# An Audio-driven Virtual Dance-Teaching Assistant

Slim Essid[*]
Institut Telecom, Télécom
ParisTech, CNRS-LTCI
Paris
France

Yves Grenier
Institut Telecom, Télécom
ParisTech, CNRS-LTCI
Paris
France

Mounira Mazaaoui
Institut Telecom, Télécom
ParisTech, CNRS-LTCI
Paris
France

Gaël Richard
Institut Telecom, Télécom
ParisTech, CNRS-LTCI
Paris
France

Robin Tournemene
Institut Telecom, Télécom
ParisTech, CNRS-LTCI
Paris
France

## ABSTRACT

This work addresses the Huawei/3Dlife Grand challenge proposing a set of audio tools for a virtual dance-teaching assistant. These tools are meant to help the dance student develop a sense of rhythm to correctly synchronize his/her movements and steps to the musical timing of the choreographies to be executed. They consist of three main components, namely a music (beat) analysis module, a source separation and remastering module and a dance step segmentation module. These components enable to create augmented tutorial videos highlighting the rhythmic information using, for instance, a synthetic dance teacher voice, but also videos highlighting the steps executed by a student to help in the evaluation of his/her performance.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

## General Terms

Algorithms

## Keywords

Grand Challenge, audio, 3DLife, music analysis, source separation, remastering.

## 1. INTRODUCTION

The Huawei/3DLife grand challenge considers the scenario where online dance lessons are given by an expert Salsa-dance teacher through the Internet. In this scenario, both the teacher's and students' performances are to be automatically analyzed and rendered using their respective avatars in an online virtual dance room. One of the main originalities of this grand challenge is linked to the dance scenes acquisition which was done using a multitude of sensors, thus enabling novel means for multimedia scene analysis, interpretation and interaction. For instance, the recording set up includes a network of video cameras, multiple microphones (both overhead and onfloor microphones), inertial measurement units, etc.

In such a scenario, two main goals emerge, namely automatically enriching the teacher's performance or the music itself by appropriate metadata that will facilitate the learning and make it more lively; and automatically analyzing (and rating) the student's dance performance compared to the teacher's performance on the same choreography. For example, dancing in rhythm on Salsa music is known to be rather difficult for beginners mainly because it is not straightforward to locate the different beats in the music. However, the student will experience much less difficulty if the dance learning system integrates an automatic metronome or even better can synthesize the voice of the teacher counting in rhythm on the music.

Obviously, all sensors can be jointly exploited to reach the best possible analysis performance but it is also important to consider situations where only part of the sensors is available. Indeed, the capture devices available to a student joining the virtual dance lesson from his/her place may not be as sophisticated as the ones used by the teacher. In this work, we consider that only a subset of signals is available (namely all acoustic signals captured by the overhead and on-floor microphones) although all tools developed can be jointly used with other modalities for an enhanced system.

In this paper, we thus introduce a dance-teaching assistant system which exploits a number of tools that permit to enrich the student's learning experience and to help evaluate his/her performance.

The paper is organized as follows. An overview of the

[*]Authors are in alphabetical order

*MM'11* November 28 - December 1, 2011, Scottsdale, Arizona, USA

complete system is given in next section. Then, the different components and modules are described in section 4 and 5 respectively for the music analysis and for the dance performance evaluation. Finally, some conclusions and future work directions are suggested in section 6.

## 2. ARCHITECTURE OF THE DANCE TEACHING ASSISTANT SYSTEM

An overview of the proposed system is provided in Figure 1. The system is organized around three main components, namely a *source separation* module, a *music analysis* module and a *dance performance analysis* module.

The source separation module relies on echo cancellation techniques (see section 3) to separate the teacher's voice and dance step sounds from the audio mixture consisting of all the overlapping audio signals that are captured by the overhead microphones. This then allows the student (joining the virtual dance room from a distant site) to re-create a new audio mix of the music, teacher's voice and step sounds using the mixing gains that best suit him/her, for example having the teacher's voice sound much louder than the music.

The music analysis module is built around a *musical beat extraction* block and a *beat controlled synthesis* block. The goal of the musical beat extraction is to provide an estimation of the quasi regular beat locations in the musical signal. This information is then used by the beat controlled synthesis component to generate a music track which is augmented with audio effects meant to explicitly provide musical timing information that is essential to help the dance students synchronize their movements to the rhythm of the music. These effects are here chosen to be either hand clapping in rhythm on the extracted musical beats or voice counting those beats using the teacher's voice templates. The latter is a particularly useful alternative to the previously mentioned teacher's voice enhancement component on recordings where the teacher is not actually giving rhythm-related voice instructions across the whole duration of the choreography (which is the case with a number of recordings of the grand challenge dataset).

