ANALYSIS AND SYNTHESIS OF HAND CLAPPING SOUNDS BASED ON ADAPTIVE DICTIONARY

Wasim Ahmad, Ahmet M. Kondoz

I-Lab, Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey, Guildford, GU2 7XH, United Kingdom
{w.ahmad, a.kondoz}@surrey.ac.uk

ABSTRACT

In the past few years, there has been an increased demand for new tools and methods to design and generate high quality sounds effects and natural sounds for gaming, and virtual reality applications. A number of analysis-synthesis methods have been developed to synthesize musical and everyday sounds using predefined analysis techniques. In this paper, we present a new dictionary-based analysis, parameterization, and synthesis method for the generation of clapping sounds. The main objective is to use audio grains to create finely-controlled synthesized sounds which are based on recordings of clapping sounds. During the analysis stage, sequences of pre-recoded clapping sounds are initially segmented into individual claps, and are then decomposed into multi-level time-scale components or grains. The extracted audio grains are optimized using K-SVD dictionary learning algorithm, which forms the basis for an adaptive dictionary. For the parameterization, each recorded individual clap is projected onto the trained dictionary, which produces the synthesis pattern. During the generation of a clap sound, the synthesis pattern and atoms from the dictionary are selected and tuned according to the parameters received from the physical interaction or via the graphical user interface. A method for expressive synthesis is also presented, which generates non-repetitive clapping sounds.

1. INTRODUCTION

Clapping sound of hands is one of the simplest percussive sounds because its production does not involve any tools or musical instruments. It is used virtually in everyday aspect of human life, and this form of human expression is found in almost every culture, and in particular as a rhythmic musical instrument. Clapping is also used to indicate agreement and appreciation where it is repeated for few seconds, and often replicated by the group. From an acoustician's viewpoint, clapping can be studied individually as a sound generated by the act of hitting both hands together, or collectively, as claps generated by a group of people.

Sound synthesis is used extensively in the music and film industry, as well as in the development of games and virtual reality applications. For example, synthetic sounds of hand clapping can enhance or even replace live recordings in sport games by allowing the modeling of the applause generated by a large audience. It can also be used to provide the performer with direct audio cues on his/her performance, such as producing an enthusiastic applause for an outstanding performance. Hand clapping has also been used as a substitute for language. For instance, Hanahara *et al*. [5] developed a human-robot communication method based on hand clapping sound, where formal language specifications were defined to represent syllables, spoken words and syntax.

Despite the significant and widespread use of clapping in our everyday life, very few researchers have addressed the subject of analysis and synthesis of clapping sounds. Repp [18] was one of the first who studied hand clapping sound and presented a number of important results. He observed that the shape of the average clap spectra of each subject varies considerably between individuals. To explore the source of this variability, Repp investigated the possible linkage between average clap spectra and the clappers' sex, hand configurations (clapping style), and hand size. He discovered that clappers' sex and hand size had no significant influence on the spectrum of clap but hand configuration of the clappers was the major source of variability in the clap spectrum. Using auditory information, Repp also studied the subjects' ability to recognize clapper identity, sex, hand size, and the hand configuration. He observed that about half of the subjects were able to recognize themselves but overall recognition was very poor. From subjects' feedback, Repp also found that subjects were not very successful in identifying the sex and the hand size of the clappers but they were very good at recognizing the different hand configurations. Jylhä and Erkut [6] presented a technique that can classify the clapping styles of a clapper from synthetic and recorded hand clapping sounds. On the synthesis side, Peltola *et al*. [15] presented physics-based synthesis algorithms and various control methods to generate sounds of single and a group of clappers. In the first synthesis system, the modes of vibration present in the clapping sound were modeled using resonator filters, and their parameters were derived from the recoded clapping sounds. The second system was devised to synthesize the sound of various clapping styles using a number of derived parameters from the experimental measurement.

In this paper, we propose an analysis based synthesis algorithm, which parameterizes the pre-recoded clapping sounds in the form of atoms and synthesis patterns, and generate a sequence of claps from these parameters on the fly. During the analysis process, the recorded sequences of claps are initially segmented into individual claps, and then stationary wavelet transform (SWT) is used to decompose them into sound grains. These sound grains form the basis of the initial dictionary, which is further optimized to produce a compact and adaptive version of dictionary. The segmented claps are projected onto the adaptive dictionary which generates the synthesis patterns for them. During the synthesis of clapping sounds, these patterns are tuned according to the reported synthesis parameters either from the physical interaction or via the graphical user interface (GUI). This technique generates realistic and expressive claps sound for interactive and multimedia applications.

