DOI: 10.61782/fa.2025.0588

FORUM ACUSTICUM
EURONOISE 2025

MULTI-CHANNEL PROCESSING FOR PATIENT IDENTIFICATION AND AUDIO ANALYSIS IN THE CONTEXT OF GROUP MUSIC THERAPY

Arina Epure[1*] **Thomas Dietzen**[2] **Toon van Waterschoot**[1]

[1] Department of Electrical Engineering, ESAT-STADIUS, KU Leuven, Leuven, Belgium
[2]KU Leuven, ESAT-PSI, EAVISE, Sint-Katelijne-Waver, Belgium

ABSTRACT

Active group music therapy is an effective way to treat patients with several psychopathologies. However, it can be challenging to extract quantitative data from music played by patients which are sensitive to external factors that may have an impact on the set therapeutic goals. For this reason, the music played by a group of patients and therapists was recorded using a circular microphone array placed at the centre of the room. This data will serve for music signal analysis in order to track various musical features for individual patients, such as the notes or instruments each patient was playing. In this paper, we explore methods to obtain a sufficient separation of the music played by each patient and extract note onsets. The scope of the analysis is build an algorithm that will eventually be used to identify patterns and follow the progress of each patient during the therapy, alongside qualitative data. The preliminary results presented in this paper show that the reverberation time has no impact on the onset detection for the actual room measurements. For a small dataset, the beamformer should be able to identify a source despite a low accuracy of the steering towards the source. Finally, for multiple instruments played at the same time, the algorithm pipeline requires fine-tuning for a better separation and patient identification.

Keywords: *multi-channel music analysis, feature extraction, source identification*

1. INTRODUCTION

There are several types of music therapies that were found very effective in promoting social-emotional interactions of patients with different psychopathlogies such as depression, borderline personality disorder, dementia [1], or for patients who spend extensive time in hospital care. In this research project, the target patients are adults with autism spectrum disorder that have developed compensating mechanisms. The therapy is active, meaning that the patients play musical instruments themselves, and, in this case, in a group of four or five people, not including the therapist(s). During one session, the participants play music and then discuss, repeatedly. Thus, the therapy creates an environment in which communication is firstly promoted through music before it is done verbally. The music played by patients can transmit relevant information embedded in these short improvisations, which is part of the analysis done by a therapist during and after each session. The focus of this research is to facilitate the analysis of the music improvisations.

In the past, some tools were developed to aid the therapists in analysing the music created [2–5]. However, for the target group involved in this project, these tools are not sufficient for the following reasons. First, the amount of people playing at the same time was not previously considered, thus, there is a need for separating what each patient is playing. Second, often the therapists need to use MIDI instruments or find a way of separating what the patient plays (such as selecting some instruments only or dividing a piano keyboard, for example). This limits the patients and affects the therapy progress. Third, the recording methods need to be less intrusive than placing microphones in front of the patients due to their higher

*__Corresponding author__: arina.epure@esat.kuleuven.be.

sensitivity to external factors so, separation of the music played by patients has to be adapted. Forth, during a session, the patients are free to move around the room and choose which instruments they want to play and/or place it in a different location. The combination of all four limitations in this project requires an alternative to perform the analysis needed by the therapists.

In order to overcome the challenges imposed by the set therapeutic goals for the target group, the solution was to use a circular array [6] to record data, placed at the centre of the room. This alternative, provides a less intrusive method of collecting data and, as opposed to the studies mentioned previously [2–5], allows for the therapy to proceed without limitations and without affecting the therapeutic process.

However, the alternative used to collect the data, has certain challenges in terms of audio processing. One of the main expected issues is that the musical instruments are played by different patients at the same time, with overlapping note onsets, which could bring difficulty in identifying the two notes separately. For this reason, it is crucial to exploit the spatial location of each sound source. To this end, beamforming is applied to the multi-channel recordings. The intention is to achieve sufficient separation so that the note onsets played by an individual instrument can be detected and assigned to a particular patient. For this purpose, a complete source signal separation may not be necessary as long as it is possible to extract onset information from the sound coming from one pre-selected direction.

In this paper, only simulated data is presented, aiming at building a base for future analysis of recorded music therapy sessions. The simulations consist of three main parts divided the following way: room modelling and acoustic parameters, beamforming, and onset detection.

