A Bilingual Kazakh-Russian System for Automatic Speech Recognition and Synthesis

Olga Khomitsevich$^{1,2}$, Valentin Mendelev$^{1,2}$, Natalia Tomashenko$^{1,2}\,$\textsuperscript{(3)}, Sergey Rybin$^2$, Ivan Medennikov$^{2,3}$, and Saule Kudubayeva$^4$

$^1$ Speech Technology Center, Saint Petersburg, Russia
{kholmitsevich, mendelev, tomashenko-n, medennikov}@speechpro.com
$^2$ ITMO University, Saint Petersburg, Russia
rybin@speechpro.com
$^3$ STC-innovations Ltd., Saint Petersburg, Russia
$^4$ Kostanay State University named after A. Baytursynov, Kostanay, Kazakhstan
saule.kudubayeva@gmail.com

Abstract. The paper presents a system for speech recognition and synthesis for the Kazakh and Russian languages. It is designed for use by speakers of Kazakh; due to the prevalence of bilingualism among Kazakh speakers, it was considered essential to design a bilingual Kazakh-Russian system. Developing our system involved building a text processing and transcription system that deals with both Kazakh and Russian text, and is used in both speech synthesis and recognition applications. We created a Kazakh TTS voice and an additional Russian voice using the recordings of the same bilingual voice artist. A Kazakh speech database was collected and used to train deep neural network acoustic models for the speech recognition system. The resulting models demonstrated sufficient performance for practical applications in interactive voice response and keyword spotting scenarios.

Keywords: Speech recognition · Speech synthesis · ASR · TTS · Kazakh

1 Introduction

This paper describes an Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) system designed for Kazakh speakers. Due to the fact that most Kazakh speakers are bilingual, and Kazakh texts and speech often contain fragments in Russian, we decided to create a bilingual Kazakh-Russian system. The TTS system we designed includes text processing and transcription modules capable of dealing with both Kazakh and Russian text, in particular with Russian borrowings in Kazakh. These modules are also used by the ASR system. We also created two TTS voices using the same bilingual female voice artist: a Kazakh voice and a Russian voice.

A Kazakh speech database, containing read and spontaneous speech, was collected and used to train a bilingual ASR system for Interactive Voice Response
(IVR) and Keyword Spotting (KWS) scenarios. The ASR system employs
the TTS transcription module to generate phonetic transcriptions. A Context-
Dependent Deep Neural Network Hidden Markov Models (DNN-HMM) archi-
tecture is implemented for acoustic model training.

This paper is organized as follows. Section 2 describes the challenges that
the Kazakh language presents for the ASR and TTS tasks. Section 3 describes
our TTS system. An overview of our ASR system and the results of recognition
experiments are given in Sect. 4. Section 5 concludes the paper.

2 The Kazakh Language

Kazakh presents several challenges to Natural Language Processing (NLP) tasks
and the development of speech technologies for this language. One is that few
language resources, such as lexicons, corpora or NLP software, are available for
Kazakh. It can thus be considered an underresourced language, and applications
such as speech recognition and synthesis have to be built virtually from scratch
(for instance, we could only find very basic Kazakh ASR systems described in
the literature [1,2]).

Another, more important problem is the problem of bilingualism and inter-
ferece with the Russian language. Kazakh is a Turkic language that is the state
language of the former Soviet republic of Kazakhstan. The country gained inde-
pendence in 1991 following the collapse of the Soviet Union, however Russian
remains the second official language of the Kazakh state, and its role is still very
prominent. Russian continues to be widely used in education, especially higher
education, in the media, by state institutions, etc., and the majority of Kazakh
speakers are bilingual [3,4].

The consequences of this sociolinguistic situation from the NLP perspective
are twofold. Firstly, code switching and language interference are very common
in Kazakh speech. It is typical for Kazakh texts or conversational speech to
incorporate phrases in Russian or to switch from one language to the other,
sometimes mid-sentence. This is an important challenge for speech recognition:
an efficient ASR system for Kazakh designed to work in real-life situations needs
in fact to be a bilingual system, recognizing both Kazakh and Russian speech
without the need for special tuning in either case.

