УДК 004.934

V. Ju. Shelepov, A.V. Nicenko
Public institution «Institute of Problems of Artificial intelligence», Donetsk
283048, Donetsk, Artema st., 118 b

RECOGNITION OF RUSSIAN CONTINUOUS PHRASES WITH SOME SPECIAL VOCABULARIES

В. Ю. Шелепов, А. В. Ниценко
Государственное учреждение «Институт проблем искусственного интеллекта», г. Донецк
283048, г. Донецк, ул. Артема, 118 б

РАСПОЗНАВАНИЕ РУССКИХ СЛИТНО ПРОИЗНОСИМЫХ ФРАЗ С НЕКОТОРЫМИ СПЕЦИАЛЬНЫМИ СЛОВАРЯМИ

В. Ю. Шелєпов, А. В. Ниценко
Державна установа «Інститут проблем штучного інтелекту», м. Донецьк
83048, м. Донецьк, вул. Артема, 118 б

РОЗПІЗНАННЯ РОСІЙСЬКИХ РАЗОМ ВИМОВЛЕНИХ ФРАЗ З ДЕЯКИМИ СПЕЦІАЛЬНИМИ СЛОВНИКАМИ

We suggest the recognition of two kinds of Russian phrases: continuously speaking complex cardinal numerals and continuously speaking Russian names with patronymics. The recognition of numerals is based on the fact that each numeral can be divided into two parts. The first part is any numeral from 1 to 999, the second part is a numeral from the same set preceded by an appropriate Russian form of the word “thousand” (“тысяча”). Either part can be empty. We select every part by spotting the fragment “тыся” of the word ”“тысяча” (“thousand”) as the key word in speech stream and we recognize these parts separately without breaking the sound sequence into separate words. To recognize the continuously pronounced name with patronymic we use generic transcription, which describes the alternation of voiced and unvoiced fragments selected promptly after recording. All names with patronymics are divided into vocabularies with identical generic transcriptions. When the generic transcription of the recorded phrase has been found we recognize the signal without dividing it into separate words within the appropriate vocabulary.
Key words: continuously speaking complex numbers and Russian names with patronymics, key word (fragment), recognition of the entire signal, generic transcription.

Мы будем говорить о распознавании русских фраз двух видов: слитно произносимое сложное количественное числительное и слитно произносимое русское имя и отчество. Распознавание числительных основано на том, что каждое из них может быть разбито на две части. Первая – числительное от 1 до 999, вторая – числительное из этого же ряда, которому предшествует словоформа слова «тысяча» (одна из частей может быть пустой). Каждая из этих частей выделяется путем нахождения звукосочетания «тыся» как ключевого фрагмента в потоке речи и распознается отдельно и без разбиения на участки, соответствующие составляющим словам. Распознавание слитно произнесенного имени и отчества проводим, используя обобщенную транскрипцию, отражающую чередование звонких и глухих фрагментов, которые быстро и надежно выделяются непосредственно после записи сказанной фразы. Множество всех имен и отчеств программно разбивается на словари с одинаковой обобщенной транскрипцией упомянутого вида. Определив обобщенную транскрипцию сказанного, мы в пределах соответствующего словаря ведем распознавание сигнала также в качестве сплошного речевого отрезка.
Ключевые слова: слитно произносимые сложные числительные и русские имена-отчества, ключевое слово (фрагмент), распознавание сигнала целиком, обобщенная транскрипция.

The recognition of numerals is considered in the research papers [1-5]. The research papers [6-8] should be mentioned as resent ones devoted to the key words search in the continuous speech stream.

1 Recognition of numerals (from 1 to 999999)

Let us programmatically create a digital list of numerals from 1 to 999 and substitute them for their verbal expressions represented by phrases of one, two or three words. First we shall work with the vocabulary consisting of these phrases plus phrases ending with Russian equivalent of the word “one” (“один”), where it is replaced by its Russian feminine gender form (“одна”). We also include a fragment “тыся”of Russian equivalent of the phrase (word) “thousand” (“тысяча”). We call this vocabulary “Vocabulary 0”.

