REAL-TIME DETECTION AND VISUALIZATION OF CLARINET BAD SOUNDS[1]

Aggelos Gkiokas, Kostas Perifanos, Stefanos Nikolaidis

Voice and Sound Technology Department,
Institute for Language and Speech Processing,
R.C. "Athena"
Athens, Greece
agkiokas@ilsp.gr, kperi@ilsp.gr,
snikol@ilsp.gr

ABSTRACT

This paper describes an approach on real-time performance 3D visualization in the context of music education. A tool is described that produces sound visualizations during a student performance that are intuitively linked to common mistakes frequently observed in the performances of novice to intermediate students. The paper discusses the case of clarinet students. Nevertheless, the approach is also well suited for a wide range of wind or other instruments where similar mistakes are often encountered.

1. INTRODUCTION

The growing use of computer software for music teaching led research to a new evolving domain, Music Visualization. So far, this term has not been given a strict definition. Various approaches have been presented in the recent years in the literature regarding music visualization.

Hiraga [1] presented a system that visualizes a whole music piece. The proposed model reads music pieces from MIDI files, and no waveform analysis is made. At the same year Hiraga proposed a more advanced system in [2], however with same properties as in [1]. McLeod in [3] developed a system that computes and visualizes the pitch of a music performance in real-time. The visualization scheme is a 2 dimensional pitch-time graph. The system was tested by an expert violin teacher and the feedback was promising. Toivinianen in [4] provided a system that visualizes the tonal content of a musical piece using SOMs. Ferguson [5] proposed a very interesting work on visualizing music performance in real-time. The developed system provides visualization of important acoustic features, such as harmonic content, tuning discrepancy and noisiness.

Music visualization can serve different purposes in the context of music education. As an offline tool, it can offer students a way to examine different aspects of their performance. Such aspects can include information about their timing, rhythm, stability and overall quality. However, as a real-time tool, the visualization of the sound is displayed *during* the student performance.

This paper presents work carried out in the context of the VEMUS project. VEMUS (Virtual European Music School) is a project funded by the European Commission under the Information Society Technologies (IST) Programme of the Sixth Framework Programme (FP6). The VEMUS project aims to design, develop and evaluate an open, highly interactive, and networked multilingual music tuition framework for popular instruments and a set of innovative pedagogically-motivated e-learning components addressing different learning settings [6].

The aim of work described in this paper is to provide a real-time music 3D-visualization tool for clarinet sound in the context of music education. The feedback provided in real-time should be short and simple, avoiding to distract the students and helping them to go on despite any errors. Furthermore, the tool must help the students to gain a perception of their progress as the time goes by. To meet real-time requirements, the system needed to operate only based on rather simple spectral features, such as pitch, RMS energy and the partials amplitudes. The performance error detection machine also had to be kept simple.

The rest of the paper is organized as follows. Section 2 describes the basic clarinet errors that are common for students of the target levels. Section 3 provides the overall system architecture and a brief description of each component. Section 4 describes the method used for detecting performance errors. Section 5 presents the proposed visual model and the way the sound quality is mapped to an image. Conclusion and directions for further work are provided in section 6.

2. CLARINET BAD SOUNDS

In a related work, Zlatintsi [7] presented a classification of "bad" clarinet notes. The separation of these classes was made by taking into account two criteria: the cause of a mistake and the resulting sound quality. These classes are described in summary below:

Hollow notes: The main cause of a hollow note is the bad airflow in the clarinet. The main attribute of a hollow note is that the energy of the individual harmonics is lower than the normal.

[1] This work has been carried out under the VEMUS project which has been partially supported by the European Community under the Information Society Technologies (IST) priority of the 6th Framework Programme for R & D.

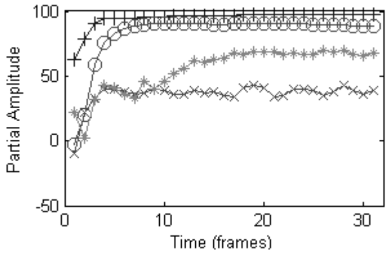

Figure 1: *The first four harmonics amplitudes over time for a good quality note. Index: "+":1st, "x": 2nd, "o": 3rd, "*": 4th. The same index will be used in Figures 2 and 3.*

However this cannot be described explicitly. "Hollowness" is related somehow with timbre. This leads us to the conclusion that it is preferable to address hollowness as a scalar property rather than crisply characterizing a note as hollow or not.

Squeak notes: The main cause of a squeak is saliva (into the clarinet or when the reed gets calcified) or when students press and bite the reed resulting no free vibrations. In a squeak note all the partials amplitudes become much higher than the normal. A graph of the harmonics amplitude over time for a good note, a hollow and a squeak note are shown in Figures 1,2 and 3 respectively.

