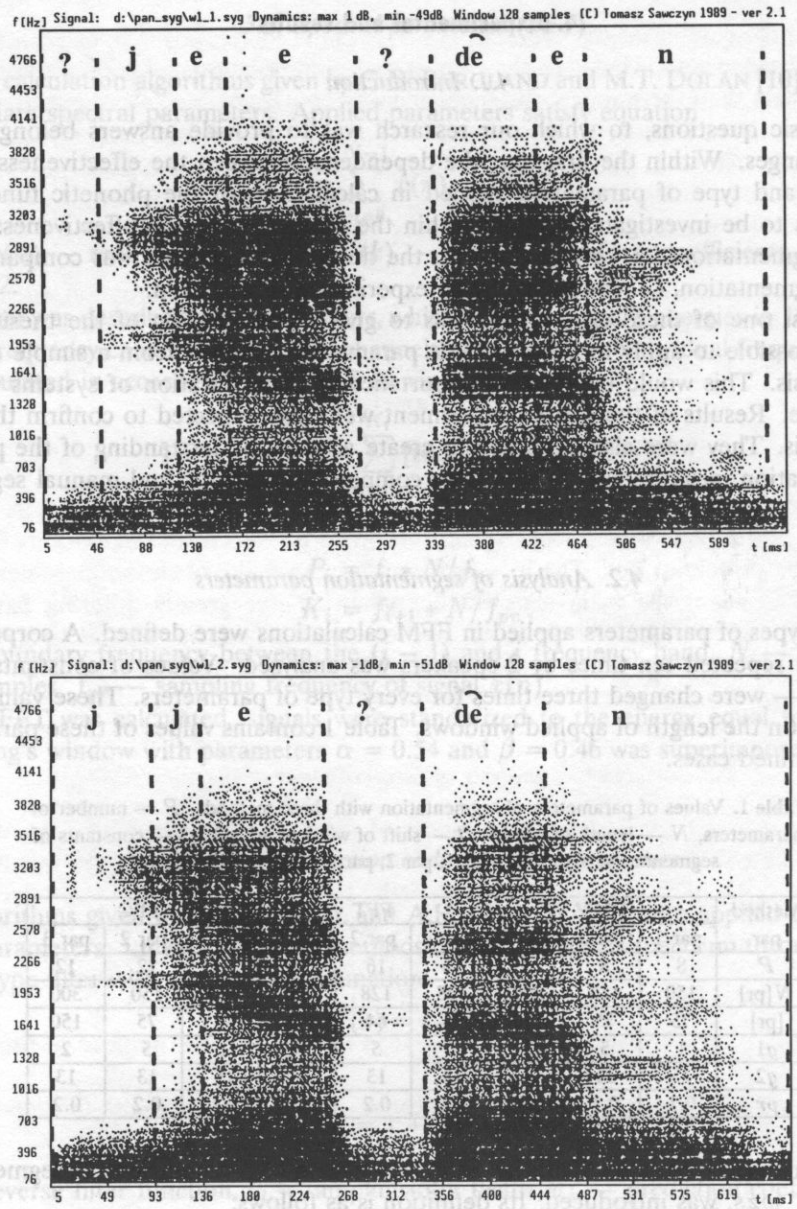


FIG. 5. Two repetitions of word /jeden/ (one) with marked segments.

Table 2 presents $wzs$ coefficients calculated for all 10 words, three parametrization methods and various values of parameters, while in Fig. 5 we have two repetitions of the word "jeden" with marked segments. Except for some cases, $wzs$ values considerably exceed the value of 0.5. High segmentation repeatibility within repetitions of the same words was stated on the basis of the analysis of mean values of the $wzs$ coefficient, contained in the range from 0.59 to 0.8.

**Table 2.** Values of coefficient "$wzs$" for 10 words (10 Polish digits from "one" to „zero") and three parametrization methods. par 1, par 2, par 3 — sets of parameters as in Table 1

| | RICZ | | | FFT | | | LPC | | |
|---|---|---|---|---|---|---|---|---|---|
| | par 1 | par 2 | par 3 | par 1 | par 2 | par 3 | par 1 | par 2 | par 3 |
| "1" | 0.88 | 0.92 | 0.85 | 0.85 | 0.82 | 0.82 | 0.88 | 0.85 | 0.85 |
| "2" | 0.76 | 0.85 | 0.65 | 0.72 | 0.65 | 0.58 | 0.79 | 0.55 | 0.65 |
| "3" | 0.52 | 0.67 | 0.55 | 0.62 | 0.51 | 0.49 | 0.85 | 0.48 | 0.68 |
| "4" | 0.65 | 0.68 | 0.50 | 0.65 | 0.58 | 0.58 | 0.85 | 0.65 | 0.65 |
| "5" | 0.71 | 0.75 | 0.55 | 0.75 | 0.68 | 0.45 | 0.78 | 0.55 | 0.62 |
| "6" | 0.85 | 0.85 | 0.62 | 0.65 | 0.52 | 0.35 | 0.75 | 0.30 | 0.45 |
| "7" | 0.85 | 0.88 | 0.65 | 0.75 | 0.78 | 0.75 | 0.85 | 0.68 | 0.72 |
| "8" | 0.79 | 0.82 | 0.75 | 0.82 | 0.65 | 0.65 | 0.78 | 0.55 | 0.62 |
| "9" | 0.56 | 0.65 | 0.65 | 0.92 | 0.85 | 0.85 | 0.84 | 0.72 | 0.78 |
| "0" | 0.65 | 0.75 | 0.72 | 0.89 | 0.75 | 0.68 | 0.65 | 0.55 | 0.68 |
| $wzs$ | 0.722 | 0.782 | 0.629 | 0.762 | 0.679 | 0.620 | 0.800 | 0.590 | 0.670 |
| | 0.11 | 0.09 | 0.07 | 0.09 | 0.11 | 0.15 | 0.06 | 0.15 | 0.1 |

