UDC 004.522(045)

K. I. Prokopenko

ADAPTIVE ALGORITHM OF SPEECH SIGNALS SEGMENTATION

Institute of Air Navigation, National Aviation University, Kyiv, Ukraine
E-mail: kprok78@gmail.com

Abstract— *New adaptive algorithm of speech signals segmentation on allophones is proposed. Algorithm based on non-stationary random process disorder detection.*

Index Terms—Speech signal, segmentation, phonemes, non-stationary random process disorder.

I. INTRODUCTION

Speech signal is a non-stationary random process, which statistical characteristics are changing at such key points as starting speech from pause, sound stressing or transition between syllables or phonemes. The problem of segmentation consists in splitting of speech signal into phonemes in order to further recognition of these phonemes.

II. PROBLEM FORMULATION

The main point of speech segmentation algorithms is the fact that phoneme is a segment of stationary random process, which does not changing its statistical properties, including spectral characteristics. Thus, analyzing of spectral characteristics changes can provide detection of phonemes transition time moments.

Today the task of speech signal disorder points detection is mostly solving in case of small number of its states [1], [2]. In Fig. 1 the classification of the known methods of segmentation is shown. They were divided into two groups – amplitude and spectral. On the first stage of analysis, speech flow is dividing into uniform frames (windows) of approximately 50 ms [3], [4]. On the second stage, using frames amplitude algorithms frames classified into two main classes – class "Speech" and class "Pause". On the third stage "language"-frames are classifying into classes "Voice", "Hissing", "Transition". In this paper, the problem of segmenting of speech signal belongs to a class of problems of non-stationary process disorder determining and should be formulated and solved in terms of statistical analysis [5].

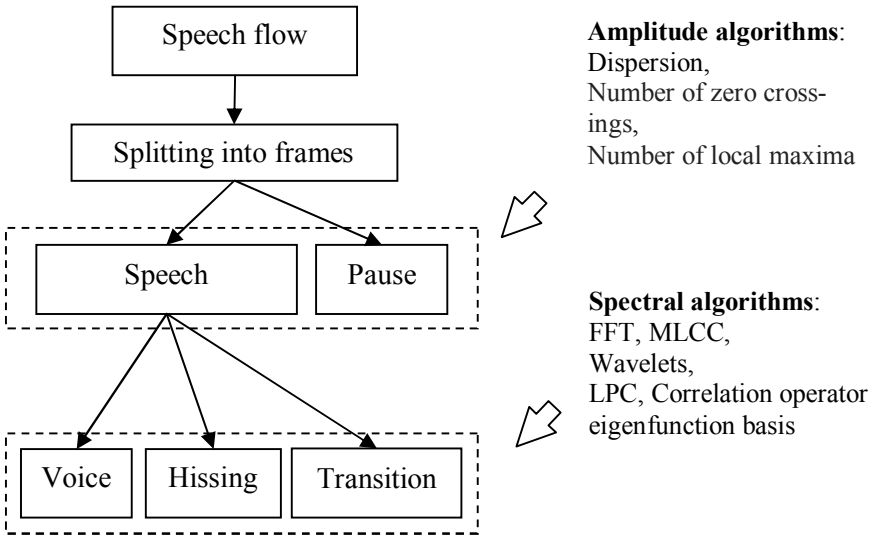

Fig. 1. Classification of segmentation algorithms

The complexity of phonemes change-detecting problem is a priori uncertainty of speech signal statistical characteristics, as a native speaker can be anyone who has its own timbre and spectral characteristics of the voice [1].

This paper proposes a new adaptive algorithm for transition points detecting, based on an evaluation the statistical characteristics of the spectrum on the uniform partition of the speech signal at time intervals.

III. DESCRIPTION OF SEGMENTATION ALGORITHM

The task of speech signal segmentation is to evaluate the time moment of phonemes change, as moment of transition of non-stationary process $l(t)$ (which represents a speech flow) from one state to

another. This point corresponds to a change of local signal characteristics, namely, modified covariance matrix n-dimensional density distribution [5] – [7].

