Overview of end-to-end speech recognition

Song Wang¹, Guanyu Li⁰
¹Key Laboratory of National language Intelligent Processing Gansu Province, Northwest Minzu University, Lanzhou, China
*guanyu-li@163.com

Abstract. In the 1960s, automatic speech recognition has been widely studied. In the past, HMM has been the mainstream of the acoustic model. With the development of machine learning, neural network is introduced to the speech recognition, relying on neural network’s strong learning ability. Thus the acoustic model of DNN has significantly improved the voice recognition rate, compared to the HMM model. In order to simplify the traditional speech recognition system, the end-to-end speech recognition method is proposed. This paper mainly introduces and analyzes the end-to-end system, and the main two models of CTC and attention, as well as the prospect of future speech recognition research.

1. Introduction
Automatic speech recognition has been a hot topic of research. In the 1980s, after IBM applied HMM to speech recognition, HMM has been playing an important role in speech recognition, and HMM-GMM has become the mainstream acoustic model. In 2006, after Li Deng and Hinton[1] proposed the use of deep learning in speech recognition, the neural network became a research upsurge of speech technology, which turned from the ANN to the DNN. The research included the neural network composed of restricted Boltzmann machine stacking and conducted layer by layer training for the network. Improve the over-fitting problem of neural network by using dropout. DNN-HMM became the main acoustic model, showing strong recognition capability in the recognition of large vocabulary. With the further development of technology, more neural networks began to invest in the field of speech recognition, such as CNN and RNN.

However, establishing a speech recognition system is a complicated process, which requires a lot of professional knowledge. Various attempts have been made in recent years to reduce the complexity of ASR, in the hope of directly mapping speech to tags. End-to-end speech recognition has been proposed. Now there are two main structures for end-to-end speech recognition: attention model and CTC. End to end technology has been applied in many aspects and has achieved remarkable results.

In this paper, I will introduce the CTC and attention model. The section 2 introduces the difference between traditional speech recognition and end-to-end speech recognition then introduce the CTC and attention model.

2. End-to-end speech recognition
This part mainly introduces what is end-to-end speech recognition and the difference between end-to-end speech recognition and traditional speech recognition.
2.1. End-to-end speech recognition

End-to-end is a system which directly maps a sequence of input acoustic features into a sequence of grapheme or words. A system which is trained to optimize criteria that are related to the final evaluation metric that we are interested in (typically, word error rate).

For Conventional ASR, most ASR systems involve separately trained acoustic, pronunciation and language model components which are trained separately. Curating pronunciation lexicon, defining phoneme sets for the particular language requires expert knowledge, and is time-consuming. Figure 1 depicts its structure.

It can be seen that end-to-end speech recognition greatly simplifies the complexity of traditional speech recognition. There is no need to manually label information, in the neural network can automatically learn language or pronunciation information, as shown in Figure 2. Now there are two main structures for end-to-end speech recognition: attention model and CTC.

2.2. Connectionist Temporal Classification (CTC)

CTC was proposed by Graves et al.\(^2\), as a way to train an acoustic model without requiring frame-level alignments. In early work, using CTC to output target phonemes is not really end-to-end, and it still requires language models. CTC allows for training an acoustic model without the need for frame-level alignments between the acoustics and the transcripts.

The acoustic model training using CTC as the loss function is a end-to-end training, which does not need to align the data in advance, but only needs an input sequence and an output sequence to be trained. Structure as shown in Figure 3. In this way, there is no need to align and label the data one by one, and the probability of CTC's direct output sequence prediction needs no external post-processing. The CTC introduces blank \(^3\)(which has no predicted value for this frame), where each classification of the prediction corresponds to a spike in the whole speech, and the other locations that are not spikes are considered blank. For a speech, the CTC ultimately outputs a sequence of spike, regardless of the duration of each phoneme.

