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

D. 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 the creation of automatic speech recognition systems for the Tatar language. We developed a speech analysis platform to work with under-resourced languages and used this tool to create a baseline speech recognition system. Additionally, some changes have been made to this language-independent system taking into account the 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 a specific speech recognition system for the Tatar language.

The structure of the rest of this paper is as follows: in Section 2 we give an overview of the proposed platform, including the description of its features and language-independent tools. In Section 3 we describe the aspects of using the proposed platform to build the Tatar speech recognition system. Finally, Section 4 deals with experimental results achieved for the 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 are 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 this fact, the proposed platform consists of two main elements: modules (which allow re-using

Gönderme ve kabul tarihi: 3.09.2014-25.10.2014

standard parts of algorithms) and projects (which consist of modules and are focused on solving a specific analysis problem).

Each module deals with some subtask and can be repeatedly used without code duplicating. In order to provide the possibility of enhancing the quality of model's work without losing any information about its settings and relations with other modules, the platform provides a 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, a continuous speech recognition system can use information from the speaker identification system in order to increase the effectiveness of its work. To implement this possibility into the platform, each module has a 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 the universal mechanism of version control and the possibility of exchanging the information between modules, the platform provides several tools to facilitate and automate the common steps of speech analysis system's creation:

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

The developed modules are language-independent, so they can be easily configured to work with a 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, the first five modules do the initial work of building the language, pronunciation and acoustic models. These models are used by the "Speech recognition" module in order to analyze the input speech utterance and calculate the recognition accuracy.

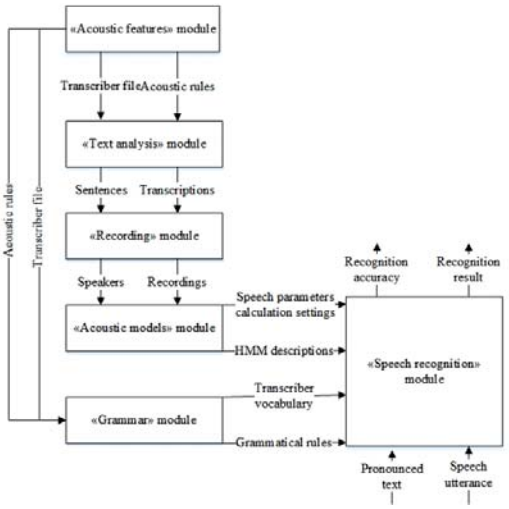

Fig. 1. Baseline speech recognition system structure

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 the Tatar language.

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

The first application is a baseline speech recognizer built on the basis of the 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 the initial data for the Tatar language. These data were used to build the necessary acoustic, pronunciation and language models.

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

3.1 Acoustic features of the Tatar language

Obviously, the acoustic features of a specific language is 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 at the next stages of analysis. The main result of this stage is the automatic phoneme transcribing tool.

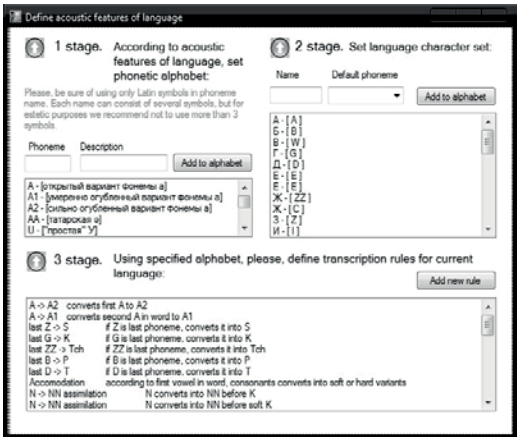

Fig. 1. "Acoustic features" module

The "Acoustic features" module is used to automate and provide the necessary visual forms and the formal type of acoustic rules representation, Fig. 2. This module consists of three main parts, each specialized to work with the character alphabet, the phoneme

alphabet and the 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-һ), 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 a phonetically rich and balanced speech corpus, we need to create a text corpus with similar features. Therefore, we used the automatic phoneme transcription subsystem and the statistical analysis of the resulting transcriptions in the "Text analysis" module, which is shown in Fig. 3.

Basing on the mentioned tools, we have created a text corpus, which consists of separate parts differentiated by the text source types: news, literature, separate words, spontaneous spoken sentences. The total amount of sentences is 776, the number of words – 6913; all the chosen phonemes are presented in sentences' transcriptions.

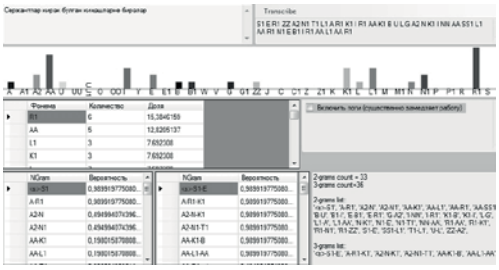

Fig. 3 "Text analysis" module

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

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

3.3 Speech corpus and acoustic models

The creation of the multi-speaker speech corpus for the Tatar language is currently in progress. So far, it contains the voices of 251 speakers, whose average age is 18.2. Each of the speakers read the set of 36 sentences from the text corpus: 13 sentences from the literature part, 7 – from the news part, 15 – from the words part, and 1 – from the 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 the spontaneous part – by one speaker. The total number of sentences in the corpus is equal to 8638. The result features of the currently available speech corpus are shown in Table 1.

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

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

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

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

3.4 Pronunciation model

To evaluate the quality of the developed system, we used the task grammar that allows the speakers to pronounce every possible word sequence. The vocabulary for this task is medium-size (1135 words) and it consists of words that are present in the test subcorpus, so we have simulated a rather compact task domain.

The last step in preparing the data for the decoding stage is creating a pronunciation model. This kind of model is a bridge between the phonemes and the words level of the recognition system. Each word has to be represented by a sequence of appropriate phonemes; this will make it possible to solve the inverse task of decoding words from a sequence of phonemes. Phoneme transcription of all words has been carried out using the developed grapheme-to-phoneme tool.

3.5 Adapted speech recognition system

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

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

- All words in the existing text corpus are divided into morphemes.
- The last morphemes of each word are provided with an additional '#' sign that means 'the end of the word'.
- Statistical 3-gram models are built for the 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 the context in a concrete word.

4 Experimental results

We used the testing 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 the baseline in both the correctness and accuracy coefficients; that proves the fact that adding the morphological

level helps to build models and execute recognition in a 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.