# Speech Analysis and Synthesis Systems for the Tatar Language

Aidar Khusainov
Institute of Applied Semiotics
of the Tatarstan Academy of Sciences
Kazan, Russia
khusainov.aidar@gmail.com

Alfira Khusainova
Kazan (Volga region) Federal University
Kazan, Russia
alfirahamzovna@gmail.com

***Abstract*—In this paper we describe our recent work of creation speech human-machine interface for the Tatar language. Our work consists of three main elements: speech recognition system, speech synthesizer and language identification system. These systems will be used in mobile and desktop applications, for instance, machine translation system, smart assistant.**

## I. 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 the functionality of automatic dictation, voice search, etc. Speech human-machine interface has some significant advantages comparing to standard interfaces: it is natural for a person to use voice to exchange information; it is comfortable to use and fast in some conditions. Nevertheless, the main reason of applying speech interface to many new applications is the increased quality of solving speech analysis, synthesis and text understanding tasks.

This paper describes recent results in the creation of three speech systems for the Tatar language:

- continuous speech recognition system with large vocabulary,
- parametric speech synthesis system,
- automatic language identification system.

All of the built systems are based on corpus approach. We have used the multispeaker speech corpus for Tatar (58 hours of read speech) [1] and two single-user speech corpora (male and female voice) for speech synthesis task.

Automatic language identification system works with Tatar, Russian and English languages; for Russian and English languages we use VoxForge [2] and TIMIT [3] corpora.

## II. Continious speech recognition system for the Tatar language

Automatic speech recognition system consists of 4 elements:

1) *Acoustic models:* represent the relationship between an audio signal and the linguistic units (for instance, phonemes).

2) *Lexical model (phonetic transcriptions of vocabulary words):* transcriptions of words based on phonetic alphabet; often produced automatically via grapheme-to-phoneme systems.

3) *Language model:* a probability distribution over sequences of words in a given language; estimated based on a big amount of text data.

4) *Decoder:* uses models to calculate the most probable word sequence according to input speech signal.

### *A. Speech corpus*

The most modern systems use speech corpora with the total duration of hundreds and thousands hours to create robust acoustic models. The robustness in this case means relatively equal recognition accuracy for male and female speakers, speakers of different sex and age, etc.

Building and annotating the multi-speaker speech corpus for the Tatar language is currently in progress [1]. Nowadays it consists of two main annotated parts: “Core” and “Reading”. The first part aims to cover all the possible phonemes pronounced by the large number of speakers. “Reading” part is focused on increasing the duration of corpus via recording 30-minutes audio files.

The corpus contains additional meta-information about speakers (gender, age, mother tongue) and expert's score of speakers’ proficiency in Tatar. In addition, we plan to continue recording and annotating the “spontaneous” part of this corpus.

The main characteristics of speech corpus are presented in Table I.

Context-dependent acoustic models have been trained using HTK toolkit [4]. Models are 3-state left-right Hidden Markov model-based models; the number of Gaussians in mixtures varies from 2 to 30.

Context dependency is introduced with left and right adjacent phonemes. Therefore, each context-dependent (CD) phoneme is presented with a triple of context-independent phonemes designated as *a-b+c* where *b* is a central phoneme name, *a* and *c* are names for left and right context phonemes respectively. Phonemes names are taken from the basic phoneme alphabet for Tatar (a, ae, b, ch, d, dzh, e, f, g, h, i, j, k, kh, l, m, n, ng, o, oe, p, r, s, sh, t, ts, u, ue, v, y, z, zh)

ISBN 978-952-68397-8-3 (PDF)

accomplished with a phoneme-pause pau, which makes total 32 items.

TABLE I. THE CHARACTERISTICS OF MULTI-SPEAKER SPEECH CORPUS FOR THE TATAR LANGUAGE

| Parameter | Value |
|---|---|
| Number of speakers | 377 |
| Duration | 57:55:09 |
| Average duration per speaker | 9:13 |
| Number of speakers in “Core” part | 251 |
| Duration of “Core” part | 8:12:16 |
| Average duration per speaker in “Core” part | 1:58 |
| Number of speakers in “Reading” part | 126 |
| Duration of “Reading” part | 49:42:53 |
| Average duration per speaker in “Reading” part | 23:40 |
| Number of speakers (unlabeled spontaneous speech part) | 5:19:33 |
| Duration (unlabeled spontaneous speech part) | 168 |
| Average duration per speaker (unlabeled spontaneous speech part) | 1:54 |

*B. Acoustic models*

Shared states are valid only within CD-phonemes having the same central phoneme name. Total count of shared states is limited to 8000. Transition probabilities between states are retained equal within all CD-phonemes for a fixed central phoneme.

