TOWARDS AUTOMATIC SPEECH RECOGNITION FOR THE TATAR LANGUAGE

A.F. Khusainov
Institute of Applied Semiotics, Tatarstan Academy of Sciences
Kazan (Volga region) Federal University, Kazan, Russia

khusainov.aidar@gmail.com

Dz. Sh. Suleymanov
Institute of Applied Semiotics, Tatarstan Academy of Sciences
Kazan (Volga region) Federal University, Kazan, Russia

dvdt.slt@gmail.com

ABSTRACT

In this paper we describe an approach to create automatic speech recognition systems for the Tatar language. We developed speech analysis platform to work with under-resourced languages and used this tool to create baseline speech recognition system. Additionally, some changes have been made to this language-independent system to take into account specific Tatar morphological structure. The resulting adapted system showed 75% accuracy on testing audio records.

1 Introduction

Using speech as a tool for manipulating electronic devices is becoming more and more common. This fact can be proved by lots of desktop and web-based services that provide functionality of automatic dictation, voice search, etc. Nevertheless, while these kinds of systems successfully work for main world languages such as English, French, Spanish, there are many languages for which speech analysis systems are not so developed.

According to Ethnologue project's statistics, more than 7100 languages are spoken in the world [1]. The significant part of these languages suffers from absence of speech services on their native languages, therefore people have to learn and use other languages in order to communicate with modern information technologies.

In this paper, we aimed to develop a platform that can be used for building baseline language-independent speech analysis systems and to use this platform to create specific speech recognition system for the Tatar language.

The structure of the rest of this paper is as follows: in Section 2 we give overview of proposed platform, including the description of its features and language-independent tools. In Section 3 we describe the aspects of using proposed platform to build the Tatar speech recognition system. Finally, Section 4 deals with experimental results achieved for continuous speech recognition task.

2 The architecture of the platform

Speech analysis systems differ by their final goal (speech recognition, speaker identification, etc.), by the language they built for and especially by the conditions under which they work properly and can be successfully used. Nevertheless, most speech analysis systems use several common blocks and similar tools. According to that fact proposed platform are built consisting of two main elements: modules (which allow re-using standard parts of algorithms) and projects (which consist of modules and focused on solving specific analysis problem).

Each module deals with some subtask and can be repeatedly used without code duplicating. In order to provide possibility to enhance quality of model's work without losing any information about its settings and relations with other modules platform provides simple version control system. In addition, it can be used to compare different realizations of some algorithm by running it two times choosing different versions of module.

Speech analysis systems not only use several common subsystems like feature calculating, but also use information from other speech analysis systems. For instance, continuous speech recognition system can use information from speaker identification system in order to increase effectiveness of its work. To implement this possibility into platform each module has list of input and output parameters. Parameter's value can be equal to a simple value or can be a reference to other module's parameter.

In addition to universal mechanism of version control and possibility to exchange information between modules platform provides several tools to ease and automate common steps of speech analysis system's creation:

1. "Acoustic features" – allows user to define phoneme and character alphabets of language and to formulate main grapheme-to-phoneme rules.
2. "Text analysis" – provides functionality of automatic phoneme transcribing (based on rules constructed in "Acoustic features" tool) and statistical analysis of the result transcription (2- and 3-gramm calculation, plotting histogram, etc.). Allows constructing text corpus with associated transcription file.
3. "Recording" – automate basic operations of constructing speech corpus, contains special visual form for saving information about speakers (age, gender, mother tongue, dialects, noise conditions), helps with creating distribution of sentences between speakers and recording corpus based on this distribution.
4. "Acoustic models" – this module allows to create acoustic models based on Gaussian mixtures models.
5. "Grammar" – automate process of creating named group of words and allows to create file which contains grammar rules for specified recognition task.
6. "Speech recognition" – execute decoding procedures according to acoustic models and given task grammar.

Developed modules are language-independent, so they can be easily configured to work with specific language. Together these modules form the skeleton of the baseline speech recognition system for any language, Fig. 1. As can be seen in Fig. 1 first five modules do the initial work of building language, pronunciation, acoustic models. These models are used by "Speech recognition" module in order to analyze input speech utterance and calculate recognition accuracy.

