Sub-band identification of speech signals word fragments according to given sample

Sub-band identification of speech signals word fragments according to given sample

E G Zhilyakov[1], S P Belov[2], A S Belov[2] and A A Medvedeva[1]

[1] Belgorod State National Research University, 85, Pobedy, st., Belgorod, 308015, Russia
[2] Belgorod University of Cooperation, Economics and Law, 116a, Sadovaya st., Belgorod, 308023, Russia

E-mail: belovssergei@gmail.com

Abstract. The article considers the problem of the speech signals fragments, generated when pronouncing a certain word form, emphasis in the recordings, which is of the interest from the position of the speech information exchange analysis solvable applied problem. In this case, it is assumed that initially the class of the desired word form is given by the fragment available in the real speech signal recording. Therefore, it is natural to term the problem under consideration as precedent identification. The relevance of the methods and algorithms development for automatic precedent identification of voice recordings fragments in such situation is determined by the breadth of their possible applications, for example, in information and analytical security systems. The variability of the speech signals fragments properties, even generated when pronouncing the same word form by the same person, and the need to learn from one precedent in determining the critical areas of decisive functions are the main factors that determine the complexity of solving this problem. It is shown in this paper that sub-band analysis is an adequate basis for solving the problem under consideration, and an original mathematical apparatus was developed for its implementation. On the basis of original sub-band representations, the decision procedures for identification of speech signals recordings fragments, including the selection of pauses between word fragments, were developed. In particular, training procedures were proposed for one precedent while maintaining its original sub-band properties.

1. Introduction

Oral speech for a person is one of the most natural forms of information exchange. Therefore, speech computer technologies are developing quite intensively [1]. Among them, an important place is taken up by the technologies of automatic speech recognition [2-6]. In this case, the analysis of speech signals samples is carried out, which are digitally recorded at the outputs of the microphones. In this article, we consider the direction of detection in the recordings of speech signals fragments samples generated by pronouncing certain word forms that are of interest from the standpoint of applications. In particular, such situations arise in information and analytical security systems. In this case, it is assumed that the speech signals records are checked for the key words presence, which are carried out for some rather significant time. Sections of speech signals samples records that are carried out when pronouncing individual word forms within the framework of the work are called word fragments.

Let's clarify the problem statement. Let it be known that the vector of speech signal samples:

$$\vec{x} = (x_1,...,x_N)', \quad (1)$$

where the prime symbol means transposition:

$$x_k = x((M+k)\Delta t), k = 1,..., N, \quad (2)$$

is registered when pronouncing a given word form.

Here, the symbol *M* denotes some beginning of the word fragment recording, the dimension of the vector is determined by its duration, and Δt is the equidistant sampling with a step:

$$\Delta t = 1/v_d, \quad (3)$$

where v_d is the sampling rate.

It is necessary to determine other word fragments in the rest of the speech dialogue recording, which are registered when pronouncing similar word forms.

Bearing in mind that the speech signal word fragments formed at other times are not a clear copy of the original vector (1), therefore, the described procedure, naturally, is termed identification. It should be noted that in a certain sense this term is the synonymous with the concept of recognition.

In accordance with the methodologies, we are talking about testing the validity of the following initial (null) hypothesis:

H_0 is the analyzed word fragment identical to the original one (registered under the influence of the given word form).

Obviously, the alternative hypothesis should have the form.

H_1 - the compared speech signals word fragments are not identical.

At the basis of automatic identification, a certain decision function is used, it determines the computational procedure for processing empirical data. In this case, it is assumed that there is a so-called critical range of decision function values, which are taken as evidence of the initial hypothesis with empirical data inconsistency. Thus, it is necessary to develop a measure of the compared word fragments proximity and to conduct training to determine the critical area.

It is important to note that within the framework of the formulated task for training, it is possible to use only one precedent in the form of a highlighted word fragment. Therefore, it is necessary to resort to modeling the training sample. The adequacy of such modeling is determined, among other things, by the properties of the proximity measure. Within the framework of this work, it is proposed to apply proximity measures that are little responsive (invariant) to the characteristics of voices pronouncing the word form - precedent.

Another significant point lies in the differences in the speech signals fragments generated by the same word forms, among which, first of all, it should be noted the differences in the lengths of the time intervals of their sounding and changes in the characteristics of the voice during the dialogue, for example, in the energy characteristics. Thus, it is necessary to ensure the invariance of the proximity measure characteristics to such word fragments nonstationarities.