To construct an algorithm, which detects the time moment of the phoneme change, two consecutive sliding windows with length $n\Delta t$ are organizing. Those windows are represent two samples – $x = x(t_1),...,x_n = x(t_n)$ and $y = x(t_1 + n\Delta t),..., y_n = x(t_n + n\Delta t)$, where Δt is the discretization interval. Let $\mathbf{R_x}$ and $\mathbf{R_y}$ are covariance matrices of n-dimensional distribution of x and y samples. For these samples statistical hypothesis H_0 of equality of two covariance matrices $\mathbf{R_x}$ and $\mathbf{R_y}$ is verified:

$$H_0 : \mathbf{R_x} = \mathbf{R_y}.$$

Invariant algorithm of checking of hypothesis H_0 is calculating the ratio of geometric- and arithmetic-means of the determinants of sample covariance matrices $\mathbf{R}_\mathbf{x}^*$ and $\mathbf{R}_\mathbf{y}^*$, constructed by samplings $x_1,...,x_n$ and $y_1,...,y_n$:

$$\lambda(x_1,...,x_n, y_1,..., y_n) = 2\frac{\sqrt{\left|\mathbf{R}_\mathbf{x}^*\right|\left|\mathbf{R}_\mathbf{y}^*\right|}}{\left|\mathbf{R}_\mathbf{x}^* + \mathbf{R}_\mathbf{y}^*\right|}. \quad (1)$$

Well-known fact is that sample estimations of covariance function are sufficient statistics [6], [7]. Values of covariance function are one to one connected with energy spectrum of signal by direct and reverse Fourier transformation. Thus, estimations of power spectrum, obtained from samples $x_1,...,x_n$ and $y_1,...,y_n$, are sufficient statistics for hypothesis H_0 verifying task either.

Power spectrum estimations are obtaining by equations:

$$S^2_{x,i} = \left| \frac{1}{n}\sum_{k=0}^{n-1} x_{k+1} e^{-j\omega\frac{2\pi}{n}ik} \right|^2, i = 0,...,n-1;$$
$$S^2_{y,i} = \left| \frac{1}{n}\sum_{k=0}^{n-1} y_{k+1} e^{-j\omega\frac{2\pi}{n}ik} \right|^2, i = 0,...,n-1. \quad (2)$$

Because for power spectrum estimations condition $S^2_{x,i} = S^2_{x,n-i}$; $S^2_{y,i} = S^2_{y,n-i}, i = 1,..n/2$ is true, so far it is enough to use only $n/2+1$ of spectral points, obtained from samples $x_1,...,x_n$ and $y_1,...,y_n$.

In order to synthesize the algorithm of speech signal segmentation, the Gaussian model of distribution of power spectrum estimations logarithms is used. Thus, two competitive statistic hypotheses about spectral distribution coefficients are must be verified:

$$H_0 : f_Z(Z_0,..,Z_{n/2}) = f_Q(Q_0..,Q_{n/2});$$
$$H_1 : f_Z(Z_0,..,Z_{n/2}) = f_Q(Q_0 + \Delta,..,Q_{n/2} + \Delta), \Delta \neq 0, \quad (3)$$

where Δ is the difference of shifting parameters of $f_Z(*)$ and $f_Q(*)$, $Z_i = \ln(S^2_{x,i})$, $Q_i = \ln(S^2_{y,i})$, $i = 0,..,n/2$ – logarithms of sample's $x_1,...,x_n$ and $y_1,..., y_n$ power estimations, $f_Z(Z_0,..,Z_{n/2})$ and $f_Q(Q_0..,Q_{n/2})$ – multi-dimensional distribution density of power spectrum coefficient's logarithms of $S^2_{x,i}$ and $S^2_{y,i}$ (2). Hypothesis H_0 means that shifting parameters of distributions $f_Z(*)$ and $f_Q(*)$ are the same. Otherwise, hypothesis H_1 means that difference between shifting parameters of distributions $f_Z(*)$ and $f_Q(*)$ is not equal to zero: $\Delta \neq 0$ (3).

According to generalized Bayes rule, the decision rule's structure is evaluated, as likelihood functions' ratio of joined statistic $Z_0,...,Z_{n/2},Q_0,...,Q_{n/2}$ for competitive hypotheses, which were integrated on signal's and noise's domains of a priori unknown parameters' values.

