ALGORITHM OF FORMING SPEECH BASE UNITS USING THE METHOD OF DYNAMIC PROGRAMMING

[1] YERZHAN N. SEITKULOV, [1] SEILKHAN BORANBAYEV
[2] HENADZI V. DAVYDAU, [2] ALEKSANDR V. PATAPOVICH
[1] L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
[2] Belarusian state university of informatics and radioelectronics, Minsk, Belarus
E-mail: [1] Seitkulov_y@enu.kz, [2] nil53@bsuir.edu.by

ABSTRACT

The article is dedicated to the development of an algorithm of autonomous speech segmentation of the given text using dynamic programming method and creation of base unit bases of a certain announcer. Base units are used to form speech-like signals in systems of active speech information protection. The algorithm is based on the use of text, spoken by one of the announcers, marked and divided manually into structural units, which is the model for an automated speech segmentation of other announcers for the same text. For the speech segmentation of other announcers and formation of structural base units of their speech and to be used with the method of dynamic programming according to the model as done manually before that. The formation of speech-like signals is conducted using compilation method of speech synthesis of the text, which is automatically synthesized including linguistic, phonetic and prosodic features of the given language.

Keywords: *Structural Unit Bases; Speech Segmentation; Dynamic Programming; Compilation Speech Synthesis; Indication Vector.*

1. INTRODUCTION

In the active systems of speech protection in rooms for talks as a masking signal, combined masking signals are often used, which consist of 'white' noise and speech-like signals [1], [2]. It is advised to use speech sequences as speech-like signals, formed using linguistic features of a language and statistical characteristics of phonemes in the given language, as well as length of words and length of sentences. Formation of speech-like signals is used by the compilation method in the structural speech unit bases. As a result, speech-liked signals formed in such a way retain all shades of speech of a particular announcer and it is rather difficult to distinguish from informational signals of the same announcer [3].

In order to create structural base unit bases, the methods of speech segmentation are used which can be divided into two categories. Segmentation methods based on the use of a priori information about the segmented speech signal [4]-[7], and speech segmentation methods which do not use information about the segmented signal and designed for speech detection, announcer verification by voice, verification of the language spoken by the announcer [8]-[12]. Together with the speech segmentation method we can highlight the methods of segmentation into phonetical elements, which are needed for the tasks of speech detection as well as its synthesis.

With the high accuracy of speech signal segmentation into phonemes for the automatic speech detection, identification and verification of the announcer, detection of bilinguals would have been carried out with higher accuracy and speed due to a small number of phoneme sequences. However, if during speech segmentation into such structural elements as sentences, phonetic paragraphs, words of segment limits can be detected precisely not depending on the expert during the manual segmentation and using the methods of automatic speech segmentation and not employing complex computations and small number of parameters in the indication vector or during speech segmentation into phonemic limit elements, these elements are difficult to establish.

Human vocal apparatus is created in such a way that it is impossible to detect its form and dynamics from the articulation of connected speech. Some phonemes transform into others without any distinct limits. When one phoneme transforms into another, the vocal apparatus must switch and take such a position, that the next phoneme can be formed; this is why it is impossible to detect the border between the phonemes. A more precise detection of such borders will be done according to the tasks, the solution of which requires speech segmentation into structural phonemic elements. The precision borders under automatic speech segmentation into structural phonemic units in a number of works is shown according to the border of structural phonemic speech units acquired through manual segmentation carried out by professional linguists [13]. However, this doesn't have the aim and the solution to which tasks the speech segmentation is carried out.

During the speech synthesis, carried out by the compilation method using structural phonemic speech units cannot always provide qualitative speech synthesis or speech-like sequences, even though the prosody formation methods are used. To increase the quality of speech synthesis is possible using exponential spline functions on the borders of transformation of one phonemic structure into another and applying endings of one phonemic structure to the beginning of another phonemic structure. At the same time, a kind of amplitude damping of frequencies of ending of one phonemic structure and increase of the same amplitude of the next phonemic structure. In order to execute such compilation speech synthesis a database of phonemic structures is required, containing increased segments in time, as it is during synthesis the imposition of ending of one phonemic structure on the next one occurs.

The method speech analysis has shown, that in order to form phonemic structural base unit bases for the synthesis carried out by compilation method the most comfortable segmentation method is segmentation which uses dynamic programming. For this, it is important to have phonetic recording of connected speech, manually marked and divided into structural phonemic elements and which contains all of the structural phonemic units required for the basis. Usually, it is around 300-400 allophones for Russian, Kazakh and Belarussian speech and 1200 structural phonemic units for Chinese, as this language is tonal.

The aim of the work was the development of an algorithm for automatic segmentation of speech on given text by the method of dynamic programming and the creation of the base of structural units of speech.

2. REQUIREMENTS FOR MODEL SPEECH AND ITS MARKING INTO STRUCTURAL PHONEMIC UNITS MANUALLY

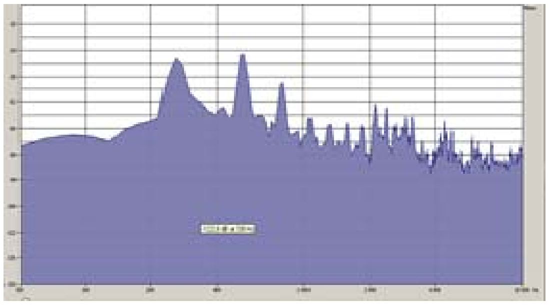
The model speech, supposed for manual segmentation and used for creating model basis of structural phonetic elements, should be connected speech. Apart from that, the model speech must have all phonetic elements, important for the creation of the basis. The works of V. Sorokin [14]-[18] show, that speech signal segmentation must carry out the search for borders of quasi-stationary and transient processes, based on the correlation between short-term spectrum of equispaced in time parts of the signal. As quasi-stationary parts for Russian speech we can single out six types of phonemic elements, which have various speech-formation mechanisms and articulation during the pronunciation of such elements. These parts, which contain vowel-like, nasal, fricative voiceless, fricative voiced, occlusive voiceless and occlusive voiced speech sounds. Such segment classification of speech, proposed in the work [18], based on the characteristical meanings of the minimal values of the cross-section area of the vocal tract and passage into the nasal cavity and for the presence of vocal impulses.

Vocal sounds of the speech spread out on that part of the tongue, which is raised when pronouncing this sound and are characterized as frontal, middle and back. Apart from that, they are distinguished according to the level of the elation of this or that part of the tongue and are classified as upper, middle and lower. With that, they are divided into labial (lips are used to form this sound, like o) and non-labialised (lips were not used to form this sound). The formation of vowels is characterized by the absence of obstacles in the speech tract, when at the same time, during the formation of consonants there is an obligatory full or semi full joint which is created by the tongue or the lips. However, precise pronunciation of the vowels happens only when they are stressed, in all other cases there is a reduction (they change). This imposes definite difficulties when choosing an element for the structural phonemic base unit bases.