In this work we have created acoustic models for rather good quality recordings: 16 bits per second, 16 kHz. We could use them to recognize speech in offices, in front of home PC, to analyze speeches in not very noisy conditions. In future, we are planning to create separate acoustic models for a speech transmitted over telephone, TV and radio channels.

*C. Lexical model*

We can create Tatar phoneme recognition system using information from acoustic models. To create word recognitions system we need to create automatic phonetic transcription algorithm. This algorithm works based on grapheme-to-phoneme rules.

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

There is no definite answer on the question about phonetic alphabet for the Tatar language. Therefore, we have used currently available results of phonetic research. In addition, we have taken into account features of speech recognition task: we need to enumerate basic phones of the language that can change the meaning of the utterance, grouping them into classes with similar sounding.

As the result of our analysis, we have identified 39 alphabet characters (Russian alphabet plus 6 specific Tatar characters Ә-ә, Ө-ө, Ү-ү, Җ-җ, Ң-ң, Һ-һ), 56 Tatar phonemes (43 consonants and 13 vowels) and 37 rules of grapheme-to-phoneme conversion [5].

*D. Language model*

Language model creation task arises in many applications from spellchecking to machine translation systems. In all cases, language model has to describe language grammar rules and has the ability to estimate probabilities of word sequence in specified language.

The Tatar language belongs to agglutinative language family. Thus, its main characteristic is rich morphology. If we try to use standard approaches to create language model for agglutinative language we will face the problem of very large vocabulary size. Due to a large number of possible affixal chains that can follow stems, it becomes impossible to create vocabulary with adequate number of words and OOV (out of vocabulary) rate at the same time.

To solve this problem researches often use sub-word units as base ones to create statistical model. We have selected these sub-word units to analyze language model quality for the Tatar language:

- word,
- morpheme,
- stem plus affixal chain,
- morphs (statistically selected morphemes),
- syllables,
- letters.

We have built language model for the Tatar language using SRILM toolkit (Speech Technology and Research (STAR) Laboratory) [6]. This tool has the functionality to create n-gram models, can interpolate different models and estimate the quality of built models. Common way to use SRILM is as follows:

1) Executing ‘ngram-count’ function to calculate the count of n-grams.
2) Executing ‘ngram-count’ function to build language model based on the results of the first step. Smoothing algorithm has to be specified.
3) Model quality estimation using ‘ngram’ function with ‘ppl’ parameter.

Moreover, some tools have been developed for text corpus processing and automating of the language model creation. These tools include the following core modules:

1) Corpus preprocessing (filtering, dividing into train and test parts).
2) Splitting words into chosen sub-word elements.
3) Automation tool that can build the set of language models with different settings, estimate the quality of built models and create result report file.

To split words of text corpus into sub-word elements we used several tools. Splitting word into stem and morphemes has been implemented using ‘MorphAn’ morphoanalyzer [7]. Selection of morphs – statistically selected part of words – using Morfessor tool [8]. We divide words into set of syllables using own algorithm that has knowledge of 6 possible types of

syllables in the Tatar language (V, CV, VC, CVC, VCC, CVCC).

Texts that have been used to create language models are from the Tatar National Corpus [9]. The main characteristics of text corpus that we have got after preprocessing step are presented in Table II.

TABLE II. THE CHARACTERISTICS OF THE TEXT CORPUS

| Parameter | Value |
|---|---|
| Number of files | 217 294 |
| Number of words | 69 810 033 |
| Number of words in learning part | 64 629 794 |
| Number of words in test part | 5 180 239 |
| Number of syllables | 186 014 478 (2,66 per word) |
| Number of morphemes | 110 280 448 (1,58 per word) |
| Number of morphs | 93 458 542 (1,34 per word) |
| Number of stem plus affixal chains | 97 461 218 (1,4 per word) |
| Number of letters | 434 636 548 (6,23 per word) |
| Size | 901 MB |

According to the limit of the corpora size, the developed language models cannot be complete. Thus, there will be unseen n-grams with zero probability. As the probability of the entire speech utterance is calculated as the multiplication of separate n-grams, this can lead to the situation, in which even one unseen n-gram zeroes out the total utterance probability. To overcome this drawback we used several smoothing algorithms.

