Machine Learning Engineer Nanodegree

Capstone Proposal

Waleed Algadhi

March 15, 2018

Domain Background

Speaker identification is the task of recognizing speakers based on their voice. It is used to decide whether an unknown speaker is a specific person or belongs to a given group of persons (Bakmand et al., 2007).

In the past, multiple approaches were used to solve this problem. These include: Pattern Recognition approaches (e.g. Hidden Markov Model), Knowledge based approaches that use Vector Quantization and Learning based approaches (e.g. Artificial Neural Networks) (Gupta et al., 2014).

The latter, Artificial Neural Networks, are currently the most effective models for speaker recognition. All research explored in a review of the literature has provided support for their superior performance (Omar, 2017).

The goal is to build a model that serves as a “shazam” for Holy Quran reciters. It identifies the reciter by name given an audio clip of him reciting verses from Quran.

My motivation for building such a model is that Quran is recited by so many reciters around the world, each having a different way of reciting and voice characteristics. Some people might find themselves listening to a recitation they like on the radio or YouTube that doesn’t include the reciter name. This model will help them identify the reciter in that case.

Problem Statement

In this project, I will build a model that identifies a Quran reciter’s name given an audio clip of his recitation. This could be not straight forward since a lot of old Quran recitations and recitations shared on social media are noisy or of poor sound quality.

The model is supposed to run live on a server receiving queries in real-time and therefore time efficiency is a concern.
Finally, I found a paper that aims to build a similar model (Asda et al., 2016).

Datasets and Inputs

The recitation clips for all reciters are downloaded from Quran2y website. The website has a link for each reciter to download a zip file that contains multiple of his recitation clips. I chose a random set of 50 reciters, and I'm looking to use about 10 audio files for each reciter to extract the features from and then train the model on. Testing will include 5-6 audio clips for each reciter. Consequently, the labels are balanced in distribution.

Quran2y web page: https://quran2y.blogspot.com/p/full-quran-mp3-download-zip.html

Solution Statement

To build a model that is able to identify Quran reciters, I will use Mel-frequency cepstrum coefficients to extract sound features from each recitation. These features are going to be used to train a recurrent neural network to extract the deep sound features for each reciter.

Those extracted features are going to be fitted into a K-nearest neighbor's algorithm so that any new recitation clip can be matched to the most similar sound features in the algorithm.

Benchmark Model

This paper (Asda et al., 2016) is the only paper I found that tries to tackle the same problem. The author reports that they achieved an accuracy of 91% using a dataset that contains sound clips for 5 different reciters.

However, in section 4.2 it seems that their evaluation was on training, validation and testing set all together (?!).

Consequently, results from this paper can not be an accurate benchmark for my model. Therefore, I used a simple logistic regression with the default parameters of sklearn library (0.19.1) as my benchmark model on the same training and testing sets. the Accuracy score on this model was 0.53 and this is going to be my benchmark model.

Evaluation Metrics

The model will be evaluated on accuracy metric which is going to be applied on the testing set. Accuracy metric is applicable here because labels are balanced in distribution as mentioned above. Accuracy score can be mathematically defined as:

Accuracy = Number of true predictions/Number of all predictions

To evaluate the model, each reciter has a unique name that labels every sound clip of his. Those labels are going to be compared to the predicted labels to compute the accuracy score.

Project Design

1. Data acquisition: finding adequate sound clips for numerous Quran reciters and unify each reciter's label.
2. Data preprocessing: convert all sound clips to WAV format, segment them into 40 milliseconds. Since there is going to be a huge amount of small segmented clips for each reciter, a random sample from is going to be taken from each reciter.
3. Training: different types and architectures of artificial neural networks are going to be used for training to explore which ones are going to produce better results. A validation set is going to be provided to monitor the performance of the algorithm and make sure it does not overfit.
4. (to experiment): use the artificial neural network only to extract features and leave the decision for K-nearest neighbors to label the reciter.
5. Testing: last step is to apply the chosen algorithm on the testing set and measure its performance using the metric specified above.

References

1. Gupta, Shikha, et al. A Study on Speech Recognition System: a Literature Review. International Journal of Science, Engineering and Technology Research (IJSETR), Aug. 2014.
2. Bakmand-Mikalski, Dan. Speaker Identification. Master thesis, Oct. 2007.
3. Omar, Najiya. Speaker Identification System Enhanced by Optimized Neural Networks. Master thesis, Feb. 2017.
4. Asda, Tayseer, et al. Development of Quran Reciter Identification System Using MFCC and Neural Network. TELKOMNIKA Indonesian Journal of Electrical Engineering, Jan. 2016.