# DETECTING MUSICAL SOUNDS IN BROADCAST AUDIO BASED ON PITCH TUNING ANALYSIS

*Yongwei Zhu, Qibin Sun and Susanto Rahardja*
Institute for Infocomm Research, A*STAR
21 Heng Mui Keng Terrace, Singapore 119613
E-mail: {ywzhu,qibin,rsusanto}@i2r.a-star.edu.sg

## ABSTRACT

Detecting the presence of musical sounds in broadcast audio is important for content-based indexing and retrieval of auditory and visual information in radio and TV programs. In this paper, we propose a novel approach for musical sounds detection in broadcast audio based on the analysis of the characteristic feature of musical tones, pitch tuning. A spectral analysis method is presented for detecting the evidence of pitch tuning in the audio signal. Unlike the existing methods for discriminating speech and music, the proposed technique is not limited by inadequate training data, and it can deal with the case of music mixed with speech. In addition, the technique can be efficiently implemented for real-time application. Experiments based on TRECVID data set have shown good performance of the proposed technique.

## 1. INTRODUCTION

Broadcast audio in radio and TV programs consists of different types of sounds, primarily speech and music. Detecting the presence of musical sounds is thus important for automatically indexing the contents of the broadcast information. Such detection can be very useful in identifying commercials for broadcast monitoring.

Researchers have developed some techniques for classifying audio samples into music and speech [1-7]. These methods basically adopted a supervised classification approach consisting of two steps: (1) feature extraction to characterize the acoustic content of the audio; and (2) classification based on the extracted features using classifiers such as Gaussian Mixture Models, Support Vector Machine or Hidden Markov Models, which are essentially trained by training data samples. These discriminative methods have three problems. First, a large number of features are used, which requires heavy computation. Second, the limited training data does not have full representation capabilities for all kinds of musical and speech sounds in the world thus limiting the performance of the system. And third, they have not addressed the ambiguous cases of speech over the music background which are not uncommon in broadcast audio such as commercials and news programs. Minami et al. [8] presented a method for detecting music and speech separately from TV audio based on an edge detection technique. However it assumes an ideal case in which the music volume is significantly higher than the speech volume and only TV drama video data is used in the experiment.

To address the above-mentioned problems, we explore the unique and generic characteristic feature of musical audio, pitch tuning of musical tones for the detection of musical sounds in audio. Pitch tuning is a common practice in music performance, in which the pitches of the notes of the instruments are tuned (adjusted) so that the notes played by all those instruments can sound harmonically. As a result, the pitches of all the notes of a piece of music have a common reference pitch. The proposed method analyzes the spectrum of the signal and detects the evidence of pitch tuning, or the reference of a common pitch, which indicates the presence of music tones. This method is robust against the presence of non-musical signals, including speech. Using this method, the musical sounds can be detected for speech over music background even though the music has a lower volume such as the news briefing at the beginning of news program. This method is also very efficient, thus it can be adopted for real-time applications.

This paper is organized as follows: Section 2 presents our proposed approach for detecting musical sounds in audio. Section 3 presents experimental results based on video data from TRECVID05. And section 4 presents the conclusion.

## 2. THE PROPOSED APPROACH

### 2.1. Background

The basic element of music is note, which corresponds to musical tones in audio. The primarily properties of music tones are pitch, timbre, volume and duration. The absolute pitch of a particular tone of an instrument is determined by two factors: (1) the pitch tuning scale and (2) the reference pitch. Equal temperament is the de facto standard of tuning scale, which ensures the key can be transposed freely to any

pitch of the 12 pitches in an octave. And the standard pitch or concert pitch A440 (440Hz for note A) is now the universal frequency that all instruments are set to, though there are numerous historical reference pitches (e.g. A439, A452, and etc.). In reality, the musical pieces may adopt any reference pitch as long as all the pitches are tuned to a common reference pitch. Because the pitch distance between any two adjacent notes is 1 semitone (1/12 octave), so a reference pitch can be maximally half semitone away from the standard pitch A440.

In musical audio, a tone exhibits a number of harmonic series or peaks in its frequency domain, and the energy is mostly concentrated on the beginning several peaks including the fundamental frequency. The peaks besides fundamental peak are almost precisely in tune with the fundamental peak. For example, the second and fourth harmonics have identical tuning pitch with the fundamental peak, since they are exactly 1 and 2 octaves above the fundamental frequency. While the third harmonic is only 2 cents (0.02 semitone) away from the tuning pitch of fundamental peaks.

### 2.2. Pitch tuning analysis

We propose a novel spectral analysis technique for analyzing the presence of pitch tuning in audio signals. The basic idea is: if most of the peaks of high energy in the spectrum have a same frequency tuning distance to a common reference pitch for a period of time, then it is very likely that the sound contains musical tones that have the tuned pitches.

