Design and Research of Japanese Computer Continuous Speech Recognition System Based on HTK

Li Shan\(^1\)^*
\(^1\)Dalian Polytechnic University, Dalian, Liaoning, 116034, China

*Corresponding author e-mail: shanli@dlpu.edu.cn

Abstract. The main meaning of speech recognition is reflecting the human voice in the form of text through computer technology, it is to convert speech into text. HTK is used in speech recognition, speech synthesis, character recognition and many other fields, and its application is more and more extensive and mature.

Keywords: Speech Continuous Recognition, Dynamic Range Adjustment, HTK

1. Introduction
This paper mainly introduces the constructed process of continuous speech recognition system. It analyzes the construction method of continuous speech recognition system and the basic method of using phoneme to construct the acoustic model. It describes a continuous speech recognition system, which was built with Hidden Markov Toolkit. At the same time, it provides a system evaluation method. The system uses the Japanese standard JNAS database for modeling, and it uses the pronunciation of Japanese students of Hokkaido University to conduct experiments and verify the effectiveness of the model.

2. Basic principles of speech recognition
Speech recognition is essentially a type of pattern recognition, which is to allow machines to understand human speech, which is a very important human-computer interaction technology. At present, most speech recognition systems use the principle of statistical model matching. Generally speaking, the speech recognition system can be divided into two core parts: one is to use the parameters extracted from the feature to train a set of models according to a suitable algorithm; the other is to use the recognition algorithm and the trained model. The second is to recognize unknown speech according to the recognition algorithm and the trained model [1].

A typical continuous speech recognition system consists of four parts: preprocessing, feature extraction, training and recognition. Pre-processing is to perform operations such as pre-emphasis, framing, windowing and endpoint detection on the voice signal; feature extraction is to extract the feature parameters of the voice signal to obtain a set of feature vectors for subsequent processing; training is based on the extracted data preparation stage, mainly including corpus preparation, phonetic annotation, dictionary creation, language model establishment, etc.
2.1. Corpus to prepare
In speech recognition, in order to fully train the acoustic model, the selected corpus must be representative. A good corpus should cover speech and language phenomena as much as possible. The choice of corpus directly affects the recognition rate of speech recognition. After preparing the corpus, the first task in speech recognition is to process the corpus. First use the prepared dictionary to segment the sentence script file [2]. The sentence segmentation is based on the words appearing in the dictionary as the basic unit to segment the sentence, and then save it as a word list format file, and then use the dictionary to split the sentence Each word is labeled at the syllable or phoneme level to generate a syllable or phoneme list labeling file, and then a part of the corpus is labeled with a period of time as the initial training corpus.

2.2. Voice annotation
Before the acoustic model training, the voice waveform file needs to be annotated. The labeling of waveform files can be either with or without time periods; manual labeling methods (generally suitable for speech recognition systems in isolated words in small vocabulary) can be used, or automatic and average The method of labeling (generally suitable for continuous speech recognition systems with large vocabulary), or a combination of automatic and manual methods. HTK provides manual recording and annotation file tool HSLAB. This annotation method is generally used in the speech recognition system of small vocabulary isolated words, which can achieve good results. In large vocabulary continuous speech recognition, HSLAB is used for The efficiency of manual labeling is very low, and it is difficult to distinguish the boundaries between words when labeling with time periods. In engineering, an average labeling method is generally adopted [3].

3. Feature extraction of speech recognition system
Figure 1 describes the general process of the speech recognition system. The upper part of the figure represents the modeling process, and the lower part represents the recognition process. A complete continuous speech recognition system mainly consists of four parts: preprocessing, feature extraction, acoustic model building and recognition of speech signals." The acoustic model building mainly uses HTK, which is specifically used to establish and process hidden Markov models (Hidden Markov Model, HMM) toolkit is currently widely used in the field of speech recognition applications and research. ” The recognition engine uses the Julius open source platform. Julius is a speech recognition engine for continuous speech recognition related research and development with large vocabulary." It is a high-performance, dual-channel based on N-gram and context-sensitive HMM The decoder software can decode the input continuous speech in real time [4].