Secondly, Kazakh has accumulated a large amount of borrowings from
Russian. That has especially important consequences for Kazakh TTS applica-
tions. Kazakh words follow the laws of vowel harmony, and for the overwhelming
majority of word forms the stress is fixed on the final syllable of the word. Bor-
rowed Russian words differ from Kazakh words in many respects. The most
important one for us is that Russian does not have fixed word stress, in fact
stress can fall on any syllable of the word. The unstressed vowels undergo both
qualitative and quantitative reduction: in particular, the phonemes /o/ and /e/
are not pronounced in unstressed positions and are replaced by /a/ and /i/,
respectively, though they are still written as /o/ and /e/. In contrast, Kazakh
vowels do not undergo this type of reduction. It follows that Russian borrowings
in Kazakh text need to be detected and transcribed accordingly. Russian words also contain phonemes not present in original Kazakh words, resulting in an increase in the number of phonemes needed for transcription. The way we deal with these problems is described in the next section.

3 Speech Synthesis and Transcription for Kazakh

Building a Kazakh TTS system involves a number of stereotypical steps. The text needs to be normalized, which means detecting words and sentences, processing non-standard words such as numbers, abbreviations, Latin script, etc. The resulting normalized text is then transcribed, and synthesized speech is formed. The Kazakh TTS system described here is a hybrid TTS system based on Unit Selection (US) algorithms and HMM intonation modeling [5,6], and is part of the VitalVoice TTS developed at Speech Technology Center Ltd. The text processing modules built for TTS, namely, text normalization and transcription, are also used in the ASR application. The tools we developed for these modules are described below.

3.1 Dictionary and POS Tagging

Kazakh belongs to Turkic languages and displays their typical properties [7,8]. It is an agglutinative language with highly regular morphology, where a sequence of affixes is added to a stem in strict succession in order to construct numerous inflectional forms. To help solve the task of text processing for Kazakh, we considered it essential to build a dictionary including a vocabulary of stems and an inventory of inflectional affixes. These tools are used in several ways. First, the affixes are added to non-standard words in writing (for example, 17-mi stands for on жеринми “seventeenth”) so we need to detect them in order to process those words. Second, in most Kazakh words, stress falls on the final syllable, however a few affixes, such as the negative verb affix ма/ме/бә/бе/на/не and its combinations, or the adjectival affix қай/ке/қай/ке do not take stress. Consequently, to predict stress placement, we need to detect affixes and also to resolve Part-of-Speech (POS) homonymy. Finally, common Russian borrowings need to be included in the dictionary; since they are routinely used with Kazakh affixes, it is also important to detect the affix for correct transcription.

Our Kazakh dictionary is organized in the following way. An entry in the vocabulary of stems contains the stem of a word together with its POS label or labels and, if applicable, the “Russian word” label, as well as the stressed syllable number for Russian words. As for affix detection, we should note that in Kazakh each grammatical function is represented by its own affix. The affix sequences can grow very long, and it is virtually impossible to compile a complete set of all possible combinations. However, for the sake of computational simplicity, we did not opt for a full morphological analysis, instead we made a list of common affix combinations and treated each as a separate affix. The affix lexicon thus contains a list of about 6700 affixes together with their POS labels. During text processing,
each word form is split into stem and affix (if any) and assigned a POS label based on stem and affix label matching. Several basic context-based homonymy resolution rules are added to deal with remaining POS disambiguation.

3.2 Building Transcription Rules and Synthesizing Speech

Kazakh writing is based on the Cyrillic alphabet, with the addition of several specific letters [8]. It is a relatively straightforward phonemic orthography, so we decided to opt for rule-based transcription of Kazakh words. We developed a set of letter-to-phoneme rewriting rules, taking into account assimilation laws. It is important to note that these rules are only applicable to original Kazakh words. For Russian borrowings, a separate letter-to-phoneme transcription rule set had to be added, resulting in virtually two rule sets in one algorithm. The set of phonemes used in our TTS/ASR system consists of 59 phonemes, including both original Kazakh phonemes and those that only occur in Russian borrowings.

