Stanislaw GMYREK[1]

6. VOICED FRAME DETECTION IN AUTOMATIC SPEECH RECOGNITION

6.1. Introduction

A speech signal can be divided into voiced and unvoiced fragments. Voiced speech is associated with air impulses produced by vocal cords vibrating with frequency f_0. One of the major issues in automatic speech recognition systems is the estimation of this frequency. With this aim, a number of methods have been developed, the most effective of which are correlation algorithms. In these algorithms, frame voicing is decided based on the normalized correlation coefficient and then pitch frequency is determined [1]. Such an approach entails high computational complexity since the autocorrelation function must be determined for each frame. In order to reduce this computational complexity, it is worth determining voicing in advance and then calculating the autocorrelation only for the selected voiced frames. Literature describes numerous methods of classifying speech signal frames [1, 5], including efficient and popular methods based on energy and zero-crossing rate [4]. They are not flawless though, and this is why this paper proposes an approach in which energy is determined based on four envelopes, and averaging and normalizing guarantees independence from signal level and the loudness of the speaker.

6.2. Speech signal parametrization

The analysed method employs quasi-quadrature filter banks, and then calculates envelopes for the obtained narrowband signals [3]. In the next step, envelopes are averaged in four frequency bands and energy is determined in a particular subband for each signal frame. A flowchart demonstrating this procedure is shown in Fig. 1.

[1] Department of Acoustics, Multimedia and Signal Processing, Wroclaw University of Science and Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland, stanislaw.gmyrek@pwr.edu.pl

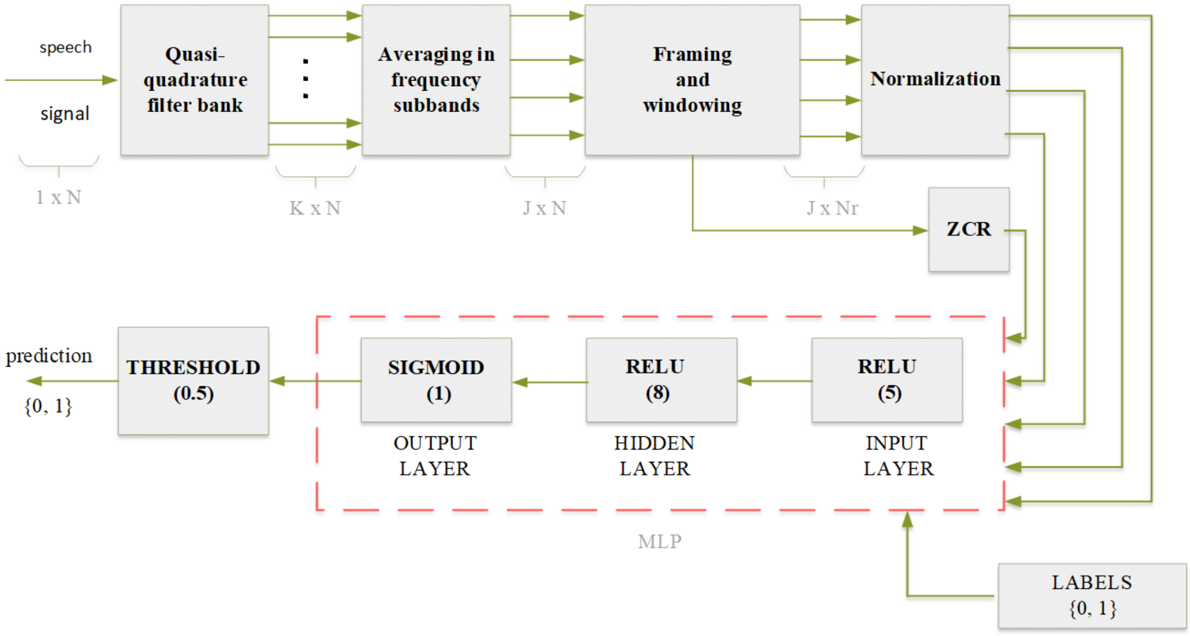

Fig. 1. The algorithm of speech signal parametrization aimed at determining frame voicing
Rys. 1. Schemat blokowy algorytmu parametryzacji sygnału mowy w celu podjęcia decyzji o dźwięczności ramki