The detailed introduction of dictionary-based signal analysis and representation methods is presented in section 2. In section 3, the different blocks of the proposed analysis-synthesis model are explained in detail. The need for an expressive synthesis model and its importance in the natural sound synthesis context are discussed in Section 4. In section 5, the summary of achievements are recapitulated and the future directions of the research are highlighted.

2. DICTIONARY-BASED SIGNAL REPRESENTATION TECHNIQUES

Conventional signal analysis and representation techniques, such as Fourier transform (FT) and short-time Fourier transform (STFT), use Fourier basis to represent the signal as a superposition of fixed basis functions i.e. sinusoids. The Fourier basis functions provide a useful representation when considering stationary signals, but most real-world everyday signals are nonstationary and transient. Therefore, these analysis and representation techniques are inadequate for such signals because of their poor localization both in time and frequency.

Over the last few years, researchers have been investigating new techniques which can represent signals in a compact form and are specialized to the signal under consideration. As a result, a number of basis functions and representation techniques [19, 10, 2] have been developed, so that any input signal can be represented in a way that is more compact, efficient and meaningful. One of such techniques, which has gained a lot of recognition in recent years, is the dictionary-based method as it offers compact representation of the signal and is highly adaptive. Dictionary-based methods have been used in many signal processing applications including analysis and representation of audio signals [4, 20] and music [7].

In dictionary-based methods, an input signal is represented as a linear combination of *atoms*. These atoms are prototype discrete-time signals, and a collection of K such signals is referred to a *dictionary*. Let $\mathbf{s}$ be a discrete-time real signal of length n, i.e. $\mathbf{s} \in \Re^n$, and $\mathbf{D} = [\phi_1, \phi_2, \ldots, \phi_K]$ be a dictionary, where each column ϕ represents an atom and its length is n, i.e. $\mathbf{D} \in \Re^{n\times K}$. Using dictionary-based methods, the aim is to represent $\mathbf{s}$ as a weighted sum of atoms, which can be written as,

$$\mathbf{s} = \sum_{k=1}^{K} \phi_k \, u_k \tag{1}$$

where $\mathbf{u}$ is a column vector in $\Re^K$ and represents the expansion coefficients or weights. Generally, an overcomplete dictionary $\mathbf{D}$ ($n < K$) is used, which means the matrix $\mathbf{D}$ has a rank n and the weights vector $\mathbf{u}$ in Eq. (1) will not have a unique solution. Therefore, some additional constraints need to be introduced to determine a unique or particular decomposition.

The representation of a signal given in Eq. (1) is usually approximated instead of solving for an exact solution. An adequate and commonly used approximation of Eq. (1) is the one where the signal $\mathbf{s}$ is represented by selecting only j number of atoms from the dictionary $\mathbf{D}$ corresponding to highest weights u_k. Such representation shows that the signal energy is predominated in few atoms that have highest weights. Therefore, the approximation of the signal $\mathbf{s}$ can be represented as,

$$\mathbf{s} \;=\; \sum_{k=1}^{j} \phi_k \, u_k + \mathbf{r} \;=\; \mathbf{Du} + \mathbf{r} \tag{2}$$

where j is the number of selected atoms ($j < K$), and $\mathbf{r} \in \Re^n$ is residual or approximation error. The selection of atoms and their numbers are controlled by limiting the value of approximation error. By applying such criterion, the approximation solution given in Eq. (2) can be redefined as,

$$\mathbf{s} \approx \mathbf{Du} \;\text{ such that }\; \|\mathbf{s} - \mathbf{Du}\|_2 \le \varepsilon \tag{3}$$

where ε is a given small positive number. The approximation solution with the fewer number of atoms and corresponding weights is certainly an appealing representation. Sparse or compact approximation of a signal $\mathbf{s}$ is measured using the ℓ_0 criterion, which counts the number of non-zero entries of the weights vector $\mathbf{u} \in \Re^K$. Finding the optimally sparse representation consists in finding the solution of

$$\min_{\mathbf{u}} \|\mathbf{u}\|_0 \;\text{ such that }\; \|\mathbf{s} - \mathbf{Du}\|_2 \le \varepsilon \tag{4}$$

where $\|\mathbf{u}\|_0$ is the ℓ_0-norm, which counts the number of non-zero coefficient in weight vector $\mathbf{u}$. The problem of finding the optimally sparse representation, i.e. with minimum $\|\mathbf{u}\|_0$, is in general a combinatorial optimization problem. Constraining the solution $\mathbf{u}$ to have the minimum number of nonzero elements creates an NP-hard problem [13] and cannot be solved easily. Therefore, approximation algorithms, such as basis pursuit (BP) [2], matching pursuit (MP) [10], and orthogonal matching pursuit (OMP) [14], are employed to compute an optimal approximation solution of Eq. (4). The MP and OMP algorithms are classified as greedy methods, because they approximate the signal iteratively where at each iteration an

