Singer Traits Identification using Deep Neural Network

Zhengshan Shi
Center for Computer Research in Music and Acoustics
Stanford University
kittyshi@stanford.edu

Abstract

The author investigates automatic recognition of singers' gender and age through audio features using deep neural network (DNN). Features of each singing voice, fundamental frequency and Mel-Frequency Cepstrum Coefficients (MFCC) are extracted for neural network training. 10,000 singing voice from Smule's Sing! Karaoke app is used for training and evaluation, and the DNN-based method achieves an average recall of 91% for gender classification and 36% for age identification.

1 Introduction

Music exhibits some similarity of structural regularity like natural language, inspired by speech recognition, a good model of musical language can help the music transcription and classification problem. Male and female have different singing pitch range as well as timbre, recognizing singer traits can help improve vocal quality over phone, as well as help collecting user information. In this project, the author applied techniques about deep learning in natural language processing into analysis of musical language. This project aims to identify singer traits (gender/age/race/etc.) through singing voice of the popular songs based on acoustic features.

First, 30,000 recordings of singing voices were collected as training and evaluation dataset, second a deep neural network model was trained. Results show that our best model outperforms traditional method using conventional acoustic features.

2 Background

Speech recognition is a major branch in natural language processing. Recent years, techniques on deep neural networks have been used on audio classification. Lee et al (2009) extensively apply deep learning approaches on auditory data, using convolutional deep belief networks for various audio classification tasks especially for speech data on phones.

As auditory signals, music exhibits some similarity as speech. So many methods applied on speech recognition have been migrated into music data recognition. Music Information Retrieval (MIR) is an interdisciplinary science between audio signal processing and music information analysis, extracting information from music including audio or meta-data. MIR is as paralinguistic speech processing in speech world. Common tasks in MIR includes cover song identification, music melody extraction, song chord recognition, music recommendation, etc. Commercialized software such as Shazam[1] and Soundhound[2] automatically recognize the song being played. They match audio

[1] http://www.shazam.com/
[2] http://www.soundhound.com/

fingerprints to existed songs in the database and pick one that maximizes the audio alignment. However, automatically recognizing meta data from recorded music is still an unexploited area. Few research have been conducted on identifying singer traits given that no pre-stored database is provided. Weninger et al. (2011) investigated automatic extraction of singers' gender, age, height and race from recorded popular music. In their approach, they identify beat-wise information using Bidirectinoal Long Short-Term Memory (BLSTM) recurrent neural networks, with two hidden layers, reaching an unweighted accuracy in unseen test data of 89.6 % for gender, and 57.6 % for age.

3 Approach

The overall architecture is summarized in the following chart.

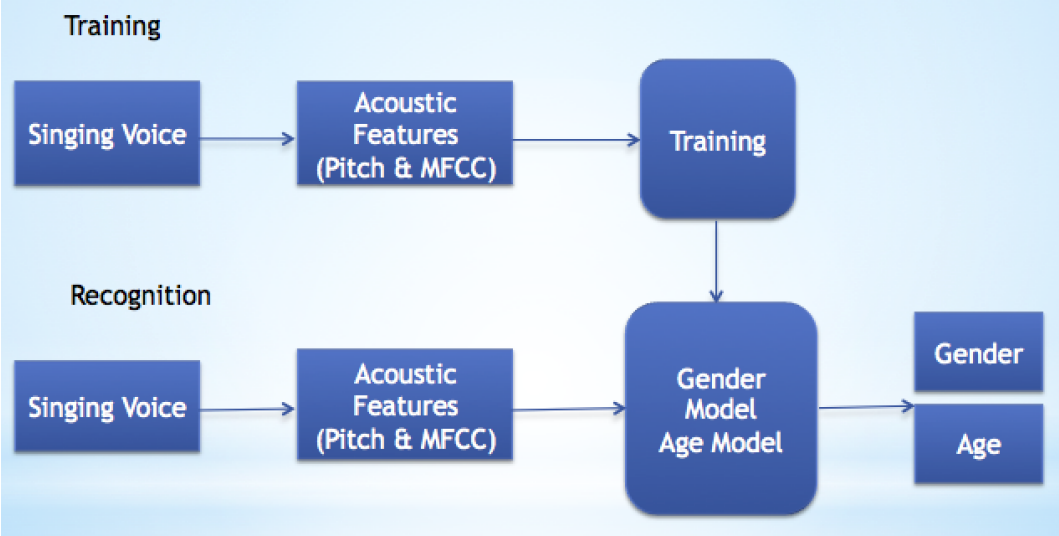

Figure 1: Architecture

3.1 Acoustic Feature Extraction

Extracting acoustic features from such raw audio is a first challenge. Mel-Frequency Cepstrum Coefficients (MFCC) extracted from the raw audio data can be used for timbre feature representation, which is also commonly used in speech recognition tasks. It represents low-level energy change as shown in the figure.

