A MULTI-SENSOR APPROACH FOR REAL-TIME DETECTION AND CLASSIFICATION OF IMPACT SOUNDS

Nikolaos Stefanakis[*] *and Athanasios Mouchtaris*[*,†],

[*] Foundation of Research and Technology - Hellas,
Institute of Computer Science,
70013 Heraklion, Crete, Greece

[†]University of Crete,
Department of Computer Science,
71409 Heraklion, Crete, Greece

ABSTRACT

We present a method for real-time detection and classification of impact sounds — relying solely on spatial features— that exploits the difference in the location of each impacted structure. Using a compact sensor array, we formulate the classification problem in terms of an undetermined source separation process where we assume that the linear mixing model can be learned through a training phase. The recovered source amplitudes are exploited for estimating the source activity in time, and the detection and classification decisions are derived based on simple energy criteria. Experimental results with two sensors demonstrate the efficiency of the method in an application scenario which considers the use of a simple object as a real-time control interface for triggering a percussion synthesizer.

Index Terms— impact sound, sensor array, gesture recognition

1. INTRODUCTION

The sounds generated by humans are gaining more and more attention in the context of control interfaces and applications. Among the many different types of sonic gestures that we are able to perform, impact sounds represent excellent acoustic cues for conveying information. Moreover, our bodies and our physical environment provide uncountable ways for producing such sounds; we can clap our hands, snap our fingers and we may hit almost any solid object around us. Using appropriate sensors and machine learning techniques, we may then train a computer to recognize our gestures in order to perform predefined actions.

Undoubtedly, a great number of techniques which are required for the detection and classification of such gestures has evolved through the need to analyse and extract information from percussive sounds in audio signals. The transcription of drum sounds in musical pieces can be seen as a low level descriptor of musical content, which can then be exploited for tasks related to audio queries and content based management systems (see [1] and references therein). Nevertheless, most of the applications in this category do not obey to the time constraints that Human-Computer Interaction (HCI) systems are obliged to meet. Interactive systems using impact sound as the input information have been developed in both musical and non-musical context. For example, a system able to track percussion gestures was presented in [2], showing a good adaptability to different instruments and acoustic conditions, while other examples include an automatic accompaniment system [3] and a rhythmic tutoring system [4]. On the other hand, Jylhä and Erkut developed a hand clap interface for sonic interactions with the computer [5], while Vesa and Lokki presented an interface which utilized two microphones integrated to the headphones of the user and was capable of detecting finger-snaps occurring on the left or right side of the users head or in front of it [6].

Detection and classification of percussion sounds may be also seen as a by-product of source separation, and several works have considered the case of drum sounds separately from other musical sources. Fitzgerald in [7] used the DUET algorithm to separate percussive sounds from a stereo track while Battenberg et al. used probabilistic spectral clustering to separate drum sounds from mono recordings [8]. Finally, "Drumatom" is a newly released commercial product capable of separating percussion sources in multi-track recordings (http://www.accusonus.com).

Gestural control of sound synthesis [9] is an additional field of research which benefits from human-generated sound input. Here, the user's gestures need to be recognized and automatically encoded to the control stream which is required to trigger a sound synthesizer. The synthetic signal which is produced may then be used in order to augment or to completely replace the physical sound of the object(s) being hit. In the context of percussion gestures, two commercial products which can be found are "Mogees" (http://mogees.co.uk) and "TableDrum" (by Dohi Entertainment), both of which exploit timbral features of the single input signal to associate different gestures to different synthetic sounds. In a similar way, Stefanakis et al., exploited the variability in the spectra of the same object being hit at different locations in order to

This research has been co-financed by the European Union and Greek national funds through the National Strategic Reference Framework (NSRF), Research Funding Program: "Cooperation-2011", Project "SeNSE".

recognize different events [10]. We note however that none of these approaches is able to handle simultaneous events, which is often the case in real percussion performance.

