SONG-SPECIFIC BOOTSTRAPPING OF SINGING VOICE STRUCTURE

George Tzanetakis

Department of Computer Science
Faculty of Engineering, University of Victoria
PO BOX 3055 STNCSC
Victoria, BC V8W3P6
Canada
gtzan@cs.uvic.ca

ABSTRACT

One of the most important characteristics of music is the singing voice. Although the identity and characteristics of the singing voice are important cues for recognizing artists, groups and musical genres, these cues have not yet been fully utilized in computer audition algorithms. A first step toward this direction is the identification of segments within a song where there is a singing voice. In this paper, we present some experiments in the automatic extraction of singing voice structure. The main characteristic of the proposed approach is that the segmentation is performed specifically for each individual song using a process we call bootstrapping. In bootstrapping a small random sampling of the song is annotated by the user. This annotation is used to learn the song-specific voice characteristics and the trained classifier is subsequently used to classify and segment the whole song. We present experimental results on a collection of pieces with jazz singers that show the potential of this approach and compare it with the traditional approach of using multiple songs for training. It is our belief that the idea of song-specific bootstrapping is applicable to other types of music and audio computer-supported annotation.

1. INTRODUCTION

The identity and characteristics of the singing voice are important cues for the recognition of artists, groups, and musical styles. Despite it's potential, singing voice information has only started being utilized in music analysis and retrieval. Locating the segments of a song where there is a singing voice is an important step toward utilizing voice information and enabling exciting new applications such as singer identification, query-by-lyrics on audio signals and extraction of musical structure. The two main problems with locating singing voice segments are : 1) the variability of singer and instrumentation characteristics 2) the difficulty of obtaining ground truth annotations to train classifiers.

In order to address these problems, a semi-automatic approach to the problem of locating singing voice segments is proposed in this paper. In this approach, which we term "bootstrapping", a small random sampling of the song is annotated manually and the information learned is used to automatically infer the singing voice structure of the entire song. More specifically the computer plays few snippets that are each 2-seconds long, and asks the user to annotate them. The main intuition behind this approach is that although it is easy for a human user to decide whether a particular snippet of music contains a singing voice or not, finding the exact boundaries and structure over the entire song can be time-consuming. Another important observation is that although singing voice and music characteristics vary a lot between different songs, they remain relatively stable within a particular song. In order to test the validity of the proposed song-specific bootstrapping approach a series of experiments were conducted. These experiments were done using a set of jazz songs with different singers. We explore the effect of the number of snippets required, necessary smoothing and compare the performance of different classifiers.

2. RELATED WORK

Music Information Retrieval (MIR) is a growing area of research that deals with the extraction of music content information for the purposes of indexing, analysis and retrieval. An overview of recent MIR activity can be found in [1]. The features used in this work are based on the feature set described in [2] for the task of automatic musical genre classification. A classic reference for the related problem of speech/voice discrimination is [3]. To the best of our knowledge the problem of locating singing voice segments within music signals, which is the main focus of this paper, was first described in [4]. In that paper the output of a neu-

ral network trained on linguistic categories was used as a feature vector for classification. The authors report classification accuracy of approximately 80% on the frame level. Singer identification in popular music recordings based on voice coding features is described in [5] and the use of voice segments to improve artist classification is explored in [6]. In all of these approach the training of the classifier is done using examples from multiple songs. In contrast our approach trains a different classifier for each song using the bootstrapping annotation information for training.

3. FEATURE EXTRACTION

Features are computed at two levels. The lowest level corresponds to approximately 20 milliseconds and forms the basic spectral analysis window over which audio features are calculated. The duration of this window is small so that the audio signal characteristics remain stationary during that window. Statistics of these audio features (means and variances) are calculated over a large size texture window, approximately 2 seconds. This texture window captures the statistical longer-term characteristics of complex audio textures such as singing or music that possibly contain a variety of different spectra [2]. Features are computed every 20 milliseconds, however the actual information used for their computation spans the 2 previous seconds. For each feature vector, classifiers are trained and a binary classification decision is made every 20 milliseconds. The low level audio features are all based on the magnitude spectrum calculated using a Short Time Fourier Transform.

We experimented with various features proposed in the literature such as spectral shape features (Centroid, Rolloff, Relative Subband Energy) [12], Mel Frequency Cepstral Coefficients (MFCC) [13] and Linear Prediction Coefficients (LPC) [14]. The final feature set we used consists of the following features: Mean Centroid, Rolloff, and Flux, Mean Relative Energy 1 (relative energy of the subband that spans the lowest 1/4th of the total bandwidth), Mean Relative Subband Energy 2 (relative energy of the second 1/4th of the total bandwidth), Standard Deviation of the Centroid, Rolloff, and Flux. More details about the definitions of these features can be found in [2]. In addition to these features, the mean and standard deviation of pitch was calculated. The pitch calculation is performed using the Average Magnitude Difference Function (AMDF) method [7] which proved to be more robust to background noise and music than other methods. This feature set showed the best singing voice classification performance using a variety of different classifiers for our data collection. It is our belief that the idea of bootstrapping can easily be applied to other alternative feature sets such as phonetic neural networks [4].

