EURASIAN JOURNAL OF MATHEMATICAL AND COMPUTER APPLICATIONS
ISSN 2306–6172
Volume 5, Issue 4 (2017) 70 – 79

THE REFINED IDENTIFICATION OF BEGINNING-END OF SPEECH; THE RECOGNITION OF THE VOICELESS SOUNDS AT THE BEGINNING-END OF SPEECH. ON THE RECOGNITION OF THE EXTRA-LARGE VOCABULARIES.

Shelepov V.Ju., Nitsenko A.V.

Abstract The present paper belongs to the diphone DTW-recognition strategy developed by the authors. Voiceless plosives, as well as energetically weak hard and soft [f] constitute a problem for recognition when they occur at the beginning or end of speech, owing to their similarity to neighboring silence stretches. The article opens up a description of some refined methods for specifying the beginning and the end of a spoken word or phrase. This is the basis for the proposed methods of recognizing the mentioned sounds beginning or concluding a spoken word or phrase. We introduce a concept of the final quasifricative fragment as well as the algorithms for its selection and use to classify voiceless plosives in the final position. The results obtained together with an insignificant increase in the number of basic speech units, makes it possible to advance in solving the difficult problems of recognizing short speech segments as well as extra-large vocabularies.

Key words: continuous-speech recognition, speech segmentation, large vocabulary speech recognition, voiceless fragment, diphone, dynamic time warping (DTW).

AMS Mathematics Subject Classification: 68T10, 68T50.

1 Introduction

Various approaches to solving the problems mentioned above have been discussed within academia for quite a few years. Articles [1-7] deal with the problem of distinguishing speech in audio recordings. In some of them the recording is assumed in the environment with significant noise. Word recognition, implemented by recognition their grammatical components, is discussed in [8-15].

All the results reported in the present article are software implemented. At the same time we establish the relationship between selecting row in the list representing the numerical arrays we work with (see for example Figure 2), and the cursor position in the signal visualization window. A similar connection is established between the selection of several rows and some part of the signal.

The recording parameters are 22050 Hz as the sampling frequency, 8 bits is the quantization. When the recording finished, we perform the amplitude normalization of the signal, i.e. divide each sample by the maximum amplitude of the entire signal and multiply it by 255.

On transcription. Russian consonants are opposed as hard or soft. We shell transcribe vowels and hard consonants with the help of corresponding Russian letters. The

soft consonants are transcribed using corresponding Latin letters (with the exception of @ for soft П).

In principle, the problem of allocating sounds from speech is being solved by speech signal segmentation. We are applying our own system of a priori segmentation, which places boundaries between neighboring sounds and simultaneously classifies them in terms of broad phonetic classification: W – vowel, C – consonant, F – class of sibilant sound [с], [s], [ш], [щ], or part of an affricate [ц], [ч], P refers to voiceless plosives (see [16,17]). Sounds [ф], [f] and [x], [h] can be associated with both F- and P-segments, as well as their combinations.

The elements of the signal corresponding to the voiceless plosives at the beginning of the word are too short, hence these sounds may be segmented only in combination with subsequent voiced sound. In the middle of the word, they are characterized by a pause-like P-segment. At the end of the word this segment ought to be present but it is hard to separate it from the subsequent silence. Therefore the recognition of these sounds at the beginning and the end of the word is not an easy task. The same applies to an energetically weak sound [f] in its both hard and soft variants.

2 The refined identification of the beginning of speech

Recognition should begin with an accurate automatic detection of the starting point in speech. Our mechanism that has been used before is described in §2, chapter 1 and §7, chapter 7 of the book [17]. We record a signal with an initial and final silence margin and segment it in the above-mentioned manner (see [16]). Considering the above mentioned silence margin, segmentation starts with a pause-like segment P. The end of the initial pause has been considered a true beginning of the speech. However, in case the word begins with a hard or soft sound [f], the latter may be included in the initial silence pause and be lost for speech. In that case the word "FACT" will be recognized as "ACT", i.e., a recognition error will occur. Here we propose a different method to find the starting point of speech. The operations described below will be performed for a speech signal that is denoised using spectral substraction method (see [18,19]).

