NOTE ONSET DETECTION BASED ON HARMONIC CEPSTRUM REGULARITY

Hoon Heo, Dooyong Sung and Kyogu Lee

Graduate School of Convergence Science & Technology
Seoul National University
151-742, Seoul, Republic of Korea
{cubist04; mrbilly; kglee}@snu.ac.kr

ABSTRACT

This paper presents a novel onset detection algorithm based on cepstral analysis. Instead of considering unnecessary mel-scale or any interests of non-harmonic components, we selectively focus on the changes in particular cepstral coefficients that represent the harmonic structure of an input signal. In comparison with a conventional time-frequency analysis, the advantage of using cepstral coefficients is that it shows the harmonic structure more clearly, and gives a robust detection function even when the envelope of waveform fluctuates or slowly increases. As a detection function, harmonic cepstrum regularity (HCR) is derived by the summation of several harmonic cepstral coefficients, but their quefrency indices are defined from the previous frame so as to reflect the temporal changes in the harmonic structure. Experiments show that the proposed algorithm achieves significant improvement in performance over other algorithms, particularly for pitched instruments with soft onsets, such as violin and singing voice.

Index Terms— Onset detection, cepstral analysis, music information retrieval

1. INTRODUCTION

For a couple of decades, onset detection of musical notes has been a major issue in the music information retrieval community. Being a fundamental low-level task in this field, precise note onset detection can lead to solving a lot of problems for advanced music analyses, including pitch estimation, tempo estimation, automatic transcription, and many commercial applications of music and audio processing.

Masri proposed a well-known algorithm for pitched non-percussive (PNP) onset detection applying linear weight on high frequency content (HFC) [1]. Klapuri's sub-band energy change method used a filter bank model to approximate the human cochlea [2]. A more general approach is introduced by Duxbury, where the changes in the spectrum called spectral difference or spectral flux, are used to indicate musical onsets [3].

With all the contributions made so far, however, according to the Music Information Retrieval Evaluation eXchange (MIREX) 2012, onset detection still remains a challenging problem, particularly for soft onsets. The best result for onset detection of solo singing voice is 55.9% in terms of a F-measure. In case of solo sustained strings, averaged F-measure for all participants is only 52.8%. Since most of traditional approaches use spectral energy and its difference via time-frequency analysis, detection function must be solely affected by the original source. A soft onset generally does not rapidly occur because it has a long attack interval or indistinguishable envelope shape, and becomes the main reason of difficulties in the peak-picking procedure.

Supervised machine learning-based approaches are also proposed to handle this problem. Toh *et al.* viewed onset detection as a classification problem, and used two Gaussian mixture models (GMMs) to classify audio features into onset and non-onset frames [4]. Also, they derived a fusion of four different types of acoustic features, including mel-frequency cepstral coefficients (MFCCs), linear predictive cepstral coefficients (LPCCs), and equal loudness phon values along critical bands. However, as in other machine learning algorithms, this approach is supposed to rely heavily on the training data. Furthermore, it is extremely time-consuming and laborious to annotate large amount of audio data at the frame level. Finally, MFCCs, which is the main feature of learning-based approaches, are hard to represent all kinds of characteristics of various musical instruments.

To resolve these issues on detecting soft onsets, we propose a novel approach based on *harmonic cepstrum regularity* (HCR), by focusing on the changes in the harmonic structure. HCR is expected to yield better results for sustained string instruments and singing voice which usually contains soft onsets and unexpected energy flow. Especially, it is robust to irregular changes in the formant structure caused by different singing styles. In addition, unlike spectral difference or spectral flux, HCR gives a steady detection function regardless of note strengths since only harmonic components in the cepstral domain are considered. Overview of the proposed system is illustrated in Fig. 1.

This work is supported by Samsung Electronics Co., Ltd. and by the Seoul R&BD Program.