Consequently, an important step in our system is to determine which set of transcription rules we need to apply to a particular word, that is, whether or not it belongs to Russian borrowings. As noted in Sect. 3.1, frequently used Russian borrowings are included in the lexicon with the corresponding label. For out-of-vocabulary (OOV) words, this decision is made based on the letters and letter combinations occurring in the word. Thus, we look for specific letters that do not occur in original Kazakh words, such as φ, ρ, ρ̃, for sequences of consonants at the beginning of the word, for consonant clusters typical for Russian, such as ǝк, ǝтб, ǝтр, and so on. Once the borrowed word is detected, a special tag is assigned to it for use in the subsequent processing modules.

An important issue is detecting the correct stress in Russian borrowings. We were not able to find any reference to this problem in the literature, which only describes stress in original Kazakh words. However, from our analysis and our interviews with Kazakh speakers we were able to conclude that the majority of Russian borrowings retain the stress they have in the source language. However, when a Kazakh affix is added to the Russian word, the rule that the stress shifts to the final syllable also applies. What happens is that the main stress remains on the stem, while the final syllable of the affix receives auxiliary stress, except for words that have the stress on the final syllable of the stem: in those cases the main stress shifts to the affix while the stem retains an auxiliary stress. We took this observation as the rule of thumb to be used in our TTS system, although this problem calls for further phonetic study. A remaining issue is detecting stress placement in OOV words detected as Russian borrowings. We developed a set of rules for stress placement in the detected Russian words based on the final combinations of letters in these words.

Finally, a few words need to be said about the recording process of the US database. We used a phonetically balanced Kazakh text that included a number of common Russian borrowings, and we also added a phonetically balanced Russian text to be read by the same voice artist (who was bilingual). The Russian borrowings and all the words in the Russian part of the voice database were then labeled with the special “ru” label. During synthesis, if a word in the input text
Table 1. Contents of recordings

<table>
<thead>
<tr>
<th>Content</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credit card numbers</td>
<td>Reading</td>
</tr>
<tr>
<td>Telephone numbers</td>
<td>Reading</td>
</tr>
<tr>
<td>Cities</td>
<td>Reading</td>
</tr>
<tr>
<td>Countries</td>
<td>Reading</td>
</tr>
<tr>
<td>Street names</td>
<td>Reading</td>
</tr>
<tr>
<td>Surnames</td>
<td>Reading</td>
</tr>
<tr>
<td>Commands</td>
<td>Reading</td>
</tr>
<tr>
<td>Dates</td>
<td>Reading</td>
</tr>
<tr>
<td>Phonetically rich sentences</td>
<td>Reading</td>
</tr>
<tr>
<td>Bio</td>
<td>Spontaneous</td>
</tr>
<tr>
<td>Current date and time</td>
<td>Spontaneous</td>
</tr>
<tr>
<td>General questions</td>
<td>Spontaneous</td>
</tr>
</tbody>
</table>

is detected as a Russian borrowing, the diphones from the “ru”-labeled words are preferably used for it in the Unit Selection process. In addition to that, a separate TTS voice was built using only the Russian part of the voice database. This voice can be used in conjunction with the Russian VitalVoice TTS engine if a purely Russian text needs to be synthesized using the same voice as in the Kazakh TTS.

4 Automatic Speech Recognition for Kazakh

4.1 The Speech Database

The Kazakh speech database used for ASR training was collected at Speech Technology Center Ltd. It comprises 780 sessions recorded over GSM and landline telephone networks in Kazakhstan. Each session is approximately 15 min long and contains speech of one native Kazakh speaker. Each speaker participated in exactly one session, so the total number of speakers is 780, of which 392 were female and 388 were male. The contents of the recordings are described in Table 1. The total duration of the database is 147 h. About 120 h from this database were chosen for acoustic model training. The word “Bio” in Table 1 denotes the situation when the speaker was asked to produce a short spontaneous description of his or her biography.

4.2 Acoustic Models

In this section we present our setup for the acoustic models. We trained Deep Neural Network Hidden Markov Models (DNN-HMM) acoustic models [9], and followed the concept of cross-language knowledge transfer [10,11], in which we
used the Russian speech corpus to improve the performance of Kazakh acoustic models. More specifically, the target language (Kazakh) network was initialized with an existing source (Russian) network. Our preliminary results showed that using both Russian and Kazakh datasets significantly improved the performance of acoustic models in comparison with the system trained only on the Kazakh dataset.