2. Materials and methods

2.1 Basics of sub-band analysis

The effectiveness of the decision procedures is mainly determined by the degree of the analyzed fragments properties reflection adequacy from the standpoint of their proximity measure reaction to differences in characteristics. In the case of speech signals, it is proposed to use the differences in the energy distribution in the frequency domain. The speech sounds energies concentration that generate speech signals in small fractions of Fourier transforms definitions domains is a motive for using sub-band analysis based on the energy part concept falling into a given sub-band. Meaning the partitioning of the domain:

$$z \in [-\pi, \pi) \quad (4)$$

Fourier transforms (spectrum) of the fragment (1):

$$X(z) = \sum_{k=1}^{N} x_k \exp(-jz(k-1)) \tag{5}$$

to adjoining sub-bands:

$$V_r = [-V_{2r}, -V_{1r}) \cup [V_{1r}, V_{2r}), r = 1,..., R, \tag{6}$$

the specified parts of the energy are determined by using the equation:

$$P_r(\vec{x}) = \int_{z \in V_r} |X(z)|^2 \, dz / 2\pi. \tag{7}$$

In this case, it is assumed that the conditions for covering the Fourier transforms domain definition by sub-bands of the same width are fulfilled:

$$\Delta z = V_{2r} - V_{1r} = \pi / R, \tag{8}$$

$$V_{11} = 0, \; V_{2R} = \pi \tag{9}$$

Substitution of definition (5) into (7) gives the representation:

$$P_r(\vec{x}) = \vec{x}' A_r \vec{x}, \tag{10}$$

so that sub-band matrices are the mathematical basis for further constructions:

$$A_r = \{a_{ik}^r\}, i, k = 1,..., N \tag{11}$$

with elements:

$$a_{ik}^r = (\sin(V_{2r}(i-k)) - \sin(V_{1r}(i-k)))/\pi(i-k), a_{ii}^r = \Delta z / \pi. \tag{12}$$

When comparing fragments, it is proposed to take into account only sub-bands, which in the case of:

$$P_r(\vec{x}) \geq \| \vec{x} \|^2 \, \Delta z / \pi, \tag{13}$$

where $\| \vec{x} \|^2 = \sum_{k=1}^{N} x_k^2$.

They can be termed informational ones.

2.2 Sub-band decision function

It is proposed to use a quadratic form as the sub-band proximity measure (it is assumed that the vectors have the same dimension):

$$F_G(\vec{x}, \vec{y}) = u' A_G \vec{u}, \tag{14}$$

where

$$\vec{u} = (P_G(\vec{y}))^{1/2} \vec{x} - (P_G(\vec{x}))^{1/2} \vec{y}; \tag{15}$$

$$A_G = \sum_{r \in G(\vec{x})} A_r; \tag{16}$$

$$P_G(\vec{x}) = \vec{x}' A_G \vec{x}; \tag{17}$$

$G(\vec{x})$ is the set of sub-band indices satisfying inequality (13).

Note that form (14) is invariant to changes in amplification or attenuation of voice intensities. It can be termed a sub-band similarity measure of compared fragments. It is clear that the fragments must be aligned according to the number of samples, for example, by adding zero values.

Obviously, the values of measure (14) satisfy the inequality:

$$0 \leq F(\vec{x}, \vec{y}) \leq 2 \tag{18}$$

where the left boundary is reached when the segments of the Fourier transforms are within the set of information sub-bands:

$$z \in B(\vec{x}), B(\vec{x}) = \cup V_r, r \in G(\vec{x}) \tag{19}$$

agree within a positive multiplier:

$$Y(z) = b \cdot X(z), b > 0 \cdot, \tag{20}$$

and the right boundary is reached when this multiplier is negative:

$$b < 0. \tag{21}$$

It is this property that makes it possible to construct the critical area of limited size of the view:

$$D_F = \{0 \le F_G > h_\alpha\}, \tag{22}$$

which upper boundary corresponds to the prior probability of the first type errors:

$$Ver(F_G > h_\alpha) \le \alpha. \tag{23}$$

A natural method for determining this boundary is training on the sample of fragments generated by the same word form. In this case, there is only one precedent fragment, which leads to the need to use artificial propagation of identical objects (augmentation).

