A METHOD FOR SEPARATING DRUM OBJECTS FROM POLYPHONIC MUSICAL SIGNALS

Wendong Huang

School of Computing
National University of Singapore
Singapore 117543
huangwd@comp.nus.edu.sg

Ye Wang

School of Computing
National University of Singapore
Singapore 117543
wangye@comp.nus.edu.sg

ABSTRACT

An additional coding of auditory objects for packet loss concealment has been proven to be effective in music streaming applications. This paper describes a new extension to our previous method in separating drum objects from polyphonic music signals with improved performance. After a simple time domain separation method employed in our early system, we propose in this paper a novel frequency domain technique, a *Tonal-components Tracking and Attenuation (TTA)*, to suppress quasi-stationary auditory objects such as singing voice in the separated drum objects. Experimental results show that the new method is an effective pre-processing step to separate drum objects from polyphonic music signals. This method helps to improve accuracy of drum clustering and to mitigate the pitch and harmonic structure mismatch problem when applied in packet loss recovery in music streaming.

1. INTRODUCTION

Object oriented coding is a foundation of MPEG-4 audio coding standard. We have shown that encoding only the percussive sound in a music stream as secondary data (metadata) can achieve good perceptual quality in packet loss concealment with minimal redundancy [1][2][3]. Our methods achieve much better perceptual results when compared to traditional error concealment methods. From our subjective evaluation, listeners seem to be very sensitive to errors which distort beat pattern of music.

The key idea of our packet loss mitigation scheme is to send receiver a small amount of metadata as a “header” segment prior to streaming the audio data. The content-based codebook is constructed from all detected drum objects, which are simply separated from the music clip with a time window of a fixed duration. The time window approximates the contour of drum objects [1]. However, the drum objects obtained with the time window alone are usually contaminated with other sustaining sound such as singing voice. This can cause two problems. First, the tonal components mixed in the drum objects make the extracted feature noisy, resulting in inaccurate clustering during the drum codebook generation. Second, transmitting a contaminated drum codebook is not economic in channel utilization and can also create *spectral fine structure disruption*, when applied in packet loss recovery in the receiver [4]. This problem is illustrated in Figure 1, where Frame Y is lost and our scheme is designed to recover the harmonic structure (tonal components) from the neighboring frames via interpolation or extrapolation and the drum object (codeword) from the codebook *without any harmonic structure*. However, if the codeword is contaminated with a different pitch and harmonic structure during the process of codebook construction in the sender side, our error mitigation scheme will be less effective. For example, if the codeword is contaminated with a female singing and the lost frame contains only a male singing, we will have a mixture of female and male singing after the error mitigation algorithm which is not desirable. Therefore, it is necessary to suppress tonal components mixed in the separated drum objects during the codebook construction phase at the sender side.

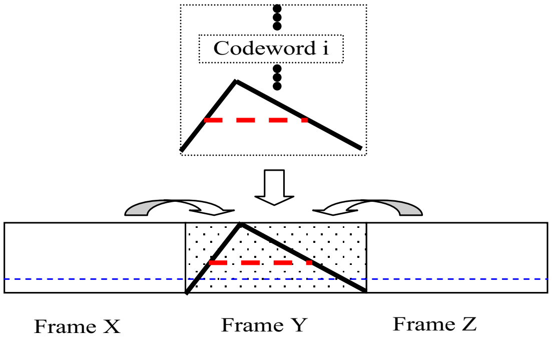

Figure 1. Pitch mismatch in packet loss recovery. Blank rectangles indicate correctly received frames and the shaded rectangle indicates the lost frame. The dashed lines indicate different pitches. The triangle represents the drum object taking from the codebook.

In this paper, we propose a novel pre-processing step in the sender side to process all detected drum objects before sending them for clustering [1]. Our pre-processing algorithm is a frequency domain method and is designed to suppress the harmonic structure in the drum objects separated with a time window alone.