Taking into account that this is the first research of language modelling for the Tatar language, we focused on obtaining the maximum information on the impact of different factors on resulting language model quality. Thus, statistical language models have been built for all possible combinations of these categories:

*1) Basic element type*: 6 types. Word, syllable, morpheme, morph, stem plus affixal chain, letter.

*2) N-gram size:* bigram, trigram, 4-gram (5-gram for letter-based models).

*3) Smoothing algorithms:* 5 types. Absolute smoothing, Good-Turing, Kneser-Ney, Witten-Bell, modified Kneser-Ney algorithm.

The quality of built models was evaluated on the following parameters: log probability calculated on test subcorpus, perplexity (model confidence level in analysis of the test subcorpus), OOV (the number of found elements that do not exist in vocabulary), model size (number of n-grams).

As a result of experiment we can make a conclusion that Kneser-Ney and modified Kneser-Ney algorithms showed the best results. Word-based language model has the best log probability value among 95 built models, Table III.

As mentioned above, one of the main problems of statistical modelling languages with rich morphology is very large vocabulary required to cover the entire lexicon. It leads either to reduce the speed of large vocabulary systems, or to increase the number of OOV words while reducing size of the vocabulary. From this point of view, sub-word based language models showed significant reduction of OOV words.

TABLE III. LANGUAGE MODEL COMPARISON

| Base element | Log probability, thous. |
|---|---|
| Word (4-gram) | -12 209,0 |
| Stem plus affixal chain (4- gram) | -12 386,7 |
| Morpheme (4- gram) | -12 638,7 |
| Morph (4- gram) | -12 772,4 |
| Syllable (4- gram) | -14 282 |
| Letter (5- gram) | -20 741,5 |

For experiments we have chosen 20k, 50k and 200k vocabularies for each type of modeling unit. The smallest number of elements in the vocabulary for the complete coverage of the test subcorpus lexicon has been shown by syllable and morph-based models. The results are shown in Table IV.

TABLE IV. LANGUAGE MODEL COMPARISON WITH 20K, 50K AND 200K VOCABULARIES

| Base element | Vocabulary size | OOV |
|---|---|---|
| Word, 3-gram | 20 thous. | 17% |
| Morpheme, 3- gram | 20 thous. | 7% |
| Morph, 3- gram | 20 thous. | 3% |
| Syllable, 3- gram | 20 thous. | 0% |
| Stem plus affixal chain, 3- gram | 20 thous. | 5% |
| Word, 3- gram | 50 thous. | 10% |
| Morpheme, 3- gram | 50 thous. | 5% |
| Morph, 3- gram | 50 thous. | 0% |
| Syllable, 3- gram | 50 thous. | 0% |
| Stem plus affixal chain, 3- gram | 50 thous. | 2% |
| Word, 3- gram | 200 thous. | 5% |
| Morpheme, 3- gram | 200 thous. | 3% |
| Morph, 3- gram | 200 thous. | - |
| Syllable, 3- gram | 200 thous. | - |
| Stem plus affixal chain, 3- gram | 200 thous. | 1% |

In the final experiment, we have built word class-based language model using 20k vocabulary. We used Brown algorithm [10] to define word classes. The developed model has no out of vocabulary words, has small size, but characterized with poorer quality. The result of decomposition of the 20k words into classes has some interest: automatically selected classes unite words with similar meanings. For example, separate classes for city names, numbers, years, surnames, country names, professions have been constructed.

As a part of the automatic speech recognition system for the Tatar language we used word-based 3-gram model with 100k most frequently used Tatar words in vocabulary.

### *D. Continuous speech recognition system for Tatar*

We used Julius toolkit as a decoder [11]. Speaker-independent continuous speech recognition system for the Tatar language has been built based on words models. The system can work in console mode, providing maximum of the service information, and in window mode, showing only recognition result. For user convenience, we have included additional speech activity detection algorithm, so user can

speak several phrases without the need of manipulating mouse or keyboard.

The developed system will be used in applications, such as speech-to-speech vocabularies, machine translation, smart assistant. During 2016 speech recognition service will be available on the site [12].

For the experiment, we used the vocabulary of 100 000 words, test subcorpus consists of 18117 sentences read by 52 speakers. Experimental results are shown in Table V.