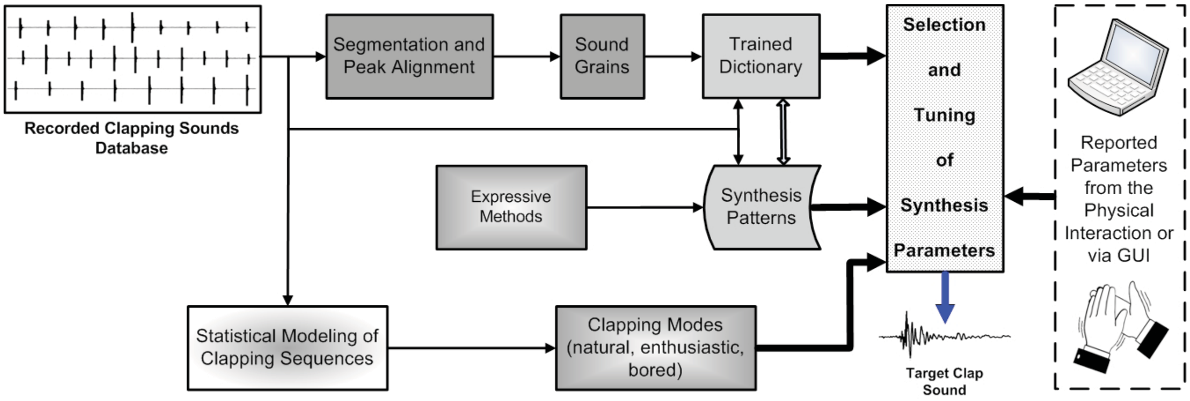
Recorded Clapping Sounds Database
Segmentation and Peak Alignment
Sound Grains
Trained Dictionary
Selection and Tuning of Synthesis Parameters
Expressive Methods
Synthesis Patterns
Reported Parameters from the Physical Interaction or via GUI
Statistical Modeling of Clapping Sequences
Clapping Modes (natural, enthusiastic, bored)
Target Clap Sound

Figure 1. Overview of proposed analysis-synthesis algorithm for clapping sounds.

atom is selected from the defined dictionary which maximally reduces the residual signal. These algorithms converge rapidly, and exhibit good approximation properties for given criterion [10].

The sparse approximation of Eq. (4) can also be improved by using a specialized dictionary, which is trained from the signals under consideration. Instead of using predetermined dictionaries, dictionary learning methods [1, 7] can be used to refine them. This topic is addressed in detail in section 3.4.

3. ANALYSIS-SYNTHESIS ALGORITHM

A dictionary-based synthesis algorithm presented here generates the clapping sounds using the parametric representation modeled from the recorded sounds. Fig. 1 depicts the building blocks of the proposed analysis-synthesis scheme. The algorithm takes recorded continuous clapping sounds and split them into sound grains. During parameterization phase, the clapping sounds are represented by synthesis patterns and an adaptive dictionary, which is trained from these sound grains. The target clap is generated at the synthesis stage where a pattern is selected and adjusted according to the reported parameters. In the following sections, each part of the algorithm is discussed in detail.

3.1. The Database

The proposed synthesis scheme models the hand clapping sounds through the analysis of generated sounds. Therefore, a set of hand clapping sounds were recorded. The recordings of hand clapping sounds were made in an acoustical booth (T60 $<$ 100 ms) at a sampling rate of 44.1 kHz. The recording was made with two males and one female subject, all aged between 20 and 35. Each subject was seated in the acoustical booth alone and a microphone was placed about 100 cm away from their hands. Each subject was asked to clap at their most comfortable or natural rate (i.e. clapping mode) using his/her conventional hand configuration (i.e. clapping style). Then the subject was asked to clap at very enthusiastic and very bored rates using the same clapping style. A sequence of 20 claps was recorded in each clapping mode from each subject.