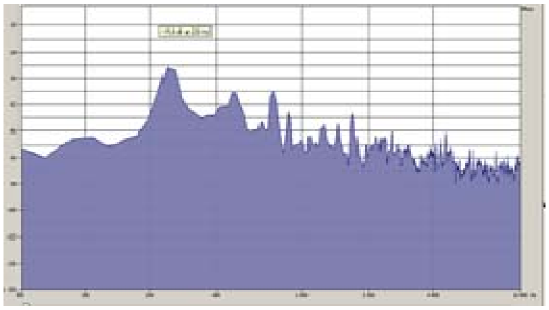
B. Lobanov in his work [19] proposes to divide vowel phonemes according to the formation criterion into the groups of sonorous, fricative, plosive and affricatives, which is very comfortable when forming phonemic base units for its synthesis. For the groups of sonorous consonants, it is significant to have a wide gap in the vocal apparatus. Fricative consonants are divided into voiced and voiceless depending on whether or not vocal chords were used in their formation. Voiced consonant phonemes are characterized by the presence of impulses, created by the vocal chords. Plosive consonants are created with the complete joint in the vocal apparatus with the sharp opening. Just as fricative consonants, plosive consonants are divided into voiced and voiceless. Phonemes of voiceless affricatives are formed by the change of joint phase by the noise – producing gap. In his work [20] it is stated, that the affricative energy lies in the range of more than 3000 Hz and for their detection in the connected speech it is advised to use S-transformation. Fricative sound segment was detected by the means of detecting frequency of spectrum's center of gravity, length of the voiceless part and the level of irregularity of the smoothed spectrum, calculated as a correlation of the max. Frequency derivative on fricative middle third to the max. energy across all of the sound combination.

Due to the reduction effects and carticulation, it is rather difficult to establish borders of the phonemes. Coarticulation is when a consonant phoneme, to a considerable extent, takes on the shade of the following vowel phoneme and vowel phoneme – takes on the shade of the preceding consonant.

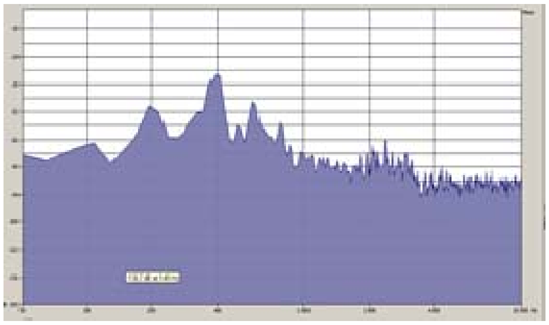
Proof of this is a spectrogram of phonetical unit 'li' from the beginning, middle and end of the word is shown in figure 1.



a)



b)



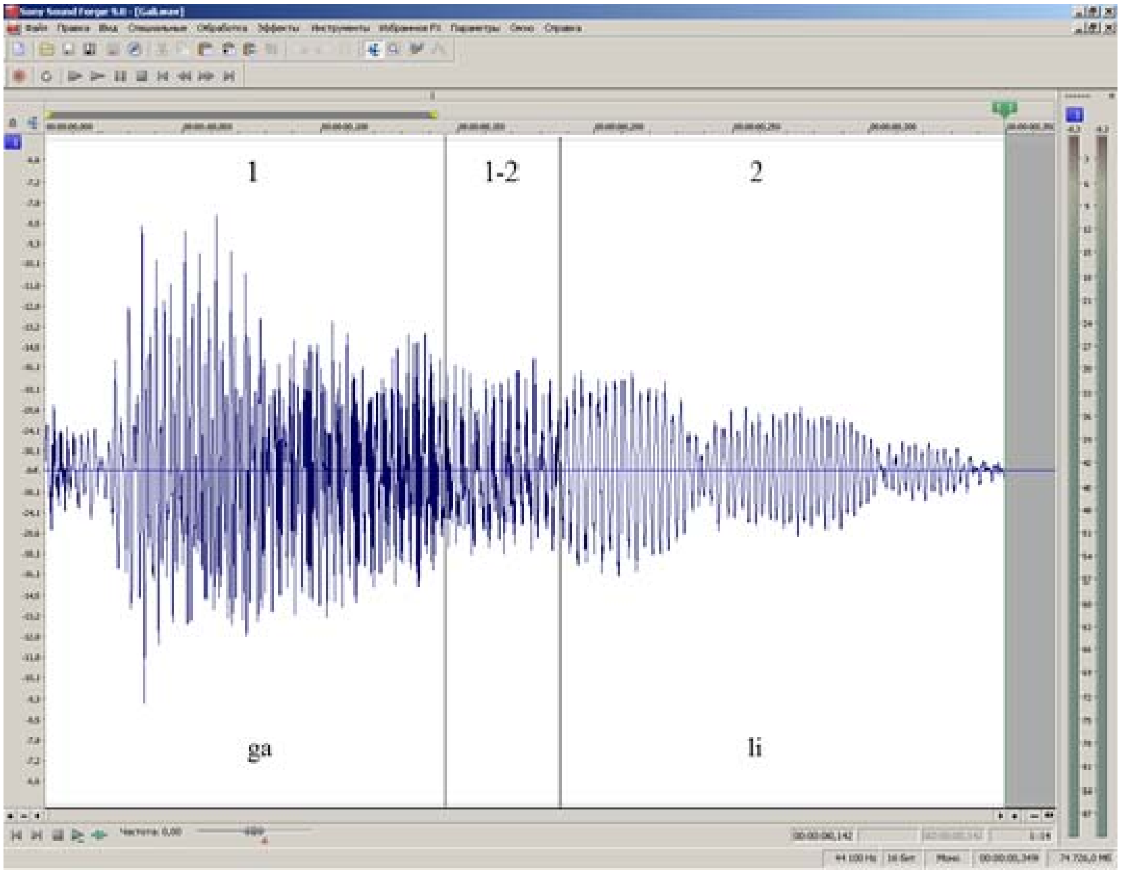
c)

Figure 1: Spectograms of phonetical speech unit "li": a) for the middle of the word; b) for the beginning of the word; c) for the end of the word.

As seen from the spectrograms shown in figure 1, depending on the position in the word, phonetical speech unit 'li' has significant differences as in formant frequencies, as well as their amplitudes. This shows that during the formation of structural phonemic speech units for the synthesis of speech-like signals and speech synthesis, the basis should include phonetical realisations of one phonemic unit its depending on the environment. It is necessary to count in the preceding phonetical elements and the following.

In the majority of works dedicated to the speech segmentation algorithm begins with the detection of energy parameters of the segments and their sharp changes, next – we calculate the spectral parts of the signal and detect the part with changes, something that is usually connected with the change in articulation [5], [8], [9]. Next, we can conduct analysis of static characteristics of the segment by averaging its spectrum and comparing the specters in neighbouring frames, calculation of cepstral coefficient. In the works [21], [22] it is proposed to use the differences in phases of neighbouring frames in order to detect borders of the phonemes for various frequency areas.

As the structure analysis of phonetical units of connected speech shows, it is impossible to detect clearly and distinctly the phoneme borders, as there is a part where the phoneme is clearly visible and there is a transient area. These areas are present before the phoneme and after it. Depending on the environment, the length of transient areas can vary. figure 2 shows time relisation of phonetical structure gali, taken from the connected speech of Russian-speaking announcer.



ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

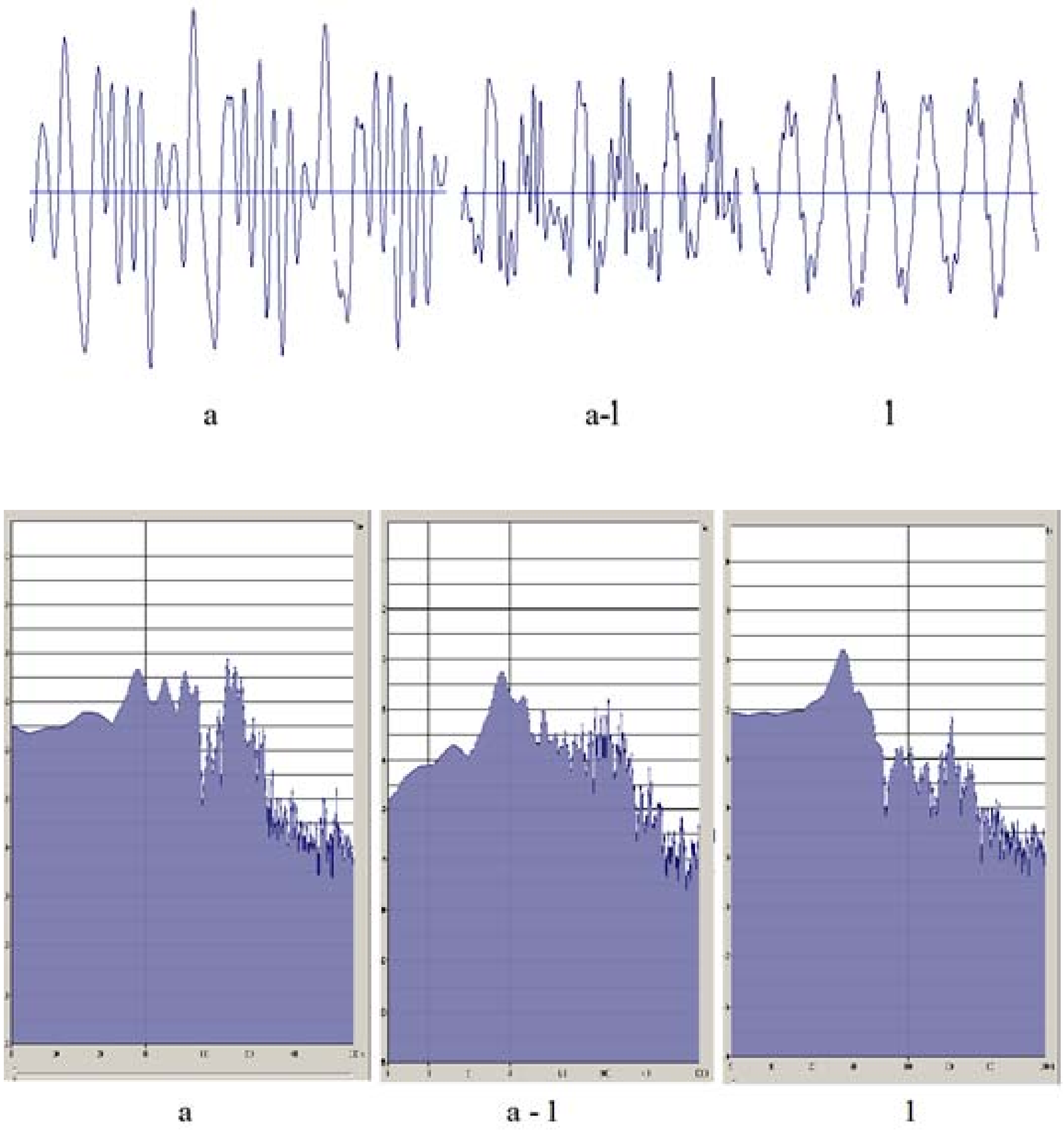


Figure 2: Time realization of phonetical structure gali.

This realization marks three time parts: part 1 – phonetical structure (syllable) ga; part 2 – phonetical structure li; part 1-2 is a transient area of phonetical structure ga into phonetical structure li. Lower, the parts with realization with lengths of 17 ms each are shown: a – part of the phoneme a; a-l-transition of phoneme a into phoneme l; l – realization part of phoneme l. The realization part of phoneme a shows clear and three repeated parts (in their form). We called these parts domains. The same domains can be marked on the transient part of the phoneme a into phoneme l and on the realization part of the phoneme. Their number stays the same (three) with the length of 17 ms, which means, that the length of the domain is 5,65 ms. This means that the frequency of the main tone was around 177 Hz. Practice showed, that with manual segmentation, using the information on domains changing forms, it is possible to quite clearly to set transition borders of one phoneme into another. For the realization, shown in the picture, the length of the transition part from a to l is 38,7 ms. The same picture shows, at the bottom, spectrum for corresponding phoneme parts and transition areas are shown. It should be noted, that the use of wavelet transforms (DWT)) as basic functions, the form of separate domains is less dependent on the language and more on the announcer.

For other 30 announcers the size of the transient area from phoneme a to l for phonetical structure gali was in the range from 25 to 45 ms. The value of these area between other phoneme combinations can be different and fluctuate from 11 to 50 ms. This is defined by the speech rate of the announcer, clear pronunciation and peculiarities of speech apparatus. During the speech segmentation into phonemes or allophones it is necessary to include transient areas into its length, i.e. the transient area preceding the given phoneme, the area of the phoneme and the transient area after the phoneme – transition to another phoneme. In this way, B.Lobanov proposed allophone basis for speech synthesis of the Russiana text to form from 498-552 elements [19], as well as index system to define allophones during the automatic speech synthesis. During the creation of multiphone basis the number of the elements increased from 6000 to 7000 [19]. Fort the synthesis of speech-like signals, allophone basis can be reduced to 320 elements for Russian and 342 for Belarussian. For Kazakh – 245 elements. During the creation of reduced phonetical structure speech unit bases, designed for the synthesis of speech-like signals, the bases did not include elements, which have small probability of occurrence in the given language.

Requirements for the model texts for manual segmentation need to contain all the base elements from the most frequently used words in the given language. Apart from that, speech must be connected.

Since the phonetical structural speech, element bases in the beginning and in the end contain transient areas, then during the speech synthesis, it is advised to use the so-called ‘stitching’ of the allophones during the compilation speech synthesis [19]. Transient area of the ending of the preceding allophone, multiplied by the decreasing function, changing from 1 to 0, is imposed on the transient area of the following allophone, multiplied by the increasing function form 0 to 1. If the lengths of the imposed transient areas are not equal to each other, then the length of the formed transient area is chosen to be equal to the length of the longer transition. The missing part of the smaller transient areas is complemented by 0 amplitude values.

