MODEL OF SELECTION OF CHARACTERISTIC PARAMETERS OF A VOICE SIGNAL FOR ANNOUNCER IDENTIFICATION TASKS

Yana Bielozorova

Abstract: *A method for isolating the parameters of the voice signal characteristic of the speaker by selecting the features of the extremums of the surface of the scalogram of the voice signal, built on the basis of wavelet coefficients. The structural model of the speech signal is described, the sequence of mathematical transformations of the information channel of speech communication is considered and generalized, on the basis of which the approach to selection of features of language structures is offered. An algorithm for selecting the characteristics of speech signal structures has been developed. The selection uses a complex Morlet wavelet, which most effectively describes the structure of the speech signal at the first level of decomposition. By synthesizing the speech signal with the specified characteristics, the rational parameters of the wavelet transform settings for the implementation of the method are selected.*

Keywords: *voice signal, Morlet wavelet, special structures, fundamental frequency, speaker identification.*

DOI: https://doi.org/10.54521/ijita29-03-p03

Introduction

The accuracy of speaker identification is based on the correct choice of identification features and the correct presentation of the voice signal. One of the most important characteristics and often determining is the frequency of the fundamental tone. Currently, there is no single approach to determining the fundamental frequency, so effective identification of the announcer is possible only on the basis of a comparison of the frequencies of the fundamental tone, performed on the basis of the same method of determining the frequency of the fundamental tone.

The basis for improving the accuracy of identification of the speaker by the voice signals left by him is a qualitative determination of the frequency of the fundamental tone, which in turn depends on the correct interpretation of the characteristics of the speaker in the voice signal.

Thus, the construction of a model for the selection of characteristic parameters of the voice signal, providing high accuracy in describing the characteristics of the speaker in the voice signal is an important scientific task.

Related works

Nowadays, it is well known that vocalized fragments of a person's speech signal maintain the periodicity of voice oscillations. The speech signal should be considered in the form of pulsating oscillations of the air flow, where the period of repetition of the pulses of the speech signal is called the frequency of the fundamental tone. It has been studied that the fundamental tone determines the structure of the speech signal, and is also the main parameter of the speech signal [Mallat, 1999]. The intonation contour of a person is a trajectory of changing the melody of the fundamental tone frequency.

Prosody of speech formation, one of the components of which is intonation, significantly distinguishes the acoustic signal of speech from written language. Therefore, the fundamental frequency carries a significant amount of information contained in the speech signal of the person. From the point of view of presenting this information, the process of allocating the fundamental frequency is of independent interest [Mallat, 1999, Muzy et al., 1991]. When calculating the fundamental frequency, it is also necessary to take into account both slow changes in the trajectory of the fundamental frequency, and its rapid changes and the very moments of inclusion / deactivation of the speech signal. However, when constructing systems for analysis, synthesis, compression or identification of a speech signal, it should be borne in mind that the fundamental frequency is usually used as one of the main features needed for improved description of the speech signal. When it is necessary to perform calculations of the fundamental frequency from a real speech signal, it is necessary to pay attention to the following points [Muzy et al., 1991]:

1) the ability to work at different noise levels;

2) estimation of the fundamental frequency should be performed with minimal error;

3) the minimization of error in a wide range of changes in the frequency of the fundamental tone, emotional and other changes in the language of the person, leading to variations in the frequency of the fundamental tone;

4) must perform all previous signal transformations;

5) effectively divide the signal into vocalized/unvocalized fragments.

To get the maximum amount of information contained in the fundamental frequency loop, you should also pay attention to the additional point when you need to perform instantaneous calculations of the fundamental frequency value.

Materials and methods

Model of speech signal representation based on wavelet analysis and determination of characteristics of its self-similar structures.

Let us present a speech signal in the form of components with different structure:

$$f(t) = x_1 f_1(t) + x_2 f_2(t) + \cdots + x_s f_s(t)\# \tag{1}$$

In the case of correlation of coefficients $x_1, x_2, \ldots, x_s$ it is difficult to draw a conclusion about the type of approximation functions. The main way to solve this version of the signal is to present it as components

$$\sum_i g_i f_j(t_i) f_k(t_i) = \delta_{jk}\,\# \tag{2}$$

where $g_i = 1/\sigma_i^2$.

