Speaker Recognition using Deep Neural Network

Xiaoyi Gu, Yue Li, Zongge Liu, Yufei Yi
Department of Statistics and Data Science, Carnegie Mellon Univerisity

Abstract

The Deep Neural Networks (DNNs) have become popular in many research fields, where they have generated results comparable or even superior to human experts. In the last few years, DNNs have been tremendously successful at speech recognition, and they have also been used in speaker recognition as well. In this project, we are seeking for designing novel deep learning architecture which would achieve low error rate on a standard National Institute of Standards and Technology (NIST)-provided data set and task. We will also try to find a optimal loss function and extract new feature representation at the same time.

Dataset and Preprocessing

We get our dataset which all NIST SRE competition from 2004 to 2008 in WAV format. Each WAV file contains a unique label that is associated with a particular speaker.

Since we have limited computational resources, we made a smaller dataset which contains 100 speakers. To convert the WAV files to mel spectrograms,

- Cut off the silences in the wav files. We define any time intervals with value 0 which are longer than 1000 timeticks (1/8 second) as silences, cut them from the original file and concatenate the rest of the segments.
- Convert the precessed sequences to mel spectrograms using mfcc function iPython Speech Features package. We use the length of the analysis window as 0.075s, the step between successive windows as 0.03s, FFT size as 1024, the number of cepstrum as 80 and the number of filters in the filterbank as 80.

Models and Methods

Model 1 — Naive Model: develop two identical networks for the two samples and obtain a fixed-length feature vector for each recording. Then pass the two feature vectors through a logistic regression layer to obtain match and mismatches.

Model 2 — Classification Model: Given an utterance, the model identifies the which speaker this utterance is produced from.

Model 3 — Speaker Identification Model: identify whether 2 utterances are from the same speaker. It consists of a feature extractor and a score calculator. Feature extractor is a non-linear PCA by CNN, and the score calculator is just dot product. Figure 1 shows the training process of the Speaker Identification Model.

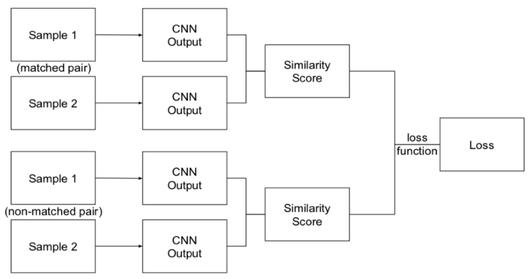
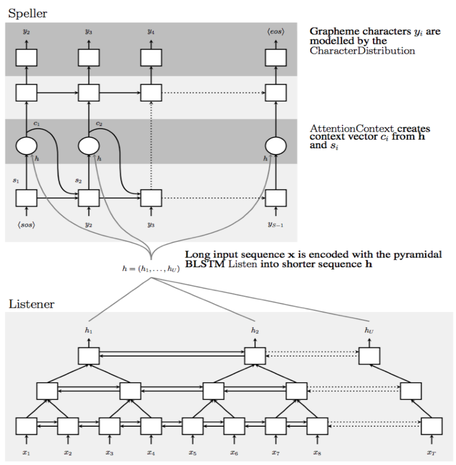


Figure 1. Training Process of Speaker Identification Model

Model 4 – Attention Model : The attention model implemented similar algorithm in the Listen, attend, spell paper. The structure is shown in figure 2.



Here, instead of using transcripts, we use the speaker id as an output. We averaged over the temporal domain to calculate the overall probability of all speakers, therefore the unique speaker is identified with the maximum probability.

Figure 2. The Structure of Attention Model.

Model 5 — Adversarial Autoencoders(AAE): The last model we use is a GAN-type model performing inference by matching the aggregated posterior of the hidden code vector of the autoencoder with an prior

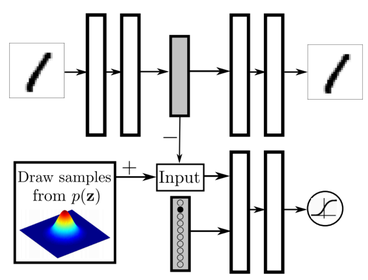


Figure 3. Diagram of AAE

The top row is a standard encoder reconstructing speech data from hidden states. The bottom left is a second network discriminatively predict whether a sample is from “true” or “fake” distribution. For semi-supervised task, we provide some additional dimensions representing label information (bottom right).

Results

Model 1 — Naive Model: To ensure the discovery rate of true similarity, we control the false negative rate at under 20%. At this level, the best accuracy Model 1 can achieve is 19.28%(accuracy on different pair is 17.98%). See figure 4 for the loss details during training.