6.2.1. Band filtering of speech signals

There is a speech signal *x(t)* with length *N*, registered digitally with sampling frequency f_p = 12 kHz. At the first stage it is filtered into narrow frequency subbands with the use of a complex filter bank. Each filter has a single pole located on a complex variable plane *z* at point $z_k = \rho_k e^{j2\pi f_k}$, where f_k: $k = 1, \ldots, K$ are middle frequencies in bands, and $\rho_k = \rho : k = 1, \ldots, K$; $\rho < 1$ is the pole module determining the width of the filter passband. The closer the value of ρ is to 1, the narrower the filter passband is. The value of ρ = 0.97 was adopted in subsequent calculations. The middle frequencies in filters are spaced at intervals of 50 Hz in the range from 200 Hz to c. 4500 Hz. The filtration process in subband *k* is recursive in character [3], i.e.

$$x_{\mathrm{r},k}(u) = x(u) + \mathrm{a}_1 x_{\mathrm{r},k}(u-1) - \mathrm{a}_2 x_{\mathrm{i},k}(u-1);\ u = 1, \ldots, N, \tag{1}$$

$$x_{\mathrm{i},k}(u) = \mathrm{a}_1 x_{\mathrm{i},k}(u-1) + \mathrm{a}_2 x_{\mathrm{r},k}(u-1);\ u = 1, \ldots, N \tag{2}$$

with initializations

$$x_{\mathrm{r},k}(0) = x(0),\ x_{\mathrm{i},k}(0) = x(0), \tag{3}$$

where $\mathrm{a}_1 = \rho\cos(2\pi f_k)$, $\mathrm{a}_2 = \rho\sin(2\pi f_k)$, while $x_{\mathrm{r},k}(u)$ and $x_{\mathrm{i},k}(u)$ are respectively the real and imaginary parts of the signal obtained at the complex filter output. The filter bank is presented in Fig. 2. Filter enhancement for middle frequencies f_k is stable and equals $1/1 - \rho^2$. The filter also introduces a slight time delay associated with the transient state [3].

