Turkish Journal of Electrical Engineering and Computer Sciences

Volume 26 | Number 6 Article 10

1-1-2018

Improved method of heuristic classification of vowels from an acoustic signal

JOSEF KROCIL

ZDENEK MACHACEK

JIRI KOZIOREK

RADEK MARTINEK

JAN NEDOMA

See next page for additional authors

Improved method of heuristic classification of vowels from an acoustic signal

Authors
JOSEF KROCIL, ZDENEK MACHACEK, JIRI KOZIOREK, RADEK MARTINEK, JAN NEDOMA, and MARCEL FAJKUS

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2018) 26: 2900 – 2914

doi:10.3906/elk-1801-292

Improved method of heuristic classification of vowels from an acoustic signal

Josef KROCIL[1*], Zdenek MACHACEK[1], Jiri KOZIOREK[1], Radek MARTINEK[1]
Jan NEDOMA[2], Marcel FAJKUS[2]
[1]Department of Cybernetics and Biomedical Engineering, Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, Ostrava, Czech Republic
[2]Department of Telecommunications, Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, Ostrava, Czech Republic

Received: 12.02.2018 • **Accepted/Published Online:** 08.06.2018 • **Final Version:** 29.11.2018

Abstract: This paper describes research in the field of the improved methodology of the classification of vowels /a, a:/, /ɛ, ɛ:/, /ɪ, i:/, /o, o:/, and /u, u:/ (vowel symbols according to IPA, i.e. International Phonetic Alphabet). The aim is to develop an improved method enabling the automatic allocation of vowel symbols to the corresponding time segments of acoustic recordings of an undisturbed speech signal. The combined classification method is based on finding frequencies of the first two local maxims (formants) in a smoothed linear predictive amplitude spectrum (LPC, linear predictive coding) and zero-crossing values of each speech active voiced short-term segment of the recording. Based on these monitored values, simple heuristic conditions are arranged for the classification of the respective vowel. Implementation of the algorithm was realized using the MATLAB environment and its Graphical User Interface (GUI) was used for the user interaction. Verification of the success rate of vowel classification was done using recordings of forty speakers (twenty men and twenty women), where each speaker repeated the vowels repeatedly with short successive pauses. The success rate of recognizing vowels is classified and evaluated based on results obtained from our designed method.

Key words: Speech recognition, classification, linear prediction, cepstrum, formants, vowels, acoustic signal

1. Introduction

Human speech is represented by a set of sounds, with a voiced and unvoiced character. The voicing of speech is created by periodic oscillation of tense glottis caused by a flow of air from the lungs. Unvoiced speech is generated by free air flow through a relaxed glottis. Voiced speech sounds can be used to distinguish the gender of the speaker or specific person. Based on the spectral properties of vowels a specific vowel can be identified, usually according to the position of the first two or three local maxims in the LPC amplitude spectrum, so-called formants [1]. Chougala and Shridar [2] proposed a method for the calculation of formant frequencies, which is based on a combination of results of searching formants in the LPC cepstral envelope and finding the roots of the denominator polynomial of the transfer function of the vocal tract. In the paper for recognition of vowels in fluent speech, the authors described an improved method based on searching the positions of the first two formants using an LPC amplitude spectrum and a retroactive check of the time period of the recognized vowel; see [3] and [4].

Vowel automation recognition is based on detection of significant acoustic part of accents in the distributions of the formants for both vowels and diphthongs. Formants are the resonant frequencies of the vocal tract

*Correspondence: josef.krocil@vsb.cz

and additionally they also affect speaker and accent characteristics. Acoustic models are realizable by speaker characteristics for consistent performance and in language learning tools where detection can provide feedback to the user. Accent refers to a pattern of pronunciation in the use of vowels or consonants and intonation. For automatic speech recognition, in particular, the dimensional analysis is more attractive than the formant analysis.

Vowels in speech are nearly periodic segments of the speech signal and they are approximately equal to the sum of sinusoids with harmonically related frequencies. There is a possibility for examining vowels in the frequency domain. There are placed peaks that correspond to the harmonics that make up the periodic signal. The peaks are spaced more closely in some plots than others, corresponding to a lower fundamental frequency and thus a longer fundamental period. The fundamental frequency is independent of the vowel being produced. The overall shape of the frequency spectrum is different between the two vowels but remains relatively constant. These peaks are called formants and are known as the position primary feature that distinguishes one vowel from another. See [5], [6], [7], [8], [9], [10], [11], [12], and [13]

This paper presents a new approach in vowel classification. Furthermore, this paper describes a newly proposed efficient method for distinguishing voiced and unvoiced sections in speech signals. The algorithm for the classification of vowels is made up of several linked parts. First, the discrete speech signal is preprocessed, i.e. modified appropriately (normalization of amplitude, etc.). Short-term equidistant sections of the recording, so-called frames or segments, in which speech activity was detected by the intensity detector, are then subject to a voicing analysis (cepstral analysis and zero-crossing calculation). Then, for a given voiced segment, an LPC model is created in whose amplitude spectrum the frequencies of the first two formants are searched for. Based on the formant frequencies and zero-crossing values, a certain vowel is allocated to the segment using heuristic rules. The individual steps of the algorithm are described in detail in the following sections; see also [14] and [15].

