A Text-independent Speaker Verification System Using Filterbank Cepstral Parameters and Machine Learning
Class Project for ECE 591-02

David C. Anchieta

Spring 2019

1 Introduction

This work describes the theory and implementation of a text-independent speaker verification system. Initially, this system was intended to be part of a security application like a voice activated padlock or some telephone system. However, during the development it was realised that the reliability level required for a security application couldn't be guaranteed. In spite of that, the method could still be combined with speech recognition systems to tag text from live TV or unscripted media. The tagged text could be used in captions or to generate a database of everything that is said on a live news channel, for example.

The parametrization of speech is based on the one proposed in [1] and is described in sections 2 and 3. Section 4 describes the machine learning prediction model using TensorFlow. Section 5 present the results and discussion.

2 Preparing speech for processing

In order to train or to be evaluated by a prediction model, a piece of speech must be converted into a set of feature vectors. According to Bimbot et al [1], most of the speaker verification systems rely on cepstral representation of speech. This section describes the parametrization of the utterances via filterbank cepstral analysis and how it was used to generate the feature vectors. The method used to generate the feature vectors for this project is summarized in Figure 1

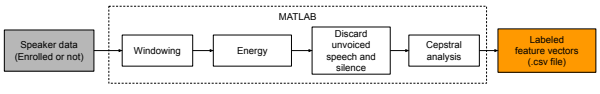

Figure 1: Block diagram for the generation of the feature vectors.

The utterances used in this project were obtained from the VoxCeleb1 dataset availabe in `http://www.robots.ox.ac.uk/~vgg/data/voxceleb/` [2]. This dataset contains over 100,000 utterances from 1,251 celebrities extracted from interview videos on YouTube.

Each of the selected utterances was split into snapshots of 30 milliseconds with a 10 milliseconds overlap between adjacent snapshots. The short-time energy for one of the utterances calculated using the snapshots is shown in Figure 2. The short-time energy (STE) will be used to decide which segments of the utterance are useful to train and evaluate the model.

The voiced speech such as vowels make the high energy snapshots, while silence and unvoiced speech such as fricatives make the low energy snapshots. For each utterance only the snapshots with total energy greater than 15% of the maximum of the STE were kept for processing. This was enough to discard all silence and unvoiced speech.

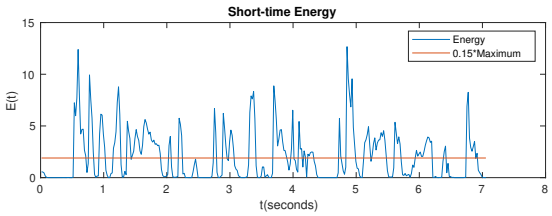

Figure 2: Short-time energy plot of one of the samples. This also shows the threshold used to decide discard silence and unvoiced speech.

3 Feature vectors with filterbank cepstral coefficients

After the segmentation and isolation of voiced speech, the power spectrum of each snapshot was calculated. This was done via a 4096 point FFT. Since the spectrum is symmetric, the latter 2048 points of the FFT were discarded. The power spectrum is the square of this FFT, which is shown in Figure 3.

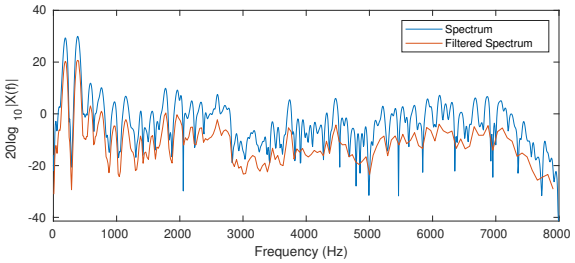

Figure 3: 2048 point spectrum was filtered with a 256 window filterbank.

As noted by Bimbot [1], the spectrum of the signal has several oscilations that are not useful for the cepstral analysis proposed here. So a mel-scale filterbank analogous to a moving average was applied to the spectrum of the signal to obtain an envelope of it. Figure 4 shows an example of a mel-scale filterbank with 20 filters. The samples for this project were filtered with a 256 window filterbank and the result of this filtering is shown in Figure 3.

