IRSTI 28.23.24

Sh. Rashid[1], D. Kuanyshbay[2], A. Nurkey[3]
[1,2,3]Suleyman Demirel University, Kaskelen, Kazakhstan

DEVELOPMENT OF METHODS AND ALGORITHMS TO BUILD A SPEAKER VERIFICATION IN KAZAKH LANGUAGE

Abstract. Speaker verification interfaces are gaining more and more popularity in both academic and commercial industries. It's connected with the latest advances in this area, which can be seen firsthand in our daily life: voice interfaces in computers, robots, cell phones, Internet browsers, and even household appliances. The relevance of the development of Kazakh speech recognition systems arises in connection with the growing needs in the field of public services, provided within the framework of electronic government (e-gov). Availability voice interfaces will open access to government services for people with disabilities, as well as people living in remote regions and having the only way of access in the form of telephones.

Keywords: speech verification, recognition system, processing signals, interference.

***

Аннотация. Интерфейсы проверки говорящего становятся все более популярными как в академической, так и в коммерческой сферах. Это связано с последними достижениями в этой области, которые можно увидеть воочию в нашей повседневной жизни: голосовые интерфейсы в компьютерах, роботах, сотовых телефонах, интернет-браузерах и даже бытовой технике. Актуальность развития казахстанских систем распознавания речи возникает в связи с растущими потребностями в сфере государственных услуг, предоставляемых в рамках электронного правительства (e-gov). Наличие голосовых интерфейсов откроет доступ к государственным услугам для людей с ограниченными возможностями, а также для людей, живущих в отдаленных регионах и имеющих единственный способ доступа в виде телефонов.

Ключевые слова: проверка речи, система распознавания, обработка сигналов, помехи.

***

Аңдатпа. Динамиктерді тексеру интерфейстері академиялық және коммерциялық салаларда танымал болып келеді. Бұл күнделікті өмірде өз көзімізбен көретін осы саладағы соңғы жетістіктермен байланысты: компьютерлердегі, интерактивті роботтардағы, ұялы телефондардағы, интернеттегі браузерлердегі, тіпті тұрмыстық техникадағы дауыстық интерфейстер. Қазақша сөйлеуді тану жүйелерін дамытудың өзектілігі

электронды үкімет (e-gov) шеңберінде көрсетілетін мемлекеттік қызметтер саласындағы қажеттіліктердің өсуіне байланысты туындайды. Дауыстық интерфейстердің болуы мүмкіндігі шектеулі жандарға, сондай-ақ шалғай өңірлерде тұратын және телефон түрінде қол жетімді жалғыз мүмкіндігі бар адамдарға арналған мемлекеттік қызметтерге қол жетімділікті ашады.

Түйін сөздер: сөйлеуді тексеру, тану жүйесі, сигналдарды өңдеу, интерференциялар.

Introduction

Today's automatic speech verification systems have stepped up forward at a rapid pace over the past decades, from simple speaker-dependent applications to speaker-independent automatic systems transcriptions of news, telephone conversations, lectures, etc. Despite the widespread use of such systems, the task of speech recognition is far from solved in view of problems related to noise, distortion on the line, foreign accent, speed and manner of speech, etc. However, research on speech recognition is being actively pursued in the light of recent achievements, and there is a huge amount of literature on this topic. To date, a unified approach to the creation of systems has formed continuous speech recognition, which includes four main components: signal preprocessing, acoustic model, language model and hypothesis search. Despite this, there is ongoing research on each of the components. Signal preprocessing requires filtering and extraction of acoustic parameters resistant to noise and other distortions of the speech signal. Existing Chalk Frequency Cepstral Coefficients (MFCC), the coefficients of perceptual linear predictions (PLP) or wavelet coefficients do not provide the proper recognition quality even when noise level of 10 dB. Hidden Markov Models with Mixtures of Normal Distributions of Steel the standard when creating an acoustic model for speech recognition systems.

Language models based on n-gram have also long been an indispensable tool in speech recognition systems, but their attachment to specific topics of speech and the inability to describe long-term language phenomena have become the reason for the search for other methods and algorithms for describing structure of the language.

Search for the most plausible hypotheses for a given acoustic the signal is conducted using weighted end transducers, which are able to combine and process various sources of information in speech recognition systems such as acoustic and language models, topology of hidden Markov models, etc. However, this in turn is also a disadvantage of such systems, since the combination of such a large amount of information (thousands of HMM states, hundreds of thousands of words and n-grams in language model) requires, in turn, a huge time (up to 35 hours and more) to create the necessary description of the transducer network.

