EURASIAN JOURNAL OF MATHEMATICAL AND COMPUTER APPLICATIONS
ISSN 2306–6172 (p); ISSN 2308-9822 (o)
Volume 4 , Number 4 (2016) 54–59

RECOGNITION OF THE CONTINUOUS-SPEECH RUSSIAN PHRASES USING THEIR VOICELESS FRAGMENTS

Shelepov V.Ju.,Nicenko A.V.

Abstract The author's a priori speech segmentation is applied for recorded signal as the basis to single out its voiceless fragments with increased reliability. We propose a continuous-speech recognition method by words partial lists generation from our full vocabulary using the number of their voiceless fragments. The appropriate parts of the signal are recognized in these dictionaries by evaluation of DTW-distance to the word patterns which are synthesized from diphone patterns. Experiments were performed on vocabularies with several hundreds of nouns and adjectives of highest frequency, recorded by the Russian language frequency dictionary. In these experiments, the correct recognition was over 90%.

Key words: continuous-speech recognition, speech segmentation, voiceless fragment, partial lists, diphone, dynamic time warping (DTW)

AMS Mathematics Subject Classification: 68T10

1 Introduction

The list of recent works on continuous speech recognition is in [1-7]. When dealing with large and overlarge vocabularies, we are to use the initial recognition engine for the appropriate vocabulary of quasi-word stems described in article [8] for separately spoken words. However, in this article we describe the continuous speech recognition algorithm assuming that the work is carried out with inflected words from the start.

2 Method description

The proposed algorithm is based on the authors' segmentation method which consists of the automatic splitting the signal into sections corresponding to individual Russian speech sounds, classifying them simultaneously within the broad phonetic classification (W - vowel, C - voiced consonant, F - voiceless fricative, P - voiceless plosive sound). The segmentation tool uses a numerical analog of the total variation, calculated as a consecutive order for segments of 256 samples:

$$V = \sum_{i=0}^{254} |x_{i+1} - x_i|$$

(For the segmentation algorithms, see for example [9, 10]). Note that possible segmentation errors have little effect on the accuracy of isolated word recognition because there is a signal matching the whole-word patterns which are synthesized from diphone

patterns. Possible segmentation errors are more essential difficulties for continuous speech recognition. Note that the segmentation problems have been discussed recently in [11-18].

Further on we use the principle of minimum DTW-distance, stated in [19].

Since the speech signal under consideration is automatically divided into segments corresponding to individual sounds in the segmenting process, the boundaries between the words are to be searched for among a finite set of intersound markers.

In the course of recognizing a segment from the beginning to its first marker, and then on to the second marker, etc., a sequence of hypotheses obtains for the first spoken word. Then a list of hypothetical words indicating a DTW-distance to each of them obtains.

The Minimum principle is this: It appears that for vocabularies with absence of the inflected forms a hypothesis corresponding to the true first word (and its true corresponding segment from the beginning) has a DTW-distance, close to the minimal.

The meaning of the Principle is obvious if we recall that the DTW algorithm aims at minimizing the distance between the spoken word and the pattern of the same word. The remaining words from the resulting list actually did not sound which makes an idea of a longer distance between them and their respective patterns, look natural.

The absence of the inflected forms means the following. The vocabulary must not contain word pairs such that the transcription of one word is derived from the transcription of the other by adding symbols. Otherwise, while uttering a longer word the DTW distance to the word with shorter transcription may be smaller.

It is clear that to recognize the second word in the phrase we must apply the method described above to the portion of the signal beginning from the end of the first word up to the end of the speech segment, etc.

Obviously, the recognition will benefit greatly in the quality and speed if the properties of the recorded signal, which allow prior limiting the recognition vocabulary at each stage are employed. This property could be a number of segmentation bits, compared to the number of word transcription symbols, i.e. a check of the length. However possible segmentation errors make a weak point of this procedure. Yet errors of this type are extremely rare in allocating the voiceless fragments (i.e, voiceless plosives and fricatives) in a speech signal. Therefore, the latter fragments will be further relied on.

We shall start our discussion with the case of non-voicing voiceless sounds in the phrase, compared to the way they are pronounced in separately spoken words. For each word from the recognition vocabulary we can determine the number of voiceless fragments using automatic transcription. All words containing n voiceless fragments are to be put into file *n.txt*. So, we have files

$$0.txt, 1.txt, ..., N.txt \tag{1}$$

the corresponding vocabularies with the same names.

