International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

Karakalpak Speech Recognition with CMU Sphinx

Mamatov Narzillo, Samijonov Abdurashid, Nurimov Parakhat, Niyozmatova Nilufar

Abstract: ***One of the main problems of speech recognition systems is the diversity of natural languages. Most of the existing systems can recognize the verbal information of some natural languages. But speech recognition on many languages is not introduced into these systems due to objective or subjective reasons. Including Uzbek, Tajik, Karakalpak and other languages.***

Keywords: amplitude, criteria, feature, frequency, phoneme, probability, segment, signal, speech.

I. INTRODUCTION

The creation of natural language man-machine interfaces, and, in particular, automatic speech recognition systems, has recently become one of the urgent problems in the field of artificial intelligence and recognition [1].

Speech is a sequence of superposition's (superposition) of sound vibrations (waves) of various frequencies. The wave, as we know from physics, is characterized by two attributes - amplitude and frequency.

Speech recognition is a multi-level pattern recognition problem in which acoustic signals are analyzed and structured into a hierarchy of structural elements (phonemes), words, phrases, and sentences.

In recognition systems, the speech signal is divided into short superimposed segments, and each segment is converted into a feature vector, with the result that the input signal is represented by a sequence of feature vectors. The process of calculating feature vectors is called feature extraction or parametric representation [1].

II. FORMULATION OF THE PROBLEM

The task of the speech recognition system is to correctly recognize the sequence of words spoken by the speaker using a speech signal. This corresponds to the optimal criterion, which can be expressed as [3]:

$$\hat{w} = \underset{w \in W}{\operatorname{argmax}} P(w \mid O), \qquad (1)$$

where $\hat{w}$ is the output hypothesis of the phrase; w - any possible sequence of words; W is a set of all possible sequences of words (hypotheses); O is the sequence of vectors of features calculated from the input speech signal.

After applying the Bayes formula, formula (1) takes the following form:

$$\hat{w} = \underset{w \in W}{\operatorname{argmax}} \frac{P(O \mid w) P(w)}{P(O)}$$

$P(O)$ obviously does not change depending on the sequence of words w. Neglecting $P(O)$, we get a posteriori maximum probability criterion:

$$\hat{w} = \underset{w \in W}{\operatorname{argmax}} P(O \mid w) P(w) \qquad (2)$$

where $P(O \mid w)$ is the probability that the current feature vector O is observed if the sequence of words w is pronounced. This expression is called acoustic probability and is calculated using the acoustic models of the recognizer. $P(w)$ is the a priori probability of the appearance of a word in a phrase, which is calculated using language models.

III. CMU SPHINX

CMU Sphinx is a modern and very popular package for developing speech recognition systems, where you can implement both high-precision voice command control systems and continuous speech recognition systems with a large vocabulary [2].

The CMU Sphinx system is a voice-independent continuous speech recognition system that uses a hidden Markov acoustic model and an n-gram static model. CMU Sphinx demonstrates the feasibility of recognizing continuous, speaker-independent speech with a voluminous dictionary, the feasibility of which has been in doubt until today.

This package includes a set of tools that solve various problems and relate to various applications, such as:

Pocket sphinx is a "lightweight" C-based speech recognizer designed for mobile platforms;

Sphinx 4 is a flexible, modifiable recognizer written in the Java programming language and focused on stationary devices;

Sphinx train - a package of tools for adapting and teaching acoustic models based on hidden Markov models (SMM);

Sphinx base is the library required for running Sphinx train.

In fig.1 shows the generalized structure of the Sphinx system, the interaction of the components of which is demonstrated by arrows.

Manuscript published on 30 August 2019.
*****Correspondence Author(s)
Mamatov Narzillo*, Scientific and Innovation Center of Information and Communication Technologies at TUIT named after Al-Kharezmi, Tashkent, Uzbekistan, m_narzullo@mail.ru
Samijonov Abdurashid, Bauman Moscow State Technical University, Russia Federation
Nurimov Parakhat, Scientific and Innovation Center of Information and Communication Technologies at TUIT named after Al-Kharezmi, Tashkent, Uzbekistan
Niyozmatova Nilufar, Scientific and Innovation Center of Information and Communication Technologies at TUIT named after Al-Kharezmi, Tashkent, Uzbekistan

