Teaching Computers to Identify Singers

Kai Hayashi
Northwestern University
k.hermitian@gmail.com

Cary Lee
Northwestern University
carylee@gmail.com

Daniel Myers
Northwestern University
dmritard96@gmail.com

Rebecca Nevin
Northwestern University
rlnevin@gmail.com

ABSTRACT

This research attempts to use machine learning techniques to identify singers based on short single-pitch samples of their singing voice. Composite transfer functions and Mel frequency cepstral coefficents were used to process audio signals which were then classified according to idividual, voice type and sex. Classification by sex was very successful while voice type and individual identification were mostly successful with noticeable misclassification trends. This gave rise to a study of how spectral density and vocal maturity are added challenges when differentiating female voices. Future work could focus on assisting voice teachers with determining students' voice types or the classification of more specific fachs.

1. INTRODUCTION

There is significant interest in teaching machines to automatically identify singers similar to the way that comes naturally to human listeners. Since each voice is unique, there are many applications of an automated identifier such as source differentiation in audio recordings or the identification of a singer's presence in a corpus of unlabeled tracks.

1.1 Background

When a human recognizes a singer's voice, the primary distinguishing feature is the singer's timbre. One of the most important components of what a human listener perceives as timbre is the spectral envelope of the singer's voice. The peaks of that spectral envelope are called formants. Using the spectral envelope and the formants to identify a singer is particularly useful because, as Khine explains in [2], "studies suggest that timbre is invariant with an individual singer." This feature will greatly facilitate the machine's ability to learn different singers.

1.2 Overview

We will expand on the work of Wakefield, Bartsch and others by labeling voice samples with the singers who produced them using machine learning techniques. Section 2 presents the methods used for data collection, data processing, and the learning algorithms. Section 3 discusses our findings and section 4 suggests future work.

2. METHODOLOGY

2.1 Data Collection

Data was collected in a similar manner to [4]. A total of 11 classically trained singers from the Northwestern University Bienen School of Music were recorded. The distribution of singers was 7 females and 4 males which can be further divided by voice type. The female group consisted of 3 sopranos and 4 mezzo-sopranos and the male group had 2 tenors and 2 baritones. Every vocalist sang the first 5 notes of 3 major scales originating in their lower, middle and upper ranges for each of the 5 common Italian language vowels. This resulted in a total of 75 1-second-per-pitch samples for each vocalist and a total data set of 825 single-note sample recordings. In addition, a contextual sample was recorded for each singer. Recordings were made on a Zoom H4 Handy Recorder with a sampling rate of 44.1kHz and a bit rate of 16 bits/sample.

2.2 Signal Processing

Formants, as discussed above, were the key feature in the learning algorithm and were calculated for each sample. Composite Transfer Functions (CTFs), as used in [4], and Mel Frequency Cepstral Coefficients (MFCCs), as described by [3], are different methods for caclulating the formant of a vocal sample from that sample's spectrogram. An autoregressive filter model of the voice tract was written in Matlab to approximate the CTF algoritm found in [4]. **melfcc.m** from Ellis was used to compute the MFCCs. In both cases

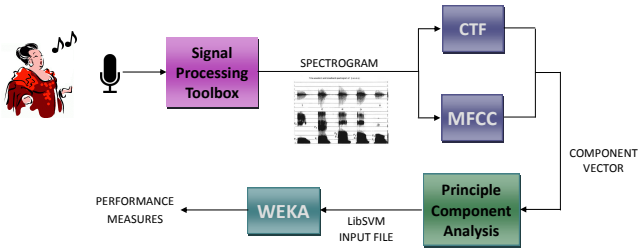
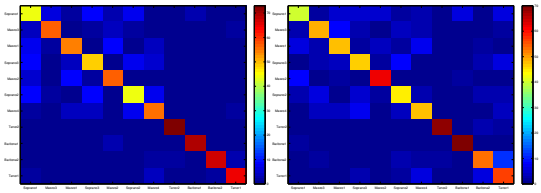


Figure 1: Block diagram of sample processing methodology

a spectrogram computed by the Matlab Signal Processing toolbox was used as the input and the results were output as a vector. MATLAB was used to perform principle component analysis (PCA), as explained in [1] and the MATLAB documentation, on the output of the MGCC calculator. This produced a single-row vector for each sample where each component was the amplitude at the transformed frequencies. A similar result was obtained for the CTF path by sampling the output of the autoregressive filter. These vectors were combined into a text file that met the input requirements for LibSVM.

