HARMONICITY AND DYNAMICS–BASED FEATURES FOR AUDIO

S H Srinivasan

Applied Research Group
Satyam Computer Services Ltd, Bangalore
SH_Srinivasan@satyam.com

Mohan Kankanhalli

Department of Computer Science
National University of Singapore
mohan@comp.nus.edu.sg

ABSTRACT

Features are very important for audio processing. Tasks like speech recognition and instrument identification are based on features. Most low-level features currently used are based on LPC and cepstral analysis. In this paper we propose a class of features based on dynamics and harmonicity. In particular, we define the notion of *harmonic derivative*. The efficacy of the features is demonstrated for music genre classification and instrument family classification. In particular, the features are shown to be cepstrum-equivalent.

1. INTRODUCTION

There are several interesting audio processing tasks considered in research literature: speech recognition, speaker identification, instrument identification, musical genre identification, etc. The processing is usually performed in two stages: feature computation and classification. The widely-used features for audio are based on LPC and cepstral analysis. Since the analysis is performed on a frame basis, these features capture intra-frame information. Dynamics is captured in the form of "delta coefficients": These are given by difference between feature values of successive frames.

Audio signals have a harmonic structure which is not exploited in the above features. In this paper we propose a class of features which are based on the harmonic structure of audio.

This paper is organized as follows. Section 2 lists some principles of auditory scene analysis from which our work takes inspiration. Section 3 describes the proposed features. These features are tested on music genre identification. Section 4 describes the classification tasks and the datasets used. Section 5 lists the results. The paper concludes with a discussion of results.

2. AUDITORY SCENE ANALYSIS

Auditory scene analysis studies the mechanisms the ear uses to analyze audio signals. See [1] for a recent review. Just as the eye segments visual scenes into objects, ear segments the audio signal into *auditory objects*. Hence the term "auditory scene analysis". There are several principle of auditory scene analysis: harmonicity, dynamicity, continuity, and simultaneity. Here we use only harmonicity and dynamicity. The following description of these principles is taken from [2]. (In [2], these go under the names Regularity 3 and Regularity 4.)

Harmonicity: When a body vibrates with a repetitive period, its vibration give rise to an acoustic pattern in which the frequency components are multiples of a common fundamental.

Dynamics: Many changes that take place in an acoustic event will affect all the components of the resulting sound in the same way and at the same time.

These principles have been shown to be extremely useful in audio separation [3]. It is natural to ask if we can define features based on these principles. There are several ways to define audio features based on these principles. The most commonly used features are

Ratio of AM and HM: The ratio of arithmetic and harmonic means measures the flatness of the spectrum. This is taken to be an approximate measure of the harmonicity of the spectrum. Because of the polyphonic nature of music signals, this is not a reliable index.

Delta features: The difference between the features (energy, cepstrum, etc.) of successive frames is taken as a measure of signal dynamics. Such delta features have yielded good performance benefits in speech recognition.

The next section describes some alternative features to describe the harmonic and dynamic structure of audio.

3. FEATURES FOR HARMONICITY AND DYNAMICS

In this section, we describe features based on harmonicity, dynamics, and combined harmonicity and dynamics. All the features are based on short-time Fourier transform spectrum. This denoted by $E_t(f)$ where t is the time/frame index and f the frequency index. When discussing features for a single frame, the time index t will be dropped.

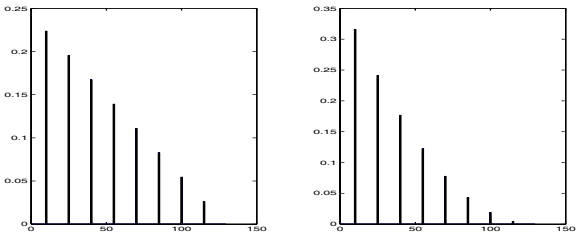

Fig. 1. Two harmonic signals with the same total energy.The energy distributions are different. We need a measure to characterize the difference.