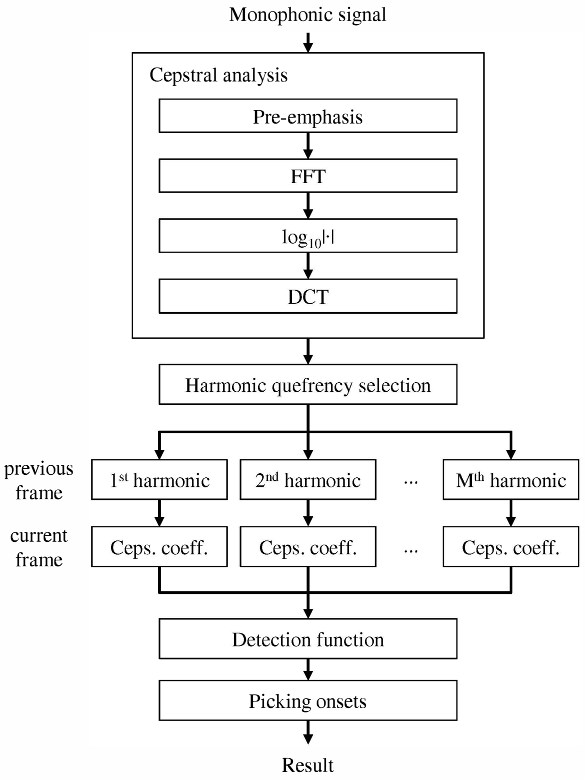

Fig. 1. System overview.

The rest of this paper is organized as follows. In section 2 we explain the reason why we applied cepstral analysis to this task. Section 3 describes the four main procedures to get positive onsets, including harmonic quefrency selection, sub-harmonic regularity function, adaptive thresholding and peak-picking. In section 4, the experiments we performed to evaluate the improvement of the proposed algorithm are presented, and we finally draw our conclusions and present directions for future work in section 5.

2. CEPSTRAL ANALYSIS

The cepstrum is originally defined as the inverse Fourier transform of the log-magnitude Fourier spectrum. For image and audio signals, discrete cosine transform (DCT) is more meaningful transformer than inverse Fourier transform, because its “energy compaction” property helps not to lose the signal information too much while most of the information are concentrated in low-frequency range of the DCT. Cepstral coefficients are determined as the results of the DCT which is the last step of cepstral analysis. These are widely used as a feature to represent timbral characteristics in many audio processing tasks, including speech recognition, speaker identification, and genre classification.

It is well known that pitch is proportional to frequency. Strictly speaking, pitch is related to frequency logarithmically rather than linearly. When pitch increases by an octave, frequency would be doubled. Mel-scale reflects this characteristic of the human auditory system. MFCCs can be derived by first taking the Fourier transform of a windowed signal, mapping the log-amplitudes of the resulting spectrum into the mel-scale, then performing DCT on the mel log-amplitudes. The amplitudes of the resultant cepstrum are the MFCCs. Similarly, LPCCs can be derived as linear predictive coding coefficients transformed into cepstra.

Almost every algorithm using MFCCs generally truncates higher coefficients. In many cases, only the first 13 coefficients are said to be enough to store the signal characteristic. However, we use a sufficient number of cepstral coefficients which are the same size as the frame length, because we would need higher quefrency resolution to get the coefficients corresponding to harmonic components more precisely. In our system, we use a Hamming window of length 2048, and therefore we obtain 2048 cepstral coefficients for every analysis frame. With a 44100 Hz sampling rate and a 87.5% overlap, we get a frame rate of 5 ms which is short enough to detect the shortest possible musical note.

Another difference between the MFCCs and our cepstral analysis is that we use a linear scale in frequency rather than a mel-scale. Mel-scaling compensates differences between frequency and subjective pitch. This psychoacoustic knowledge definitely makes some advantages when human perception (i.e. timbre, masking, etc.) is an important issue, but in this situation we do not need to consider this because we are only interested in whether the harmonic structure is noticeable or not. It is shown that the n-th order harmonic frequency for three types of musical instruments is derived as follow [5]:

$$f_n = \begin{cases} nf_1 & \text{open tube \& string instruments} \\ (2n-1)f_1 & \text{closed tube instruments} \end{cases} \quad (1)$$

where f_1 means the fundamental frequency. For singing voice, very complex calculation is required to derive the exact harmonic frequency but we can simplify vocal tract as a closed-tube type instrument. While harmonic frequencies are linearly related to the order, any scaling along frequency bands is not necessary.