TABLE V. THE RESULTS OF TESTING SPEECH RECOGNITION SYSTEM FOR THE TATAR LANGUAGE

| Parameter | Value |
|---|---|
| Correctness | 82.31% |
| Accuracy | 76.21% |
| Total number of words | 100060 |
| Deletion error | 900 |
| Substitution error | 16797 |
| Insertion error | 6108 |

## III. AUTOMATIC SPEECH SYNTHESIS SYSTEM FOR THE TATAR LANGUAGE

The goal of speech synthesis system is to produce audio signal according to the input text phrase. Most of the approaches use concatenative approach (diphone-based synthesis [13], unit selection [14]). The initial information for these approaches is small audio fragments from speech signal. Starting from 2002, parametric approach is gaining popularity. Base elements in parametric approach are not audio fragments, but statistical models of phones.

We use parametric approach based on hidden Markov models to develop synthesizer for the Tatar language (HMM-based speech synthesis, HTS [15]). Similar to acoustic models for recognition task, acoustic models for synthesis task have to be trained on annotated speech corpus.

The difference of creating speech corpus for synthesizer is that we need high-quality recording, so we need professional sound-recording equipment and soundproof room. Moreover, we recorded professional theatre actors (male and female).

Recorded files have been manually annotated by the experts. Experts have annotated all intonational groups, tagged all loanwords and accented words. The resulting annotation has been converted into phonetic transcription using the grapheme-to-phoneme converter.

In addition to manual annotation of the text, we have created a script to automatically extend the annotation using following features:

*1) Phoneme level:* current phoneme, two previous and two succeeding phonemes.

*2) Syllable level:* syllable type (V, VC, CV, CVC, VCC, CVCC); phoneme position in syllable; number of phonemes in previous, current and next syllable; current syllable position in the word; vowel in current syllable.

*3) Word level:* part of speech, number of syllables in previous, current and next word; number of previous and next words in phrase.

*4) Phrase level: number of words in previous, current and next phrase; number of syllables in previous, current and next phrase.*

## IV. LANGUAGE IDENTIFICATION SYSTEM

In practical tasks, we need to correctly identify language that is spoken before applying speech analysis system. In the context of using the Tatar speech analysis systems, we need to identify between three languages: English, Russian and Tatar.

Schematically, the process of the proposed system is shown in Fig. 1.

PPRLM (Parallel Phone Recognition followed by Language Modeling) approach for language identification was used. This approach requires speech corpora for all the three languages: English, Russian and Tatar.

As the training information for English part of the system we used TIMIT corpus, for Russian – VoxForge corpus, and for Tatar – speech corpus described in II.A.

The general scheme of proposed system work assumes that three languages distinguished based on information from English and Tatar recognizers.

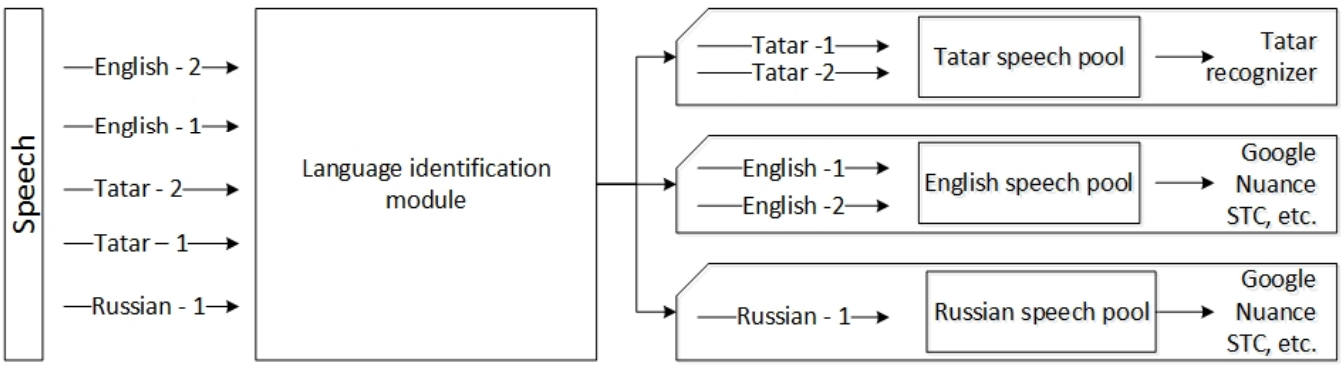


Fig. 1. The structure of the language identification system for Russian, English and Tatar languages