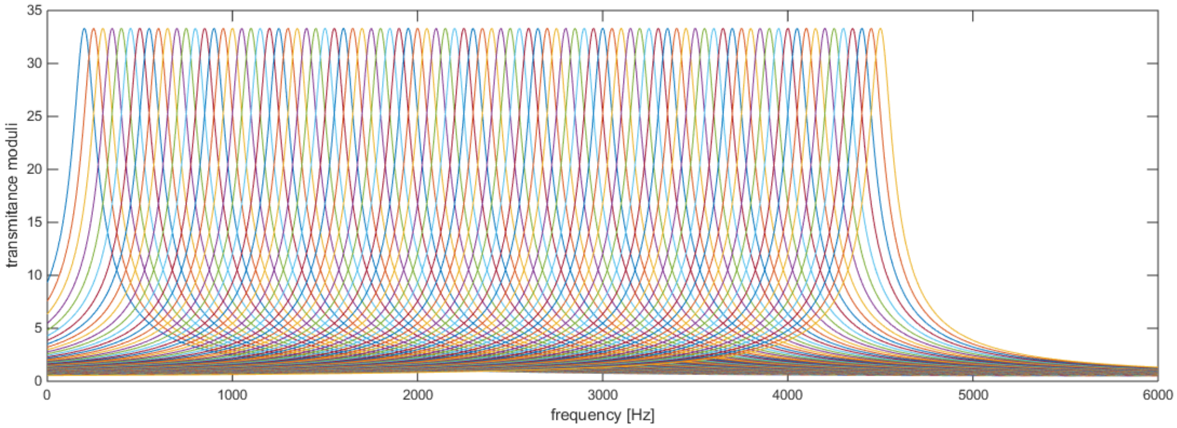

Fig. 2. Quasi-quadrature filter banks – filter transmittance modules

Rys. 2. Bank filtrów quasi-kwadraturowych – moduły transmitancji filtrówThe signal $x_{i,k}(u)$ can, with high accuracy, be treated as the Hilbert transform of a narrowband signal $x_{r,k}(u)$. Hence the signal pair $x_{r,k}(u)$ and $x_{i,k}(u)$ forms an analytic signal for a particular subband, making it possible to effectively determine the square of the envelope:

$$e_k(u) = x^2_{r,k}(u) + x^2_{i,k}(u). \tag{4}$$

Examples of envelopes for four selected frequency bands in the uttered word „ćwikła” are shown in Fig. 3.

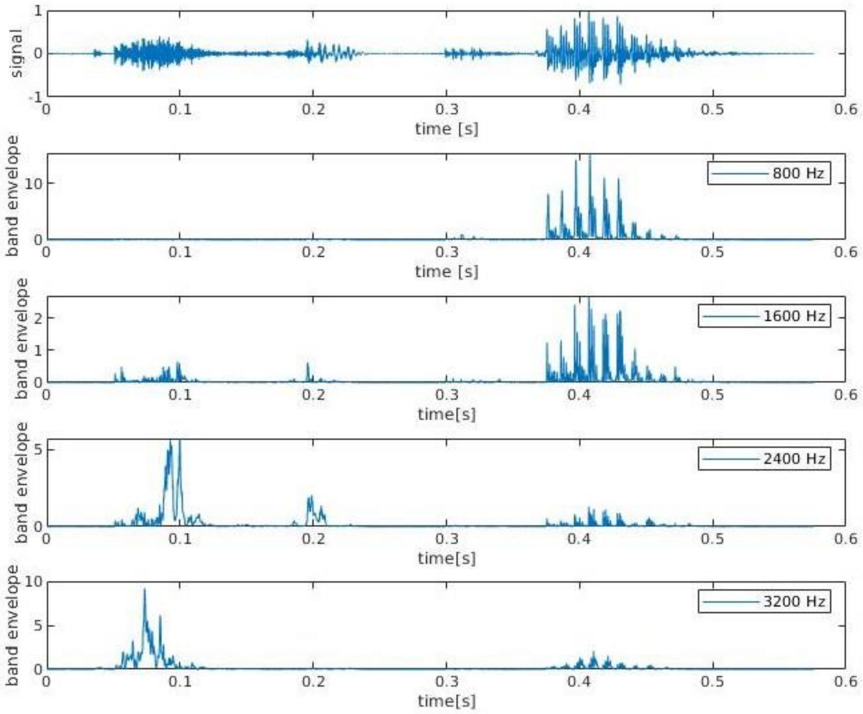

Fig. 3. Speech signal (word „ćwikła”) and the envelopes in four frequency bands

Rys. 3. Sygnał mowy (słowo „ćwikła”) i obwiednie w czterech pasmach częstotliwości

6.2.2. Envelope summation in frequency bands and averaging in speech signal frames

In order to reduce the extensiveness of the problem given the properties of the human hearing organ, the envelopes obtained from relations (4) are ascribed to one of $J = 4$ mel-frequency bands and then averaged within a particular band according to the relation:

$$E_j(u) = \frac{1}{n_j}\sum_{k=k_l}^{\mathrm{k_h}} e_k(u), \tag{5}$$

where k_l is the index of the lowest frequency band envelope ascribed to mel band j, k_h – the highest frequency index, and n_j – the number of components in a given band. The signals obtained in this way are shown in Fig. 4.

In the next step, the signal $x(t)$ and the envelopes $E_j(u)$ were divided into frames with the length N_r = 25 ms and the step N_s = 10 ms, and then windowed with a Hamming window. As a result, signals $E_{j,\,t}(m)$ were obtained, where $m = 1, \ldots, Nr$, j is the frequency subband index, $t = 1, \ldots, T$ is the frame number and T – the number of frames.

