The Key Technology of Speech Interaction Based on Deep Learning

Liang Chen, Yugang Dai*, Bingjun La, Mengli Sun, Zhuangzhi Xiong
Northwest Minzu University Information Technology Academy, LanZhou China
* 49666594@qq.com

Abstract. Language recognition is essentially a pattern recognition problem, and all kinds of pattern recognition methods are used for language recognition research. For language recognition studied in this project, the highly effective speech feature parameters, acoustic model and language recognition classifier, the Chinese, Tibetan, Mongolian, uygur and training of acoustic model are established, so as to identify the voice input languages. In recent years, the deep learning theory has been applied in the field of speech recognition, and the DNN-HMM model has become the mainstream of acoustic modeling. This project studies Tibetan, Mongolian, Uygur emotion modeling of acoustic model and language model, training methods, studies the decoding algorithm, for continuous speech recognition of input speech, converts it to ethnic Chinese words.

1. Introduction
High-quality corpus is the key to training high-precision and robust acoustic models. This project requires a large number of labeled Tibetan, Mongolian and uygur language materials. First of all, we should study the phoneme balance algorithm, design the emotional articulation text of phoneme balance, and record the emotional speech corpus. The study establishes the Tibetan, Mongolian and uygur pronunciation dictionary, and then annotates the pronunciation of the corpus.

There are many voice input sources, including fixed and mobile phone, mobile client, other Internet voice input, etc. Background noise and channel are different. It is necessary to study the speech enhancement, noise reduction and endpoint detection of low signal-to-noise ratio (SNR), and research channel adaptive method to reduce the influence of different channels.

Language recognition is essentially a pattern recognition problem, and all kinds of pattern recognition methods are used for language recognition research. For language recognition are studied in this project, the highly effective speech feature parameters, acoustic model and language recognition classifier, the han, Tibetan, Mongolian, uygur and training of acoustic model is established, so as to identify the voice input languages.

This project study Tibetan, Mongolian, Uygur emotion modeling of acoustic model and language model, training methods, studies the decoding algorithm, for continuous speech recognition of input speech, converts it to ethnic Chinese words.

2. The construction of national language emotional corpus
Must first study phonemes balancing algorithm, design of phonemes, balance the emotional pronunciation text recording emotional speech corpus, the research set up the Tibetan, Mongolian, uygur pronunciation dictionary, and then to the pronunciation of corpus tagging, specific include: the corpus, the segmentation of text design, recording corpus and corpus.
2.1 Corpus text design
Training acoustic model with the basic principles of corpus building is to be able to use the corpus of less as far as possible to cover as much of the natural language phenomenon, including sound and super sound sound phenomenon, at the same time avoid the data sparseness problem. The combination of speech recognition and synthesis algorithm with the selection algorithm of corpus makes the constructed corpus have certain typicality and representativeness. The selection algorithm of text corpus is divided into two stages: the primary stage and the optimization stage based on the Greedy algorithm.

2.2 Corpus recording
Recording environment requires the use of a professional recording studio, recording environment condition are constant, all kinds of parameter variation range is small, the resulting data recording to achieve higher signal-to-noise ratio, can make the quality requirements of corpus. The recording hardware equipment mainly includes desktop computer, external sound card, mixer, microphone. The recording software USES Adobe Audition V3.0 software. The sampling rate of the recorded sample is 40KHz and the sampling accuracy is 16bit. The record file is stored in a PCM encoded *.Wav format.

2.3 The syncopation of corpus
In this study, a speech synthesis algorithm based on HMM was proposed, and phonemes were used as synthetic primitives in speech synthesis. On the basis of the hidden markov model, the program automatically completes the syncopation of phonemes.

3. Speech preprocessing in complex environment
Complex multi-ethnic voice source, voice pretreatment under complicated environment need to study low SNR speech signal endpoint detection of speech enhancement, noise reduction, etc., research channel adaptive method, to reduce the effects of different channel.