![Figure 1. Frame structure of speech recognition system](image)

The preprocessing is mainly to perform basic operations on the speech signal. Generally, formula (1) is used to compensate the high frame part of the speech signal, and the value range of \( \mu \) is generally between 0.94 and 0.97.
This system uses traditional MFCC as the method of speech feature extraction. The noise robustness algorithm is introduced while extracting the MFCC features, and the noise robust voice feature extraction is performed by recompiling the HCopy command in the HTK to obtain continuous speech feature parameters that are robust to extracting noise.

4. Modeling of speech recognition system

Acoustic model is the most basic unit and core part of speech recognition system. Using HMM to build acoustic model is the method adopted by most speech recognition systems. In Japanese continuous speech recognition, phonemes are used as the modeling unit of the acoustic model. The number of phonemes is not equal to the number of letters, it is the smallest phonetic unit divided from the perspective of sound quality. There are 40 standard phonemes in Japanese, plus the pause (sp) caused by ventilation, thinking, etc., the start-silence segment (SilB) and end-silence segment (SilE) of continuous speech [5]. A total of 43 phonemes were finally modeled. Compared with isolated word recognition, in continuous speech, the speech is also affected by nearby sounds. This form of influence is called co-articulation. Therefore, in the continuous speech model, it is also necessary to consider the co-articulation produced by related phonemes in the context. This context-sensitive phoneme model captures co-pronunciation by considering several phonemes before and after a certain phoneme, and improves the recognition performance of the system. The more relevant phonemes considered, the higher the complexity of the model. The commonly used method is to consider only the phonemes adjacent to the left and right of the current phoneme. This method is called the riphone model. The training of the triphone model directly comes from a single phoneme related to the context. This training method causes the number of HMM models in the system to increase by a multiple of 3, which affects the processing efficiency and recognition accuracy of the system. In order to avoid this problem, the same type of triphones are bundled according to the conditions obtained by experience and experiment. This operation uses HEED tool functions and decision tree files to train all triphone models for many times [6].

In the modeling training, the continuous speech features are first used to generate a monophone model. This article uses HCompV and HERest to create a training monophone model. In order to solve the co-pronunciation, it is also necessary to create a triphone model and re-evaluate the model. Generate triphone model according to Net file. For training with monophone and triphone models, it is also necessary to introduce Gaussian mixture numbers for model reestimation classification training. The purpose of this step is to train an efficient and stable acoustic model that stabilizes the accuracy of recognition [7]. There are two purposes of classification. The first is to reduce the phoneme model categories, so that the models trained on fewer data sets are more reliable: second, to reduce the overlap between models and increase the discrimination, factor training classification. The process is shown in Figure 2.

![Figure 2. Phoneme training classification process](image-url)
Create the original HMM model, which is composed of a mean vector and a covariance moment Chen, represented by 5 states, 26-dimensional coefficients and a state transition matrix. Then generate the HMM of each phoneme level according to the phoneme table and the phoneme feature file. Finally, the phoneme-level HMM is trained based on all the training speech data to form a monophone model. In order to enhance the anti-interference ability of the HMM model, a silent part (for phoneme sp) is added for state optimization. This process is completed by HHEd in HTK. For multi-pronounced words, use HVite tool function combined with corpus to perform repeated training to complete [8].

5. Data source of speech recognition system
The voice data is divided into two parts: training library and test library. The training speech database comes from the JNAS (Japanese news article sentences) database. The speech in the training database comes from the content in the newspaper "Daily News", with more than 23,000 sentences read by 153 men. In order to better detect the robustness of the model, the test library data not only comes from the data selected by JNAS from the training library, but also the sentences read by Japanese students from Hokkaido University that are completely unrelated to the training library data.