2. SYSTEM DESCRIPTION

A block diagram of our system is depicted in Figure 2, where the sub-system in the dashed rectangle highlights novel contributions of this paper. Other parts are the same as in [1].

0-7803-9154-3/05/$20.00 ©2005 IEEE

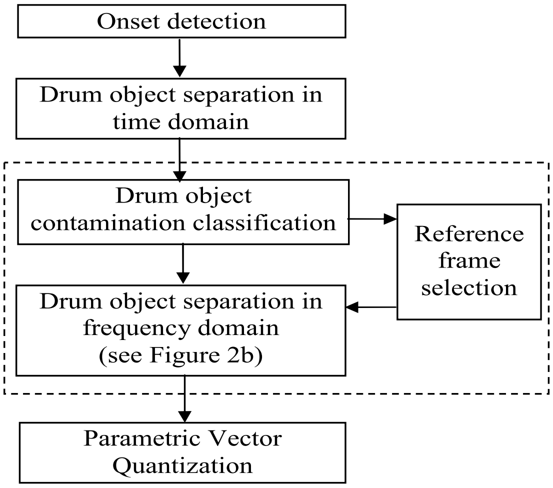
Onset detection
Drum object separation in time domain
Drum object contamination classification
Reference frame selection
Drum object separation in frequency domain (see Figure 2b)
Parametric Vector Quantization

Figure 2a. System overview

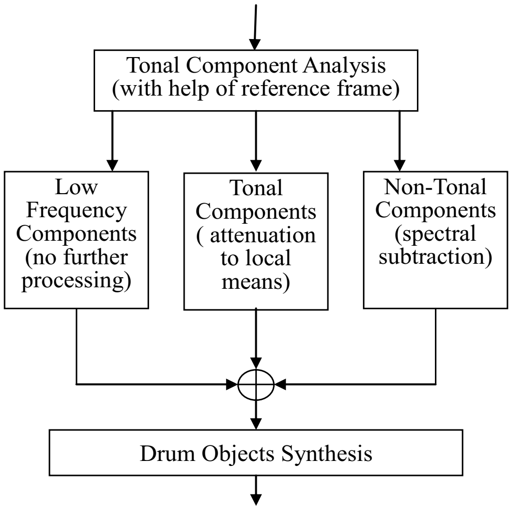
Tonal Component Analysis (with help of reference frame)
Low Frequency Components (no further processing)
Tonal Components (attenuation to local means)
Non-Tonal Components (spectral subtraction)
Drum Objects Synthesis

Figure 2b. Block diagram for drum object separation from polyphonic music in the frequency domain

2.1. Drum Object Contamination Classification

To enable better drum separation in frequency domain, we roughly classify all detected drum objects into three classes as described in [2]. Our simple threshold based classification is shown in Figure 3.

Let us denote X, Y and Z to represent the pre-drum frame, drum frame and post-drum frame, respectively. The binary classification at node A is based on signal harmonicity of frame X, Y and Z. If any of the three frames has a high level of harmonicity, it goes to node B, otherwise, it belongs to class c1. This step is designed to find drum objects without contamination (class c1). The classification at node B is based on frequency domain similarity s between frame X and Z according to (1).

$$s = \frac{\sum_k |X(k)| \cdot |Z(k)|}{\max\left(\sum_k |X(k)|^2, \sum_k |Z(k)|^2 \right)} \tag{1}$$

where $x(n) \leftrightarrow X(k)$, $y(n) \leftrightarrow Y(k)$ and $z(n) \leftrightarrow Z(k)$ represent the three frames in the time and frequency domains. When s is greater than a predefined threshold, frames X and Z are considered similar, thus belonging to class c2. Other objects are classified to c3.