In order to make the discussed algorithm independent of signal level, total energy was calculated in each signal frame.

$$E_t^{\mathrm{c}} = \sum_{j=1}^{J}\sum_{m=1}^{Nr} E_{j,\,t}(m) \tag{6}$$

and normalized energies of the speech signal $\varepsilon_{j,t}$ were determined for each frame and frequency subband, i.e.:

$$\varepsilon_{j,t} = \frac{\sum_{m=1}^{Nr} E_{j,\,t}(m)}{E_t^{\mathrm{c}}}. \tag{7}$$

The parameter E_t^{c} from relation (7) can be interpreted as energy, as the signals $E_{j,\,t}(m)$ were determined based on the square of the envelope $e_k(u)$ from the output of the quasi-quadrature filter. The parameters obtained in this way are representative of the frame in the process of learning and deciding about voicing. Each number $\varepsilon_{j,t}$ is contained in the range [0,1] and $\sum_{j=1}^{J} \varepsilon_{j,t} = 1$.

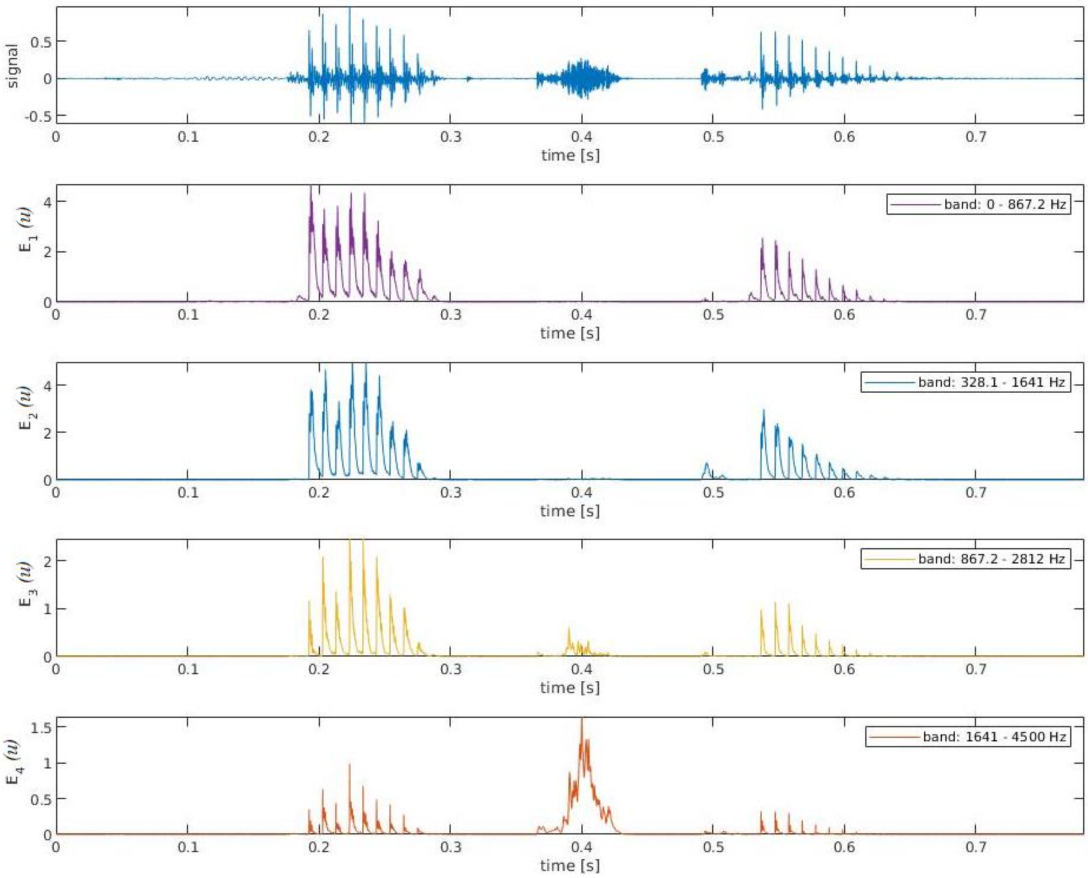

Fig. 4. A fragment of speech signal (word „ćwikła”) and four identified envelopes $E_j(u)$ averaged in mel frequency bands