We also take a pre-emphasis step by sending the input signal through a highpass filter. This process is to compensate the high-frequency part and emphasize the high-order cepstral coefficients that we want to concentrate on. Pre-emphasized signal s'_n is derived as

$$s'_n = s_n - \alpha s_{n-1} \quad (2)$$

where s_n is the original signal and the value of α is usually defined between 0.9 and 1, and we set this value to 0.97.

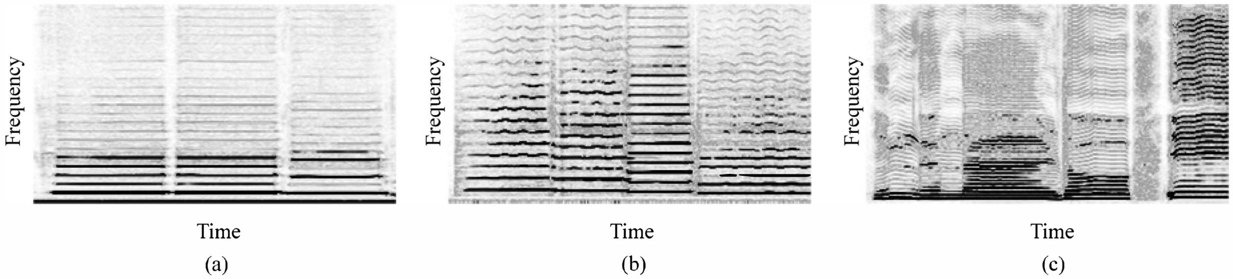

Fig. 2. Spectrogram of (a) a clarinet and (b) a violin and (c) a singing voice signal.

3. HARMONIC CEPSTRUM REGULARITY

Fig. 2 shows the spectrograms of some musical signals – a clarinet, a violin, and a singing voice signal, respectively. It is not difficult to recognize the onsets from the spectrogram simply by our intuition, because we tend to pay visual attention to several imaginary vertical lines. These are easily distinguishable due to the discontinuity of many horizontal lines, which indicate the energy of the harmonic components. In order to determine how regularly this harmonic structure of the input signals is maintained, we need to examine the amount of temporal change of the harmonic components.

To this end, we first extract the harmonic quefrencies from their cepstral coefficients, and then check how much the energy changes in the selected quefrencies. An important point is that the harmonic structure of the previous frame is applied to the current frame. In other words, cepstral coefficients of the previous harmonic quefrencies are selected to build the detection function of the current frame. The cepstral coefficients of the current frame would not be different much from those of the previous frame if the harmonic structure is stable, and thus the peak locations would remain the same. On the other hand, if the harmonic structure is changing rapidly, these peak locations will also change, resulting in smaller amplitudes for the coefficients found in the previous frame. As shown in Fig. 3, this causes the pronounced difference between a sustain and a transient.

3.1. Harmonic quefrency selection

Harmonic quefrency selection is an essential procedure to make the results reliable. Pitch estimation can be regarded as an advanced concept of this process, in a sense, so we can also extend this algorithm not only to note onset detection but also to monophonic music transcription while as long as the harmonic quefrencies are selected correctly. Therefore, conventional pitch estimators can be used to find the exact harmonic quefrencies. For example, YIN is a well-known pitch estimator [6] and the correntropy method is known to give the best result for singing voice [7]. The relationship between the fundamental frequency f_1 and its corresponding quefrency q_1 follows as below:

$$q_1 = \frac{2f_s}{f_1} \tag{3}$$

In this paper, however, we simply take an autocorrelation method to pick the harmonic peaks because pitch estimation is not our main goal and the autocorrelation function (ACF) is also enough to yield the reasonable result. 1-D Gaussian kernel function is applied to make peaks more salient, and the kernel size is set to 64 via experiments. Then we compute the ACF of the cepstral coefficient and derive the fundamental quefrency q_1 from the index of the maximum ACF value. For high order harmonics we approximated their possible ranges based on the integer harmonics assumption, and the local maxima are chosen within the possible ranges.