Based on above classification, the reference frame can be selected for drum object separation. The reference frame should have similar harmonic structure as the drum frame. We select pre-drum frame as the reference frame for class c2, and post-drum frame as the reference frame for class c3. The reason for such selection is explained as follows. Our drum codewords have a fixed duration of 2048 PCM samples, which is about 46 ms for the sampling rate of 44.1 kHz. This duration is usually shorter than that of a drum sound which is more than 100 ms. However, our drum codewords contain the most significant part of the drum and are sufficient for our purpose of error mitigation. Nevertheless, our algorithm always chooses pre-drum frame as the reference for the class c2 due to the fact that post-drum frame usually has considerable amount of drum energy.

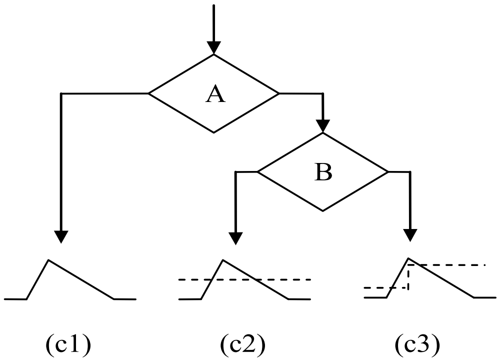

Figure 3. Two-step classification of detected drum objects into three categories: (c1) represents drum objects without contamination; (c2) represents drum objects mixed with sustaining sound; (c3) represents drum objects mixed with note onset. The dashed lines indicate pitch lines and the solid lines indicate loudness.

2.2. Drum Object Separation in Frequency Domain

Drum object separation in frequency domain is the key contribution of this paper. Our objective is to maintain drum object while removing other sustaining sound, especially singing voice, as much as possible.

In the course of our research, we have found that singing voice is the major contamination in the detected drum objects for two reasons: 1) singing voice is often the most dominant component in popular music in addition to drums; 2) the fundamental frequency and formants of singing voice are within the most sensitive area of human auditory perception.

We have tried to remove singing voice from contaminated drum objects using a conventional spectral subtraction algorithm and different variants such as [5], but failed. The reason for the failure is explained as follows.

The spectral subtraction algorithm is originally proposed for noise suppression in speech communication systems with the assumption that the noise spectrum is relatively stable and can be estimated quite accurately from the previous frames. This assumption is not valid in our case. The "noise" in our drum objects is singing voice, which has totally different characteristics in comparison to white noise.

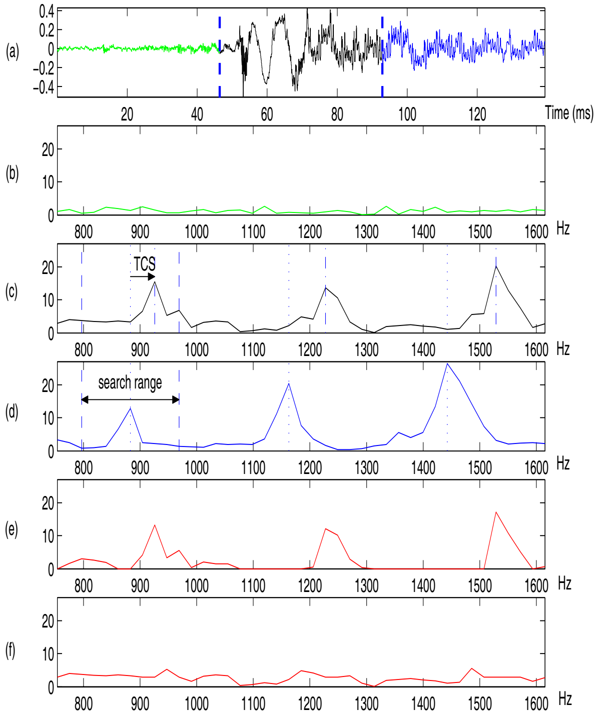

Figure 4. (a) waveform of the pre-drum, drum and post-drum frames; (b) spectrum of the pre-drum frame; (c) spectrum of the drum frame; (d) spectrum of the post-drum frame which serves as the reference frame in this case; (e) residual after conventional spectral subtraction between drum frame and reference frame; (f) residual after TTA between drum frame and reference frame.

