The experiments has shown that this system's performance of translation from Chinese to Kazakh language in satisfactory level. In future, it is planned to improve the translation of verb tenses as well as use of the plural nouns, since compared with English language their rules are much more and complicated depending on the vowels and consonants.

Defining the tenses is also considered as a complex process , because it is necessary to perform analysis. Afterwards using taken analysis the program synthesizer has to be created.

Creating the synthesizer include several difficulties since the results needs to be matched with the tense used in original text. Otherwise the sentence meaning can be modified and making the translator confused. All of this abovementioned cases is fully linked to the distinct features of Kazakh language compared with other languages.

.

**References**

1. Gulila Altenbek. and Dawel,A. and Muheyat,N. 2009. A Study of Word Tagging Corpus for the Modern Kazakh Language, Journal of Xinjiang University, 26(4):323–326

2. http://check.cnki.net/user/,

UDC 004.8

## SPOKEN TERM DETECTION FOR KAZAKH LANGUAGE

***Zhanibek Kozhirbayev National Laboratory Astana,****Nazarbayev University, Kazakhstan e-mail: zhanibek.kozhirbayev@nu.edu.kz*
***Muslima Karabalayeva*** *National Laboratory Astana,Nazarbayev University,Kazakhstan, e-mail: zhyessenbayev@nu.edu.kz*
***Zhandos Yessenbayev*** *National Laboratory Astana,Nazarbayev University, Kazakhstan e-mail: muslima.karabalayeva@nu.edu.kz*

The paper presents a spoken term detection system for Kazakh language in which significant improvements are obtained through modifying speech-to-text process used for generating word-based lattices. These lattices are indexed and used for the keyword search later. Spoken Term Detection systems quickly discover the occurrence of a term, which might be just a word or sequence of words, in a large audio set of heterogeneous speech records. The paper provides an overview of a speech-to-text and keyword search system architecture built primarily on the top of the Kaldi toolkit and expands on a few highlights. Our aim was to develop a general system pipeline which could be advanced regarding phonological and linguistic features of Kazakh language in order to detect OOV keywords.

**Keywords:** Speech Retrieval, Lattice Indexing, Spoken Term Detection, Speech Recognition, Keyword Search

## ПОИСК РАЗГОВОРНОГО ТЕРМИНА НА КАЗАХСКОМ ЯЗЫКЕ

***Кожирбаев Жанибек Мамбеткаримович*** *Национальная лаборатория Астаны, Назарбаев Университет, Казахстан e-mail: zhanibek.kozhirbayev@nu.edu.kz*
***Карабалаева Муслима Хизятовна*** *Национальная лаборатория Астаны, Назарбаев Университет, Казахстан e-mail: muslima.karabalayeva@nu.edu.kz*
***Есенбаев Жандос Аманбаевич*** *Национальная лаборатория Астаны, Назарбаев Университет, Казахстан e-mail: zhyessenbayev@nu.edu.kz*

В статье представлена система для обнаружения разговорного термина на казахском языке, в котором получены значительные улучшения путем модификации процесса речи в текстовый формат, используемый для создания слов на структурной (lattice) основе. Эта

структура индексируется и позже используется для поиска ключевого слова. Система обнаружения разговорного термина быстро обнаруживает возникновение термина, которым может быть просто слово или последовательность слов в большом аудио наборе разнородных речевых записей. В статье дается обзор речи в текстовый формат и архитектура системы поиска ключевых слов, построенной на основе платформы Kaldi и расширяемой на несколько основных моментов. Наша цель состояла в том, чтобы разработать общую систему, которая может продвинуться вперед в отношении фонологических и лингвистических особенностей казахского языка с целью выявления OOV ключевых слов.

**Ключевые слова:** Поиск речи, Индексация структур, Поиск разговорного термина, Распознавание Речи, Поиск по ключевым словам

## 1. Introduction

Data processing is nowadays an essential activity of IT industry and recorded materials is considered as core source of that data. Hence, together with increasing volume of audio information, a possibility grew for developing an efficient information retrieval system for audio data storage. Spoken term detection (STD) attempts to set up a specific term (considered as a sequence of single or multiple words) speedily and thoroughly in major heterogeneous audio storages, in order to be utilized solely as input to better compounded information retrieval techniques.