Unstable notes: Unstable notes have many causes such as insufficient amount of airflow for the specific tone or not firm embouchure. Instability can be either pitch instability or RMS energy instability. Both can be easily detected by calculating the standard deviation of pitch and RMS-energy within a note (or part of note)

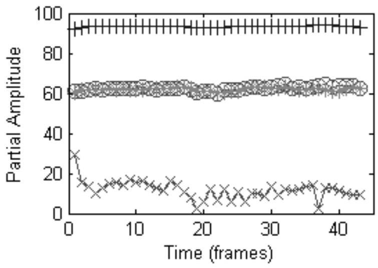

Figure 2: *The first four partials of a hollow note.*

3. SYSTEM OVERVIEW

The individual components of the proposed system and the flow diagram are shown in Figure 4. Real-Time Audio Recognizer (RTAR) reads streamed data from the microphone as the student performs the musical piece. With a conventional front-end processing scheme RTRA process a window of 25 ms long every 10 ms with a 60% overlap. For every window it processes, RTAR writes output data to the Audio Buffer and sends a message to the synchronizer module that a new frame is processed. Output data consists of pitch, RMS energy, the partials up to the 6th and the MIDI value of the tone played. The synchronizer activates the Error-Detection (ED) module. ED reads the Audio Buffer and computes the "Quality" or "Hollowness/Squakness" value as will be described in the next section. Every 4 iterations of this procedure (i.e. 40 ms) the synchronizer sends a message to the 2-Dimensional curve generator which produces the 2D curve. Finally the 2D curve is fed to the 3D-Curve Generator which draws

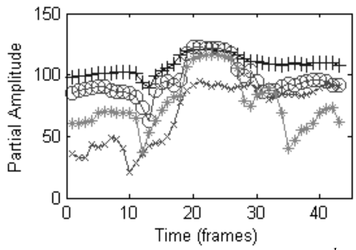

Figure 3: *A squeak note appearing at about 35th frame.*

the final shape. These rates are adjustable, to adapt to machines of different computing power.

4. BAD SOUND DETECTION

The intuition that a classic ML approach (labeling sounds as hollow, squeaks etc, training a machine, categorizing) would not succeed the desirable results led us to a different approach. Our system is based on the concept that a machine should be flexible on how it judges the students' performance. The teacher should be able to adjust the strictness of the module.

The limitation of the real-time processing led us to use only the partials extracted by the Audio Recognizer Tool, without any extra signal processing. Our approach is based on the fact that the mean values of the harmonics of a music piece performance can represent the clarinet's "sound". Any divergence from these values can be considered as a bad sound.

4.1. Features Used

The basic features used for the detection of clarinet bad sounds are the partials of each frame up to the 6th. Specifically, because of the fact that the first harmonic is proportional to RMS energy, we used partial values from the 2nd to the 6th, divided by the amplitude of the first. Thus, system uses 5 features denoted by $\{f_i\}, i=1..5$.

4.2. Training

In the training phase we fit a Gaussian distribution to data from recordings where a clarinet teacher or professional player performed.. During this process, and accordant to the bibliography we measured that the partial amplitudes depend strongly on the pitch. Thus, for different pitch values, we fit a different Gaussian distribution to the relative partials. Specifically, for each individual musical tone (MIDI value), we train a different model. This process results N Gaussian distributions for each feature, a total of $5N$.

$$p_{ij} = p_{Gauss}(\mu_{ij}, \sigma^2_{ij}, f_j), i = 1..N, j = 1..5 \qquad (1)$$

where index i is referred to the tone identity and j to the feature identity.

4.3. Error Detection

The error analysis presented in section 2 led us to the following admission. If the relative partial j is greater than the mean value calculated in the training phase for a specific tone, then contributes to the sound to be heard more squeak. Reversely if it's smaller, contributes to the sound to be heard hollower. The

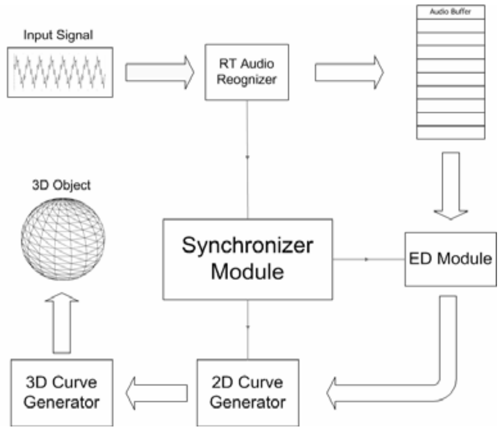

Figure 4: *The overall architecture of the system*

measure of this contribution is a quantity somewhat inverse proportional or decreasing to p_{ij}. The final characterization will be a sum of these quantities.