2. Properties of vowels

In terms of acoustics, all vowels are voiced. It is possible to see a quasiperiodic profile with high amplitude in their acoustic signals. During the passing of the basic vocal cord tone F0 through the vocal tract, areas with a higher concentration of acoustic energy are produced, which are represented in the frequency profile by so-called formants, which are visible as local maxima in the amplitude spectrum. The position of formant frequencies is affected by the shape and cross-section of the oral and throat cavity. A change of the dimensions of the anatomical components, such as lips and tongue, will change the formant structure of sound and thereby create various vowels. With respect to time duration, vowels can generally be divided into short and long, which is evident from Czech vowels. In the case of short vowels (/a/, /ɛ/, /ɪ/, /o/, /u/) the mean duration time ranges from 40 ms to 160 ms, and in the case of long vowels (/a:/, /ɛ:/, /i:/, /o:/, /u:/) it is 80 ms to 320 ms (vowel symbols according to IPA, i.e. International Phonetic Alphabet). As was mentioned above, the monitored parameters are mainly formants Fn, where n is the index of formant sequence. Typically, the first two to three formants are analyzed [16]. The values of formant frequencies are shown in Table 1.

3. Methods of formant analysis

3.1. Linear predictive analysis

The production of a speech signal over a short time section (typically ranging from 15 ms to 32 ms) can be expressed using Eq. (1). The aim is to determine the parameters of the speech production model under the

Table 1. Frequency values of first two formants of Czech vowels [3].

Vowels (IPA)	Formant F1 (Hz)	Formant F2 (Hz)
/ɪ/, /i:/	from 300 to 500	from 2000 to 2800
/ɛ/, /ɛ:/	from 480 to 700	from 1560 to 2100
/a/, /a:/	from 700 to 1100	from 1100 to 1500
/o/, /o:/	from 500 to 700	from 850 to 1200
/u/, /u:/	from 300 to 500	from 600 to 1000

presumption that the kth sample of discrete signal $s(k)$ can be described by a linear combination of Q previous samples in time and excitation $u(k)$. Here, the signal $u(k)$ is the equivalent of excitation impulses created by vocal cords.

$$s(k) = -\sum_{i=1}^{Q} a_i s(k-i) + Gu(k). \tag{1}$$

Coefficients a_i and gain G are the searched parameters of the model. Application of the Z transform to Eq. (1) will lead to Eq. (2) for the transfer function of model $H(z)$.

$$H(z) = \frac{G}{1 + \sum_{i=1}^{Q} a_i z^{-1}}. \tag{2}$$

Term $Gu(k)$ in Eq. (1) is unknown and therefore an equation for the short-term function of energy E of prediction error $e(k)$ must be introduced.

$$E = \sum_{k} e^2(k) = \sum_{k} [s(k) - \hat{s}(k)]^2. \tag{3}$$

Prediction error $e(k)$ is defined as the difference of the predicted sample value $\hat{s}(k)$ from the actual signal sample $s(k)$. By leaving out term $Gu(k)$ from Eq. (1) and inserting it into Eq. (3), the following equation is obtained:

$$E = \sum_{k} \left[s(k) + \sum_{i=1}^{Q} a_i s(k-i) \right]^2. \tag{4}$$

Energy function E of prediction error $e(k)$ has a minimum at a point where the partial derivative of energy E is minimal.

$$\frac{\partial E}{\partial a_j} = 0, \ 1 \leq j \leq Q. \tag{5}$$

After finding the relation for the minimum of function E according to Eq. (5) and the appropriate modifications, we get a set of linear equations (Eq. (6)), whose solutions are coefficients a_i. Eq. (6) pertains to the so-called autocorrelation method of the calculation of coefficients a_i because $R(\ldots)$ are the coefficients of

autocorrelation function of signal $s(k)$.

$$\sum_{i=1}^{Q} a_i R(|j-i|) = -R(j),\ 1 \le j \le Q. \tag{6}$$

Gain G can be obtained by the following equation:

$$E = R(0) + \sum_{i=1}^{Q} a_i R(i) = G^2. \tag{7}$$

The frequency response of the model of speech signal production can be obtained from Eq. (2), using substitution:

$$z = e^{j\omega}. \tag{8}$$

Usually the calculation of coefficients a_i and gain G is done using the iterative algorithm designed by Durbin and Levinson [3]. After inserting the linear prediction coefficients and substitution according to Eq. (8) into Eq. (2), the estimated smoothed amplitude or power spectrum can be plotted, as shown in Figure 1.

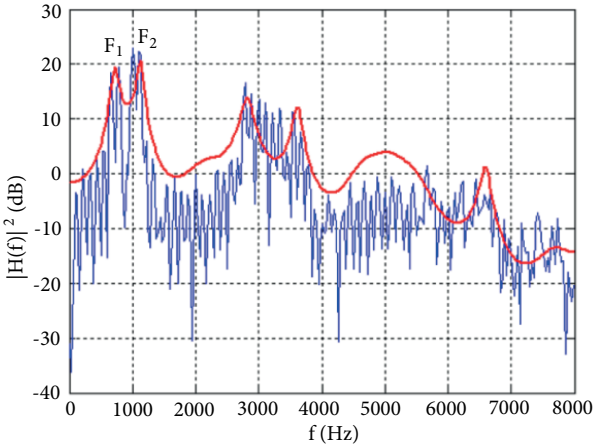

Figure 1. Power spectrum (blue) and power LPC spectrum of the voiced speech (red).