2. METHODOLOGY

The methodology is divided the following way: section 2.1 describes the modelling process of the room and the chosen acoustic parameters, section 2.2 explains the signal model and what type of beamforer was used. Section 2.3 is dedicated to onset detection. These three sections describe a pipeline of algorithms that eventually allow onset detection analysis to be performed. The topics investigated in this paper are the effect of reverberation time, beamforming implementation and source differentiation on onset detection.

2.1 Room acoustics parameters for simulations

In order to generate audio data that reflects a realistic scenario of musical instruments being played in a room, the environment itself was defined and used to create room impulse responses (RIRs) paths from sources to each microphone. The method used for the generation of RIRs was the randomized image method (RIM) as described by [7]. The acoustics of the room were based on previous work presented in [6]. The room dimensions are 8x6.6x5.4m giving a total volume of approximately $285m^3$. The average reverberation time (RT) measured previously is 0.8s over 125-4000Hz octave bands.

However, the RT of the room simulations was modified in order to investigate its effect on the signal analysis done. By recreating the environment, a ground truth could be set for future analysis of the recorded music, as well as open the possibilities to create any desired space in which the music therapy could take place.

The audio data for the simulations was created using dry audio recordings of singular notes of different musical instruments from the McGill University Master Samples [8], together with the MixNotes algorithm presented in [9] which generates audio files by mixing individual recordings of notes (either separated or overlapping). At the same time, the MixNotes algorithm generates a file containing the timestamps of the note onsets. The method used bypasses the need for manually labelling the onsets of audio data, and it was considered the ground truth to test the onset detection algorithms.

2.1.1 The noise source

In the scenario where additional sources were used, the signal received by each microphone is a mixture of the desired source and interference, $x(n) = s(n) + v(n)$. In this case, the other instruments played are considered noise sources or undesired signals for a particular direction of the beamformer. As baseline, no noise was added to the signal received by the microphones. When a noise source was introduced, a similar type of signal was used. Another audio file containing three notes was created with MixNotes and it was combined with the desired signal. The choice of noise source was intentionally selected to be correlated with the signal $s(n)$ because it is considered closest to the real scenario of having multiple instruments played at the same time.

2.2 Signal model and beamforming algorithm

A circular microphone array consisting of 12 omnidirectional sensors was used to record music therapy sessions, with a radius of 14cm. The design of the array was previously discussed in [6]. Thus, in the simulations done using Matlab, an equivalent array was reproduced to test the beamforming algorithm. The signal received by each microphone is defined as [10, 11]

$$x_m(n) = \sum_{i=0}^{L} h_{m,i}(n)s(n-i)+v_m(n) = s_m(n)+v_m(n), \tag{1}$$

where m takes values from 1 to the number of microphones, M, n is the time index of the signals and L is the length of the RIRs generated from a source to each microphone, $h_{m,i}$. This is equivalent to the two components $s_m(n)$, the source of interest, and $v_m(n)$, the noise sources as described above.

The short-time Fourier transform (STFT) is applied to the microphone signal giving the stacked vector $y(l) \in C$

$$y(l) = (y_1(l)...y_M(l))^T, \tag{2}$$

where l is the frame number and each signal is composed of

$$y(l) = x(l) + v(l), \tag{3}$$

the frequency domain equivalents of the source $s_m(n)$, and noise $v_m(n)$ components.

The generated signals, $y_m(l)$, were then used as input to a delay and sum (DAS) beamforming algorithm. For this type of beamformer, the signals received by each microphone are delayed in order to compensate for the time difference of arrival of the sound to the sensor. This is equivalent to a phase shift, $w(\omega)$, with ω the angular frequency, and as a function of the angle of incidence, θ.

$$w(\omega) = \frac{1}{M}\begin{bmatrix} 1 \\ e^{-j\frac{\omega d\cos\theta_0}{c}} \\ \vdots \\ e^{-j\frac{\omega(M-1)d\cos\theta_0}{c}} \end{bmatrix} \tag{4}$$

where θ_0 is the angle to which the beamformer is steered, d is the distance between two microphones, and c is the speed of sound, $340m/s$.