The technique has 4 steps as shown in Fig. 1.

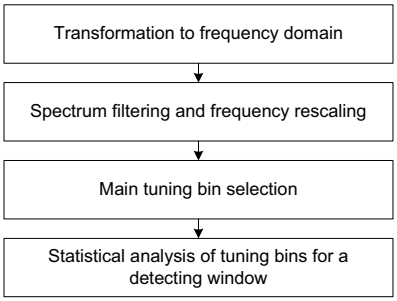



**Figure 1: Four steps in pitch tuning detection**

In step 1, the signal is framed and transformed into the frequency domain. We choose a frame size of 2048 samples and stepping size of 1152 samples (the frame size of MPEG-1 Layer 2 and Layer 3 audio). The Constant Q Transform (CQT) [9] is ideal for signal representation in frequency domain, in which frequency distance is proportional to pitch distance in music scale (e.g. semitone). However, CQT involves heavy computation and can hardly be executed in real-time. We have used FFT for the spectrum computation, which can be done very efficiently in PC platform. The order of the transform is 15 ($2^{15}$ points), thus it has adequate resolution in the lower band for the typical audio sampling frequency, e.g. 32kHz.

In step 2, the spectrum obtained by FFT is filtered and rescaled into music scale, such that frequency distance is proportional to pitch distance. This is done by the following processing.

A frequency band starting from point 256 (about 125 Hz for sampling frequency of 32kHz) and with a bandwidth of 5 octaves is selected in the spectrum. This band is chosen because it well covers the energy of musical tones. The spectrum in the frequency band is then filtered spectral-temporally: a local peak frequency sample is selected if its energy is higher than any samples in the previous, current and following frames within its ½ semitone frequency range excluding the peak's own frequency. This peak sample is then assigned to a new spectrum with resolution of 1/10 semitone. It can be noted that many samples in the new spectrum have zero values, because the filtering bandwidth ½ semitone is larger than frequency resolution 1/10 semitone. The obtained spectrum has 5x12x10=600 frequency samples and has a musical (logarithm) frequency scale.

In step 3, we analyze how the energy in the spectrum is distributed in terms of tuning frequency over a range of 1 semitone. Since the frequency resolution in the spectrum is 1/10 semitone, we naturally group and sum the frequency samples into 10 bins, where each bin contains the samples of frequency index with same modulus after being divided by 10. In another word, 1 bin contains the samples that are integer numbers of semitones away (precisely in tune) to each other. In the case of music tones, the energy will tend to concentrated into 1 of the 10 bins. In this step, the main tuning bin with maximal energy is picked to represent the current frame.

In step 4, a detecting window (Fig. 2) corresponding to 1 second time is analyzed based on the frames in the window. For sampling frequency 32kHz, each detecting window has 28 (overlapped) frames. A tuning histogram is constructed based on the main tuning bins of each frame. Two values are then computed for the detecting window. First, the *prime pitch tuning bin* is decided based on the bins with maximal population. Second, the pitch *in tune ratio* is computed as the population of the prime tuning pitch bin divided by the number of frames in a detecting window.

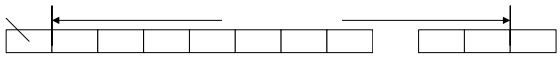


**Figure 2: Frame and detecting window**

Fig. 3 and 4 illustrate two examples of the spectra in step 1, 2 and 3. Fig. 3 shows the spectra of a frame of musical signal, and Fig. 4 shows those of a speech signal. The top subfigure is the spectrum after FFT; the second subfigure is the spectrum after filtering in step 2; and the third subfigure is the spectrum in musical scale with

resolution 120 points per octave. Note that each spectrum shown is only part of actual spectrum. The fourth subfigure illustrates the energy distribution over tuning bins (10 bins for a semitone). It can be seen that for musical signal, the energy is concentrated in one tuning bin, whereas it is not the case for the speech signal.