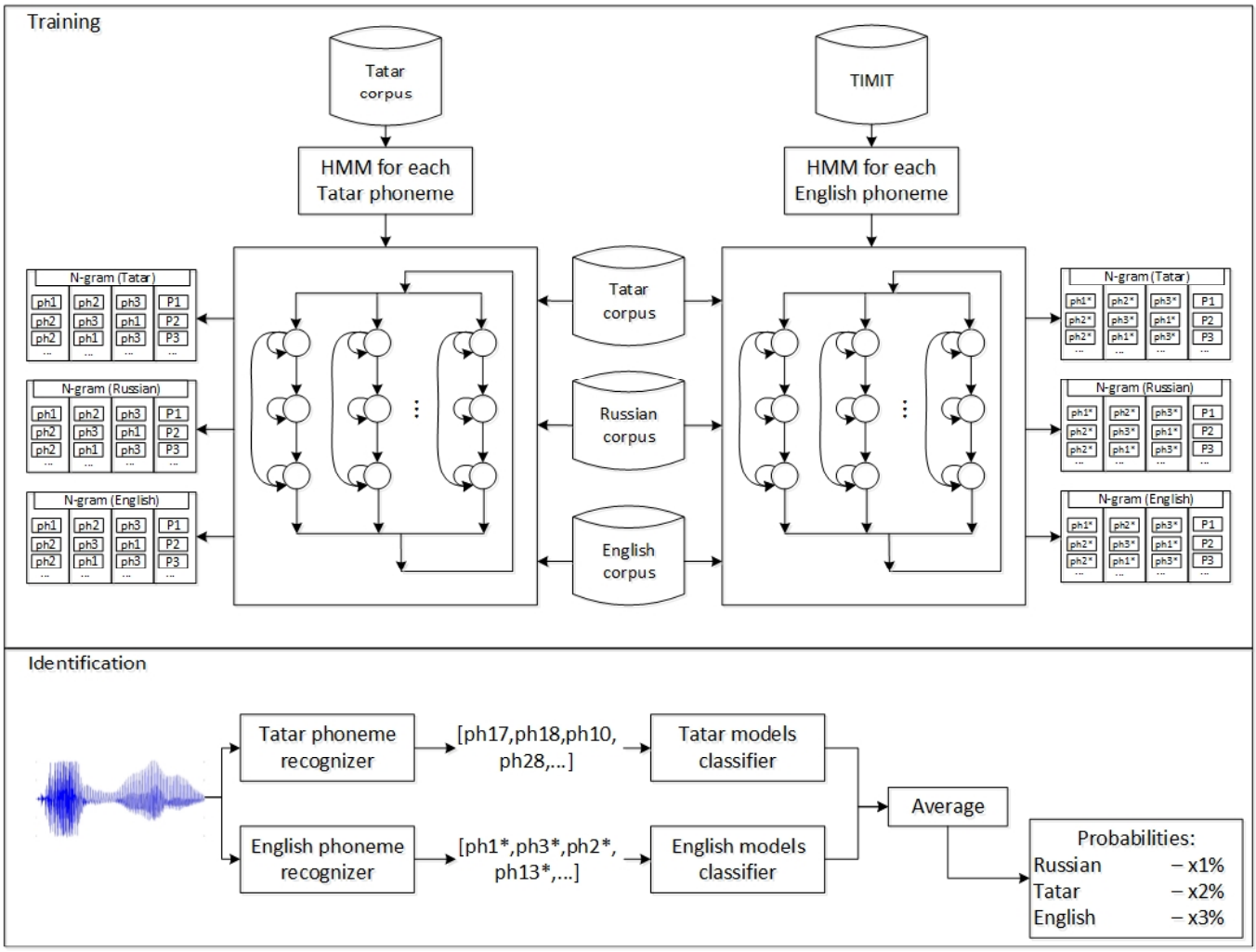


Fig. 2. Structure of PPRLM approach-based system

As the term suggests, PPRLM approach has two main parts: phone recognition and language modeling (generally n-gram models). Records from the training parts of speech corpora in each language are tokenized by English and Tatar phone recognizers. The resulting phone sequences are used to estimate the language model pairs (English phone and Tatar phone-based) for each language.

During the recognition, the same recognition procedure take place. Then the system calculates the probability that this phone sequence belongs to each of the three languages. The language model with the highest probability is selected as the result of the identification process. The structure of a PPRLM-based system is shown in Fig.2.

The main parts of the proposed language identification system are following:

1) Speech corpora of the English, Tatar and Russian languages.
2) English and Tatar acoustic models and phoneme recognizers.
3) Phonotactic language models of the English, Tatar and Russian languages.
4) Classifier and decision maker.