Given that the functions f_i in (1) have a different structure, which is subject to changes at random moments of time, the most effective way to describe them is to use approximation methods based on basic decomposition

$$f_i(t) = \sum_n c_{in} \varphi_{in}(t)\,\# \tag{3}$$

where $f_i \in L^2(R), \varphi_{in}$ – basic spatial functions $L^2(R)$.

To create models that adapt to the structure of the signal, it is proposed to use nonlinear approximation schemes. In this case, the approximation f is performed by M vectors depending on the structure of the signal

$$f_M = \sum_{m \in I_M} \langle f, \varphi_m \rangle \varphi_m\,\# \tag{4}$$

where I_M –a set of indices that is determined by the properties of the function f.

The mathematical construction (1) taking into account the introduced properties (1) - (4) in relation to the description of the signal structure is called the structural model of the speech signal.

Given the search for self-similar structures in speech signals, their different shape and length, the most appropriate space for their representation is the space of wavelet bases. Wavelet coefficients $c_{j,n} = \langle f, \psi_{j,n}\rangle, where\{\psi_{j,n}\}_{(j,n)\in Z^2}$, are considered as a result of mapping the function f into a space with a resolution j.

Consider the process of selection of self-similar structures in the speech signal. Naturally, to obtain the characteristics of the signal and the selection of any structures in the signal, wavelet transform algorithms are used, which allow to decompose the signal by the operation of shift and tension of the wavelet ψ. The peculiarity of the wavelet is that its average value is zero, and its integral has the form

$$\text{Wf}(a,b) = \int f(t)\frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)\text{dt}\# \tag{5}$$

And allows us to estimate the behavior of f in the vicinity of point b, which is directly related to *a*. A feature of the wavelet coefficients obtained from (5) is the effective representation of the property of the function f at the proximity of the scale a to zero in the vicinity of point b. We write the Taylor polynomial for some neighborhood v, if the condition of differentiability of the function fm times in the interval $[v-l; v+l]$

$$\rho_v(t) = \sum_{k=0}^{m-1}\frac{f^{(k)}(v)}{k!}\# \tag{6}$$

With this kind of polynomial and evaluation of the quality of its description $\varepsilon_v(t) = f(t) - \rho_v(t)$ the following condition must be fulfilled

$$\forall t \in [v - l; v + l] |\varepsilon_v(t)| \leq \frac{|t - v|^m}{m!} \sup_{u \in [v-l;v+l]} |f^m(u)| \# \tag{7}$$

The estimate of the maximum error $\varepsilon_v(t)$ in the pursuit of t to v is determined by the order of differentiation f in the vicinity of v. To clarify the maximum value of the error, we use Lipschitz smoothness by adding the following indicator [Pavlov A.V. et al., 2007]:

The function f satisfies the Lipschitz condition $\alpha \geq 0$ at point v, if there exists $K > 0$ and the polynomial ρ_v of degree $m = \lfloor \alpha \rfloor$ such that

$$\forall t \in R; |f(t) - \rho_v(t)| \leq K|t - v|^\alpha \# \tag{8}$$

If the function f satisfies (4) for all $v \in [a, b]$ with a constant K, independent of v, then it is assumed that it satisfies the Lipschitz condition α on $[a, b]$. It is known that Lipschitz indicators give the most accurate representation of smoothness, which is used throughout the range.

In [Lardiès J. et al., 2004] it is shown if the wavelet ψ has n zero moments, i.e.

$$\int_{-\infty}^{+\infty} l^k \psi(t) \mathrm{dt} = 0, k = \overline{0; n - 1} \# \tag{9}$$

and n derivatives, then there is $A > 0$ for $f \in L^2(R)$, which satisfies the uniform exponent Lipschitz $\alpha, \alpha \leq n$ on $[a, b]$

$$\forall (s, u) \in R^{+\times[a,b]|\mathrm{Wf}(s,u)| \leq \mathrm{As}^{\alpha+1/2}} \# \tag{10}$$

We can conclude that f will satisfy the Lipschitz condition α on $[a+\varepsilon, b-\varepsilon]$ for any $\varepsilon > 0$, if f is bounded, and wavelet coefficients Wf(s,u) will satisfy (10) for $\alpha < n$.

Thus, condition (10) makes it possible to say that the decrease in the amplitude of the wavelet transform of the speech signal depending on the scale is associated with the uniform and point smoothness of Lipschitz.

