MADClassifier: Content-Based Continuous Classification of Mixed Audio Data

Shariq J. Rizvi[*], Lei Chen and M. Tamer Özsu [†]

Technical Report CS-2002-34 October 2002

[*]This work was done in part while the author was visiting University of Waterloo in summer 2002

[†]School Of Computer Science, University of Waterloo, Waterloo, ON

Abstract

Content-based classification of audio data is an important problem for the overall analysis of audio-visual streams. Tho- ugh the classification of audio into pure classes, such as music, speech, environmental sound and silence, is well studied, classification of mixed audio data, such as clips having speech with music, is still considered a difficult problem. We present *MADClassifier* (Mixed Audio Data Classifier), a system for the classification of audio onto a continuous scale. We introduce the notion of *continuity of audio features*, which makes the feature vary with the type of audio in a manner that is representative of its composition. We use these features to build two versions of our classifier, one is based on the simple k-nearest neighbor classifier and the other is a neural network classifier. Unlike the previous research that has gone into mixed audio classification, we do not generate classification-time thresholds empirically. This ensures that the classifier can be specifically trained for a focussed domain where it is intended to be used. The performance of the proposed system is validated against representative real and synthetic data.

1 Introduction

In general, simple classification with just the "speech and music" category may not be satisfactory to determine whether an audio clip is speech with background music or predominantly music with some speech. For example, in many speech processing applications, only clips having a relatively high speech content should percolate through the data cleaning step while music processing applications will consider too much of speech in the input as noise. Therefore, in this paper, we address the problem of mixed audio type classification. Unlike all other previous approaches [10, 12], we classify an input clip having music and speech content on a continuous scale, based on the relative significance of the speech and music components. Our classifier can label an input clip as, say "10% speech and 90% music". This kind of information may provide additional help in the overall analysis of an audio-visual stream, of say a movie. A typical application can be the validation of the quality of a musical concert recording based on the amount of speech distortion caused by voices from the audience.

Another issue that we address is that of reliability of the classifier. The previous approaches ([10, 12]) to classify mixed audio type have reported

empirically derived values for the threshold parameters used in their rule-based models. Our model presents a classifier which can be trained with training data that is, possibly, restricted to a focussed domain. For example, if a movie's audio stream predominantly contains music only from the violin, then the classifier can be better trained with training data that focuses on violin pieces.

The rest of the paper is organized as follows: Section 2 presents some related works on audio classifications. Section 3 explains the audio features that we use for continuous classification of mixed audio. The idea behind the classifier is discussed in Section 4. Section 5 reports the results of experiments done on extensive data sets. We conclude and point out some future work directions in Section 6.

2 Related Work

Saunders [8] addressed the issue of pure type audio classification for FM radio. The idea was to allow automatic switching of channels when music is interrupted by advertisements. Zero crossing rate (ZCR) and short time energy were used to classify input audio into two classes: speech and music. Scheirer and Slaney [9] used thirteen features in time, frequency, spectrum and cepstrum domains and achieved better classification, they got the conclusion that not all the audio features are necessary to perform an accurate classification. Besides that, they claimed that they improved the error rate to 1.4 % for a 2.4s window compared to 2.8% of Saunders' approach. Based on Scheirer's conclusion, Carey et al. [1] made a comparison study on audio features for speech and music discrimination. They figured out that simple audio features, such as pitch and amplitude, have significant differences between music and speech. Since then, many approaches have been proposed to classify pure type audio using different audio features and classifiers [1, 2, 5, 11, 4, 6].

Mixed type audio consists of more than one pure types, like speech with music background, and music with noise. The simple audio features, such as ZCR, which are used to classify pure types, have been proved insufficient in classifying mixed type audio [10]. Srinivasan et al [10] proposed a method to classify mixed audio into discrete classes. They used a rule based model with empirically determined thresholds. Zhang and Kuo [12] proposed a method for classification into speech, music, song, environmental sound, speech with

background music, silence, etc. Again, empirically determined values for classification thresholds have been proposed.

3 Continuous Nature of some Audio Features

A number of audio features provide separability between the different classes involved in pure audio classification, for eg., the variance of ZCR is relatively higher for speech than for music. This is because of the considerable difference between the ZCR values of voiced and unvoiced speech. Intuitively, we expect that the ZCR variance for an audio clip having both the components in a particular composition, will be between the variance values for the pure components. This, as our experiments have shown, is true to a large extent. Informally, We say that an audio feature is *continuous* if, for mixed audio, it takes values intermediate to the ones it takes for its pure audio components. In the following subsections, we discuss the continuous nature of the audio features that we will later use to build our classifier.