3.1 Speech enhancement
The interference of background noise inevitably causes the speech to be polluted, thus reducing the accuracy of speech recognition. This project USES wavelet transform algorithm to improve the speech signal processing. Wavelet transforms belongs to the time/frequency analysis method, it has the characteristics of multi-resolution, namely in the low frequency part with high frequency resolution and low time resolution, the high frequency part has higher time resolution and lower frequency resolution, very suitable for detecting voice in the midst of abnormal changes of transient signal, and can show its composition.

3.2 Endpoint detection
Accurate endpoint detection can improve the accuracy of recognition, and it is used to estimate the accurate noise model in speech enhancement system. Spectral entropy method of endpoint detection of detailed steps is as follows: in the early stages in the first 10 frames as no sound clips used to estimate the noise, speech signal spectrum entropy is obtained by using the spectral entropy method for noise reduction processing, according to the spectral entropy method for speech endpoint detection, when the data is fetched, ends speech endpoint detection, or continue to operate.

3.3 Channel self-adaptation
The recognition process of speech signals is to map the feature space to the model space and process the processing according to the likelihood of the mapping, as shown in figure 1.
Therefore, training and test environment can not match several links in the channel compensation, thus forming based on the characteristics of compensation, compensation based on the model, and based on the logarithmic likelihood score for several types of compensation. The feature compensation
method is widely used because it does not involve the model training, and this project adopts the inverse spectral mean reduction (CMS) to carry out feature compensation.

Figure 1. The impact of training and test environment mismatch

4. National language recognition based on GMM

In the process of national language language recognition using GMM (gaussian mixture model) model, the model of language recognition algorithm using at the bottom of the acoustic information, according to the probability statistical distribution characteristics of feature vector space build language related parameters.

The GMM-UBM system framework is shown in figure 2.

Figure 2. GMM-UBM system framework

4.1 Feature extraction

The MFCC parameters were extracted for speech data in various ethnic languages, and the MFCC coefficient (c0-c6) was obtained. MFCC is extended to 49 d SDC by 7-1-3-7 (n-d-p-k). The 7th-order MFCC coefficient and 49 SDC are combined to obtain the 56-dimensional feature. The VAD algorithm was used to remove the silent frame, and the characteristic parameters were also used to remove the channel convolution noise by using the inverse spectral domain filter of CMS, gaussian and RASTA.

4.2 UBM model training

Firstly, the algorithm uses HMM-UBM to model speech signals in time sequence ,after obtaining the matching score of each registered speech relative to other training speech, normalizing the matching score of each registered speech, make the matching score a feature vector, then, input it to the relevance vector machine for learning, obtaining the classified information of each registered speaker's voice. Finally, establishing the HMM model for the speech waits to be identified and calculate its matching score relative to other training speech ,then, normalization the matching score of this speech, make the matching score a feature vector, then input it to the relevance vector machine for decision making.

The UBM model is adopted to train the GMM system to a high degree of mixing.

4.3 The adaptation of the language model

The adaptation of the language model to the target language model is based on the UBM, and the data of the language is adjusted appropriately, which is the adaptive method. The adaptive methods are
varied, and the maximum likelihood linear regression and the maximum posterior probability is used to achieve better performance in the case of relatively few voice data. But for language recognition, there are enough data for each language.

4.4 Gaussian back end classifier
The back end USES a gaussian classifier, and all the model scores are spliced into a scoring vector, which is trained for the data of each language. Single gaussian backend model in training, the first LDA analysis, the linear transformation of N - 1 dimension score vector, where N represents the number of languages, then score vectors on this training a single gaussian classifier.

5. National speech emotion recognition
The recognition of emotional speech is the first classification of the speech of national emotion. Then, based on the specific reliable voice emotion database, the research on the emotional information and the construction of the speech emotion recognition model was developed, as shown in figure 3.

Figure 3. Speech emotion analysis step diagram

5.1 The classification and representation of national speech emotion
There is no unified standard for the classification of emotion in the study of national speech emotion recognition, and the emotion is identified and classified by the daily language label.

