УДК 004.7

DEVELOPING SPEECH RECOGNITION APPLICATION IN KAZAKH LANGUAGE

Aitimov A.K. MSc student, Amirgaliyev Y. N. Prof
Suleyman Demirel University

Түйін

Бұл мақаланың басты мақсаты Қазақ тіліне арналған сөйлем тану бағдарламасын құру. Мақалада сөйлем тану дегеніміз не және оның құрылысы баяндалған. Сонымен қатар, бағдарламаның бұйрықтары мен әрекеттері және тексеру көрсеткіштері де қамтылған.

Резюме

Эта статья предназначена для создания программы для распознавания речи Казахского языка. Здесь рассмотрено что представляет собой APP и его архитектура. А также включены команды, действия и результаты тестов.

Abstract. *Speech Recognition (is also known as Automatic Speech Recognition (ASR), or computer speech recognition) is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program. This paper is aimed to create speech recognition application in Kazakh language. The paper includes what is ASR and its architecture. Also, commands and actions of application, and result of tests.*

Key words: ASR, Kazakh language, commands, grammar, speech recognition application.

Introduction

Speech Recognition

Research in speech processing and communication for the most part, was motivated by people desire to build mechanical models to emulate human verbal communication capabilities. Speech is the most natural form of human communication and speech processing has been one of the most exciting areas of the signal processing. Speech recognition technology has made it possible for computer to follow human voice commands and understand human languages. The main goal of speech recognition area is to develop techniques and systems for speech input to machine. Speech is the primary means of communication between humans. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to desire to automate simple tasks which necessitates human machine interactions and research in automatic speech recognition by machines has attracted a great deal of attention for sixty years [1]. Based on major advances in statistical modeling of speech, automatic speech recognition systems today find widespread application in tasks that require human machine interface, such as automatic call processing in telephone networks, and query based information

systems that provide updated travel information, stock price quotations, weather reports, Data entry, voice dictation, access to information: travel, banking, Commands, Automobile portal, speech transcription, railway reservations etc.

History of ASR Technology

The earliest attempts to devise systems for automatic speech recognition by machine were made in the 1950s. Much of the early research leading to the development of speech activation and recognition technology was funded by the National Science Foundation (NSF) and the Defense Department's Defense Advanced Research Projects Agency (DARPA). Much of the initial research, performed with NSA and NSF funding, was conducted in the 1980s.

Speech recognition technology was designed initially for individuals in the disability community. For example, voice recognition can help people with musculoskeletal disabilities caused by multiple sclerosis, cerebral palsy, or arthritis achieve maximum productivity on computers.

During the early 1990s, tremendous market opportunities emerged for speech recognition computer technology. The early versions of these products were clunky and hard to use. The early language recognition systems had to make compromises: they were "tuned" to be dependent on a particular speaker, or had small vocabulary, or used a very stylized and rigid syntax. However, in the computer industry, nothing stays the same for very long and by the end of the 1990s there was a whole new crop of commercial speech recognition software packages that were easier to use and more effective than their predecessors [2].

Architecture of ASR

An ASR contains three knowledge sources: acoustic models, a dictionary and a language model as shown in figure. These independent knowledge sources also called ASR database, are subjects to adapt to fulfil natural variations occurring in speech signals. Most research works, focuses on the dictionary adaptation. Regardless of the adaptation level, a high integration among the ASR components is required to achieve better performance.

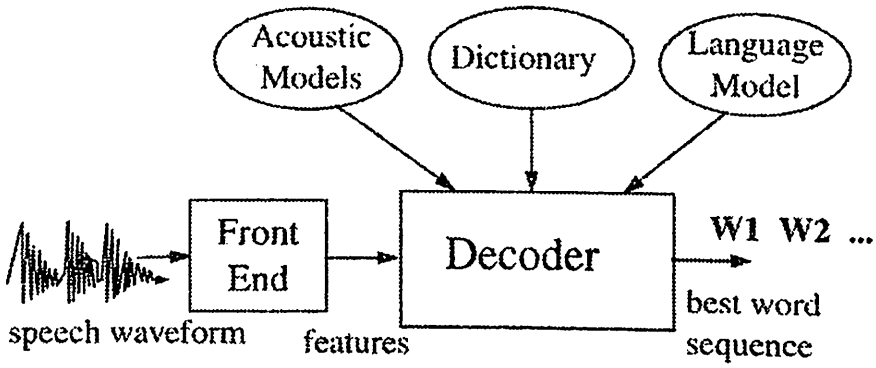

Figure 1.