6. System implementation
JNAS includes phonetic data, phonetic romaji labeling, and phoneme time segment division. First, convert the Roman characters of the phonetic information into phoneme-level annotations. The annotations are completed by the Perl script tool and Hled. The division of the phoneme time period is converted into a start frame and an end frame stored in the file [9]. Generate context-sensitive ternary phoneme network through perl gadget tr2net. The network and the MFCC features of the speech are jointly trained to obtain the system acoustic model. Finally, after adding the Gaussian mixture number and performing state classification, an HMM model with about 2000 states is obtained.

7. Results and evaluation
In continuous speech, co-pronunciation will cause changes in the pronunciation of adjacent phonemes. This change will cause corresponding errors in recognition (deletion errors, insertion errors, and substitution errors). These errors lead to a significant drop in the performance of continuous speech recognition technology. The following two formulas are evaluation rules:

\[
R_c = \frac{N - S - D}{N} \times 100\% \tag{2}
\]

\[
R_A = \frac{N - S - D - 1}{N} \times 100\% \tag{3}
\]

N represents the total number of vocabulary in a sentence, and S represents the misrecognized word, that is, the correct is recognized as wrong. D represents a word that is not selected as a word, that is, it is correctly recognized but not evaluated as a word. I represents a word that should not be a vocabulary but is recognized as a word, for example: noise or silent parts are recognized as a word. RA represents the recognition performance of the entire continuous speech. RC represents the correct recognition ratio of words in the entire continuous speech vocabulary set [10].

8. Conclusion
Speech recognition is a hot spot of current research, and it has a very broad application prospect. This paper mainly studies the continuous speech recognition system, which is based on HTK. All walks of life pay more and more attention to speech recognition research, so applying speech recognition technology to real life will be a revolutionary breakthrough. I believe that in the near future, it will be the era of speech recognition. It will completely improve people’s lifestyles and truly enable people to "talk without hands".
Acknowledgments
Dalian University of technology 2018 teaching achievement award cultivation project "Japanese translation inter university credit course reform and construction"; Dalian University of technology 2018 education and teaching reform project "Japanese Intensive Reading inter university course teaching mode research and practice" (jglx2018079)

References
[1] Design and implementation of Chinese continuous speech recognition system based on HTK [D]. Anhui University, 2011.
[2] Fan Huimin, Zheng Luan. Design and Implementation of Speech Recognition System Based on HTK [J]. Computer Programming Skills and Maintenance, 2015: 97-99.
[3] GAO Jian. Research on Continuous Small Word Speech Recognition System Based on HMM [J]. Modern Electronic Technology, 2011: 213-215.
[4] Li Yi. Research on continuous speech recognition system based on HTK [J]. Computer Disc Software and Application, 2012: 104-105.
[5] Liu Yanxiu, Fu Haidong. Construction and Research of Continuous Speech Recognition System Based on HMM [J]. Journal of Changchun University, 2015: 38-41.
[6] Peng Xi; Qian Yingjing; Melvin wong. Gong Xiao Nan; Chen Shuangting; Yang Huixian; . Research and design of speech recognition algorithm based on eigenvalue statistics [J]. Science & Technology Information, 2019: 7-9+11.
[7] Shi Xianfeng, Zhang Xuezhi, Zhang Feng. Design of Speech Recognition System Based on HTK [J]. Computer Technology and Development, 2006: 43-44+47.
[8] Sun Yiming, Liu Wei. The establishment and research of Japanese continuous speech recognition system based on HTK [J]. Computer CD-ROM Software and Application, 2013: 199-200.
[9] Wang Hongru, Yang Genke, Yang Zuhua. Research and Implementation of Continuous Speech Recognition Website System Based on HTK [J]. Microcomputer Application, 2010: 25-26+36.
[10] Xue Hui. Research and Design of Mobile Phone Order System Based on Speech Recognition [J]. Microcomputer Applications, 2017