We encounter a wide range of sounds everyday which are harmonic, transient or a mixture of both. A number of models have been developed in recent years to synthesise harmonic sounds but these models do not perform effectively on transient sounds because of the sharp attack and decay part of the transient sounds. This paper presents a new technique for analysis-based synthesis of transient sounds by adding orthogonal bases of frequency bands. The proposed technique consists of four stages. First, the sound samples taken from a group of sounds are decomposed into frequency bands using the stationary wavelet transform (SWT). A set of orthonormal bases for each frequency band is then computed using singular value decomposition (SVD). The model parameters for all sounds present in the group are derived from the orthogonal basis and their weights. Finally, the required sound is synthesised by adding the weighted orthogonal basis for each frequency band and then taking the inverse stationary wavelet transform (ISWT). The proposed technique provides a generic synthesis approach to synthesise a diverse transient set of sounds using a single analysis-synthesis model.

1. INTRODUCTION

Our environment is full of diverse types of sounds such as cracking, hitting, collision, bumping, breaking, dripping, car engine, etc. Most of these sounds are transient impact sounds with energy concentrated in a short time window. Such transient impact sounds are generally produced when two or more objects interact with each other. The resulting sounds are dependent on the physical properties of the vibrating objects such as, shape, size, elasticity, material, and the type of interaction such as, position, force of interaction, velocity. Therefore, upon hearing an impact sound, listeners can identify the general properties of the objects that generated it. Listening to everyday sounds is not just listening to these sounds per se but rather it is an experience of hearing events in the world [1].

During the last two decades a number of synthesis algorithms have been developed and applied mostly to synthesise musical instrument sounds [2]. Some of these synthesis algorithms were adapted to synthesise everyday impact sounds but the results were not effective and satisfactory [3] as the perceptual dimensions and attributes of everyday listening are different from musical listening [1]. Therefore new synthesis algorithms have been introduced in recent years to synthesise everyday impact sounds.

1.1. Related work

Physical models were initially introduced by the computer music community to simulate the existing musical instruments, but subsequently this technique was used for sound rendering in interactive applications. The physical models represent the relationship between the sound and its connection with the physical characteristics of the structure. Doel et al. [4] presented a framework to synthesise contact sounds using modal synthesis which describes the acoustic properties of the vibrating objects. Doel also [5] developed a physically-based sound synthesis model for liquid sounds. Doel identified that more complex sounds like streams, pouring water, rivers, and rain can also be generated from single water bubble model and their stochastic models. Rath et al. [6] used a modified form of modal synthesis to synthesise a number of impact sounds. Rath et al. used the proposed model to generate bouncing, breaking, rolling, and sliding sounds.

Cook [7] introduced two algorithms, PhiSAM (Physically Informed Spectral Additive Modelling) and PhiSEM (Physically Informed Stochastic Event Modelling), to analyse and synthesise musical and everyday sounds. The PhiSAM is based on modal synthesis and is suitable for objects which are struck or plucked. The PhiSEM is based on pseudo-random overlapping and adding of small grains of sound. Cook used PhiSEM algorithm to analyse and synthesise footsteps, police whistle, gravel shaking, crunching of gravel, and sandpaper sounds [8, 7, 3]. Peltola et al. [9] presented two physics-based synthesis algorithms and control methods to synthesise hand clapping sounds. These methods can synthesise single hand-clap and applause from a group of clappers.

Spectral models represent and imitate the properties of sound signals and parameterise sounds as perceived by
the listener. Aramaki et al. [10] proposed an analysis-synthesis model based on a time-varying subtractive synthesis process which acts on a noisy input signal. This model was aimed at reproducing perceptual effects corresponding to impacted materials, without simulating physical properties. Serra [11] developed a method, called spectral modeling synthesis (SMS) which can be used to synthesise the sound produced by musical instruments or by any physical system. The proposed algorithm is based on modeling sounds as stable sinusoids (i.e. deterministic component) plus noise (i.e. stochastic components). The SMS model restricts the deterministic part to sinusoidal components with piecewise linear amplitude and frequency variations. This affects the generality of the model and sounds, like transient and noisy, cannot be accurately modeled by the technique. Verma et al. [12] presented an algorithm, called transient modeling synthesis (TMS), to analyse and synthesise the transient part of the input signals. It was also proposed that the TMS can be combined with spectral modeling synthesis (SMS) to model a wide range of instrumental sounds. Misra et al. [13] introduced an analysis-synthesis system with user interface that was used for analysis, transformation, and synthesis of natural sounds. It used deterministic, transient, and stochastic models to analyse and represent the input sound. Deterministic components were extracted using the sinusoidal analysis [11] and the transient part was detected and modeled using transient modeling [12].