The beampattern is then given by

$$\Psi(\omega,\theta) = \frac{1}{M}\sum_{m=0}^{M-1} e^{-j\omega\frac{md(cos\theta - cos\theta_0)}{c}}. \tag{5}$$

The DAS beamformer was steered towards the chosen position of the source (180 degrees), giving the beampattern shown in Fig. 1. The frequency bins are created by dividing the sampling frequency (fs) equal to 48000Hz by 256. Since this is the beampattern of a circular array, the angles over which the analysis are done are from 0 to 360 degrees.

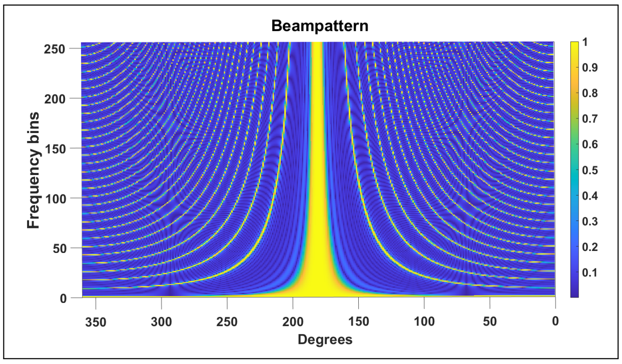

Figure 1. Beampattern steered to 180 degrees w.r.t. the reference microphone

Optionally, the input signals, $s_m(n)$ and $v_m(n)$, where $m = 1...M$, could be modified with a secondary source interference function that adjusts the levels ratio between two sources.

2.3 Onset detection algorithm

The onset detection problem is usually divided into three steps: pre-processing, defining an onset detection function (ODF) and peak-picking. In general, the pre-processing step involves noise reduction or beamforming. In the second step, a reduction of the signal is obtained by defining the best ODF for the type of audio processed. For this reason, several ODFs were tested. As shown in [12], some algorithms perform better when applied to percussive instruments, and these were selected and tested on the same dry recordings and on real data. These analysis were done without applying the beamforming algorithm beforehand, in order to assess the level of accuracy in this scenario. The methods tested were two variations of the spectral flux [13, 14], two variations of the logarithmic spectral flux, high frequency content and NINOS2, as described in [9, 12]. The results showed that the best method to use further is the logarithmic spectral flux (LSF) [15]. This method is based on the spectral flux ODF, which is defined by Eqn. (6), with L_2-norm for calculating the difference between magnitude spectra of adjacent samples:

$$SF(n) = \sum_{-\frac{N}{2}}^{\frac{N}{2}-1} H(|X_k(n)| - |X_k(n-1)|)^2 \qquad (6)$$

where $H(x) = (x + |x|)/2$ is a half-wave rectifier for the spectral difference, which outputs zero for a negative argument, and $X_k(n)$ is the N-point short-time Fourier transform (STFT) of the audio signal at frequency k and frame index n of the windowed signal in time domain, x(n),

$$X_k(n) = \sum_{i=-\frac{N}{2}}^{\frac{N}{2}-1} w(i)x(n+i)e^{\frac{-2j\pi ik}{N}}, \qquad (7)$$

and $w(i)$ is a window function of choice of length N samples and i is the index of the window.

Similar to [15], the ODF was made invariant to scaling by applying a logarithm to the magnitude spectrum,

$$Y_k(n) = log(\lambda|X_k(n)| + 1), \qquad (8)$$

where λ is a compression variable [9], here 0.5, and plus 1 is added to ensure a positive value of the logarithmic function. Therefore, the ODF defined by the LSF method is:

$$LSF(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} H(|Y_k(n)| - |Y_k(n-1)|)^2. \qquad (9)$$

In the third step of the onset detection algorithm, the peaks of the ODF are selected. In order to obtain relevant results, a few parameters were included to define a threshold of detection. First, the minimum height of a peak was calculated by taking the median of the ODF function and multiplying it by a constant [12] of choice. The minimum height of the peak is generally set above the noise floor, and can be set to adapt to the local median, rather than the entirety of the ODF. Here, the value of the constant is 2.

The second parameter adjusted was the minimum peak prominence which the defined as the height of the amplitude in its neighbourhood. This value has an impact on detecting only the first peak of a note and ignore the decay of it. The value for the minimum peak prominence was set to 0.2. In this case, the minimum peak height is superseded by the minimum peak prominence due to the small variation in the noise floor.