We apply an 8-bit recording with 22050 Hz frequency. The recognition is based on the system of indicators using relative frequencies of whole oscillation lengths (consult [9]). We also use the authors’ a priori segmentation, which is an automatic division of a speech signal into parts representing separate sounds. It takes place directly after sound recording with further classification of sounds as W (vowel), C (voiced consonant), F (voiceless fricative), P (voiceless plosive). As the basic speech unit we use the diphone, which is a symmetric segment (368×6 samples) around the mark between sounds (consult [9]). Every phrase of the Vocabulary 0 has an automatically generated transcription and a template constructed by splicing corresponding templates of the diphone base. The phrase templates form the tree. Each pronounced phrase is recognized with the help of DTW-algorithm as an entire wave file (without separating words). Such recognition provides a reliable quick result even for outdated computers with the following parameters: single-core processor with 2.4 GHz clock frequency and 1 GB RAM.

The rest numerals consist of two parts: an already considered numeral from 1 to 999 and the same numeral preceded by an appropriate form of the word “thousand” (“тысяча”). Either part can be empty. Meanwhile, the second part can be represented only by Russian wordform of “thousand” (“тысяча”). Hereafter we shall refer to the word “part” only in these specific terms. If Russian equivalent of the word “one” (“один”) precedes the word “thousand” (“тысяча”), it is replaced by its feminine gender form “одна”; by analogy the word “два” (“two”) is replaced by “две” if it precedes the word “тысячи” (“thousands”).

Either part is selected and recognized independently, if the separating (initial, final) Russian wordform of the word “thousand” (“тысяча”) is a priori recognized in the pronounced phrase as the key word of the speech stream. The latter can be achieved by the above-mentioned a priori segmentation. Namely, we search for the segments sequence WFW corresponding to the sound combination “тыся” in the initial part of the phrase, or the sequence PWFW in any other part of the phrase. In both cases the segment WFW is recognized through the Vocabulary 0. If there are several such segments, we choose the one with minimum DTW-distance to the sound combination “тыся”.

Let us consider the operations with the mentioned parts in details.

The algorithm of recognition, for the case when only the first part of the phrase is not empty, is described at the beginning of this part of the paper. We work with the Vocabulary 0.

Let the both parts are not empty. If the first part ends with Russian feminine gender form equivalents of the words “one”, “two” (“одна”, “две’) or Russian equivalents of the words “three”, “four” (“три’, “четыре’), then it should be followed by one of two

Russian wordforms of “thousand” (“тысяча” or “тысячи”). In the rest cases we have the key wordform “тысяч”. In this connection we shall use two more vocabularies. The first vocabulary begins with Russian equivalent wordform of the word “thousand” (“тысяча”) and consists of numerals from 1 to 999 preceded by this wordform. Let us call it “Vocabulary 1”. The second vocabulary begins with Russian equivalent wordform of the word “thousand” (“тысяч”) and consists of numerals from 1 to 999 preceded by this wordform. Let us call it “Vocabulary 2”. The recognition starts with the first part, from the beginning of the phrase to the beginning of the P-segment preceding the key fragment WFW. Depending on the result, we recognize the second part from the beginning of the key fragment to the end of the phrase through the Vocabulary 1 or the Vocabulary 2.

If only the second part is not empty, the recognition is realized through the Vocabulary 1. Separately pronounced numeral “thousand” (“тысяча”) can be recognized on the basis of the signal starting with the key fragment WFW followed by just two or three segments.

We should notice that the program sometimes can recognize an extraneous sequence as the key fragment, for example like in the case of Russian equivalent of the numeral “ninety seven” (“девяносто семь”) segmented as CWCWCWF-PWFW-C. To avoid the error we should carry out the recognition in assumption of presence and absence of the key fragment and compare the results (in the second case the whole signal should be interpreted as the first part). The final result is determined by the minimum value of DTW-distance.