1.2. Motivation

The diversity of everyday sounds and sources is such that it is not feasible to generate all or a group of everyday impact sounds using a single analysis-synthesis model. Therefore effective and flexible analysis-synthesis models are needed for environmental transient impact sounds which can be used to design sounds for computer music, computer games, and virtual reality (VR) applications.

Physical models are very efficient and accurate in simulating simple sounds where a simple vibrating structure is involved and a mathematical model of the interaction can be deduced. For complex structures where the vibrating mechanism can not easily be described, physical models can become infeasible. Furthermore, the analysis process in physical models is specifically related to that particular sound or musical instrument and is not applicable to any other sound source. From a perceptual point of view, the refinement of physical models is not always successful because the physical mechanisms of many environmental sounds are still not completely understood [7]. Spectral models have a broader scope than that of the physical models because they parameterise and construct the spectrum as received in the ear. Therefore, their refinement and repurposing is easier than physical models. On the other hand, spectral models (e.g. additive synthesis, spectral modeling synthesis (SMS)) perform very well on the harmonic, stationary, and noise free sounds but fail to model and synthesise transient, nonstationary, and noisy sounds accurately [11]. Therefore, generic synthesis models with better analysis and parameterisation techniques are required that can be applied to many if not all of the impact sounds.

In this paper, an analysis-synthesis algorithm is presented which has the potential to synthesise common transient sounds. The analysis process in the proposed model uses the stationary wavelet transform (SWT). The SWT analyses both the frequency and time behaviour of the input signal and has the ability to analyse transient, nonstationary, and noisy sounds. The parameterisation process uses singular value decomposition (SVD) that parameterises and represents the data by identifying the similarities and differences present in the data. Furthermore, the presented synthesis model is generic, robust, and can synthesise a single sound or a group of transient sounds.

1.3. Overview

The detailed introduction of the stationary wavelet transform and singular value decomposition is presented in sections 2 and 3 respectively. In section 4, the proposed analysis-synthesis model is presented and different blocks of the model are explained in detail. In section 5, a set of example sounds and their synthesis are presented. The need for an expressive synthesis model and its importance in the natural sound synthesis context are also discussed. In section 6, the summary of achievements and the results of the paper are reviewed.

2. STATIONARY WAVELET TRANSFORM

The wavelet transform (WT) decomposes an input sound signal into wavelet coefficients through a series of filtering operations. Similar to short-time Fourier transform (STFT) where a window is translated to analyse the input signal, WT uses wavelet function to analyse the input signal which is not only translated but also scaled (expanded or dilated) in time. A wavelet $\psi(t)$ is a function $\psi \in L^2(\mathbb{R})$ with a zero average. $\int \psi(t)dt = 0$. The wavelet $\psi(t)$ is also named as mother wavelet because it generates a new family of wavelets by scaling $\psi(t)$ by $a$ and translating it by $\tau$:

$$\psi_{\tau,a}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t - \tau}{a}\right)$$  \hspace{1cm} (1)

Usually, the mother wavelet is dilated by power of two and translated by integer. The wavelet transform of a sound signal $s(t)$ at time $\tau$ and scale $a$ can be represented as:

$$Ws(\tau, a) = \langle s(t), \psi_{\tau,a}(t) \rangle$$  \hspace{1cm} (2)

$$Ws(\tau, a) = \int_{-\infty}^{+\infty} s(t) \frac{1}{\sqrt{a}} \psi^*\left(\frac{t - \tau}{a}\right)dt$$  \hspace{1cm} (3)

$$= x \ast \tilde{\psi}_a(\tau)$$  \hspace{1cm} (4)
The singular value matrix. \( V \) is the right singular matrix, and \( U \) is the left singular matrix, and \( \Lambda \) is the singular value matrix.