3.1.1. Searching for formants in the LPC spectral envelope

The easiest and least demanding method, in terms of calculation, of determining formant frequencies is searching for local maximums in the LPC spectral envelope. The found peaks of the envelope are then represented by the corresponding values of their frequencies. As additional information, the bandwidth of the envelope peaks can also be given for the purpose of refining the search strategy. Peaks whose bandwidth exceeds 500 Hz are then excluded from the analysis [17].

3.1.2. Calculation of poles of the transfer function

The values of formant frequencies may be determined based on the calculation of roots of the denominator polynomial of the transfer function of the LPC model of Eq. (2) according to the following equation:

$$z^Q + a_1 z^{Q-1} + \cdots + a_{Q-1} z + a_Q = 0, \tag{9}$$

whereas the resultant solution is dominated by pairs of complexly conjugated roots z_i and $\bar{z}_i$. Roots are calculated using, for example, the Newton–Raphson or Bairstow method. One pair of complex conjugated roots can be expressed as:

$$z_i = |z_i|\, e^{j\varphi_i},\ \bar{z}_i = |z_i|\, e^{-j\varphi_i}, \tag{10}$$

where φ_i is the complex number argument.

$$F_i = \frac{\omega_i}{2\pi} = \frac{arg(z_i)}{2\pi T}\ (Hz), \tag{11}$$

and the bandwidth B_i can be calculated according to the following equation:

$$B_i = -\frac{ln\,|z_i|}{\pi T}\ (Hz), \tag{12}$$

where T is the period of sampling of the acoustic speech signal.

This mentioned approach of solving complex roots usually does not allow the use of this method in real time; therefore, algorithms based on searching formants in the LPC spectral envelope are applied more often.

3.2. Nonstandard methods

The predominant methods for determining formants are based on linear predictive analysis. During the research of other methods of formant analysis, methods based on the sinusoidal decomposition of speech vibrations or on modeling sounds using Markov models [18] were used. Furthermore, we can also mention the application of learning classifiers, e.g., algorithm k-NN (k nearest neighbors), where Mel cepstral coefficients (MFCC) [19] were used for the classification of vowels.

4. Design and implementation of algorithm

Analysis of the speech signal is performed on acoustic recordings in the offline mode. For selected algorithms, and mathematical relations that do not require elaboration for the actual solution, the following text shows links to the respective literature sources.

4.1. Making of recording and its preprocessing

Records of speech signals are obtained as mono-channel recordings in wav format with a sampling frequency F_s of 16 kHz and a bit depth of 16 bits per sample. After loading the vector of samples obtained from the signal its mean value is subtracted from it and then amplitude normalization (to values ranging from –1 to 1) is performed. The next step is the performance of the signal preemphasis, given by the following equation:

$$\hat{S}(k) = S(k) - \alpha S(k-1), \tag{13}$$

where $\hat{S}(k)$ is the value of the sample of signal $S(k)$ after preemphasis and α is a constant ranging from 0 to 1. During implementation into MATLAB the value was set to 0.95. In the frequency domain the preemphasis is equivalent to the characteristic of a first-order high-pass numerical filter. The purpose of the preemphasis is to highlight the formant structure in the speech signal spectrum, which has a positive effect on the results of the formants analysis. The preprocessing algorithm is shown in Figure 2.

4.2. Segmentation of the signal and detection of speech activity

4.2.1. Segmentation of speech signal

The analysis of the speech signal must be done in short time intervals for which the processed signal can be regarded as stationary. From recording $\hat{S}(k)$ short-term segments, 32 ms (512 samples per segment) are extracted and weighted by the Hamming window function. The segments do not link up together directly in time (segmentation without overlap), but they overlap each other by half of their length. The first half of the samples of the current segment therefore consists of the second half of the previous segment's samples. This is segmentation with a half overlap, in which gradual signal changes, caused for example by coarticulation, can be better detected. The maximum number of extractable segments q_{max} can be calculated from the following equation:

$$q_{max} = floor\left(\frac{M-N}{\tau}\right), \tag{14}$$

where M is the total number of recording samples $\hat{S}(k)$ and N is the required number of samples per segment. Symbol τ represents the segmentation shift, i.e. 256 samples in the case of the half overlap. The segmentation of the signal $\hat{S}(k)$ for $0 \leq q \leq q_{max}-1$ is given by the following equation:

$$s_{q+1}(k) = w(k)\hat{S}(q\tau + k), \tag{15}$$

where $s_{q+1}(k)$ is the extracted segment and $w(k)$ is the Hamming window function, whereas (further with respect to the indexation of elements in MATLAB):

$$1 \leq k \leq N. \tag{16}$$

Segmentation according to Eq. (15) is integrated into the algorithm for the detection of speech activity, where the signal intensity is monitored for a specific extracted segment.

4.2.2. Detection of speech activity in segments

For the detection of speech activity, a simple detector of signal intensity with a variable detection threshold [20] was chosen, where signal intensity I of the short-term segment can be calculated using the following equation for $0 \leq q \leq q_{max}$–1:

$$I_{q+1} = \frac{1}{N}\sum_{k=1}^{N}\left|s_{q+1}(k)\right|. \tag{17}$$

For this type of detector it is important to ensure that the start of the analyzed recording (e.g., for 1 s) contains only ambient noise without speech activity. This segment has a certain number of initial segments q_{init} (selected by the user). The calculation of the average value of intensity from the initial segments (Figure 3) gives the detection intensity I_{init}, which is used for the calculation of the threshold intensity I_r.