As a feature for speech analysis, we will use the number of constancy points. We call the discrete time moment as a point of constancy for the signal, if at the next moment the signal value does not change. Otherwise, we call this point as inconstancy point.

Next, at the beginning of the signal, we consider a sequence of speech frames of 256 samples and then calculate the number of constancy point C_i for each frame, where i is the frame number. The resulting array of values is called the Array-1. It is also convenient to use the Array-2, consisting of the numbers $N_i = 256 - C_i$ of inconstancy points and the Array-3 of differences $N_i - C_i$. Figure 2 reflects these arrays for the signal presented in Figure 1. The described feature is convenient in speech endpoint detection, since we need to separate speech from silence which is characterized by a large number of constancy points, so that the numbers N_i are small, and the differences $N_i - C_i$ will be negative. The differences become positive in the neighborhood of the transition from silence to speech. This property is violated if speech begins with the

hard or soft sound [f].

From the beginning of the Array-3, we search for the first positive value. We take a corresponding element of the Array-2 and, moving from it to the beginning of the array (bottom up in Figure 2), we mark the last nonzero element. So there are 2 marks. If the distance between them is not less than 7 frames, consider the first mark as the beginning of speech (the case characteristic of the initial [ф], [f]). If the specified distance is less than 7 frames, we specify the speech start point, moving from the first mark to the end in search for the first element exceeding the given threshold p_1 (for the authors of this article and their applied equipment the threshold $p_1 = 28$ is appropriate). It defines the beginning of the speech in this case.

Figure 1 gives an example of start point detection in the word utterance "факультет". Here the first P-segment corresponds to the [ф] sound. Its beginning is detected in accordance with the row selected in Figure 2.

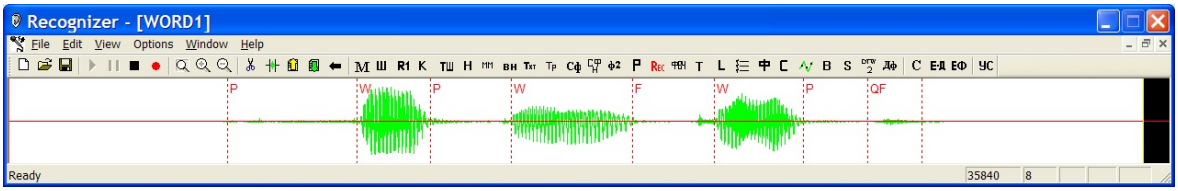

Figure 1: The waveform of the word utterance "факультет".

3 The refined identification of the end of speech

The speech signal always fades at the end slower than it grows at the beginning. Therefore, it is always advisable to determine the end with a single numerical threshold. At the same time, it is necessary to keep certain compromise: it is desirable to cut off the part associated with the aspiration at the end of the speech, ending with a vowel or voiced consonant, without losing a significant part in the case of ending with a plosive sound. In terms of the Array-2, this leads to a search for the first element that exceeds a certain threshold p_2 by moving from the end of the Array to the beginning. For the authors and their applied equipment, the threshold $p_2 = 78$ is appropriate; in terms of the Array 3 it corresponds to the -100 threshold. For example the end of the speech in Figure 1 is determined by the last line highlighted in Figure 3.

After determining the boundaries of speech, we select the entire speech segment and apply a priori segmentation to it (consult the introduction to the present article).

4 The recognition of voiceless sound at the beginning and at the end of speech

The refined identification of the beginning and the end of the speech segment described here makes it possible to recognize energetically weak sounds [ф] and [f] (soft Ф) in these positions. It also allows recognizing hard voiceless plosives[к], [п], [т] and the corresponding soft sounds [k], [@], [t] at the beginning. For this see next section.

