Kazakh Vowel Recognition at the Beginning of Words[1]

Aigerim K. Buribayeva

Master of Science in Computer Engineering, Lecturer, L. N. Gumilyov Eurasian National University, Astana, Kazakhstan
Email: buribayeva@mail.ru

Altynbek A. Sharipbay

PhD in Technical Sciences, Professor, L. N. Gumilyov Eurasian National University, Astana, Kazakhstan
Email: sharalt@mail.ru

Doi:10.5901/mjss.2015.v6n2s4p121

Abstract

This paper describes the method of recognition of Kazakh vowels at the beginning of the words using Dynamic Time Warping algorithm. This can be used for acceleration of recognition since word's first sound identification can significantly decrease the list of words-candidates during recognition. Also, the acoustic analysis of Kazakh vowels and their transcription during speech recognition are presented.

Keywords: *Dynamic Time Warping algorithm, Kazakh vowels, speech recognition, transcription*

1. Introduction

Automatic recognition of natural language verbal speech is one of important areas of development of artificial intelligence and computer science as a whole, as results in this area will allow to solve the problem of development of man's efficient voice response means with the help of computer. A principal opportunity for transition from formal languages-mediators between man and machine to natural language in verbal form as universal means of expression of man's ideas and wishes has appeared with development of modern voice technologies. Voice input has a number of advantages such as naturalness, promptness, input's notional accuracy, user's hands and vision freedom, possibility of control and processing in extreme conditions.

Literature analysis has shown that currently there is no description of phonetic system of Kazakh language containing acoustic characteristics of sounds. This is necessary for formation of automatic transcription which is an integral part of speech recognition system. In this regard the authors carry out associated work. Thus, for example, complete formant analysis of Kazakh vowels was presented in the work of Yessenbayev, Karabalayeva and Sharipbayev (2012). Also the phonetic transcription converter for speech recognition was developed and Kazakh speech phonemic segmentation algorithms were implemented. However in the course of the work it has been discovered that more detailed transcription is necessary for reliable recognition, which is presented in this article. The marking-out of word's first sound contributes to acceleration of the whole recognition process and decreases the number of words-candidates for recognition.

2. Related Work

An analysis of cues to identify Arabic vowels is provided by Iqbal, Awais, Masud and Shamail (2008). The algorithm for vowel identification has been developed that uses formant frequencies. The algorithm extracts the formants of already segmented recitation audio files and recognizes the vowels on the basis of these extracted formants. Acoustic Analysis was performed on 150 samples of different reciters and a corpus comprising recitation of five experts was used to validate the results. The vowel identification system developed here has shown up to 90% average accuracy on

[1] *The presented work is supported by "Automation of Recognition and Generation of the Kazakh Language Written and Oral Speech" Project implemented under the budget program 120 "Grant Financing of Scientific Researches", specificity 149 "Other Services and Works", by Priority 3. Information and Telecommunication Technologies.*

continuous speech files comprising around 1000 vowels.

Also the recognition of vowels by using formant frequencies in Arabic speech is described in the works of Alotaibi and Hussain (2010). They have developed the recognition system based on HMM and determined experimentally the frequencies of the first and second formant (F1 and F2). The recognition value resulted to be about 91.6%.

A simple method for recognizing the five vowels of the Serbian language in continuous speech presented by Prica and Ili'c (2010). The method they have used is based on recognition of frequencies of first three formants that are present in vowels. By using of LPC method for determining the frequencies and amplitudes of formants in speech, they have set the frequency ranges of formants F1, F2 and F3 for all vowels and defined the areas that vowels occupy in F1-F2-F3 space. When recognition is performed only to vowel speech samples, the average correct recognition rate they have obtained was 83.2%.

Kocharov has developed a system of recognition of vowels in the Russian language, which is based on synchronization with the pitch period. The recognition achieved equalled 87.7% for isolated vowels and 83.93% for the vowels within a word (Kocharov, 2004).

Malaysian spoken vowel recognition by using Autoregressive Models presented in the works of Yusof, Raj and Yaacob (2007). Detection accuracy based on recorded vowels was about 99%.

Kodandaramaiah, Giriprasad and Rao (2010) described the standard approach for the classification of English vowels based on formants. They have shown 90 to 95% of speaker recognition using Euclidian distance measure.