3.2. Segmentation and Peak Alignment

The first step during the analysis of recorded hand clapping sound is to segment each sound signal into individual *sound events* i.e. single clap, which is represented by $\mathbf{s}_i$. Each clap from a sequence of 20 is isolated by detecting and labeling its onset and offset points. Onset of each clap is labeled by using the energy distribution method proposed by Masri *et al.* [11]. This method detects the beginning of an impulsive event, such as clap, by observing the suddenness and the increase in energy of the attack transient. Short-time energy of the signal is used to locate the offset of each clap. Starting from the onset of each event, the short-time energy is calculated with overlapped frames, and compared against a constant threshold to determine the offset.

There is no constraint on the number of claps or events taken from each clapping mode. Equal or different number of events can be selected from each clapper and their clapping modes. For simplicity, an equal number of claps, i.e. 20, were taken from each clapping mode of a clapper. Once the claps are selected and segmented, they are peak aligned by means of cross-correlation such that the highest peaks occur at the same point in time. This increases the similarities between the extracted sound grains and improves the dictionary learning process. The set of collected sound events are put into a matrix form as,

$$\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m] \tag{5}$$

where each column represents a clap (sound event) and length of each clap is n. Zero padding is used for any segmented clap whose length is less than n. For this experiment, $n = 2048$ and $m = 180$.

3.3. Sound Grains

The main idea behind the proposed analysis scheme is that we want to represent the recorded clapping sounds in a way that i) the similarities and differences between

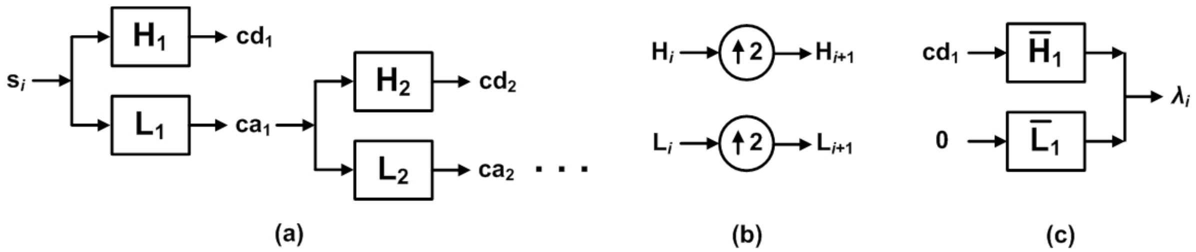

Figure 2. (a) Decomposition tree of SWT, (b) SWT filters, (c) construction of a sound grain.

clapping sounds recorded from different subjects can be observed and parameterized, and ii) this parametric representation can be manipulated in various ways to generate sound effects at synthesis stage. Clapping sounds belong to transient signal family that is non-stationary. Based on the frequency resolution properties of the human auditory system, such signals can be split into layers of grains, where the energy of each grain is present at particular frequency or scale. The information in each grain and the overall structure of these grains are analyzed and represented based on human auditory system. Such parametric representation can be used to compare the characteristics of different sounds [21]. Furthermore, during the synthesis process, the parameters representing these grains can be manipulated in various ways to control the generated sound.

The discrete wavelet transform (DWT) has gained widespread recognition and popularity due to its ability to underline and represent time-varying spectral properties of many transient and other nonstationary signals, and offers localization both in time and frequency. Stationary wavelet transform (SWT) [12, 16] is a real-valued extension to the standard DWT, which intended to solve the shift-invariance problem of the DWT. Therefore, in the proposed analysis scheme, the SWT is used to extract the sound grains from the clapping sound.

The SWT is applied to each clap $\mathbf{s}_i$ which decomposes it into two sets of wavelet coefficient vectors: the approximation coefficients $\mathbf{ca_1}$ and the detail coefficients $\mathbf{cd_1}$, where the subscript represents the level of decomposition. The approximation coefficient vector $\mathbf{ca_1}$ is further split into two parts, $\mathbf{ca_2}$ and $\mathbf{cd_2}$, using the scheme shown in Fig. 2(a). This decomposition process continues up to L^{th} level which produces the following set of coefficient vectors: $[\mathbf{cd}_1, \mathbf{cd}_2, \ldots, \mathbf{cd}_L, \mathbf{ca}_L]$. The approximation coefficients represent the low-frequency components, whereas the detail coefficients represent the high-frequency components. To construct the sound grains from coefficients vectors, the inverse SWT is applied to each coefficient vector individually by setting all others to zero, which produces the following bandlimited sound grains: $[\lambda_1, \lambda_2, \ldots, \lambda_{L+1}]$. The block diagram of the process of acquiring the sound grains from coefficient vectors is shown in Fig. 2(c). Each grain contains unique information from the sound event and its length is the same as the sound event. The entire clap matrix $\mathbf{S}$ is split into sound grains which produce grain matrix $\Lambda = [\lambda_i : i = 1, 2, \ldots, p]$, where λ_i form the columns of the grain matrix and the number of total grains are $p = m \times (L+1)$.