2 Recognition of Russian names with patronymics

We consider 150 Russian male names and as many corresponding patronymics, as well as 88 Russian female names. The total number of names is 238. The total number of names with patronymics is 35700. They can be also recognized as a whole without separating words. However, their number is already rather large, in this connection such method provides insufficient speed of recognition. To increase its speed (as well as its reliability) we use an automatic division of phrasal vocabulary into parts with the help of VF-transcription. This transcription represents the alternation of symbols V and F, where V is a (maximum) segment of neighboring voiced sounds (vowels and voiced consonants), and F is the similar segment of unvoiced sounds (fricatives and plosives). We have already used such generic transcription to accelerate the recognition for vocabulary of separately-pronounced words. This transcription can be automatically derived from the whole transcription and placed in the Transcription Tree, in its terminal vertexes corresponding to the vocabulary words. For every phrase of the vocabulary “Names with Patronymics” we have a priori transcriptions and VF-transcriptions. Then there are automatically generating partial lists of phrases with identical VF-transcription and names of corresponding files such as VmV.txt, VnF.txt, FkV.txt, FlF.txt. Initial and final name symbols V and F show the beginning and the ending of VF-transcription; m, n, k, l are the number of symbols F within the VF-transcription.

The VF-transcription is determined while recording the phrase and its a priori segmentation. Further we carry out the recognition through the corresponding partial list. The sizes of these lists, except the file V3F.txt, are appropriate for recognition of any phrase as a whole with the help of diphone recognition. The list V3F contains 6923 phrases and this quantity induces the slow speed of recognition. To increase the speed we use the initial sound identifying algorithms [10].

Список литературы

1. Speaker independent recognition of spontaneously spoken connected digits [Text] / Ramesh P., Wilpon J. G., McGee M. A., Roe D. B., Lee C. H., Rabiner L. R. // Speech Communication. – 1992. – Vol. 11, Iss. 2–3. – P. 229–235.
2. Gandhi M. B. Natural number recognition using MCE trained inter-word context dependent acoustic models [Text] / M. B. Gandhi, J. Jacob // Proceedings of the IEEE International Symposium «Acoustics, Speech and Signal Processing». – 1998. – Vol.1. – P. 457–460.
3. Imperl B. Multilingual connected digits and natural numbers recognition in the telephone speech dialog systems [Text] / B. Imperl // Proceedings of the IEEE International Symposium «Industrial Electronics» (ISIE '99). – 1999. – Vol.1. – P. 188–192.
4. Robust numeric recognition in spoken language dialogue [Text] / Rahim M., Riccardi G., Saul L., Wright J., Buntschuh B., Gorin A. // Speech Communication. – 2001. – №34. – P.195–212.
5. Santosh V. Chapaneri. Spoken Digits Recognition using Weighted MFCC and Improved Features for Dynamic Time Warping [Text] / V. Santosh // International Journal of Computer Applications. – 2012. – Vol. 40, № 3. – P.6–12.
6. Barakat M. S. Keyword spotting based on the analysis of template matching distances [Text] / M. S. Barakat, C. H. Ritz, D. A. Stirling // 5th International Conference on Signal Processing and Communication Systems (ICSPCS). – 2011. – P. 1–6.
7. Guo H. An algorithm for spoken keyword spotting via subsequence DTW [Text] / Guo H., Huang D., Zhao X. // International Conference on Network Infrastructure and Digital Content. – 2012. – P. 573-576.
8. Tetariy E. Cross-language phoneme mapping for phonetic search keyword spotting in continuous speech of under-resourced languages [Text] / E. Tetariy, Y. Bar-Yosef, V. Silber-Varod // Artificial Intelligence Research. – 2015. – Vol. 4, №. 2. – P.72–82.
9. Сегментация и дифонное распознавание речевых сигналов [Текст] / Бурибаева А. К., Дорохина Г. В., Ниценко А. В., Шелепов В. Ю. // Тр. СПИИРАН. – Т. 31. – 2013. – С. 20–42.
10. Шелепов В. Ю. О распознавании первого звука в слитном речевом отрезке [Текст] / Шелепов В. Ю., Ниценко А. В. // Проблемы искусственного интеллекта. – 2015. – № 0(1). – С. 116–122.