where \( \ast \) represents the convolution and \( \hat{\psi}_a \) is,

\[
\hat{\psi}_a(t) = \frac{1}{\sqrt{a}} \psi^\ast \left( \frac{t}{a} \right).
\]

The convolution in Eq. (4) computes the wavelet transform of the input signal with dilated band-pass filters. The wavelet transform splits the input signal \( s(t) \) into two sets of coefficients at each level: approximation coefficients, \( C_A L \), and detail coefficients, \( C_D L \), where subscript represents the level of decomposition. The WT involves decimation process at each level which makes it a shift-variant transform. It means the wavelet transform of a signal and the wavelet transform of a time-shifted version of the same signal are significantly different. The lack of shift-invariance is a well-known disadvantage of the classical WT. There are real-valued extensions to the standard wavelet transform [14, 15, 16], which are shift invariant and which have been used for different purposes.

To overcome the shift-variance problem, the stationary wavelet transform (SWT) is used to decompose the input signal into frequency bands. The SWT allows the projection of the signal to an orthonormal basis which is shift-invariant and redundant. Nason et al. [14] proposed a simple algorithm for SWT that is adopted in this paper. In classical WT, the standard high and low pass filters are applied to the input signal which produces two sequences at each level. These sequences are decimated by factor of 2 at each level but filters are not modified. In the shift-invariant algorithm, the decimation is not applied on both high and low pass filtered sequence. Therefore, the output signals have the same length as the original undecimated signal. Instead, both high and low pass filters are modified at each level by interleaving them with zeros.

3. SINGULAR VALUE DECOMPOSITION

Singular value decomposition (SVD) is a well-known statistical technique that has been widely used in signal processing and statistics [17, 18]. SVD factorises the rectangular real or complex matrix into three simple matrices: two orthonormal matrices and a diagonal matrix. The SVD of a real valued \( m \times n \) rectangular matrix \( Y \) (where \( m \leq n \)) can be written as:

\[
Y = U \Lambda V^T \tag{6}
\]

where \( U = [\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_m] \) is an \( m \times m \) orthonormal matrix (i.e. \( UU^T = I \)), \( \Lambda \) is an \( m \times n \) matrix equal to \( \text{diag} \{ \lambda_1, \lambda_2, \ldots, \lambda_m \} : 0 \}, \) and \( V = [\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n] \) is an \( n \times n \) orthonormal matrix (i.e. \( VV^T = I \)). The diagonal elements \( \lambda_i \) are called the singular values of \( Y \) and they are represented as \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0 \) for convenience. The corresponding column vectors \( \vec{u}_i \) and \( \vec{v}_i \) are also arranged accordingly which means the matrices \( U \) and \( V \) are ordered from most variation to the least. This means, the first row of \( V^T \) has the highest variance and is called the principal component. The matrix \( U \) is the left singular matrix, \( V \) is the right singular matrix, and \( \Lambda \) is the singular value matrix.

4. SOUND SYNTHESIS MODEL

In this section, a sound synthesis model is presented, which simulates the impact sounds. The proposed synthesis model addresses the two major issues encountered in the existing models: the lack of generality and the deficiency to analyse transient sounds. The presented synthesis model has the capability to analyse and synthesise all kinds of everyday impact sounds or groups of impact sounds.

A block diagram which depicts the building blocks of the proposed sound synthesis model is shown in Fig. 1. The synthesis model is made up of three stages: analysis, parameterisation, and synthesis. The analysis part takes the input sound signals and extracts the most salient features of the sound. The parameterisation block finds the relationship between these features and represents them in a compact form. In the synthesis block, the user controls the variables and the sound parameters that are used to generate the required sound signal. The performance of the synthesis block and the synthesis sound quality measures the robustness of the selected features and their parameterisation. In fact, these blocks are interlinked and the synthesis quality depends on all of them.