We employ an example in Figure 4 to show how TTA works. Figure 4(a) shows an audio clip of 138 ms which is divided into 3 frames. During the drum frame the drum component, whose spectrum is more noise-like, is quite dominant. From the post-drum frame, the drum energy decay rapidly while singing voice starts to build up. This is illustrated in Figure 4 (c)(d). The spectrum of singing voice has clear harmonic structure with major energy concentrated in the fundamental frequency and harmonics. Furthermore, the harmonic structure in neighboring frames can change in the form of *tonal components shifting (TCS)* as shown in Figure 4. This illustrates clearly why conventional spectral subtraction techniques do not work in our application. Figure 4(e) shows the residual after a conventional spectral subtraction where negative components of the magnitude spectrum are set to zero and Figure 4(f) shows the residual after TTA. Comparing the two residuals we can see the difference easily.

The block diagram of the proposed method is shown in Figure 2. We assume that drums used in a song do not have harmonic structure. We divide the spectrum of the drum frame into three parts: low frequency components below 300 Hz, tonal and non tonal components as shown in Figure 2(b). Based on our observation, in line with the results in [6], the frequency band below 300 Hz is almost free from singing voice and other instrumental sound. Therefore, we keep the low frequency components, mostly belong to drum objects, untouched.

For frequency band above 300 Hz, we employ a peak picking algorithm, similar to that in the psychoacoustic model 1 in MPEG-1 audio informative part [7], to identify tonal components in the reference frame where the drum object is not so dominant. Then we use the identified tonal components in the reference frame to track the corresponding tonal components in the drum frame via a simple neighborhood search algorithm (see Figure 4 (d)). The search range is determined empirically based on the fact that the frequency shift of the singing voice between two neighboring frames is usually very small, but sufficient to make conventional spectral subtraction algorithms to fail.

The identified tonal components are attenuated by replacing them with their local means of four neighboring frequency components excluding the tonal components themselves. Effectively, this is a spectral flattening method. It is worth mentioning that above algorithm is conceptually analogous to the motion compensation (MC) algorithm in video coding.

Although singing voice is suppressed significantly after this step, it is still perceivable due to the leakage of FFT and possible remaining tonal components. To further suppress the harmonic components, we perform a conventional spectral subtraction for the non-tonal components with the reference frame. After all these steps, we have managed to separate the drum objects quite satisfactorily from polyphonic music signals.

3. EXPERIMENTAL RESULTS

To evaluate the effectiveness of our algorithm, we have performed subjective evaluations on two sets of samples. The first set of samples is artificially generated by mixing pure drums and vocals. This set of samples allows us to compare the drum objects with and without TTA against original drum objects (ground truth). Signals from a typical drum set, including bass drum, tom drum, snare drum, crash and hi hat, are chosen to mix with male and female vocals. The second set of samples is chosen from commercially available pop songs. For this kind of samples, we don't have the original drum samples, but have only the drum objects with and without TTA.

We carry out our experiments on a group of 12 subjects (male and female graduate students). All subjects are asked to evaluate the audio quality using the mean opinion score (MOS), which is a 5-point scale (5 – excellent, 4 – good, 3 – fair, 2 – poor, and 1 – bad). We have used 10 musical programmes for the evaluation: five produced by mixing drums and vocals, and five selected from pop music clips. All 10 programmes were around 30 seconds in duration. They are all monophonic, 16 bit, 44.1 kHz PCM samples. For each audio sample in the first set, we have prepared 4 copies for testing, the original drum object as reference, the original drum object with a hidden name, the drum mixed with vocals, and the drum cleaned by TTA, also with hidden names. The 3 hidden samples were arranged in random order for evaluation. Samples in the second set do not have original drums. The subjects were asked to refer to the original drum objects without any contamination from singing voice in the first set to experience the quality at MOS scale of 5. For the sake of fairness, all test samples were arranged in random order.