Let us assume that the recognition signal is not beginning with voiceless fragment. If the first word of the phrase does not contain voiceless fragments at all, it must be recognized in the vocabulary *0.txt*, and it should be done in the interval from the start of the signal up to its first voiceless fragment, because our assumption is that the latter does not belong to the first word. It should be done by increasing the detection interval

from the beginning to the first segmentation marker, then from the beginning to the second marker, and so on, until the left boundary of the first voiceless fragment is reached.

Similarly, assuming that the first word contains one voiceless fragment we are looking for it in the *1.txt* vocabulary, sequentially adding to the recognition interval, beginning the signal, the segmentation sections until the beginning of the second voiceless fragment is reached. And so on. The result list of all these recognitions is formed, indicating DTW-distances to them. From this list the row with the minimal DTW-distance is selected. This completes the first cycle of recognition. Its result is the recognition of the first word and the place where it ends. We shall call this algorithm *"algorithm C"*. Next, then, from this place the second word of the phrase is recognized and its end defined, and so on. Figures 1 and 2 illustrate the first word recognition of the phrase "syplet cherjomuha snegom".

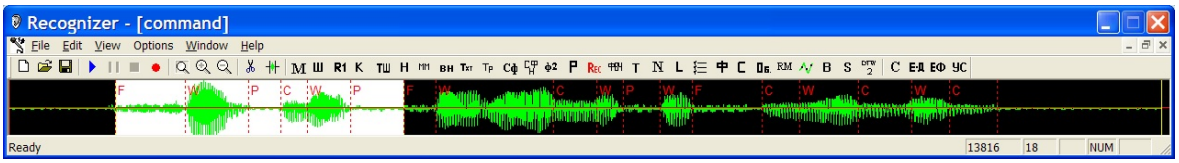

Figure 1: Recording visualization of the phrase "syplet cherjomuha snegom".

It is clear that recognition must start from vocabulary *1.txt*, if recognition signal starts with an voiceless fragment. Note that in case of the voiceless plosive acting as the first sound, it does not stand out as a separate segment. Thus, for example, the Russian word *"pogoda"* gets into *0.txt* vocabulary.

Let us introduce a qualification.

Suppose that recognition is conducted in the *N-1.txt* vocabulary. If an voiceless N-fragment has segmentation PFP, it is necessarily additionally recognize a portion of the signal from the beginning mark to the end of the first segment P in this voiceless fragment and a portion of the signal from the beginning mark to the end of the PF (in order to include the case when the first word is finished with a P-segment and the next word begins with a FP-fragment ore the first word is finished with a PF-fragment and the next word begins with a F-segment. We do similarly if voiceless N-fragment has segmentation FF, FP or PF.

It must also be noted that a word beginning with a voiceless plosive is to be searched for after the P-segment.

If a word ends with a vowel and the next one also begins with a vowel, a short pause may occur in their smooth enunciation naturally, reflected in the speech signal. Obviously, neither a pause of this type nor the appearance of excessive voiceless P-fragments will change the recognition algorithm and its result.

If a word from the *n.txt* vocabulary ends with a voiceless sound, and the next word begins with one of the sounds of the B, G, D, Z, ZH group, the voiceless sound mentioned above is voiced in continuous speech. Therefore, the vocabulary *n-1, V.txt* should be added to the vocabulary *n.txt*. It contains the same words, transcribed with the help of a modified transcriptor, replacing the voiceless sounds at the end of words with their paired voiced sounds: P - B, T - D, and so on. The quantity of voiceless