In spite of the many different applications that may benefit from the work presented in this paper, we also consider the scenario of a real-time sound synthesis controller, which is a particularly time-critical application among other HCI scenarios. Indeed, as Wessel and Wright state, the time between a gesture and its computer generated audible reaction should be below 10 milliseconds [11], which poses strict limitations to the amount of information that can be extracted from the sonic gesture before deriving a decision. Similar to [5], we present an approach based on multiple sensors rather than just a single sensor, and we rely solely on spatial features for discriminating among the different events. Moreover, we evaluate the classification performance under realistic playing conditions, including densely-played and simultaneous events. Since the observed sound signals are expected to overlap in time, we propose a simple source separation approach for assisting the classification process, relying on a linear mixture model which is learned *a-priori* during a short training phase. Experimental results are presented for the case of a wooden table which is stroked by the user with the help of two metallic rods.

2. METHODOLOGY

Assume an array with M sensors and the problem of detecting and classifying N sources of impact sound. At each time instant, the signal at the microphones will be a mixture of all the active source sound signals convolved with the impulse response of the room. Following [12], the mixing process may be approximated in the time-frequency domain (TF) as

$$\mathbf{X}(k,\tau) = \mathbf{H}(k)\mathbf{S}(k,\tau), \tag{1}$$

where τ is the time frame index, k is the frequency bin, $\mathbf{X} = [X_1, ..., X_M]^T \in \mathbb{C}^M$ is the signal at the microphones, $\mathbf{H} \in \mathbb{C}^{M\times N}$ is the mixing matrix and $\mathbf{S} = [S_1, ..., S_N]^T \in \mathbb{C}^N$ are the source amplitudes. Typically, the length of the analysis window used for the TF implementation should be long enough in order to account for the reflections and reverberation which are introduced by the acoustic environment. In this paper, we consider a similar model for the mixing process based on a shorter frame size as

$$\mathbf{X}(k,\tau_o) = \mathbf{H}_o(k)\mathbf{S}(k,\tau_o), \tag{2}$$

where τ_o denotes a time frame where an onset occurs and $\mathbf{H}_o = [\mathbf{h}_{o,1}, ..., \mathbf{h}_{o,N}]$ is different from $\mathbf{H}$ in (1) in the sense that it contains the mixing model characteristics corresponding to a very short duration after the onset of a particular event. Certainly, by doing so, important information might be disregarded, but this is something that we may afford to do because we are interested in classifying the events and not in separating them. The truncated mixing model $\mathbf{H}_o$ is estimated by using an onset detection algorithm, as described in the next section.

2.1. Mixing Model Estimation

We exploit the main assumption that both the impacted structures, the sensors and their locations are consistent in time and therefore, the “spatial signature” of each impact sound source can be estimated *a-priori* during a training phase where the user is asked to excite a particular object (or impact region) several times, ensuring that there is no overlap between successive strikes. This ensures that the mixing vector characteristics $\mathbf{h}_{o,q}$ contain the direct path (and possibly some first reflections) from the objects to the sensors but no latter part of the impulse response.

Now, there are numerous methods for onset detection which are suitable for musical signals and for the needs of this paper we actually used the energy-based approach proposed by Tan et al. [13]. We note that choice of the onset detection method plays a trivial role here, as it is only required during the training phase. Detecting onsets from a recording that contains clean transient signals shouldn’t be a problem for any kind of onset detection algorithm.