For each following segment, the signal intensity is calculated, which is compared to the threshold value I_r. If the intensity in the analyzed segment is higher than the threshold value, the segment is classified as speech active, and in the opposite case as speech inactive. Speech inactive segments are used in speech pauses for the calculation of updated threshold intensity I_r (calculated according to [20]). Information about speech activity is stored in the A_SEG_{q+1} variable, where a speech active segment is represented by the value 1,

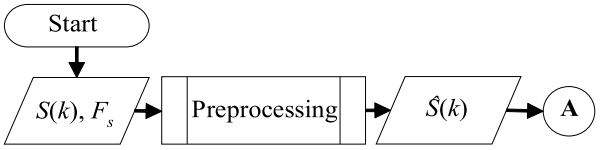

Figure 2. Preprocessing of speech signal.

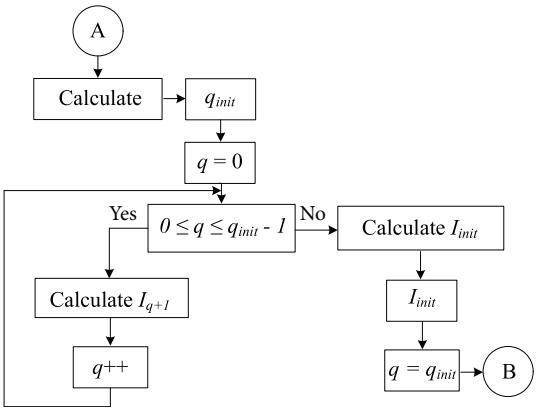

Figure 3. Initialization algorithm for the intensity detector.

otherwise by value 0. For each speech active segment the mean number of zero-crossings Z is calculated and stored, according to the following equation:

$$Z_{q+1} = \sum_{k=1}^{N-1} \left| sgn\left[s_{q+1}(k+1)\right] - sgn\left[s_{q+1}(k)\right] \right|, \tag{18}$$

where the function sgn is defined as:

$$sgn\left[s_{q+1}(k)\right] = \begin{cases} 1, s_{q+1}(k) \geq 0 \\ -1, s_{q+1}(k) \leq 0. \end{cases} \tag{19}$$

The values of Z_{q+1} are used in the next algorithm step as one of two criteria for determining the voicing of segments. The speech detection algorithm is shown in the flow chart (Figure 4).

4.3. Voiced/unvoiced analysis of speech active segments

4.3.1. Spectral analysis

Information about the presence of the basic vocal cord frequency F_0 in the speech active segment (if $A\_SEG_{q+1} = 1$) can be obtained using the real cepstrum, according to the following equation:

$$c_{q+1}(n) = Re\left\{IFFT\left\{log\left|FFT\left\{s_{q+1}(k)\right\}\right|\right\}\right\}, \tag{20}$$

where (I) FFT is (inverse) fast Fourier transform and $c_{q+1}(n)$ are cepstral coefficients of the analyzed segment. If the segment is speech passive (i.e. $A\_SEG_{q+1} = 0$), the analysis of the speech activity of the next segment is continued. The linearity of the Fourier transform together with the nonlinearity of the logarithmic function will cause the separation of frequency components of excitation (vocal cords) from the frequency components of the vocal tract (formants). This is because in the short-term segment we can consider signal $s_{q+1}(k)$ as the product of convolution of the excitation signal and impulse response of the vocal tract. The inverse Fourier transform of the logarithm module of the amplitude signal spectrum $s_{q+1}(k)$ provides a real cepstrum (time domain), where the peak index is searched for (Figure 5), which represents the basic vocal cord period T_0 after

recalculation. T_0 is given by the following equation:

$$T_0 = \frac{argmax_{40 \le n \le 224}\,[c(n)]}{F_s}, \tag{21}$$

where F_s is the sampling frequency; see [21] and [22].

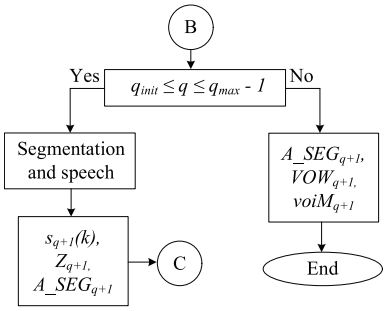

Figure 4. Simplified flowchart of speech activity detection algorithm and zero-crossing rate computation.

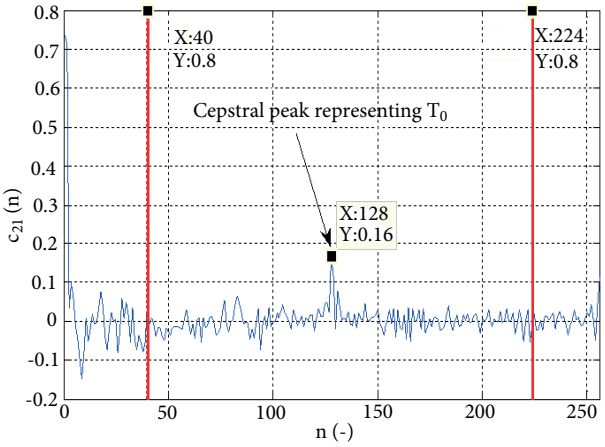

Figure 5. Real cepstrum of voiced speech segment.