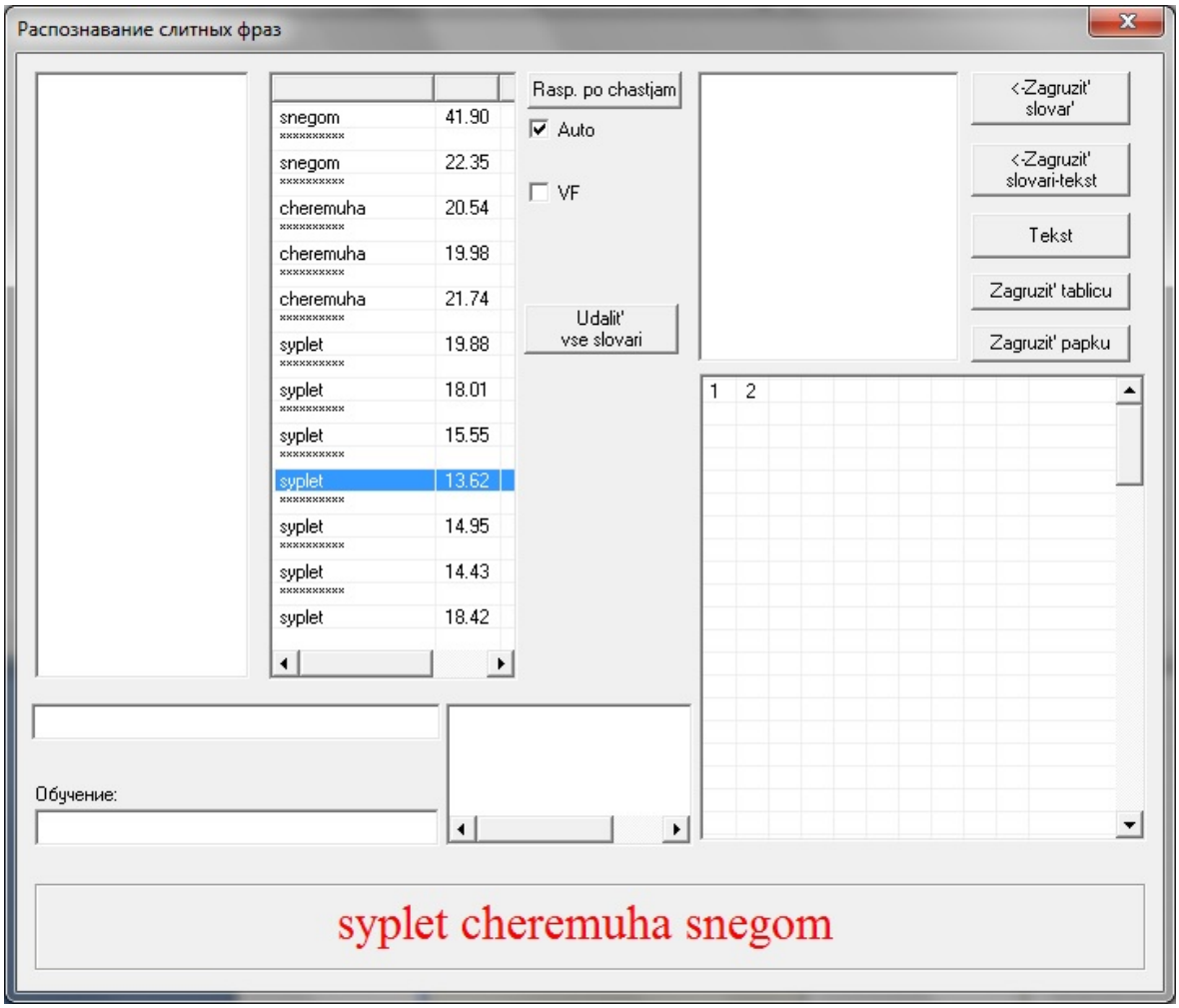

Figure 2: Dialog window with hypotheses list and recognized text.

fragments in the *n-1, V.txt* words is minus one unit compared to those of the *n.txt*. Further, each of the vocabularies *m.txt* retains the words which do not begin with B, G, D, Z, ZH, and the rest form the *mD.txt* vocabulary. The algorithm *C* must be employed on the *n-1, V.txt* after it has worked with *n-1.txt*; and after *m.txt* work with *mD.txt* accordingly. The correct result is provided by employing the principle of minimum DTW-distance. In case when the expected voicing did not happen, the corresponding recognized word will be found not in the *n-1, V.txt* vocabulary but in the *n.txt* vocabulary.

The recorded signal largely controls the selection lists for partial recognition vocabularies generated from the total vocabulary.

3 Conclusions

The proof of the reliability of the proposed method is our program (that allows typing in some phrases as texts via a special window) which automatically creates files (1), thus enabling one to recognize continuous speech, changing words and their order ar-

bitrarily, within the frame of the total composed vocabulary. As a rule, the result is a hundred percent recognition. Experimenting on larger vocabularies, we recognized ADJECTIVE - NOUN pairs up to several hundreds of nouns and adjectives of highest frequency, recorded by the Russian language frequency dictionary [20]. In these experiments, the correct recognition is over 90%.