Having detected the onset locations, the mixing model is estimated by performing a phase- and amplitude-normalization process well known in clustering approaches in blind source separation [12]; the observation mixture of the jth training instance is stored to the corresponding source collection as

$$\mathbf{y}_j^{(q)} \leftarrow \frac{\mathbf{X}}{\|\mathbf{X}\|_2} e^{-\phi_{x_1}} \tag{3}$$

where we have omitted time and frequency dependency for convenience. Here, $\|\cdot\|_2$ represents the Euclidean norm and ϕ_{x_1} is the phase at the first microphone. Now, assuming that all J samples originate from the same source, the observations should cluster around the mixing vector $\mathbf{h}_{o,q}$. The first one would thus think is to average over all samples in the same class in order to define one centroid in $\mathbb{C}^M$. However, we have observed that in several cases there is significant variance in the samples and the centroids may be inaccurately estimated due to the presence of outliers. We therefore follow a simple approach for removing such outliers. Recalling that the J available samples are normalized, we define for each training instance the *consistency measure*

$$C_j = \frac{1}{J-1} \sum_{j' \neq j} \left| \left\langle \mathbf{y}_j^{(q)}, \mathbf{y}_{j'}^{(q)} \right\rangle \right|, \tag{4}$$

where $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^H \mathbf{b}$ denotes the dot product. Obviously, $C_j \leq 1, \forall j$ with the equality to 1 holding when all samples are identical. Assuming that most of the samples have small distances from one another, outliers will appear with a small value of consistency. We eliminate such outliers using an iterative procedure as follows. We calculate $C_j(k)$ for all $j = 1, ..., J$. If all consistency values are above a given threshold $C_{min} \in (0, 1)$, we then stop and calculate the centroid as the average of all samples. If not, we then find the index j with the smallest consistency value, remove it from the collection, update $J \leftarrow J - 1$ and recalculate (4). The procedure stops when all consistency values are above the given threshold or when a minimum number of samples J_0 is reached. The mixing vector particular to the qth class is then calculated as $\mathbf{h}_{o,q} = \frac{1}{J}\sum_{j=1}^{J} \mathbf{y}_j^q$.

2.2. Source Amplitude and Source Activity Estimation

Assuming that the mixing vectors $\mathbf{h}_{o,q}$, $q = 1, ..., N$ have been estimated from the previous step, we propose recovering the source amplitudes using the W-disjoint Orthogonality (WDO) assumption which is commonly exploited in source separation and direction-of-arrival estimation problems [14, 15]. We basically assume that one impact sound source is always dominant against the others at each time-frequency point. The dominant source is found by searching for the mixing vector which exhibits the highest correlation with the observed mixture, e.g.,

$$q = \arg\max_q |\langle \mathbf{h}_{o,q}, \mathbf{X}\rangle|\,, q = 1, ..., N. \quad (5)$$

The qth source amplitude is then determined as $S_q = \mathbf{h}_{o,q}^H \mathbf{X}$ while for the other sources the amplitudes are set to zero. We recognize here that the WDO assumption is not justified for impact sounds. Considering their wideband characteristics, it is very likely that the source amplitudes will not have a disjoint support in the frequency domain, especially in the case of two simultaneous events. Nevertheless, this approach is attractive because of its simplicity and its low computational complexity.

Now, having separated the sources, the goal is to construct an one-dimensional time-sequence, one for each source, which describes each source's presence in the mixture at each time frame. We observed that a good measure for such a task is the L1 norm of the vector containing the recovered source amplitudes across a wide frequency range. We therefore define the *source activity sequence*, p_q, which for the qth source at the τth time frame can be calculated as

$$p_q(\tau) = \sum_{k=k_{min}}^{k_{max}} |S_q(k, \tau)|\,, \quad (6)$$

where k_{min} and k_{max} represent minimum and maximum frequency indexes. In comparison to the L2 norm, the L1 norm here favours sources that have a rich support in their recovered amplitude vector, rather than large amplitudes at few non-zero entries. Also, it is advantageous to disregard low frequencies, as in these frequencies the acoustic modes tend to decay slower and this may increase the overlap in the activity sequences of successive events.