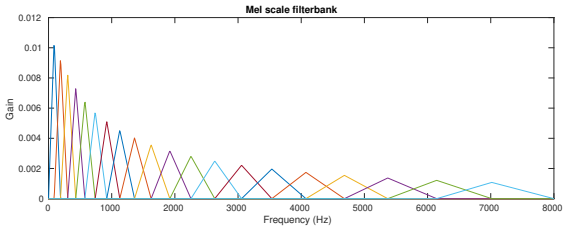

Figure 4: A 20-window mel scale filterbank. Each window is made so the area below them is equal to 1.

The cosine transformation in (1) was then applied to the filtered power spectrum. This transformation yields the cepstral coefficients [1].

$$c_n = \sum_{k=1}^{K} S_k \cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right], n = 1, 2, \ldots, L \quad (1)$$

Where K is the number of windows in the filterbank, L is the number of cepstral coefficients to be calculated and S_k is the power spectrum. Also $L \leq K$.

Some of the cepstral coefficients for one of the utterances used in this project is shown in Figure 5.

The feature vectors used to train and evaluate the prediction model are composed by the reduced and centered cepstral coefficients. Those are obtained by subtracting the mean vector from the cepstral coefficients and dividing them by their standard deviation so their mean is centered at zero and variance is normalized to one. The result of this is shown in Figure 6.

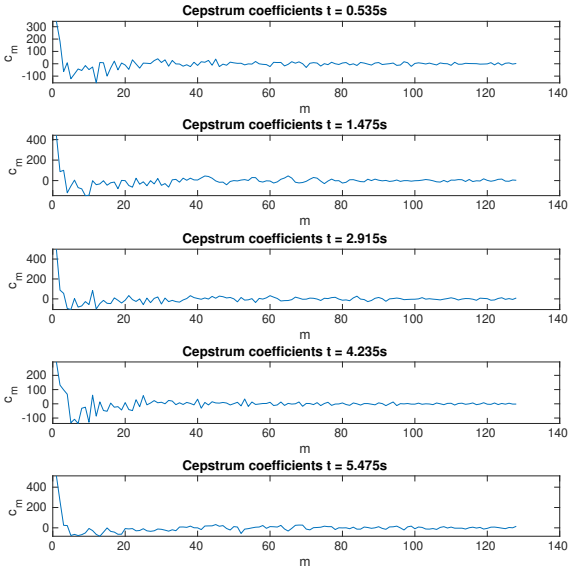

Figure 5: Cesptral coefficients obtained via cosine transform of the power spectrum.

4 Development and training of the prediction model

The previous text-independent speaker verification systems use statistical methods like Hidden Markov Models or Gaussian Mixture Models to make the decision system. However, to decrease development time, an already established machine learning library was used to develop the prediction model for this work.

The machine learning system used for this project was the TensorFlow `http://tensorflow.org`, which is a free and open-source Python library by Google. This library provides comprehensive tools to develop, train and evaluate artificial neural networks.

A general model of an artificial neural network like the one generated with TensorFlow is shown in Figure 7. The input layer represents the feature vector and its components. The output layer has only two nodes, one for probability of acceptance and one for probability of rejection.

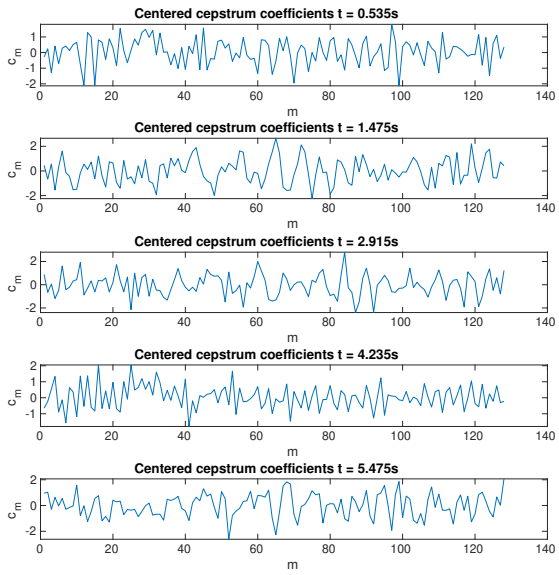

Figure 6: Cesptral coefficients are centered at mean and divided by standard deviation. Those are the feature vectors used in the predicction model.