Rys. 4. Fragment sygnału mowy (słowo „ćwikła”) i wyznaczone cztery obwiednie uśrednione w melowych pasmach częstotliwościThe parameters determined for an exemplary speech signal are shown in Fig. 5.

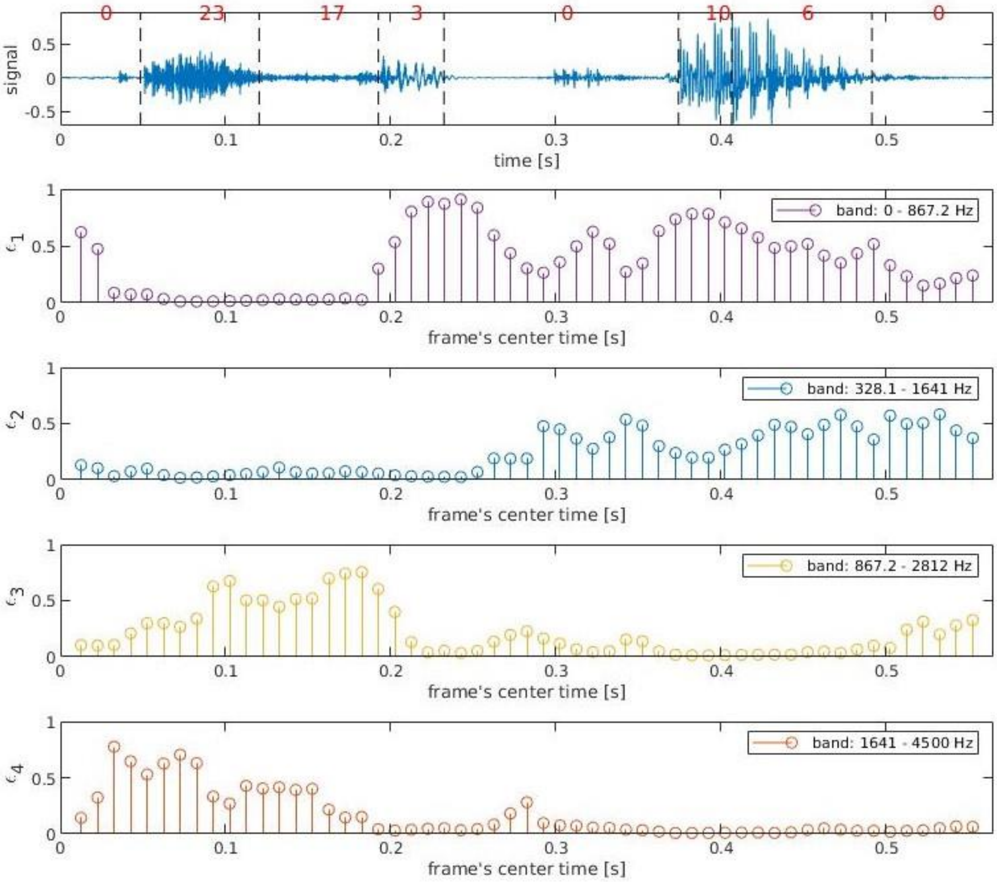

Fig. 5. Example - fragment of a speech signal (word „ćwikła”) and normalized energies in each frame for four frequency subbands

Rys. 5. Przykładowy fragment sygnału mowy (słowo „ćwikła”) i unormowane energie w każdej z ramek dla czterech podpasm częstotliwości

The first graph represents the time course of the registered speech signal – a recording of a male voice (the word „ćwikła”). Status and phoneme labels are shown at the top and values of parameters calculated for each frame according to relation (7) are plotted underneath. It is noteworthy that a high value of energy was obtained for voiced phonemes in the low frequency band ε_1, and a low value – in the high frequency band ε_4, while the situation is exactly opposite for unvoiced phonemes, which is a distinctive feature of the analysed problem.