However, the major difference between speech and music is, music contains pitch and rhythmic information. Songs are representedk in audio wave files, in which there are 44,100 samples for one second audio clip. With the help of Fourier Transform and other signal processing techniques, we are able to get frequency representation of an audio file, in which we can extract pitch information – an important attribute of musical tones with duration, loudness and timbre. In addition to fundamental frequency of each note sung by the users, the overall min frequency, max frequency, median frequency, standard deviation of frequency, as well as pitch distribution range are extracted. As in the figure 2, from the distribution (normalized) of pitch range for male and female, we see a vague boundary for gender(male and female) versus pitch range (in Hz).

OpenSmile (Open Speech and Music Interpretation by Large Space Extraction)[3] framework is used for feature extraction. Each second of the song was divided into 50 frames, and for each frame, MFCC information and pitch information is extracted. Then, these two features are stacked together as a vector.

[3]http://sourceforge.net/projects/opensmile/

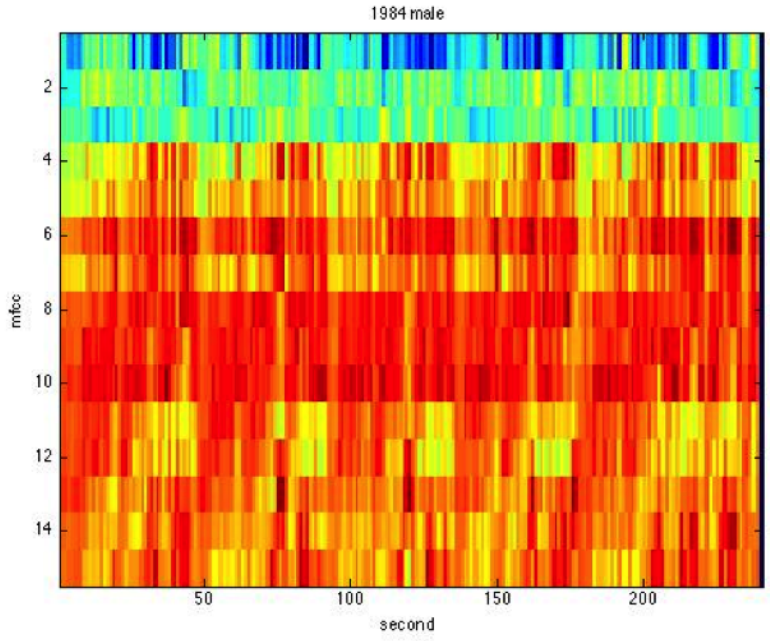

Figure 2: MFCC for a 1984-born male

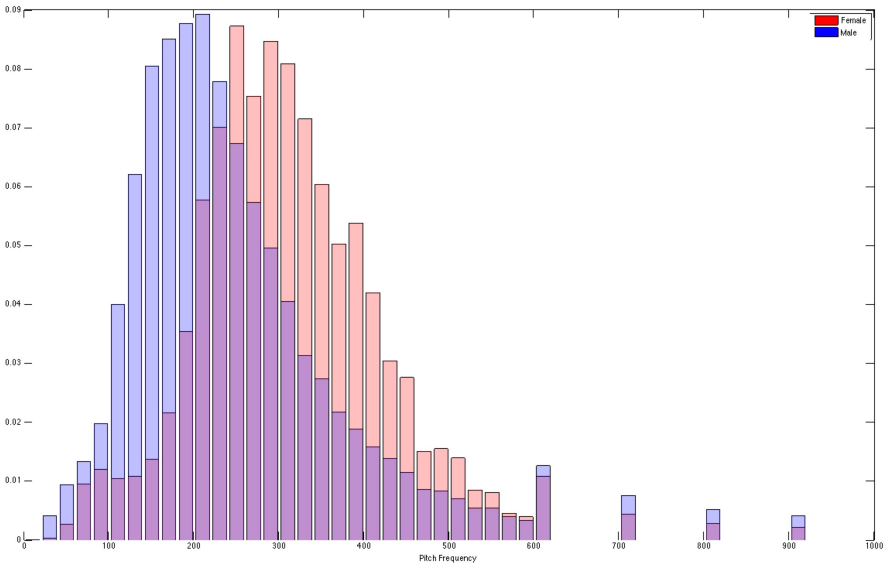

Figure 3: Pitch distribution versus gender

3.2 Baseline algorithm

A baseline algorithm of logistic regression was used for predicting singer gender and age from the two features above. For the baseline, I fit a simple logistic regression model to predict the gender as well as the birth year of the singers. Parameters were chosen using 10-fold cross validation.

