BUILDING AUDIO CLASSIFIERS FOR BROADCAST NEWS RETRIEVAL

George Tzanetakis, Ming-Yu Chen
School of Computer Science
Carnegie Mellon University
Pittsburgh PA 15218
gtzan@cs.cmu.edu, mychen@cs.cmu.edu

ABSTRACT

The process of building audio classifiers for high-level content descriptors, especially in large datasets, is not trivial. In this paper we describe the design and development of audio classification algorithms for broadcast news retrieval in the context of the TREC 2003 video retrieval evaluation. The main focus of this paper is the actual building process itself rather than the final results, although some representative results will be provided. It is our belief that the insights obtained and tools developed in order to work with real world large audio collections are important and frequently unmentioned in existing published work. An important and critical aspect of this process is obtaining ground truth annotations for training the classifiers. Therefore tools and techniques that assist the human annotation of news audio will be described.

1. INTRODUCTION

Video is a rich source of information, with visual, audio, and textual content. The TREC Video Retrieval task provides a large-scale, standardized evaluation of video retrieval systems. In video retrieval, the most common use of audio information is for automatic speech recognition and the subsequent use of the generated transcript for text retrieval. However, the audio information can also be used, more directly, to provide additional information such as the gender of the speaker, music and speech separation, and audio textures such as fast speaking sports announcers. This paper describes the process of building such audio classifiers that model the audio directly and don't perform speech recognition. These classifiers were used as part of the much larger Informedia [1] system entry to the TREC 2003 Video Retrieval evaluation. This system integrates under a common interface diverse sources of information such as video, images, OCR, speech recognition and face identification.

The focus of this paper is to describe the process of building these classifiers, the tools developed, and the lessons learned rather than providing a detailed description of the final results. Some representative results will be presented to support the proposed techniques and more details about the full system can be found in the Informedia TREC 2003 video retrieval report [2]. The TREC 2003 video retrieval task requires analyzing for retrieval, 130 hours of broadcast news that correspond to approx. 20 gigabytes of audio data (mono, 22050 Hz sampling rate). We discovered that building audio classifiers for such a large real world collection is significantly harder than working with a well labeled small data set as is the case with the majority of existing literature. It is our belief that the tools we developed and the insights we obtained are of interest to other researchers working on similar large scale problems. An important and critical aspect of building audio classifiers is obtaining human ground truth annotations for training. Therefore the tools we have developed to assist this annotation process will be described.

2. RELATED WORK

There has been a lot of work in various types of non-speech audio classification. The references in this section are representative of existing approaches and are by no means exhaustive. Probably the earliest related work is the music speech discrimination algorithm described in [3]. A comparison of different features and classifiers for the same task is provided in [4]. An hierarchical audio classification system based on individual feature heuristics is described in [5]. A review of music and audio retrieval in general is provided in [6]. Techniques for the automatic segmentation of MP3 and AAC compressed audio streams into *speech, music* and *silence* are presented in [7]. Examples of systems for video retrieval where audio information is combined with visual information include: a Hidden Markov Model (HMM) framework for video segmentation of conference meetings based on audio and image features [8], the use of a simple relative loudness feature in combination with visual features for the semantic indexing of sports sequences [9], and combining audio and visual information for video content analysis into broad categories such as *sports* or *news* [10]. An important influence to the design of our computer assisted annotation tools is SpeechSkimmer a system for interactive browsing of recorded speech [11].

3. FEATURE SELECTION

The foundation of any type of audio analysis algorithm is the extraction of numerical feature vectors that characterize the audio content. Although audio feature extraction has been extensively explored in the context of speech recognition, there are some unique aspects of general audio feature extraction. In the TREC 2003 evaluation all classification results are reported for each video shot. These variable duration shots (duration range 2-60 seconds) are calculated using image processing information and are provided as input to the audio classification subsystem. In order to classify a shot, three levels of information are used. 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 speech or music that possibly contain a variety of different spectra [12]. 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 decision for the whole shot is obtained by the majority of classified windows within the shot and the percentage of the majority windows is used as a confidence measure for classification. This approach has the advantage of dealing gracefully with the problem of shots that contain two different audio textures, which although not common, occurs sometimes in the data. In the ideal case, rather than somehow mixing the statistics of the two audio textures, this majority voting scheme will correctly classify each texture separately and calculate the confidence based on the relative durations of the two textures.

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]. More details about the feature selection process will be provided in Section 5. 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 $1/4^{th}$ of the total bandwidth), Mean Relative Subband Energy 2 (relative energy of the second $1/4^{th}$ of the total bandwidth), Standard Deviation of the Centroid, Rolloff, and Flux. More details about the definitions of these features can be found in [4, 12]. In addition to these features, the mean and standard deviation of pitch was calculated for the male/female voice discrimination. The pitch calculation is performed using the Average Magnitude Difference Function (AMDF) method [15] which proved to be more robust to background noise and music than other methods. One surprising finding was that the MFCC and LPC features did not perform as well as the ones described above. Probably the reason is that these features are designed for speech modeling and recognition and don't work as well for modeling more general audio textures.