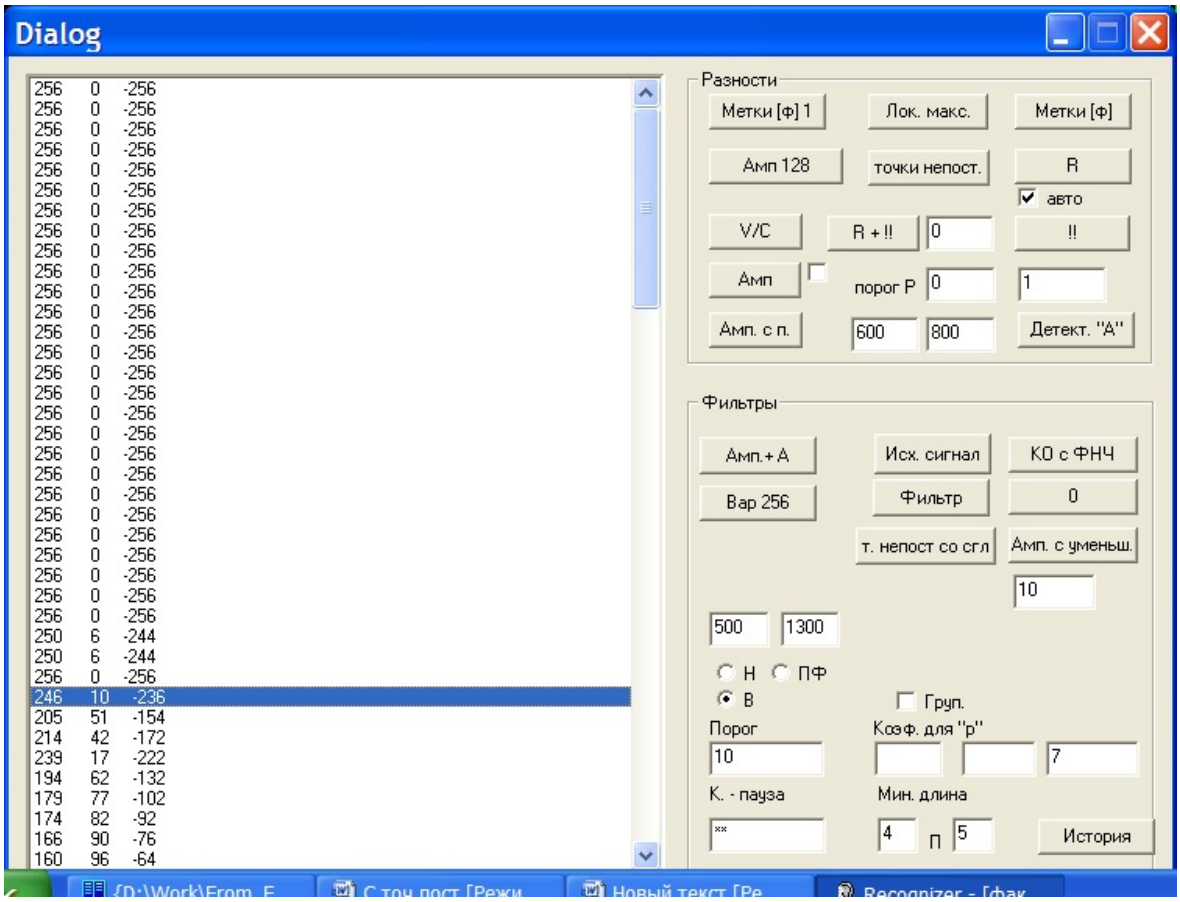

Figure 2: Beginning of the arrays calculated for speech sample in Fig. 1.

To recognize voiceless plosives at the end of a speech let's consider the case when the speech ends with a plosive sound and only P-segment corresponds to it in a priory segmentation. A more detailed analysis reveals that it includes a substantially pause-like part, corresponding to small elements of the Array-2, and the final part, which is pertinent to be called QUASI-FRICATIVE. It corresponds to sufficiently large elements of the Array-2. The search for its beginning occurs in the process of continuing the movement to the beginning of the array until an element smaller than certain threshold p_3 appears. For the authors and their applied equipment, $p_3 = 28$ (in terms of the Array-3 this is a threshold is -200). Figure 4 shows the quasi-fricative part of the final [т] in the word "факультет" (Fig. 1). It corresponds to the lines highlighted in Figure 3.