The STD task is based on the detection process, demanding every appearance to be clarified by its beginning and ending times in contradiction to spoken document retrieval. Moreover, systems have to maintain a score for every appearance and a tough decision showing its accuracy. Unprocessed audio materials and a list of search key words together make an input for the task. Despite the fact that the evaluation practically utilizes only small number of data, it is designed to initiate the biggest data condition. Thus, the systems are obliged to be deployed in two stages: indexing as well as searching. Processing of audio data is held meanwhile indexing stage going on, without knowing about search terms. In order to retrieve the terms the output index is kept and admitted in the course of the searching stage. The searching stage might be replicated several times for various terms so the effectiveness of its deployment is very significant

The remainder of this paper is organized as follows. Section 2 describes related works. The system architecture will be discussed in details in Section 3. Section 4 demonstrates the experiments and the obtained results. Finally, the last section concludes the paper and suggests further research in this area.

## 2. Related works

Languages with morphologically features like Kazakh experience challenges in large vocabulary continuous speech recognition (LVCSR) process by virtue of vocabulary increase. Enormous vocabularies, great frequency of OOV words as well as scattered data for LM are the prominent indications of word-based lexical techniques. A lot of researches have been conducted to solve these challenges regarding OOV issue.

First method to limit the OOV issue is to preventively enlarge the LVCSR dictionary. To be precise, a person augments automatically created pronunciations of a huge amount of words in the LVCSR dictionary prior to lattice creation [1].

Second method to deal with OOV words is though sub-word blocks similar to phones, syllables or morphemes. A subword index is built by producing a sub-word lattice which were presented by [2], or by modifying a word lattice to a sub-word lattice with or without the utilization of a proper phone confusion matrix described in [3].

The OOV challenge also can be solved by introducing the concept of query expansion in text retrieval. Optional words or syllables for OOV can be produced by a confusion matrix utilized in [4]. Consequently, rather than searching primary keywords, the system searches within word or syllable index. In contradistinction to concept of query expansion, in which a person adds

potentially lacking word with other terms which are corresponding semantic features, speech search involves different words which sound identically.

Job performed in this paper is based on [5], in which proxy keywords are generated utilizing a confusion matrix. Rather than looking for the proxy keywords in list produced by the LVCSR system, weighted finite state transducer (WFST) is applied on the basis of framework for directly linking several proxies versus the whole LVCSR lattices.

### 3. Kaldi-based STD system

The Kaldi-based STD system includes two different subsystems, as illustrated in Figure 1: the ASR subsystem and the STD subsystem. This section presents each subsystem in more detail regarding interprocesses and tuning parameters.

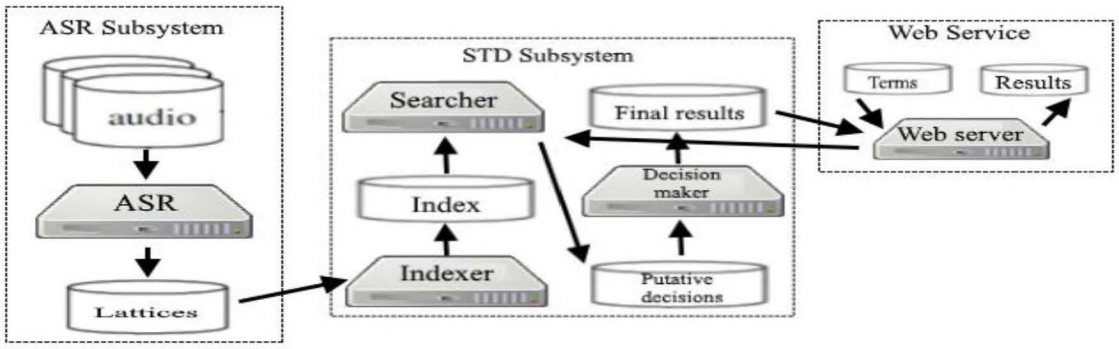


Figure 1: Spoken Term Detection System Architecture

### 3.1 ASR subsystem

The ASR subsystem utilizes the Kaldi toolkit [6] to gather word lattices from the raw audio data. It applies 13-dimensional Mel-frequency cepstral coefficients, Linear Discriminant Analysis transform as well as Maximum Likelihood Linear Transform rating. A flat-start initialization of context-independent phonetic HMMs begins the training, while speaker adaptive training of state-clustered triphone HMMs along with GMM output densities finishes it. Moreover, the ML-based acoustic model training step is run, which followed by a universal background model. It is created from speaker-transformed training data that is utilized to train an SGMM applied in the decoding step to produce the word lattices [7, 8].