References

1. Ramesh P., Wilpon J.G., McGee M.A., Roe D.B., Lee C.H., Rabiner L.R. Speaker independent recognition of spontaneously spoken connected digits. *Speech Communication*, 1992, vol. 11, issue 2-3, pp. 229-235.
2. Gandhi M.B., Jacob J. Natural number recognition using MCE trained inter-word context dependent acoustic models. *Proceedings of the IEEE International Symposium «Acoustics, Speech and Signal Processing»*, 1998, Vol.1, pp. 457- 460.
3. Imperl B. Multilingual connected digits and natural numbers recognition in the telephone speech dialog systems. *Proceedings of the IEEE International Symposium «Industrial Electronics» (ISIE '99)*, 1999, vol.1, pp. 188 – 192.
4. Rahim M., Riccardi G., Saul L., Wright J., Buntschuh B., Gorin A. Robust numeric recognition in spoken language dialogue. *Speech Communication*, 2001, no.34, pp.195-212.
5. Santosh V. Chapaneri. Spoken Digits Recognition using Weighted MFCC and Improved Features for Dynamic Time Warping. *International Journal of Computer Applications*, 2012, vol. 40, no.3, pp.6-12.
6. Barakat M.S., Ritz C.H., Stirling D.A. Keyword spotting based on the analysis of template matching distances. *5th International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2011, pp. 1-6.
7. Guo H., Huang D., Zhao X. An algorithm for spoken keyword spotting via subsequence DTW. *International Conference on Network Infrastructure and Digital Content*, 2012, pp. 573-576.
8. Tetariy E., Bar-Yosef Y., Silber-Varod V. Cross-language phoneme mapping for phonetic search keyword spotting in continuous speech of under-resourced languages. *Artificial Intelligence Research*, 2015, vol. 4, no.2, pp. 72-82.
9. Buribaeva A.K., Dorohina G.V., Nicenko A.V., Shelepov V.Ju. Segmentacia i difonnoe raspoznavanie rechevyh signalov [Segmentation and Diphone Recognition of Speech Signals]. SPIIRAS Proceedings, 2013, vol. 31, pp. 20–42.
10. Shelepov V.Ju., Nicenko A.V. O raspoznavanii pervogo zvuka v slitnom rechevom otrezke [On Recognition of the First Sound in Continuous Speech Fragment]. *Problems of Artificial Intelligence*, 2015, no. 0(1), pp. 116 – 122.

RESUME

V. Ju. Shelepov, A. V. Nicenko

Recognition of Russian continuous phrases with some special vocabularies

Background: today in the world there is a huge army of operators working on computers at various cash desks or in banks and offices. They are engaged in drafting documents, often by transferring information from other sources to necessary "forms" using the keyboard. The urgent task is to simplify this hard and substantially mechanical work.

Materials and methods: A significant part of the above information is made up of names, patronymics, and various multivalued numbers. The present paper is devoted to the problem of voice input of such data in the form of continuous speech. The authors' methods for diphone recognition of Russian speech are used.

Results: methods for recognizing co-pronounced Russian names and patronymics, as well as complex cardinal numerals as whole sound signals, without recognizing separate constituent words, are proposed.

Conclusion: the proposed methods allow quickly and reliably recognizing tens of thousands of co-pronounced names and patronymics, as well as a million of complex cardinal numerals. Thus, the number of recognizable phrases significantly exceeds one million. In this case, the found methods are based on the recognition of many times smaller lists and without recognition of separate constituent words. This enables the real-time recognition.

Статья поступила в редакцию 14.04.2017.