3.1. Harmonicity

Most natural sounds are harmonic. If frequency f has high energy, it is likely that frequencies $2f, 3f, \cdots$ also contain high energy. A large part of audio signals contain harmonic sounds. [1]

If the signal is harmonic, energy is concentrated only in these frequencies. Hence the fraction of energy concentrated in these frequencies is a measure of harmonicity. Hence the *harmonic concentration* is defined as

$$hc = \frac{\sum_k E(kf)}{E}$$

where E is the total energy of the frame and $E(f)$ is the energy at frequency f. We take f to be the frequency at which the energy is high. (Other choices for f are possible – especially the pitch.)

While harmonic concentration measures the energy in the dominant harmonic component, it does not measure the harmonic structure of that component. Figure 1 shows two signals with different energies in the harmonic bands. We use *harmonic energy entropy* – denoted he – to characterize this structure. This feature is defined as the entropy of the energy distribution in the harmonic frequencies.

Peaks in the energy spectrum are important. The above two features take the dominant peak and impose a harmonic structure based on it. It is possible to quantify the harmonic structure of actual spectral peaks. Let $f_1, f_2, \cdots, f_n$ be the frequencies of spectral peaks of a given frame. These frequencies are ordered so that $f_1 < f_2 < \cdots < f_n$. If the peaks obey a harmonic relation, then $f_k = f_1 + (k-1)\delta$ for some δ. It can be seen that $\Delta f_k \stackrel{\text{def}}{=} f_k - f_{k-1} = \delta$. The entropy of Δf_is provide a measure of the harmonic structure of the actual spectral peaks. We call this the *spectral peak structure*. Hence the entropy of the histogram of the differences

$$f_2 - f_1, f_3 - f_2, \cdots, f_n - f_{n-1}$$

gives the harmonic structure of spectral peaks.

[1] Inharmonic and nonharmonic sounds make up the rest.

3.2. Dynamics

We now move over to dynamic structure of music. Traditionally temporal derivative has been used. In addition to first derivative, the second temporal derivative can be used. The second temporal derivative is defined as

$$\sum_f |E_{t+1}(f) - 2E_t(f) + E_{t-1}(f)|$$

where $E_t(f)$ is the energy at frequency f at time t.

3.3. Combined Harmonics–Dynamics

Audio has a time-frequency structure. While temporal derivatives have been defined, frequency derivatives are seldom used. The definition of spectral derivative as per the usual definition of derivative is[2]

$$\frac{\partial E}{\partial f} \stackrel{\text{def}}{=} E(f+1) - E(f)$$

is not meaningful since frequencies f and $f+1$ are not really neighbors in an musical sense. Since music is harmonic, frequencies f and $2f$ are neighbors. Hence we define the *harmonic derivate* as

$$\frac{\partial E}{\partial_h f} \stackrel{\text{def}}{=} \frac{E(2f) - E(f)}{f}$$

Since auditory perception follows the logarithmic scale, the definition can be changed to

$$\frac{\partial E}{\partial_h f} \stackrel{\text{def}}{=} \frac{\log E(2f) - \log E(f)}{\log f}$$

We can extend this definition to mixed derivative. For example, the definition of $\frac{\partial^2 E}{\partial_h f \partial t}$ is given below.

$$\frac{\log E_{t+1}(2f) - \log E_{t+1}(f) - \log E_t(2f) + \log E_t(f)}{\log f}$$

4. CLASSIFICATION TASKS AND FEATURE SETS

To test the efficacy of the above features, we need to choose a classification task. Two classification tasks for which there has been much recent work are

Genre classification: Musical genre is basically musical style. With the preponderance of "fusion music", it is not easy to define genre. Hence genre classification is inherently subjective. A recent paper [4] reports an accuracy of around 60% using only low-level features. [3]

[2] The expression given is really the discrete approximation of the derivative.

[3] The maximum accuracy is around 90% using higher-level features like beat.