4. COMPUTER ASSISTED ANNOTATION

In order to train classifiers it is necessary to have data labeled with ground truth by a human. The quantity and quality of this ground truth data is critical to building robust classifiers that have good generalization properties. The process of annotating 120 hours of audio can be extremely time-consuming without automatic tools to assist and support it. Moreover, most existing software solutions for video annotation are based on shot key-frames and can not be used for audio annotation.

A special purpose audio editor was developed to assist users with the annotation process. The main idea is to provide a flexible semi-automatic environment that combines the abilities of the human user to make high level decisions with the computer's ability to work with large amounts of data. The editor displays the audio signal both in amplitude envelope waveform display as well as a spectrogram. The shot boundaries are imported and colors are used to represent each class label such as *Male Speech* or *Commercial*. In addition to being able to playback the whole shot, there is the option to hear a set of random 1 second snippets reducing playback time considerably while still ensuring correct annotations.

While the user is annotating the audio signal, a "fast to train" classifier is trained on the fly and used to predict the remaining shots that are not annotated. That way the user only has to confirm the annotation rather than having to make a decision. We use the term bootstrapping to describe this approach to annotation. In addition, shots that contain more than one texture can either be split manually before training or ignored so that the training data consists only of correct examples of each audio texture. This is important as our experience has shown incorrectly labeled training samples can significantly reduce classification performance.

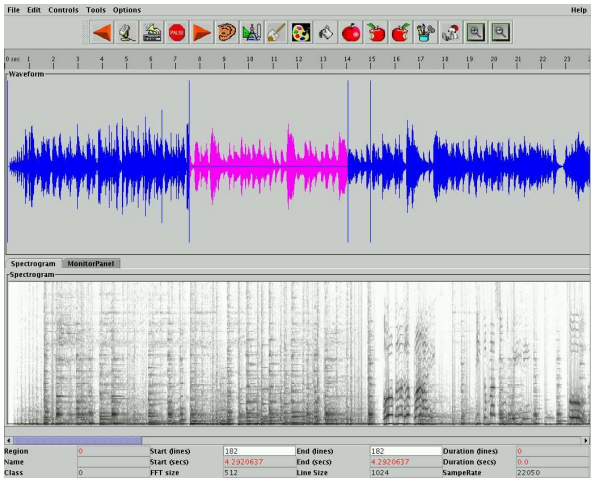

Figure 1. GUI for computer-assisted annotation

It has been shown [11] that time stretching techniques can be used to significantly reduce browsing time for speech signals without affecting intelligibility. Although simple pause removal works for clean speech unfortunately it doesn't for broadcast news as there is frequently background music or sounds. In order to reduce playback duration we have used a phasevocoder algorithm [16] that enables time shrinking without pitch shifting. Although the resulting signal is slightly distorted it is still intelligible and annotation to general audio textures can be performed without any problem. Figure 1. is a screenshot of the graphical user interface (GUI) utilized for computer-assisted annotation. Male voice is represented by blue and female voice by pink.

5. CLASSIFICATION

Once enough annotated shots have been collected, standard statistical pattern classifiers are trained and used to predict the class label of previously unseen, or more appropriately unheard, shots. One of the main problems with training using large amounts of data is that training time can become extremely long (days on current hardware). However, in many cases fast training time is required for computer-assisted annotation, feature selection and experimentation is general. In order to address this problem we used a simple fast-to-train classifier for experimentation and for the final results we employed a more powerful but much slower to train classifier. For the "fast" classifier, a single Gaussian with a full covariance matrix was used to model each class.

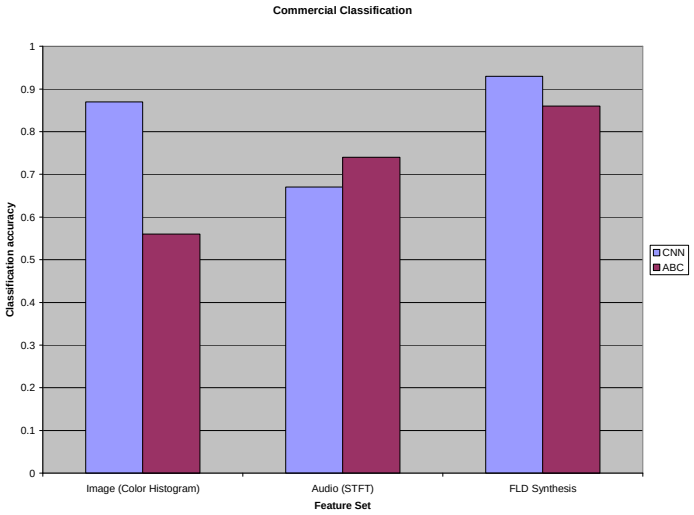

Figure 2. Classification accuracy for classification of commercials

Commercial	Image (Color Histogram)	Audio (STFT)	FLD Synthesis
CNN	0.87	0.67	0.93
ABC	0.56	0.74	0.86

Table 1. Classification accuracy for classification of commercials

The best results for the final "slow" classification, were obtained using Support Vector Machines (SVM) with Radial Basis Functions (RBF). More details about these classification methods can be found in [17].