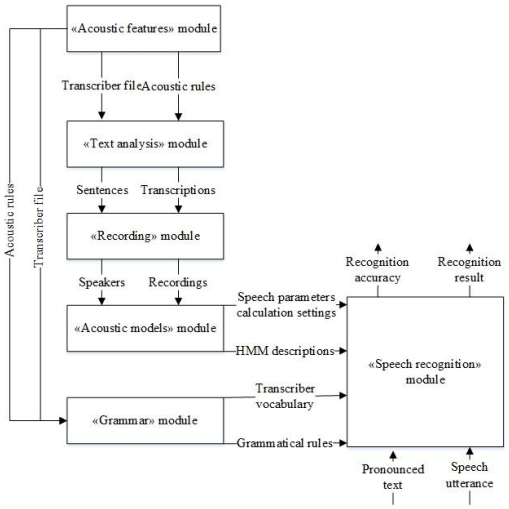

Fig. 1. Baseline speech recognition system structure

3 Continuous speech recognition system for the Tatar language

The Tatar language can be referred to under-resourced languages due to the low-level of developed information technologies and absence of well-designed text and speech corpora. At the same time, there are more than 8 million Tatar-speaking people in the world. Therefore, there is a great demand for speech technologies adapted to work with Tatar language.

To satisfy this demand and to show the potential of using the proposed platform we developed two speech recognition system for the Tatar language.

The first application is baseline speech recognizer built based on proposed analysis tools (for example, grapheme-to-phoneme conversion tool, acoustic modeling and training/decoding tools) that are encapsulated by 6 modules. Each module has been properly set up and used to create initial data for the Tatar language. These data used to build necessary acoustic, pronunciation and language models.

The second application is adapted recognition system that takes into account specific morphological features of the Tatar language. Changes have been primarily made to pronunciation and language models, details presented in Section 3.5.

3.1 Acoustic features of the Tatar Language

Obviously, acoustic features of specific language are the basic information for all types of recognition systems. These features can be described as consisting of character and phoneme alphabets and rules of conversion from character to phoneme representations. This information will be used by the next steps of analysis. The main result of this stage is automatic phoneme transcribing tool.

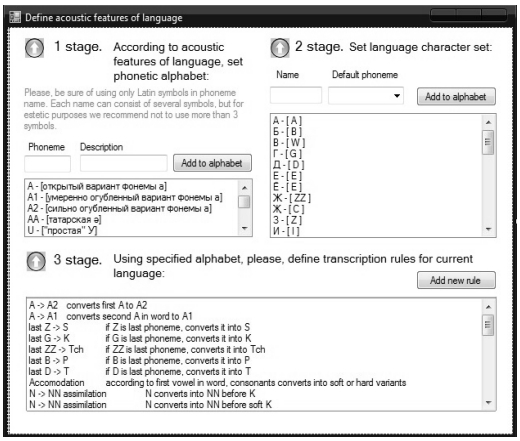

Fig. 1. "Acoustic features" module

"Acoustic features" module used to automate and provide necessary visual forms and formal kind of acoustic rules representation, Fig. 2. Module consists of three main parts, each specialized to work with character alphabet, phoneme alphabet and acoustic rules respectively. As a result, for the Tatar language we have used 39 characters alphabet (Russian alphabet plus 6 specific Tatar characters: Ә-ә, Ө-ө, Ү-ү, Җ-ж, Ң-ң, Һ-h), 56 phonemes (43 consonants and 13 vowels) and 37 rules of grapheme-to-phoneme conversion [2].

3.2 Text corpus and language model

In order to build phonetically rich and balanced speech corpus, we have to create text corpus with similar features. Therefore, we used automatic phoneme transcription subsystem and statistical analysis of resulting transcriptions in "Text analysis" module, which is shown in Fig. 3.

Based on the mentioned tools we have created text corpus, which consists of separate parts differentiating by text source types: news, literature, separate words, spontaneous spoken sentences. Total amount of sentences is 776,

number of words – 6913; all chosen phonemes are presented in sentences' transcriptions.

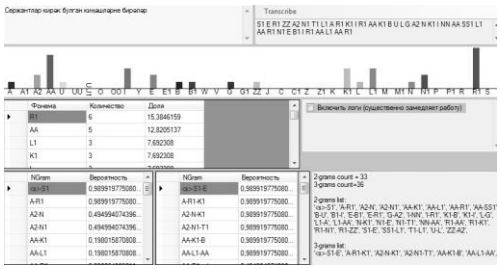

Fig. 3 "Text analysis" module

Based on collected text data language model can be constructed. We apply 3-gram language model into our speech recognition system. This model based on assumption that probability of each word depends only on previous two words, so this probability can be approximately estimated via statistical analysis of huge sequence of words.

Estimated probabilities will be used at the decoding stage to help recognition system to predict right sequence of words.