![Figure 1. Sound synthesis model.](image)

4.1. Analysis process

Parameterisation of input sounds is the final representation of sound features at the end of the analysis process. As the success of synthesis depends on the quality of these methods, the analysis part forms the core of the analysis-synthesis system. Furthermore, the analysis process is responsible for extracting the sound features that are most significant and directly related to the input sound characteristics.

The proposed analysis process uses time-scale decomposition of the input signal which is known as stationary wavelet transform (SWT) which makes the analysis process generic and robust. The stationary wavelet transform has the ability to perform local analysis, i.e. to analyse a localised area of a large signal. The SWT decomposes the signal into different frequency components and then represents each component at a different scale. The SWT extracts the sound features that are salient, localised, and
directly related to input sound signal. Therefore, SWT has the ability to underlie and represent time-varying spectral properties of the transient and nonstationary signal.

4.1.1. Pre-processing

In Figure 1, the input sound signals \{\bar{s}_i : i = 1, \ldots, m\} form one group of input sounds. The sound \bar{s}_i can be a single sound event or a sequence of same event repeated \(n\) times. For example, it can be one clapping event or a sequence of a number of claps. If \(\bar{s}_i\) is a sequence of the same sound event, then each sound event is segmented and the onset and offset points are labelled. The length of each segment in one group of sound \(\{\bar{s}_i : i = 1, \ldots, m\}\) should be the same. There is no constraint on the number of sound events required from each type of sound, \(\bar{s}_i\). Equal or different number of sound events can be selected from each type of sound but the length of all the segments should be equal. During the segmentation of the sound, all the sound segments are peak aligned i.e. same length and the heights peak occurs at the same point. This increases the similarities between the extracted sound features and improves the parameterisation process. All the segmented sound events are put into a matrix form where each row contains samples of a segmented sound event. The input sound is represented using following input matrix:

\[
X = \begin{bmatrix}
\bar{s}_1 \\
\bar{s}_2 \\
\vdots \\
\bar{s}_m
\end{bmatrix} = \begin{bmatrix}
s_{11} & s_{12} & \cdots & s_{1n} \\
s_{21} & s_{22} & \cdots & s_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
s_{m1} & s_{m2} & \cdots & s_{mn}
\end{bmatrix} \quad (7)
\]

where \(m\) represents the number of sound samples and \(n\) represents the length of each sound samples.

4.1.2. Sound features extraction using stationary wavelet transform

To extract the wavelet coefficients, the stationary wavelet transform is applied to each row of input matrix, \(X\), which decomposes it into the first level approximation coefficients matrix \(\text{CA}_1\), and the first level detail coefficients matrix \(\text{CD}_1\). The sizes of these matrices are same as that of \(X\) and rows of these matrices represent the wavelet coefficients of the corresponding row of \(X\). The first level approximation coefficients matrix \(\text{CA}_1\) is further decomposed using SWT into second level approximation coefficients matrix \(\text{CA}_2\), and the second level detail coefficients matrix \(\text{CD}_2\). This decomposition process using SWT continues up to \(L^{th}\) level as shown in Fig. 2. At the end of \(L^{th}\) level decomposition, the set of coefficients matrices \{\(\text{CD}_1, \ldots, \text{CD}_L, \text{CA}_L\)\} are the sound feature that contains all of the time-varying frequency information of the input sound matrix \(X\). The sound feature matrix \(\text{CA}_L\) represents the low frequency approximation coefficients, and sound feature matrices \{\(\text{CD}_1, \ldots, \text{CD}_L\)\} represent high frequency detail coefficients of the input sound matrix. The approximation coefficients contain the slow changing characteristics of the signal. The detail coefficients contain the rapidly changing structure of the input sounds.

The selection of wavelet and their decomposition level depend on the application area and synthesis model. The sound features changes from signal to signal and application to application, therefore the best choice of wavelet is the one which extract the salient sound features successfully. The decomposition level reveals the time-varying frequency information of the input signal, which is used to analyse the time-frequency patterns of each band which plays key role in the parameterization process. Some sound groups expose these patterns at lower decomposition levels and some requires further decomposition.