The decision rule is representing by the following expression:

$$\ln\lambda(Z_0,..,Z_{n/2},Q_0,..,Q_{n/2}) = \frac{\Delta\sum_{i=0}^{n/2}(Z_i - Q_i)}{2\sigma^2} + \frac{n\Delta^2}{8\sigma^2}. \quad (4)$$

Since Δ can take both positive and negative values, double-sided decision rule is formulated as a comparison of statistic's (4) module with decision threshold value V_p:

$$\gamma(Z_0,..,Z_{n/2},Q_0,..,Q_{n/2}) = \begin{cases} 1, \sum_{i=0}^{n/2}\left|(Z_i - Q_i)\right| \geq V_P(n,\sigma,\Delta); \\ 0, \sum_{i=0}^{n/2}\left|(Z_i - Q_i)\right| < V_P(n,\sigma,\Delta). \end{cases} \quad (5)$$

IV. Evaluating of decision threshold value

Algorithm (5) for disorder determining is to compare statistic

$$L = \sum_{i=0}^{n/2}\left|(Z_i - Q_i)\right| = \sum_{i=0}^{n/2}\left|G_i\right|. \quad (6)$$

with threshold value V_p where $Z_i = \ln S_{x,i}$, $Q_i = \ln S_{y,i}, i = 0...n/2$. Threshold value V_p depends on defined level of false alarm α. In order to calculate V_p it is necessary to know the distribution of statistic L (in case of disorder's absence), or to know distribution's moments of statistic L (6).

According to the central limit theorem of probability theory, it is possible to assume that statistic L distributed according to a law close to Gaussian distribution. Thus, in order to obtain information about distribution law of statistic L, it is necessary and sufficient evaluate values of first two moments. According to (6), L is a sum of summands $|G_i|$, each of them is a module of difference between two logarithms of spectrum coefficients. Thus, in order to calculate first two moments it is necessary to obtain first two moments of each of summands $|G_i|$.

In accordance with modeling results which was obtained from experiments, it can be assumed that logarithm of power spectrum coefficient is distributed with Gaussian distribution law. The difference between two logarithms also distributed with Gaussian distribution law. In turn, the module of normal random variable is a value distributed according to the following law:

$$f_\xi(x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad \xi = |\eta|,$$

η is distributed with Gaussian distribution law.

In order to find the moments of verification statistic L it is necessary to calculate the values of moments of Z_i, Q_i. This values is a logarithms of spectrum powers. In turn, the power of i th spectrum's component is a sum of real and imaginary parts of Fourier transformation of incoming sample $x(kT)$: $S^2{}_{x,i} = \mathrm{Re}_{x,i}{}^2 + \mathrm{Im}_{x,i}{}^2$ χ^2 with 2 degrees of freedom:

$$f_{S^2}(x) = \begin{cases} \frac{1}{2a} e^{-\frac{x}{2a}}, & x > 0; \\ 0, & x \le 0, \end{cases}$$

where a is scale parameter which depends on signal power.

Based on the fact that the variance of the logarithmic transformation of a random variable independent of changes in scale, it can be considered constant $a = 1$.

V. Finding of moments distributions of random variables Z_i and Q_i

Since in hypothesis H_0 disorder is absent (these values are distributed equally), it is enough to find the moments of one of these variables.

Mathematical expectation and dispersion of statistics L:

$$m_1\{L\} = \sum_{i=0}^{n/2} m_1\{|G_i|\} = \left(\frac{n}{2}+1\right)\mu_2\{G_i\}\sqrt{\frac{2}{\pi}} = (n+2)\sqrt{\frac{2}{\pi}}\left(\frac{\pi^2}{6} + C^2 - 2C\ln 2 + (\ln 2)^2 - (\ln 2 - C)^2\right),$$

$$\mu_2\{L\} = \sum_{i=0}^{n/2} \mu_2\{|G_i|\} = \sum_{i=0}^{n/2} \mu_2\{G_i\} = \left(\frac{n}{2}+1\right)\mu_2\{G_i\}\left(1 - \frac{2}{\pi}\right)$$

$$= (n+2)\left(\frac{\pi^2}{6} + C^2 - 2C\ln 2 + (\ln 2)^2 - (\ln 2 - C)^2\right)\left(1 - \frac{2}{\pi}\right).$$

Decision threshold value evaluating with expression

$$V_P = m_1\{L\} + \chi_{1-\alpha}\sqrt{\mu_2\{L\}},$$

where $\chi_{1-\alpha}$ is the fractile of $1-\alpha$ level of normalized Gaussian distribution; α is the level of false alarm.