Kazakh speech data from the KazSpeechDB and KazMedia corpora is employed to train the above mentioned acoustic model. The KazSpeechDB corpus as part of Kazakh Language Corpus [9] is a body of utterances consisting of 12675 Kazakh sentences recorded in a sound recording studio, uttered by speakers of different age and gender, from different regions of Kazakhstan. Every audio file is supplied with a text file that contains transcription text of the utterance. The corpus contains 22 hours of speech.

The KazMedia corpus is a body of text and audio data collected from official websites of broadcast news channels “Khabar”, “Astana TV” and “Channel 31”. The audio data is 600 wav-files, which are actually audio tracks extracted from a number of video news in Kazakh. Every wav-file is supplied with a txt-file that contains detailed transcription text of the news and a time-aligned annotation file with labels about speaker gender, language and noise. The total duration of these audio files is 13 hours of speech. The total training data amount is about 30 hours of speech.

The dictionary and the language model (LM) of the ASR subsystem were formed on the basis of cumulative text data of both KazSpeechDB and KazMedia sub-corpora. The dictionary contains over 160000 words. The LM was trained using a text database of about 3.5 million words.

A common metric of the performance of experimental speech recognition models is WER (word error rate), which is computed as the ratio of erroneously recognized words to the total number of words. It is commonly believed that a lower WER shows superior accuracy in recognition of speech, compared with a higher WER. The experimental results are given in Table 1.

Table 1: Minimum value of the WER on the train, validation and test sets

| Experiment \ Set | train | dev | test1 (Khabar) | test2 (Astana TV) | test3 (Channel 31) |
|---|---|---|---|---|---|
| Triphones | 6.19 % | 6.42 % | 6.62 % | 14.87 % | 19.52 % |
| SGMM | 5.16 % | 5.39 % | 5.56 % | 13.18 % | 16.95 % |

### 3.2 STD subsystem

The STD subsystem combines Kaldi term detector [1] that searches for the keyword terms among the lattices generated by the ASR subsystem. Firstly, word lattices of all the utterances in the speech data from individual WFSTs are converted to a single generalized factor transducer pattern that aggregates the start and end times, as well as the lattice posterior probability of each word token as a three-tuple cost in the lattice indexing approach presented by [10]. This factor transducer indicates an inverted index of all the word sequences comprised in the lattices. Hence, with the search term, an ordinary finite state machine which receives the term is built and structured with the factor transducer with an eye to gather all the appearances of the term in the audio data. The posterior probabilities of the lattice relatively to all the words of the search term are piled up, appointing a confidence score to every detection. The decision maker process merely gets rid of those detections with a confidence score which is less than a predetermined threshold.

The Kaldi STD system [1] treats OOV term search by virtue of a technique named proxy words [6]. This approach based on replacement every OOV word of the keyword term to acoustically corresponding IV proxy words, demolishing the necessity of a subword-based system for handling OOV term search.

## 4. System evaluation and results

An assessment of the STD systems performance on the basis of the term length (IV words) has been conducted (Table 2, 3). Test search queries are divided into three classes: unigrams, bigrams and trigrams. Generally, longer queries should produce greater results than shorter ones as long as these are originally more liable to be confused with speech data.

Table 2: Search term categories (IV test set)

| | Terms | | |
|---|---|---|---|
| Type | Unigrams | Bigrams | Trigrams |
| Amount | 1000 | 1000 | 1000 |
| Total | 3000 | | |

Table 3: Search terms statistics (IV test set)

| Terms | Occurence of IV quesires in test set |
|---|---|
| unigrams | 82874 |
| bigrams | 12383 |
| trigrams | 3937 |