The interval in which the maximum peak is searched is derived from the presumed physiological range of T_0 or F_0, respectively, i.e. ranging from 70 Hz to 400 Hz, which corresponds to cepstral coefficients $c_{q+1}(n)$, whose index n is $40 \le n \le 224$. Searching for the maximum in this area is supplemented by the requirement for this maximum to be at least six times higher than the mean value only of positive cepstral coefficients in the given interval (determined experimentally). If the searched maximum in the cepstrum of the given segment is found, value 1 is assigned in the $cepM_{q+1}$ variable, and in the opposite case 0 is assigned (Figure 6).

4.3.2. Determining the voicing of active segments

Determining the voicing of a segment is a criterion for initiating the formant analysis. Segment voicing is determined based on the condition, whose input parameters are values Z_{q+1} and $cepM_{q+1}$. If the value Z_{q+1} is smaller than or equal to 480 and, at the same time, if a cepstral peak, i.e. $cepM_{q+1} = 1$, is found, then the segment is declared as voiced and value 1 is assigned to variable $voiM_{q+1}$ (0 if unvoiced). If the segment voicing condition is fulfilled, formant analysis is initiated, or it is refused the speech activity analysis of the next segment can be started (Figure 7). The threshold value for zero-crossing is based on the relation between frequency f and zero-crossing Z in a short-term interval of 512 samples length, which can be expressed by the following equation:

$$f = \frac{1}{2}\frac{Z}{32ms}, \tag{22}$$

where $Z = 480$ and the ideal sine or cosine signal is the resultant value f equal to 7500 Hz. In reality, the voiced speech signal in a short-time section is rather quasiperiodic; nevertheless, in the analysis of voicing, the values for signal zero-crossing can be used as a good indicator with satisfactory results. The width of the utilizable band of the speech spectrum is approximately from 70 Hz to 7500 Hz, whereas values from approx. 300 Hz to

7500 Hz are related mainly to formant characteristics of the vocal tract, where voiced speech is involved (vowels, nasals, etc.). Frequencies above 7500 Hz are not too important with respect to the formant analysis of vowels. High Z values usually mean the presence of noise, particularly in the case of unvoiced sounds (fricatives, etc.).

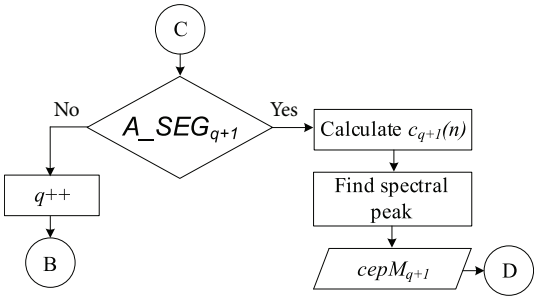

Figure 6. Real cepstrum computation and cepstral peak localization algorithm.

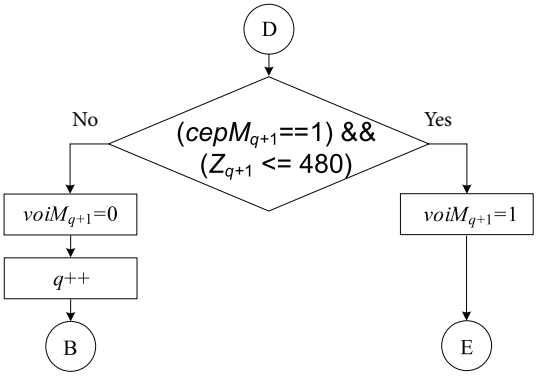

Figure 7. Voiced/unvoiced classification algorithm.

4.4. Analysis of formants and setting of classification conditions

4.4.1. Analysis of formants

The analysis of formants is based on the searching of frequencies of the first two formants in the LPC amplitude spectrum of the speech segment. The order of linear prediction Q equals 20 based on calculation [23] according to the following equation:

$$Q = \frac{F_s}{1000} + 4, \tag{23}$$

where as the Durbin–Levinson algorithm is used for the calculation of prediction coefficients and gain. The formants bandwidth is not calculated because it is presumed that during efficient classification of segment voicing and optimal prediction order the bandwidth of the first two formants will not exceed 500 Hz. After finding the frequencies of the first two formants F1 and F2, these frequency values are written to the $forreg_{q+1}$(F1, F2) variable (Figure 8). In MATLAB the $forreg$ variable is represented by the $q+1$ index at the line position, and values F1 and F2 are stored in the first and second variable columns, respectively. The values of formant frequencies together with the zero-crossing values are used as features for the classification of vowels in the next algorithm step; see [24] and [25].

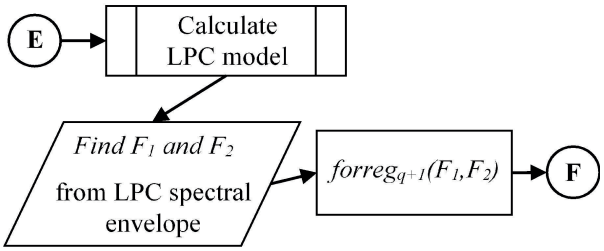

Figure 8. LPC spectral envelope computation and formant extraction algorithm.

