Short Phrase Speaker Identification in Noisy Environment

Petr V. Malinin, Viktor V. Polyakov
Altai State University, Barnaul, Russia

Abstract – **The article deals with the problem of speaker identification, which uses short phrases in noisy environment. The speaker identification in harsh conditions (short phrases, noisy environment) is an actual task, since it is in harsh conditions that the possibilities of the methods used in identification are manifested. Short phrases are characterized by a minimum content of informative elements (phonemes). Usually the duration of short phrases varies hundreds of milliseconds up to several seconds, which leads to certain difficulties in identifying a person by voice. As a method of extracting voice attributes, it is proposed to use the Morlet wavelet transforms. The speakers identification results obtained on basis of the k-nearest neighbors method are presented. Identification results were obtained on the basis of voice records from the database "Acoustic speech signals for the identity system by voice data." Noisy conditions were formed by additive overlap of different noise levels on voice records. The most important types of noise in real conditions were used white noise, speech-like noise and street noise. The approach proposed is recommended to be used in identification systems with access control of information security.**

Index Terms- **speaker identification, noisy environment, short phrase, wavelet, k-nearest neighbors.**

I. INTRODUCTION

The successful application of biometric technologies in various areas shows prospects for their further development. Biometric technologies hold a firm place in information security area as parts of access control systems. Voice identification is a part of access control techniques in biometric systems. Sometimes, voice identification and authentication are the only available techniques (for example, radio communication, telephone and mobile phone connections). Voice identification has its obvious advantages and shortcomings that need to be eliminated. However, many shortcomings continue to be relevant, such as high level of errors when using short phrases and a negative impact of external noise.

II. PROBLEM DEFINITION

A high level of errors in speaker identification is due to a number of factors. In addition to the dispersion of the parameters of voice characteristics, the environment and the shortage of informative parameters of the voice characteristics exert a significant influence. It is required to involve more advanced methods of data processing and analysis. They will allow to normalize the dispersion, to increase the informative parameters of voice characteristics and to reduce the negative impact of environment. The time-frequency wavelet transform has proved to be suited in solving mentioned above problems. It is allow to take into account additional informative features and to reduce the influence of the noise signal components.

III. FEATURES EXTRACTION

Phrases with minimum informative elements and time duration from hundreds of milliseconds up to several seconds are considered to be short phrases. It is quite difficult to determine whether a sample of a voice recording belongs to a particular person due to a limited time duration of a short phrase. Typically, time-frequency transforms are used for extracting voice specific features. The most popular ones are Fourier and wavelet window transforms that allow the extraction of the maximum number of informative features for further voice identification.

Wavelet transforms provide the optimal resolution in the time-frequency domain, for example [1]:

$$\Psi_x^{\psi}(\tau,s)=\frac{1}{\sqrt{|s|}}\int x(t)\psi^{*}\left(\frac{t-\tau}{s}\right)dt \quad (1)$$

Additional characteristic features that are not available with the Fourier window transform can be obtained from wavelet transforms. A basis wavelet function and decomposition level are required for the proper use of wavelet transforms. The Morlet wavelet functions are used as the basis functions due to their close relation to human auditory perception [2]. Optimal decomposition levels are estimated by calculation of Shannon entropy levels [3].

Shannon entropy calculation results are shown in Fig. 1. According to Fig.1, the optimal decomposition level should be set to 13.

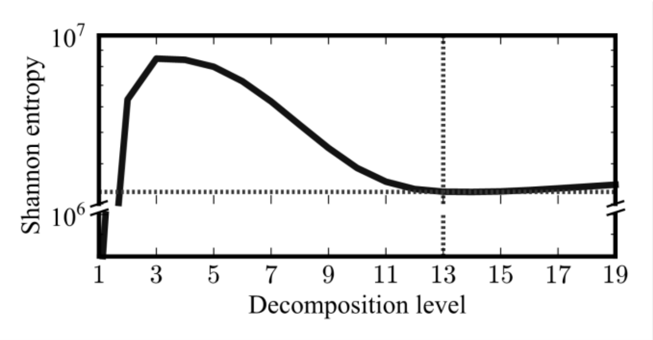

Fig. 1. Estimation of the decomposition level.