3. Materials and Methods

3.1 Acoustic analysis of Kazakh vowels

The Kazakh language belongs to the Turkic group of languages. There are 9 vowels and 19 consonants in the Kazakh language. In Cyrillic alphabet they are denoted as follows:

Vowels: *А, Ә, Е, О, Ө, Ұ, У, Ы, І,* out of which the vowels *А, О, Ұ, Ы* and *Е* are phonemes, and vowels *Ә, Ө, У, І* are allophones of phonemes *А, О, Ұ, Ы*;

The acoustic analysis of Kazakh vowels is based on the following synharmonic tones:

- *А, Ы* – hard non-labial synharmonic vowels;
- *Ә, І* – soft non-labial synharmonic vowels;
- *Ұ, О* – hard labial synharmonic vowels;
- *У, Ө* – soft labial synharmonic vowels;
- *Е* – soft non-labial synharmonic vowel.

The system of synharmonic features of Kazakh vowels is presented in Table 1. The graphic model of the system of Kazakh vowels can be presented as Figure 1.

The vowels [*а*], [*о*], [*ұ*], [*ы*] on the bottom plane form a group of "hard" (back) vowels, whereas the top vowels [*ə*], [*ө*], [*у*], [*і*] as well as [*е*] form a group of "soft" (front) vowels.

Further, the planes {[*о*], [*ө*], [*ұ*], [*у*]} and {[*а*], [*ы*], [*і*], [*ə*]} reflect the classification of the vowels according to the lip position ("rounded" and "unrounded") as well as the height of a tongue ("high" and "low"). And finally, the planes {[*а*], [*о*], [*ө*], [*ə*]} and {[*ы*], [*ұ*], [*у*], [*і*]} reflect the classification of the vowels according to the jaw position with the first group being the "open" vowels and the second — "close" vowels. The vowel [*е*] was placed on the top position, because it has the lowest F1 and the highest F2.

Geometrically, with respect to the Fig. 1, this can be interpreted as follows. No three vowels of the bottom or top plane can occur in the same word, but only those that are on the edges or single vertices (including [*е*]). Any word having three distinct letters contains the vowels that lie on the plane passing through the vertex *[е*] and any two vertices on the top plane (Yessenbayev, Karabalayeva & Sharipbayev, 2012).

Table 1. The system of synharmonic features of vowels

Vowels	Palatal tone		Labial tone	
	hard	soft	non-labial	labial
А	+	-	+	-
Ә	-	+	+	-
Ы	+	-	+	-
І	-	+	+	-
Ұ	+	-	-	+
Ү	-	+	-	+
О	+	-	-	+
Ө	-	+	-	+
Е		+	+	

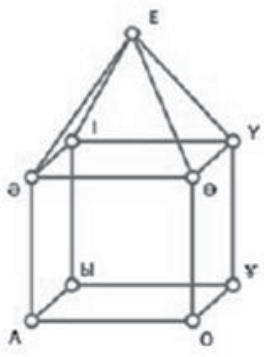

Figure 1. System of Kazakh Vowels

3.2 Kazakh Vowels Transcription

The phonetic transcription converter is performed as a software program replacing some symbols by other symbols according to the following substitution rules included in the control file:

1. *#е=ье, #о=уо, #ө=уө;*
2. *#л=ыл, #р=ыр, #l=il, #r=ir;*
3. *о^ы=о^ұ, ұ^ы=ұ^ұ, ө^і=ө^ү, ү^і=ү^ү,*
4. *ө^е=ө^ө, ү^е=ү^ө;*
5. *і^а=і^ә, ү^а=ү^ә, ө^а=ө^ә, ү^е=ү^ө;*
6. *а^у=а^ұу, о^у=о^ұу, ұ^у=ұ^ұу, ы^у=ы^ұу, ә^у=ә^үу, ө^у=ө^үу, ү^у=ү^үу, і^у=і^үу, у^а=ұу^а, у^о=ұу^о, у^ұ=ұу^ұ, у^ы=ұу^ы, у^ә=үу^ә, у^ө=үу^ө, у^ү=үу^ү, у^і=үу^і;*
7. *а^и=а^ый, о^и=о^ый, ұ^и=ұ^ый, ы^и=ы^ый, ә^и=ә^ій, ө^и=ө^ій, ү^и=ү^ій, і^и=і^ій, и^а=ый^а, и^о=ый^о, и^ұ=ый^ұ, и^ы=ый^ы, и^ә=ій^ә, и^ө=ій^ө, и^ү=ій^ү, и^і=ій^і; қи=қый, ғи=ғый, иқ=ыйқ, иғ=ыйғ.*

Each substitution rule consists of two parts separated from each other by symbol "=". To the left of this symbol are primary symbols of word's letter record, to the right – symbols by which they should be replaced in the transcription.