Instrument classification: Instrument identification has more objective ground truth. The early studies in identification are [5] [6] [7]. A recent study [8] reports a classification accuracy of 77% for instrument families using kNN classifier. (The same study reports an accuracy of 97% using Quadratic discriminant analysis. We report the results only for kNN classifier.)

We choose both the above tasks to test the features listed above. Since instrument timbres are defined by the harmonic structure of the signals, we expect the proposed features to be more effective in the instrument classification task.

4.1. Genre classification

The MARSYAS [4] software for genre classification is available in the public domain[4]. We used a database which is a subset of that used [4]. The genres represented in the database used in this study are: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock.

There are approximately 100 files for each class in the database. All clips are sampled at 22050 Hz and are of 30 sec in duration.

4.2. Instrument family classification

For the instrument family classification task, we use the McGill University Master Samples database (MUMS) [9]. We use the following families. brass, keyboard, percussion, string, and woodwind.

Each file contains a two channel (stereo) recording of a single note played by a instrument. The sampling is at 44100 Hz. The average of the two channels was downsampled to 22050 Hz and used in the experiments. Each class has approximately 100 files. Very small files were ignored.

4.3. Baseline features and test features

As a reference for comparisons, we define a baseline system. The baseline system uses a mix of spectral, temporal, and cepstral features. For ease of comparison, these are the same as those used in [4]. The features are: spectral centroid, spectral rolloff, spectral flux, zero crossings, five mel-frequency cepstral coefficients, and percentage of low-energy frames over the entire clip.

The harmonicity and dynamics features proposed in section 3 have only partial information about the signal spectrum. Hence the following two features were added to the proposed feature list.

[4]from `http://www.cs.princeton.edu/~gtzan/marsyas.html`

uniformity: The uniformity measure for a frame is the "entropy" of the energy distribution of the frame.

$$-\sum_f \left(\frac{E(f)}{\sum_f E(f)}\right) \log \left(\frac{E(f)}{\sum_f E(f)}\right)$$

bandwidth: The bandwidth is defined as $\sqrt{\frac{\sum_f (\overline{f}-\log f)^2 E(f)}{\sum_f E(f)}}$ where $\overline{f}$ is the spectral centroid defined on logarithmic frequency $\overline{f} = \frac{\sum_f E(f) \log f}{\sum_f E(f)}$

These two features have been used in [10] for music classification.

For each clip, the above features are calculated for each frame (except the percentage of low energy frames feature).

5. RESULTS

In all the experiments we used the entire duration of the clips in the database. (MUMS clips have varying duration.) The audio data is split into frames of size 512 samples with 256 sample overlap. Hamming window was used. Features are calculated for each frame. *The feature vector of the clip is given by the means and standard deviations of the frame features*. Thus the baseline system uses a 19-dimensional feature vector. The feature vector size for the proposed system is 16. The feature vectors are calculated using Matlab.

The classifier used is kNN with $k = 3$. Training was done using 90% of the samples and testing using the remaining 10%. Different splits were used for different iterations. The average results of 100 iterations are reported. We used the MARSYAS software for only this part. The following experiments were performed. For ease of reference, the results are consolidated in table 1.

1. The proposed features perform worse than the baseline features for the genre classification task. The proposed features perform equally well for instrument classification task. *It should be noted that the proposed features are more compact*.

2. We augmented the baseline features with AM/HM ratio. The AM/HM ratio is defined for each frame. The mean and standard deviation of these values are used as additional features for the clips. Hence the feature vector dimensionality increases by 2. The classification accuracy drops slightly for genre classification and increases for instrument classification.

3. We use the means and standard deviations of delta cepstral coefficients. The classification accuracy improves for both classification tasks.