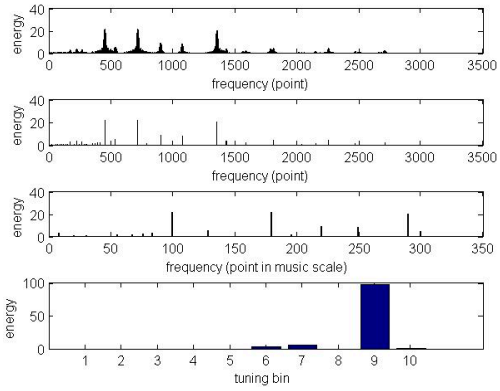



**Figure 3: Spectrum of a musical signal**

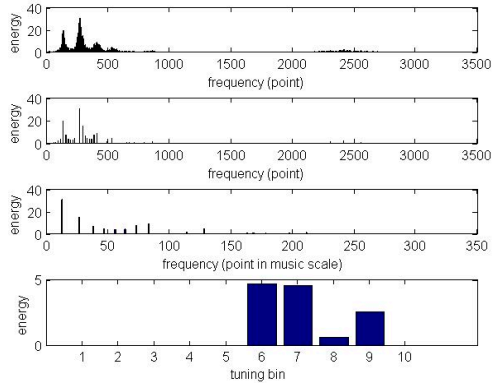
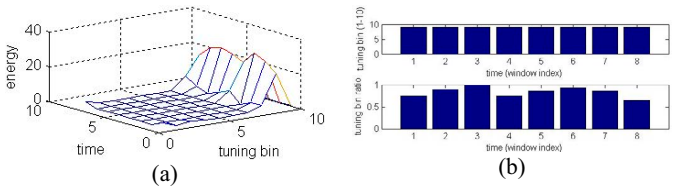



**Figure 4: Spectrum of a speech signal**

Fig. 5 illustrates the examples of pitch tuning analysis in step 4. The first example (subfigures (a) and (b)) is for a pure musical sound. It can be seen that the prime tuning bins are uniformly number 9, and the *in tune ratios* are all above 0.6. The second example (subfigure (c) and (d)) is for a speech signal. And the prime tuning bins are basically random, and the *in tune ratios* are all lower than 0.3. The third example (e)(f) is for a sound of music mixed with speech. In this case, the distribution of prime tuning bins is slightly noisy. However the tuning bins are almost uniformly number 10, and most of the *in tune ratios* are above 0.4, so the musical sound is still detectable.



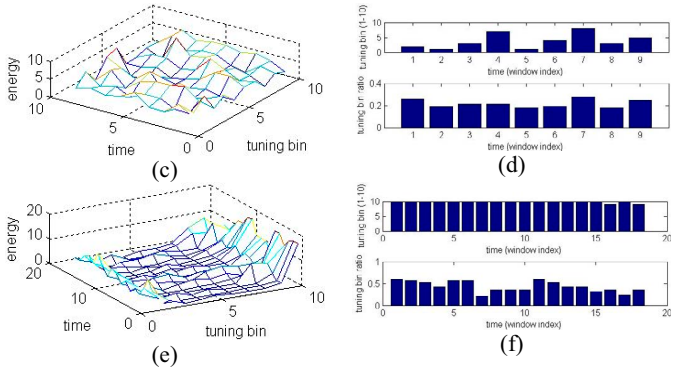




**Figure 5: Examples of music sound, speech and music mixed with speech.**

### 2.3. Musical sounds detection

Detecting the presence of musical sounds is done based on the result of pitch tuning analysis. We basically consider the temporal continuity of the prime tuning bins and the in tune ratio values of each detecting window. The details are given as follows.

The musical sound detection is conducted for $N$ consecutive windows. The presence of musical sound is claimed if two conditions are met: (1) the tuning bins of the $N$ windows are within two adjacent bins (e.g. 2 and 3, and 10 and 1), and (2) the *mean in tune ratio value* is above a threshold $T$. In the first condition, two adjacent bins are used to accommodate the sampling error of the tuning bins. In our experiment, we set $N$=5 and $T$=0.3. The detection is done repeatedly after each shifting of a detecting window.

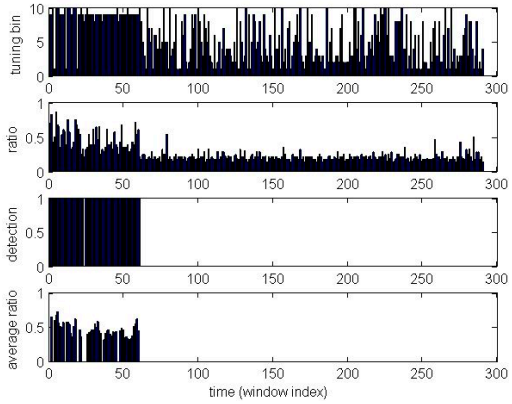



**Figure 6: News audio with pure music, speech mixed with music and speech sounds**

Fig. 6 shows the detection result of a news clip of 5 minutes. The upper two subfigures shows the *prime tuning bin* and *in tune ratio* for each detecting window. The lower two subfigures illustrate the windows with detected musical sound (1 for detected, 0 for undetected) and the corresponding *mean in tune ratio values* (only of the center window of the consecutive window sequence). In this

example, the first 6 seconds are of pure music sounds. The next 50 seconds are the news briefing that is speech mixed with music. And the rest of the signal is the speech sound.
Fig. 7 shows another example, which consists of several sound clips of advertisement, each of which contains musical sounds partly mixed with speech.

