Russian Speech Recognition System Design Based on HMM

Ma Yanzhou
Department of Basic
PLA University of Foreign Languages, Luoyang, China
myz827@126.com

Yi Mianzhu
Department of Language Engineering
PLA University of Foreign Languages, Luoyang, China
mianzhuyi@gmail.com

Abstract—Russian speech recognition is an important research focus. Starting from the basic principles of speech recognition, this article puts forward a basic HMM recognition system and analyzes the application of HMM in this system. It points out that Russian text corpus and speech corpus are the guarantee of high recognition rate. By experiments and comparison, we find some methods for further improvement, especially the optimization of recognition algorithm and its combination with Russian, which are the focus for future research.

Keywords-Russian Speech Recognition; Acoustic Model; Language Model; Russian Corpus

I. INTRODUCTION

In 1980s, HMM (Hidden Markov Model, HMM [1-2]) made the emergence, bringing about substantial breakthroughs in the field of speech recognition and greatly promoting the development of speech recognition technology. The focus of automatic speech recognition research gradually shifted from specific-person, small-vocabulary, and isolated word speech recognition to non-specific, large-vocabulary, and continuous speech recognition. The modeling of most speech systems is based on HMM. But Russian speech recognition [3] based on HMM model is still a very difficult research task. In addition to the complexity of speech recognition technology, Russian linguistic features also bring great difficulties for speech recognition applications. The research of Russian speech recognition by domestic and foreign researchers started at the beginning of this century. The recognition technology is still immature, and many key issues remain to be resolved.

II. THE PRINCIPLES OF SPEECH RECOGNITION

Speech recognition technology is also known as automatic speech recognition (ASR [4]), which converts the words content in human speech into computer-readable input, such as a key, a binary code or a character sequence. Its fundamental purpose is find a way to directly receive people's speech, understand their intentions and transform the speech signals into appropriate text or commands via computer identifying and understanding. It is an integrated technology consisting of phonetics, computer science, acoustics, information processing, and artificial intelligence.

The core of a speech recognition system is a speech sampling, recognizing and pattern matching system, whose basic principles are shown in Figure 1. First, pre-process the input speech signals, carry out appropriate amplification and gain control, and conduct anti-aliasing filtering to remove the interfering signal; then digitize the analog signal into a digital signal to be processed by computer; extract features and represent the speech signal in terms of several parameters of sound characteristics; finally, carry out speech recognition. Speech recognition is further divided into two phases: training and recognition. In the training phase, speech signals represented as characteristic parameters are processed, leading to basic recognition data which represent the common characteristics of the basic recognition unit. Based on such data, a reference template is constituted. The reference templates of all the recognizable basic units form the reference pattern database. In the recognition phase, feature extracting is carried out on the speech to be recognized, and the speech is compared with each individual template in reference pattern libraries in order to find the sound corresponding to the most similar reference template, which is the result of recognition.

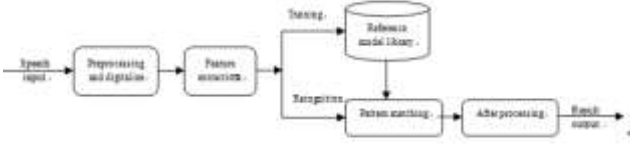

Figure 1. The basic principle of speech recognition

A. Preprocessing

Preprocessing includes sampling, quantizing and pre-emphasis of speech signals. The purpose of sampling and quantizing of speech signals is to digitize the analog speech waveform to be processed by digital computer. In sampling, the analog signals are sampled in the time domain at equal intervals. According to the sampling theorem, when the sampling frequency is greater than twice the highest signal frequency, original signals can be recovered. Usually, the sampling frequency of the speech signal is 8 KHz-10 KHz. The digitized speech signal is a time varying signal. Assuming that the speech signal is stable in the short time period of 10ms-30ms, windowing operation should be performed on the speech signal so that conventional methods can be used. Window function smoothly moves on the speech signal. The shape and length of the window function should be taken into consideration when a window function is chosen. Typically, boxcar window and hamming window will be used.

B. Endpoint Detection

In speech signal processing tasks, it is necessary to determine the speech segments and silent segments in the input signal. In speech recognition, correctly determining the starting point and end point of speech input is often very important for improving the recognition rate. For the part which has been determined as speech segments, surds and sonant should be further determined. In isolated word speech recognition system, it is necessary to correctly determine the starting point and end point of each speech input. Endpoint detection of the speech segment data can be done by using the short-time average magnitude parameter M and the short-time average zero-crossing rate Z.