The third parameter defines how close two peaks can be to each other, called minimum peak distance, and was set to 0 because the algorithm may be used in identifying small deviations in time of two separate sources. All three parameters were adjusted for the specific data set that was used to test the onset detection algorithm and were kept the same for all results presented here.

Only after identifying the best ODF to be used further, the pre-processing step was included, which was the beamforming algorithm. Additional pre-processing may be considered for analysis done on recorded music therapy data to reduce noise, if found necessary.

3. RESULTS

In this section, the results of three case studies are presented. In 3.1 is shown the effect of RT on onset detection, in 3.2 a source is placed off-axis w.r.t. the DOA of a steered beamformer to measure the level of accuracy needed for the real case scenario, and in 3.3 are investigated multiple secondary source interference ratios (SIR) and their impact on the accuracy of notes detection.

3.1 Reverberation time effect on onset detection

Firstly, the effect of reverberation time on the accuracy in detecting onsets was investigated. The RTs generated in the simulations of RIRs were: 0s (anechoic conditions), 0.3s, 0.5s, 0.8s (closest to actual room acoustic measurements) and an extreme case of 1.7s. The chosen onset detection algorithm was used in the described simulations after applying beamforming. The source was initially placed at 180 degrees, 1 meter away from the centre of the array (see Fig. 2).

In Fig. 3 are shown the results for a direction of arrival (DOA) of 180 degrees for the RTs mentioned above. From this figure it can be concluded that in the simulated conditions, the RT does not have an impact on the onset detection algorithm until the extreme case of high RT where one false positive peak was be detected.

Another study was done by moving the source at 1.75m away from the centre of the array. This distance is beyond the critical distance of all RTs except anechoic conditions. A similar behaviour to the previous study is

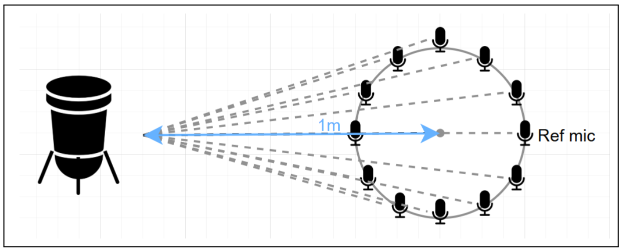

Figure 2. Study of RT values on onset detection performance for a source 1m away and DOA 180 degrees

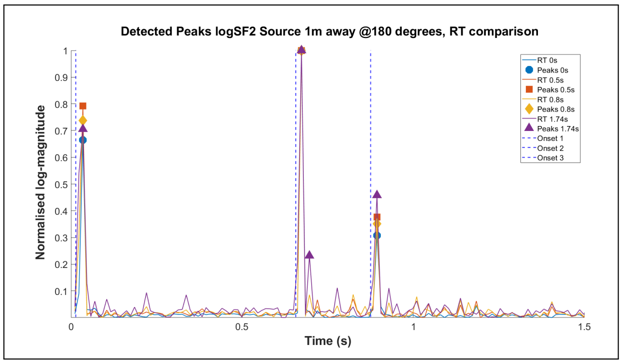

Figure 3. Study of RT values effect on onset detection performance for a source at 180 degrees, 1m away

also observed in Fig. 4. Some additional high peaks appear in this case, however, they are below the detection threshold, and only one false positive onset appears for RT=1.7s.

3.2 Off-axis source placement effect on onset detection

A source was placed at 180 degrees and the beamformer was steered to different angles to study the effect on the onset detection output. In Fig. 5 are shown the results for an RT closest to the actual parameters of the room (0.8s) and they show consistent results in terms of correct detection. It can be noticed that as the DOA mismatch increases, additional smaller peaks become apparent, however, they are below the detection threshold.

It should be noted that further analysis will be done on a more extensive data set containing a wider range of musical notes in order to examine the behaviour of the beampattern at different frequencies.