4.4.2. Arrangement of conditions for vowel classification

The classification of vowels is derived from typical values of the frequency range of formants F1 and F2 and the values of zero-crossing in the voiced segment. The analysis uses values stored in the $forreg_{q+1}$(F1, F2) and Z_{q+1} variables. Due to the variability of these values for each individual (throat cavity shape, etc.), heuristic rules for the classification of vowels for the respective segment were determined experimentally. In the case of satisfying one of the conditions, a symbol representing a specific vowel is allocated to the respective segment. Long and short vowels are not differentiated based on the range of formant frequencies (Table 1). If the analyzed segment does not satisfy the decision-making conditions for vowels, then an empty symbol is allocated to it. These symbols are stored in the VOW_{q+1} variable. If other segments are available, analysis of the speech activity of the next segments continues (Figure 9) [26].

The advantage of the heuristic approach is the absence of any training data (obtained by clustering, for example); however, the classification result is dependent on the judgment of an expert, who defines the classification rules. If necessary, these rules can be further ramified, which increases the probability of a correct result, but it also increases the complexity of the designed system. The aim is to always determine suitable representative tags and limit the ramification of conditions. The set of rules for vowel classification is defined in the following form:

$$
\begin{aligned}
&\forall q \in [0, q-1], q \in \mathbb{Z} \\
&If\ forreg_{q+1,1} \in [650, 1050] \wedge forreg_{q+1,2} \in [1050, 1550] \wedge Z_{q+1} \in [64, 320] \\
&\quad VOW_{q+1} \leftarrow \text{`}a\text{`}; \\
&Else\,If\ forreg_{q+1,1} \in [500, 750] \wedge forreg_{q+1,2} \in [1450, 2100] \wedge Z_{q+1} \in [192, 384] \\
&\quad VOW_{q+1} \leftarrow \text{`}\varepsilon\text{`}; \\
&Else\,If\ forreg_{q+1,1} \in [200, 550] \wedge forreg_{q+1,2} \in [1950, 3050] \wedge Z_{q+1} \in [154, 380] \\
&\quad VOW_{q+1} \leftarrow \text{`}\imath\text{`}; \\
&Else\,If\ forreg_{q+1,1} \in [450, 750] \wedge forreg_{q+1,2} \in [760, 1200] \wedge Z_{q+1} \in [52, 256] \\
&\quad VOW_{q+1} \leftarrow \text{`}o\text{`}; \\
&Else\,If\ forreg_{q+1,1} \in [200, 450] \wedge forreg_{q+1,2} \in [500, 1000] \wedge Z_{q+1} \in [51, 320] \\
&\quad VOW_{q+1} \leftarrow \text{`}u\text{`}; \\
&Else \\
&\quad VOW_{q+1} \leftarrow \text{`'}; \\
&End\,If
\end{aligned}
$$

5. Experimental verification of the designed method

To verify the efficiency of the heuristic classifier, recordings of ten people were made: twenty men (speakers S1 to S20) and twenty women (speakers S21 to S40), where each person has ten speech sessions of successively spoken vowels, separated by short pauses. The intensity detector then detected the respective segments of the

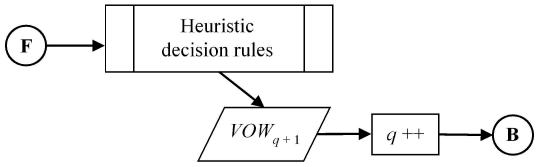

Figure 9. Heuristic vowel classification algorithm.

speech signal, and then analyzed and classified them. The results of vowel segment classification are recorded in the signal's time profile (Figure 10). The success rate of vowel classification is determined by the number of correctly classified segments in the detected section. If the ratio of successfully classified segments to the total number of segments is higher than 0.5, then the respective section is allocated to the given vowel. The results of the success rate of the implemented algorithm are shown in the occurrence rate tables (Table 2 and Table 3). The results show significantly different values of the total classification success rate between certain speakers. This is given mainly by the experimental setting between intervals of measured values (F1, F2, and number of zero-crossings), which allocate the respective vowels to speech segments. The lowest overall success rate applies to vowels /a/ and /u/, which were classified correctly in 66 of 100 measurements, i.e. 66%, and 67 times respectively, i.e. 67% of measurements. In the case of speakers S1, S2, S6, S7, S12, S13, and S18 (men) the vowel /a/ was proclaimed as vowel /o/ in the majority of the ten measurements. Vowel /u/ was incorrectly identified as vowel /ɪ/, especially in the case of speaker S24 (woman, only 1 good estimate out of 10) and speakers S5 and S9 (men, only 2 correct estimates). The success rate of the classification of the vowel /o/ (74.5% male speakers, 71.5% female speakers) is incorrectly identified as vowel /a/ especially for speakers S23, S36, and S40 (women, only 1 correct estimate). The best overall results apply to vowels /ɛ/ (estimate success rate 81.5% male speakers, 86.5% female speakers) and /ɪ/ (94.5% male speakers, 97.5% female speakers), whereas vowel /ɛ/ is successfully identified 3 times in the case of speakers S1 and S19 (men) and in the case of the remaining speakers the individual absolute occurrences are six to ten successful estimates. The occurrence rate of the correct estimates of vowel /ɪ/ is 8 to 10 for the individual male speakers and 9 to 10 for the individual female speakers, which makes it the most successful estimate of all tested vowels. The range of the overall success rate of classification of all vowels by a specific speaker is from 56 percent to 90 percent. In the case of the analysis of separately spoken vowels it is not unconditionally necessary to analyze their voicing, because these were always voiced. In the case of the analysis of vowels and generally voiced sections contained in the word, the analysis of voicing is desirable and forms the basis for calculation of the mean value of F0, and based on this also for determining the speaker's gender. Figure 11 shows the time profile of the Czech word /tfas/ (translation to