Acoustic models

Acoustic modeling of speech typically refers to the process of establishing statistical representations for the feature vector sequences computed from the speech waveform. Hidden Markov Model (HMM) is one most common type of acoustic models. Other acosutic models include segmental models, super-segmental models (including hidden dynamic models), neural networks, maximum entropy models, and (hidden) conditional random fields, etc.

Acoustic modeling also encompasses "pronunciation modeling", which describes how a sequence or multi-sequences of fundamental speech units (such as phones or phonetic feature) are used to represent larger speech units such as words or phrases which are the object of speech recognition. Acoustic modeling may also include the use of feeback information from the recognizer to reshape the feature vectors of speech in achieving noise robustness in speech recognition.

Dictionary

While the built-in language model of a recognizer is intended to represent a comprehensive language domain (such as everyday spoken English), a speech application will often need to process only certain utterances that have particular semantic meaning to that application. Rather than using the general purpose language model, an application should use a grammar that constrains the recognizer to listen only for speech that is meaningful to the application. This provides the following benefits:

- Increases the accuracy of recognition
- Guarantees that all recognition results are meaningful to the application
- Enables the recognition engine to specify the semantic values inherent in the recognized text

Language Model

Language models assign weights to word sequences to discriminate between acoustically similar sequences. Discriminative training of language models has been shown to improve the speech recognition accuracy by resolving acoustic confusions more effectively [3]. In discriminative language modeling (DLM), a speech recognizer is employed to generate a set of competing hypotheses for an utterance. Given the correct transcription of an utterance and the set of competing hypotheses (confusion set), discriminative learning techniques can be applied to make use of positive and negative examples to reward features in the correct transcription and penalize features in the competing hypotheses. However, this approach requires a large amount of transcribed speech data. Several approaches have been proposed to overcome the necessity of supervised learning for DLM. For instance, Xu et al. propose a self-supervised discriminative training method, in which an exponential language model is trained using only untranscribed speech and a large text corpus [4]. First, cohorts for words w are determined from the first-pass ASR output lattices for untranscribed speech utterances. The discriminative training is based on maximization of the likelihood ratio between the words w in the text corpus and their cohorts. In another work, Kurata et al. propose to generate the probable n-best lists that an ASR system may possibly output, for an input hypothetical utterance given a word sequence [5]. They call this process PseudoASR since they use phoneme similarities estimated from an acoustic model to generate the competing hypotheses. The discriminative training of the model is based on the generalized probabilistic descent (GPD) algorithm and more recently they applied discriminative reranking using the perceptron algorithm [6]. In another study, Tan et al. propose a system for channel modeling of ASR for simulating the ASR corruption using a phrase-based machine translation system trained between the reference phoneme and output phoneme sequences from a phoneme recognizer. Jyothi et al. have also modeled the phonetic confusions using a confusion matrix that takes into account word-based phone confusion log likelihoods and distances between the phonetic acoustic models. The confusion matrix is used to generate confusable word graphs for training a discriminative language model using the perceptron algorithm.

Application

Grammar

A grammar contains a set of rules that specify the words, phrases, or commands that a user can speak to a application. These commands are available for recognition by the application. XML elements is used to create the rules that identify the words or phrases that comprise these commands. The rules used to identify commands are represented by rule elements. In addition to identifying commands, rules provide

structure to the commands by allowing encapsulation of semantically related words or phrases. This structure provides logical organization of user commands, and also allows for reuse of rules within containing or external grammars.

List of Commands

To Do This	Say This
Starts listening	*Қосыл*
Display power(charge) of computer	*Қалайсың*
Display current time	*Уақыт*
Display today's weather	*Ауа райы*
Open web-site "nur.kz"	*Жаңалықтар*
Open Visual Studio application	*Жұмыс*
Open selected file or folder	*Аш*
Close current window	*Жап*
Open web-site "fapl.ru"	*Футбол*
Stop listening	*Өш*

Speech Recognition Training

Before start using application train speech recognition so it can better understand you. This isn't mandatory, but some training will allow the computer to better understand your voice. This training feature is clever. It will have you read a speech recognition tutorial aloud, so both you and the computer will be learning at the same time. The speech recognition feature will improve over time as it learns more about your voice.