In work [19] it is proposed to multiply transient areas on the linear function: changing from 1 to 0 and from 0 to 1. However, with this, the sensitivity of hearing has non-linear character, the use of spline function of higher order is more effective to use (till cube).

3. DYNAMIC PROGRAMMING METHOD DURING SPEECH SEGMENTATION

Theoretical bases for the method of dynamic programming during speech segmentation are rooted in the principle of Bellman equation for the task of discrete optimal control [23]. The development of this method to detect and synthesize speech can be seen in works [4], [7], [24] – [28].The method of dynamic programming allows finding optimal solution in multidimensional tasks by dividing it into stages. At each stage the task is solved according to one variable for many parameters, which describe the condition of the system not depending on the condition of the system during the previous step. With this, instead of multidimensional task, one-dimensional optimization tasks are solved at each stage. Thus, the task is only to find min. value of functional H during the system transition from the initial to final state.

A very important and determining factor is the choice of vectors of informational parameters. As it was mentioned before, for speech segmentation of a new announcer it is necessary to have the manually segmented model speech. With this, the model speech of the announcer must be the same as the segmented speech. We can use sonogram of model speech as parameter informational vector, as well as final differences of the spectrum, cepstral coefficients, frequency measure of the main tone and final differences of the main tone frequency. Due to the fact, that the rate of the model speech can differ from the segmented speech it is necessary to input controlling coefficients [4]. The values of the controlling coefficients can be corrected including the length of the model and segmented speech.

As the main informational parameter vector we use spectral signal description (sonogram). In the work [4] it is proposed to use 10 bandpass filters, which overlap the frequency range up to 5900 Hz with digitization in time of every channel in 18 ms.

Before the beginning of the segmentation process it is necessary to conduct preliminary preparations of the speech signal recordings:

- Conduct digital digitization of speech signals (reduce them to the same digitization frequency, preferably, equal to 16 kHz);

- Conduct scaling of speech signals according to energy parameters (acting values of the model speech signal must be equal to the acting values of the segmented speech signal Ux = Uy in all of the recording, where Ux – acting value of model speech signal; Uy – acting values of the segmented speech signal);

- Carry out segmentation of the model speech signal of allophones including the transient areas of one phoneme into another and mark the model speech signal.

The generalized block-diagram of the dynamic programming segmentation algorithm method is shown in Figure 3.

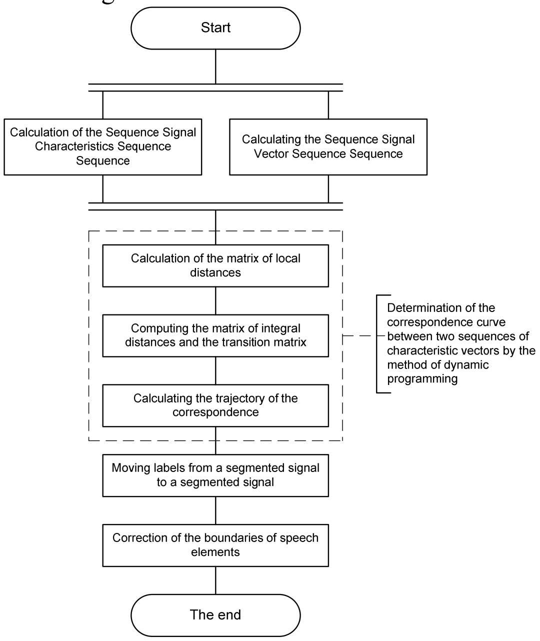

Start
Calculation of the Sequence Signal Characteristics Sequence Sequence
Calculating the Sequence Signal Vector Sequence Sequence
Calculation of the matrix of local distances
Computing the matrix of integral distances and the transition matrix
Calculating the trajectory of the correspondence
Determination of the correspondence curve between two sequences of characteristic vectors by the method of dynamic programming
Moving labels from a segmented signal to a segmented signal
Correction of the boundaries of speech elements
The end

Figure 3: Block-diagram of the dynamic programming segmentation algorithm method.

In a general case, segmentation process which uses dynamic programming with control coefficient is based on the following positions. The information indicator vector of the model signal (speech sequence, manual segmentation) is described sequentially

$$\overline{X}=\{X_1, X_2,...,X_m,...,X_M\} ,$$

where $X_1, \ldots, X_M$ – component of vector.

The information indicator vector for the segmented speech analysis is written down as

$$\overline{Y}=\{Y_1, Y_2,...,Y_m,...,Y_M\} ,$$

where $Y_1, \ldots, Y_M$ – component of vector .

It is proposed to arrange information indicator vector in the following succession:

- X_1 и Y_1 – acting value of the segmented model signal of speech signal and correspondingly, acting value of the segmented speech signal on the time interval of averaging К;

- $X_2, X_3, \ldots, X_{21}$ и $Y_2, Y_3, \ldots, Y_{21}$ – spectral components of model speech signal and segmented speech signal for frequency bands of 100 – 200, 200 – 300, 300 – 400, 400 – 510, 510 – 630, 630 – 770, 770 – 920, 920 – 1080, 1080 – 1270, 1270 – 1480, 1480 – 1720, 1720 – 2000, 2000 – 2320, 2320 – 2700, 2700 – 3150, 3150 – 3700, 3700 – 4400, 4400 – 5300, 5300 – 6400, 6400, 6400 – 7800 Hz (in order to receive spectral components of speech signal, we can use Butterworth filter bank of 3rd or 4th orders);

- $X_{22}, X_{23}, \ldots, X_{41}$ и $Y_{22}, Y_{23}, \ldots, Y_{41}$ – averaged final differences of spectral components on time interval of averaging K, accordingly X is for model segmented speech signal and Y for segmented speech signal;

- X_{42}, Y_{42} –contribution of the main tone frequency;

- X_{43}, Y_{43} – final contribution differences of main tone frequency;

- $X_{44}, X_{45}, \ldots, X_{63}$ и $Y_{44}, Y_{45}, \ldots, Y_{63}$ – cepstral coefficients for the frequency in range from 100 to 10000 Hz with 20-frequency bands, accordingly X

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

for the model speech signal segmentation and Y for segmented speech signal.

Acting value of the segmented speech signal on the K averaging interval is defined from the formula:

$$U_k = \sqrt{\frac{1}{M}\sum_{m=1}^{M} U_{k+m}^2} \ ,$$

where M – number of counts on the K averaging area; U_{k+m} - signal amplitude in point m of area K of the segmented speech signal.

Spectral components of the model speech signal and the segmented signal can be defined on the interval of 20 -25 ms.

Averaged final differences of the spectrum for each channel are set by the formula:

$$\Delta X_{fi,k} = \frac{1}{M}\sum_{m=1}^{M}\left(X_{fi,k+m} - X_{fi,k+m-1}\right),$$

where M – number of counts on the K sonogram area; $X_{fi,k+m}$ - sonogram component for K area with placement in point m for I frequency channel.

In order to define the frequency component it is necessary to filter signals in frequency range from 60 to 500 Hz from the model speech signal and segmented speech signal, since this frequency range lies the frequency of vocalized phonemes of the announcer (frequency of the main tone). With that, frequency of the main tone and its changes are set according to the methods, discussed in works [29] – [31] using linear and median filtration for the correction of frequency curve of the main tone.