Columns in Table 2 correspond with columns in Table 1, i.e. $wzs$ values in the $n$-column in Table 2 were achieved with the application of parameters given in the $n$-column of Table 1.

### *4.3. Analysis of segmentation correctness*

The analysis of segmentation correctness was performed on the basis of a comparison between manual and automatic segmentation time functions and digital spectograms of 120 words (40 words, repeated 3 times each) from a vocabulary defined in paper [4]. Manual segmentation was carried out by an experienced phonetician. For questionable boundaries, audio monitoring was also used. Automatic segmentation was done with the application of parametrization according to the method of analysis of zero-crossings with parameter values given in column 3 in Table 2. Figure 4 presents a spectrogram of the word /jeden/ (one) with segments calculated automatically and phonetical transcription resulting from such a segmentation.

As in the previous experiment, a certain measure was introduced to obtain a quantitative description of segmentation correctness. This is the coefficient of segmentation correctness — $wps$ — defined as follows:

$$wps = \frac{lsp}{ls} \tag{10}$$

where $lsp$ — number of correct segments resulting from the phonetic description, $ls$ — number of segments obtained from automatic segmentation.

These coefficients have been calculated for all phonems occurring in the tested vocabulary of 40 words. The results are gathered in Table 3.

It resulted from analysis that frequently more than one segment occurs within one phonem, generally vowels and voiced sounds. Several segments appear in those places of the signal, which do not have an interpretation in the collection of Polish phonems (e.g. silence preceeding plosive consonants), and that segments contain frequently two or more phonems (e.g. in the word "jeden" (one) /j/ merges with /e/ in "dwa" (/dva/-two) — /d/ with /v/ and "pięć" (/p'ɛɲt͡ɕ/-five) a combination of /ɲ/ with /t͡ɕ/ occurs. In order

**Table 3.** Values of coefficient "*wps*" for all phonemes occuring in the vocabulary under investigation *x* — denotes the lack of a phoneme in the vocabulary 0 — denotes the lack of segmentation conformability (phoneme not delected)

| phoneme | i | ɨ | e | a | o | u | p | b | t | d |
|---|---|---|---|---|---|---|---|---|---|---|
| *wps* | 0.47 | 0.50 | 0.68 | 0.57 | 0.68 | 0.73 | 1.36 | x | 2.81 | 3.75 |

| phoneme | c | ɟ | k | g | f | v | s | z | ʃ | ʒ |
|---|---|---|---|---|---|---|---|---|---|---|
| *wps* | x | x | 2.40 | 1.80 | 1.50 | 1.50 | 1.50 | 1.00 | 0.57 | 0.00 |

| phoneme | ɕ | ʑ | x | ts | dz | t͡ʃ | d͡ʒ | t͡ɕ | d͡ʑ | m |
|---|---|---|---|---|---|---|---|---|---|---|
| *wps* | 0.64 | x | 0 | 0 | x | 1.09 | x | 0.47 | 1.50 | 1.05 |

| phoneme | n | ɲ | ŋ | l | w | j | r |
|---|---|---|---|---|---|---|---|
| *wps* | 1.50 | 1.00 | 4.50 | 0.75 | 1.00 | 1.33 | 2.18 |

to describe these two latter phenomena the notions of: a coefficient of indeterminable segment and a coefficient of "merged" segment were introduced. They are denoted by $wsn$ and $wsz$, respectively, and defined as follows:

$$wsn = \frac{lsn}{ls}; \quad wsz = \frac{lsz}{ls} \tag{11}$$

where $lsn$ — number of all indeterminable segments, $lsz$ — number of all merged segments, $ls$ — number of all segments. The $wps$ coefficient can be interpreted on the basis of described above phenomena. If there is more than one phonetic segment within a phonem, then the $wps$ value exceeds one. The effect of merging or indetermination of a phonetic segments results in a $wps$ value below one. $wsn$ and $wsz$ values indicate the relative number of merged and indeterminable segments. The lower these values are the better the segmentation is. $wps$ values for vowel phonem range from 0.47 to 0.73, what means that on the average there was twice as many phonetic segments as phonem. There was also twice as many phonetic segments for unvoiced phonems /ɕ, ʃ, t͡ʃ/. Automatic segmentation was found to agree with manual segmentation *g* for phonem /z/ and /ɲ/. In other cases the $wps$ values exceeds 1 and for phonem /t, d, k/ it exceeds even 2. This value indicates a strong tendency to merge these phonems with order ones. $wsz$ and $wsn$ values are equal to 0.22 and 0.067 respectively.