The selection of wavelet type from the family of wavelets (i.e. Haar, Daubechies, etc.) and their decomposition level depend on the input sound signal, application area, and the representation model. This is an iterative process where the best wavelet type and optimum decomposition level are obtained by evaluating the perceived quality of the synthesized sounds generated from the different wavelet types and decomposition levels. Based on listening tests, *db4* wavelet with 5^{th}-level decomposition ($L = 5$) is used in the presented work.

3.4. Dictionary and Synthesis Patterns

The proper parameterization of the sound features extracted from the analysis part is an essential element of the synthesis systems. In the proposed scheme, a dictionary-based approach is used to create a parametric representation of the recorded sounds. The similarities and differences of the sound grains, as well as their relationships to the input sounds are preserved and reflected in the presented parametric representation. One key advantage of dictionary-based signal representation methods is the adaptivity of the composing atoms. This gives the user the ability to make a decomposition suited to specific structures in a signal. Therefore, one can select a dictionary either from a prespecified set of bases functions, such as wavelets, wavelet packets, Gabor, cosine packets, chirplets, warplets etc., or design one by adapting its content to fit a given set of signals, such as dictionary of instrument-specific harmonic atoms [7].

Choosing a prespecified basis matrix is appealing because of its simplicity, but there is no guarantee that these basis will lead to the compact representation of all type of signals. The performance of such dictionaries depends on how suitable they are to describe these signals. However, there are many potential application areas, such as transient and complex music sound signals, where fixed basis expansions are not well-suited to model this type of signals. A compact decomposition is best achieved when the elements of the dictionary have strong similarities with the signal under consideration. In this case, a fewer set of more specialized basis functions in the dictionary is needed to describe the significant characteristics of the signal [7, 17, 8]. Ideally, the basis itself should be adapted to the specific class of signals which are used to compose the original signal. As we are dealing with a specific class of transient signals, we believe that it is more appropriate to consider designing dictionaries based on learning.

Given training clapping sound and using adaptive training process, we seek a dictionary that yields compact representations of the claps matrix, **S**. Aharon *et al.* [1] proposed such a method, named as K-SVD algorithm, which trains a dictionary from the given training signals. The K-SVD algorithm is a very effective technique, and has been used in many image processing applications [3, 9]. It is based on an iterative process of optimization to produce a sparse representation of the given samples using the current dictionary, and updating the atoms until the best representation is reached. Dictionary columns are updated along with the sparse representation coefficients related to it which accelerate the convergence.

In the proposed scheme, the K-SVD algorithm is used to train an adaptive dictionary **D** which determines the best possible representation of the given clapping sounds. The K-SVD algorithm takes the sound grains matrix Λ, as initial dictionary $\mathbf{D}_0$, a number of iterations j, and a set of given example signals, i.e. claps matrix **S**. Finding sparser representation of the sound events in **S** consists in solving the optimization problem

$$\min_{\mathbf{u}_i} \|\mathbf{s}_i - \mathbf{D}\mathbf{u}_i\|_2^2 \ \text{ such that } \ \forall i \ \|\mathbf{u}_i\|_0 \leq T_0 \qquad (6)$$

where T_0 is the number of non-zero entries in $\mathbf{u}_i$. The iteration of K-SVD algorithms is performed in two basic steps: i) given the current dictionary, the sound events in **S** are sparse-coded which produce the sparse representations matrix **U**, and ii) using these current sparse representations, the dictionary atoms are updated. The algorithm consists in updating the dictionary atom, one at a time, and optimizing the defined target function associated with it.