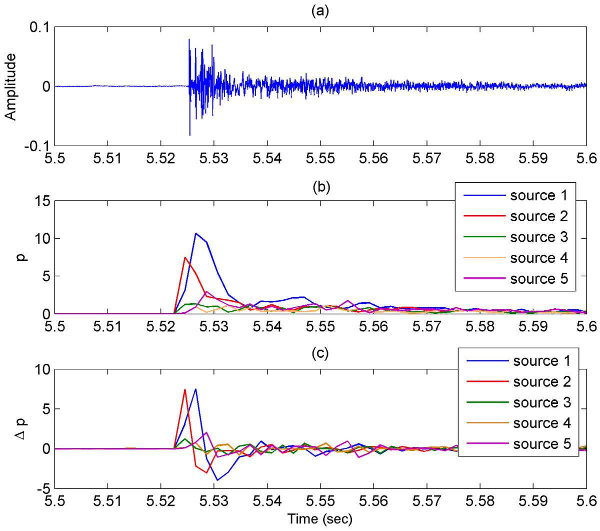

Fig. 1. Signal at one of the sensors for a simultaneous event in (a), source activity sequence in (b) and corresponding source detection values in (c).

2.3. Source Detection and Labelling

Using the measure in (6), the problem can be transformed to a binary classification task, where the presence or absence of each source may be determined at each time frame according to predefined activity thresholds. Ideally, the source activity sequence should approximate a Dirac-delta function at the onset of a particular event. In this paper, we also exploit the first-order time-difference of the source activity sequence, $\Delta p_q(\tau) = p_q(\tau) - p_q(\tau - 1)$, which we call the *detection sequence* from now on. A necessary condition to have an onset for the qth class at time τ is that the conditions $\Delta p_q(\tau) > T_{\Delta p}$ and $p_q(\tau) > T_p$ are fulfilled simultaneously, where T_{Δ_p} and T_p are empirically defined positive thresholds.

Now, some practical constraints may be set which not only improve the performance, but are also in accordance to physical limitations related to human hearing and to human percussive performance. First of all, at time τ, at most one source may be detected but not a second one. While this seems to be preventing the ability to detect two simultaneous events, tests performed with real recordings indicated that even if the intention of the user was to perform two simultaneous strikes (one with each hand), the detection sequences $\Delta p_q(\tau)$ rarely overlapped. In Figure 1 we show a typical example for the case of 5 sources and 2 microphones; while sources 1 and 2 are excited “simultaneously”, the p_q and Δp_q values are well discriminated in time in (b) and (c) respectively.

Additional temporal and amplitude constraints were found useful for allowing the system to respond to a wide dynamic range and for preventing false-detections following a strong attack. First, two successive onsets may be arbitrarily close in time, but onsets from the same class must be at least $T_{\Delta\tau}$ analysis frames apart, with $T_{\Delta\tau}$ a positive integer. This is helpful in order to avoid a "double onset" due to ambiguities in the neighbourhood of a strong attack. For similar reasons, a time-varying masking threshold is defined; assume that source q is the last one to be detected at time τ_q and let $p_q(\tau_q)$ be its activity value. We define the exponentially decaying threshold $g(\tau) = ap_q(\tau_q)e^{\sigma(\tau-\tau_q)}$ with $0 \leq a \leq 1$ and $\sigma < 0$ which sets a new lower bound for any upcoming event; at any time instant, the source activity value must be greater than both $g(\tau)$ and T_p in order to admit an onset. Potentially, this may lead to missing a weak strike following a strong attack, but this is expected to have small perceptual significance due to temporal masking phenomena of human hearing.

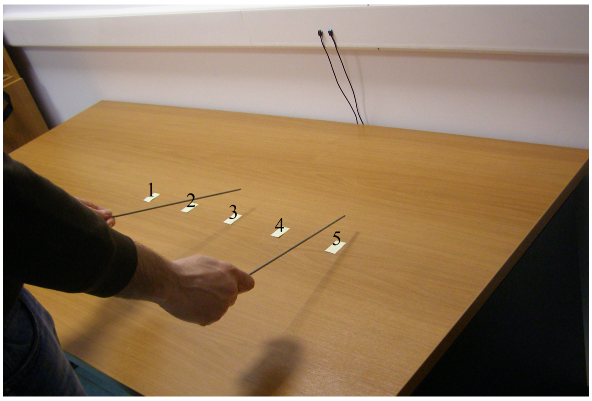

Fig. 2. Experimental setup showing the five impact regions and the two microphones.