The third component is dedicated to the evaluation of the student's dance performance. This analysis is achieved by comparing the timing of the dance steps executed by the students to the ideal step timing (given by the ground-truth annotations provided with the dataset). The students' steps timing is deduced after automatic step detection is performed using either the onfloor audio sensors or the enhanced versions of the overhead microphone signals to simulate a situation where only the latter recordings are available.

## 3. SOURCE SEPARATION AND AUDIO REMASTERING

We aim to separate the teacher's voice and dancers' step sounds from the audio mixture captured by the overhead microphones to be able to remaster the audio track in such a way to highlight the sound sources of interest.

To this end, we rely on an echo cancellation technique [5], to "cancel" the musical background, taking advantage of the fact that the original music that is played in the dance room is separately recorded on channels 5 and 6 of the audio setup.

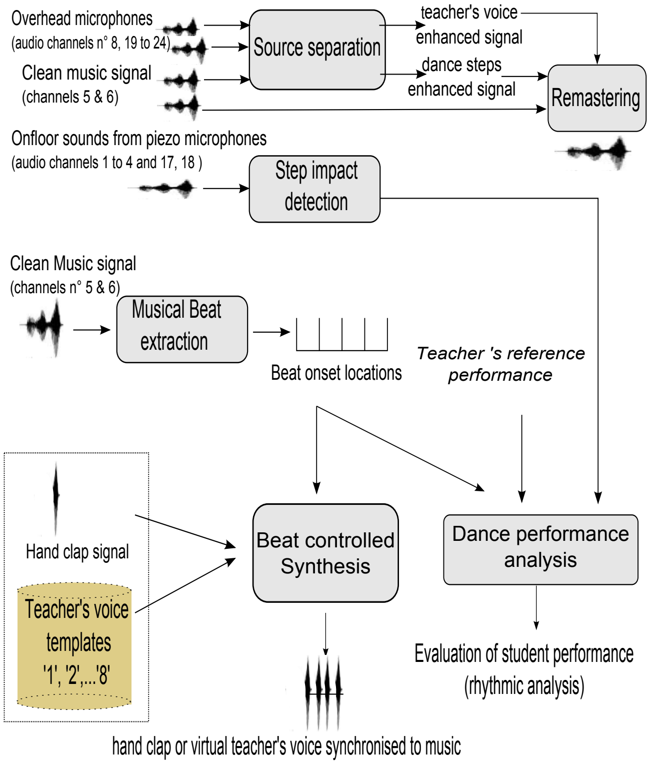

**Figure 1: Overview of the 3Dlife Dance audio analysis system**

The main challenge in real acoustic cancellation is to deal with the large impulse response present in this kind of application where thousands of FIR filter coefficients may be required to achieve the desired performance. We use an echo canceler based on a block frequency-domain adaptive filtering procedure which is the Generalized Multidelay Filter (GMDF) [5]. The key idea of this algorithm lies in the segmentation of the impulse response to be identified into small blocks and the introduction of a parameter that controls the overlap between the successive input blocks. The "echo" free signal, that is the music free signal, is reconstructed by a weighted overlap-and-add technique.

It is then possible to remix this signal with the "clean" music at a higher loudness level, thus realizing the desired remastering of the original audio. The reader is referred to [1] for some demonstrative examples.

## 4. MUSIC ANALYSIS AND AUDIO AUGMENTATION

### 4.1 Musical beat extraction

In many musical pieces, and especially in dance music, listeners can easily feel the *beat*, and tap their foot or clap their hands to the rhythm of the music. In most music, this tapping, or this beat sequence, is regular and this overall regularity is often summarized in the so-called *tempo*. Beat extraction from music signals or tempo estimation have been extensively studied in the past years and the interested reader is referred for the example to the following works [2],[4],[3],[7].