Estimating the self-similarity of f at point v on the basis of Lipschitz indicators can be quite difficult, due to the fact that they can change arbitrarily in the vicinity of point v. We use Jaffar's theorem [Xuedong, 2001] to impose the sufficiency condition on the wavelet transform for estimating the Lipschitz smoothness of the function f at the point v.

Let the wavelet ψ have n zero moments and n derivatives. If $f \in L^2(R)$ satisfies the Lipschitz condition $\alpha \leq n$ at point v, then there exists A such that

$$\forall(s,u) \in R^{+\times R|\mathrm{Wf}(s,u)|\leq \mathrm{As}^{\alpha+1/2}\left(1+\left|\frac{u-v}{s}\right|^{\alpha}\right)} \# \tag{11}$$

Conversely, if $\alpha < n$ — not whole, but exists $A, \alpha < \alpha$ such that

$$\forall(s,u) \in R^{+\times R|\mathrm{Wf}(s,u)|\leq \mathrm{As}^{\alpha+1/2}\left(1+\left|\frac{u-v}{s}\right|^{\alpha}\right)} \# \tag{12}$$

then f satisfies the Lipschitz condition α at the point v.

Therefore, when killing the scale s, the calculated amplitudes of the wavelet coefficients based on speech signals have a rapid decrease to zero in areas where the signal is smooth and has no self-similar structures.

Therefore, if |Wf(s,u)| has no local maxima on a small scale, it is assumed that the function f describing the speech signal is locally smooth and the process of isolating self-similar structures of the speech signal function f can be constructed by determining the maximum values of the functions $|\mathrm{Wf}(s,u)|$ on a small scale. It is taken into account that the scale parameters are limited by the parameters of the segmentation of the speech signal and its step.

After the selection of a self-similar structure, the next task is its classification. We use the approach proposed in [Chui, 1992, SounGen]. Expression (6) can be written in a similar form

$$\log_2|\mathrm{Wf}(s,u)| \leq \log_2 A + \left(\alpha + \frac{1}{2}\right)\log_2 s\# \tag{13}$$

Therefore, the smoothness parameters at point v are determined by the slope of the function $\log_2 s$ (and accordingly $\log_2|\mathrm{Wf}(s,u)|$) along the maxima line. The peculiarity of the lines of maxima is its construction on the basis of the points of maxima of the module, which is the curve $s(u)$ in the coordinates (s,u).

The classification of self-similar structures of the speech signal will be performed using (13) as follows.

We introduce the notation $O_v(s,u)$ – as a line of maxima of the wavelet transform of the speech signal converging to the point $u=v$, at $s \to 0$. For each point v we define the slope $\log_2 O_v(s,u)$ as a function $\log_2 s$ for $s \to 0$:

$$\log_2 O_v(s,u) = \log_2 A + \left(\alpha' + \frac{1}{2}\right)\log_2 s\# \tag{14}$$

We will consider that at the point $u=v$ we have a self-similar structure α'.

An effective solution to the problem of classifying the self-similar structure of the speech signal depends on the characteristics of the basic function of the wavelet transform ψ.

For example, if the wavelet basis ψ has n zero moments, then there is a function θ such that

$$\psi = (-1)^n \theta^{(n)} \int_{-\infty}^{+\infty} \theta(t)\mathrm{dt} \neq 0\, \# \tag{15}$$

The wavelet transform is defined in the form

$$\mathrm{Wf}(s,u) = s^n \frac{d^n}{\mathrm{du}^n} \left(f * \overline{\theta}_n\right)(u) \# \tag{16}$$

where $\theta_s(t) = s^{-1/2} \theta\left(\frac{-t}{s}\right)$ [Chui, 1992]. Given this, if the wavelet ψ is represented by only one zero moment, then

$$\psi = -\theta`, \mathrm{Wf}(s,u) = s \frac{d}{\mathrm{du}} \left(f * \overline{\theta}_s\right)(u) \# \tag{17}$$

In (17) the maxima |Wf(s,u)| – represent the smoothed by the function $\overline{\theta}_s$ maxima of the first derivative function of the speech signal f. These multiscale maxima determine the location of the breakpoints and differences of the function of the speech signal f, and, accordingly, allow us to describe the location of self-similar structures. If the wavelet ψ has two zero moments, the representation of the maxima of the modulus of the function of the speech signal f will look like

$$W_2 f(s,u) = s^2 \frac{d^2}{\mathrm{du}^2} \left(f * \overline{\theta}_s\right)(u) \# \tag{18}$$

And they themselves will correspond to the local features of self-similar structures of the speech signal. Thus, it was found that performing a wavelet transform allows to obtain a set of wavelet coefficients that describe the speech signal. The most popular method of presenting the results of wavelet transform is a scalogram (Fig. 1), which allows you to visually represent and assess the location of the surface extremums, built on the basis of wavelet coefficients $W(a,b)$.