The final quasi-fricative part is denoted as QF-segment. We should note that the quasi-fricative part is frequently determined as F-segment during the a priori segmentation, in this case additional operations are not needed.

For recognition of unvoiced plosive sounds in the end of the speech we create templates for their quasi-fricative parts. According to these templates these sounds are recognized with the help of the DTW-algorithm.

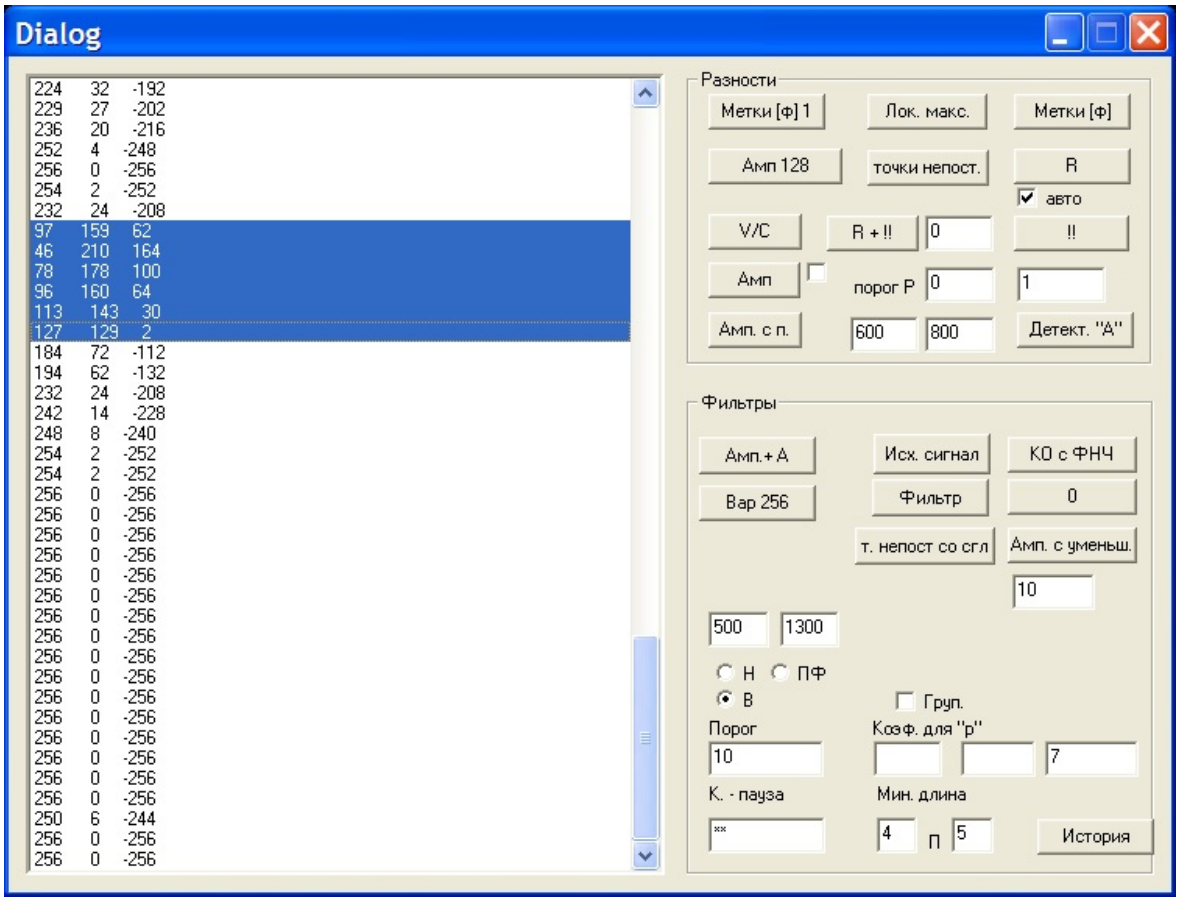

Figure 3: End of the arrays detected for the signal shown in Fig. 1.