References

[1] T.Andreas, P.Ghosh, P.Georgiou, S.Narayanan, *Robust Word Boundary Detection in Spontaneous Speech Using Acoustic and Lexical Cues*, IEEE International Conference on Acoustics, Speech, and Signal Processing,Taipei, 2009, 4785-4788.

[2] O.Zhijian, X.Ji, *A study of large vocabulary speech recognition decoding using finite-state graphs*,Chinese Spoken Language Processing (ISCSLP), 7th International Symposium, 2010, 123-128.

[3] Hong Kai Sze, Tan Tien Ping, Tang Enya Kong, Cheah Yu-N, *Linguistic stem concatenation for malay large vocabulary continuous speech recognition*,IEEE Student Conference on Research and Development (SCOReD), 3, 2010, 144-148.

[4] A.Karpov, I.Kipytkova, A.Ronzhin, *Very large vocabulary ASR for spoken Russian with syntactic and morphemic analysis*, Proceedings of INTERSPEECH' 2011, Florence, 2011, 3161вЂ“3164.

[5] D.Susman, S.Kopru, A.Yazici , *Turkish Large Vocabulary Continuous Speech Recognition by using limited audio corpus*, Signal Processing and Communications Applications Conference, 2012, 1-4.

[6] G.Saon, Jen-Tzung Chien, *Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances*, Signal Processing Magazine, Vol.29, No. 6, 2012, 18-33.

[7] J.Stas, D.Hladek, J.Juhar, D.Zlacky, *Analysis of morph-based language modeling and speech recognition in Slovak*, Information and communication technologies and services, Vol.10, No. 4, 2012, 291-296.

[8] V.Ju Shelepov, A.V.Nicenko, *O raspoznavanii sverhbol'shih slovarej russkih slovoform s ispol'zovaniem kvaziosnov*, Izvestija Juzhnogo federal'nogo universiteta,tehnicheskie nauki, No. 4, 2016, 82-92.

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

[10] V.Ju Shelepov, A.V.Nicenko, *Segmentacija i difonnoe raspoznavanie rechi*, Donetsk: GUIPII, 2015, 231 p.

[11] I.Mporas, T.Ganchev, N.Fakotakis, *Speech segmentation using regression fusion of boundary prevocabuls*, Computer Speech and Language,Vol. 24, No. 2, 2010, 273-288.

[12] J.A.GГimez, M.Calvo, *Improvements on Automatic Speech Segmentation at the Phonetic Level*, Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications,Vol. 7042, 2011, 557-564.

[13] V.A.Natarajan, S.Jothilakshmi, *Segmentation of Continuous Speech into Consonant and Vowel Units using Formant Frequencies*, International Journal of Computer Applications,Vol. 56, No. 15, 2012, 24-27.

[14] J.Yuan, N.Ryant, *Automatic phonetic segmentation using boundary models*, Proceedings of Interspeech 2013, 2013, 2306-2310.

[15] Z.Patc, P.Mizera, P.Pollak, *Phonetic Segmentation Using KALDI and Reduced Pronunciation Detection in Causal Czech Speech*, Text, Speech, and Dialogue, 2015, Vol. 9302, 433-441.

[16] M.Kalamani, Dr.S.Valarmathy, S.Anitha, *Hybrid Speech Segmentation Algorithm for Continuous Speech Recognition*, International Journal on Applications of Information and Communication Engineering, 2015, Vol. 1, Iss. 1, 39-46.

[17] H.Kamper, A.Jansen, S.Goldwater, *Unsupervised Word Segmentation and Lexicon Discovery Using Acoustic Word Embeddings*, Transactions on Audio, Speech, and Language Processing, 2016, No. 24, 669-679.

[18] V.Ju Shelepov, A.V.Nicenko, *Segmentacija rechevogo signala na osnove predpolozhenija o ego foneticheskom sostave*, Problemy iskusstvennogo intellekta, 2016, No. 1, 73-81.

[19] V.Ju Shelepov, A.V.Nicenko, *K probleme raspoznavanija slitnoj rechi*, Iskusstvennyj intellekt, 2012, No. 4, 272-281.

[20] O.N.Ljashevskaja, S.A.Sharov, *Chastotnyj slovar' sovremennogo russkogo jazyka (na materialah Nacional'nogo korpusa russkogo jazyka)*, Moskva: Azbukovnik, 2009, 1112 p.

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

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

Received 12.11.2016, Accepted 30.11.2016