![Figure 2. L^{th}-stage analysis tree of SWT.](image)

The extracted wavelet coefficient matrices \{\(\text{CD}_1, \ldots, \text{CD}_L, \text{CA}_L\)\} represent the time-frequency pattern of the input sound matrix \(X\), where \(i^{th}\) row of \(X\) (a sound event) is decomposed into six frequency bands which are placed in the \(i^{th}\) row of each feature matrix \{\(\text{CD}_1, \ldots, \text{CD}_L, \text{CA}_L\)\}. The sound events present in the input matrix, \(X\), can be different transient sounds and the differences and similarities between these sounds are reflected in these features matrices. The correlation and variation present in these sound features are parameterised by singular value decomposition in the next section.

4.2. Parameterisation process

The sound features extracted during the analysis process contain all the characteristics and information in the input sounds. The parameterisation process represents these sound features in a very compact way such that their similarities, differences, and interconnection with the input sound signals are preserved.

4.2.1. Sound parameters using singular value decomposition

The purpose of the singular value decomposition is to reduce the dimensionality of the dataset, identify existing patterns in data, and highlight the similarities and differences. SVD identifies and orders the dimensions in which input data exhibit the greatest variation. Once it is identified where the most variation is, it is possible to find the
best approximation of the original data points using fewer dimensions.

The coefficient matrices \{CD_1, \ldots, CD_r, CA_L\} represent the time-frequency pattern where in each matrix some of the rows have very strong correlation. The SVD can be used to factorise each coefficient matrix and then find the best way to represent them. Let us take any one feature matrix \( C \in \{CD_1, \ldots, CD_r, CA_L\} \) and factorise it using SVD. Expanding Eq. (6) using matrices and vectors, we obtained:

\[
\begin{bmatrix}
\vec{c}_1 \\
\vec{c}_2 \\
\vdots \\
\vec{c}_m
\end{bmatrix} = \begin{bmatrix}
\vec{u}_1 \\
\vec{u}_2 \\
\vdots \\
\vec{u}_m
\end{bmatrix}^T \begin{bmatrix}
\lambda_1 & 0 & 0 & \cdots & 0 & s_{nj} \\
0 & \lambda_2 & 0 & \cdots & 0 & s_{nj} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \lambda_m & \cdots & 0
\end{bmatrix} \begin{bmatrix}
\vec{v}_1 \\
\vec{v}_2 \\
\vdots \\
\vec{v}_m
\end{bmatrix}
\]

(8)

\[
\begin{bmatrix}
\vec{c}_1 \\
\vec{c}_2 \\
\vdots \\
\vec{c}_m
\end{bmatrix} = \begin{bmatrix}
\vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_m
\end{bmatrix} \begin{bmatrix}
\lambda_1 \vec{v}_1 \\
\lambda_2 \vec{v}_2 \\
\vdots \\
\lambda_m \vec{v}_m
\end{bmatrix}
\]

(9)

\[
\begin{bmatrix}
\vec{c}_1 \\
\vec{c}_2 \\
\vdots \\
\vec{c}_m
\end{bmatrix} = \begin{bmatrix}
\vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_m
\end{bmatrix} \begin{bmatrix}
\Phi_1 \\
\Phi_2 \\
\vdots \\
\Phi_m
\end{bmatrix}
\]

(10)

where \( \{\vec{u}_i : i = 1, \ldots, m\} \) are the column vectors, \( \{\Phi_j : j = 1, \ldots, m\} \) are the row vectors, and \( \vec{c}_i \) is the \( i\)th coefficient vector. Eq. (10) can be expanded as:

\[
\begin{bmatrix}
\vec{c}_1 \\
\vec{c}_2 \\
\vdots \\
\vec{c}_m
\end{bmatrix} = \begin{bmatrix}
u_{11} & u_{21} & \cdots & u_{m1} \\
u_{12} & u_{22} & \cdots & u_{m2} \\
\vdots & \vdots & \ddots & \vdots \\
u_{1m} & u_{2m} & \cdots & u_{mm}
\end{bmatrix} \begin{bmatrix}
\Phi_1 \\
\Phi_2 \\
\vdots \\
\Phi_m
\end{bmatrix}_{m \times n}
\]