Contribution of the component frequency of the main tone in the spectrum of speech signal is proposed to be defined by the formula:

$$X_{42,K} = 1 - \min F_{0,K,m} \, ,$$

where $F_{0,K,m}$ - reduced to the range from 0 to 1 contribution frequency value of the main tone in the spectrum of speech signal for k window.

Final frequency contribution differences of the main tone are defined by:

$$\Delta X_{42,k} = \frac{1}{M}\sum_{m=1}^{M}\left(F_{0,k+m} - F_{0,k+m-1}\right),$$

where $F_{0,K,m}$ - reduced to the range from 0 to 1 frequency contribution value in the spectrum of speech signal for k window.

Cepstral coefficients can be defined using discrete cosine transform:

$$X_{c,fi,K} = \sum_{m=1}^{M} S_m \cos\left(\frac{\pi(2m+1)}{2M}\right),$$

where Sm – logarithm energy at the filter output.

Let‘s have a look at the juxtaposition procedure of the model (segmented) and analyzed (segmenting) signals using dynamic programming method with control coefficients.

The first step in this procedure would be detecting matrix of local distances $d_{n,m}$ between model vectors and current speech signals:

$$d_{n,m} = \frac{1}{L}\sum_{l=1}^{L}\left|X_{n,l} - Y_{m,l}\right| ,$$

where L – dimension of model and current speech signal vectors.

Integral distances matrix $D_{n,m}$ and transition matrix $Tr_{n,m}$ are calculated according to the following initial conditions an d recurring formulas:

$$D_{1,1} = d_{1,1} \ ; \ Tr_{1,1} = TrEnd \ ; \quad Q = \sqrt{N^2 + M^2} \ ;$$

$$Dh_{n,m} = D_{n-1,m} + k_H d_{n,m} + \frac{k_T}{Q}\left|N \cdot m - M \cdot (n-1)\right| ;$$

$$Dv_{n,m} = D_{n,m-1} + k_V d_{n,m} + \frac{k_T}{Q}\left|N \cdot (m-1) - M \cdot n\right| ;$$

$$Dd_{n,m} = D_{n-1,m-1} + k_D d_{n,m} + \frac{k_T}{Q}\left|N \cdot (m-1) - M \cdot (n-1)\right| ;$$

$$D_{n,m} = \min\left[Dh_{n,m}, Dv_{n,m}, Dd_{n,m}\right] \ ;$$

$$Tr_{n,m} = \begin{cases} TrHoriz, & \text{если} \quad D_{n,m} = Dh_{n,m} \\ TrVert, & \text{если} \quad D_{n,m} = Dv_{n,m} \\ TrDiag, & \text{если} \quad D_{n,m} = Dd_{n,m} \end{cases} ,$$

where k_T – time weight ratio; k_H, k_V, k_D – transition coefficients horizontally, vertically and diagonally accordingly.

The scheme of the dynamic programming method is shown in the following picture, using juxtaposition method. To make it easier, figure 4 shows time weight ratio which equals 0.

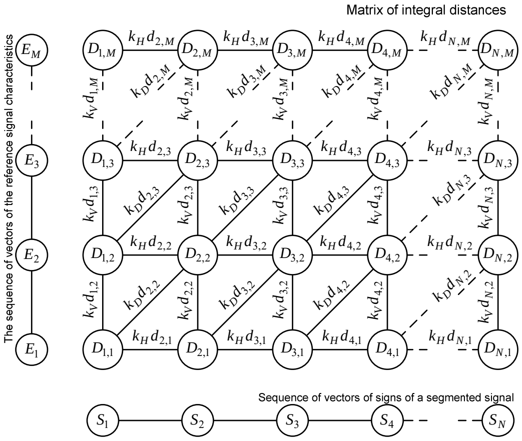

Matrix of integral distances
The sequence of vectors of the reference signal characteristics
E_M
E_3
E_2
E_1
$D_{1,M}$ $D_{2,M}$ $D_{3,M}$ $D_{4,M}$ $D_{N,M}$
$D_{1,3}$ $D_{2,3}$ $D_{3,3}$ $D_{4,3}$ $D_{N,3}$
$D_{1,2}$ $D_{2,2}$ $D_{3,2}$ $D_{4,2}$ $D_{N,2}$
$D_{1,1}$ $D_{2,1}$ $D_{3,1}$ $D_{4,1}$ $D_{N,1}$
$k_H d_{2,M}$ $k_H d_{3,M}$ $k_H d_{4,M}$ $k_H d_{N,M}$
$k_H d_{2,3}$ $k_H d_{3,3}$ $k_H d_{4,3}$ $k_H d_{N,3}$
$k_H d_{2,2}$ $k_H d_{3,2}$ $k_H d_{4,2}$ $k_H d_{N,2}$
$k_H d_{2,1}$ $k_H d_{3,1}$ $k_H d_{4,1}$ $k_H d_{N,1}$
$k_V d_{1,M}$ $k_V d_{2,M}$ $k_V d_{3,M}$ $k_V d_{4,M}$ $k_V d_{N,M}$
$k_V d_{1,3}$ $k_V d_{2,3}$ $k_V d_{3,3}$ $k_V d_{4,3}$ $k_V d_{N,3}$
$k_V d_{1,2}$ $k_V d_{2,2}$ $k_V d_{3,2}$ $k_V d_{4,2}$ $k_V d_{N,2}$
$k_D d_{2,M}$ $k_D d_{3,M}$ $k_D d_{4,M}$ $k_D d_{N,M}$
$k_D d_{2,3}$ $k_D d_{3,3}$ $k_D d_{4,3}$ $k_D d_{N,3}$
$k_D d_{2,2}$ $k_D d_{3,2}$ $k_D d_{4,2}$ $k_D d_{N,2}$
Sequence of vectors of signs of a segmented signal
S_1 S_2 S_3 S_4 S_N

Figure 4: Scheme of juxtaposition of successions of dynamic programming method.

As seen from the picture shown above, control horizontal coefficients (k_H),vertical (k_V) and diagonal (k_D) are locally weight coefficients. Their use allows setting priority 'mode' of transitions between the points in integral distances matrix: horizontally, vertically and diagonally accordingly. The less the value of the chosen coefficient is to other coefficients, the smaller integral mistake is accumulated with the given 'mode', therefore, the more priority is the chosen 'mode' transition. Time weight coefficient k_T is part of global limitations of the correspondence trajectory search between the compared successions.

The optimal values of the control coefficients are defined during the stage of setting the segmentation system up. With this, as initial values we recommend using control coefficients k_T= 0, k_H= 1, k_V =1, k_D =1.

The result of the succession juxtaposition using dynamic programming method is correspondence curve, in according with which the marking is transferred from the segmented recording to the processed recording and further, into the structural speech units bases for the given announcer. With this, in order to receive the best result while used in compilation synthesis systems the desired activity would be overlapping borders of vocalized segments with position of main tone impulses. This is carried out to decrease distortions during the speech signal synthesis. To solve this task, we define vocalized areas of speech on the processed recording and mark them according to the excitation pulses of the vocal apparatus.