After several initial rounds of training and testing it was found that MFCCs produced better overall results and thus were used as the primary method for the remaining investigations. The confusion matrices in figures 2a and 2b show that the MFCC method performed significantly better than the CTF approximation used.



(a) MFCCs Method (b) CTF Method

Figure 2: Confusion matrices for calculation methods.

2.3 Learning Algorithms

The Weka software package was used to perform the learning and testing. Specifically, LibSVM using the quadratic kernel and ten-fold cross-validation was used. Three levels of classification were attempted. Initially singers were classified individually. To further investigate specific classification challenges, the learner was generalized to classify by voice type. A third, even more general level then identified samples based on sex. Performance was measured as the percentage of correctly classified samples. The results of the learning are discussed below.

3. RESULTS AND DISCUSSION

Figure 4 summarizes the performance at each of the three classification levels. In addition to further proving the strength of the MFCCs relative to autoregressive filtering the CTF, it is clear that the learner was very successful at classifying samples based on sex (on average, 95.3% correct). However, voice type (84.7% correct) and individuals (77.2% correct) were more challenging. The quality of our results is supported by p-values in the range of 2e-27 indicating extremely consistent replicatibility.

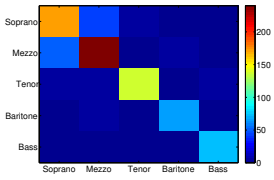
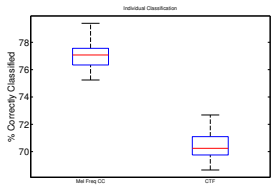


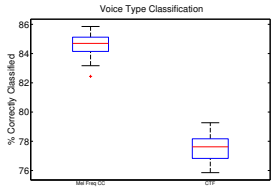
Figure 3: Voice type confusion matrix

The following discussion will focus on the two lower-level classifications where there was a noticeable misclassification trend. Figure 3 reveals that at the voice type level the majority of misclassifications occured amongst the female samples (sopranos and mezzo-sopranos). One explanation for this is the sparse spectra at higher frequencies. As Wakefield and Bartsch explain, the female voice typically has a higher fundamental frequency, such that the spectral envelopes generated by female singers are sparser than those of male singers. These sparse spectra likely contribute to the learner's difficulty in distinguishing female voices.

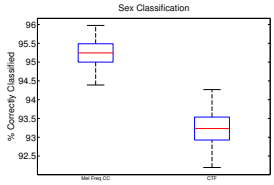
The age and maturity of the sample set must also be considered. A young woman's voice can continue to change and develop into her 30's. As such, it is not uncommon for young female singers to change voice types early in their career. The singers used in this study were all under the age of 30 and the youngest singer, Mezzo3, was currently considering a transition to soprano repertoire. Thus the results are more vulnerable to voice type discrepancies than a sample set of fully developed singers.



(a) Individuals



(b) Voice Type



(c) Sex

Figure 4: Performance at each classification level.

4. CONCLUSION

Although better performance for individual classification was desired, the results obtained present interesting opporunities for further research. For example, could machines trained on matured singers assist voice teachers in determining the voice type of young students, particularly pre-mature females? Also, how would an automated singer identifier perform when the sample set contained more specific fachs such as coloratura soprano, lyric soprano, spinto soprano, and soubrette?

5. REFERENCES

[1] I.T. Jolliffe. *Principal component analysis*. Springer series in statistics. Springer-Verlag, 2002.

[2] Swe Zin Kalayar Khine, Tin Lay Nwe, and Haizhou Li. Exploring perceptual based timbre feature for singer identification. *Computer Music Modeling and Retrieval Sense of Sounds*, pages 159–171, 2008.

[3] Beth Logan. Mel frequency cepstral coefficients for music modeling. In *In International Symposium on Music Information Retrieval*, 2000.

[4] Gregory H. Wakefield and Mark A. Bartsch. Where's caruso? singer identification by listener and machine.