5 The Recognition of Russian adjectival full word forms

One of the consequences of the results presented above is an opportunity to recognize the short beginnings and endings of Russian word forms, which in turn makes it possible to organize recognition of superlarge vocabularies in real time.

We will rely on a technique using voiceless fragments of the speech signal proposed in [20]. At the same time we add it as follows. We will use the generalized VF-transcription introduced in [21] (the V-transcriptional symbol for the voiced, F-transcriptional symbol for the voiceless fragment). It is written in the vertices corresponding to the ends of words when constructing a tree of full transcriptions. When recognizing the next segment of the signal, we will take into account the sequence of voiced and voiceless fragments in it, and look for the result only among words with the same alternation, that is, with the corresponding VF-transcription. It will save us a lot of irrelevant calculations.

Let us consider the set of all Russian adjectives from A.A.Zaliznyak's dictionary [22], of the amount is 20387. Their full word form vocabulary consists of 262160 units and belongs to superlarge ones. Let us call it D_0 vocabulary. It is loaded into the recognition program in the form of a tree, which allows very fast search in it.

The number of word stems is the same as the number of lemmas, i.e. an order of

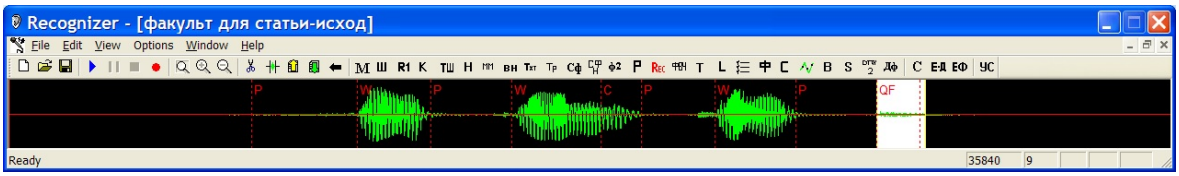

Figure 4: The quasifricative part of the final [т] is highlighted in the waveform of the word utterance "факультет".

magnitude smaller than the number of word forms. The set of all endings consists of 69 units:

а, ам, ами, ах, ая, аяся, е, его, егося, ее, еся, ей, ейся, ем, емся, ему, емуся, ею, еюся, ие, иеся, ий, ийся, им, ими, имися, имся, их, ихся, о, ов, ого, \ого, ое, \ое, ой, \ой, ом, \ом, ому, \ому, ою, \ою, у, ую, уюся, ы, ые, ый, ым, ыми, ых, ье, ьего, ьей, ьем, ьему, ьею, ьи, ьим, ьими, ьих, ька, ьки, ько, ью, ья, ью, яя,

(some of them include a sign "ь" for convenience, it will be absent in the corresponding stems).

Let us recognize the word form by methods of recognizing continuous speech, treated as the result of two "quasi-words" fusion, which are the stem and the ending.

The set of all stems is divided into vocabularies consisting of stems with the same number of voiceless fragments. For each of these dictionaries a transcription tree is created and all these trees are loaded into the program. The tree of transcriptions of endings is loaded there, too.

Note, please, that endings must be transcribed in a special way. Thus the end of the genitive case of "ero" should have a "ева" transcription. At the same time, though, general rules will give the transcription"јева". Therefore, a separate transcription file "trans_except-inflection.txt" is created for the endings.

Stems and endings are recognized independently by methods of [20] with the addition specified above. The result is two recognized "quasi-words" that should be software glued together into one word form. To combat possible errors associated with phonetic attachments, in recognizing the stem, three DTW-distance closest candidates are taken. From the end of the segment of each candidate, the recognition is carried out using the vocabulary of the endings, and three ending hypotheses appear.