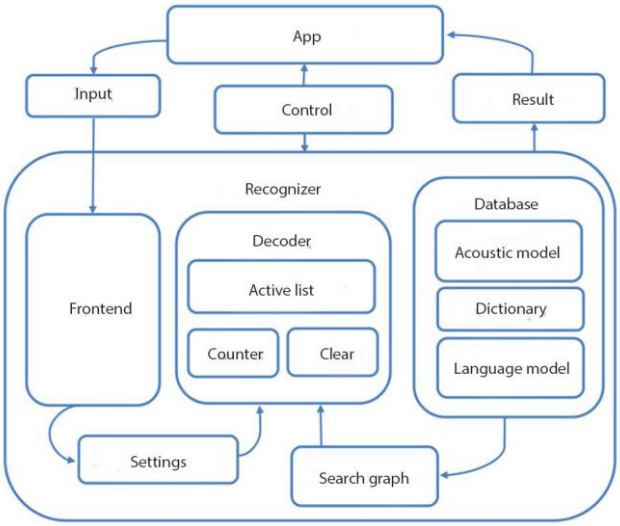

Fig. 1. Sphinx 4 architecture

The speech recognition system consists of three blocks: the frontend (external representation), the decoder and the knowledge base. These parts are controlled by an external application. The frontend block accepts a speech recording and characterizes it. Inside the frontend, the end-of-speech indication module highlights the beginning and end of speech, divides the audio stream into speech and non-speech, and removes areas with no speech. It can perform these operations either on a speech stream, directly, or on a sequence of feature descriptions (featured vectors) calculated on its basis. Block decoder directly produces recognition. It contains a graph design module (linguist), which converts any type of standard language model provided by the application to the knowledge base, into an internal format, and together with information from the dictionary and one or several acoustic models, builds a hidden language Markov model (HMM). The latter is further used by the search module to determine the structure of the vocabulary grid that should be found.

The acoustic model is the main part of the training. Acoustic model provides a display of acoustic information and phonetics. It uses speech signals from the training database. Various models are available for acoustic modeling. The Hidden Markov Model (HMM) is a widely used and accepted model [8] for learning and recognition.

The language model is also part of the training. The language model contains structural constraints, available in the language, for generating the probabilities of a word, followed by a sequence of *n-1* words [1]. The speech recognition system uses bigram, trigram and n-grammatical language models. The language model distinguishes between a word and a phrase with a similar sound.

IV. KARAKALPAK SPEECH RECOGNITION SYSTEM

In this speech recognition system, data is initially prepared, in which 150 words are collected from each of 10 Karakalpak speakers. Using the phonetic transcription, a phonetic dictionary is compiled and an acoustic and language model is developed.

The base of the Karakalpak words is used in this work and contains a corpus of speech and their transcription. The corpus contains 150 words collected from each of the 10 speakers. To facilitate the labeling of speech signals, audio files were generated by words in alphabetical order. The sampling rate of the recording is 16 kHz with a resolution of 16 bits.

Next, a speech rule file (transcription file) was created, which contains transcriptions for each sentence of a specific audio recording (the file is named asr5_train.transcription). For each transcription, the beginning and end of the sentence are indicated. At the end are the file names (without the extension) of the audio recording of this offer. Example:

The transcription file contains:

<s> BELGILE </s> (p_belgile)

<s> BIYKARLA </s> (p_biykarla)

Files:

p_belgile.wav

p_biykarla.wav

The next step is to create a dictionary. The dictionary contains all the words that are in the transcription file in alphabetical order without repetition. After each word phonetic analysis of the word is registered. Phonemes were specified by a certain rule.

Example: asr5.dic file

ALDI`G`A	A L D Y GH A	(forward)
ALDI`N`G`I`	A L D Y NG GH Y	(previous)
ALI`W	A L Y W	(to take)
ALPI`S	A L P Y S	(sixty)
ALTI`	A L T Y	(six)
AQI`RI`	A KH Y R Y	(end)
AQI`RI`NA	A KH Y R Y N A	(finally)

V. TRAINING

Training is the process of training the acoustic and language model along with a pronunciation dictionary to create a database for use in a recognition system. Acoustic model training is performed using CMU Sphinx tools.

VI. ACOUSTIC MODEL

In the acoustic model, the observed features of phonemes (basic speech units) are compared with the hidden Markov model (HMM). Words in the dictionary are modelled as a sequence of phonemes, and each phoneme is modelled as a sequence of model states.

VII. LANGUAGE MODEL

In this system, the n-gram language model is used to find the correct sequence of words. The search is performed by predicting the probability of the nth word using n - 1 preceding words. Commonly used n-gram models: unigram, bigram and trigram. The language model is created by calculating the number of unigrams of the word, which are converted into a vocabulary of tasks with word frequencies. Bigrams and trigrams are generated from the educational text based on this dictionary. In this paper, the Cambridge Modeling Language Modelling Tool (CMUCLMTK) is used to create a language model for this system.

Retrieval Number: J95240881019/19©BEIESP
DOI: 10.35940/ijitee.J9524.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

VIII. TESTING

Testing is the next stage after model training. Testing is a very important part, as it makes it possible to assess the quality of the generated database, for subsequent improvement and optimization.

IX. RESULT AND DISCUSSION

Implementation of the proposed work can be estimated by the percentage of recognition, determined by the following formula:

$$W = \frac{S + D + I}{N}$$

where S is the number of replaced words, D is the number of deleted words, I is the number of words inserted; N is the number of words.

The system showed the best rate of 87.88%. Table 1 shows the results of the experiments.

Table 1: The overall level of recognition of the system for the model

SER (%)	*WER(%)*	*Recognition (%)*
11,0	12,2	87,8

X. CONCLUSION

The developed voice control system can be used in a large number of tasks: control of computer applications, mobile platforms, or robotic devices, such as loader robots. The presented approach allows you to create speech recognition systems based on open technologies and a personal computer equipped with a microphone.

REFERENCES

1. Kipyatkova, I. S., Ronzhin, A. L., and Karpov, A. A. 2013. Automatic processing of colloquial Russian. St. Petersburg.
2. https://cmusphinx.github.io/wiki/tutorialoverview/
3. Rabiner, L. and Juang, V.N. 1993. Fundamentals of Speech Recognition. Prentice Hall.