2.3 Sub-band augmentation of a precedent fragment

It follows directly from the equation (7) and relation (12) that, in terms of symmetry and positive definiteness, the matrix (16) can be represented in the form:

$$A_G = Q_G L_G Q_G^{'}, \tag{24}$$

where $Q_G = (\vec{q}_1^G ... \vec{q}_N^G)$ is the orthogonal matrix of eigenvectors:

$$A_G Q_G = Q_G L_G, Q_G Q_G^{'} = Q_G^{'} Q_G = diag(1,..,1); \tag{25}$$

L_G is the diagonal matrix of positive eigenvalues, arranged by descending order:

$$L_G = diag(\lambda_1^G,..., \lambda_N^G), \tag{25}$$

$$\lambda_1^G \ge \lambda_2^G \ge ... \ge \lambda_N^G \ge 0. \tag{26}$$

In the paper [8], the validity of the equation:

$$1 \ge \lambda_k^G = \int\limits_{z \in B(\vec{x})} |W_k^G(z)|^2 \, dz / 2\pi, \tag{27}$$

where W_k^G is the Fourier transform of the corresponding eigenvector.

In this case, with the sufficient degree of accuracy, the equalities are:

$$\lambda_k^G = 0, \quad J_G < k \le N, \tag{28}$$

when (taking into account (8)):

$$J_G = 2[MN / 2R] + 4, \tag{29}$$

the square bracket means the integer part of the number, and M is the number of terms in the sum (16).

Therefore, by building:

$$Q1_G = (\vec{q}_1^G ... \vec{q}_{J_G}^G), \tag{30}$$

it is easy to show that the vector:

$$\vec{s}_G = Q1_G Q1_G^{'} \vec{x} \tag{31}$$

satisfies the equality:

$$F_G(\vec{x}, \vec{s}_G) = 0, \tag{32}$$

that is, in the indicated sense, it is identical to the original one.

Therefore, in the process of augmentation, it is proposed to form the following training sequence of vectors:

$$\vec{x}_k = \vec{s}_G + d_k \vec{v}_k, \; k = 1,...., K, \tag{33}$$

where $\vec{v}_k = (v_{1k},...,v_{Nk})^{'}$ is the normalized vector,

$$\| \vec{v}_k \|^2 = 1, \tag{34}$$

consisting of Gaussian pseudo-random numbers with zero mean; the multiplier d_k ensures the equality of the Euclidean norms with the original vector:

$$\| \vec{x}_k \|^2 = \| \vec{x} \|^2 . \quad (35)$$

Its value is determined by the corresponding quadratic equation, any of the solutions of which can be used, for example:

$$d_k = ((\vec{s}_G, \vec{v}_k)^2 + \| \vec{x} \|^2 - \| \vec{s}_G \|^2)^{1/2} - (\vec{s}_G, \vec{v}_k), \quad (36)$$

where the symbol $(,)$ stands for the scalar product of vectors in Euclidean space.

The number of vectors generated is determined by the desired error probability of the first one:

$$K = 1/[\alpha] + 1. \quad (37)$$

Then the boundary of the critical area satisfies the equation:

$$h_\alpha = \max F_G(\vec{x}, \vec{v}_k), 1 \le k \le [1/\alpha] + 1. \quad (38)$$

3. Conclusion

The article formulates the urgent problem of speech signals word fragments detecting in the recordings due to the pronunciation of the same word form (precedent), which is determined in advance. The analysis of the features of its solution is carried out and the approach to the solution based on the division of the Fourier transforms definition domains into sub-bands is justified. Sub-band decision functions were constructed and a training procedure was developed by using sub-band augmentation of the original fragment - the speech signal precedent.

4. Acknowledgments

The research was carried out with the support of the RFBR grant No. 20-07-00215a

References

[1] Shelukhin O I and Lukyantsev N F 2000 *Digital processing and transmission of speech* (Moscow: Radio and communication) 456
[2] Nitsenko A V and Shelepov V Yu 2004 Algorithms for phonemic recognition of words in a predetermined dictionary *Artificial intelligence* **4** 633-639
[3] Kipyatkova I S, Ronzhin A L and Karpov A A 2013 *Automatic processing of colloquial Russian speech*: monograph (St. Petersburg: GUAP) 314
[4] Gerasimov A V, Morozov V R and Fidelman V R 2005 Application of the modified linear prediction method to the problems of identifying acoustic features of speech signals Radio *Engineering and Electronics* **50(10)** 1287-1292
[5] Agranovsky A V and Lednov D A 2004 *Theoretical aspects of algorithms for processing and classification of speech signals* (Moscow: Radio and communication) 164
[6] Rabiner L R and Shafer R F 1981 *Digital processing of speech signals* (Moscow: Radio and communication) 496
[7] Gantmakher F R 1967 *Matrix Theory* (Moscow: Nauka) 575
[8] Zhilyakov E G 2015 Optimal sub-band methods for analysis and synthesis of finite-duration signals *Autom. Remote Control* **76(4)** 589-602