4. BOOTSTRAPPING AND CLASSIFICATION

The basic idea in bootstrapping is to use a small random collection of short duration snippets to train a machine learning classifier that is subsequently used to classify/segment the entire song. Important bootstrapping parameters are: the duration of each snippet, the number of snippets used, and the classifier.

A choice of two second snippet duration was made based on practical experience and empirical evidence about the amount of sound required to characterize audio texture [2]. Shorter segments tend to not contain enough information and longer segments tend to contain more than one type of audio texture. The number of annotated snippets is also important as it is directly correlated with the amount of user time. It is important that the classifier used for bootstrapping has good generalization properties i.e it's classification performance over the entire song is good despite the limited amount of data provided for training. The following classifiers were tried out for this purpose: *naive bayes* with a single multidimensional Gaussian distribution modeling each class, *nearest neighbor*, and *backpropagation artificial neural network*. More details about these classifiers can be found in [8]. In addition, we tried a *decision tree classifier* based on the well-known C4.5 algorithm [9], a *support vector machine* trained using the Sequential Minimal Optimization (SMO) [10] and *logistic regression* [11].

5. EXPERIMENTS

In this section, we describe a set of experiments that were conducted to investigate different choices in the number of snippets and classifier used for song-specific bootstrapping. Another goal of the experiments was to compare the performance of song-specific boot-strapping with the traditional multiple song training and classification for the same dataset.

5.1. Data collection

A collection of 10 jazz songs from the Ray Brown Trio compact disk titled "Some of my best friends are the singers" were used. Six different singers (5 female and 1 male) are featured. In some songs, in addition to the jazz trio (piano, bass, drums) additional instruments ,such as saxophone and electric guitar, are used. The songs have different tempi and styles. Each song was annotated by hand into singing and instrumental sections. The segmentation was done coarsely as a human would expect i.e small instrumental breaks of short duration where not labeled. Therefore typical segment lengths range from 30 seconds to few minutes. The minimum song duration is 2.25 minutes and the maximum is 7.17 minutes.

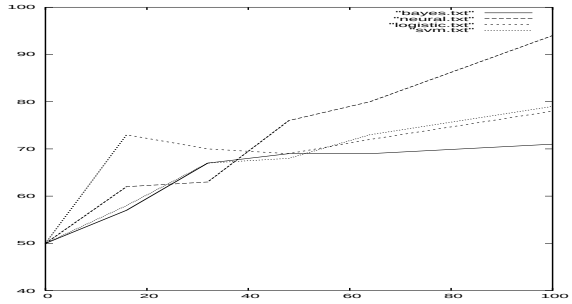

Fig. 1. Classification accuracy percentage as a function of time (seconds) for track 1

	RND	16	32	48	64	FULL
Bayes	50	57	67	69	69	71
Logistic	50	73	70	69	72	78
Neural	50	62	63	76	80	94
SVM	50	58	67	68	73	79

Table 1. Classification accuracy percentage as a function of time (seconds) for track 1

5.2. Results

An important parameter in our bootstrapping method is the number of 2-second snippets used for training. Figure 5.1 and 5.1 shows the effect of training size to frame-based classification accuracy for the first track. The RND column refers to classification without any prior knowledge or training and the full column refers to the extreme meaningless case of annotating the entire song. As can be seen even with 16 seconds and 32 seconds (8 and 16 snippets respectively) the classification accuracy is reasonable. It is important to note, that these numbers are underestimating the actual classification accuracy as the ground truth was obtained coarsely and therefore many of the small segments are not correctly accounted for. Similarly for the snippets used for training only a binary decision is provided by the user. For the few ambiguous cases the user can select to skip that particular segment. Despite these two facts, the frame-based numbers provided still are good indicators of relative performance under various parameter settings.

Table 5.2 compares the performance of different classifiers using 16 and 32 seconds of bootstrapping. These results are averaged across the 10 songs of the data collection. As can be seen, the best generalization performance is obtained using the *Logistic Regression* classifier and the *Neural Network*. The first and third column shows the classification obtained using the same classifier and the average classification performance of classifier trained using all the 16/32 second snippets. The results of the first column and third column are indicative of the standard use of multiple song examples for training the classifier. It is clear that for this particular dataset, the song-specific bootstrapping approach outperforms the traditional approach with the obvious additional burden of the bootstrapping annotation. One important detail is that in order to provide balanced training data the computer prompts the user for annotated snippets until equal number of snippets are annotated for each particular class. In practice, most song were relatively balanced so only a 2-3 additional prompts were needed. Using the bootstrapping approach a 5-minute song that would require at least 5 minutes to be manually annotated can be annotated in 16 seconds. The results with some additional smoothing are good for most practical purposes.

In some of the songs the class distributions were definitely multimodal. For example the non-singing parts contained both a piano and a saxophone solo or there was soft slow singing as well as loud and fast singing. This didn't seem to significantly affect bootstrapping performance except a few cases were only representative examples of only one of the "modes" were included in the training set. Another observation is that there is an asymmetry in classification decisions. It is much more common for a non-vocal frame to be misclassified as vocal than a vocal-frame to be misclassified as non-vocal.