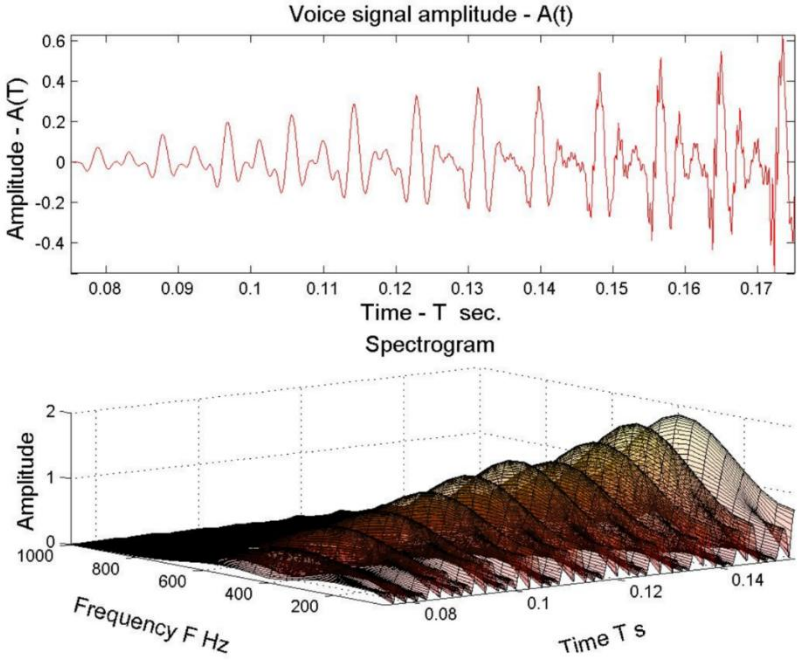

Figure 1: Scalogram of a speech fragment

Of particular interest are the local extrema of the surface coefficients. Theoretically, the analysis of self-similar structures can be performed on the basis of scalogram parameters, but there are a number of statistical functions that allow the evaluation of spectrum characteristics more efficiently.

The general view of such statistical measures of the measure can be represented as

$$M(q,a) = \sum_{l \in L(a)} \left| W\left(a, t_l(a)\right) \right|^q \# \qquad (19)$$

where l – local maximum line, L – a set of lines of maxima of wavelet coefficient modules, $t_l(a)$ – maxima of wavelet coefficients related to lines l scale a.

According to [Turiel and al., 2006], the dependence is executed

$$M(q,a) \sim a^{\tau(q)} \# \qquad (20)$$

where $\tau(q)$ is determined for the value of q by calculating the slope $ln\big(M(q,a)\big)$ from lna, which is called the scaling exponential. Setting the value of q in (9) we obtain the dependence $\tau(q)$. The dependence $\tau(q)$ allows to obtain a multifractal spectrum of the speech signal based on the wavelet transform [Oświęcimka and al., 2020], which allows to describe the main characteristics of self-similar structures. The following dependence is used to obtain the multifractal spectrum

$$\begin{cases} D(h) = \min\limits_{q}[qh - \tau(q)] \\ h = \partial\tau/\partial q \end{cases} \# \qquad (21)$$

The stability of this method of obtaining the characteristics of self-similar structures is to use the frequency-time window, which automatically performs averaging operations, as well as to obtain modules of wavelet coefficients in the calculation. It should be noted that from the point of view of the energy approach to the analysis of self-similar structures, the maxima of the wavelet transform coefficients at different levels of decomposition are the most significant.

The presented approach allows to expand the possibilities of speech signal analysis through the use of fractal and wavelet analysis. Unlike existing methods, this allows the analysis of non-stationary and short-lived signals.

Experiments

Selection of parameters and study of the effectiveness of the method of increasing the informativeness of the fundamental frequency.