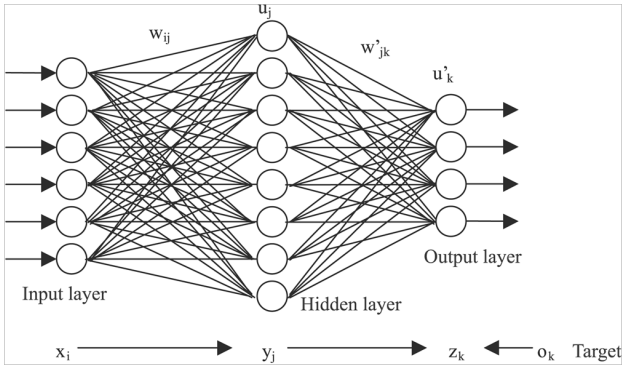

Figure 7: General model of a neural network. Each branch is associated with a weight. In the training step, the weights are adjusted iteratively to fit to the training set. Source: `www.extremetech.com`

The Figure 8 shows the block diagram for the training and evaluation of the prediction model. First, a training set with labels for enrolled and not enrolled is fed into TensorFlow, which will generate a prediction model that fits this set. This prediction model will then be used to evaluate the candidate samples.

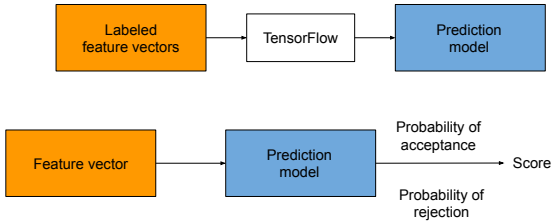

Figure 8: Block diagram of the training and evaluation steps of the prediction model.

The output of the prediction model is a probability of acceptance and a probability of rejection. The base 10 logarithm of the ratio of those probabilities is the score. The higher score, the more likely that the evaluated utterance belongs to the enrolled speaker.

5 Results and discussion

A small set of utterances from 3 speakers was used to train the prediction model. The objective is to train the system to verify if a candidate voice belongs to the enrolled speaker (Chris Martin) or not. To achieve this, the model was trained with about 50 seconds of speech from the enrolled speaker and about 1 minute and 20 seconds of speech from two non enrolled speakers (one male and one female). The Table 1, shows some details of this training set.

Subject	Utter.	Time	Vectors
Chris Martin (Enrolled)	5	54.1s	627
Cillian Murphy	5	54.8s	243
Tilda Swinton	4	31.9s	183

Table 1: Features of the samples used to train the prediction model. Table shows number of utterances, total time of speech and number of feature vectors obtained from the utterances. Samples were obtained from the VoxCeleb dataset.

The model was trained with 50 iterations of the fitting algorithm. This was enough for the model to fit to the training set with maximum accuracy.

The resulting prediction model was tested with two utterances from the enrolled speaker, one utterance from each of the same non enrolled speakers in the training set and one utterance from a non enrolled speaker which wasn't in the training set. The results are summarized in Table 2.

```
Enrolled speaker
================

Chris Martin's speech:
Probability of acceptance:      0.940666
Probability of rejection:       0.05933421
Score:                          1.2001303


Chris Martin's speech 2:
Probability of acceptance:      0.8600377
Probability of rejection:       0.13996238
Score:                          0.78850615


Unenrolled speaker
==================

Cillian Murphy's speech:
Probability of acceptance:      0.3224028
Probability of rejection:       0.6775971
Score:                          -0.32257274


Tilda Swinton's speech:
Probability of acceptance:      0.24345122
Probability of rejection:       0.7565487
Score:                          -0.49242496


Jesse Eisemberg's speech:
Probability of acceptance:      0.4761958
Probability of rejection:       0.5238043
Score:                          -0.041383505
```

Table 2: Output result of the prediction model.

As seen in the results, the prediction model was successful in distinguishing The utterances of the enrolled speaker from the ones of the non enrolled speakers. Although the results are positive, more research must be done to verify the effectiveness of the voice as a mean of biometric verification.

All the code wrote for this project is hosted in the GitHub repository `https://github.com/danchieta/ece591speech_project`.

References

[1] Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Advances in Signal Processing*, 2004(4):101962, 2004.

[2] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.