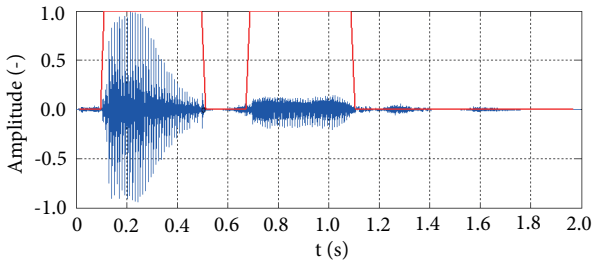

Figure 10. Speech signal of separately spoken vowels /a/ and /u/ (blue) and speech activity detection results (red).

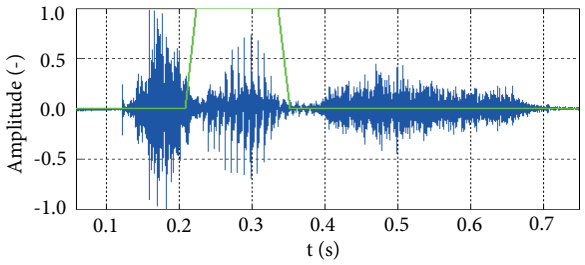

Figure 11. Speech signal of word /tfas/ (blue) and its voiced part /a/ (green).

English: time) with a green-highlighted voiced section (result of voicing analysis), which pertains to the voiced segments of vowel /a/. The results of vowel classification in segments are also shown for this section. The future aim is to increase the algorithm's efficiency, especially for its use in fluent speech, where worse classification results are achieved due to coarticulation than in the case of separately spoken vowels.

Table 2. Results of vowel classification (male speakers).

Male speakers	**Vowels**					(%)
	/a/	/ɛ/	/ɪ/	/o/	/u/	
S1	0	3	9	10	10	64
S2	0	10	9	7	10	72
S3	8	10	10	7	10	90
S4	10	10	10	4	4	76
S5	10	10	10	10	2	84
S6	0	10	9	8	10	74
S7	1	9	9	6	10	70
S8	5	8	10	6	6	70
S9	10	10	10	10	2	84
S10	8	6	10	7	9	80
S11	7	6	8	7	9	74
S12	0	10	8	7	6	62
S13	0	10	9	6	6	62
S14	8	8	10	5	5	72
S15	8	10	10	5	10	86
S16	10	6	10	10	6	84
S17	6	8	10	7	7	76
S18	0	6	8	10	6	60
S19	10	3	10	10	4	74
S20	8	10	10	7	4	78
(%)	54.5	81.5	94.5	74.5	68.0	

6. Conclusion

This paper describes the design and implementation of an improved combined method for the classification of vowels in a speech signal. The algorithm is based on short-term segmentation and the detection of the speech signal, from which only segments of a voiced character are extracted based on the results of spectral analysis and zero-crossing values. Based on the monitored values (first two formants in the LPC spectral envelope and zero-crossing values) the segments are then allocated the significance of a specific vowel, using a simple heuristic classifier. The efficiency of the algorithm was tested on separately spoken vowels. The future aim is its modification to achieve better results when applied to fluent speech. During testing, repeated recordings of vowels from ten people (twenty men, twenty women) were made. The highest classification success rate was achieved for vowel /ɪ/, i.e. 97.5%, and the lowest for /a/, i.e. 54.5%. In the case of the remaining two vowels the success rate was 66% to 94.5%. Errors of vowel classification in segments were caused mainly by

Table 3. Results of vowel classification (female speakers).

Female speakers	**Vowels**					(%)
	/a/	/ɛ/	/ɪ/	/o/	/u/	
S21	9	7	10	10	8	88
S22	10	9	10	9	6	88
S23	5	10	10	1	9	70
S24	5	6	10	7	1	58
S25	9	10	10	8	7	88
S26	9	10	10	8	6	86
S27	10	7	10	10	7	88
S28	7	9	10	10	8	88
S29	10	9	10	8	7	88
S30	6	10	9	7	6	76
S31	5	7	10	3	7	64
S32	8	10	9	8	5	80
S33	5	7	9	8	8	74
S34	9	10	10	9	6	88
S35	10	10	10	10	4	88
S36	5	6	9	1	7	56
S37	9	10	10	8	8	90
S38	5	6	9	10	6	72
S39	7	10	10	7	7	82
S40	5	10	10	1	9	70
(%)	74.0	86.5	97.5	71.5	66.0	

the allocation of another vowel or the allocation of an empty symbol, which can be caused by the merging of formants or background recording noise (effect on values of zero-crossing). The possible disadvantage can lie in the setting of the classifier rules, which is designed based on the subjective judgment of an expert; nonetheless, these rules can be easily modified for the further improvement of results. The advantage of the algorithm is mainly the simplicity and possible implementation in applications operating in real time. The quality of achieved results corresponds to the presumptions and possibilities of the presented newly designed combined method. In comparison to other familiar methods the newly developed combined method is comparable to currently used methods in terms of classification correctness; however, it is substantially simpler. This fact enables an increase in calculation speed and the use of the method for digital equipment with lower performance.

