AUDIO-BASED GENDER IDENTIFICATION USING BOOTSTRAPPING

George Tzanetakis

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

ABSTRACT

Annotation of audio content is an important component of modern multimedia information retrieval systems. Automatic gender identification is used for video indexing and can improve speech recognition results by using gender-specific classifiers. Gender identification in large datasets is difficult because of the large variability in speaker characteristics. Bootstrapping is an approach that attempts to combine minimal user annotations with automatic techniques for audio classification. In bootstrapping a small random sampling of the training data is annotated by the user and this annotation is used to train a classifier that annotates the remaining data. This technique is useful when the training set is too large to be fully annotated by the user. Experimental results showing that bootstrapping is effective for automatic audio-based gender identification are provided.

1. INTRODUCTION

Annotating audio data is becoming increasingly important for multimedia information retrieval. The focus of this paper is time-based audio annotations of gender (i.e identifying audio segments that contain only male or female speech). Manually locating the exact time boundaries and categories of each segment is tedious. On the other hand, existing automatic approaches rely on training classifiers using large collections of data. Manually annotating these collection (necessary for training) can be very time-consuming. In order to address this problem, a semi-automatic approach called “bootstrapping” can be used. In this approach, a small random sampling of the training data is annotated manually and the information learned is used to train “bootstrapping” classifiers that are then used to annotate the entire set of training data. The resulting semi-automatic annotation can then be used, in the usual fashion, as ground truth to train classifiers using the entire data set. Obviously the ground truth obtained will not be as good as a complete manual annotation but will still be very useful if the boostrapping classifier has good performance.

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 the category of a snippet, finding the exact boundaries and structure over a large file can be time-consuming. Another observation is that although sound characteristics vary a lot between different files, they remain relatively stable within a particular file (for example a news broadcast). We explore this observation by comparing broadcast-specific bootstrapping with bootstrapping using all the training data. A series of experiments using the TREC03 video retrieval collection were conducted to test the validity of the proposed approach.

Gender identification based on a general audio classifier is described in [1]. An earlier work on the same topic is [2]. The features used in this work are based on the feature set described in [3] for the task of automatic musical genre classification. A classic reference for the related problem of speech/voice discrimination is [4]. Locating voice segments within music signals have been explored in [5] where the output of a neural network trained on linguistic categories is used as a feature vector for classification. Bootstrapping was proposed for signing voice structure detection in [6].

2. METHODOLOGY

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. Statistics of these features (means and variances) are calculated over a “texture” window, approximately 2 seconds long. This window captures the statistical longer-term characteristics of complex audio textures such as singing or music [3, 1]. After training, a binary classification decision is made every 20 milliseconds based on 2 seconds of “texture” data.

After experiments with various features proposed in the literature the following features were chosen: 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 (for details [3]). In addition, a pitch detection algorithm based on the average magnitude difference function (AMDF) is used. This pitch detector was shown to be more robust to background music and noise.

The basic idea in bootstrapping is to use a small random collection of snippets to train a machine learning classifier that is subsequently used to classify/segment the entire training data. Important bootstrapping parameters are: the duration of each snippet, the number of snippets used, and the classifier. The classifier used for bootstrapping must have good generalization properties (i.e its performance using a small amount of representative data should be representative of its performance on a large set of unknown data). The following classifiers were tried out for this purpose: Naive Bayes, Nearest Neighbor, Backpropagation Artificial Neural Network, Decision Tree, Support Vector Machine and Logistic Regression.

3. RESULTS

A collection of 10 news broadcast excerpts (each 5 minutes long) from the TREC 2003 video retrieval data set was used for the experiments. Each excerpt contains typically 3-4 female and male speakers and was fully annotated by hand to evaluate the bootstrapping performance. Table 1 compares the performance of different classifiers using 16 seconds (8 snippets) and 32 seconds (16 snippets) of bootstrapping. The mean and standard deviation of the percentage of correctly classified 20-millisecond frames is shown. The choice of two seconds as the snippet size was guided by experiments in speaker identifiaction and musical genre classification that show that the short term memory required for these tasks is approximately 2-3 seconds.

As can be seen, the best generalization performance is obtained using the Neural Network. The first and third column (A16 and A32) are based on using snippets for all the broadcasts in the training set. The second and fourth column (16, 32) show results that only use the corresponding news broadcast for bootstrapping (file-specific). As can be seen there is some slight improvement (due to better fitting of the particular speaker characteristics) but it is not significant. The last column (ALL) shows the results obtained using all the annotated information. All classification accuracy results are calculated by running the trained classifier over the entire data collection and are at the frame level. Using the bootstrapping approach a 5-minute file that would require at least 5 minutes to be manually annotated can be annotated in 16 seconds. Using additional smoothing these results are good for most practical purposes.

	A16	16	A32	32	ALL
Bayes	78 ± 6	80 ± 9	80 ± 7	82 ± 7	84 ± 5
J48	81 ± 7	76 ± 6	83 ± 6	82 ± 7	83 ± 3
NN	84 ± 7	86 ± 6	85 ± 5	89 ± 5	86 ± 4
Logistic	81 ± 8	84 ± 7	84 ± 8	88 ± 6	84 ± 5
Neural	85 ± 7	86 ± 4	88 ± 7	90 ± 4	90 ± 3
SVM	82 ± 9	86 ± 4	85 ± 9	89 ± 5	87 ± 6

Table 1. Classification accuracy (mean,std) of different classifiers (10 news excerpts -5 minutes long) (the columns 16, 32 use file-specific bootstrapping and the columns A16,A32 use the annotated snippets from all the broadcasts)

The audio analysis is implemented using Marsyas (`http://marsyas.sourceforge.net`). The classification experiments were performed using Weka:
(`http://www.cs.waikato.ac.nz/ml/weka/`). The calculated features and ground truth data used in this study are avialable upon request via email.

We showed that bootstrapping is an effective technique for semi-automatic gender annoation. The best classifier generalization for bootstrapping was obtained with a Neural Network classifier.

4. REFERENCES

[1] H. Harb and L. Chen, "Gender identification using a general audio classifier," in *Proc. Int. Conf. on Multimedia and Expo (ICME)*, Baltimore, July 2003, IEEE.

[2] Eluned Paris and J. Carey, Michael, "Language independent gender identifacation," in *Proc. Int. Conf. on Audio Speech and Signal Processing (ICASSP)*. IEEE, July 1996, pp. 685–688.

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

[4] 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.

[5] 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.

[6] George Tzanetakis, "Song specific bootstrapping of singing voice structure," in *(to appear) Proc. Int. Conf. on Multimedia and Exposition (ICME)*, TaiPei, Taiwan, July 2004, IEEE.