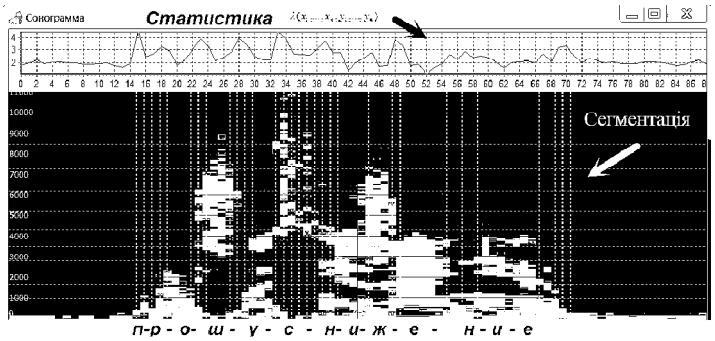

Fig. 2. Segmentation's result on phrase «Прошу снижение» (on Russian). Top chart show values of statistic λ. Vertical dotted lines shows the moments of phoneme changes

CONCLUSIONS

1. For synthesis of sound signals segmentation methods and algorithms, that are invariant to power, it is advisable to use the normalized spectral power.

2. The invariant algorithm of audio signal's segmentation, based on Gaussian distribution model of logarithms of phonemes' spectral characteristics, was synthesized.

3. Suggested algorithm demonstrates high quality of speech signal's separating into allophones and it can be effectively used for automatic recognition of speech messages.

REFERENCES

[1] George Saon and Jen-Tzung Chien. "Large-Vocabulary Continuous Speech Recognition Systems." *IEEE Signal* Processing Magazine. vol. 29, no. 6, November 2012, pp.18–33.

[2] Tsiplihin, A. I. Analis i avtomaticheskaya segmentatsiya rechevogo signala: dis. kand. teh. nauk / IPPI RAN. Moscow, 2006. 149 p. (in Russian).

[3] Sorokin, V. N.; Tsiplihin, A. I. "Segmentatsiya i raspoznavanie glasnyh." *Informatsionnie process*. 2004. vol. 4, no. 2. pp. 202–220. (in Russian).

[4] Dorohin, O. A.; Starushko, D. G.; Fedorov, E. E.; Shelepov, V. Iu. "Segmentatsiya rechevogo signala." *Iskusstvenniy intellect*. no. 3'2000, pp. 450–458. (in Russian).

[5] Rabiner, L. R.; Juang, B.-H. Fundamentals of Speech Recognition. Prentice Hall PTR, 1993. 507 p.

[6] Anderson, T. W. An introduction to multivariate statistical analysis, Wiley (1958).

[7] Kendall, M. G.; Stuart, A. The advanced theory of statistics, 3, Griffin (1983).

Received 28 May 2014.

Prokopenko Kostiantyn. PhD, post-doctoral.
National Aviation University, institute of Air Navigation; aero navigation systems chair, Kyiv, Ukraine.
Education: T. G. Shevchenko Kyiv National University (2001).
Research interests: signal and data processing in aero navigation systems.
Publications: 25.
E-mail: kprok78@gmail.com

К. І. Прокопенко. Адаптивний алгоритм сегментації мовних сигналів

Запропоновано адаптивний алгоритм сегментації мовних сигналів на алофони за допомогою вирішення задачі визначення моменту розладки нестаціонарного процесу.

Ключові слова: мовний сигнал; сегментація; фонеми; алофони; розладка нестаціонарного випадкового процесу.

Прокопенко Костянтин Ігорович. Кандидат технічних наук, докторант.
Кафедра аеронавігаційних систем, інститут аеронавігації, Національний авіаційний університет, Київ, Україна.
Освіта: КНУ ім. Т. Г. Шевченка, факультет кібернетики. Київ, Україна (2001).
Напрям наукової діяльності: обробка сигналів і даних.
Кількість публікацій: 25.
E-mail: kprok78@gmail.com

К. И. Прокопенко. Адаптивный алгоритм сегментации речевых сигналов

Предложен адаптивный алгоритм сегментации речевых сигналов на аллофоны при помощи решения задачи определения момента розладки нестационарного процесса.

Ключевые слова: речевой сигнал; сегментация; фонема; аллофон; розладка нестационарного случайного процесса.

Прокопенко Константин Игоревич. Кандидат технических наук, докторант.
Кафедра аэронавигационных систем, институт аэронавигации, Национальный авиационный университет, Киев, Украина.
Образование: КНУ им. Т. Г. Шевченко, факультет кибернетики. Киев, Украина (2001).
Направление научной деятельности: обработка сигналов и данных.
Количество публикаций: 25.
E-mail: kprok78@gmail.com