The orthogonal matching pursuit (OMP) is used to find the decomposition pattern of the input claps over the trained dictionary. The OMP is a greedy step-wise regression algorithm, and it is applied in the proposed scheme to approximate the solution of the sparsity-constrained sparse coding problem given in Eq. (6), where the dictionary atoms have been normalized. This algorithm selects at each stage the dictionary atom having the maximal projection onto the residual signal. The selected atom is subtracted from the current signal and the residual is recomputed. The algorithm stops after a predetermined number of steps, selecting a fix number of atoms T_0 in every iteration. Now the claps matrix **S** can be fully represented as a dictionary **D** and synthesis patterns **U**. The information about the subjects and clapping modes are labeled onto each synthesis pattern for future reference and for the possible use during synthesis process.

3.5. Synthesis

To synthesize the target clap sound, the reported parameters and the controlling variables are employed to select the best sound parameters. During synthesis process, a clap sound from the claps matrix **S** can be synthesized by selecting and adding the corresponding synthesis pattern and the dictionary atoms, which can be written as,

$$\hat{\mathbf{s}}_i \cong \sum_{j \in J} \phi_j \, \mathbf{u}_i(j) \qquad (7)$$

where vector J contains the T_0 number of indices of the non-zero entries in $\mathbf{u}_i$. The perceptual quality of the synthesized clap sound $\hat{\mathbf{s}}_i$ is directly related to the number of non-zero entries in $\mathbf{u}_i$. The quality of synthesized clap sound $\hat{\mathbf{s}}_i$ improves sharply for the first few atoms but become imperceptible after a particular value of T_0.

The synthesis Eq. (7) is used to generate a set of three synthesized claps taken from three different subjects at $T_0 = [5, 10, 15]$. It can be observed from Figs. 3, 4, and 5 that as the number of non-zero entries T_0 increases in the weights vector $\mathbf{u}_i$, the synthesized clap closely approximate the original sound with the best approximation achieved at the least value of $T_0 = 15$, after which the gain in the approximation is insignificant.

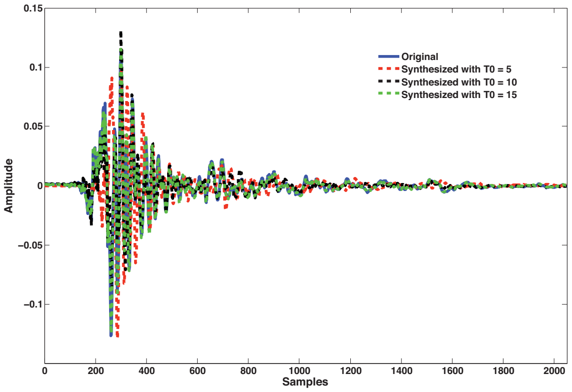

Figure 3. Original and synthesized claps from subject-1 at $T_0 = [5, 10, 15]$.

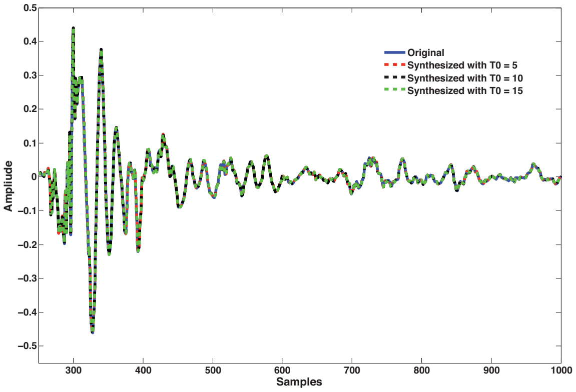

Figure 4. Magnified portion of original and synthesized claps from subject-2 at $T_0 = [5, 10, 15]$.

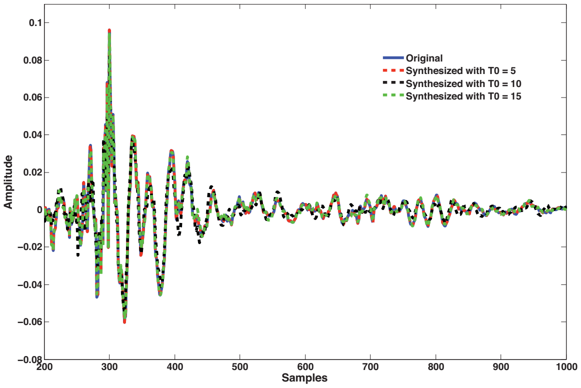

Figure 5. Magnified portion of original and synthesized claps from subject-3 at $T_0 = [5, 10, 15]$.

