USE DIFFERENT ACOUSTIC MODEL DATA MINING METHODS, COMPARING THE RECOGNITION RESULTS ON THE SAME DATABASE

E.I. Gubin Ph.D., Associate Professor
Hongshuai Sun, student 8PM0I
Tomsk Polytechnic University
E-mail: hunshuay1@tpu.ru

Introduction

Speech recognition is the use of computers to convert speech signals into text corresponding to their content. Speech recognition is a complex interdisciplinary subject, and its application is also manifested in many aspects. In the Internet era, speech recognition has become more and more widely used and it has gradually become the entrance to the Internet. Therefore, in speech recognition, it is very important to study different acoustic model data mining methods.

Research methods

Using feature extraction techniques. The original speech waveform signal is converted into a sequence of feature vectors by means of signal processing, as input to a speech recognition system. The speech signal is usually sampled at 16 kHz/8kHz, often with a frame length of 25ms and a frame shift of 10ms. Then build different acoustic models and compare the recognition results.

The steps we need to do: 1) Prepare data and feature processing. 2) Build a basic acoustic model. 3) Performance comparison of DNN-HMM model and GMM-HMM model (word error rate, WER). 4) Comparison of recognition results of CNN-HMM-based models on SWB-300 database. 5) Comparison of recognition results of RNN-HMM-based models on SWB-300 database. 6) Comparison of recognition results of the model based on LSTM-RNN-HMM on a 2000-hour speech retrieval database. Compare different acoustic model data mining methods, compare the performance of the models and select the most suitable and accurate model.

Results

First, we prepare the data and perform feature extraction. Common features: Short-term spectral features such as MFCC, PLP, fbank, etc.

Then we build the basic acoustic model. Since the 1980s, there have been three cornerstones of speech recognition: Hidden Markov Model (HMM), Gaussian Mixture Model (GMM) and MFCC/PLP short-term features. In 2011, the DNN-HMM acoustic modeling technology based on deep neural network appeared, which greatly improved the performance of speech recognition system. The deep neural network uses a unified objective function, and multiple hidden layers perform feature extraction and model classification at the same time, and the model parameters are more effective. Compared with single-layer neural network theory, deep neural network has no breakthrough in itself, but it has excellent data modeling ability in practical application. Performance comparison of DNN-HMM model and GMM-HMM model (word error rate, WER).

Table 1. DNN-HMM model and GMM-HMM model

acoustic model	decoding method	RT03S FSH	RT03S SWB	Hub5'00 SWB
GMM 40-mix,ML,SWB 309h	One pass	30.3	40.9	26.6
GMM 40-mix,BMMI,SWB 309h	One pass	27.5	37.7	23.7
CD-DNN 7 layers * 2048,SWB 309h	One pass	18.6	27.6	16.2
GMM 72-mix,BMMI,Fisher 2000h	Multi pass	18.7	25.3	17.2

Advanced application of deep neural network in speech recognition.

1) Convolutional Neural Network Acoustic Model
   Comparison of recognition results of CNN-HMM-based models on SWB-300 database.

Table 2. CNN-HMM model

acoustic model	SWB-300
GMM-HMM	14.6
DNN-HMM	12.6
CNN-HMM	11.6

The error rate is relatively reduced by 10%.

2) Recurrent Neural Network
Comparison of recognition results of RNN-HMM-based models on SWB-300 database.

Table 3. RNN-HMM model

acoustic model	SWB-300
GMM-HMM	14.6
DNN-HMM	12.6
RNN-HMM	12.1

The error rate is relatively reduced by 5%.

3) Long Short-Term Memory Recurrent Neural Networks
Comparison of recognition results of the model based on LSTM-RNN-HMM on a 2000-hour speech retrieval database.

Table 4. RNN-HMM model

model training	DNN	LSTM
Cross-Entropy	11.2	10.1
Sequence	10.1	9.0

The error rate is relatively reduced by 11%.

Conclusion

By understanding the application of deep neural networks in speech recognition. We compare the performance of the DNN-HMM model and the GMM-HMM model. The DNN-HMM model greatly improves the performance of speech recognition systems. Compared with the convolutional neural network acoustic model, Recurrent Neural Networks have worse recognition results. This is because standard recurrent neural networks are not ideal for modeling long-term contextual relationships. So the LSTM-RNN-HMM model appeared. The long-short-term memory recursive structure can better solve the gradient disappearance problem in the time dimension. So the LSTM-RNN-HMM model works better. Different neural network modules can be combined to build more flexible and powerful acoustic models.

References

1. Huang Shan, Gubin E. Data cleaning for data analysis // Молодежь и современные информационные технологии: Труды XVI Междунар. научно - практической конференции студентов, аспирантов и молодых ученых. Томск, 2018г. - С. 387-389.
2. What Role Does an Acoustic Model Play in Speech Recognition? [Electronic resource]. – URL: https://www.rev.com/blog/resources/what-is-an-acoustic-model-in-speech-recognition
3. Acoustic Modeling for Speech Synthesis [Electronic resource]. – URL: https://static.googleusercontent.com/media/research.google.com/zh-CN//pubs/archive/44630.pdf