For transcription of specified word sequential search of inclusion of the left part of regular rule is looked for, and if such is present, then it is replaced by the right part of the given rule.

As transcription signs for vowels, mainly corresponding Kazakh letters are used. Hard Kazakh consonants are also transcribed by Kazakh letters, and corresponding soft consonants are transcribed by similar Latin letters.

Symbol "#" denotes beginning of a word or end of a word depending on its location: if "#" is in front of symbols, then it is the beginning of a word; if "#" is after symbols, then it is end of a word.

Symbol "^" denotes any symbols in any quantity between two sounds.

For the reader's convenience rules presented in this article are divided into groups which are numerated. It is recommended to include these groups in the order of numbers into the control file without changing the order of rules in

groups as it is obvious that the substitution order is important.

1. If a word in the Kazakh language begins with vowel "*е*", sound "*й*" is heard before it during pronunciation, if a word begins with vowels "*о*", "*ө*", short insertion of "*у*" is formed before them during pronunciation, for example, "*ет*" – "*йет*", "*он*" – "*уон'*, "*өнер*" – "*уөнер*".
2. If a word begins with consonants "*р*" or "*л*", vowels "*ы*", "*і*" are heard before these sounds during pronunciation, depending on hardness or softness of consonants "*r"*, "*l*" here mean soft analogues of "*р*" and "*л*", for example, "*рас*" – "*ырас*", "*рет*" – "*ірет*", "*лас*" – "*ылас*", "*лезде*" – "*ілезде*".
3. Vowels *"ұ", "ү", "о", "ө"* at the beginning of a word or word's first syllable change vowels *"ы", "і"* to vowels "*ұ", "ү*" accordingly in the following syllables during pronunciation. For example, "*қолтық*" – "*қолтұқ*", "*құлын*" – "*құлұн*", "*күлкі*" – "*күлкү*", "*көлік*" – "*көлүк*";
4. Vowels "*ү", "ө"* at the beginning of word or word's first syllable change vowel "*е*" to vowel "*ө*" in the following syllables during pronunciation, for example, "*үлкен*" – "*үлкөн*", "*өнер*" - "*өнөр*".
5. Vowels "*ә*", "*ү*", "*і*" at the beginning of a word or word's first syllable change vowel "*а*" to its allophone "*ə"* in the following syllables during pronunciation, for example, "*ләззат*" – "*ләззәт*", "*діндар*" – "*діндәр*".
6. During pronunciation of diphthong "*у*" in the word composition sounds "*ұу*", "*үу*" are heard depending on hardness or softness of vowels in the rest syllables. For example, "*туыс*" – "*тұуыс*", "*күту*" – "*күтүу*".
7. During pronunciation of diphthong "*и*" in the word composition sounds "*ый*", "*ій*" are heard, depending on hardness of softness of vowels in the rest syllables. For example, "*ине*" – "*ійне*", "*жина*" – "*жыйна*". If before or after "*и*" there are consonants "*қ", "ғ*", sound "*ый*" is always heard during pronunciation of sound "*и*". For example, "*қиын*" – "*қыйын*", "*қиғаш*" – "*қыйғаш*".

The given phonetic transcription converter can be used for speech synthesis as well.

3.3 Recognition of Vowels at Word Beginning

3.3.1 Determination of speech beginning

At first it is necessary to make a reliable determination of speech beginning. We have used the algorithm of V.Y. Shelepov (Shelepov & Nitsenko, 2012). Here is the brief description of this algorithm.

There is an 8-bit record with frequency of 22,050 Hz used. When the record button is pressed sequential sound sections by 300 counts (windows) are recorded. For each of them the relation of V/C is calculated where

$$V = \sum_{i=0}^{298} |x_{i+1} - x_i|$$

is numerical analogue of full variation, C is the number of constancy points i.e. such moments of time when the signal value remains the same at the next moment. By first ten windows an average of this relation is taken. We will name this value as "current StartPorog". It characterizes the upper threshold of "silence". Then we wait for the moment when this threshold is exceeded by not less than 5 times. Then we go back by 20 windows (initial stock) and, beginning from this moment we will enter recorded counts into buffer 1. Thereby the record of what we suppose to be speech starts. We will determine "current EndPorog" as fivefold current StartPorog. The filling of buffer 1 continues up to the moment after which values V/C in the course of 10 thousand counts will be less than current EndPorog. The mentioned 10 thousand counts (end stock) are also entered into it. Thus, the record of supposed speech section stops. We will note that during each record new values of "current StartPorog" and "current EndPorog" are calculated.