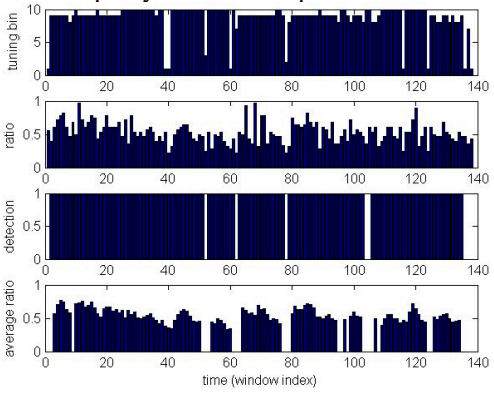



**Figure 7: Advertisement audio with music and music mixed with speech**

## 3. EXPERIMENTAL RESULTS

We evaluate the proposed method using the video data in TRECVID 2005 data set. The data set is the VCD quality recordings of TV program of different channels in several languages, including English, Chinese and Arabic. In our experiments, we choose 8 videos totaling 4 hours from CCTV channel. The videos consist of mainly news programs, and some commercials and drama. We manually separate the data into the three categories and label the ground truth of the presence of musical sound.

In evaluation, a detected music is considered correct if the detecting window is within 3 seconds of the labeled ground truth. It will be a false positive if no labeled ground truth is within 3 second. And it will be a false negative if all the detecting windows within 3 second are negative, whereas the current detecting window has a label of music.
The detection result is shown in Table 1. It can be seen that the result for drama video is quite good. For news video, some of the false positives are caused by the musical sound presented in the reporting site, which are not labeled in the ground truth. In commercial, the false negative is a bit higher. This is mainly due to the pure percussion sounds in the music which is considered musical but has no pitch (or tones).

**Table 1: Musical sounds detection result**

| Category | Recall | Precision |
|---|---|---|
| News | 96.5% | 94% |
| Commercial | 91% | 99% |
| Drama | 98% | 98.5% |
| Total | 95.5% | 96% |

Our detection algorithm is implemented using Visual C++ on PC platform, with Intel signal processing library and MPEG audio decoding library. On a PC with Pentium IV 1.6GHz, the speed of the algorithm is 4 times of the audio playing speed, thus the musical sound detection can be performed in real time.

## 4. CONCLUSION

We have presented an effective and efficient technique for detecting musical sounds in broadcast audio. The technique is based on the analysis of the characteristic pitch tuning feature of the musical tones in the audio. This method is shown to be robust against the presence of other sound mixed with musical sounds. The proposed technique is also very efficient, which can be implemented in real time.

In our approach, the detection of speech sound is independent from music sound detection. The separation of pure music and music mixed with speech is our future work. Since the proposed method is based on analyzing pitch tuning, percussion sounds that have no pitch cannot be correctly detected as musical sound using this method. Detection of such sounds shall involve timbre and rhythm analysis. However, in broadcast audio, usually the percussion sounds are shortly followed by pitched musical sounds, which can help identify the percussion musical sounds. This task would also be a subject of our future work.

## 5. REFERENCES

[1] J. Sounders, "Real-Time Discrimination of Broadcast Speech/Music", *Proc. ICASSP 1996*, pp993-996.
[2] E. Scheier and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator", *Proc. ICASSP 1997*, pp1331-1334..
[3] M.J., Carey, E.S., Parris, and H., Lloyd-Thomas, A comparison of features for speech, music discrimination, *Proc. ICASSP 1999*.
[4] E.S., Parris, M.J., Carey, and H., Lloyd-Thomas, Feature fusion for music detection, *European Conf. Speech Comm. Technology*, pp. 2191-2194. 1999.
[5] G., Williams, and D., Ellis, Speech/music discrimination based on posterior probabilities, *European Conf. Speech Comm. Technology*, pp. 687-690. 1999.
[6] K., El-Maleh, M., Klein, G., Petrucci, and P., Kabal, Speech/ music discrimination for multimedia application, *IEEE Internat. Conf. Acoust., Speech, Signal Process.*, pp. 2445-2448. 2000.
[7] J., Ajmera, I., McCowan, and H., Bourlard, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," Speech Communication 40, pp. 351-363, 2003.
[8] K., Minami, A., Akutsu, H., Hamada, and Y., Tonomura, "Video handling with music and speech detection," In *IEEE Multimedia*, vol. 5, no. 3, pp. 17–25, 1998.
[9] J.C. Brown, "Calculation of a constant Q spectral transform". In *J. Acoust. Soc. Am.*, 89(1):425-434, 1991.