In the case of given \(x\), the probability that the output is a label sequence \(y\) is formula(1):

\[
P(y | x) = \sum_{y \in B(y, x)} \prod_{i=1}^{T} P(y_i | x)
\]

2.3. Attention model

Attention-based Encoder-Decoder Models emerged first in the context of neural machine translation\(^4\). Attention Mechanism is mainly used to solve the problems of traditional RNN-based Seq2Seq model\(^5\[6\]. Seq2Seq model is an end-to-end machine translation model built based on an encoder and a
decoder. Encoder encodes input X into a fixed length hidden vector Z, Decoder decodes target output Y based on Z. There are two obvious problems with this model:

- All information of input X is compressed into a fixed length hidden vector Z, ignoring the length of input X. When the input sentence length is very long, especially longer than the original sentence length in the training set, the performance of the model drops sharply.
- It is unreasonable to encode input X into a fixed length and assign the same weight to each word in the sentence. For example, in machine translation, the input sentence and the output sentence are usually one or more words corresponding to the output word or words. Therefore, assigning the same weight to each word input makes no distinction and often results in model performance degradation.

The traditional Seq2Seq model lacks discrimination of input sequence X, therefore, in 2015, Kyunghyun Cho introduced the attention mechanism to solve the problem\(^{[6]}\). In 2015, were first applied in ASR\(^{[7]}\). The basic idea of Attention mechanism is that it breaks the limitation of traditional encoder-decoder structure, which is dependent on a fixed length vector inside. Figure 4 depicts the structure.

- **Encoder (analogous to AM):** Transforms input speech into higher-level representation.

\[
\hat{h}_i^{enc} = Encoder(x_i)
\]

The structure of encoder is usually PLSTM, BLSTM, CNN+LSTM.

- **Attention (alignment model):** Identifies encoded frames that are relevant to producing current output.

Attention module computes a similarity score between the decoder and each frame of the encoder.

\[
e_{u,t} = \text{score}(\hat{h}_u^{att}, \hat{h}_t^{enc})
\]

\[
a_{u,t} = \frac{\exp(e_{u,t})}{\sum_t \exp(e_{u,t})}
\]

\[
c_u = \sum_{t=1}^{T} a_{u,t} \hat{h}_t^{enc}
\]

- **Decoder (analogous to PM, LM):** It has an auto regressive operation by predicting each output token as a function of the previous prediction.

Output the hidden state \(\hat{h}_t^{dec}\)

\[
\hat{h}_t^{dec} = Decoder(y_{t-1}, c_{t-1})
\]

Generate the label prediction

\[
y_t = \text{softmax}(U[\hat{h}_t^{dec}, c_t])
\]

### 3. Conclusion

At present, the end-to-end speech recognition technology based on end-to-end technology has achieved remarkable results, but the end-to-end speech recognition based on CTC still needs language model to get better results, and how to further realize the real end-to-end speech recognition is worth paying attention to in the future.

### Acknowledgments

Supported by the Fundamental Research Funds for the Central Universities (Research on Unified Acoustic Models of Mandarin and Tibetan, 31920170145)

### References

1. Muda L, Begam M, Elamvazuthi I. “Voice recognition algorithms using Mel-frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques.”(2010)
[2] Graves et al. "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks." (ICML 2006)
[3] Alex Graves, Navdeep Jaitly. "Towards End-to-End Speech Recognition with Recurrent Neural Networks." (ICML 2014)
[4] Chan et al. "Listen, Attend and Spell." (2015)
[5] Ilya Sutskever et al. "Sequence to Sequence Learning with Neural Networks." (2014)
[6] Kyunghyun Cho et al. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation." (2015)
[7] Chorowski et al. "Attention-Based Models for Speech Recognition." (2015)
[8] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, “Speech recognition with deep recurrent neural networks.” (2013)
[9] Graves, A., Jaitly, N., and Mohamed, A.-r. “Hybrid speech recognition with deep bidirectional LSTM.” (2013)
[10] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups.” (2012)