The recorded data is checked for speech presence by means of quasi-periodicity (Shelepov, 2009). If speech presence is revealed, the contents of buffer 1 are transferred to buffer 2.

3.3.2 Selection of features for recognition by model

We will consider the system of features used by us at DTW recognition (Dynamic Time Warping) (Shelepov, 2009). There are 10 thousand numbers entered into the corresponding buffer:

$$y_1, y_2, \ldots, y_{10000} \quad (1)$$

of tension values at microphone output in sequential time moments (We will name these time moments as counts). The row of numbers itself (1) and corresponding function

$$y(i) = y_i \quad (2)$$

will be named by us as a signal. Thus, the numbers (1), in the final analysis, reflect the change of pressure on

microphone membrane as a time function. The signal graph as a time function can be displayed on the monitor (signal visualization).

As signal smoothing we name signal processing by three-point sliding filter

$$y_i = \frac{y_{i-1} + y_i + y_{i+1}}{3}, \qquad i = 2,3,...,9999 \tag{3}$$

Further work is carried out with pointwise difference of primary and tenfold-smoothed signal. This allows to "purify" it to some extent from the speaker's individual tone and thus advance toward speaker independent recognition system. Further except as otherwise stipulated under the signal we understand the specified difference and in order to avoid complicated denotations we will consider that (1) and (2) correspond exactly to it.

Let l be the number of counts between two adjacent local maximums of function (2) (we will name function contraction to corresponding interval as full oscillation). If maximums are not strict, then under l we understand the number of counts from first maximum beginning to second maximum beginning. We will define value z:

$z=l$ at $2 \leq l<20$; $z=20+\frac{l-20}{6}$ at $20 \leq l<50$;

$z=25+\frac{l-50}{10}$ at $50 \leq l<90$; $z=29$ at $l \geq 90$.

We will name the nearest integral number not exceeding z as length of corresponding full oscillation. In this way the length of full oscillation is considered the more accurately the shorter the oscillation is. We will mark out signal section and specify through n the total number of full oscillations at the given section, through n_1 - the number of full oscillations of length 2,..., through n_{28} - the number of full oscillations of length 29.

We will assign the vector to the specified section

$$(x_1,...,x_{28},\varepsilon) \tag{4}$$

where $x_k = n_k/n$, $k = 1,2,...28$, ε is amplitude relation (difference of largest and smallest values) of considered signal section to the whole signal amplitude. The value ε is entered in order to separate reliably the pause from significant signal part and its normalization is carried out in order to abstract from loudness of the pronounced.

We will divide the recorded signal making 10 thousand counts into sections making 368 counts each (duplicated quasi-period of major tone for male voice of average pitch). For each of 27 full sections we will calculate the vector (4). The last non-full section will be simply discarded by us. As a result we present the signal in the form of path, i.e. sequence of 27 points in 29-dimensional space:

$$A = (a_1, a_2, ..., a_{27}).$$

The presentation of the whole word is described above, and three first vectors are enough for the vowel definition at the beginning of word, i.e.:

$$A = (a_1, a_2, a_3).$$

Further, for recognition we apply the algorithm of T.K.Vintsyuk which has already become classical, also known as DTW algorithm (Vintsyuk, 1987).

4. Results

17 words for each sound with different sound combinations were selected for testing. For example, for sound "***ө***" words: "***ө**кше*", "***ө**нер*", "***ө**гей*", "***ө**ширгіш*", "***ө**шпес*", "***ө**зен*", "***ө**збек*", "***ө**рік*", "***ө**ркеш*", "***ө**лке*", "***ө**жет*", "***ө**сек*", "***ө**сиет*", "***ө**мір*", "***ө**тірік*", "***ө**тем*", "***ө**бектеу*" were selected. In all versions except for word "***ө**гей*" the program reliably recognized the sound "***ө***".

On Figures 2 and 3 the visualization of word "***ө**кше*" and recognition of sound "***ө***" at its beginning are presented.