As mentioned earlier, we will use the complex Morlet wavelet as a basic wavelet to isolate self-similar structures. An important feature of this wavelet is the ability to determine the instantaneous frequency using the analytical form of the wavelet [SoundGen]. It is believed that the Morlet is wavelet is a wavelet of small oscillations, ie it provides a center frequency of about 1 Hz, has good localization in time and frequency resolution. It is this set of properties that makes the Morlet wavelet one of the best for speech signal analysis. An analytical representation of the Morlet wavelet can be recorded as

$$\psi(t)=\left[exp\left(\frac{-t^2}{2\sigma_t^2}\right)-\sqrt{2}exp-\left(\frac{\omega_c^2\sigma_t^2}{4}+\frac{t^2}{\sigma_t^2}\right)\right]exp(j\omega_c t)\# \tag{22}$$

where $\omega_c=2\pi F_c$, F_c – central wavelet frequency.

$$\sigma_t=\frac{1}{2\pi\sigma_f}\# \tag{23}$$

where σ_t – standard Gaussian deviation ($4\sigma_f$ – width of the wavelet).

The product $\omega_c\sigma_t$ relates the width of the Gaussian wavelet to the frequency of its oscillations. For the Morlet wavelet, this product must take on fairly large values ($\omega_c\sigma_t\geq 5$). The most frequently used range is $5\leq\omega_c\sigma_t\leq 10$, when $0{,}8\leq F_c\leq 1$ [Markel and al., 1982].

Given that $\omega_c\sigma_t$ provides a link between the width of the wavelet and the frequency of its oscillations, to determine the rational parameters of the wavelet, it is necessary to analyze the combination of parameters $(F_c,\omega_c\sigma_t)$, o more accurately identify features and structures speech signal.

It is known [Solovyov and al., 2014] that the activities of phonetics as an analysis of vowel sounds, the frequency range of which is quite wide, but it can be divided into the following subbands depending on the specific sound or set.

In view of this, the total frequency range of the study was divided into subbands of 100-500 Hz, 500-1500 Hz, 1500-4000 Hz, from 4000 Hz. These ranges will be used to fix the scales of the wavelet transform settings.

In order to select the parameters of the wavelet, a speech signal with predefined characteristics was synthesized. The open source software SoundGen [SoundGen] was used for the synthesis. This is a library in the R programming language that allows you to synthesize a language with specified parameters. The estimateVTL function was used to generate the formant frequencies, and the Soundgen function was used to generate the speech signal. A signal with frequencies of 450 Hz, 1400 Hz, 2800 Hz and 4300 Hz was created (Fig. 2).

Results

During the selection of parameters, the central frequency of the wavelet changed in 0.5 Hz increments in the range of 0.8-1 Hz. For each value of the center frequency, the parameter $\omega_c\sigma_t$ changed in steps of 0,5 in the range from 5 to 10. For each combination of parameters, the wavelet transform modulus was calculated, the scales corresponding to the maximum of the coefficient modulus were found, and the frequencies were calculated, deriving the phase wavelet transform coefficients corresponding to these scales. The obtained parameters of the wavelet transform settings are presented in table 1.

The next step of testing the developed methods was to study the effectiveness of determining the frequency of the fundamental tone. Comparative tests to determine the frequency of the fundamental tone on the basis of the proposed method to determine the frequency of the fundamental ion with the Pitch method for pure signal, signal with added noise, signal limited by the telephone channel band.

