Speaker Identification with VoxCeleb DataSet

Yifan He, Zhang Zhang
CS 229 Machine Learning, Stanford University

Motivation

Speaker identification determines the identity of the speaker from a pre-known speaker set. It could be applied to areas such as voice based criminal investigations or fine tuning smart devices setting according to family member identities.

In this project, we perform a text independent speaker identification experiment with a newly released data set, VoxCeleb[1]. An deep neural network (DNN) model is trained to serve as our baseline, and it is compared with a support vector machine (SVM) model with k-mean vector quantization.

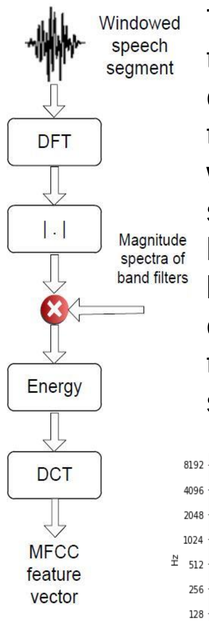
VoxCeleb Dataset

VoxCeleb is a newly released dataset from an Oxford group. It directly extracts sound clips from Youtube celebrities' interviews. It's a challenging dataset in the sense that there are complex background sounds and sometimes there are multiple speakers in the same clip.

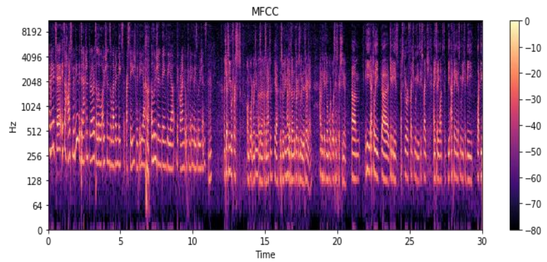
Due to computing power limitation, we only use a 30 seconds section from the original clips. This makes this data set even more challenging. In some of our samples, it may be the interviewer or some movie clips making an introduction for the target speaker, who makes few to none utterance.

Our test data set consists of 190 audio clips from 8 different celebrities.

MFCC Feature Vectors

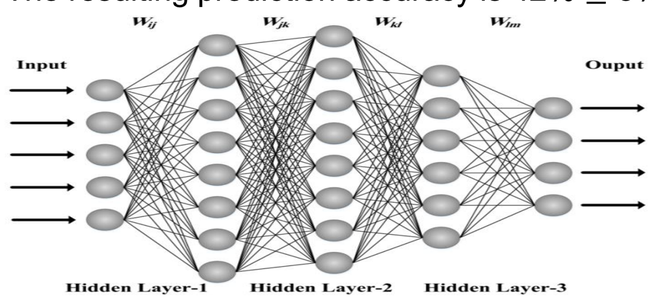


The features vector selected are the Mel-frequency Cepstrum coefficients (MFCC). It transforms the frequencies to a mel scale, which has two set of filters: one spaced linearly for frequencies below 1000 Hz and one spaced logarithmically above 1000 Hz. we extract 20 MFCC features for each time frames that lasted 93 ms and spaced 23 ms apart.



DNN Model

The DNN consists of 3 fully connected hidden layers of neuron numbers of 32, 32 and 16 respectively, with RELU as activation function. The resulting prediction accuracy is 42% $\pm$ 5%.



SVM Results and Discussions

For multi-class classification problem, there are two popular strategies to extend the binary SVM, which are one-against-all (OAA) and one-against-one (OAO). We explore both of these methods.

SVM algorithm scales super-linearly with number of samples and our sample size greatly exceeds our computation power. To solve this problem, we reduces our training sample size by performing a k-mean vector quantization (VQ) procedure first. MFCC feature vectors belonged to the each classification label are reduced to their 150 k-mean cluster center vectors.

Model	Accuracy %	Error %
SVM (OAA)	58	7
SVM (OAO)	52	8
DNN	42	5

We think the fact that the DNN model trained on the whole set is worse than the SVM on a reduced set is caused by the VQ process. It smoothed out the noise introduced by other sources such as the interviewer and other speakers. Therefore, VQ maybe a helpful procedure for training on multi speaker audio samples.

Future Work

We wish to apply the Total Variability Model (TVM) on our MFCC features and calculate the i-vectors. It separates the variability from the speakers to that from the sessions and hopefully can reduce some of the errors due to other vocal sources from the environment.

[1] Arsha. N., Joon S. C., Andrew Z. (2017) VoxCeleb: a large-scale speaker identification dataset. arXiv:1706.08612.