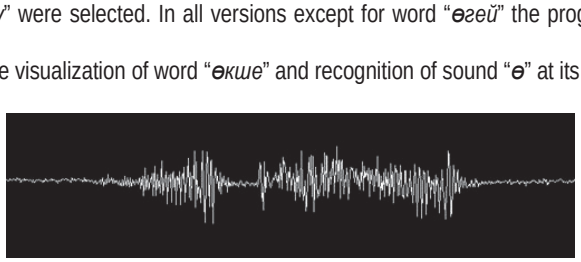

Figure 2. The visualization of word "***ө**кше*"

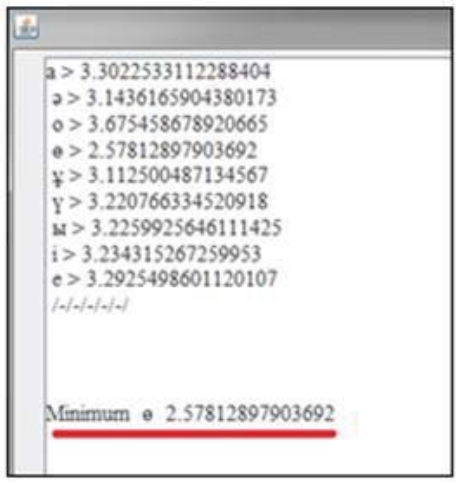

Figure 3. Recognition of sound "*ө*"

The complete description of the experiment's result is presented in Table 2. As you can see from the table, sounds "*a*" and "*e*" are recognized without mistakes, during recognition of sounds "*ұ*", "*o*", "*ө*" the program was wrong once for each of them. With sounds "*ə*", "*ү*", "*ұ*" it's worse as the program mixes up sound "*ə*" with sound "*a*", and sounds "*ұ*", "*ы*", "*i*" among themselves. More reliable recognition can be achieved by means of additional training as during each training sound models are averaged.

Table 2. Experiment results

Sound	Recognition
A	100%
Ə	88,23%
Ы	88,23%
І	88,23%
Ұ	88,23%
Ү	94,11%
О (yo)	94,11%
Ө (yө)	94,11%
Е (йе)	100%

5. Conclusion and Perspectives

The authors of this given work received the following results:

- Realization of acoustic analysis of Kazakh vowels;
- development and performance of phonetic transcription converter of Kazakh vowels for speech recognition as a software program;
- development and performance of algorithm of vowel recognition at the beginning of word as a software program.
- further it is planned to realize:
- the phonetic transcription converter's expansion and adaptation for continuous speech;
- development and performance of algorithm of consonant recognition at the beginning of word as a software program;
- development of algorithms and software programs for recognition of continuous Kazakh speech on the basis of interphoneme transitions.

References

Alotaibi, A., & Hussain, A. (2010). Comparative Analysis of Arabic Vowels using Formants and an Automatic Speech Recognition System International. *Journal of Signal Processing, Image Processing and Pattern Recognition, 3 (2).*

Iqbal, H., Awais, M., Masud, S. & Shamail, S. (2008). On vowels segmentation and identification using formant transitions in continuous recitation of quranic Arabic. *New Challenges in Applied Intelligence Technologies, ser. Studies in Computational Intelligence, 134,* 155–162.

Kocharov, D. A. (September, 2004). *Automatic vowel recognition in fluent speech.* In Proceedings of the 9th Conference of Speech and Computer. St. Petersburg, Russia.

Kodandaramaiah, G. N., Giriprasad, M. N. & Rao, M. M. (2010). Independent speaker recognition for native English vowels. *International Journal of Electronic Engineering Research, 2.*

Prica, B., Ili´c, S. (2010). Recognition of Vowels in Continuous Speech by Using Formants. *Facta universitatis, 23 (3),* 379-393.

Shelepov, V.Y. (2009). *Lectures on speech recognition*. Donetsk, Ukraine: IPSHI Nauka i osvita.

Shelepov, V.Y. & Nitsenko, A.V. (2012). A new approach to the definition of the boundaries of the speech signal. Problems of the signal end. *Speech technology, 1,* 74-79.

Vintsyuk, T.K. (1987). *Analysis, recognition and interpretation of speech signals*. Kiev, Ukraine: Naukova dumka.

Yessenbayev, Zh., Karabalayeva, M. & Sharipbayev, A. (2012). Formant Analysis and Mathematical Model of Kazakh Vowels. In *14th International Conference on Computer Modelling and Simulation (UKSim)*, London, pp. 427-431.

Yusof, S. A. M., Raj, P. M. & Yaacob, S. (June, 2007). Speech recognition application based on Malaysian spoken vowels using autoregressive model of the vocal tract. In *Proceedings of the International Conference on Electrical Engineering and Informatics.* Bandung, Indonesia: Institut Teknologi Bandung.