Next, checking the consistency of the stems and endings follows: from the three "stem + end" sequences only those, that are truly word forms from the D_0 vocabulary, are selected. It is closed with the final DTW-recognition among these word forms. If there are no word forms of this type, all word forms with recognized 3 stems are selected from the D_0 and the final DTW-recognition is carried out among them.

The recognizing program will work fastest and efficiently if the start of the recording is performed by pressing the key for the first letter (or first few letters) before pronouncing a word form. As a voice analogue of such an operation, let us make a preliminary recognition of the initial sound of the spoken word form or class it belongs to. A version of the preliminary recognition is proposed in [23], [24]. Now we use a slightly different option.

To avoid errors associated with segmentation, we will work with the recognition segments tied to the beginning of speech. The first is a segment of 3 windows with 368 samples to the right of the initial label (segment-3), the second - a segment of 5 windows like these (segment-5).

1) Let us consider the first segment of a priori segmentation. If it is an F-segment or a P-segment, the recognition is performed on segment-3 with templates

$$\text{с, ц, ч, ш, щ, ф, f, x, h} \tag{1}$$

The result of the recognition is considered the result of recognizing the first sound of the word form.

2) If the first segment is a W-segment or a C-segment, then the recognition is performed on a segment-3 with the templates

$$\text{к, k, п, @, т, t,} \tag{2}$$

$$\text{notP.} \tag{3}$$

(It is possible that for some of them several copies constructed using averaging will be required (see [16, 17])). If the recognition result is one of the sounds (3), it is considered the initial sound of the word form. In case the recognition is determined by one of the templates (4), a transition to recognition takes place along segment-5 with templates

$$\text{a, j, ж, з, z, и, л, l, p, r, o, y, э, D+N} \tag{4}$$

If one of the sounds (5) is recognized, it is the result of recognizing the initial sound.

Class D includes sounds [б], [в], [г], [д] and their soft analogs, class N includes sounds [м], [н] and their soft analogs. D+N is the union of these classes. If one of the sounds (2), (3) or (5) is recognized, the program recognizes the transcription of the stems by the trees, described above, selecting for recognition just transcriptions beginning with the recognized first sound. This is an analogue of recognition with the indication of the first letter of the word form on the keyboard.The number of the transcriptions is not big, so sufficient reliability and speed are ensured.

A different situation is in the case of recognizing class D+ N. We will talk about it now. 6431 stems correspond to it and the recognition of the tree of stem transcriptions becomes too long and insufficiently reliable. In this situation an additional initial recognition of the first 3 transcription symbols is suggested.

We have developed a simple program that cuts the transcriptions of a given vocabulary up to 3 first characters and indicates the number of transcriptions with this beginning. Here are the first 60 members of the corresponding list for words with the first sound from D+N, ordered in decreasing order of the number of transcriptions mentioned:

bез 183, дву 160, мал 150, неп 146, bec 146, мна 100, nea 94, nec 76, гра 65, ney 64, мар 60, вад 58, бла 54, выс 53, гар 51, пер 49, бар 47, пед 46, мел 46, бра 46, ман 46, нев 45, гал 44, пез 44, над 44, vep 43, мет 42, фse 42, меж 42, бal 41, неи 37, вел 37, баг 36, нав 34, нат 34, дра 33, вас 32, бал 32, вал 31, gид 31, нар 31, нас 31, вне 30, vel 29, дал 29, пек 29, дIи 29, меш 29, ваз 28, гла 28, вну 27, гру 26, вal 25,

neb 25, mal 25, мак 24, дге 24, неб 24, дас 23, вар 23.

The number following transcription symbols is the number of stems whose transcriptions have this beginning. Summing up these numbers, it is easy to see that these 60 beginnings cover 3003 stems. The set of words with the initial sound from D+N, but other initial transcriptions within 3428, will be called the "remainder".