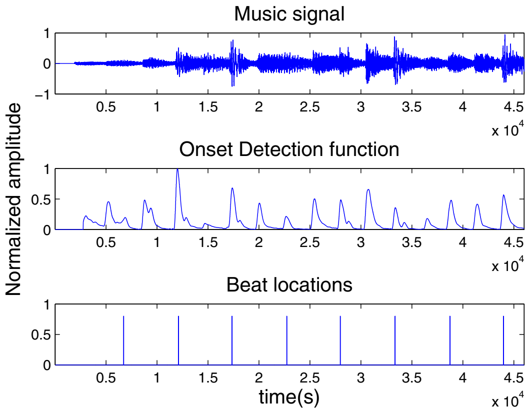

**Figure 2: Illustration of the Musical beat extraction system: (top) the music signal; (middle) the corresponding onset detection function and (bottom) the beat positions.**

Our approach is mainly based on [3] but integrates a specific module for the extraction of the beat locations. The first step of the musical beat extraction system is then computing an onset detection function (ODF). This ODF is obtained by summing over all frequencies the values of the spectral flux (*i.e.* a derivative of the signal spectrum amplitude) computed for each analysis window. The ODF is then further processed and an estimation of the tempo variations across the music excerpt is computed and smoothed using a Dynamic Programming approach (see [3] for more details). Then, based on this tempo curve the beat positions are estimated as a pseudo-periodic sequence (the period being given by the tempo at the corresponding time). To be well synchronized in time with the beats, the initial phase (or equivalently initial sequence delay) is also estimated.

The output of the module is illustrated in Figure 2.

## 4.2 Beat controlled synthesis

Once an estimation of the beat locations is obtained, it is possible to generate different types of signals that can enrich the original music signal and facilitate learning to the dance student. The simplest example is to generate a hand clap on each beat location, which can be regarded as the audio rendering of the scene corresponding to the teacher's virtual agent clapping his/her hands in rhythm to the music. A more sophisticated example would illustrate the situation where the virtual teacher is counting the beats by sequence of eight successive beats[1]. In this case, the synthesis is done by generating the successive teacher's speech templates on the beat locations. Note however, that for a fully automatic processing, it is necessary to automatically locate the downbeat, *i.e.* the strongest beat of the measure, (see for example [6]). However, on Salsa music, this is a particularly difficult task which is still today an open problem. In this work, one of the beat onset locations (found automatically) is supposed to be manually labeled by the teacher as a downbeat to improve the robustness of the system. Sound examples of beat controlled synthesis on the Huawei/3Dlife Grand challenge data are given on line [1].

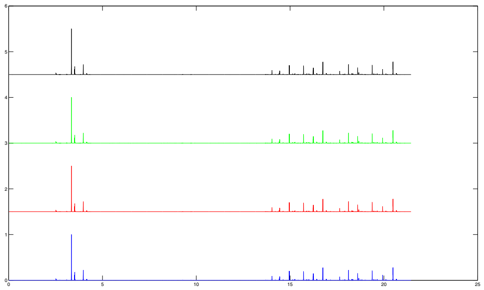

**Figure 3: Audio ODF features extracted from signals captured by the onfloor Piezo sensors.**

## 5. DANCE PERFORMANCE ANALYSIS

### 5.1 Step impact detection

Step impacts are detected using the signals captured by the onfloor piezoelectric transducers (audio channels 1, 2, 17 and 18). The impacts are located by:

- first, extracting onset detection functions $f_c(t)$ from every signal (as done in the beat extraction stage), where $c$ is a channel number ($c \in \{1, 2, 17, 18\}$ or $c \in \{8, 19, 20, 21, 22, 23, 24\}$), and $t$ is a time index;
- then forming feature vectors $\mathbf{x}_t$ which are, at every time instant $t$, merely the concatenation of the coefficients of every onset detection function, that is $\mathbf{x}_t = [f_1(t), f_2(t), f_{17}(t), f_{18}(t)]$;
- applying a one-class Support Vector Machine (SVM) [8] to those feature vectors (as a way of achieving a fusion of the onset detection functions) and taking the SVM output values which are below a negative threshold to be indicators of step impact instants.

To understand the rationale behind the one-class SVM approach let us examine Figure 3 depicting the onset detection functions $f_c(t)$ for a particular recording (identified as `bertrand_c1_t1` in the dataset). Most samples of these feature values are close to 0 except at a few time indices, expected to indicate the step impacts, where higher values are observed. The latter can then be considered as outliers[2] of the feature vectors distributions which motivates the use of the single-class SVM technique for its ability to robustly detect outliers [8]. These outliers are merely the feature vector $\mathbf{x}_t$ observations where the SVM output is negative.

[1] Western style music is, most often, rhythmically structured in segments that are built on a small number of measures, themselves containing a predefined number of beats. In Salsa music, a segment traditionally contains two measures of four beats each, which justifies a count between one to eight

[2] The reader should bear in mind that the word "outliers" is here used in a statistical sense, and that these extreme observations are actually the ones to be selected.