Test and evaluation

To test the application 12 men and 16 woman were participated. First, half of users test application without training. Then, other users test after training. Expected that results will almost same, because there are just ten commands and they sounds different. But, result was unbelievable as shown in figure 2.

Result of the first test:

- Application recognized 4/14 users command [*Аш, Жап, Өш*]
- Application recognized 2/14 users command [*Уақыт, Ауа райы, Футбол*]
- Application recognized 1/14 users command [*Қосыл*]
- Application recognized totally just 13.55%

Result of the second test:

- Application recognized 14/14 users command [*Аш, Жап, Өш, Футбол, Уақыт, Ауа райы*]
- Application recognized 12/14 users command [*Қосыл*]

- Application recognized 9/14 users command [*Қалайсын*]
- Application recognized 5/14 users command [*Жұмыс*]
- Application recognized 4/14 users command [*Жаңалықтар*]
- Application recognized totally just 81.41%.

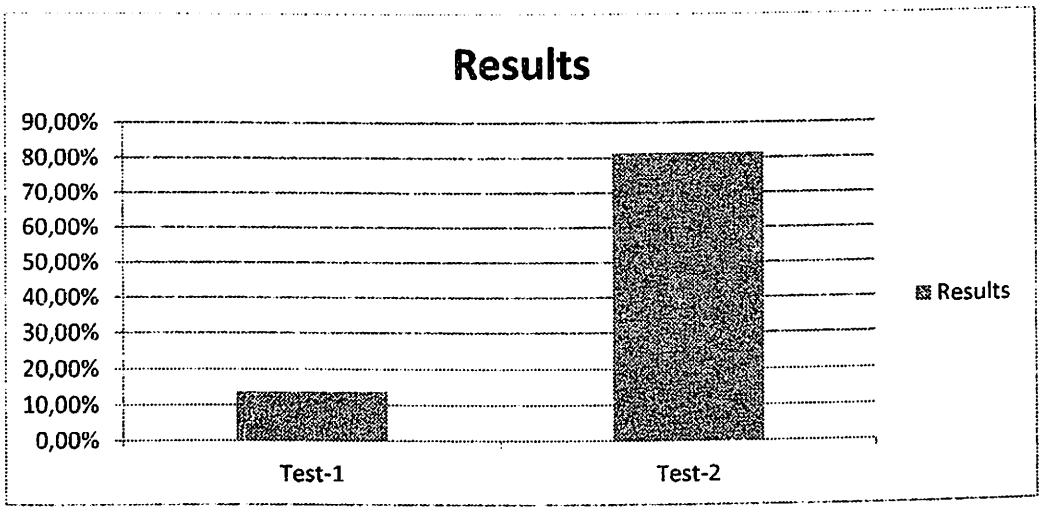

Figure 2. Test results

Conclusion

In this paper we present speech recognition application which consist of three independent sources. The application was tested by 28 users. Half of the users tested without training and others after training. Result showed that the efficiency of application 6 times greater when user has training.

References

[1]. Dat Tat Tran, Fuzzy Approaches to Speech and SpeakerRecognition, A thesis submitted for the degree of Doctor of Philosophy of the university of Canberra.

[2]. Global Security.Org

[3] B. Roark, M. Sarac¸lar, and M. Collins, "Discriminative n-gram language modeling," Computer Speech and Language, vol. 21, no. 2, pp. 373–392, April 2007.

[4] P. Xu, D. Karakos, and S. Khudanpur, "Self-supervised discriminative training of statistical language models," in ASRU, 2009, pp. 317–322.

[5] G. Kurata, N. Itoh, and M. Nishimura, "Acoustically discriminative training for language models," in ICASSP, 2009, pp. 4717–4720.

[6] G. Kurata, N. Itoh, and M. Nishimura, "Training of errorcorrective model for ASR without using audio data," in ICASSP, 2011, pp. 5576–5579