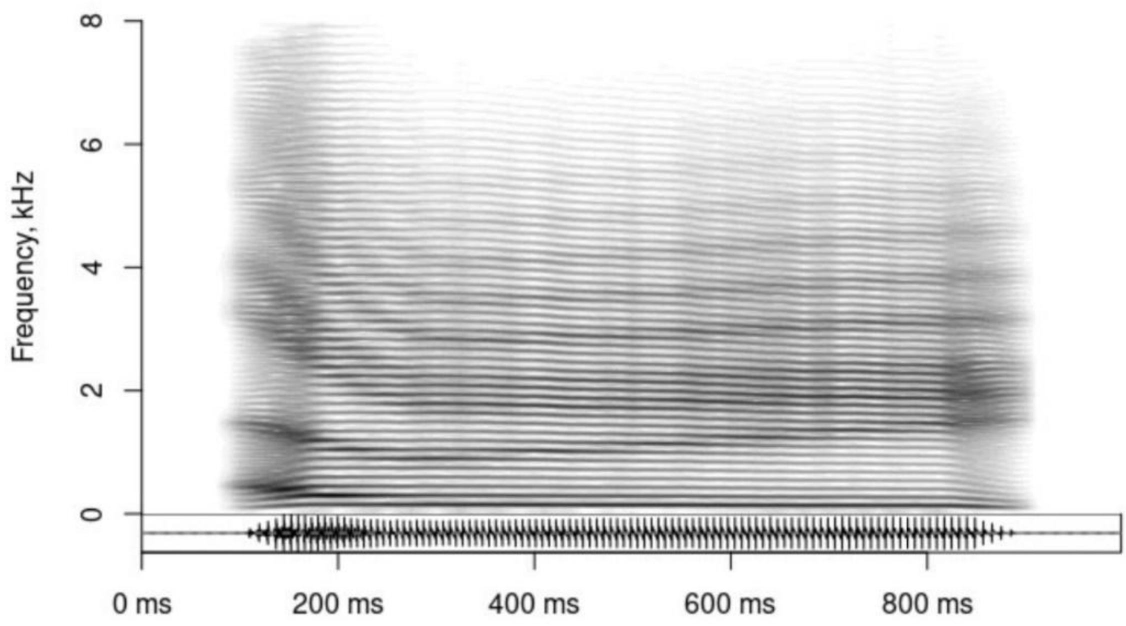

Figure 2: Synthesized speech signal with specified frequency parameters (picture representation is given without changes - as obtained in the software)

Table 1

Configuration parameters $(F_c, \omega_c\sigma_t)$ for selected frequency bands

Frequency range, Гц	F_c	$\omega_c\sigma_t$
100-500	0,8	5
500-1500	1	7
1500-4000	0,9	8
>4000	1	9

Adjustment of method parameters was performed by selection so that the generalized error of allocation of the frequency of the fundamental tone for the test speech material would be minimal. The total test results by estimating the generalized error are presented in Fig. 3, ten men, ten women, for 300 language fragments for each person. The generalized error was calculated from the normalized correlation coefficient with a single delay for calculations by each method, followed by summation.

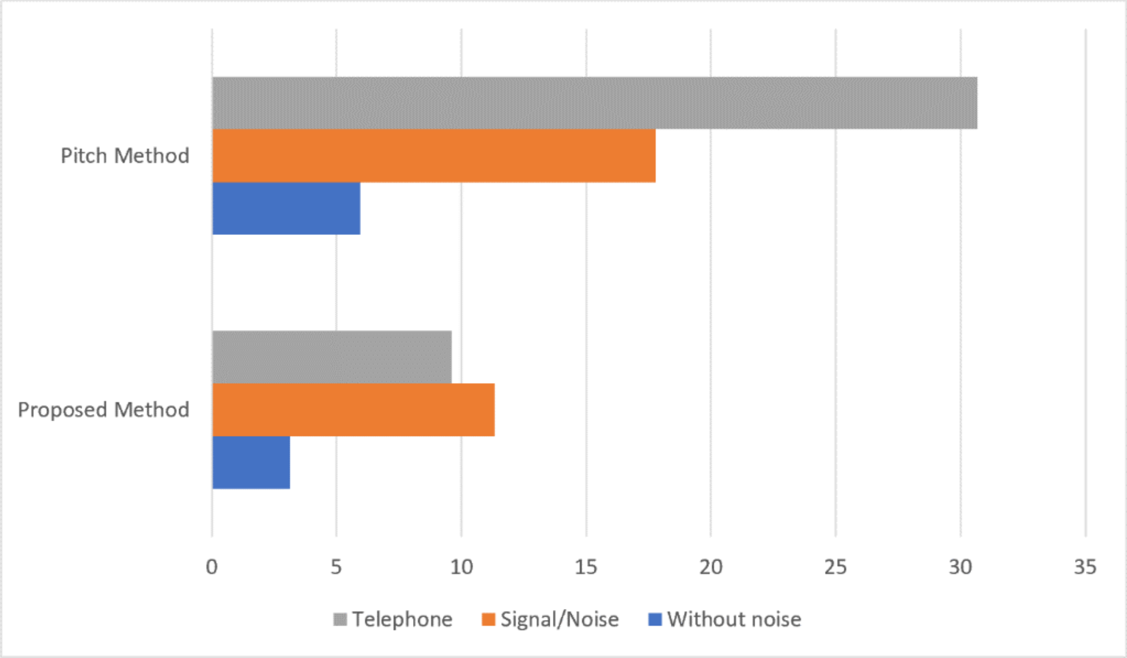

Figure 3: Test results of fundamental tone selection methods (%)

The developed method of determining the fundamental frequency showed the best average results in determining the frequency of the fundamental tone and provided proper tracking of the trajectory of the fundamental frequency throughout the proclamation, even at the ratio Signal/noise = 5 dB. Therefore it is possible to consider that the method is competitive in comparison with other considered methods of allocation of the basic tone which demand manual adjustment of parameters.