We trained all models using our proprietary tools and Kaldi speech recognition toolkit [12]. For each language, we first built two standard maximum-likelihood (ML) trained GMM-HMM systems, using 39-dimensional Mel-frequency cepstral coefficients (MFCC) features (C0-C12, with delta and acceleration coefficients). The number of context-dependent triphone states was 1500 for Russian, and 4100 for Kazakh. Then, using the obtained state tying [13] we trained two DNN-HMM models, as shown in Fig. 1. Input features for these DNNs were 13-dimensional MFCC with Cepstral Mean Normalization (CMN), spliced in time taking a context size of 31 frames (i.e., ±15). The resulting 403-dimensional features were used to train Russian and Kazakh DNNs.

![Diagram](image)

**Fig. 1.** DNN-HMM acoustic model training for Kazakh based on transfer learning using Russian speech data

The first DNN was trained using the data only from the Russian corpus. It had 5 hidden layers with 1000 neurons in each layer and a softmax layer with 1500 senons, corresponding to tied states of the Russian GMM-HMM. The DNN was trained with pre-training using the standard back propagation algorithm with
cross entropy error criteria. Then the softmax layer of this DNN was replaced with the softmax layer (with 4100 senones) corresponding to the Kazakh language, and the resulting DNN was finetuned on the Kazakh corpus using several iterations of sequence-discriminative training with state-level Minimum Bayes Risk (sMBR) criterion [14].

4.3 Experiments

In this section we present the experimental results for the IVR and KWS scenarios. To test the system, 40 recordings from the recorded speech database were chosen randomly and were not used in training. In-grammar (IG) sets were composed from the test recordings for “Yes/No”, “Cities” and “Surnames” grammars. All the grammars had a linear list structure and contained 2, 135 and 230 items respectively. Out-of-grammar (OOG) sets were made artificially to mimic real conditions. All sets except “Yes/No” included both Kazakh and Russian words. The results for the IVR scenario in terms of equal error rate (EER) are presented in Table 2. The network described in Sect.4.2 is referred as DNN-2 in the tables below. For comparison purposes, we also trained another DNN acoustic model (denoted as “DNN-1”) using only the Kazakh corpus. It had a topology similar to DNN-2, and was trained in a similar way, except the pre-training stage: layer-wise pre-training based on Restricted Boltzmann Machines (RBM) [15] was performed on Kazakh speech data.

<table>
<thead>
<tr>
<th></th>
<th>Yes/no</th>
<th>Cities</th>
<th>Surnames</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN-1</td>
<td>–</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>DNN-2</td>
<td>2,3</td>
<td>15</td>
<td>10</td>
</tr>
</tbody>
</table>

For the keyword spotting task, spontaneous parts of the test recordings were used with target word lists consisting of 10 and 100 items. The KWS results are presented in Table 3.

<table>
<thead>
<tr>
<th></th>
<th>10 words</th>
<th>100 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN-1</td>
<td>18</td>
<td>29</td>
</tr>
<tr>
<td>DNN-2</td>
<td>13</td>
<td>20</td>
</tr>
</tbody>
</table>

As evident from Tables 2 and 3, DNN-2 achieves significantly lower error rates compared with DNN-1. It should be noted that the observed Kazakh language ASR performance is weaker than that of our Russian language system but is high enough for practical applications.
5 Conclusions and Future Work

In this paper we presented a bilingual Kazakh-Russian TTS and ASR system. The TTS system includes both a Kazakh and a Russian voice based on the voice of the same bilingual voice artist, and tools for processing Kazakh and Russian text. The ASR system was trained on approximately 147 h of Kazakh speech and tested on IVR and KWS scenarios. The system is well suited for applications targeted at Kazakh speakers, who often use both Kazakh and Russian languages interchangeably.

In the future, we plan to record a second (male) Kazakh TTS voice for our system, and to continue enhancing the text processing and transcription modules of the TTS system in order to get rid of remaining processing errors. As for ASR, the development of a large vocabulary continuous speech recognition system for Kazakh language is the next obvious goal, and work in this direction is now in progress.

Acknowledgments. The work was financially supported by the Government of the Russian Federation, Grant 074-U01.

References