C. *Feature Extraction*

After a variety of computing, a feature vector can be extracted from the audio data to be processed, which serves as a characteristic parameter (i.e. the speech recognition parameter) distinguishing this audio from other audios. Selecting characteristic parameters is a fundamental problem in speech recognition. The purpose of feature extraction of speech signal is to analyze and process the speech signal, remove redundant information irrelevant to speech recognition and obtain important information affecting speech recognition. At the same time, the speech signal is compressed. The speech signal compression ratio is generally between 10 and 100. Since the speech signal contains a large variety of information, many factors should be considered in order to determine what information should be extracted with what ways. The factors include cost, performance, response time, computation, and so on.

Linear prediction (LP [5]) analysis technology is widely used feature extraction technology. Many successful applications use cepstrum extracted by LP technology. However, linear prediction model is a pure mathematical model, which doesn't take into account the features of human auditory system in speech processing.

Mel parameter and perceptual linear prediction cepstrum extracted by perceptual linear prediction (PLP) analysis, to a certain extent, simulates the characteristics of the human ear's processing of speech and utilizes some achievements in the studies of human auditory perception. Using this technique, the speech recognition system performance has been improved. Judging from the practical case, Mel cepstral parameters have gradually replaced the originally popular cepstral parameters derived from linear prediction coding, because they takes into account the characteristics of human sound producing and receiving, which results in better robustness.

III. RUSSIAN SPEECH RECOGNITION SYSTEM

Russian speech recognition system is designed based on HMM, using the currently popular HTK (HMM Tools Kit, an experimental toolkit developed by the University of Cambridge for HMM establishing and processing. It is widely used in the field of speech recognition, but also can be applied to other fields such as speech synthesis and character recognition) for Russian speech acoustic modeling [6] and language modeling [7], as well as the analysis and processing of the recognition results.

A. *Hidden Markov Models*

HMM (Hidden Markov Model, HMM), a statistical model for speech signals, is a doubly stochastic process: one stochastic process describes how each short-time smooth segment transits to the next, i.e. the dynamic characteristics of short-time statistical features; the other one is used to describe the statistical characteristics of short-time smooth segments of non-stationary signals. Based on these characteristics, HMM can effectively track both short-time stationary segments of these two different parameters and identify the transition among them. Russian pronunciation process is also a double random process; so HMM can precisely describe the production of Russian speech signal.

HMM model is an important research result achieved in the field of speech recognition. It calculates output probability of speech parameters to the HMM model using the probability density function, searches the best state sequence and obtains the recognition result using the criteria of maximum posteriori probability; Meanwhile, it describe changes in pronunciation in terms of state residing and transferring and make the implied states correspond to the relatively stable pronunciation units. HMM models are usually represented as $\lambda = \{$л$, A, B\}$. For a HMM, define a combination of basic elements shown in Table 1. HMM model is a relatively complete representation of the acoustic model of speech. By using statistical training methods, it incorporates the upper language model and the underlying acoustic models into speech recognition search algorithm, and gets better results.

TABLE I. THE COMBINATION OF ELEMENTS OF A HMM

Model parameters	Description
λ	State number of model
$A = \{A_{IJ}\}$	State transition probability matrix
$\pi = \{\pi_{IJ}\}$	The initial state probability distribution
$B = b_{j(o)}$	Output probability density function

B. *Russian speech recognition system*

Russian speech recognition system is based on the basic principles and methods of speech recognition system, including the establishment of the Russian acoustic model and language model, as shown in Figure 2.

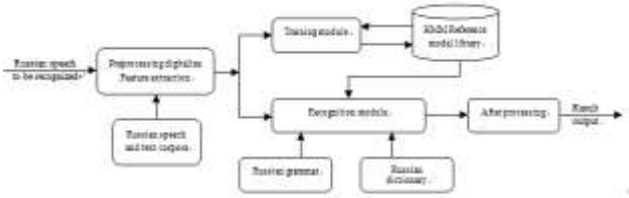
Russian speech to be recognized
Preprocessing digitalize Feature extraction
Russian speech and text corpora
Training module
HMM Reference model library
Recognition module
After processing
Result output
Russian grammar
Russian dictionary

Figure 2. Russian speech recognition system structure

Russian speech recognition system consists of two phases: training and recognition. In the training phase, the HMM model is established. By re-evaluating the parameters and adjusting the model parameters, we can obtain a model with better robustness. In the recognition phase, we use an existing HMM model library, data dictionary and grammar to form a recognition network, and use algorithms to find the best matching process. The process is as follows: first, sample the Russian speech signal to be processed, convert it into an electrical signal, and pre-process the electrical signal such as pre-emphasis and the endpoint detection; then, extract the feature vector data; combine them according to requirements of HMM models and dictionary in the recognition module, complete the recognition, match the results with the language model, results match, and output the final recognition results which meet the criteria. When establishing the HMM model database, in order to obtain vocabulary eigenvalues from the corpus to serve as reference values, it is necessary to store the data before the system uses them and to obtain the best HMM model via module training. To meet the recognition requirements of the system, it is necessary to select the recognition type. After some calculation and comparison, the system models conforming to the rules are kept, while those not conforming to the rules are discarded, so as to obtain good recognition results. Among them, eigenvalues, HMM model, grammar and data dictionary are important factors affecting the recognition rate, which is a difficult problem in the speech recognition nowadays.

