Acoustic Modelling for Spontaneous Speech Recognition

Rashmi Kethireddy

Speech Processing Laboratory
International Institute of Information Technology, Hyderabad, India
rashmi.kethireddy@research.iiit.ac.in

1. Introduction

Based on the type of speech used, speech recognition has classified into isolated word recognition,connected word recognition, continuous speech recognition, spontaneous speech recognition [1]. The objective of this paper is to review the challenges in modeling a system used for labelling the acoustic speech vectors for continous and spontaneous speech recognitions. Sequencing and labelling is a major challenge for the speech recognition systems. The goal of automatic speech recognition (ASR) systems is to recognize and understand the speech as human does. ASR systems generally has two stages - feature extraction, classification. This paper mainly discuss the challenges faced during classification of continuous speech and methods used in segmentation and labelling of the speech vectors to overcome the challenges over a period of time. Section 2 discusses about key challenges and how they are solved till now. In section 3, discussion of the challenges which still exists.In section 4, probable solution for existing challenges.

2. Literature Survey

In this section, the structure and the performance of few state-of-art models for recognizing speech from 1970 till now. The Dragon system was developed at Carnegie-Mellon University [2] which is one of the state-of-art model for connected word recognition.This model has two stages of speech processing - one extracting the acoustic information from the signal and the other extracting already known facts from different knowledge sources (such as syntactic, semantic of language) and match with the acoustic information. The knowledge source is arranged in hierarchical fashion where the top layers are not frequently changed. Each knowledge source is a generative model. For searching for the optimal path a markov model is used. The performance of this model is very low considering the accuracy. The foundation for hidden Markov model (HMM)[3] based ASR systems was done by Carnegie-Mellon, IBM and Bell Labs in the 1970's. This classic generative probablistic model gave better performance for continuous speech recognition. In this ASR model there are four components of modelling where acoustic speech vectors has to undergo to get the transcriptions for continous speech. HMM's are statistical representations of each basic unit of sound. Basic units can be phones, biphones, triphones. Using triphone models worked better where it captured the transition between the phones. But, this lead to many triphones models where searching through them increased the complexity. Solution to this was a tied-mixture based triphone HMM models. First each HMM model is trained using Gaussian mixture model (GMM) then the system extracts the features from the unknown speech utterances and match this acoustic vector sequence to the few of most probable source text. Decoding component uses the pronunciation dictionary and language model to choose the most probable sequence which is the transcription for given unknown speech signal. Pronunciation dictionary is a look-up table where it maps each word to a sequence of phonemes. The classic language model which always worked well is the N-gram model. There were few other challenges which were not solved such as speaker adaption, environmental robustness,spontaneous speech recognition.
As the trend of recurrent neural network for sequential inputs a new model which used HMM-RNN [4] based acoustic model. This requires some preprocessing of acoustic speech vectors (segmenting) and post processing of results (to get back continuous transcripts). There are few other approaches where each component in generative model is replaced by a neural nerwork and trained independently. But, the errors in one of the component may not work well with others.
This lead a way towards end-to-end models for speech recognition using deep neural networks. Connectionist Temporal Classification (CTC) [5][6] is one of the method used for directly mapping acoustic speech vectors to phonetic sequences. This method takes a feature vector sequence as input to RNN and output layer uses softmax distribution to give the probability of each phone in pronunciation dictionary. There are certain rules imposed on transition such as a transition can be to itself or to a blank. Once a sequence of transition path is obtained, the dynamic programming techniques are used to find final output transcription. RNN transducers [7] combines CTC with RNN to predict the next phone given present phone. This gave state-of-art performance in continuous speech recognition when neural network is trained heavily.

3. Existing Challeges

There are few still existing challenges in ASR systems.

- Enormous amount of speech data is available in the internet from which neural network can be trained, however manual labelling of such poorly labelled data is an impossible or very costly task.
- Spontaneous speech varies from the speech which is used for training and testing the models. Spontaneous speech has many variations such as noise variations, code switching and mixing.
- Speech recognition of slurred speech is a still existing problem. Slurred speech occurs either by neural damage or damage to articulatories.
- These models are task dependent which purely depends on the data used for training. Any unknown or new data to train data during testing might not give better performance as the present system does for task dependent models.

4. Future Plans

The first challenge can be explored by using better unsupervised methods which can work inpar with the currect models which will have tremondous amounts of data. Next three problems can be solved by training the models by collecting the data from natural environments rather than controlled situations and also collecting it from various types of sources.

5. References

[1] S. Boruah and S. Basishtha, "A study on HMM based speech recognition system," in *2013 IEEE International Conference on Computational Intelligence and Computing Research*, 2013, pp. 1–5.

[2] J. Baker, "The dragon system–an overview," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, pp. 24–29, 1975.

[3] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, 1989.

[4] A. J. Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Transactions on Neural Networks*, vol. 5, pp. 298–305, 1994.

[5] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*. New York, NY, USA: ACM, 2006.

[6] G. Alex, *Supervised Sequence Labelling*. Springer Berlin Heidelberg, 2012.

[7] A. Graves, "Sequence transduction with recurrent neural networks," *CoRR*, 2012.