3.2. Sub-harmonic regularity function

As a detection function, the harmonic cepstrum regularity function (HCR) is derived by the summation of all harmonic

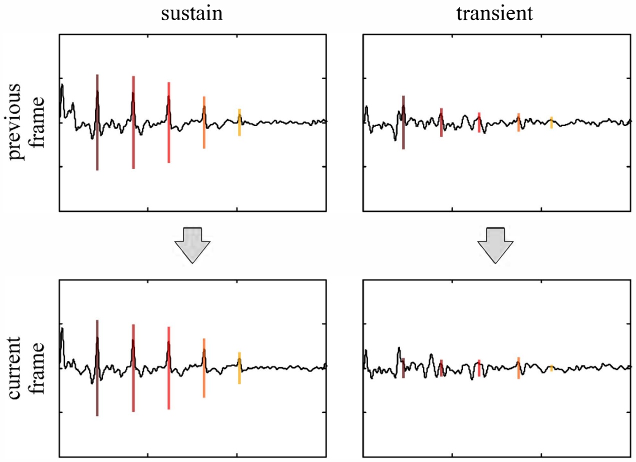

Fig. 3. Comparison between a sustain and a transient. Each vertical line represents its relative amplitude of cepstral coefficient.

cepstral coefficients, where harmonic quefrencies represent the harmonic structure of the previous frame. That is to say,

$$d_n = \sum_{k=1}^{M} C_{q_{k,n}, n-1} \tag{4}$$

where C is the cepstral coefficient matrix whose rows and columns represent quefrency and the frame index, respectively. k is the harmonic order up to M (normally $M = 5$), which depends on the instrument type and the degree of preemphasis. In general, the more harmonic components were used, the better result we would get as the detection function would fully describe the harmonic structure.

3.3. Adaptive thresholding

We now describe the adaptive thresholding procedure, which picks local minima of the detection function d_n. Since our HCR function represents how regular the harmonic structure is and this regularity is disrupted when onsets occur, local minima are supposed to be picked instead of maxima. There are several adaptive thresholding methods such as low-pass FIR filtering, median filtering, and low-pass filtering of the square of the detection function. We choose a method based on the local median because it is known to be robust by minimizing the effects caused by the outliers [8]. In this paper, adaptive threshold δ_n is determined not to miss local minima as

$$\delta_n = \delta + \text{median}\{d_{n-T}, \ldots, d_{n+T}\} \tag{5}$$

where T, the size of the median filter, is set to 40 in our experiments.

A fixed thresholding is also used to detect silences of the input signal. δ_c is a constant value separating silence from non-silence frames, and it is relevant to the signal-to-noise ratio of the signal. For general recordings, 20 percent of the maximum of d_n will be proper.

3.4. Picking onsets

Before picking onsets, a post-processing is performed to discard multiple false positive onsets adjacent to a true positive onset. Since the shortest note in our experiment data is longer than 15 ms, we dismiss multiple onsets whose duration are shorter than this interval.

Unlike many other approaches where the peaks in the detection function directly indicate onsets, we first compute the 'transient sections', which means the interval between different notes. Frames where the detection function d_n is greater than the adaptive threshold δ_n or less than the fixed threshold δ_c are classified as the transient section. Then, positive onsets are defined on the end of each transient section. Offsets can also be simply found (if needed) on the beginning of each transient section, as long as transient sections are well-defined.

Fig. 4 shows the graphical summary of procedures described in this section. An excerpt from a solo violin performance of "Ach Gott und Herr" from Bach10 database[1] [9] was used in this figure. Triangle markers in the first two plots indicate the ground-truth onsets and detected onsets, respectively. In plot (b), detected onsets are located at the end of each transient section which is depicted as a gray-shaded area. Solid line indicates a detection function d_n, dotted line an adaptive threshold, and dashed line a fixed threshold, respectively. The detection function in (b) is obtained by summing across five sub-harmonic cepstral coefficients which are shown in plot (d). We can observe in plot (b) that the detection function is mostly stable within a note regardless of the waveform amplitude, only exception being the fluctuation in the middle of the input signal due to the vibrato of violin. In plot (c) are illustrated five harmonic quefrencies that correspond to five cepstral coefficients shown in (d).