3.1 Variance of Zero Crossing Rate

Due to the sharp difference between the ZCR values for voiced and unvoiced components, speech tends to have a high variance in its ZCR values. Because of the absence of any such phenomenon, music tends to have a lower variance. We calculate a measure of the ZCR variance for a window of N frames, as the ratio of frames having ZCR value more than 1.5 times the average ZCR of the window (*high zero-crossing rate ratio* of [6]):

$$HZCRR = \frac{1}{2N}\sum_{n=0}^{N-1}[sgn(ZCR(n) - 1.5avZCR) + 1]$$

$$avZCR = \frac{1}{N}\sum_{n=0}^{N-1} ZCR(n)$$

Where *sgn* is the sign function and *ZCR(n)* is the ZCR value at the n^{th} frame. In our experiments, we divide the 1-second window into 100 frames.

Figure 1 shows the variation of average HZCRR with the composition of audio when a particular speech and music clip pair is combined in different ratios. Roughly, the y-axis readings increase continuously as the composition

goes from 0% speech to 100% speech. This trait of HZCRR makes it a good choice for use in a composition predictor, like ours.

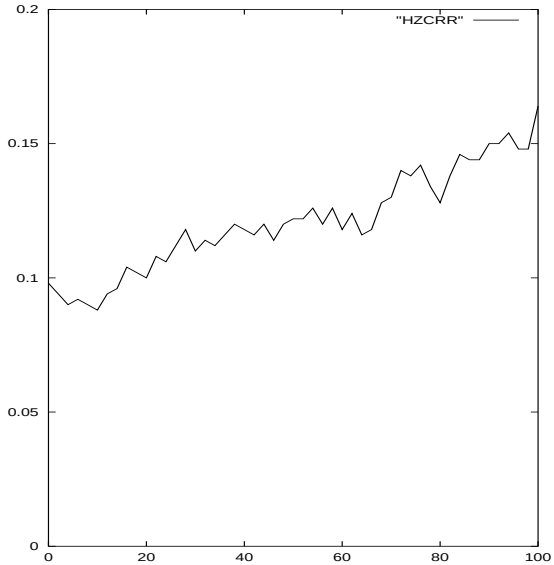

Figure 1: Variation of average HZCRR with the percentage of speech in audio

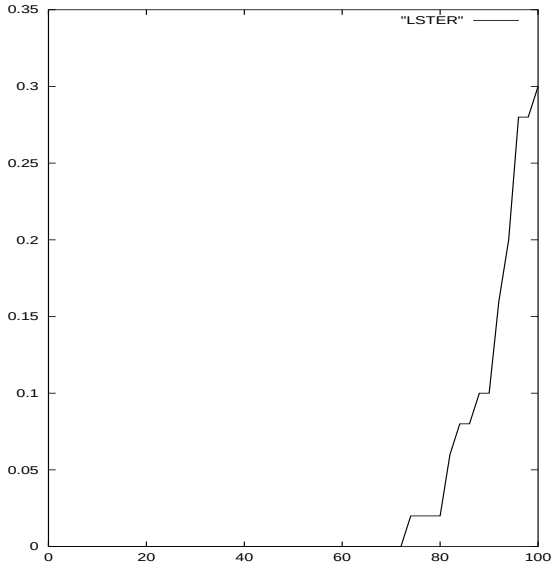

Figure 2: Variation of average LSTER with the percentage of speech in audio

3.2 Low Short Time Energy Ratio

This is a measure of the number of frames in a window that have their short-time energies lower than a particular fraction of the average energy of the whole window. Speech tends to have higher values for this feature as it has more silence frames than music.

We calculate the feature as done in [6]:

$$LSTER = \frac{1}{2N} \sum_{n=0}^{N-1} [sgn(0.5avSTE - STE(n)) + 1]$$

$$avSTE = \frac{1}{N} \sum_{n=0}^{N-1} STE(n)$$

Figure 2 shows the variation of average LSTER with the composition of audio when a particular speech and music clip pair is combined in different ratios. Again, LSTER is a non-decreasing function of the speech fraction in the mixed clip.

3.3 Fundamental Frequency

As stated in [7], spectrum analysis shows that pure music is more harmonic than speech, since pure speech contains a sequence of tonal (vowels) and noise (consonants). Harmonic sound is defined as one which contains a series of frequencies which are derived from the fundamental or original frequency as a multiple of that. We compute fundamental frequencies of the audio clips using the algorithm in [7].