Then, the border correction of allophones in the marker position of the main tone can be carried out according to the criteria of min. distance.

4. FORMATION ALGORITHM OF SPEECH UNIT BASES FOR THE KAZAKH LANGUAGE AND SYTHESIS OF SPEECH-LIKE SIGNALS

The word-formation process in the Kazakh language follows the rules of vowel harmony. The word-formation happens by adding suffixes to the root of the word first and then adding multiplicity endings, possessive endings, case endings and conjugation endings. There are no words with one letter int the Kazakh language. Word-formation of complex words in the Kazakh language can be carried out by two main ways: by adding bases; by transforming word combination s into complex words.

It should be noted that letters v, f, ch, e are used in Kazakh only to write down words of foreign origin.

Letters v, f, h, ch, sh, e, are not used in originally Kazakh words. Out of those the letters ch, sh, e are used to pronounce the words borrowed from the Russian language. Letter h is used in words, borrowed from Arabic and Persian languages and is pronounced as voiceless 'x'. An important feature of the Kazakh language is the presence of the law of sound harmony (vowel harmony). With this law, all vowels in the words must be either front vowel or back vowel and all the consonants, correspondingly, must be soft or hard. As a result, the pronounced words in Kazakh can be either soft or hard.

When forming allophones, which correspond to the letter of the alphabet, including their phonetical characteristics the following scheme was used - a, o, u, I were hard and a - á, e, i, ó, ú – soft.

The adjoining vowel sounds would show whether the sounds were soft or hard.

Stressed or unstressed vowels are defined by their place in the word. The stressed vowels were in the last syllable of the word. All the rest of the vowels were unstressed.

It was proposed to form structural speech unit bases for the synthesis of speech-like signals in the Kazakh language to be formed on the basis of allophones and the basis, which includes suffixes and endings. The use of two bases would allow to decrease the number of calculations with automated synthesis of speech-like signals and make the speech-like signals word-formation process easier. Using the allophone basis the root of the word is formed and the rest of the word (suffixes and endings) are added from the base of suffixes and endings. The total number of allophones of the Kazakh language for the synthesis of speech-like signals is 283. The allophone base includes letters of the Kazakh alphabet with their placement and environment in the text.

If the first digital index after the vowel allophone 0, then it means that the vowel allophone is not stressed. If the first digital index of the vowel allophone 1,then the vowel allophone is stressed. For consonant allophones the first digital index is 1 and it means that the consonant allophone is hard. If the first digital index for consonant allophone is 2, then it means that the consonant allophone is soft.

The second index of the allophone characterizes its environment on the left. If the second index is 0, then it means that the environment is absent, i.e. this allophone starts a new word. If the second index is 1, then on its left it is preceded by vowel phoneme. If the second index is 2, then on its left it is preceded by the consonant phoneme.

The third index of the allophone characterized its environment on the right. If the third index is 0, it means that its environment on the right is absent, i.e. this allophone finishes the word. If the third index is 1, then on its right a vowel phoneme follows it. If the third index is 2, then on its right a consonant phoneme follows it. The list of the allophones of the Kazakh language can be seen in Table 1.

Table 1: List of allophones in the Kazakh language.

a001	v10	e001	i022	l22	o022	r22	u110	f22	sh21
a002	v11	e002	i102	l20	o102	r20	u112	f20	sh22
a012	v12	e012	i110	m01	o110	s01	u120	h01	sh20
a021	v20	e021	i112	m02	o112	s02	u122	h02	y001
a022	v21	e022	i120	m10	o120	s10	ú001	h10	y002
a102	v22	e102	i122	m11	o122	s11	ú002	x11	y012
a110	g01	e110	i001	m12	ó001	s12	ú012	x12	y021
a112	g02	e112	i002	m21	ó002	s21	ú021	x21	y022
a120	g10	e120	i012	m22	ó012	s22	ú022	x22	y102
a122	g11	e122	i021	m20	ó021	s20	ú102	x20	y110
á001	g12	j01	i022	n01	ó022	t01	ú110	s01	y112
á002	g20	j02	i102	n02	ó102	t02	ú112	s02	y120
á012	g21	j10	i110	n10	ó110	t10	ú120	s11	y122
á021	g22	j11	i112	n11	ó112	t11	ú122	s12	
á022	ǵ01	j12	i120	n12	ó120	t12	u001	s21	
á102	ǵ02	j20	i122	n21	ó122	t21	u002	s22	
á110	ǵ10	j21	k01	n22	p01	t22	u012	s20	
á112	ǵ11	j22	k02	n20	p02	t20	u021	ch01	
á120	ǵ12	z01	k10	ң01	p10	t10	u022	ch02	
á122	ǵ20	z02	k11	ң02	p11	t11	u102	ch10	
b01	ǵ21	z10	q12	ń10	p12	t12	u110	ch11	
b02	ǵ22	z11	q21	ń11	p21	t21	u112	ch12	
b10	d01	z12	q22	ń12	p22	t22	u120	ch21	
b11	d02	z20	q20	ń21	p20	t20	u122	ch22	
b12	d10	z21	l01	ń22	r01	u001	f01	sh20	
b21	d11	z22	l02	ń20	r02	u002	f02	sh01	
b22	d12	i001	l10	o001	r10	u012	f10	sh02	
b20	d20	i002	l11	o002	r11	u021	f11	sh10	
v01	d21	i012	l12	o012	r12	u022	f12	sh11	
v02	d22	i021	l21	o021	r21	u102	f21	sh12	

Allophones which are not solely Kazakh and are borrowed form other languages are less in numbers since they are rarely used in the Kazakh language. The suffix and ending basis is shown in Table 2.

Table 2: Suffix and ending base for the synthesis of speech-like signals.

mın	sız	mız	dın	tar	bak	sızdar	túr	ba	ta
min	siz	miz	din	ter	bek	sizdar	júr	be	te
bın	nız	dan	tın	lar	pak	ndar	dı	pa	nda
bin	niz	den	tin	ler	pek	nder	di	pe	nde
pın	bız	tan	nın	ben	diki	nızdar	tı	ma	na
pin	biz	ten	nin	pen	mek	nizdar	ti	me	ne
sın	pız	nan	dar	men	sındar	otır	nı	da	niki
sin	piz	nen	der	mak	sindar	jatúr	ni	de	tiki

Synthesis algorithm of speech-like signals using allophone, suffix and ending bases include the root of the word according to the allophone base and complemented by its suffix and ending from the corresponding basis depending on the part of speech that is being synthesized. There are 9 parts of speech in the Kazakh language: nouns, adjectives, numerals, verbs, pronouns, adverbs, conjunctions, postpositions, interjections and five of them change in number and person.

The synthesis algorithm of speech-like signals based on the structural speech units are shown in figure 5.