4. If the baseline features are augmented with the proposed features, the classification accuracy increases for both tasks. *The augmented feature set provides the best classification accuracy for both the tasks.* It should be noted

Features	feat dim	Accuracy
Baseline	19	60.81 ± 4.27
Proposed	16	55.29 ± 4.44
Baseline + AM/HM	21	59.95 ± 4.23
Baseline + Δ cep	29	62.9 ± 4.56
Baseline + Proposed	35	64.32 ± 4.06

Genre classification

Features	feat dim	Accuracy
Baseline	19	83.67 ± 4.72
Proposed	16	83.23 ± 5.49
Baseline + AM/HM	21	84.31 ± 4.70
Baseline + Δ cep	29	86.13 ± 4.24
Baseline + Proposed	35	88.29 ± 4.46

Instrument classification

Table 1. Summary of results. The table lists the features used, dimension of the feature vector, and the classification statistics averaged over 100 runs.

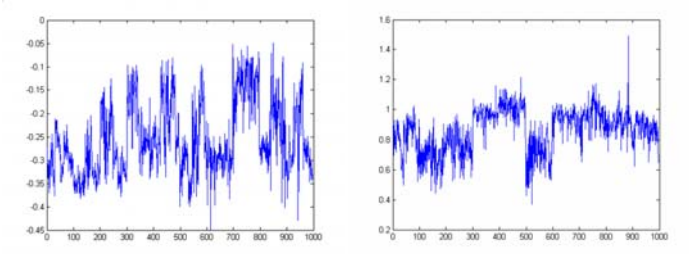

Fig. 2. Plot of harmonic and second derivatives for genre files. Each genre is represented by 100 files. X-axis shows the file number. Hence each interval of 100 roughly corresponds to one genre. It can be seen that the derivative values reflect the underlying genres.

that the genre classification accuracy for humans is 70% [4] where as the augmented features provide an accuracy of 65%. Human performance figures for instrument classification are not available since the instrument files contain a single note only.

6. DISCUSSION

Harmonicity and dynamics are powerful grouping principles used by the human auditory system. Features are at the heart of pattern recognition tasks. In this paper we have defined features based on the principles of auditory scene analysis. While the delta features are widely used in speech and audio processing, the notion of harmonic derivative is novel. These features provide a good description of musical timbre.

The features provide an increase in classification accuracy for both genre and instrument classification. It is encouraging to note that the features are "cepstrum-equivalent". Cepstral features have provided good accuracy in instrument classification [6].

The interaction between different features needs to be studied. It is important to arrive at the optimum combination of features for the task considered here.

In [8], it has been shown that Quadratic Discrimination Analysis (QDA) achieves an accuracy of around 20% more than classifiers like SVM and kNN. It will be interesting to use the features proposed here in conjunction with QDA. These directions are being pursued.

Acknowledgements The authors thank Tan Choon Woei for help in simulations. The authors also thank Dr. George Tzanetakis for making the MARSYAS database available.

7. REFERENCES

[1] M Cooke and D P W Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication*, 2001.

[2] A S Bregman, "Auditory scene analysis: hearing in complex environments," in *Thinking in sound: The cognitive psychology of human audition*, S McAdams and E Bigand, Eds., pp. 10–36. Clarendon Press, 1992.

[3] S H Srinivasan and M Kankanhalli, "Harmonicity and dynamics based audio separation," in *ICASSP*. 2003.

[4] G Tzanetakis and P Cook, "Musical genre classification of audio signals," *IEEE SAP*, vol. 10, no. 5, 2002.

[5] K D Martin and Y E Kim, "Musical instrument identification: A pattern-recognition approach," in *Meeting of Acoustical Society of America*. 1998.

[6] J C Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," in *JASA*, number 105.

[7] A Eronen and A Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," in *ICASSP*. 2000.

[8] G Agostini, M Longari, and E Pollastri, "Content-based classification of musical instrument timbre," in *International Workshop on Content-Based Multimedia Indexing*. 2001.

[9] F Opolko and J Wapnick, "McGill University Master Samples (CDROM)," 1987.

[10] S Golub, "Classifying recorded music," 2000, MSc in Artificial Intelligence, Division of Informatics, University of Edinburgh.