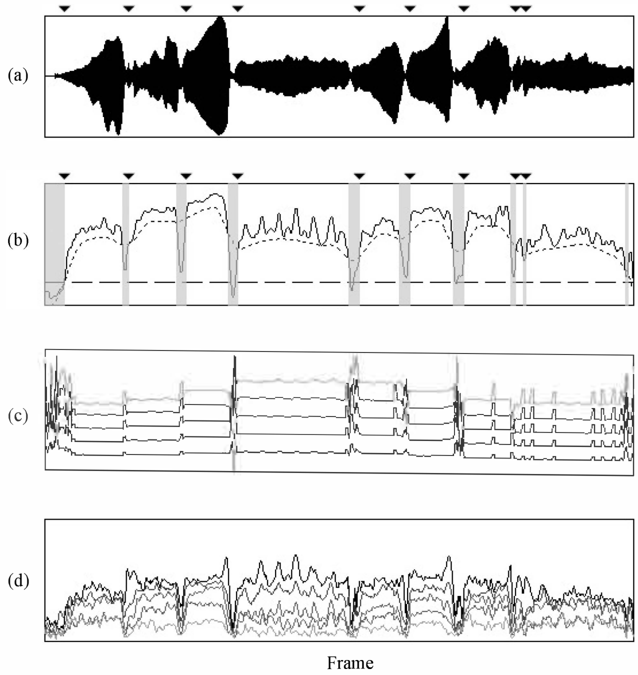

Fig. 4. (a) Waveform of a violin signal. (b) Detection function and adaptive threshold. (c) Five harmonic quefrencies. (d) Five sub-harmonic cepstral coefficients.

[1] Details of the database can be found online at URL http://music.cs.northwestern.edu/data/Bach10.html.

4. EXPERIMENTS

4.1. Database description

The experiments were performed on Bach10 database [9] and both commercial and noncommercial singing voice recordings, which contain 3474 onsets in total. The Bach10 database is accompanied by the ground-truth onsets. For singing voice, 13 male and 2 female recordings are used, which contain more than 1500 onsets. Onset labeling for singing voice recordings was cross-validated by three professional musicians. Ambiguous musical articulations such as glissando and non-pitched notes were excluded in the experiments. All data were preprocessed to be monaural signals sampled at 44.1 kHz. The detailed information of the data set is reported in Table 1.

Table 1. Data set details.

Class	Instrument	Reference	Duration	# of Onsets
Sustained Strings	Violin	Bach10	5m 34s	425
Woodwind	Saxophone	Bach10	5m 34s	500
	Clarinet	Bach10	5m 34s	475
	Bassoon	Bach10	5m 34s	507
Singing Voice	Male	Commercial Recordings	11m 46s	1533
	Female	Non-commercial Recordings	24s	34

4.2. Evaluation results

Like many other approaches, we regarded an onset to be correctly detected (CD) if ground-truth and detected onsets are within a 50-ms interval. Because of inaccuracy found on the annotation of the Bach10 database, a 70-ms tolerance window was used instead for some clips. We applied the same tolerance to all the comparison groups. If a detected onset is not within this interval, it is regarded as a false positive (FP). If a ground-truth onset is not detected (i.e., missing onset), there is a false negative (FN). Precision (P), Recall (R), and the F-measure (F) are used to evaluate the performance. These measures are defined as follows:

$$P = \frac{CD}{CD + FP} \tag{6}$$

$$R = \frac{CD}{CD + FN} \tag{7}$$

$$F = \frac{2PR}{P + R} \tag{8}$$

Table 2. Performance of the proposed algorithm.

Instrument	# of Onsets	Precision (%)	Recall (%)	F-measure (%)
Violin	425	87.4	94.9	91.0
Saxophone	500	93.4	96.0	94.7
Clarinet	475	87.8	93.3	90.4
Bassoon	507	80.9	82.9	81.9
Singing Voice(male)	1533	71.3	73.7	72.5
Singing Voice(female)	34	78.4	82.9	80.6
Total	3474	80.2	83.6	81.9

The overall results of our algorithm are summarized in Table 2. We can see that there is no large difference in performances between soft and hard onsets. Particularly for singing voice, although not directly comparable, F-measure was significantly improved by about 30% over the best performing algorithm of the MIREX 2012.