6.2.3. Complementation of the parameter vector

The parameter vector was complemented with the zero crossing rate (ZCR) of the signal. The energy of voiced fragments of speech signals is concentrated in the low frequency band, and that of unvoiced fragments – in the high frequency band. The voiced fragments are related to a small number of zero crossings, while the unvoiced ones – to a large number [4]. Thus, in recognition terms, this parameter is an indicator enabling effective classification. The ZCR is calculated from the following formula:

$$\mathrm{ZCR} = \frac{\sum_{m=1}^{N} \operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]}{N}, \tag{8}$$

where N represents the length of the signal, and

$$\operatorname{sgn}[x(n)] = \begin{cases} 1, x(n) \geq 0; \\ -1, x(n) < 0. \end{cases} \tag{9}$$

6.3. Research results

The aim of the work is to select voiced frames of a speech signal by determining the values of detection thresholds. As a result, it is necessary to find a relationship between the input data and frame labels. With this aim, artificial neural networks (ANN) were used.

6.3.1. Neural network architecture

The database used for computational experiments consists of more than 10 800 recordings comprising individual words uttered by female and male voices. There were 725 000 speech signal frames, 70% of which constitute a training and validation set, while the others make a test set. The first architecture that was tested was the MLP (Multilayer Perceptron) neural network composed of three layers. The input layer and the hidden layer consisted of 8 neurons and a non-linear activation function RELU. As the problem discussed here concerns binary classification, the output layer was composed of one neuron with a sigmoidal-type activation function. This layer returns a number in the range from 0 to 1, which could be interpreted as an indicator of the probability of frame voicing. Subsequently, the value obtained from the output layer is compared with the detection threshold, whose value has been set as 0.5. Every value equal or larger than 0.5 is ascribed the label „1” (voiced frame), and every value below 0.5 - the label „0” (unvoiced frame). Binary mutual entropy was adopted as a loss function, as it produces good results in binary classification problems [2]. The input was a vector composed of four numbers, which were normalized energies $\varepsilon_{j,t}$. This enabled recognition efficiency of 90%. The result was unsatisfactory, so the input vector was complemented with the ZCR parameter in accordance with chapter 2.3. The architecture of the employed neural network is shown in Fig. 6.

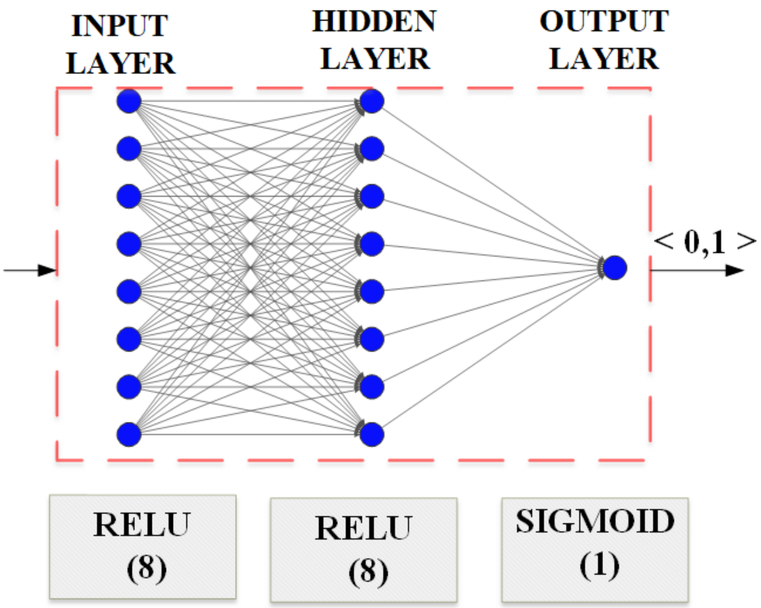

Fig. 6. Architecture of a neural network with an observation vector supplemented with a ZCR parameter
Rys. 6. Architektura sieci neuronowej z wektorem obserwacji uzupełnionym o parametr ZCR