C. *Speech and text corpus*

The most challenging issue of Russian speech recognition research [8-9] is large-vocabulary, speaker-independent continuous speech recognition. Despite the difficulties, there are some breakthroughs, which can be attributed to the unified framework of HMM algorithm in the recognition system and the delicate introduction of acoustic, phonetic and linguistic knowledge into this framework. In this framework, each of the basic unit of the bottom layer of the acoustic structure corresponds to a set of HMM structure and parameters, and constitutes phonemes, words, sentences, etc. The advantage is the unified structure, which facilitates computer processing.

Russian words can be divided into syllables [10-11]. A syllable can be constituted by a vowel, or by a vowel and one or more consonants. The number of syllables equals to that of vowels. Words with only on syllable are called monosyllabic words, such as тот. Words composed of two syllables are called two-syllable words, such as один; Words with more than two syllables are called multi-syllable words, such as комната. When a word contains two or more syllables, the vowel of one syllable needs to be read with some force, which is called a stress. Syllables with stresses are called stressed syllables. Vowels in stressed syllables are called stressed vowels. Russian pronunciation rules are shown in Table 2:

TABLE II. RUSSIAN PRONUNCIATION RULES

Vowel	**Consonant**			
	Voiceless	***Voiced***	***Hard***	***Soft***
а о у и е ё ы э ю я	п ф к т ш с х ч щ	й л м н р б в г д ж з	б в г д з л м н р х ж ш ц	к л м н п р с т ф х ч щ й

In Russian voice data collection, the voice recording program HSLAB in HTK is used for recording and manual tagging. The recorded Russian speech is standard news speech from the Russian State Television. The sampling frequency is 11025Hz, and the sampling precision is mono 16bit. The completely natural speed is 4 to 5 syllables / sec.

Wav format is adopted for speech storage. In order to construct a set of HMM models, a series of speech files and associated tag files are required. The speech can be tagged as different word-level transcripts and the associated tag files are converted into phonemes consistent with application requirements or correct formats of word-level tagging. The tagging of speech file can be manual or automatic or a combination of both. In this system, a small portion of the speech data is manually tagged, and then is used to correct those automatically tagged training data. In this way, a corpus is created for speech model training and language model testing [12-13].

As for Russian text corpus building, web crawler program is written to download texts in different fields the from the RIA Novosti news agency website. Then, the labels are deleted, special symbols are removed or replaced, duplication removal is conducted, and finally a plain text file is obtained and stored in Unicode format.

D. *Create a system using HTK tool*

HTK toolkit provides developers with rich and powerful tools, including attribute sampling, model parameter training, parameter self-tuning, dialect habilitation training, dialect identification, thesaurus construction, and so on. HTK toolkit has good maneuverability. Its debugging function provides good support for system development and shortens the system development cycle. It also provides application interfaces to support discrete density hidden Markov models (DHMM) and continuous density hidden Markov models (CDHMM).

Tools, such as speech data preparation, HMM training, recognition and data analysis, are used to train Russian speech acoustic model. According to a left-to-right HMM structure, a HMM model is created for each recognition unit. Like in the training process, different tools are used to recognize Russian speech. First, the speech to be recognized is preprocessed and a digitReg.mlf is generated in advance. This file records text content corresponding to the audio (including audio tag names and audio content). The audio tag name must be the same with the audio name, and audio content may be randomly generated by the system before being recognized. Second, the acoustic model digitReg.mnl file is generated by the HLEd tool. Characteristic parameters are extracted from it. And then, referring to the training process, initialize the original model, generate HMM file using HCompV function, revaluate it using HERest function, and complete internal recognition and external recognition.