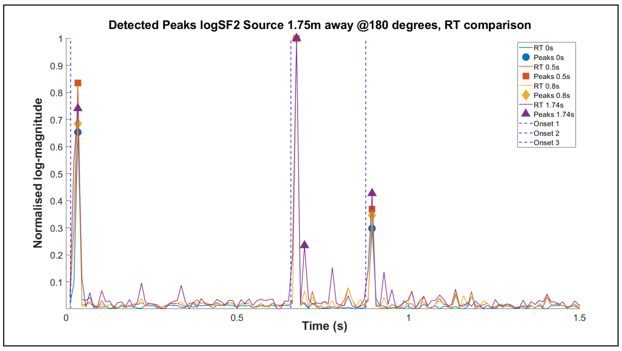

Figure 4. Study of RT values on onset detection performance for a source 1.75m away and DOA 180 degrees

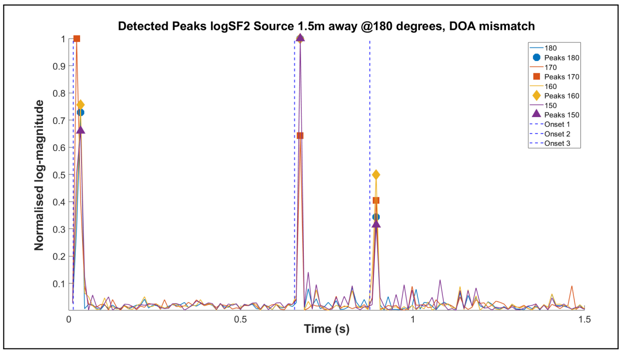

Figure 5. Study of DOA mismatch effect on onset detection

3.3 Signal to noise ratio effect on onset detection

A secondary source was added as described in 2.1.1. The sources were placed 1m away at 180 degrees and 90 degrees. This study was done in order to investigate the SIR necessary for an acceptable performance of the onset detection. In other words, what is the effect of the difference in levels between two sources.

Firstly, the general effect of adding beamforming in the pre-processing step is compared to the signal received at microphone 7, the closest to the primary source at 180 degrees. The results are shown in Fig. 6 and it can be concluded that this is a necessary step in improving the output of the onset detection algorithm.

In Fig. 7 can be observed that an SIR of 20dB would be necessary in order to have a perfect result in the onset detection algorithm. All other values of the SIR detect one onset from the source place at 90 degrees. For an SIR of

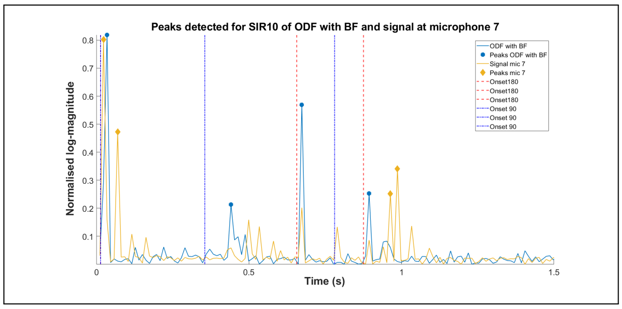

Figure 6. Comparison of ODF with beamforming and the signal received at microphone 7

6dB, there is no detection of the third note, but only one false positive peak is detected. For an SIR of 10dB, all peaks are detected, but also one false positive as well. This implies that issues may appear when the desired source is not loud enough compared to the noise sources. Also, when two impulses overlap (the first peak in Fig. 7, there is no distinction between them as they combine into one higher peak.

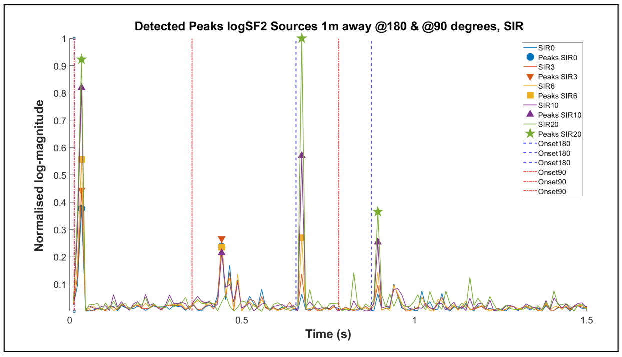

Figure 7. Study of SIR values on onset detection with two sources 1m away at 180 and 90 degrees

4. CONCLUSIONS AND FURTHER RESEARCH

In this paper, preliminary results were shown for a pipeline of algorithms that were built to solve onset detection in the context of group music therapy. An alternative to existing tools for analysis of music improvisations had to be developed due to the particularity of the research project. This, however, increases the difficulty of signal processing required to extract relevant information for therapists.