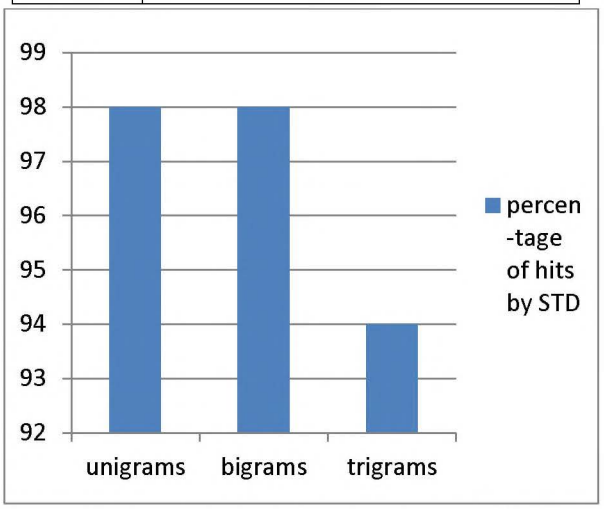


Figure 2: The percentage of hits by STD

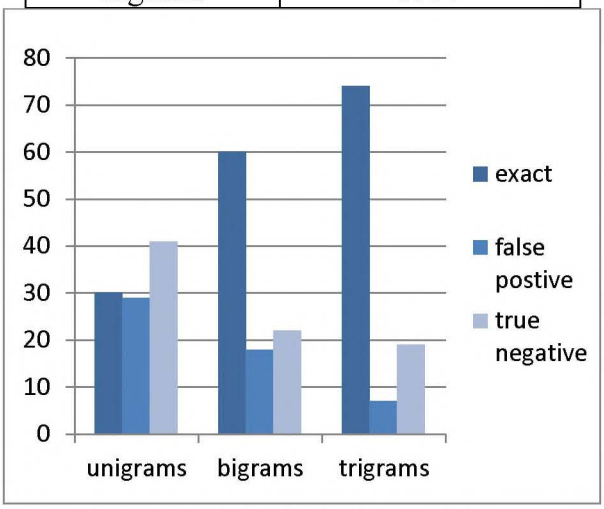


Figure 3: Hits out of 1000 queries for each category

The above figures show that STD system performs in searching IV keywords well. It can be seen that trigrams and bigrams are detected more accurately rather than unigrams as expected (Figure 2). Figure 3 presents the results of exact, false positive and true negative hits out of 1000 queries for each category. As the decision maker process depends on the posterior probability of the occurred term, there might be a false alarm or accurate decision for query terms. Therefore, the exact hits mean that the frequency of query matches with frequency obtained by the STD system, the false positive and true negative hits show that original frequencies are more and less respectively than the gathered ones.

The test set with 295 OOV words was created in order to evaluate the STD system for OOV queries. More precisely, the words are not in the main lexicon but are in the manual transcripts of the test audio set. In order to do fuzzy search, the phone confusions are gathered firstly and a G2P model is trained using G2P Sequitur [11]. Afterwards, this model was applied to the OOV keywords and KWS data directory was built. The system searches the words from the main lexicon with the minimum edit distance score. For example: "ӘЖЕМДІ" → "ӘЖІМДІ", "АДЫМДАРЫН" → "АДАМДАРЫН", "ДӘПТЕРІМ" → "ДӘПТЕРІН". The results of the experiment are shown in Table 4.

Table 4: OOV test set results

| OOV queries | Hits by STD | # Yes decisions | # No decisions |
|---|---|---|---|
| 295 | 112 | 259 | 114 |

## 5. Conclusion and future works

In this paper we present an STD system that benefit significantly from tuning the speech-to-text process and applying lattice indexing. Making applied models more diverse, indexing and searching methods produce more advanced results for our system. Even though the proxy keywords approach applied for detecting the OOV terms and it does not show the expected results, there are other approaches for handling the keywords which are not in lexicon. In future, phone, syllable, morph bases techniques and even feature based method will be utilized to enhance this research for Kazakh language.


## Acknowledgments

This work was supported by the Ministry of Education of Science of the Republic of Kazakhstan under the research program number 349//049-2016.