3.3 Speech corpus and acoustic models

Building multi-speaker speech corpus for the Tatar language is currently in progress. Currently it contains voices of 251 speakers, average age – 18.2. Each of speakers read the set of 36 sentences from the text corpus: 13 sentences from literature part, 7 – from news part, 15 – from words part, and 1 – from spontaneous part. At the same time, each sentence from literature has been read by 20 different speakers, from news and words – by 10 speakers, from spontaneous part – by one speaker. The total number of sentences in corpus is equal to 8638. The result features of currently available speech corpus are shown in Table 1.

Table 1. Features of multi-speaker speech corpora for the Tatar language

Parameter	Value
Number of files	8638
Total duration	8:14:24
Number of files in training subcorpus	8125
Duration of training subcorpus	7:48:12
Number of files in testing subcorpus	513
Duration of testing subcorpus	0:26:12

Corpus contains additional information about speakers (gender, age, mother tongue) and expert's score of speakers' proficiency in Tatar.

Automatic phoneme alignment approach realization has been built in "Acoustic module". This module allows to create acoustic models using two different types of input data: speech records from corpus and corresponding texts. Based on this data 57 acoustic models (56 – for phonemes, 1 – for silence model) were trained by "Acoustic module" using the HTK toolkit [3]. Models are 3-state left-right Gaussian mixture models. Number of Gaussians in mixtures varied from 1 to 170, the best phoneme recognition accuracy was showed by the models with 31 Gaussians in each mixture.

3.4 Pronunciation model

For evaluating the quality of the developed system, we used task grammar that allow speakers to pronounce every possible word sequence. Vocabulary for this task is medium-size (1135 words) and consists of words that occur in the test subcorpus, so, we have simulated rather compact task domain.

The last step in preparing data for the decoding stage is creating pronunciation model. This kind of model is a bridge between phonemes and words level of the recognition system. Each word has to be represented by sequence

of appropriate phonemes, this will make possible to solve the inverse task of decoding words from sequence of phonemes. Phoneme transcription of all words have been defined using developed grapheme-to-phoneme tool.

3.5 Adapted speech recognition system

The second application differs with the approach used to build language and pronunciation models. The idea is practically the same: we have to estimate statistics of 3-grams and to build phoneme transcriptions for all elements. The difference is that these elements are not whole words but sub-words units. The Tatar language is agglutinative language (words are constructed by concatenating of several morphemes) with rich morphology. Using sub-words units is profitably for this kind of languages, because it helps to reduce the number of units in vocabulary, but at the same time to widen the amount of covered words [4].

This approach called particle-based and requires implementing additional morpheme level into recognition system. Considering this fact, the process of building adapted language model is as follows:

- All words in existing text corpus are divided into morphemes.
- Last morphemes of each word are provided with additional '#' sign that means 'the end of the word'.
- Statistical 3-gram model are built for morphemes and '#' sign.

The pronunciation model also needs to be changed, because not words but morphemes have to be constructed from phonemes. This leads to the multiple transcription model, because some morphemes can be pronounced differently depending on context in concrete word.

4 Experimental results

We used the test part of the speech corpus for the purpose of continuous speech recognition experiments. Overall, the speech recognition systems have shown good accuracy rates near 70 percent.

As can be seen in Table 3, the adapted system outperformed baseline in both correctness and accuracy coefficients; that proves the fact that adding morphological level helps to build models and execute recognition in more accurate manner.

Table 2. Continuous speech recognition results

Parameter	Baseline system	Adapted system
Correctness	77%	83%
Accuracy	67%	75%
Number of words in all sentences	3368	3368
Substitution errors	735	533
Deletion errors	50	39
Insertion errors	316	269

4 References

[1] Lewis, M. Paul, Gary F. Simons, Charles D. Fennig (eds.). "Ethnologue: Languages of the World", Dallas, Texas: SIL International, 2013.

[2] Khusainov A.F. "Automatic phoneme recognition system for the Tatar language". In: The 1st International Conference "TurkLang", Astana, 2013, pp 211–217.

[3] Young S., Kershaw D., Odell J., Ollason D., Valtchev V., Woodland Ph. The HTK Book [Electronic resource]. URL: http://nesl.ee.ucla.edu/projects/ibadge/docs/ASR/htk/htkbook.pdf.

[4] Kurimo M, Puurula A., Arisoy E., Alumae T., Saraclar M.. "Unlimited vocabulary speech recognition for agglutinative languages". In: HLT-NAACL, NY, USA, 2006, pp 487–494.