For each audio clip, we take the ratio of 1-second windows having a dominant frequency to the total number of windows as a measure of the harmony in the clip. Figure 3 shows the variation of this quantity with the composition of audio when a particular speech and music clip pair is combined in different ratios. We observe that as the music component decreases and the audio tends to become predominantly speech, our measure of harmony deceases in a representative fashion. It should be noted that this measure of harmony can take only discrete values depending on the choice of window size and clip length.

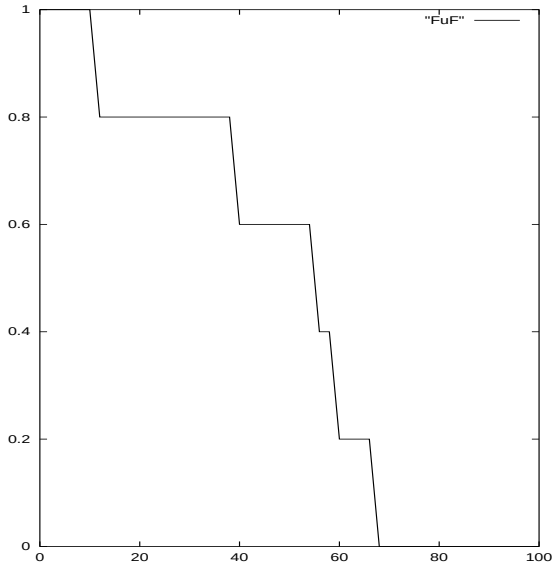

Figure 3: Variation of our FuF measure with the percentage of speech in audio

4 The Classifier

We experimented with two versions of the continuous classifier. The first one is based on the simple k-nearest neighbor concept while the second one is based on neural networks.

In a k-nearest neighbor classifier [3], each training sample is a point in an n-dimensional space (based on its n numeric attributes). When given an unknown sample, such a classifier finds the appropriate class by searching the pattern space for the k training samples that are closest to the unknown sample. We apply such a classifier for mixed type audio classification. The proposed classifier return the average value of the real-valued labels associated with the k nearest points of the unknown sample. In our case, the space is 3-dimensional, with a dimension each to the three audio features described in Section 3.

Nearest neighbor classifiers are also called *lazy learners* [3] since they store all of the training samples and do not build a classifier until a new (unlabeled) sample needs to be classified. This increases the classification-time computational costs. We also implemented an *eager* version MADClassifier based on a neural network [13] which approximates a linear function. *Eager learners* construct a generalization model before receiving new samples

to classify, which becomes indispensable for real-time applications like automatically switching channels by audio-content analysis.

5 Experimental Evaluation

In this section, we evaluate the efficiency of our proposed classifier through extensive experiments with data synthesized from real world audio. We selected pure music and pure speech audio clips from different movies [1] and some music resources on the web, which were transformed into 16 bit, 44kHz, single channel raw audio clips of 5 second duration. With these pure clips, we generated a database with a size of around 1200 clips, as follows: We used a random number generator operating on uniform distribution to generate the ratios in which the pure clips were combined (a $(90, 10)$ mixed clip refers to a clip synthesized from a pure speech and music clip in such a way that is has them in a relative ratio of 9:1 by energy (volume)).

We partitioned the above data set to get training and test data, with sizes in an approximate ratio of 70:30. Its important to make sure that the sets of pure speech (and pure music) clips used in the training and test data are disjoint. This ensures that the test data points are not illegitimately cognate with the training data.

5.1 MADClassifier-Lazy

We run our *lazy* version MADClassifier with the parameter k (representing the number of nearest neighbors to be used for prediction) set to 40. Table 1 reports the average composition errors ($\overline{\delta}$) resulting from the use of classifier with different audio features in action (if a clip having a $(30, 70)$ composition is predicted to be $(35, 65)$, then we say that the error is 5%). As we expected, the error increases when one or more features are "turned off".

Figure 4 shows the variation of average error when different values of k are used along with all the three audio features activated. From the figure, we can see that the value of k does not affect the performance to a large extent. Based on the experiment data, we conclude that *lazy* version MADClassifer can place audio clips on the percentage composition scale with an average error of around 11%.

[1] 1. "Crouching Tiger Hidden Dragon", 2000; 2. "Gladiator", 2000; 3. "Patch Adams" 1998.

HZCRR	LSTER	FuF	$\overline{\delta}$
√	√	√	11.00%
	√	√	12.08%
√		√	13.16%
√	√		12.75%
√			16.49%
	√		16.74%
		√	14.37%

Table 1: Average composition errors resulting with different combinations of audio features

5.2 MADClassifier-Eager

We used a neural network implementation provided by the YALE project [13] to build our *eager* version MADClassifier. With the *eager* version MADClassifier, we were able to obtain a similar average error of around 11-12% as what we got by using *lazy* version MADClassifier.