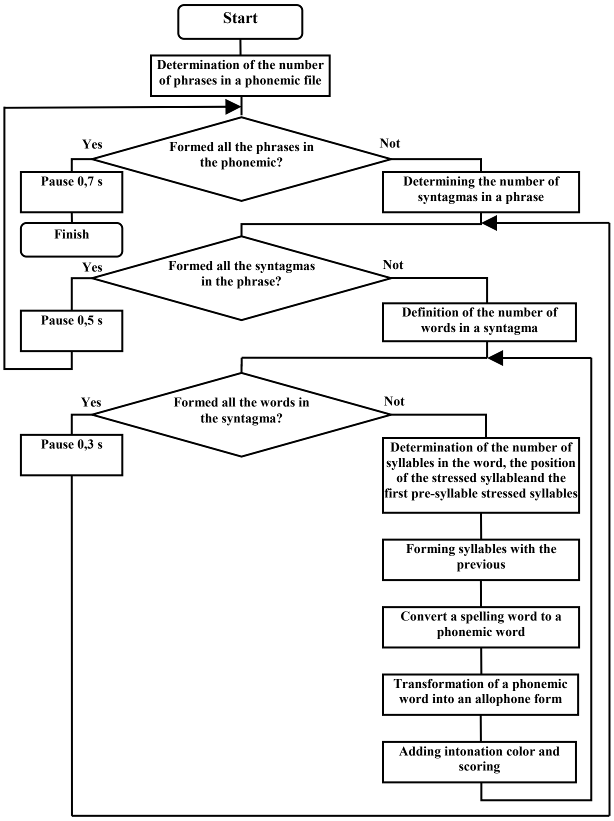

Start
Determination of the number of phrases in a phonemic file
Yes
Formed all the phrases in the phonemic?
Not
Pause 0,7 s
Determining the number of syntagmas in a phrase
Finish
Yes
Formed all the syntagmas in the phrase?
Not
Pause 0,5 s
Definition of the number of words in a syntagma
Yes
Formed all the words in the syntagma?
Not
Pause 0,3 s
Determination of the number of syllables in the word, the position of the stressed syllableand the first pre-syllable stressed syllables
Forming syllables with the previous
Convert a spelling word to a phonemic word
Transformation of a phonemic word into an allophone form
Adding intonation color and scoring

Figure 5: Synthesis algorithm of speech-like signals in Kazakh based on structural speech units.

The work of the synthesis algorithm of speech-like signals in the Kazakh language. Random number generator forms a complete number in the give number range. According to the value of this number and according to the number probability table in the phonetic paragraph with the corresponding division of the paragraph of random numbers into sub-ranges the values of the number of phrases of the given phonetic paragraph. Then, for each phrase we need to define the number of syntagmas from which it consists. For this, we generate a new random number. According to the given value from the number probability table of syntagmas in the phrase and the corresponding division of random numbers into sub-ranges, we define the sub-range under which the random number falls abd the corresponding sub-range values of the number of syntagmas in the second phrase. The process goes on until the number of the syntagmas will not be defined for each phrase of the phonetic paragraph.

After that we form new random number, the values of which defines the number of syllables int he first syntagma according to the number probability table of the syntagmas and the corresponding division of the random number range into sub-ranges. Thus, we defined the number of syllables in the first syntagma. Similarly to this, we defined the number of words in each syntagma.

After that we form the next random number, the value of which defined the number of letters in the first words according to the probability number table of letter in the word and the corresponding division of the random number range into sub-ranges. Similarly to this we define the number of letters in each words of all the syntagmas.

After this we form another random number the values of which defined the first letter in the first word according to the probability number table of letters in the word and the corresponding division of random number range into sub-ranges. Similarly to this, we define the following letters for each word. It is necessary to count in the peculiarities of the Kazakh language mentioned above. If the chosen letter does not correspond to the peculiarities of the Kazakh language, then we form a new random number to choose another letter, which does not come into conflict with the peculiarities of the Kazakh language.

The peculiarities of the Kazakh language are taken into account in the synthesis algorithm of speech-like signals in Kazakh. In the block ‘Syllable formation including the previous one’ a choice of vowels occurs according to the first vowel in this language. If the first vowel is soft, then the following vowels in the word must be soft.

Thus, a text with speech-like signals is formed. Text is acoustically reproduced on the basis of structural speech unit bases by successive reproduction as wav. files which are present in the structural speech unit basis.

5. EXPERIMENTAL RESULTS

Experimental studies of the accuracy of speech segmentation on structural units were evaluated by comparing the results of automatic segmentation with the results of segmentation performed by an expert. To perform a comparative analysis of the accuracy of speech segmentation by various methods, a parameter in the quantitative sense is necessary. As a parameter characterizing the segmentation accuracy, it is proposed to use the effective value of the segmentation error e_{rms}, which is determined from the expression:

$$e_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2},$$

where e_i is value of the segmentation error for i structural unit of the speech, ms.

The accuracy of segmentation was estimated from a database of 40 records of four speakers (two men and two women). Data in the database was recorded at a sampling frequency of 22050 Hz and quantization of 16 bits/sample.

The experimental analysis of the accuracy of speech segmentation on the basis of dynamic programming with control coefficients, carried out on the sound database of four speakers for the feature vector, consisting of the spectrum, fraction of the tone component and their time-averaged finite differences. It made it possible to determine the optimal parameters of the system operation. The effective value of the segmentation error was 6.9 ms. Comparison of the segmentation error of the proposed algorithm with the results presented in [32] shows that for the best Adapted CDHMM method for the tolerance of 8 ms, the percentage of error in this interval is 44 %, and after refinement of the algorithm 67 %

By means of experimental studies for the algorithm of segmentation of speech by the method of dynamic programming with the known beginning and end of the phonogram, the optimal control coefficients are determined: the horizontal transition is 1; diagonal transition is 1; the vertical transition is 1, and for the segmentation algorithm by the dynamic programming method with an arbitrary start and end of the phonogram, the optimal control coefficients are: the horizontal transition is 0.3; the diagonal transition is 0.7; the vertical transition is 1.1.

6. CONCLUSION

In this paper, an algorithm for automatic segmentation of speech on a given text by the method of dynamic programming using control coefficients was proposed, which made it possible to reduce the segmentation error to 6.9 ms. Formation of the base of structural units of speech in the form of wav files is considered on the example of the Kazakh language. The given bases can be used for the speech synthesis by the voice of a certain announcer, as well as in the systems of active protection of speech information by forming combined signals which mask the speech. The masking signals include ‘white noise’ and speech-like signals which are formed automatically on the basis of structural speech units including possibilities of structural texts characteristics for the given language. The work studies the formation mechanism of structural speech unit bases of the Kazakh language and synthesis of speech-like signals in the Kazakh language.

7. ACKNOWLEDGEMENTS

The given work is carried out with the grant financial support from KS MES RK, № AP05130293.

REFRENCES:

[1] H.V. Davydau, V.A. Papou, A.V. Patapovich, Y.N. Seitkulov, Li Ye, Fan Yanhong, Jiang Jingsai, Bi Xiaoyan, “Method for protecting speech information”, Doklady BSUIR, No.8(94), 2015, pp. 107–110.