IV. RESULTS ANALYSIS

By experiment and comparison, we found that Russian recognition results are not too satisfactory. Russian pronunciation varies greatly among different groups of people. Under the same condition, a large gap exists between the recognition rates of internal recognition and external recognition. The reason may be that the test speech used for internal recognition is the speech used for model training, so it is easier to be recognized by the system, thus the relatively high recognition rate. But external test speech recognition rate is low. To increase the recognition rate of Russian, improvements should be achieved in the following areas:

Optimize and improve the training and recognition algorithms. The computing amount of HMM algorithms is large. More efficient recognition training algorithms are needed for practical purpose. In this paper, the recognition tool HVite in HTK Toolkit is adopted for the speech recognition system. It is a token Viterbi-Beam search algorithm. The speed of this search algorithm is very slow and the performance is not very satisfactory, resulting in low recognition rate. Therefore, it is necessary to seek recognition search algorithms with better performance and create a recognition system with high response speed and user-friendly interface.

Matching and recognition reliability depends on the accuracy of the selected eigenvectors. When the quality of the audio file is not high, it is difficult to guarantee the accuracy. Some of the speech files collected is only part of the actual audio files, so it may not be possible to determine the features completely and accurately for characteristic parameters, which greatly reduces the recognition rate. Due to the large dynamic range of speech, different speakers' speeches, even the same person's speeches at different times and places, are quite different. Therefore, in HMM training, giving full consideration to the speaker's influence is very important for better estimation of HMM parameters.

Increase the amount of speech corpus data for training. An HMM model contains a lot of parameters to be estimated, so in order to get a satisfactory model, you must have a lot of training data, which is difficult in practice. On the other hand, adopting the smaller model, i.e. reducing the number of states and the number of mixed Gaussian components on each state, is also difficult in practice. We should base on real Russian data, analyze and sort massive texts, establish large-scale speech corpus containing linguistic phenomena and improve representativeness and accuracy of speech recognition.

Optimize and improve the Russian language model. To improve recognition rates, especially for large-vocabulary speech recognition, it is necessary to integrate N-gram statistical language model and rule-based language model. Binary and ternary models should be adopted to increase the amount of training, enhance complexity of the statistical language model and improve the performance of context constraint.

V. CONCLUSIONS

Russian speech recognition is a relatively new research focus. Domestic and foreign researches have just started and are still in infancy. There is still a lot of work to do to improve the efficiency, processing speed and user experience feedback. Among them are the size and quality of the speech corpus, the flatness of the text corpus, speech recognition algorithm optimization and adjustment, and so on. This article provides only a reference for Russian speech recognition research. Further theoretical and experimental studies are needed to transform the results into practical application. We hope that a mature system with a higher recognition rate and faster recognition speed can be developed in future studies.

[1] Biing Hwang Juang,Laurence R Rabiner.Hidden Markov Models for Speech Recognition[J].Technometrics,1991,33(3):251-272.

[2] Lawrence Rabiner.A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition[J].Proceedings of the Ieee,1989,77(2):257-286.

[3] Andrey L Ronzhin,Rafael M Yusupov,Izolda V Li, et al.Survey of Russian Speech Recognition Systems[C]//Specom,2006:54-60.

[4] Stephen E Levinson,Lawrence R Rabiner,Man Mohan Sondhi.An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition[J].Bell System Technical Journal,1983,62(4):1035-1074.

[5] John Makhoul.Linear Prediction: a Tutorial Review[J].Proceedings of the Ieee,1975,63(4):561-580.

[6] Tanja Schultz,Alex Waibel.Language-independent and Language-adaptive Acoustic Modeling for Speech Recognition[J].Speech Communication,2001,35(1):31-51.

[7] Edward WD Whittaker.Statistical Language Modelling for Automatic Speech Recognition of Russian and English[J].Daktaro Disertacija, Cambridge University Engineering Department, Cambridge,2000:1-141.

[8] A. Karpov,K. Markov,I. Kipyatkova, et al.Large Vocabulary Russian Speech Recognition Using Syntactico-statistical Language Modeling[J].Speech Communication,2014,56:213-228.

[9] C. Tillmann,S. Hewavitharana.A Unified Alignment Algorithm for Bilingual Data[J].Natural Language Engineering,2013,19(1):33-60.

[10] Daria Vazhenina,Konstantin Markov.Phoneme Set Selection for Russian Speech Recognition[C]//Natural Language Processing Andknowledge Engineering (nlp-ke), 2011 7th International Conference on,Ieee,2011:475-478.

[11] Andrey L Ronzhin,Alexey A Karpov.Implementation of Morphemic Analysis for Russian Speech Recognition[C]//9th Conference Speech and Computer,2004:1-6.

[12] A. V. Savchenko.Phonetic Words Decoding Software in the Problem of Russian Speech Recognition[J].Automation and Remote Control,2013,74(7):1225-1232.

[13] S. Zablotskiy,A. Shvets,M. Sidorov, et al.Lrec 2012 - Eighth International Conference on Language Resources and Evaluation[M].[S.l.]:European Language Resources Assoc-elra,2012:3374-3377.