In this regard, in addition to the "stem" and "ending" objects, we introduce the "beginning" object. The alphabetic records of the beginnings with the most common transcriptions are introduced into the program (we will continue to talk about the 60 mentioned above). For each letter recording the corresponding words are selected from the vocabulary D_0 and, through it, the corresponding transcriptions are constructed. The tree is created from these transcriptions and transcriptions of remainder. When recognizing the stem with the first sound from D+N, just transcriptions of the stems with the recognized beginning from the number of the specified 60-s are allowed. To this the recognition of the remainder stems transcriptions are added. So, if the first sound turned out to be a sound of class D+N, no more than 3428 + 183 stems are recognized.

For any other result for the first sound, the number of stems that are further recognized is less numerous.Thus, with the initial sound [п] (P hard), we have 2343 stems , with the initial sound [a] (the words start with A or unstressed O) we have 1622 stems, and so on. The start recognition described here is a voice analogue of pre-pressing several initial characters on the keyboard.

Note, that the aforementioned first sounds of words, beginnings and endings are realized by short speech segments, the recognition of which is objectively hampered by a small number of discriminators. Their successful recognition is based on the refined identification of the beginning of the continuous speech segment, and also on the inclusion of stationary parts of sounds into the number of basic speech units. Their templates are also used when creating the templates of words or their necessary parts through transcriptions. For example, the template for the word “ЗИМА” will be constructed as follows:

зима → [zима] → z0-z-zи-и-им-м-ма-а-a2.

The transcription of the word is in square brackets. The templates of diphones will correspond to the couples ЗИ, ИМ, МА, the templates of pure sounds – to the symbols Z, I, M, A, the initial and final semi-diphones – to the symbols Z0, A2.

Note also, that perfect recognition of all word forms with a recognized stem is in a certain sense less promising than using the recognition of endings: the differences between endings are relatively larger than the differences between word forms. One has to resort to recognition of whole word forms with a recognized stem only in exceptional cases in the above algorithms.

6 Conclusions

The proposed algorithms allows high speed isolated words recognition with the extremely large size of D_0 vocabulary even on a low performance computer with a single-core 2.4 GHz processor and 1 GB of RAM.

References

[1] Denilson C. Silva, *A Robust Endpoint Detection Algorithm Based on Identification of the Noise Nature*, ITRW on Nonlinear Speech Processing (NOLISP 07) Paris, France, 2007,108 -111.

[2] R.B. Blazek, Wei-Tyng Hong, *Robust Hierarchical Linear Model Comparison for End-of-Utterance Detection under Noisy Environments* , International Symposium on Biometrics and Security Technologies (ISBAST), 2012,126-133.

[3] O.Zh. Mamyirbaev, M.N.Kalimoldaev, R.R.Musabaev, *Metodyi primeneniya VAD v sistemah raspoznavaniya kazahskoy rechi*, Problemyi informatiki: nauchn.-tehn. zhurnal IVMiMG SO RAN,No.1(18),2013, 63-68.

[4] V.A.Volchenkov, V.V.Vityazev, *Metodyi i algoritmyi detektirovaniya aktivnosti rechi*,Tsifrovaya Obrabotka Signalov, No.1, 2013, 54-60.

[5] T. R.Sahoo, S.Patra, *Silence Removal and Endpoint Detection of Speech Signal for Text Independent Speaker Identification*, International Journal of Image, Graphics and Signal Processing, Vol. 6, No. 6, 2014, 27-35.

[6] S.Graf, *Features for voice activity detection: a comparative analysis*, EURASIP Journal on Advances in Signal Processing, No. 1, 2015, 1-15.

[7] Z.H.Ji, H.Yang, R.Li, Y.Jin , *A speech endpoint detection algorithm based on short-term autocorrelation and zero-crossing*, Electronic Science and Technology, No. 9, 2016, 52-55.

[8] J.Kneissler, D.Klakow, *Speech recognition for huge vocabularies by using optimized subword units*, Proc. Eurospeech 2001, Aalborg, Denmark, 2001, 69-72..