[2] Y. Seitkulov, S. Boranbayev, B. Yergalieva, G. Davydov, A. Patapovich, “Rationale for the method of formation of the combined speech masking signals”, in 2014 IEEE 8th International Conference on Application on Information and Communication Technologies (AICT), Astana, Kazakhstan.

[3] V.I.Vorobev, A.G.Davydov, B.M.Lobanov “Speechlike signals synthesis using allophones” in XIII Session of the Russian Acoustical SocietyMoscow, August 25-29, 203 pp.527-530.

[4] Hiroake Sakoe and Seibi Chiba “Dynamic Programming Algorithm Optimization for Spoken Word Recognition”, in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-26, no. 1, 1978, pp. 43–49.

[5] Odette Scharenborg, Vincent Wan, Mirjam Ernestus“Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries“ in The Journal of the Acoustical Society of America, vol. 127, no. 2, 2010, pp. 1084–1095.

[6] Jon Ander Gomez and Marcos Calvo “Improvements on automatic speech segmentation at the phonetic level,” in 16th Iberoamerican Congress,CIARP 2011 Progress in Pattern Recognition, Image Analysis, Computer Vision and Applikations. Springer, 2011, pp. 557–564.

[7] Donald J. Bemdt James Clifford “Using Dynamic Time Warping to FindPatterns in Time Series“ in AAAI Proc. knowledge discovery in databases, 1994, pp. 359–370.

[8] Alaa Ehab Sakran, Sherif Mahdy Abdou, Salah Eldeen Hamid, Mohsen Rashwan “A Review: Automatic Speech Segmentation“ in International Jornal of Computer Science and Mobile Computing, IJCSMC, vol. 6, no. 4, 2017, pp. 308–315.

[9] Ryszard Makowski, Robert Hossa “Automatic speech signal segmentation based on the innovation adaptive filter,” in International Journal of Applied Mathematics and Computer Science, vol. 24, no. 2, 2014, pp. 259–270.

[10] Juozas Kamarauskas “Automatic Segmetation of Phonemes using Artificial Neural Networks“ in Elektronika ir Elektrotechnika, vol. 72, no. 8, 2006, pp. 39–42.

[11] Syed Akhter Hossain, Nusrat Nahid, Nasia Nuzhat Khan Dick Carol Gomes, Sabah Mohammad Mugab “Automatic Silence/Unvoiced/Voiced Classification of Bangla Velar Phonemes: New Approach“ in 8th ICCIT, Dhaka, 2005.

[12] Andreas Stolcke, Neville Ryant, Vikramjet Mitra, Jiahong Yuan, Wen Wang, Mark Liberman “Highly accurate phonetic segmentation using boundary correction models and system fusion“ in 2014 IEEE

International Conference on Acoustic, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 5552–5556.

[13] K. Ströber, W. Hess "Additional use of phoneme duration hypotheses in automatic speech segmentation" in Spoken Language Processing (ICSLP 98): Proceeding of the 5th International Conference Sydney, Australia, 30 November – 4 December 1998. – Sydney, 1998, pp. 595-598.

[14] V. N. Sorokin "Segmentation of the period of the fundamental tone of a voice source" in Acoustical Physics, vol. 62, no. 2, 2016, pp. 244–254.

[15] V. N. Sorokin, I.S. Makarov "Gender recognition from vocal source" in Acoustical Physics, vol. 54, no. 4, 2008, pp. 571–578.

[16] V. N. Sorokin, A.A. Tananykin, V.G. Trunov "Speaker recognition using vocal source model" in Journal Pattern Recognition and Image Analysis, vol. 24, no. 1, 2014, pp. 156–173.

[17] V. N. Sorokin, A.S. Leonov, V.G. Trunov "Speaker Recognition Regardless of Context and Language on a Fixed Set of Competitors" in Journal Pattern Recognition and Image Analysis, vol. 26, no. 2, 2016, pp. 450–459.

[18] A.I. Cyplihin, V. N. Sorokin "Segmetation of Speech cardinal elements" in Russian Information Processes, vol. 6, no. 3, 2006, pp. 177–207.

[19] B.M. Lobanov, L.I. Tsirulnik "Computer Synthesis and Cloning of Speech" in Russian, Minsk, "Belarusian Science", 2008, 316 p.

[20] A.M.A. Ali, J.V. Spigel "Acoustic-phonetic features for the automatic classification of fricatives" in J. Acoust. Soc. Am., 2001, vol. 109, no. 5, pp. 2217–2235.

[21] Y.J. Wu, H. Kawai, J. Ni, R.H. Wang "Discriminative training and explicit duration modeling for HMM-based automatic segmentation" in Speech Communication, 2005, vol. 47, no. 4, pp. 397–410.

[22] Hema A Murthy and B. Yegnanarayana "Group delay functions and its applications in speech technology" in Sadhana, Indian Academy of Sciences vol. 36, part 5, 2011, pp. 745–782.

[23] Richard E. Bellman and Stuart E. Dreyfus "Applied Dynamic Programming" New Jersey: Princeton Univ. Press, 1962, p. 363.

[24] Chin-Hui Lee "Applications of dynamic programming to speech and language processing" in AT&T Technical Journal, 1989, vol. 68, no. 3, pp.114 – 130.

[25] Ozgul Salor "Dynamic programming approach to voice transformation" in Speech Communication, 2005, vol.48, no. 10, pp.510 – 513.

[26] Naoto Iwahashi, Nobuyoshi Kaiki, Yoshinori Sagisaka "Speech Segment Selection for Concatenative Synthesis Based on Spectral Distortion Minimization" in IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 1993, vol. E76-A, no. 11, pp. 1942 – 1948.

[27] H. Ney "Dynamic programming parsing for context-free grammars in continuous speech recognition" in Journal IEEE Transactions on Signal Processing, 1991, vol. 39, no. 2, pp. 336 – 340.

[28] Abdelkader Chabchoub and Adnan Cherif "High Quality Arabic Concatenative Speech Synthesis" in Signal & Image Processing : An International Journal (SIPIJ), 2011, vol. 2, no. 4, pp. 27 – 36.

[29] J. W. Tukey "Nonlinear (Nonsuperposable) Methods for Smoothing Data" in Electronics and Aerospace Systems Convention, 1974 (EASCON74): Proceedings of Conference, Washington, DC, USA, October 1974 - Congress record, 1974 p.673.

[30] Advances in speech signal processing, edited by S. Furui, M.M. Sondhi – New York: Marcel Dikker, 1991, p.872.

[31] C. Wang, S. Seneff "Robust pitch tracking for prosodic modeling in telephone speech" in Acoustics, Speech, and Signal Processing, 2000 (ICASSP 2000): Proceedings of IEEE International Conference, Istanbul, Turkey, 5 - 9 June 2000, IEEE Computer Society. 2000, vol. 3, pp.1343 – 1346.

[32] A. Sethy, S. Narayanan "Refined speech segmentation for concatenative speech synthesis" in Spoken Language Processing (ICSLP 2002 – Interspeech 2002): Proceedings of International Conference, Denver, USA, 16–20 September 2002. – pp. 149–152.