5.3 Performance with Real World Audio

Besides using synthesized data, we also tested our classifier using the real mixed type audio clips which extracted from movies. With those clips, the predictions of the classifier are in tune with human intuition. The set of real mixed type audio clips along with their predicted compositions are available at: *http://db.uwaterloo.ca/∼l6chen/MADClassifier/*

6 Conclusions and Future Work

A lot of research has gone into audio classification. However, none of the approaches classifies audio data onto a continuous scale, which is quite important for audio-visual analysis and audio editing. In this paper, we presented MADClassifier, a system for discriminating audio data on the basis of its speech-music composition. Using three audio features - ZCR Variance, Low Short Time Energy Ratio and Fundamental Frequency, we were able to achieve a considerably low error mark of 11% in composition on data synthesized from real-world audio. With real audio data from movies, the

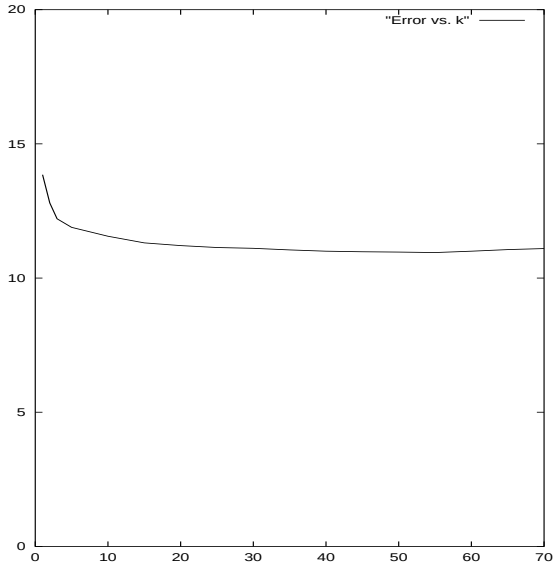

Figure 4: Variation of average composition error with k

composition predicted by our classifier was quite compatible with human intuition. The proposed classifier can be easily modified to work as a discrete type classifier by giving the range of composition scale that each class locates. In future, we plan to extend the number of features used by MADClassifier and experiment with other classification models, such as multi-layered neural networks, especially more *eager learners* for real-time applications. We also plan to extend our data set to include more audio phenomena like noise and study its effect on the variation of audio features.

Acknowledgments

This research is funded by Intelligent Robotics and Information Systems (IRIS), a Network of Center of Excellence of the Government of Canada.

References

[1] M.J. Carey, E. S. Parris, and H. Lloyd-Thomas. A comparison of features for speech, music discrimination. In *Proceedings of IEEE International*

Conference on Acoustics, Speech, and Signal Processing, pages 149–152, April 1999.

[2] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal. Speech/music discrimination for multimedia applications. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2445–2448, June 2000.

[3] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.

[4] Stan Z. Li. Content-based audio classification and retrieval using the nearest feature line method. *IEEE Transaction on Speech and Audio Processing*, 8(5), pages 619-625, 2000.

[5] Guojun Lu and Hankinson Templar. An investigation of automatic audio classification and segmentation. In *Proceedings of 5th International Conference on Signal Processing (WCCC-ICSP)*, pages 776–781, 2000.

[6] Lie Lu, Hao Jiang, and HongJiang Zhang. A robust audio classification and segmentation method. In *Proceedings of ACM International Conference on Multimeida*, pages 203-211, 2001.

[7] Silvia Pfeiffer, Stephan Fischer, and Wolfgang Effelsberg. Automatic audio content analysis. In *Proceedings of ACM International Conference on Multimeida*, pages 21-30, 1996.

[8] J. Saunders. Real-time discrimination of broadcast speech/music. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 993–996, May 1996.

[9] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 21–24, April 1997.

[10] S. Srinivasan, D. Petkovic, and D. Ponceleon. Towards robust features for classifying audio in the cuevideo system. In *Proceedings of the seventh ACM international conference on Multimedia*, pages 393-400, 1999.

[11] E. Wold, T. Blum, D. Keislar, and J. Wheaten. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, 3(3):27–36, 1996.

[12] Tong Zhang and C. C. Jay Kuo. Audio content analysis for online audiovisual data. *IEEE Transaction on Speech and Audio Processing*, 9(4):619–625, 5 2001.

[13] Fischer, Simon and Klinkenburg, Ralf and Mierswa, Ingo and Ritthoff, Oliver. Yale: Yet Another Learning Environment - Tutorial. *CI-136/02, Collaborative Research Center 531, University of Dortmund*, 2002, ISSN 1433-3325.