Conclusion

An algorithm for selecting special structures in a speech signal has been developed and an informative feature for linguistic identification of a person who, unlike existing ones, uses the values of wavelet transform coefficients of speech signal on segments where extremes of fundamental frequency correlation are observed. who are responsible for the individuality of the speech signal, and achieve high accuracy of identification. An experiment on the selection of rational parameters of the wavelet transform for the implementation of the method is performed, the coefficients of the wavelet transform adjustment for the frequency bands are obtained.

As an assessment of the effectiveness of the study, a comparative experiment was performed to determine the frequency of the fundamental tone based on the proposed approach and the Pitch method. The option to determine the frequency of the fundamental tone, based on the proposed method, reduces the error of determination in all analyzed variants of the voice signal.

Thus, the Model of allocation of characteristic parameters of a voice signal for speaker identification tasks presented in the article solves the problem of increasing the accuracy of speaker identification based on the fundamental frequency by improving the accuracy of determining the structural features of the voice signal.

As a continuation of the study, it is proposed to consider the relationship between the structural elements of the voice signal with its fractal characteristics.

Bibliography

[Mallat, 1999]. Mallat S. A wavelet tour of signal processing, Courant Institute, New York University, 1999, 671 pp.

[Muzy et al., 1991]. Muzy J.F., Bacry E., Arneodo A. Wavelets and multifractal formalism for singular signals: application to turbulence data // Phys. Rev. Lett. 1991. V.67. P.3515−3518.

[Pavlov A.V. et al., 2007]. Pavlov A.V., Anishenko V.S. Multifractal signal analysis based on wavelet transform //Saratov University, 2007. №1. 3-25 pp.

[Lardiès J. et al., 2004]. Lardiès J., Ta M.N., Berthillier M. Modal parameter estimation from output-only data using the wavelet transform, Archive of Applied Mechanics, Vol. 73. 2004. 718-733 pp.

[Xuedong, 2001]. Huang Xuedong. Spoken language processing: a guide to theory, algorithm and system development. – New Jersey: Prentice Hall PTR, 2001. 910 p.

[Chui, 1992]. Chui C. An introduction to wavelets. Academic Press. 1992. p. 278

[SoundGen]. A set of tools with a clear code for voice synthesis, manipulation and analysis. http://cogsci.se/soundgen.html

[Markel and al., 1982]. Markel J.E., Gray A.H. Linear Prediction of Speech. New York, NY: Springer. (1982).

[Turiel and al., 2006]. Turiel, A. & Pérez-Vicente, Conrad & Grazzini, Jacopo. (2006). Numerical methods for the estimation of multifractal singularity spectra on sampled data: A comparative study. Journal of Computational Physics. 216. 362-390. 10.1016/j.jcp.2005.12.004.

[Oświęcimka and al., 2020]. Oświęcimka, P., Drożdż, S., Frasca, M. et al. Wavelet-based discrimination of isolated singularities masquerading as multifractals in detrended fluctuation analyses. NonlinearDyn 100, 1689–1704 (2020). https://doi.org/10.1007/s11071-020-05581-y

[Solovyov and al., 2014] Solovyov Victor, Byelozorova Yana. Multifractal approach in pattern recognition of an announcer’s voice / Victor Solovyov, Yana Byelozorova // TEKA. Commission of motorization and energetics in agricultire. – 2014. – Vol. 14, № 2.– pp. 164 – 170.

Authors' Information

Yana Bielozorova *– PhD, Associate Professor of Software Engineering Department, National Aviation University, Kyiv, Ukraine.*

E-mail: *bryukhanova.ya@gmail.com*

Major Fields of Scientific Research*: Speech Recognition Models, Wavelet analysis, Software Architecture*