(11)

The rows of matrix \( V^T \) are orthonomal and form a linearly independent basis which spans the input matrix \( C \). Therefore, the matrix product \( UV^T \) produces a matrix \( \Phi \) whose rows are linearly independent and forms the orthogonal basis which spans the input matrix \( C \). Equation (11) shows that each row vector of coefficient matrix can be written as a linear combination of orthogonal bases, \( \{\Phi_j : j = 1, \ldots, m\} \), and weight vectors, \( \{\vec{u}_i : i = 1, \ldots, m\} \). Therefore, Eq. (10) can be written as:

\[
\vec{c}_i = \sum_{j=1}^{r} u_{ji} \Phi_j
\]

(12)

Equation (12) reveals that any coefficient vector can be perfectly reconstructed using the orthogonal bases matrix \( \Phi \), and the corresponding weight matrix \( U \), obtained using SVD. For perfect reconstruction of \( C \), \( r = m \) and for approximation of \( C \), \( r < m \). Therefore, a set of orthogonal bases \( \Phi \) and weight \( U \) matrices are calculated to parameterise each coefficient matrix. A set of \( \{U, \Phi\} \) are the parametric representation for each coefficient matrix.

It is known that the singular values \( \lambda_i \) are in decreasing order and the rows of matrix \( V^T \) are ordered from most variation to the least in Eq. (6) and (10), therefore the rows of the orthogonal basis matrix \( \Phi \) are also in order from most variation to the least. This means, first few basis vectors of \( \Phi \) contain most of the information about the sound event set. Hence, to find the approximation of the input sound coefficient matrix, variation below a particular threshold can simply be ignored to massively reduce the data.

4.3. Synthesis process

The synthesis process is mainly controlled by input control variable from the user. The control variables are employed to select the best sound parameters for synthesis process.

In the presented synthesis model, any sound signal from the set of input sounds \( \{s_i : i = 1, \ldots, m\} \) can be synthesised by taking the inverse stationary wavelet transform (ISWT) of the set of approximation and detail coefficients which are the linear combination of orthogonal bases \( \Phi \) and their weights \( U \) (sound parameters). For example, the synthesis of a sound signal \( \hat{s}_k \in \{s_i : i = 1, \ldots, m\} \) using all the orthogonal basis (perfect reconstruction with \( r = m \)) can be obtained by using Eq. (13), where ISWT stands for inverse stationary wavelet transform. The approximation synthesis of the sound signal \( \hat{s}_k \) using first two orthogonal basis is described in Eq. (14), when \( r = 2 \).

\[
\hat{s}_k = \text{ISWT} \begin{cases}
CD_{kl}^1 = \sum_{j=1}^{r} u_{(1,k)}^{CD} \Phi_j^1 \\
\vdots \\
CD_{kl}^r = \sum_{j=1}^{r} u_{(r,k)}^{CD} \Phi_j^r \\
CA_{lk}^1 = \sum_{j=1}^{r} u_{(1,k)}^{CA} \Phi_j^1 \\
\vdots \\
CA_{lk}^r = \sum_{j=1}^{r} u_{(r,k)}^{CA} \Phi_j^r
\end{cases}
\]

(13)

\[
\hat{s}_k = \text{ISWT} \begin{cases}
CD_{kl}^1 = u_{(1,k)}^{CD} \Phi_j^1 + u_{(2,k)}^{CD} \Phi_j^2 \\
\vdots \\
CD_{kl}^r = u_{(r,k)}^{CD} \Phi_j^1 + u_{(2,k)}^{CD} \Phi_j^2 \\
CA_{lk}^1 = u_{(1,k)}^{CA} \Phi_j^1 + u_{(2,k)}^{CA} \Phi_j^2 \\
\vdots \\
CA_{lk}^r = u_{(r,k)}^{CA} \Phi_j^1 + u_{(2,k)}^{CA} \Phi_j^2
\end{cases}
\]