Divide emotional types into four categories: happiness, anger, surprise and sadness, and try to include all emotions in these four emotional states.

5.2 Extraction of speech emotion characteristic parameters of national language
The phonetic and acoustic characteristics of national speech related to emotion can be classified into three categories: prosodic features, sound quality characteristics and tonal characteristics. The main characteristic parameters are: base frequency, amplitude of speech signal, energy characteristic, resonance peak.

5.3 National speech emotion recognition based on hidden markov model (HMM)
GMM emotion recognition, first of all, according to the characteristics of the emotional speech database and emotional speech emotional speech model was constructed, and voice to voice of speech database model, make its performance to the best approximation. In GMM, it can be regarded as a continuous HMM with a state number of 1.

Through EM algorithm, the model parameters of GMM are obtained and the emotional dimension model is constructed. When recognizing the input speech, the characteristic parameters of the input speech are extracted, and the model with the highest matching score is calculated and selected as the identification result.
6. **Continuous speech recognition based on deep learning**

The acoustic models and language models are trained through Tibetan, Mongolian, uyghur speech and language materials. In the recognition, 39-dimension MFCC feature parameters were extracted for the pre-processed national speech, and the acoustic contrast and language decoding were performed, and the final output of the text was shown in figure 4.

![Continuous speech recognition framework](image)

**Figure 4. Continuous speech recognition framework**

The training of acoustic model is divided into two stages. First, the GMM-HMM model is trained, then the GMM is replaced with DNN, and the DNN-HMM model is obtained, and DNN-HMM is further trained and optimized.

6.1 **National language GMM model training**

The initial model is GMM-HMM, which establishes the initial model for the phoneme of each national language, and its structure is shown in figure 5.

![Five-state HMM model topology](image)

**Figure 5. Five-state HMM model topology**

In the training process of HMM model, the sequence of characteristic parameters of each speech is the observation sequence, which is connected with the first and end of the phonetic model corresponding to this sentence, forming a compound HMM model. On the basis of single-tone submodels, three-tone sub-model is generated.

6.2 **DNN-HMM acoustic model of national language**

In GMM-HMM training model, on the basis of the GMM is replaced by within DNN, get within DNN-HMM model transformation, retains the state transition probability, the bottom-up training within DNN models of each layer, and estimate the state transition probability. Depth within DNN-HMM (hidden markov model, neural network figure 6), the upper of the HMM is used to describe the temporal change of speech signals, each senone (i.e., binding of three sound sub-state) observation probability within DNN is used to describe.
Acknowledgements
The Research and Application Demonstration Project of Ethnic Minorities Comprehensive Analysis of Network Public Opinion and Cloud Service Key Technologies (No. 2014BAK10B03) 2016 Science and Technology Support Plan 《Large Data Based Ethnic Language Public Information Intelligent Service System and Key Technologies and Applications》 Project Number: 1604JCCA168

References
[1] Wang mingsheng, dong cuiling. Build call center upgrade service level [J]. Cable TV technology, 2005-11: 74-81.
[2] 2010 China (Asia Pacific) best call center grand award (EB/OL). Tianjin: China outsourcing network, In the 2010-11-26 s.
[3] CTI BBS. 2010 China call center industry development research report [EB/OL]. Beijing: doudin net, 2010-5.
[4] Jun-de song. Development and future of CTI technology in China [J]. Chinese computer user, 2003(6) : 22-23.
[5] New operator operational guidelines for operational guidelines - call center [EB/OL]. Chengdu: small and medium-sized enterprise IT network, In the 2007-10-16.
[6] Hui-chun Chen. The composition design and application of call center [D]. Jilin: jilin university, 2007.
[7] Liu Jing. China call center development trend and development trend [EB/OL]. Beijing: China software network, In the 2007-8-10.
[8] The severn.2006 China call center industry development trend research report officially released [EB/OL].Beijing: sina Network, 006-7-11.
[9] Call center function extension direction exploration [EB/OL]. Shanghai: call center and BPO industry information network, The 2005-11-25.