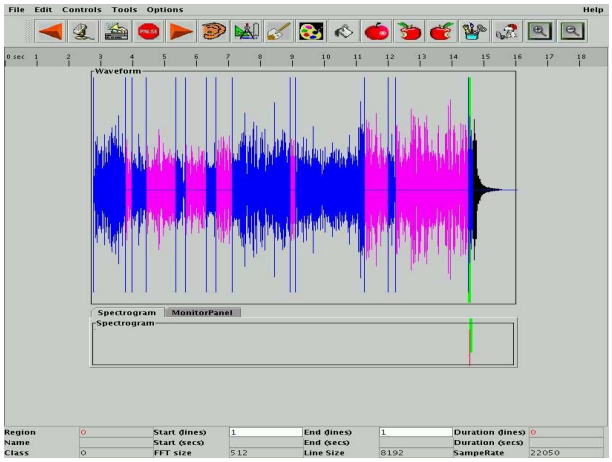

Fig. 2. User interface showing singing voice structure

	A16	16	A32	32
Bayes	58	65	59	67
J48	61	67	64	71
NN	63	65	67	73
Logistic	70	72	68	72
Neural	66	70	66	75
SVM	61	68	61	72

Table 2. Average classification accuracy of different classifiers (10 songs)

6. IMPLEMENTATION

The feature extraction and graphical user interface that was used for the annotation were implemented using Marsyas (`http://marsyas.sourceforge.net`), a free software framework for Computer Audition research described in [12]. Figure 5.2 shows a screenshot of the user interface used for annotation and experimenting with singing voice structure. The dark light colored segments correspond to singing. The classification experiments were performed using Weka: (`http://www.cs.waikato.ac.nz/ml/weka/`) a free software machine learning framework described in [13]. The calculated features and ground truth data used in this study are avialable upon request via email.

7. CONCLUSIONS - FUTURE WORK

We showed that bootstrapping is an effective technique for the semi-automatic annotation of singing voice structure. The best classifier generalization for bootstrapping was obtained with a Logistic regression or a Neural Network classifier. Using bootstrapping the singing voice structure of a 5 minute song can be automatically discovered using only 16 seconds of user time. It was also shown that for a dataset of different jazz singer, song-specific bootstrapping outperforms the traditional approach to classification that uses multiple songs.

For the future we plan to explore more powerful feature front ends such as the phonetic neural network described in [4]. In addition we want to investigate bootstrapping in more genres of music and also whether it is applicable for instrument annotation (locate the saxophone solo). The results can also further be improved using a temporal classification approaches such as Hidden Markov Models. We believe that singing information is very important for automatically understanding music and bootstrapping provides a practical way to semi-automatically extract singing voice structure. Finally, it is our hope that the idea of bootstrapping is applicable to other areas of audio and multimedia annotation in general.

8. REFERENCES

[1] Joe Futrelle and Stephen J. Downie, "Interdisciplinary Communities and Research Issues in Music Infomration Retrieval," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, Paris, France, Oct. 2002, pp. 215–221.

[2] George Tzanetakis and Perry Cook, "Musical Genre Classification of Audio Signals," *IEEE Transactions on Speech and Audio Processing*, July 2002.

[3] Eric Scheirer and Malcolm Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing ICASSP*. IEEE, 1997, pp. 1331–1334.

[4] Adam L. Berenzweig and Daniel P.W. Ellis, "Locating singing voice segments within musical signals," in *Proc. Int. Workshop on Applications of Signal Processing to Audio and Acoustics WASPAA*, Mohonk, NY, 2001, IEEE, pp. 119–123.

[5] Yongmoo E. Kim and Brian Whitman, "Singer Identification in Popular Music Recordings Using Voice Coding Features," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, Paris, France, Oct. 2002, pp. 164–169.

[6] Adam L. Berenzweig, Daniel P.W. Ellis, and Lawrence S., "Using voice segments to improve artist classification of music," in *Proc. AES Int. Conf.*, Espoo, Finland, 2002.

[7] W. Hess, *Pitch Determination of Speech Signals*, Springer Verlag, 1983.

[8] Richard Duda, Peter Hart, and David Stork, *Pattern classification*, John Wiley & Sons, New York, 2000.

[9] R. Quinlan, *C4.5 Programs for Machine Learning*, Morgan Kaufmann, 1993.

[10] S.S. Keerthi, S.K Shevade, Bhattacharya, and K.R.K. C., Murthy, "Improvements to platt's smo algorithm for svm classifier design," *Neural Computation*, vol. 13(3), pp. 637–649, 2001.

[11] S. le Cessie and J.C. van Houwelingen, "Ridge estimators in logistic regression," *Applied Statistics*, vol. 41(1), pp. 191–201, 1992.

[12] George Tzanetakis and Perry Cook, "Marsyas: A framework for audio analysis," *Organised Sound*, vol. 4(3), 2000.

[13] I.H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques with Java Implementations*, Morgan Kaufmann, 1999.