## References


1. Chen, G, Khudanpur, K, Povey, D, Trmal, J, Yarowsky, D and Yilmaz, O 2013, "Quantifying the value of pronunciation lexicons for keyword search in low resource languages", *in Proceedings of ICASSP 2013*, pp. 8560–8564.
2. Saraclar, M and Sproat, R 2004, "Lattice-based search for spoken utterance retrieval", *in Proceedings of HLTNAACL 2004*.
3. Chaudhari, U & Picheny, M 2007, "Improvements in phone based audio search via constrained match with high order confusion estimates", *in Proceedings of ASRU 2007*, pp. 665–670.
4. Karanasou, P, Burget, L, Vergyri, D, Akbacak, M and Mandal, A 2012, "Discriminatively trained phoneme confusion model for keyword spotting", *in Proceedings of Interspeech 2012*.
5. Chen, G, Yilmaz, O, Trmal, J, Povey, D, Khudanpur, K 2013, "Using proxies for OOV keywords in the keyword search task", *in Proceedings of ASRU*, pp. 416–421.
6. Povey, D, Ghoshal, A, Boulianne, G, Burget, L, Glembek, O, Goel, N, Hannemann, M, Motlicek, P, Qian, Y, Schwarz, P, Silovsky, J, Stemmer, G, Vesely, K 2011, "The KALDI speech recognition toolkit", in Proceedings of ASRU.
7. Yessenbayev, Zh and Karabalayeva, M 2016, "A baseline system for Kazakh broadcast

news transcription", *in Proceedings of the V International scientific-practical conference information society*, pp. 48-50.

8. Kozhirbayev, Zh and Islam, Sh 2016, "A distributed platform for speech recognition research", *in Proceedings of the V International scientific-practical conference information society*, pp. 38-40.

9. Makhambetov, O, Makazhanov, A, Yessenbayev, Zh, Matkarimov, B, Sabyrgaliyev, I and Sharafudinov, A 2013, "Assembling the Kazakh Language Corpus", *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1022–1031.

10. Can, D and Saraclar, M 2011, "Lattice indexing for spoken term detection", *IEEE Trans. Audio Speech Lang. Process*, pp. 2338–2347.

УДК 681.5

## ТАБИГЫЙ ТИЛДЕГИ ТЕКСТТЕРДИ ОРУС ТИЛИНЕН КЫРГЫЗ ТИЛИНЕ МАШИНАЛЫК КОТОРУУДА СӨЗДӨРДҮ АНАЛИЗДӨӨНҮН АЛГОРИТМИН ТҮЗҮҮ

***Кочконбаева Буажар Осмоналиевна,*** *старший преподаватель, ОшТУ им. академика М.М. Адышева, 714018, г. Ош, ул. Исанова 81, e-mail: buajar@mail.ru*

Морфологиялык талдоо жана сөз түзүү процесстери бири бирине жакын болуп эсептелет. Машиналык которуу процессинде бир эле учурда сөздөрдү морфологиялык жактан талдап аны жаңы тилге которуу үчүн сөз жасоого туура келет. Бул процесстер бүгүнкү күндө актуалдуу проблемалардан болуп эсептелет. Анткени бардык эле улут өз тилинин унутулбай автоматташтырылышын жана интернет сайттарына кошулуусу үчүн аракет кылышууда. Ошондуктан мен бул макалада табигый тилдеги тексттерди орус тилинен кыргыз тилине которуучу программаны түзүүнүн алгоритмин карап чыктым.

**Ачкыч сөздөр:** машиналык которуу, морфологиялык анализатор, аффикстер, синтаксистик анализатор,семантикалык анализатор, программа, маалыматтар базасы

## DEVELOPMENT OF ALGORITHM ANALYSIS OF WORDS IN NATURAL TEXT MACHINE TRANSLATION FROM RUSSIAN INTO KYRGYZ

***Kochkonbaeva Buazhar Osmonalievna,*** *senior lecturer, OshTU after named acad. M.M. Adishev, 714018,c. Osh, Isanov st. 81, e-mail: buajar@mail.ru*

Morphological analysis of words and word-formation are very similar in how to conduct and are closely related to each other. Today these processes are very actual. Because all the nations seek to automate their language and add to the online translators. In this article I will consider the algorithm of a program algorithm machine translation of texts from Russian to Kyrgyz language.

**Keywords:** machine translation, Morphological analyzer, affixes, syntactical analyzer, semantic analyzer, program, database.

### Киришүү

Жасалма интеллект багытындагы адам баласынын изилдоолору кун санап өсүп барат. Азыркы күндө компьютер бир гана эсептөө техникасы эмес, ар кандай багыттагы суроолорго жооп берүүчү эксперттик система, бир тилден экинчи тилге тексттерди которуучу каражат же болбосо интеллектуалдык оюндарды ойноодо экинчи тарап катары жасалма интеллекттин ролун жаратып келүүдө. Бул макалада жогоруда айтылган багыттардын ичинен табигый