For comparison with other approaches, we implemented several algorithms aforementioned in Section 1 [1, 3] plus energy-based method, which was first introduced by Schloss [10]. Klapuri's psychoacoustic knowledge-based approach [2] was implemented based on MIR toolbox 1.4 [11]. Parameters for adaptive thresholding and peak-picking were fixed to the same values in every experiment.

We classified all recordings into the soft onset class and the hard onset class by the instrument type. Violin and singing voice are categorized into the typical soft onset class. All other instruments were classified as the hard onset class. As depicted in Fig. 5, HCR shows the remarkable improvement for the soft onset class. While an F-measure of other algorithms is below 50%, HCR achieves an F-measure of 76.7%. Considering that every algorithm yields a good performance for the hard onset class, it is obvious that the performance for the soft onset class makes the overall improvement.

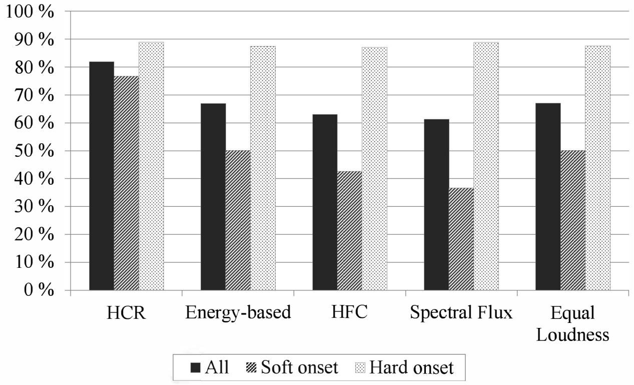

Fig. 5. F-measure comparison for different classes of onset.

5. CONCLUSIONS

In this paper, we have proposed an automatic note onset detection algorithm for pitched instruments including singing voice signals. The presented algorithm is simple and yet achieves a significant improvement, especially for soft onsets. Using the cepstral analysis, sub-harmonic regularity functions were derived from the changes in harmonic cepstral coefficients. The experiments were performed on over about 3500 onsets on multi-instrument recordings from the Bach10 database and 15 singing voice recordings. The results showed that the proposed algorithm not only achieved performance comparable to other conventional algorithms for hard onsets, but also outperformed significantly for soft onsets.

Since the proposed algorithm is able to locate the transient sections whose beginning and end position indicate note offset/onset, respectively, and also to find the fundamental quefrency which is related to pitch, it has a potential to be applicable to an integrated automatic music transcription system. Future research will cover offset detection at the beginning of the transient section we already obtained. We also plan to extend it for the polyphonic music transcription.

6. REFERENCES

[1] P. Masri, *Computer modeling of sound for transformation and synthesis of musical signal*, Ph.D. thesis, Univ. of Bristol, Bristol, U.K., 1996.

[2] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP-99)*, Phoenix, AZ, 1999, pp. 115–118.

[3] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection," in *Proc. Digital Audio Effects Conf. (DAFX,02)*, Hamburg, Germany, 2002, pp. 33–38.

[4] C. C. Toh, B. Zhang, and Y. Wang, "Multiple-feature fusion based onset detection for solo singing voice," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Kobe, Japan, 2009.

[5] R. E. Berg and D. G. Stork, *The physics of sound*, Upper Saddle River, NJ: Prentice Hall, 3rd edition, 2005.

[6] A. de Cheveign and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[7] J. W. Xu and J. C. Principe, "A pitch detector based on a generalized correlation function," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 8, pp. 1420–1432, Aug. 2008.

[8] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in musical signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 13, no. 5, pp. 1035–1047, 2005.

[9] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 8, pp. 2121–2133, 2010.

[10] A. W. Schloss, *On the automatic transcription of percussive musicfrom acoustic signal to high-level analysis*, Ph.D. thesis, Dept. Hearing and Speech, Stanford Univ., Stanford, CA, 1985, Tech. Rep. STAN-M-27.

[11] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *Proc. Digital Audio Effects Conf. (DAFX,07)*, Bordeaux, 2007.