Background of Literature Review

The best result in modern speaker-independent systems of speech recognition for the task of recognizing telephone conversations is in English language which achieved at 15.2% (word error rate, WER), for the task of transcribing news in Arabic - 7.4% (level word errors, WER), for transcribing news in Chinese - 6.2% (character error rate, CER).

Research in the field of automatic recognition of Kazakh speech was conducted relatively recently. The first works appeared only in 2006-2007 years. In which the author limited himself to phonemic recognition of individual words. Unfortunately, the author did not provide information on the quality of phoneme and word recognition. On the other hand, a word recognition algorithm based on interdiphone transitions with the quality of word recognition is up to 91%. There were approaches to isolating speech areas in the signal for further use in the recognition process Kazakh speech. In 2009, there were attempts to develop commercial systems recognition of continuous Kazakh speech. However, in the literature there is no detailed information about the applied methods and algorithms.

Aim and objectives of research

The purpose of this thesis is a comprehensive research and solution of the problem of automatic speaker-independent recognition of Kazakh speech according to a certain vocabulary. To achieve this goal, it is necessary to decide the following theoretical and practical tasks:

- development of methods and algorithms for speaker-independent recognition Kazakh speech;
- development of methods and algorithms for recognizing Kazakh speech in noise conditions;
- software implementation of the application system using developed methods and algorithms for recognizing Kazakh speech.

The object of research is the process of automatic verification Kazakh speech according to a certain vocabulary.

Methods and Materials

The development and implementation of speech recognition systems in real conditions is associated with certain difficulties. The main problem is the presence of various kinds of noise arising from the specifics of the environment, channels and communication devices. Also, completely different people can use the system. by gender, age and regional characteristics, which requires consideration of all kinds of speech variations arising from a large circle of users. All this directly affects the quality of recognition.

This work is dedicated to solving these problems and is unique in its kind, since currently there are no examples of practical use in Kazakhstan speech recognition systems, as well as there are no scientific developments in the field recognition of continuous Kazakh speech in really noisy conditions. Worth note, however, that there have been attempts at development by commercial

companies, which were subsequently suspended due to the lack of scientific base and proper funding.

Language model - permits to decide the highest probable word sequences. From that unpredictability to build of this model relies to a great extent upon the particular language. For the english words, they are sufficient to utilize measurable models. On the other hand for agglutinative dialects by a moderately rich network, the factual model isn't appropriate and half breed model is utilized.

The speech model s(z) allows the earlier likelihood by the word grouping w. Fundamentally it demonstrates how acceptable a word arrangement should be articulated, in view of syntactic standards of a language. Considering this speech model just relies upon content and an autonomous of audio information, accordingly huge measures of words accessible on the albums, diary, papers and so on, may be utilized like an info origin. Furthermore, we need the speech model to catch the subject explicit data to exceptional Automatic Speech Recognition frameworks. For catching several attributes related with human discourse for example certain linguistic blunders regular in talking, redundancies, waverings and so on, records of spoken content are likewise a helpful information origin. After all the absolute quantity of conceivable speech successions is limitless, disentangling suspicions should be made to get dependable incomplete appraisals. The basic step to compute the speech model anticipations is over collection based on tallies like neighboring words. It accepts that the likelihood of the current speech n relies just upon the past m - 1 word n - 1 …. n - m + 1.

Speech verification includes various segments, for example, highlight expression, audio displaying, language demonstrating and Deep Neural Networks, that appeared from Figure 1.

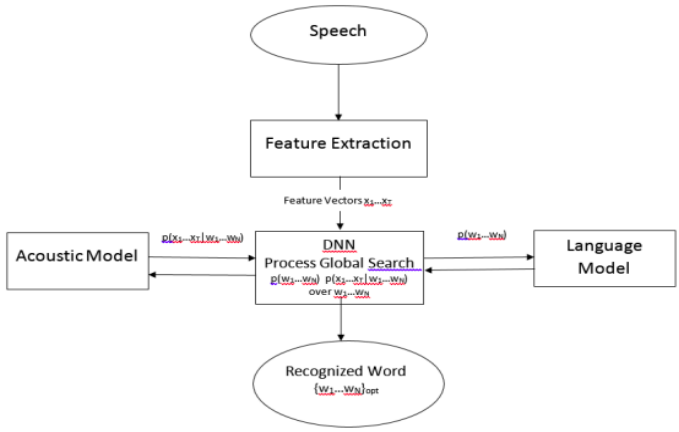

Figure 1. Diagram of an ASR framework.