[9] R Singh, B Raj, R.M. Stern, *Automatic generation of subword units for speech recognition systems*, Speech and Audio Processing, IEEE Transactions on, Vol. 10, No. 2, 2002, 89-99.

[10] V. Siivola, T. Hirsimäki, M. Creutz, M. Kurimo, *Unlimited Vocabulary Speech Recognition Based on Morphs Discovered in an Unsupervised Manner*, Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003), Geneva, Switzerland, September 2003, 2293-2296.

[11] S. Maucec, Z. Kacic, B. Horvat, *Modelling highly inflected languages*, Information Sciences, Vol. 166, No. 1-4, 2004, 249-269.

[12] A. Hagen, B. Pellom, *Data Driven Subword Unit Modeling for Speech Recognition and its Application to Interactive Reading Tutors*,Interspeech 2005, 2757-2760.

[13] M.Kurimo, M.Creutz, M.Varjokallio, E.Arisoy, M.Saraclar, *Unsupervised segmentation of words into morphemes - Morpho challenge 2005 application to automatic speech recognition*, Proc. Interspeech 2006, Pittsburgh, USA, 2006, 1021-1024.

[14] T. Rotovnik, M.S. Maucee, Z. Kacic, *Large vocabulary continuous speech recognition of an inflected language using stems and endings*, Speech Communication, Vol. 49, 2007, 437-452.

[15] A.L.Ronzhin, *Topologicheskie osobennosti morfofonemnogo sposoba predstavleniya slovarya dlya raspoznavaniya russkoy rechi*, Vestnik kompyuternyih i informatsionnyih tehnologiy, No. 9, 2008, 12-19.

[16] A.K.Buribaeva, G.V.Dorohina, A.V.Nicenko, V.Ju.Shelepov, *Segmentacija i difonnoe raspoznavanie rechevyh signalov*, Trudi SPIIRAN, No. 31, 2013, 20-42.

[17] V.Ju.Shelepov, A.V.Nicenko, *Segmentacija i difonnoe raspoznavanie rechi*, Donetsk, GU IPII, 2015, 231 p.

[18] S.F.Boll, *Suppression of Acoustic Noise in Speech Using Spectral Subtraction*, IEEE Transactions on Acoustic, Speech and Signal Processing, No. 27, 1979, 113-120.

[19] M.Karam, H.F.Khazaal, H.Aglan, C.Colel, *Noise Removal in Speech Processing Using Spectral Subtraction*, Journal of Signal and Information Processing, No. 5, 2014, 32-41.

[20] V.Ju.Shelepov, A.V.Nicenko, *Recognition of the continuous-speech russian phrases using their voiceless fragments*, Eurasian journal of mathematical and computer applications, Vol. 4., No. 4, 2016, 19-24.

[21] V.Ju. Shelepov, A.V.Nicenko, *O nekotoryh voprosah, svjazannyh s difonnym raspoznavaniem i raspoznavaniem slitnoj rechi* , Artifical Intelligence, No. 3, 2013, 209-216.

[22] A.A.Zaliznjak, *Grammaticheskij slovar russkogo jazyka*, Russkij jazyk, 1977, 879 p.

[23] V.Ju. Shelepov, A.V.Nicenko, *O raspoznavanii pervogo zvuka v slitnom rechevom otrezke*, Problemy iskusstvennogo intellekta, № 0(1), 2015, 116-122.

[24] V.Ju. Shelepov, A.V.Nicenko, *O raspoznavanii sverhbolshih slovarej russkih slovoform s ispolzovaniem kvaziosnov*, Izvestija Juzhnogo federalnogo universiteta, tehnicheskie nauki, №4, 2016, 82-92.

Shelepov V.Ju. ,
Institute of Artifical Intelligence,
118-b, Artyom st, 83048 Donetsk, Ukraine
Email: vladislav.shelepov2012@yandex.ua,

Nitsenko A.V. ,
Institute of Artifical Intelligence,
118-b, Artyom st, 83048 Donetsk, Ukraine
Email: nav_box@mail.ru,

Received 15.09.2017, Accepted 09.11.2017