3. EXPERIMENTAL RESULTS

Experimental results are presented for the case of five classes and two sensors in an application scenario which considers a real-time percussion controller. In this experiment there was a single object, a wooden table, and the user was able to strike it at 5 different impact regions by holding one metallic rod in each of his hands, as shown in Fig. 2. Five small pieces of yellow paper were placed as visual landmarks just below each impact region. Two omnidirectional microphones (Shure 93) were glued with blue-tack at the closest wall in front of the table. The distance between the two sensors was 2.8 cm while the distance between adjacent impact regions was 14 cm. As seen from the center of the two sensors, the impact regions spanned an angle from -40 to 40 degrees and the distance from impact region 3 was approximately 40 cm. The recordings took place in a small office with an estimated reverberation time of 0.4 seconds.

During both the training and the testing phase, the acoustic events were recorded with a sampling frequency of 44.1 kHz. About 20 strikes per impact region were used for training. For onset detection, we used a running Hanning window of 128 samples length and 50% overlap. On the other hand, the length of the analysis window which was used for obtaining the source spatial images in (3) was 180 samples. The consistency threshold C_{min} was set to the value of 0.94 and the mixing vectors for each impact region were calculated according to the analysis in subsection 2.1.

We observed that the classification performance depends a lot on the rhythmic context. For example, isolated and sparsely-played events have less chances of not being detected or being miss-classified, as there is little overlap between consecutive events. In an attempt to test more realistic conditions, we decided to perform two separate tests; one scenario where two different impact regions are excited alternately and one where two impact regions are excited simultaneously. In both cases, each region is excited with the same rod and the events are played close in time so that there is non-trivial overlap between consecutive strikes. We recorded phrases for all 10 pairwise combinations between the five classes for each scenario. For the TF implementation we used an analysis window of 180 samples duration and 50% overlap. The thresholds related to the classification process were the same for both scenarios. In particular, T_p and $T_{\Delta p}$ were equal to 5 and 3 respectively, $T_{\Delta\tau}$ was set to 5 analysis frames, a was equal to 0.5 and σ was equal to -0.041. The k_{min} and k_{max} values used in (6) corresponded to frequencies of 1000 and 16000 Hz respectively.

As metrics of performance we use the precision rate and recall rate, calculated as

$$\text{Precision} = \frac{\sharp \text{ correctly recognized events}}{\sharp \text{ detected events}}, \qquad (7)$$

$$\text{Recall} = \frac{\sharp \text{ correctly recognized events}}{\sharp \text{ total events}}, \qquad (8)$$

where "total events" is the true number of events to be detected. The classification scores are shown for each class in Table 1 and 2 for the alternately-played and the simultaneous events respectively. As expected, simultaneously played events represent a much more difficult challenge for the presented method, but the scores obtained are still satisfactory. We observed that almost in all cases where an error occurred, the classifier picked an impact region lying in between the two locations that were actually excited. For example, when regions 1 and 4 were excited, regions 2 and 3 had a lot of chances of being falsely detected, which also explains why class 3 has the lowest precision rate in Table 2. To our opinion, this is an indication that the WDO assumption explained in section 2.2 is not always fulfilled, which is somehow expected because the different events originate from the same object and therefore exhibit little variability in their spectra.

Of additional concern in this experiment is the calculation time required for applying the proposed method at each time frame. Implemented in Matlab with an Intel Core i7 @3.4 GHz CPU, the average computation time required for pro-

Source	Total events	Precision	Recall
1	62	1.00	0.98
2	66	1.00	1.00
3	69	1.00	1.00
4	71	1.00	1.00
5	75	0.99	1.00

Table 1. Classification scores for alternately-played events.