$v^l = f(z^l) = f(W^l v^{l-1} + b^l)$, for $0 < l < L$

where $z^1 = W^1 v^{l-1} + b^1 \in R^{N_1+1}, W^1 \in R^{N_1 * N_1 - 1}, b^1 \in R^{N_1+1}$ and $N_1 \in R$ separately, the excitation vector, the enactment vector, weight framework, removal vectors and the quantity of neurons in layer 1 $v_0 = 0 \in R$, $v_0 = 0 \in R$, R^{N_0+1} observation vector, $N_0 = D$ that is the element size and $f(\cdot): R^{N_1+1} \to R^{N_1+1}$. Activation function in relation to the output vector element wise. In a lot of examples, the Sigmoid function is used $\sigma(z) = \frac{1}{1+e^{-z}}$,

as an Activation function. Then, We build the algorithm for our model.

Algorithm direct Deep Neural Network estimating.

1. start(F) -> Each column F is a measurement vector;
2. $V^0 \leftarrow$ F;
3. For $l \leftarrow 1$; $l < L$; $l \leftarrow l + 1 > l$ is the sum quantity of our hidden layer;
4. $Y^1 \leftarrow W^1 V^1 - 1 + B^1 >$ Each column B^1 is bl;
5. $V^1 \leftarrow f(Z^1) > f(\cdot)$ can be sigmoid, ReLU or other activation functions;
6. End for;
7. $Y^1 \leftarrow W^1 V^{1-1} + B^1 >$ one more iteration;
8. if regression > regression part;
9. $V^1 \leftarrow Y^1$;
10. else {;
11. $V^1 \leftarrow softmax(Y^1) >$ function softmax;
12. End if;
13. return $V^1 >$ return of our V;
14. End procedure > end of our procedure.

During preparing, we utilize the calculation of single-stage determination by the Monroe-Carlo strategy in the Markov chain. RBM have Gauss-Bernoulli units and are prepared at an underlying learning pace of 0.009 units. Preparing was not monitored, the number of emphasis was 4, the quantity of hidden layers taken as 6, the quantity of units in this hidden layer was up to 1024.

Research results

Over the span of this research, highlight expression strategies, for example, audio, speech model, Deep Neural Networks were examined. The outcomes were assessed by the speech mistake percentage for old style models. Outcomes demonstrating vertical hubs are rate, and the even pivot: preparing

mono models , the entry of the principal and second and third thyryphon. The top outcome was 35.35% Word error rate. The outcomes were by the utilization of Deep Neural Networks utilizing 0 to 5 shrouded(hidden) layers. The ideal aftereffect of 30.06% Word error rate was acquired for 5 shrouded(hidden) layers, and it was a development through the old style model.

Note that the result is increased when we make the volume of the corpus to prepare high. The best outcomes have been obtained utilizing Deep Neural Networks calculation.

Conclusion

According to the results of scientific research, at present, universal speech recognition systems do not have a sufficiently high recognition accuracy at the human level. We have formulated the main criteria of existing disadvantages of speech recognition systems. Optimal architecture of building such systems, to overcome the indicated disadvantages, including among those currently used by leading IT companies. Features described the use of multi-tier architectures in speech recognition and segmentation systems. Practically implemented and considered a system for recognizing speech commands of the Kazakh language based on DNN technology.

References

1 Amirgaliev, E.N., Musabaev, R.R., Musabaev, T.R. The speech signal segmentation algorithm using pitch synchronous analysis. *Open Comput. Sci.* 7 (2017): pp. 1–8.
2 Bazhenova, I.Yu. Delphi 7 Programmer's tutorial. 2003.349 p.
3 Glushkov, V.M., Amosov, N.M., Artemenko, I.A. Encyclopedia of Cybernetics. 1974. – 590 p.
4 Anusuya, M.A., Katti, S.K. Speech Recognition by Machine: A Review. *(IJCSIS) International Journal of Computer Science*, 6 (3), (2009): pp. 181-205.
5 Big data what is Hadoop. URL: //http://singhvikash.blogspot.co.uk/2013/12/big-data-what-is hadoop.html
6 Nguyen ,Zh. A distributed platform for parallel training of disann artificial neural networks. *International Journal of software products and systems*, 3 (2013): pp. 99-103.
7 What is Siris architecture in Apple's data center. URL: https://www.quora.com/What-is-Siris-architecture-in-Apples-data-center.