## 5. Conclusions

This research is related in the sense of applied methods to an earlier paper by BASZTURA, JURKIEWICZ and TYBURCY [4]. Authors of this paper applied the procedure of the phonetic function of speech to segmentation in spectral parametrization. Owing to the dissemination of modern calculation techniques, including A/D conversion and new analysis methods of the speech signal [3], it was possible to widen and continue this research. A similar conception of segmentation can be also found in the mentioned previously paper [1].

The radically different parametrization methods were used in investigations:

— spectral, one based on FFT calculations,

— temporal, based on the analysis of zero-crossings,

— linear prediction coding.

As it results from Table 2, among others, comparably good segmentation results were

achieved with all these methods. This means that the $P(t)$ calculation algorithm and division into phonetic segments is effective, and that the parameters are evaluated from such requirements, as the rate and complexity of the procedure. In respect this, the distribution of time intervals between zero-crossings of a signal meets accepted effectiveness postulates, expressed by coefficients $wzs$ and $wps$, and calculation rate, at best.

Other parameters, such as spectral or prediction for example used in the parametrization block (see Fig. 1) can be applied at the same time to determine boundaries of segments. This conclusions finds confirmation in Table 2.

Results grouped in Table 3 are of statistic, quantitative as well as qualitative meaning. The complexity of the problem of automatic segmentation can be exhibited on the basis of physical parameters. A wide span of values of the $wps$ coefficient indicates how difficult it is to establish uniform criteria, which would lead to a reduction of the effect of phoneme separation into several segments ($wps < 1$), or merging of two phonems into one ($wps > 1$). These tendencies are also quantitatively defined by introduced here coefficients of merged and indeterminable segments.

On the basis of results achieved from our research a thesis can be formulated that this method of segmentation can advantageously influence the solution to the problem of continuous speech recognition. This thesis should find confirmation in experiments of automatic recognition of phonetic segments and recognition of words and continuous speech on this basis. The topic will be included among others, in further research which will apply "explicite" segmentation in automatic recognition of voices independently of the content of the statement. This is voice recognition with initial recognition of model linguistic elements, applied as key elements [11].

## References


[1] R. ANDRE-OBRECHT, *A new statistical approach for the automatic segmentation of continuous speech signals*, IEEE Trans. on Acoustics, Speech and Signal Processing, **36**, 1, 29–39 (1988).

[2] Cz. BASZTURA, J. JURKIEWICZ, E. TRYBURCY, *Phonetic function of speech F.F.M. Applied in continuous speech signal segmentation* (in Polish) Archiwum Akustyki, **4**, 4, 121–130 (1979).

[3] Cz. BASZTURA, *Acoustic sources signals and images* (in Polish), WKiŁ, Warszawa 1988.

[4] Cz. BASZTURA, W. MAJEWSKI, W. BARYCKI, *Effectiveness estimation of parameters of a global description of words in simple systems of automatic speech recognition* (in Polish), (placed for printing in Archives of Acoustics).

[5] A.M.L. DIJK-KAPPERS, *Comparison of parameter sets for temporal decomposition*, Manuscript no. 62 Institute of Perception Research, Eindhoven, The Netherlands, T.b.s.t.: Speech Communication 1988.

[6] A.M.L. DIJK-KAPPERS, S.M. MARKCUS, *Temporal decomposition of speech*, Manuscript no 608/11 Institute for Perception Research Eindhoven. The Netherlands. T.b.s.t.: Speech Communication 1988.

[7] D.J. HERMES, *Vowel-onset detection*, Manuscript no 601/11 Institute for Perception Research Eindhoven, The Netherlands, T.b.s.t: Speech Communication 1988.

[8] J. KACPROWSKI, R. GUBRYNOWICZ, *Automatic recognition of Polish vowels using a method of spectrum segmentation* In: Speech Analysis and Synthesis Ed. W. Jassem vol.2 PWN, Warszawa 1970 pp. 51–170.

[9] W.A. LEA [Ed], *Trends in speech recognition*, Prentice Hall Inc. Englewood Cliffs 1980.

[10] *Programs fordigital signal processing* „IEEE Press 1979.

[11] J. ZALEWSKI, Text dependent speaker recognition in noise. Proc. Eurospeech 89 vol.1 Paris 1989 p. 287–289.


*Received November 28, 1989; English version September 17, 1990*