The results (coefficients) of the Morlet wavelet transform are used as characteristic features for further analysis.

External acoustic noise superimposed with an identifiable voice signal in real life is the main source of voice identification errors [4,5].

Voice identification is impossible if the noise level is above a certain threshold. Noise reduction techniques are commonly used in voice identification systems on different steps of voice signal processing. Typically, noise reduction is applied on a hardware level and a preprocessing step.

An additive noise overlay is performed to estimate the influence of different types of noises on the voice record used for identification. Several types of noises are overlaid – the white noise, speech-like noise, and street noise. The street noise here is the street traffic noise. The speech-like noise is composed of voice records of several speakers (10 different speakers overlaid). The white noise is generated by the audio editor software (Audacity).

Therefore, short phrases with overlaid noises are used for voice identification approach using 13-level continuous Morlet wavelet transform.

IV. PROPOSED APPROACH

The simplest method of classification is the method of k-nearest neighbors. This method provides a quite simple quality estimation of classification of compared data sets (coefficients of the wavelet transform).

Below is a mathematical description of the nearest-neighbor method [6]. Let $X \in R^n$ – the set of objects (voice signals), n – the number of variables, which are the coefficients of transform; Y – the set of permissible responses, i.e. numbers of speakers classes. Let the calibration sample $\{(x_i, y_i)\}_{i=1}^{l}$ that represented by the calibration database of voice data (l – the number of records). A set of objects $X^m = \{x_i\}_{i=1}^m$ (m – the number of test sample records) consisting of the voice signals of unknown speakers (one or several). It is required to find the set of responses $\{y_i\}_{i=1}^m$, that show relations of unknown signals with known speakers for objects $\{x_i\}_{i=1}^m$.

Objects are compared using multidimensional Euclidean distance between objects, which is expressed by the coordinates x and x ':

$$\rho(x,x') = \sqrt{\sum_{i=1}^{n}(x_i - x'_i)^2} \, . \qquad (2)$$

For an arbitrary object $x \in \{x_i\}_{i=1}^m$ it is necessary to place objects x_i in the training sample in the order of increasing of their distances to the object with the coordinate x:

$$\rho(x, x_{1;x}) \leq \ldots \leq \rho(x, x_{m;x}) \qquad (3)$$

where $x_{i;x}$ – the training sample object that is the i-th neighbor of the object x, $y_{i;x}$ – the response to the i-th neighbor. Thus, an arbitrary object generates its renumbering of the sample. In the most general form, the affiliation of the nearest neighbor to the corresponding response (speaker number) can be written in the following form:

$$a(x') = \arg\max_{y \in Y} \sum_{i=1}^{m} [x_{i;x} = y] w(i,x). \qquad (4)$$

where $w(i, x)$ – a given weight function that estimates the degree of importance of the i-th neighbor for the classification object.

Training and testing samples constructed from combinations of voice records (50 speakers) from the speech database [7] are used to train and evaluate the model. Short phrases are composed of combinations of pronounced numbers from 0 to 9. The average duration of a single voice record for one speaker is about 3 s. There are unique sets of training and testing samples used for each studied condition. 80 % of voice records from the initial database are used for the construction of training samples, while 20% of voice records are used for testing samples. The number of calculations for each experiment is determined by the number of combinations randomly selected from the database according to the quantity of the initial data.

V. IDENTIFICATION RESULTS

Reliability of the proposed approach, as well as its applicability limitations and the most significant features important for identification are estimated with quantitative calculations of accuracy:

$$Ac = \frac{\text{Number of correct predictions}}{\text{Number of correct predictions}} 100\% \, .$$

Comparison of accuracy for voice signals of different durations (less than 3 s and more than 10 s) overlaid by different types of noises (white noise, speech like noise, street noise) are shown in Fig. 2. The white noise causes a significant influence on identification of voice signals of different durations (Fig. 2; LW, SW – for long and short phrases respectively). There are two families of essentially similar curves that can be clearly distinguished (Fig. 2; LW, LSp, LSt – for long phrases, SW, SSp, SSt – for short phrases). They demonstrate that the percentage of correct identification is greater for longer phrases than for shorter phrases.