Acknowledgments

This work was supported by project SP2018/170 and SP2018/160 VSB-TU Ostrava. This work was also supported by the European Regional Development Fund in the Research Centre of Advanced Mechatronic Systems project, project number CZ.02.1.01/0.0/0.0/16_019/0000867 within the Operational Programme Research, Development, and Education.

References

[1] Young-Giu J, Mun-Sung H, Sang L. Development of an optimized feature extraction algorithm for throat signal analysis. ETRI J 2007; 29: 292-299.

[2] Uribe A, Gomez A, Bastidas M, Quintero OL, Campo D. A novel emotion recognition technique from voiced speech. In: IEEE 2017 Automatic Control; 18–20 October 2017; Cartagena, Colombia. pp. 1-4.

[3] Stanek M, Polak L. Algorithms for vowel recognition in fluent speech based on formant positions. In: IEEE 2013 Telecommunications and Signal Processing; 2-4 July 2013; Rome, Italy. pp. 521-525.

[4] Sung JL, Byung K, Hoon Ch, Yunkeun L. Intra- and inter-frame features for automatic speech recognition. ETRI J 2014; 36: 514-517

[5] Lobanov B M. Classification of Russian vowels spoken by different speakers, J Acoust Soc Am 1971; 17 49: 606-608.

[6] Yan Q, Vaseghi S, Rentzos D, Ho CH. Analysis and synthesis of formant spaces of British, Australian, and American accents. IEEE T Audio Speech 2007; 15: 676-689.

[7] Adank P, Smits R, Van Hout R. A comparison of vowel normalization procedures for language variation research. J Acoust Soc Am 2004; 116: 3099-3107.

[8] Zahorian SA, Kelkar S, Livingston D. Formant estimation from cepstral coefficients using a feedforward memoryless neural network. In: IEEE 1992 International Joint Conference on Neural Networks; 7–11 June 1992; Baltimore, MD, USA. pp. 673-678.

[9] Mousmita S, Kandarpa S. Segmentation and classification of vowel phonemes of Assamese speech using a hybrid neural framework. Applied Computational Intelligence and Soft Computing 2012; 2012: 871324.

[10] Kaladharan N. A review of different speech coding methods. International Journal of Electrical and Electronic Engineering and Telecommunications 2017; 6: 96-103.

[11] Radova V, Psutka J, Muller L, Byrne W, Psutka JV, Ircing P, Matousek J. Czech Broadcast News Speech and Transcripts. 1st ed. Philadelphia, PA, USA: Linguistic Data Consortium, 2004.

[12] Stanek M, Sigmund M. Speaker distinction using vowel polygons: experimental study. In: IEEE 2015 International Conference Radioelektronika; 21–22 April 2015; Pardubice, Czech Republic. pp. 125-128.

[13] Mishra S, Bhowmick A, Shrotriya MCh. Hindi vowel classification using QCN-MFCC features, Perspect Sci 2016; 33: 28-31.

[14] In-Chul Y, Dongsuk Y. Robust voice activity detection using the spectral peaks of vowel sounds. ETRI J 2009; 31: 451-453.

[15] Martinek R, Kelnar M, Vanus J, Bilik P, Zidek J. A robust approach for acoustic noise suppression in speech using ANFIS. J Electr Eng 2015; 66: 301-310.

[16] Martinek R, Kelnar M, Vanus J, Koudelka P, Bilik P, Koziorek J, Zidek J. Adaptive noise suppression in voice communication using a neuro-fuzzy inference system. In: IEEE 2015 Telecommunications and Signal Processing; 9–11 July 2015; Prague, Czech Republic. pp. 382-386.

[17] Stanek M, Sigmund M. Finding the most uniform changes in vowel polygon caused by psychological stress. Radioengineering 2015; 24: 604-609.

[18] Huang X, Acero A, Hon H. Spoken Language Processing. A Guide to Theory, Algorithm, and System Development. 1st ed. Upper Saddle River, NJ, USA: Prentice Hall, 2001.

[19] Amami R, Ayed D, Ellouze N. An empirical comparison of SVM and some supervised learning algorithms for vowel recognition. Int J Intell Inform Process 2012; 3: 1-8.

[20] Prasad R, Sangwan A, Jamadagni H, Chiranth M. Comparison of voice activity detection algorithms for VoIP. In: IEEE 2002 Computers and Communications; 1–4 July 2002; Taormina-Giardini Naxos, Italy. pp. 530-535.

[21] Machacek Z. Analysis and elimination of dangerous wave propagation as intelligent adaptive technique. Lect Notes Artif Int 2011; 6592: 482-491.

[22] Vanus J, Smolon M, Koziorek J, Martinek R. Voice control of technical functions in smart home with KNX technology. IFIP Adv Inf Comm Te 2015; 455-462.

[23] Lawrence R, Schafer W. Introduction to digital speech processing. Found Trends Signal Process 2007; 1: 1-194.

[24] Vanus J, Smolon M, Martinek R, Koziorek J, Zidek J, Bilik P. Testing of the voice communication in smart home care. Lect Notes Comp Sci 2015; 5: 15.

[25] Youngjik L, Kyu-Woong H. Selecting good speech features for recognition. ETRI J 1996; 18: 29-40.

[26] Sandeep K. Performance evaluation of novel AMDF-based pitch detection scheme. ETRI J 2016; 38: 425-434.