The following binary high-level classifiers were trained based purely on audio features: male voice, female voice, noise, music, and silence. In addition audio features were used in combination with other features for the following classifiers: commercials, anchors, weather and sports. Although in some cases such as sports the results were not particularly promising, in most cases good classification performance was obtained. Figure 2 and Table 1 show some representative results for commercial classification of the CNN and ABC news broadcast using image information, audio information and combining the results. FLD synthesis refers to a feature synthesis technique based on Fisher Linear Discriminant Analysis.

Some additional indicative results are 78% for female voice classification and 74% for male voice. More detailed results can be found in [1]. It is important to mention that the quality of the training data was a more important factor than the exact feature set or classifier used. The annotated samples used for training must be representative and must have the necessary variability without on the other hand introducing confusing outliers.

The audio feature extraction, as well as the GUI for user annotation, was implemented using Marsyas (http://marsyas.sourceforge.net/) [18], a free software framework for Computer Audition research. Support Vector Machine classification was performed using libsvm (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) [19].

7. CONCLUSIONS AND FUTURE WORK

Building general audio classifiers for large "real-world" datasets is challenging. Various factors such as human annotation, training speed and quality control, that are typically not important in smaller datasets, become crucial for designing and developing effective audio classification algorithms for large datasets. It is our hope that the techniques and tools described in this paper will provide ideas and inspiration to researchers working with automatic content analysis for large audio datasets.

In the future we plan to explore source enhancement and separation techniques so that audio textures are treated more specifically based on the audio sources they contain. In addition, we plan to develop additional tools for computer-assisted annotation such as unsupervised clustering of shots based on audio information, speaker identification, and audio similarity calculation between shots. We would also like to be able to classify specific sounds such as explosions, gunshots, helicopters etc.

8. REFERENCES

[1] A. Hauptman, et al., "Informedia at TREC 2003: Analyzing and Searching Broadcast News Video" in *Proceedings of (VIDEO) TREC 2003*, Gaithersburgh, MD, November 2003

[2] A. Hauptmann and M. Witbrock, "Informedia: News-on-demand Multimedia Information Acquisition and Reetrieval", in *Intelligent Multimedia Information Retrieval*, chapter 10, pp. 215-240, MIT Press, Cambridge, Mass. 1997

[3] J. Saunders, "Real time Discrimination of Broadcast Speech/Music", in *Proceedings of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 1996, pp. 993-996

[4] E. Scheirer and M. Slaney, "Construction and Evaluation of a robust multifeature speech/music discriminator" in *Proc.of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 1997, pp. 1331-1334.

[5] T. Zhang and J. Kuo, "Audio Content Analysis for online Audio-Visual Data Segmentation and Classification", *Transcactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 441-457, May 2001.

[6] J. Foote, "Content-based Retrieval of Music and Audio," in *Multimedia Storage and Archiving Systems II*, 1997, pp. 138-147.

[7] S. Kiranyaz, M. Aubazac, M. Gabbouj, "Unsupervised Segmentation and Classification over MP3 and AAC Audio Bitstreams", in *Proc. 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2003, pp. 338-344.

[8] J. Boreczky and L. Wilcox, "A Hidden Markov Model framework for Video Segmentation using Audio and Image Features", in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 6, pp. 3741-3744, 1998.

[9] R. Leonardi, P. Migliorati, M. Prandini, "A Markov Chain Model for semantic indexing of sport program sequences", in *Proc. 4th European Workshop on Image Analysis and Multimedia Interactive Services (WIAMIS)*, 2003, pp. 20-26.

[10] Y. Wang, Z.Liu, J.C. Huang, "Multimedia Content Analysis Using Audio and Visual Information", *IEEE Signal Processing Magazine*, vol 17, no. 6, pp. 12-36, Nov. 2000.

[11] B. Arons "SpeechSkimmer: A System for Interactively Skimming Recorded Speech", *ACM Transactions Computer Human Interaction*, vol. 4, pp. 3-38, 1997.

[12] G. Tzanetakis, P. Cook, "Musical Genre Classification of Audio Signals", *IEEE Trans. On Speech and Audio Processing*, vol. 10, no. 5., July 2002.

[13] S. Davis and P. Mermelstein, "Experiments in syllable-based recognition of continuous speech," *IEEE Trans. On Acoustics, Speech and Signal Processing*, vol. 28, pp. 357-366, Aug. 1980.

[14] J. Makhoul, "Linear Prediction: A tutorial overview", *Proc. IEE*, 63: 561-580

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

[16] M. Dolson, "The Phase Vocoder: A Tutorial", *Computer Music Journal*, 10(4), 14-27, 1986.

[17] R. Duda, P. Hart, and D. Stork, "Pattern Classification", John Wiley & Sons, New York, 2000

[18] G. Tzanetakis and P. Cook, "Marsyas: A Framework for Audio Analysis", Organized Sound, vol 4(3), 2000.