4. EXPRESSIVE SYNTHESIS

Two sound events generated consecutively by the same sound source will be similar but not identical. For example, when a person claps twice in the same way with the same applied force, the generated clapping sounds will be similar but not identical. The proposed algorithm can synthesize example claps sound approximately from the represented parameters, i.e. synthesis patterns $\mathbf{U}$ and a dictionary $\mathbf{D}$. A limited sequence of clapping sounds can be generated from this representation, as the numbers of synthesis pattern vectors are limited and fixed. Therefore, the same set of claps' sound will be repeated during long impact sound sequences, which will make it perceptually artificial in the ears of the listeners.

To generate more natural and customized sounds, an expressive synthesis process is presented here. The proposed method modifies the synthesis process given in Eq. (7). This equation uses the represented parameters, $\mathbf{U}$ and $\mathbf{D}$, to synthesize a clap sound. Every time Eq. (7) is executed to synthesize a clap sound $\hat{\mathbf{s}}_i$, a weights vector $\mathbf{u}_i$ is used to combine the dictionary atoms to generate a clap. For expressive synthesis, when a clap sound $\hat{\mathbf{s}}_i$ is generated, a small random vector α is added to the selected weights vector $\mathbf{u}_i$ such that the overall time-varying spectrum of the clap sound is unchanged. The value of α is generated randomly over a hypersphere of radius R with the origin at the weights vector of the generated clap sound. Different α vector is generated for every clap and the length of α is equal to T_0 because only non-zero entries in $\mathbf{u}_i$ are changed. Hence, The synthesis equation given in Eq. (7) is modified for the expressive synthesis process and can be rewritten as,

$$\hat{\mathbf{s}}_i \cong \sum_{j \in J} \phi_j \, [\mathbf{u}_i + \alpha](j). \tag{8}$$

The clapping sequence generated using Eq. (8) will be similar but not identical, and they will also not be exact copies of claps matrix $\mathbf{U}$.

To generate example expressive claps sound, the expressive synthesis model defined in Eq. (8) is used. Two different settings of α are generated and used to modify the selected $\mathbf{u}_i$. The original, approximated, and two expressive claps sounds are synthesized using $T_0 = 15$. The modified weights vectors and corresponding synthesized claps are plotted in Figs. 6 and 7. It may be observed that the synthesized claps using the expressive model are not identical to the originals but perceptually similar. The listeners gave the same feedback when these samples were played back.

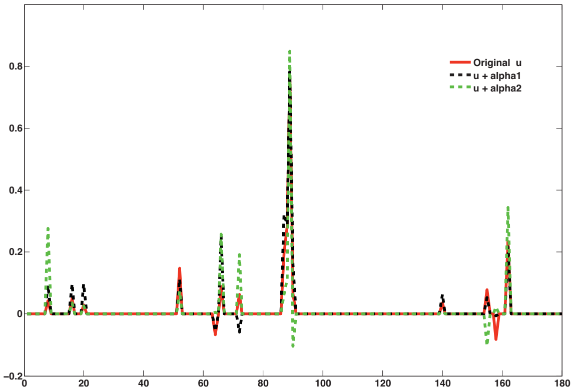

Figure 6. Original and modified $\mathbf{u}_i$ at $T_0 = 15$.

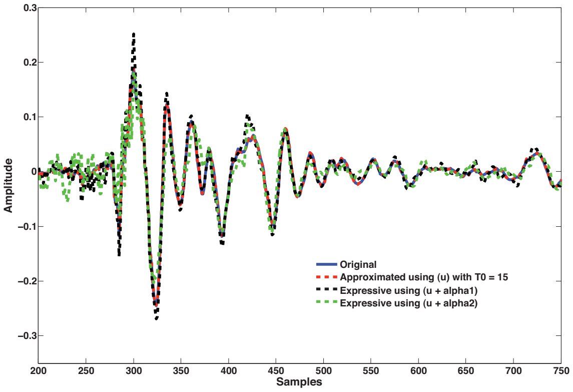

Figure 7. Magnified portion of original, synthesized approximated and expressive claps sound at $T_0 = 15$.