The obtained results are shown in table 1 and 2.

Drum type		Bass drum	Tom drum	Hi hat	Crash	Snare drum
Original	Mean	4.58	4.31	4.64	4.59	4.41
	Stddev	0.42	0.98	0.64	0.44	0.89
Mixed	Mean	1.73	2.34	2.23	2.48	2.77
	Stddev	0.82	0.92	0.90	0.94	1.01
Cleaned	Mean	2.82	2.64	3.10	2.63	3.29
	Stddev	0.84	0.92	0.80	0.98	1.09

Table 1. Evaluation results with artificial audio samples

Sample index		1	2	3	4	5
Mixed	Mean	3.00	2.72	1.86	2.02	2.22
	Stddev	0.50	0.99	0.83	0.88	1.10
Cleaned	Mean	3.85	4.05	3.19	2.72	3.88
	Stddev	1.05	0.56	1.03	1.04	0.74

Table 2. Evaluation results with commercial audio samples

The experimental results from both sets of test samples show that drum objects after TTA have significantly less undesirable harmonic structure.

As per Table 1 and 2, TTA achieves different quality improvements on different audio data. With the artificial audio samples, the best separation occurs with the bass drum sample mixed with vocals. The worst separation occurs with the crash sample. This is not very surprising. Bass drum and vocal are clearly separated in the frequency domain, which makes the separation task easy. However, the spectra of crash and vocal overlap with each other and make the separation task difficult. A further reason for the poor result with the audio sample mixing crash and vocal can be explained from the perceptual viewpoint. Crash components spread over a broad spectrum which masks vocal sound quite well. In this case, spectral subtraction only changes the spectral characteristics of the original crash, thus making the re-synthesized crash different from the original. This shows that our non-tonal component processing can be improved by making it adaptive to signal characteristics.

In comparison with the artificial samples, TTA achieves better performance on audio clips from commercial CDs mainly due to the fact that subjects are usually more critical if the original sample is available as reference.

4. CONCLUDING REMARKS

We have presented a new technique, TTA, to separate drum objects from polyphonic music for error robust music streaming with promising results. We plan to integrate the proposed method into our packet loss mitigation scheme and examine the performance improvement in terms of human perception.

5. REFERENCES

[1] Wang, Y., Tang, J., Ahmaniemi, A., Vaalgamaa, M., “Parametric Vector Quantization For Coding Percussive Sounds in Music,” IEEE ICASSP2003, Hong Kong, China (presented at ICME2003 in Baltimore, USA)

[2] Wang, Y., Ahmaniemi, A., Isherwood, D., Huang, W., “Content-based UEP: A New Scheme for Packet Loss Recovery,” ACM Multimedia 2003, pp. 412-421, Berkeley, California, USA, November 4-6, 2003

[3] Wyse, L., Wang, Y., Zhu, X., “Application of a Content-Based Percussive Sound Synthesizer to Packet Loss Recovery in Music Streaming,” ACM Multimedia 2003, pp. 335-338, Berkeley, California, USA, November 4-6, 2003

[4] Wang, Y., Streich, S., “A Drumbeat-Pattern based Error Concealment Method for Music Streaming Applications,” IEEE ICASSP2002, Orlando, Florida, USA, May 13-17, 2002

[5] Ephraim, Y., Malah, D., “Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator,” IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-32, No.6, December 1984

[6] Alghoniemy, M., and Tewfik, A.H., “Rhythm and Periodicity Detection in Polyphonic Music,” IEEE third Workshop on Multimedia Signal Processing, pp. 185-190, Copenhagen, Denmark, September 13-15, 1999.

[7] ISO/IEC 11172-3 International Standard, “Information Technology – Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s, Part 3: Audio,” 1993