(14)

The quality of the synthesised sound \( \hat{s}_k \) is directly related to the number of bases used during the synthesis process but this proportionality is very sharp for the first few basis as they contain the most information of the feature matrices (in order from most variation to the least). Therefore, the change in the synthesis quality will be imperceptible after a particular value of \( r \) (threshold value) which can be found heuristically or by using subjective listening tests. During synthesis process, same or different value of \( r \) can be used for each feature vector band \( \{CD_{1,1}, \ldots, CD_{1,r}, CA_{1,1}, \ldots, CA_{1,r}\} \).

5. SYNTHESIS EXAMPLES

The proposed synthesis model is implemented on a group of common impact sounds. The group contains six different impact sounds which were recorded in an acoustical
booth (T60 < 100 ms) at CCSR, University of Surrey. These impact sounds are: tennis ball, taped tennis ball (tennis ball covered with PVC insulation tape), football, basketball bumping sounds on laminate floor, male, and female clapping sounds.

5.1. Synthesis results

The sound database was recorded at sampling rate of 44.1 kHz. A sequence of sound events was recorded for each sound source. One sound event $s_k$ from each sound source is taken for analysis purposes. Highest peaks of all the sound signals are aligned and their onset and offset points are labelled. The length of each sound signal present in one group should be the same. If a sound sample is not long enough, zeros are padded at the end of the signal. The length of each sound sample used here is 8192 samples. The segmented, peak aligned, and energy normalised samples of sounds are plotted in Fig. 3 and 4. These sound samples are put into a matrix $X = [s_1, \ldots, s_6]^T$, where each row represents a sound sample and the column represents the length of each sample.

![Figure 3](image3.png)

Figure 3. The input tennis and taped tennis balls, football, and basketball bumping sounds samples.

![Figure 4](image4.png)

Figure 4. The input male and female clap sound sample.

An extensive simulation was performed to select the best wavelet basis and decomposition level for the input sound matrix $X$. It was found that ‘db4’ at decomposition level five outperformed all others. It was also observed that the synthesis quality does not change by increasing the decomposition level further. Therefore, the SWT using ‘db4’ wavelets as applied to the sound matrix $X$ decomposed the signal up to the fifth level. This transform decomposed the sound matrix $X$ into six coefficient matrices $\{CD_1, \ldots, CD_5, CA_5\}$ where the $i^{th}$ row of each of these matrices is a coefficient vector belonging to the $i^{th}$ row of sound matrix $X$. The coefficients matrices are then parameterised using SVD where a set of orthogonal basis $\{\Phi_i : i = 1, \ldots, 6\}$ and the corresponding weight matrix $\{u_{ji} : j = 1, \ldots, 6; i = 1, \ldots, 6\}$ are obtained for each coefficient matrix.

During synthesis process, any sound signal $s_i \in X$ can be synthesised either perfectly by using Eq. (13), or approximately by using Eq. (14). For example, the tennis ball sound is synthesised from full set of orthogonal basis $r = m = 6$ (perfect reconstruction) using Eq. (13). The original and synthesised signal is shown in Fig. 5 below. Furthermore, the basketball sound and the female clap sound are also synthesised from full set of orthogonal bases where $r = m = 6$ (perfect reconstruction), and from sub-set of orthogonal bases when $r = 5$ and $r = 4$ (approximation). These original and synthesised signals are plotted in Fig. 6 and 7 respectively.

![Figure 5](image5.png)

Figure 5. Original and synthesised tennis ball sound using full set of basis (perfect reconstruction).

![Figure 6](image6.png)

Figure 6. Original and synthesised basketball using full set ($r = 6$) and subset of basis ($r = 5$, $r = 4$).
5.2. Expressive synthesis using interpolation of weights

When an everyday sound source generates two sound events consecutively, they both may be similar but not identical. For example, when a person claps twice in the same way with the same applied force, a ball is dropped from the same height twice etc., they generate similar sounds but not identical.