From the results presented, it can be concluded that RT is not a parameter that impacts the final results, even in the case where the source is placed beyond the critical distance. From the study on the DOA mismatch with respect to the beamformer steering angle, no notable changes have been observed on the small data set used. Thus, further analysis are crucial.

In the final study, two competing sources were added to investigate the achievable accuracy when they are equidistantly placed from the centre of the microphone array, but at different, well-separated angles. It was concluded that an SIR of 10dB is able to detect all notes, but also one false positive from the secondary source. For an SIR below 10dB, the third note is no longer detected. Thus, this is considered a minimum limit of SIR, unless other pre-processing steps are added which could improve the reduction of the secondary source. One option would be to improve on the current design of the beamformer by further enhancing the amplitude of the notes.

Based on the conclusions mentioned above, further research is necessary on several aspects. Firstly, the beampattern should be improved to provide a better separation of the notes. One option to explore would be to deduce the direction of a source from differences in the amplitude levels when steering the beamfomer towards each source.

Secondly, in the case where two notes are very close to each other, a better resolution should be achieved to avoid false negative peak identifications. This could be solved by implementing a pitch identification algorithm which would be able to differentiate between two different sources.

Additionally, it is desired to test the algorithm on extensive data to calculate the F1-score and compare with other relevant algorithms. Finally, the algorithms pipeline should be tested on recorded data to assess the real scenario performance.

5. ACKNOWLEDGMENTS

This research work was carried out in the frame of the FWO Research Project: ”UNMUTED: Understanding Music-Based Interventions to Encourage Social- Emotional Development in Groups of Adults with Autism Spectrum Disorder” (G0A2721N). The research leading to these results has received funding from the European Research Council under the European Union’s Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors’ views and the Union is not liable for any

use that may be made of the contained information.

6. REFERENCES

[1] T. Wigram and J. de Backer, *Clinical Applications of Music Therapy in Psychiatry*. Jessica Kingsley Publishers, 1999.

[2] E. Streeter, M. E. Davies, J. D. Reiss, A. Hunt, R. Caley, and C. Roberts, "Computer aided music therapy evaluation: Testing the music therapy logbook prototype 1 system," *The Arts in Psychotherpy*, vol. 39, pp. 1–10, 2012.

[3] A. Gilboa, "Testing the map: A graphic method for describing and analyzing music therapy sessions," *The Arts in Psychotherapy*, vol. 34, pp. 309–320, 2007.

[4] J. Erkkilä, E. Ala-Ruona, and O. Lartillot, "Technology and clinical improvisation – from production and playback to analysis and interpretation," *Music, Health, Technology and Design*, vol. 8, p. 209–225, 2014.

[5] A. Hunt, R. Kirk, M. Abbotson, and R. Abbotson, "Music therapy and electronic technology," in *Proceedings of the 26th Euromicro Conference. EUROMICRO 2000. Informatics: Inventing the Future*, vol. 2, pp. 362–367 vol.2, 2000.

[6] A. Epure, T. Dietzen, K. Foubert, J. de Backer, and T. van Waterschoot, "Room acoustic treatment and design of a recording setup for music therapy," in *Proc. of Forum Acusticum*, (Torino, Italy), 2023.

[7] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 774–786, 2015.

[8] F. Opolko and J. Wapnick, "Mcgill university master samples." 2006.

[9] M. Mounir, *Acoustic Event Detection: Feature, Evaluation and Dataset Design*. PhD thesis, KU Leuven, 2020.

[10] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, *Handbook on Array Processing and Sensor Networks*, ch. Acoustic beamforming for hearing aid applications. Wiley, 2010.

[11] G. Huang, J. Benesty, and J. Chen, "On the design of frequency-invariant beampatterns with uniform circular microphone arrays," vol. 25(5), p. 1140–1153, 2017.

[12] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," vol. 13, p. 1035–1047, Sep 2005.

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

[14] P. Masri, *Computer Modeling of Sound for Transformation and Synthesis of Musical Signal*. PhD thesis, Ph.D. dissertation, Univ. of Bristol, Bristol, UK, 1996.

[15] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *in Proc. 1999 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '99)*, vol. 6, p. 3089–3092, March 1999.