Model 2 — Classification Model: this model gives a better result than the naive model. In the end, we are able to successfully predict 88.6% of the matched pairs and 97.8% of the unmatched pairs. The true negative rate and false positive rate are, respectively, 1.71% and 1.88%. Figure 5 shows the decrease in validation loss during training.

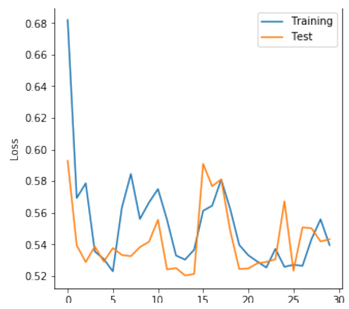


Figure 4: loss for naive model.

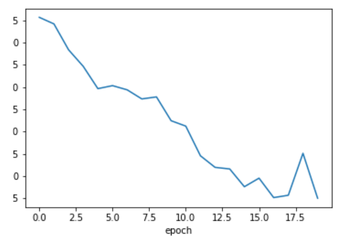


Figure 5: loss for classification model

Model 3 — Speaker Identification Model: We picked the (1 – loss/batch) quantiles of the non-match core as the threshold, which is P(non-match > match) based on the way we define loss. The scores can be visualized in Figure 6. The final accuracy for this model is 93.36%.

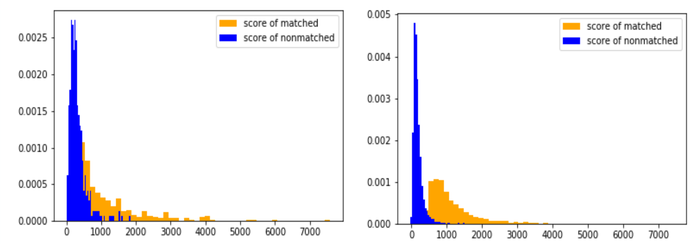


Figure 6. Scores of the Test Set Before(left) and After(right) Model 2.

Model 4 — Attention Model : The results of the attention model is shown as following figure. The training loss and validation loss keeps dropping in the first 20 epochs. The final accuracy of speaker recognition reaches to ~96%.

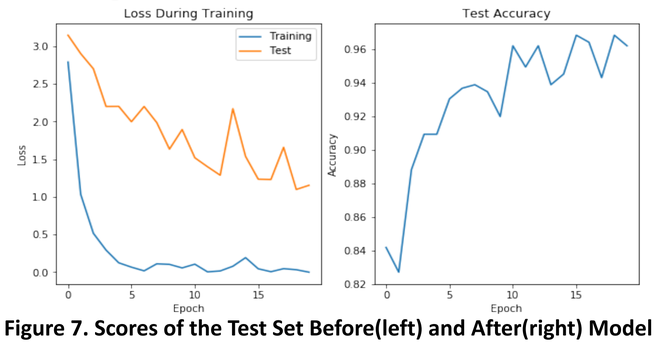


Figure 7. Scores of the Test Set Before(left) and After(right) Model 2.

Model 5 -- AAE : For decoder network, we use three-layer 1d-CNN with kernel_size=3. Encoder has inverse structure of encoder network. Mean square error loss is used for reconstruction error; cross entropy loss is used for the other two losses. In the semi-supervised learning part, we assume half of the labels are known. We plot the hidden state for each class with their mean and standard deviation in each axis; there should be 96 class in total, and we only show ⅓ of all to illustrate the method; this ⅓ sample is selected uniformly on the hidden phase plane. We should be aware that it is impossible to classify human speech data with only two dimensions; the visualization result of AAE is already surprisingly good.

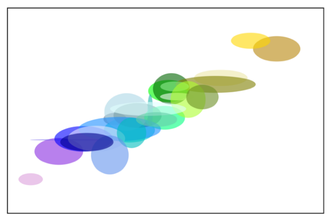


Figure 8. Visualization of AAE.

Conclusion

in summary, we build five models in order to identify whether two recordings are from the same speaker. The naive model gives a similarity metric for any two recordings by developing two identical networks, and gives an overall 19% accuracy. The classification identifies which speaker each recording comes from, and can achieve 88% accuracy. The identification model uses the classification model as a pre-training model, calculates the similarity score using the dot product of the output, and gives an overall accuracy of 93%. Out of the four models, the attention model achieves the highest accuracy of 96%. We also visualizes the hidden state of the recordings using the AAE model.