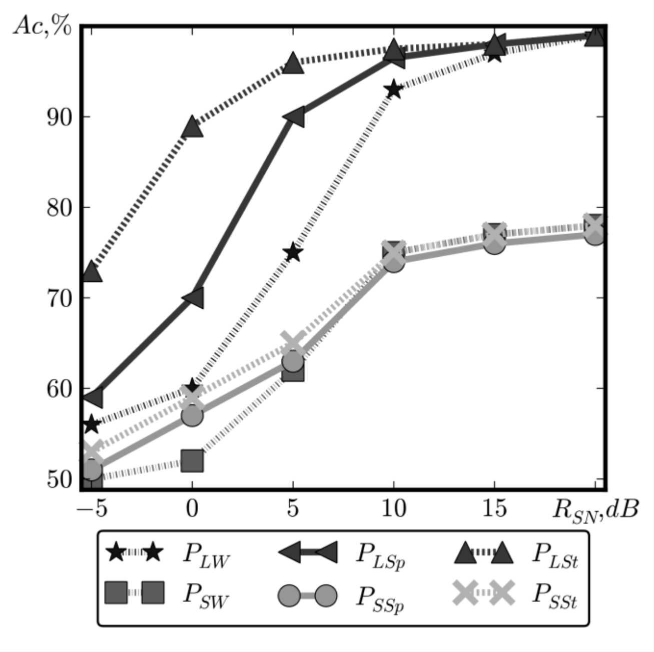

Fig. 2. Identification results.
L - long records(duration more than 10 s);
S - short records (duration less than 3 s);
W - white noise; Sp - speech-like noise; St - street noise.

VI. SUMMARY AND CONCLUSIONS

The proposed approach provides a sufficiently reliable identification of speakers using a short phrase. This approach can be recommended for further implementation in access control identification systems and information security systems.

The work was prepared within the framework of the scientific project 16-33-01160-OGN, supported by the Russian Humanitarian Scientific Foundation.

REFERENCES

[1] Wavelets and Signal Processing: An Application-Based Introduction, Stark H.G. - Berlin: Springer, 2005, p. 150.

[2] Daugman J.G. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters // Journal of the Optical Society of America A. - 1985. 2(7). pp. 1160-1169.

[3] Sang Y, Wang D., Wu J. Entropy-Based Method of Choosing the Decomposition Level in Wavelet Threshold De-noising // Entorpy. - 2010. Vol. 12, Issue 6. pp. 1499-1513.

[4] Malinin P.V., Polyakov V.V. Influence of the type of acoustic noise on voice identification of a person // Izvestiya of Altai State University. - 2013. - № 1/2. - pp. 168-169. (in Russian)

[5] Graciarena M., Kajarekar S., Stolcke A., Shriberg E. Noise robust speaker identification for spontaneous arabic speech // ICASSP 2007. IEEE International conference. - 2007. - pp. 245-248.

[6] Ryazanov V.V., Senko O.V., Zhuravlev Yu.I. Recognition. Mathematical methods. Software system. Practical applications. - Moscow: PHASIS, 2006. - p. 176. (in Russian)

[7] Malinin P.V. Acoustic speech signals for the system of identification by voice. Certificate of state registration of the database No. 2013620132, January 9, 2013. (in Russian)

Brief professional biography of the authors

	Malinin Petr Vladimirovich, PhD, Associate Professor. In 2015, at the Dissertational Council at the Tomsk State University of Control Systems and Radioelectronics, he defended his thesis on the topic: "Technology of voice identification of the person on the basis of projection methods of analysis of multidimensional data". . In 2016 he became the executor of the grant of the Russian Humanitarian Scientific Foundation, on the topic of: "Criminalistic features of crimes in the sphere of computer information" (2015-2018). At present time he conducts classes in the academic disciplines: "Technical Security of Information", author's special courses for undergraduates: "Technologies of information security", "Expert systems and audit of information security".

	Viktor Vladimirovich Polyakov, Doctor of Physical and Mathematical Sciences, Professor, Head of the Department of Applied Physics, Electronics and Information Security; Dean of the Faculty of Physics and Technology, Honorary Worker of Higher Professional Education of Russian Federation. 4 Science and Research Conferences were prepared under his supervision. More than 500 publications, including 6 monographs and 15 textbooks and study guides.