Step impacts are hence detected by thresholding the SVM decision function and applying a heuristic post-processing avoiding the detection of steps that would be too close in time based on the prior that two consecutive steps should be at least 0.25s away one from another.

Figure 4 shows the result of the automatic step impact detection on the same recording `bertrand_c1_t1` along with ground-truth annotations. It can be seen that the approach successfully detects all steps on this example.

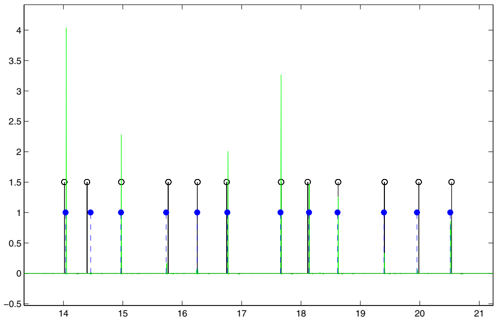

**Figure 4: Step detection results on recording `Bertrand_c1_t1`. Negative SVM decision values across time (in green) along with detected steps (dashed blue), and ground-truth annotations in black.**

## 5.2 Evaluation of students' performances

On the basis of the dance step segmentation previously exposed two types of evaluations of a student's performance are proposed:

- a means for a subjective evaluation of the dance performance timing that consists in the augmentation of the sound track of the feet video with audio effects under the form of synthetic beeps placed at time instants where steps are detected (see [1] for examples);
- an objective evaluation consisting in musical-timing ratings of the dance performance.

The automatic ratings aim to characterize two rhythmic aspects of each performance:

- the general timing precision of the dancer;
- the regularity of steps.

The first gives a general rating with respect to the ideal timing of the choreography (given by the automatic beat detection module or the ground-truth annotations in the dataset). The second assesses a dancer with respect to his/her own body rhythm, that is trying to determine whether he/she is arrhythmic or not. The two ratings refer to two different notions : an arrhythmic person may get an acceptable overall rating owing to the fact that he/she often steps at the right musical time instants, and yet be unable to be regular.

The general precision is computed as the arithmetic mean of the differences between step times and their nearest groundtruth counterparts. As for the regularity rating, we compute the standard deviation of the time-differences between every two consecutive steps (assuming a constant music tempo). We actually process only the first Salsa patterns of each choreography. Indeed, these patterns present the particularity of having a constant time interval between two consecutive steps : either 1 or 2 beats.

# 6. CONCLUSION

In this paper we have presented a set of audio tools composing a virtual dance-teaching assistant for the 3DLife/Huawei challenge centered on Salsa dance scenes. These tools are essential in teaching the dancers the sense of Salsa rhythm which is not easy to pick-up by beginners. Thanks to musical beat analysis, source separation and dancer's steps segmentation modules we are able to create videos of both tutorial and dance performance assessment content highlighting the rhythmic information of the music and the timing of the steps executed by the dancers.

Future work will look at improving the dance steps assessment component, in particular by tackling the problem of feet tracking using the onfloor sensors.

# 7. ACKNOWLEDGMENTS

This work was supported by the EU FP7 project 3DLife, ICT-247688.

# 8. REFERENCES

[1] http://huawei-3dlife-gc-submission.blogspot.com, Aug. 2011.

[2] M. Alonso, G. Richard, and B. David. Extracting note onsets from audio recordings. *Proc. of IEEE-ICME*, 2005.

[3] M. Alonso, G. Richard, and B. David. Accurate tempo estimation based on harmonic+noise decomposition. *Eurasip Journal on Advances in Signal Processing*, Jan. 2007.

[4] J. Bello, L. Daudet, and M. Sandler. Automatic piano transcription using frequency and time-domain information. *IEEE Transactions on Audio, Speech and Language Processing*, 14(6):2242–2251, aug 2010.

[5] E. Moulines, O. Ait Amrane, and Y. Grenier. The generalized multidelay adaptive filter: structure and convergence analysis. *Signal Processing, IEEE Transactions on*, 43(1):14 –28, jan 1995.

[6] H. Papadopoulos and G. Peeters. Joint estimation of chords and downbeats from an audio signal. *IEEE Transactions on Audio, Speech & Language Processing*, 19(1):138–152, 2011.

[7] G. Peeters. Beat-tracking using a probabilistic framework and linear discriminant analysis. *Proc. DAFX*, 2009.

[8] B. Shölkopf and A. J. Smola. *Learning with kernels.* The MIT Press, Cambridge, MA, 2002.