The presented sound synthesis model produces the input sound accurately. The synthesis model can generate only one sound event for every sound during the synthesis process and exactly the same sound event is generated if the synthesis for that particular sound is repeated \( n \) times. However, this is not natural and a listener can easily perceive that the same sound is repeated \( n \) times.

To generate more natural expressive sounds, an expressive synthesis process is presented here. The expressive synthesis process modifies the synthesis process proposed in Section 4. The calculated sound parameters \( \{U, \Phi\} \) are used in Eq. (13) to generate any sound signal \( \tilde{s}_k \in \{ \tilde{s}_i : i = 1, \ldots, m\} \). Every time this equation is repeated for sound signal \( \tilde{s}_k \), it uses the same set of weights to combine the orthogonal bases and generate one sound event of \( \tilde{s}_k \).

In the proposed expressive synthesis process, these set of weights are modified by adding a small random vector \( \tilde{\epsilon} \) such that the overall time-varying spectrum for the target sound will not change. The value of \( \tilde{\epsilon} \) is generated randomly over a hypersphere of radius \( R \) with the origin at the weight vector of the generated sound. Different \( \tilde{\epsilon} \) are generated for each frequency band. The size of \( \tilde{\epsilon} \) is same as of weight vector and the radius \( R \) is controlled by the user. Thus, the Eq. (13) is modified for the expressive synthesis process as given in the Eq. (15).

\[
\begin{align*}
\tilde{s}_k = \text{ISWT} & \left( \sum_{j=1}^{r} (U_{CD}^{1}(j,k) + \epsilon_{1(j)}^{CD}) \Phi_{j}^{CD} \
= \vdots \quad \vdots \quad \vdots \
C_{DL}^{i} & \sum_{j=1}^{r} (U_{CDL}^{i}(j,k) + \epsilon_{1(j)}^{CDL}) \Phi_{j}^{CDL} \
C_{AL}^{i} & \sum_{j=1}^{r} (U_{CDL}^{i}(j,k) + \epsilon_{1(j)}^{CDL}) \Phi_{j}^{CDL}
\right)
\end{align*}
\]  

When any sound signal \( \tilde{s}_k \in \{ \tilde{s}_i : i = 1, \ldots, m\} \) is synthesised using Eq. (15), different set of \( \tilde{\epsilon} \) are generated that update the weights and produces expressive sounds, which are similar but not identical. The expressive sound synthesis model, presented in Eq. (15), is used to synthesise taped tennis ball and football sounds. The original and the three expressive sound samples are plotted in Figs. 8 and 9. It may be observed that the sound events synthesised using the expressive model are not identical to the originals. The listeners gave the same feedback when these samples were played back. These sound signals are perceptually similar.

6. CONCLUSIONS

A large proportion of existing musical synthesis models is not suitable for everyday impact sounds as their analysis and parameterisation processes are specific to harmonic sounds. We addressed both the analysis and parameterisation processes and proposed a model which is more general and which can synthesise common transient sounds effectively. We used SWT in the analysis process, which gives better time-frequency varying information of the nonstationary and transient signals as compared to STFT. SVD is used to parameterise the sound features, which represents each sound features matrix as a linear combination of orthogonal basis. The expressive synthesis is also presented, which generates natural sounds.
The proposed model is evaluated using a set of six everyday transient sounds. The simulation results showed that the perfect reconstruction of the input sound is possible using a full set of orthogonal bases. The sounds synthesised using a subset of orthogonal basis is also very similar to the original and perceptually convincing. An expressive synthesis model was presented which generates sounds that are more natural and convincing. This enhances the realism of the generated sounds and has potential applications in computer music, computer games, and virtual reality.

In future, we want to further investigate the expressive synthesis model and analyse the distribution of weight vectors in real life sound events and their possible statistical or mathematical modeling. The quality of the synthesis models and the synthesised sound quality will be evaluated using subjective tests.

7. ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under Research Grant GR/S72320/01 Portfolio Partnership Award in Integrated Electronics.

8. REFERENCES