3.3 Deep Neural Network Model

The proposed deep structured acoustic model is trained by maximizing the likelihood of the gender/age given a short sound clip. The model is evaluated on a dataset consisting 1,600 hours of sound clips. Results show that our best model outperforms traditional method using conventional features.

The input (raw audio features) to the DNN is a 15xN-dimensional vector, e.g., the first 13 coefficients are MFCC data, and the last two are pitch. More formally, if we denote x as the input vector, y as the output vector, h_i as the intermediate hidden layers, W_i as the i-th weight matrix, and b_i as the i-th bias term. The overall formula are as following:

$$
\begin{aligned}
l_1 &= W_1 * x \\
l_2 &= ReLU(W_1 * l_1 + b_1) \\
l_3 &= ReLU(W_2 * l_2 + b_2) \\
y_1 &= Sigmoid(W_3 * l_3 + b_3) \\
y_2 &= W_3 * l_3 + b_3
\end{aligned}
$$

where we use ReLu as the activation funciton at the hidden layers and sigmoid at the output layer. Two models were trained separately for age and for gender, as we see in Figure 4. The final layer is linear for age identification and is a sigmoid function for gender classification since the second task is a binary classifier. The models are illustrated in next figure. The model was trained in mini-batch

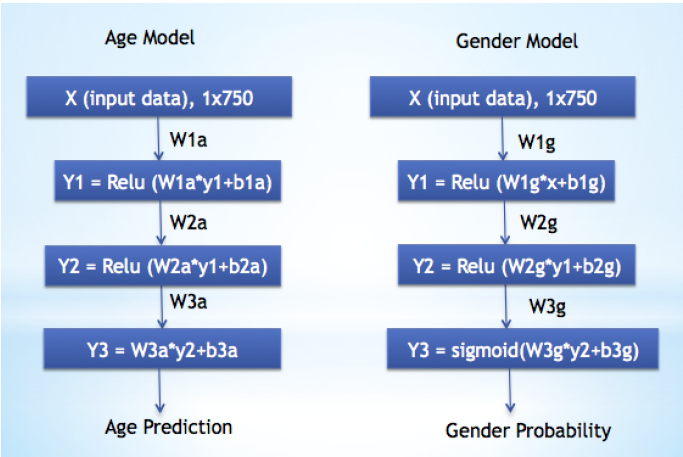

Figure 4: Model for gender and age identification

fashion with batch size of 1000 examples.

4 Experiments and Results

4.1 Dataset

The dataset used in this project consists of 30,000 raw audio singing performance from Sing! Karaoke by Smule[4], in which 2,000 are for training and 10,000 are for testing. The female-to-male ratio is 2 to 1, and age distribution is from 16 to 65. So the testing set is approximate 1600 hours of music. Sing! Karaoke is a music social app that encourages people to sing songs and upload their songs to the server. The specific data used in this project are recorded singing voice that the

[4]http://www.smule.com/

TRAIT	Baseline	DNN
Gender	89%	91%
Age	23%	36%

Table 1: Model Accuracy

users upload to the server in Sep. 2013. This dataset also includes the metadata of the performances including the player ID of the user who sang the performance, the song ID, the performance key, and some other fields including user gender and age. In this work, only age and gender info are used.

4.2 Evaluation and Result

The main result is summarized in Table 1, where we compared the model with a base-line model using logistic regression. Our DNN model outperforms the baseline. For baseline, the average accuracy is 89% for gender prediction, and 23% for age prediction. For DNN, the model converges after 40 iterations, and the average accuracy is 91% and 36% for gender and age prediction, respectively. A live demo in is also provided with pre-trained data information. Using Audacity[5], we can do live recording of the singer's voice, and load them into the GUI interface to see the final prediction result. Demo graphs are attached below.