Source	Total events	Precision	Recall
1	112	1.00	0.94
2	116	0.96	0.94
3	112	0.87	0.96
4	116	0.96	0.96
5	116	0.99	0.93

Table 2. Classification scores for simultaneous events.

cessing a single time frame was 0.4 ms, which is well below the actual duration of the time frame, equal to 4 ms in this experiment. This verifies that the method is appropriate for a real-time application.

4. CONCLUSION

A lot has been said and written about how to discriminate percussive events based on their timbral differences but little on how to classify them based on their spatial variability. In this paper, the detection and classification of human-generated impact sounds is accomplished simultaneously from the solution to a source separation problem. We have proposed using a simple training phase for learning the instantaneous linear mixing model, assuming a multiplicity of spatially distributed impact sound sources and sensors whose locations are consistent during the training and testing phase. An extension of this work would be to explore more sophisticated techniques for source amplitude recovery than the presented WDO approach, as well as to investigate approaches for estimating the strike strengths. Finally, one of the major priorities of the authors is to apply and evaluate the method in the case of a real drum kit and also in more realistic "live" conditions where other musical sources are also present.

REFERENCES

[1] K. Yoshii, M. Goto, and G. Okuno, "Automatic drum sound description for real-world music using template adaptation and matching methods," in *Proceedings of 5th Int. Conf. on Music Information Retrieval*, 2014, pp. 184–191.

[2] U. Şimşekli, A. Jylhä, C. Erkut, and T. Cemgil, "Real-time recognition of percussive sounds by a model-based method," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, pp. 1–14, 2011.

[3] G. Weinberg and S. Driscoll, "Toward robotic musicianship," *J. Computer Music*, vol. 30, no. 4, pp. 28–45, 2006.

[4] A. Jylhä, I. Ekman, C. Erkut, and K. Tahiroglu, "ipalmas - an interactive flamenco rhythm machine," in *Proc. Audio Mostly*, 2009, pp. 69–76.

[5] A. Jylhä and C. Erkut, "A hand clap interface for sonic interaction with the computer," in *Proc. Human Factors in Computing Systems*, 2009, pp. 3175–3180.

[6] S. Vesa and T. Lokki, "An eyes-free user interface controlled by finger snaps," in *Proc. 8th Int. Conf. Digital Audio Effects*, 2005, pp. 262–265.

[7] D. Fitzgerald, *Automatic drum transcription and source separation*, Dublin Institute of Technology, 2004, PhD Thesis.

[8] E. Battenberg, V. Huang, and D. Wessel, "Toward live drum separation using probabilistic spectral clustering based on the itakura-saito divergence," in *AES 45th Int. Conf.*, 2012.

[9] M. Wanderley and P. Depalle, "Gestural control of sound synthesis," *Proceedings of the IEEE*, vol. 92, no. 4, pp. 632–644, 2004.

[10] N. Stefanakis, Y. Mastorakis, and A. Mouchtaris, "Instantaneous detection and classification of impact sound; turning simple objects into powerful musical control interfaces," in *Proc. joint ICMC-SMC conference*, 2014, pp. 1178–1184.

[11] D. Wessel and M. Wright, "Problems and prospects for intimate musical control of computers," *Computer Music*, vol. 26, no. 3, pp. 11–22, 2002.

[12] S. Winter, W. Kellerman, H. Sawada, and S. Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l1-norm minimization," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007, Article ID 24717.

[13] H. Tan, Y. Zhu, L. Chainsorn, and S. Rahardja, "Audio onset detection using energy-based and pitch-based processing," in *Proc. 2010 IEEE Int. Symposium on Circuits and Systems*, Paris, 2010, pp. 3689–3692.

[14] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Process.*, vol. 52, pp. 1830–1847, 2004.

[15] M. Swartling, B. Sällberg, and N. Grbić, "Source localization for multiple speech sources using low complexity non-parametric source separation and clustering," *Signal Process.*, vol. 91, pp. 1781–1788, 2011.