6.3.2. Simulation results

The model training process was carried out with the use of 20 epochs and the batch size of 64 samples. The efficiency of the trained model reached c. 95%. The normalized confusion matrix is shown in Fig. 7.

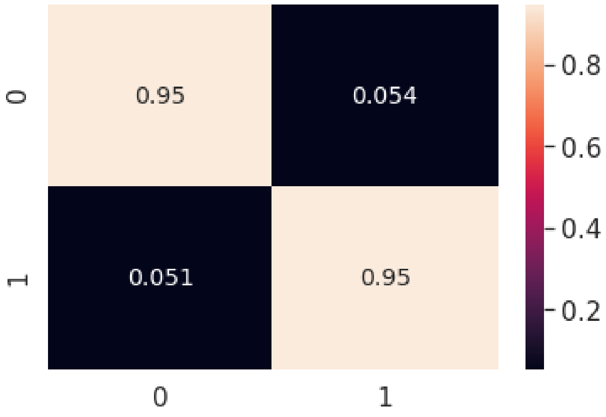

Fig. 7. The normalized confusion matrix
Rys. 7. Znormalizowana macierz pomyłek

This matrix is an effective method of verifying neural network efficiency. Its rows correspond to real labels, and the columns – to the labels calculated by the network. To assess the effectiveness and the correctness of the classification, precision and recall metrics were adopted. Their calculation requires the following parameters: TP (True Positive) – the value defining the number of frames correctly classified as voiced, FP (False Positive) – the number of frames incorrectly recognized as voiced, TN (True Negative) – the number of frames correctly classified as unvoiced, and FN (False Negative) – the number of frames incorrectly classified as unvoiced. The precision and recall measures are expressed by the following relations:

$$\text{precision} = \frac{TP}{TP + FP}; \tag{10}$$

$$\text{recall} = \frac{TP}{TP + FN}. \tag{11}$$

Additionally, the balanced measure of score coefficient F_1 was introduced, defined as the harmonic average of the precision and recall values:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2TP}{2TP + FP + FN}. \tag{12}$$

The classification accuracy measures obtained in simulations are presented in Table 1.

Table 1Parameters of neural network efficiency assessment

Parameter	value
Precision	0.949
Recall	0.950
F1 score	0.949

Fig. 8 presents the result of the algorithm's operation for an exemplary recording of a male voice. The first graph represents the time course of the speech signal for the word „ćwikła", the second one comprises true labels determined in the process of manual segmentation and labelling, while the last one depicts the decision about frame voicing taken by the neural network. The first phoneme „ć" is voiceless, so this fragment of speech signal was labelled as „0". In the signal, one can distinguish voiced phonemes such as „i", „ł", „a", associated with periodical stimulation of vocal cords. They were labelled with „1".

6.4. Conclusion

The paper proposes an efficient method of determining the voicing of speech signal frames. Compared to classic methods based on energy, averaging and normalization guarantees independence of signal level. The high energy value in the low-frequency band, and low in the high-frequency band indicate the voicing of the speech signal frame. In the first step, the classification accuracy reached 90%. In order to improve the recognition quality, zero-crossing rate was added to the parameter vector. The value of this parameter is low in voiced frames, and high in unvoiced frames. Thanks to this approach, the learning and classification efficiency rose to c. 95%. The proposed algorithm may be useful in many operations in automatic speech recognition systems. It saves time and reduces computational complexity in operations like estimating pitch frequency through preliminary selection of voiced frames and calculating the autocorrelation function only for selected frames.