To develop language identification system according to the scheme described above, we need English and Tatar phoneme recognition system. TIMIT speech corpus contains speech utterances and text phonetic annotation needed to train acoustic models. For the Tatar language we use part of developed continuous speech recognition system.

Language model training is as follows: all speech files from the three corpora used as input for English and Tatar recognizers. As the result of recognition process, we can associate each language with the long sequence of English and Tatar phonemes. These sequences used to construct two 3-gram models for each language.

Similar to language modeling for continuous speech recognition task, we need to smooth our phoneme 3-gram models in order to give non-zero probabilities to unseen sequence of phonemes. In this case, we used Katz's back-off algorithm. It accomplishes the estimation of unseen n-grams by "backing-off" to models with smaller histories.

The resulting statistical models describe patterns of sound sequence in these languages and provide the initial data to determine the language of the speaker.

To evaluate the quality of the developed language identification system, we have worked out a testing subcorpus for each of the analyzed languages. This testing corpus contains three hundred records with duration from 3 seconds to 2 minutes each.

Overall, the language identification system has shown 94 percent correctness on the testing corpus. Nevertheless, this quality varies from language to language. As can be seen in Table V, the average quality for the Russian and English languages identification exceeds the same value for the Tatar language.

The results of testing language identification system are shown in Table VI.

TABLE VI. THE RESULTS OF TESTING LANGUAGE IDENTIFICATION SYSTEM

| Parameter | English | Russian | Tatar |
|---|---|---|---|
| Identification correctness | 96% | 97% | 91% |
| Overall | 94,7% | | |

## V. CONCLUSION

In this paper, we present three speech systems for the Tatar language: speaker-independent continuous speech recognizer, parametric speech synthesizer and language identification system. These systems allow us to start working on inclusion human-machine speech interface in the Tatar language.

Further development of developed systems makes possible the joint use of the results in semantic and speech analysis of the Tatar language to create intellectual systems. We plan to develop mobile and desktop applications for different dictionaries, machine translation, tools for dictation, for visually impaired, etc.

## REFERENCES


[1] A. Khusainov, "Design and creation of speech corpora for the Tatar speech recognition and synthesis tasks", *Proc. of the 3rd International Conference on Turkic Languages Processing*, Kazan, 2015, pp. 475-484.

[2] VoxForge official website, Web: http://www.voxforge.org/.

[3] LDC official website, Language resources, Web: https://catalog.ldc.upenn.edu/ldc93s1.

[4] HTK speech recognition toolkit, Web: http://htk.eng.cam.ac.uk/.

[5] A.F. Khusainov, "Automatic phoneme recognition system for the Tatar language", *Proc. Of the 1st International Conference on Turkic Languages Processing*, Astana, 2013, pp 211–217.

[6] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit", *Proc. Intl. Conf. on Spoken Language Processing*, vol. 2, Denver, 2002, pp. 901-904.

[7] D. Sh. Suleymanov, R. A. Guilmoulline, A. A. Guilmoulline, "Tatar phonological rules as a base of two-level morphological analyzer", *in Proceedings of LP'2000*, Prague: The Karolinum Press. – P. 495–504.

[8] M. Creutz, K. Lagus, "Unsupervised discovery of morphemes", *In Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, Philadelphia, Pennsylvania, 11 July, 2002, pp. 21-30.

[9] Dz. Suleymanov, O.A. Nevzorova, and B. Khakimov, "National Corpus of the Tatar Language "Tugan Tel": Structure and Features of Grammatical Annotation", *Proc. International Conference Georgian Language and modern Technology*, Tbilisi, 2013, pp. 107-108.

[10] P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai and R. L. Mercer, "Class-Based n-gram Models of Natural Language", *Computational Linguistics*, vol. 18(4), 1992, pp. 467-479.

[11] Open-Source Large Vocabulary Continuous Speech Recognition Engine, Web: https://github.com/julius-speech/julius.

[12] Programmnye produkty, lokalizovannye na tatarskiy yazyk, Web: http://tatsoft.tatar.

[13] E. Moulines, F. Charpentier, "Pitch-sinchronous waveform processing techniques for text-to-speech synthesis using diphones", *In Speech Communication*, 9 (5/6), 1990, pp. 453–467.

[14] Y. Sagisaka, "ATR v-talk speech synthesis system", *In Proc. ICSLP-92*, Banff, Canada, 1992.

[15] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-basedspeech synthesis", *In Proc. Eurospeech*, 1999, pp. 2347–2350.