March 10, 2005

Singer Similarity

Catherine Lai
lai@music.mcgill.ca

Introduction

This paper presents recent work that has been done on singer similarity. One of the most popular applications in the context of music information retrieval is singer identification, which is to determine a singer or vocalist based on extracted features of audio signal.

Background

Rapid progress in computer and network technology in recent years has resulted in multitude of audio files circulation on the Internet. In most popular music, the vocals sung by the lead singer are usually the focal point of the song. Once people have been exposed to the singing voice of an artist, the unique qualities of a singer's voice make it relatively easy for human to identify an artist's singing voice regardless of the musical context, even if they are hearing a piece for the first time. However, the presence of interfering sounds such as instruments or background noise makes the task of singer identification harder for machines.

Recent Work

A singer identification system is presented by Kim & Whitman in 2002. The approach taken by the authors makes the assumption that the song data will have strong harmonicity from the vocals, and this is employed as the main method for singer identification. Another assumption made by the authors is the use of popular music. This implies that the critical range of vocal analysis is in the range of 200-2,000 Hz, and inside this range other instrumentation will not be strong. The identification within this work is a two-step process. First, an untrained algorithm is used for automatically extracting vocal segments from within songs. Then a classification with training is applied to the vocal segments.

The first step consists of filtering all frequencies outside the vocal range of 200-2,000 Hz. This is done by the use of a Chebychev IIR digital filter. The intention is to remove as much as possible any unwanted signals that may be confused for vocals. However, as many instruments such as drums have a broad spectrum and so may still contribute energies in this critical region, a second step is taken to look for strong harmonicity in the vocal segment. The detection of harmonicity is achieved by sending this bandlimited signal through a bank of inverse comb filters with varying delays. Based on the initial assumption, the vocals should be the strongest harmonic signal in the regions. To measure the harmonicity ratio, the authors take the ratio of total signal energy to the energy of the maximally attenuated signal.

When the vocal segments of audio are known, they are presented to a singer identification system that has been trained on data taken from other songs by the same artists from a database of popular music. The first part of this requires an extraction of features from the audio signal. This is done through the use of Linear Predictive Coding (LPC). Specifically, a 12-pole linear predictor is used to find the location and magnitude of formants using the autocorrelation method.

Once the features have been extracted, two different classification techniques are used: the Gaussian Mixture Model (GMM) and Support Vector Machine (SVM). In this context, each class represents an individual singer. The Gaussian mixture model uses multiple weighted Gaussians to capture the behavior of each class of training data. SVMs work by computing an optimal hyperplane that can linearly separate two classes of data.

The data sets used in this experiment included 17 different solo singers and more than 200 songs. All songs were downsampled to 11.025 kHz to reduce the data storage and processing requirements. Half of the database was used to train the classifier and the remaining was used to evaluate the performance of the classifier. Two sets of experiments were conducted. In the first set, LPC features were extracted from entire songs and used for classification. The other set used only features from regions classified as containing vocals. A 12-pole LP analysis was performed on both linear and warped scales. Three different features sets (linear scale data, warp scale data, and both linear and warped data) were tested and two different classifiers (GMM and SVM) were used. The classification results were greater than chance. Using the linear frequency features and the warped frequency features worked the best, and in general, the linear frequency features outperformed the warped frequency features when each was used alone. One thing that puzzles the authors is that song and frame accuracy increases when using only vocal segment in the GMM, but accuracy decreases in the SVN when using the same segment.

A recognition system based on a MP3 database was presented by Liu and Huang also in 2002. The approach is to first extract coefficients from MP3 files to compute the MP3 features for segmentation. Based on these features the audio is segmented into phonemes, which represent a note in the staff or a syllable in a music sentence. Then for each MP3 phoneme in the training set, its MP3 feature is extracted and used to train an MP3 classifier, which then is used to identify the singer of an unknown MP3 file.

The kNN classifier is used to classify the unknown MP3 songs. For each unknown MP3 song, it is first segmented into a sequence of phonemes. To speed up the performance, for classification only the first N phonemes of an unknown MP3 song are used, and each of the phonemes is compared with every discriminator in the phoneme database. The k closest neighbors are found. For each of the closest neighbors, if its distance is within the threshold, it is weighted by a discriminating function. The unknown MP3 song is assigned to the singer with the largest weighted vote.

For the experiment, ten male Chinese singers and ten female Chinese singers were chosen. 30 songs were randomly picked for each of the singers. Three factors dominate the results of the MP3 music classification method: the value of k for the kNN classification, the threshold of the discrimination function, and the number of singers allowed in each class. Two experiments were conducted. The first experiment was to find the best setting for k in the kNN classifier, and it was shown that the best setting of k is 80, which will result recognition rate of 90%. The second experiment was to illustrate the effect of the threshold setting. It was found that the best setting for the threshold is 0.2. Also, three singers to a class was proven to give slightly better results than a class of two singers, and a class of two singers was shown to give better results than a class of one singer.

Another singer identification system that clusters undocumented music recordings based on their associated singers operates in an unsupervised manner (Tsai et al. 2003). The singer-based clustering is based on the singer's voice rather than the background music, musical genre, or other characteristics of the recording. The system consists of three stages: segmentation of each recording into vocal and non-vocal part, suppressing the characteristics of the background, and clustering of the recordings based on singer characteristic similarity.

For detecting singing voices, the authors constructed a statistical classifier with parametric models trained using accompanied singing voices. The classifier consists of a front-end signal process that converts digital waveforms into spectrum-based feature vectors, followed by a backend statistical processor that perform modeling, matching, and decision making.

The classifier operates in two phases, training and testing. During training, a music database with manual vocal/non-vocal transcriptions is used to form two separate Gaussian mixture models, a vocal GMM and a non-vocal GMM. In the testing phase, the recognizer takes as input the feature vectors extracted from an unknown recording and produces as outputs the frame log-likelihoods for the vocal and the non-vocal GMM. According to a decision rule, the attribute of each frame is hypothesized on the frame log-likelihoods.

To compare and cluster the singers, the authors used the k-mean algorithm, which starts with a single cluster and recursively splits clusters. The Bayesian Information Criterion (BIC) is employed to decide the best value of k.

The musical data used in the study consisted of 416 tracks from Mandarin pop music CDs. All music data were down-sampled to 20.05 kHz to exclude the high frequency components beyond the range of normal singing voices. Accuracy was computed by comparing the hypothesized attribute of each frame with the manually labeled. The best accuracy achieved was 79.8%.

Conclusion

This paper presents recent work on singer similarity systems. Although viable results have been reported in all these papers, on all the systems, more work is needed to validate for a wider variety of musical data, for example, using larger populations of singers and with different genres.

References

Kim, Y. and B. Whitman. 2002. Singer identification in popular music recordings using voice coding features. In *Proceedings of the 2002 International Symposium on Music Information Retrieval*.

Liu, C. and C. Huang. 2002. A singer identification technique for content-based classification of mp3 music objects. In *Proceedings of the 2002 Conference on Information and Knowledge Management*. 438–45.

Tsai, W., H. Wang, D. Rodgers, S. Cheng, and H. Yu. 2003. Blind clustering of popular music recording based on singer voice characteristics. In *Proceedings of the 2003 International Symposium on Music Information Retrieval*. 67–73.