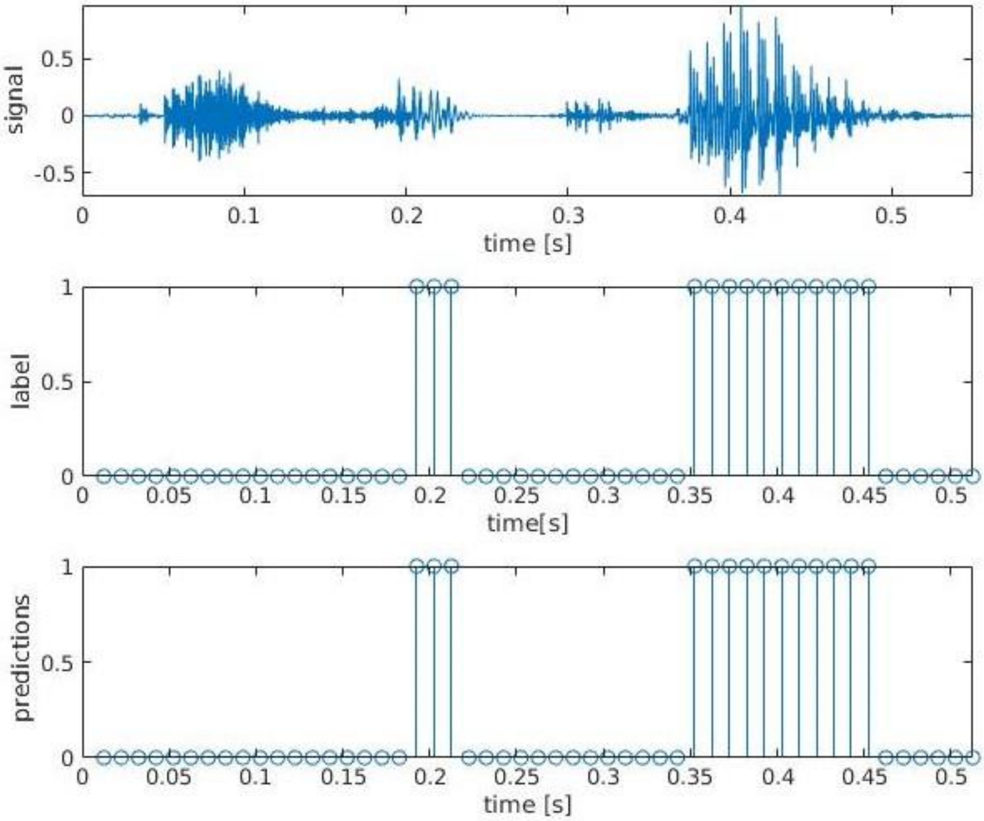

Fig. 8. Time course of the word „ćwikła” and comparison of network-generated labels with real labels
Rys. 8. Porównanie etykiet wygenerowanych przez sieć z etykietami prawdziwymi

Bibliography

1. Makowski R.: Automatyczne rozpoznawanie mowy - wybrane zagadnienia, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław 2011.
2. Chollet F.: Deep Learning with Python, Manning Publications, 2017.
3. Hossa R., Makowski R.: Voice activity detection with quasi-quadrature filters and GMM decomposition for speech and noise. Applied Acoustics, vol. 166, pp. 1-10, 2020.

4. Kopparthi S., Adapa B., Barkana B.D., Bachu R.: Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy. In: Elleithy K. (eds) Advanced Techniques in Computing Sciences and Software Engineering, Springer, Dordrecht, 2010. doi:10.1007/978-90-481-3660-5_47.

5. Rabiner L.R., Cheng M.J., Rosenberg A.E., McGonegal C.A.: A Comparative Performance Study of Several Pitch Detection Algorithms, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 24, no. 5, pp. 399-418, October 1976, doi:10.1109/TASSP.1976.1162846