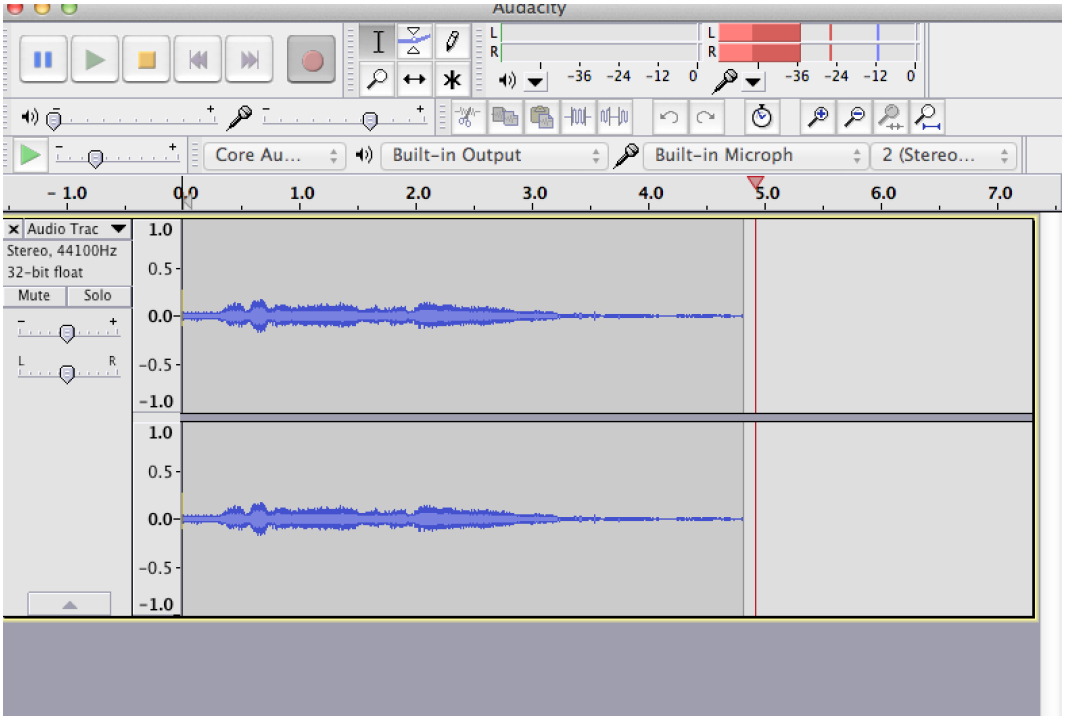

Figure 5: Recording interface

5 Conclusion

This work is an approach to apply deep neural network models into music signal processing. Singer traits such as age and gender are predicted using raw audio data with extracted MFCC and pitch as feature vectors. A 2-hidden-layer deep neural network was used, with a final linear layer for

[5]http://www.audacityteam.org/

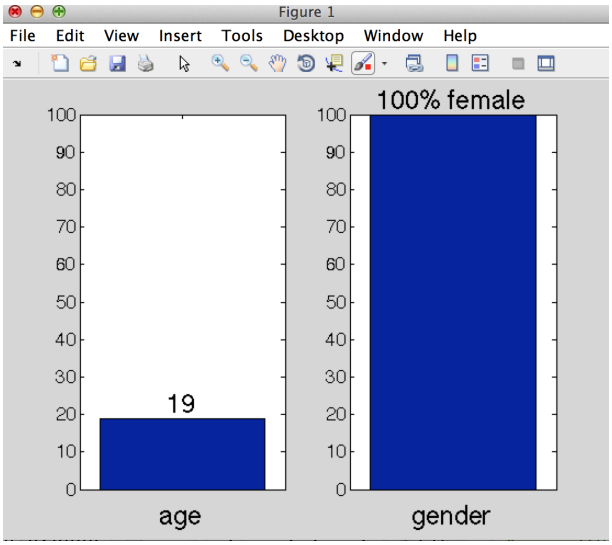

Figure 6: Prediction interface

age prediction and sigmoid function for gender prediction.This work shows very promising result using DNN-based method to predict singers' traits. In the future, we can include more singer's information in the model as well as exploring more features from raw audio data. A more complex model architecture such as deep belief network can also be explored.

6 Reference

[1] H. Lee, Y. Largman, P. Pham, A. Y. Ng. (2009) Unsupervised feature learning for audio classification using convolutional deep belief networks, NIPS 2009.

[2] Z.Fu, G. Lu, K. Ting, D. Zhang. (2011) A Survey of Audio-Based Music Classification and Annotation. IEEE transactions on multimedia, vol.13, no.2.

[3] E. Schmidt, P. Hamel, E. Humphrey (2013) Deep learning in MIR: Demystifying the Dark Art. 14th International Society for Music Information Retrieval Conference.

[4] F. Weninger, M. Wollmer, B. Schuller (2011) Automatic Assessment of Singer traits in popular music: gender, age, height and race. 12th International Society for Music Information Retrieval Conference.

[5] F. Weninger, J.-L. Durrieu, F. Eyben, G. Richard, and B. Schuller (2011). Combining Monoaural Source Separa- tion With Long Short-Term Memory for Increased Ro- bustness in Vocalist Gender Recognition. In Proc. of ICASSP, Prague, Czech Republic.

[6] A. Mesaros, T. Virtanen, and A. Klapuri. (2007) Singer identi- fication in polyphonic music using vocal separation and pattern recognition methods. In Proc. of ISMIR, pages 375?378.