5. CONCLUSIONS AND FUTURE WORK

An analysis based synthesis algorithm was presented here which can synthesize clapping sounds from the represented parameters i.e. a set of atoms and synthesis patterns. The atoms of the dictionary were first adaptively trained from the recorded clapping sounds using K-SVD algorithm, and then the synthesis patterns were generated by projecting the sound events over the trained dictionary. The target clapping sound was synthesized by selecting and tuning the synthesis patterns and their corresponding atoms from the dictionary. In addition, an expressive synthesis method was presented which can generate non-repetitive and customized clapping sounds. The simulated claps showed that as the number of non-zero entries in weights vector increased to generate the target clap, the synthesized clap sound gets closer to the original clap. In some cases, even five weights (i.e. $T_0 = 5$) and their corresponding atoms were sufficient to generate a clap with a satisfactory level of perceived sound quality. It was also observed that an approximation sound with $T_0 = 15$ was sufficient to yield an excellent perceived sound quality.

In future, we will create control models for a single clapper and for several clappers with different clapping modes. We would also like to further investigate the expressive synthesis model and analyze the distribution of synthesis patterns of real life sound events and their possible statistical or mathematical modeling. We will then evaluate the perceptual quality of the synthesis model using subjective tests.

6. REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein, “K-svd: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, Dec 1998.

[3] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736 –3745, Dec 2006.

[4] R. Gribonval and E. Bacry, “Harmonic decomposition of audio signals with matching pursuit,” *IEEE Transactions on Signal Processing*, vol. 51, no. 1, pp. 101–111, Jan 2003.

[5] K. Hanahara, Y. Tada, and T. Muroi, “Human-robot communication by means of hand-clapping (preliminary experiment with hand-clapping language),” in *Proc. of IEEE Int. Conf. on Systems, Man and Cybernetics (ISIC-2007)*, Oct 2007, pp. 2995 –3000.

[6] A. Jylhä and C. Erkut, “Inferring the hand configuration from hand clapping sounds,” in *Proc. of Int. Conf. on Digital Audio Effects (DAFx-08)*, Espoo, Finland, Sep 2008.

[7] P. Leveau, E. Vincent, G. Richard, and L. Daudet, “Instrument-specific harmonic atoms for mid-level music representation,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 116–128, Jan 2008.

[8] M. S. Lewicki, T. J. Sejnowski, and H. Hughes, “Learning overcomplete representations,” *Neural Computation*, vol. 12, no. 2, pp. 337–365, Feb 2000.

[9] J. Mairal, M. Elad, and G. Sapiro, “Sparse representation for color image restoration,” *IEEE Transactions on Image Processing,*, vol. 17, no. 1, pp. 53 –69, Jan 2008.

[10] S. G. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, Dec 1993.

[11] P. Masri and A. Bateman, “Improved modeling of attack transients in music analysis-resynthesis,” in *Proc. of Int. Computer Music Conf. (ICMC-96)*, Hong Kong, China, Aug 1996, pp. 100–103.

[12] G. P. Nason and B. Silverman, “The stationary wavelet transform and some statistical applications,” in *Lecture Notes in Statistics, 103*, 1995, pp. 281–299.

[13] B. K. Natarajan, “Sparse approximate solutions to linear systems,” *SIAM Journal on Computing*, vol. 24, no. 2, pp. 227–234, Apr 1995.

[14] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,” in *Proc. of Asilomar Conf. on Signals, Systems and Computers*, vol. 1, Nov 1993, pp. 40–44.

[15] L. Peltola, C. Erkut, P. R. Cook, and V. Välimäki, “Synthesis of hand clapping sounds,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1021–1029, Mar 2007.

[16] J. C. Pesquet, H. Krim, and H. Carfatan, “Time-invariant orthonormal wavelet representations,” *IEEE Transactions on Signal Processing*, vol. 44, no. 8, pp. 1964–1970, August 1996.

[17] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, “Sparse representations in audio and music: From coding to source separation,” *Proceedings of the IEEE*, vol. 98, no. 6, pp. 995–1005, Jun 2010.

[18] B. H. Repp, “The sound of two hands clapping: An exploratory study,” *Journal of the Acoustical Society of America*, vol. 81, no. 4, pp. 1100–1109, Apr 1987.

[19] C. Roads, “Introduction to granular synthesis,” *Computer Music Journal*, vol. 12, no. 2, pp. 11–13, 1988.

[20] B. L. Sturm, C. Roads, A. McLeran, and J. J. Shynk, “Analysis, visualization, and transformation of audio signals using dictionary-based methods,” in *Proc. of Int. Computer Music Conf. (ICMC-2008)*, Belfast, Northern Ireland, Aug 2008.

[21] S. Tucker and G. J. Brown, “Classification of transient sonar sounds using perceptually motivated features,” *IEEE Journal of Oceanic Engineering*, vol. 30, no. 3, pp. 588–600, Jul 2005.