We tested various such functions from simple linear combination, to more complex ones. We found that a sum of powers of $1\text{-}p_{ij}$ worked well enough. The final formula we used is:

$$Q = \sum_{i=1}^{5} [sng(f_j - \mu_{ij})(1 - p_{ij})^4] \tag{2}$$

The polynomial power of 4 is used to smooth small variations of $1\text{-}p_{ij}$ around zero. The sign function in each clause has the role to define if the corresponding feature contributes for the sound to be heard more as a squeak or hollow. If Q is positive, means that we probably have a squeak sound, if negative a hollow sound. Closest to zero is Q, the better the sound is.

4.4. Time Averaging

Because in time space, one visual frame corresponds to more acoustic frame, it is not desirable to characterize the sound between two subsequent visual frames from only one acoustic frame. Thus, we take the average value of Q between these two visual frames.

The extreme case, where between two visual frames exist both almost equal high valued squeak and hollow sounds, resulting an average Q close to zero is almost impossible.

4.5. Onset and Offset Discarding

The sound modelling described before does not correspond to the case where the sound data processed is a part of the onset of a note. Onsets have very different statistical properties between the partials, thus it is ineffective to try characterizing such frames. In a very simple fashion, we discard onset frames, by ignoring the first frames of each note.

The same stands for the offset of each note. Higher partials decay faster than lower. In contrast with onset, we do not have prior knowledge when a note will end. Therefore such discarding is impossible.

We handled this situation in terms of smoothing, as will be described in section 5.4. The basic idea is when we have decaying in RMS energy on the signal, implying the note ends; we limit Q from changing value greater than a certain ratio. This worked well enough.

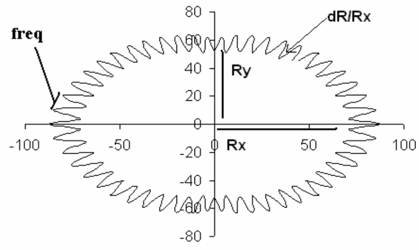

Figure 5: *The 2-Dimensional visual model*

4.6. Coping with the Different Level of Students

Visualizer module must have different behavior in different levels of students. For the same waveform, produced by a beginner and an intermediate student, visualization feedback must be stricter for the latter. This can be easily adopted adding one more parameter to the model described. In equation 1 we substitute standard deviation by a multiple of it by a value α.

Parameter α is global for all density components. The lesser α is, the more sensitive is system to mistakes. This extension allows the teacher by adjusting this value to personalize visualization according to the student level.

5. THE VISUAL MODEL

The Q value is fed to the Visualizer. The main idea is to represent a note as circle. This circle has four attributes to control (plus the color, a total five). These attributes can be shown in Figure 5. Changing the values according to the student's performance produces a meaningful shape evolving over time.

Attribute R_y/R_x is controlling the capability of the shape to become more or less elliptic. On the surface of the shape a sinusoidal disturbance is added. The amplitude of the disturbance is controlled by the attribute dR/R_x (we use radius R_x as reference as the ratio R_y/R_x changes), and the frequency is labeled as freq. Finally the size of the shape is represented by R_x.

In the next sections we describe how the circle's attributes values depend on the errors made by the performer. The choice of this relationship is made using intuitive criteria, in accordance with clarinet teachers' opinions.

5.1. Visualizing a Squeak Frame

When a frame is classified as a squeak, the shape is drawn as "craggy" or "rough". As more squeak a frame is, the rougher the circle should be. The rules that determinate the circle's attributes values are the following: $R_y/R_x=1$ and attribute *freq* is high valued, and increases as squeakness increases. Also dR/R_x is proportional to squeakness and R is proportional to RMS energy of the frame.

5.2. Visualizing a Hollow Frame

A Hollow note is represented as a more "flabby", "sleazy" shape, as shown in Figure 7. R_y/R_x is decreasing as hollowness increases and attribute *freq* is low valued and decreases as hollowness increases. dR/R_x is proportional to hollowness and R is proportional to RMS energy.

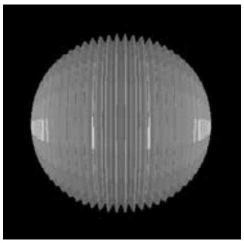

Figure 6: *The 3-Dimensional output for a squeak note*

5.3. Pitch and RMS Instability

The RMS instability is directly related with the sphere shape, because of the proportional relationship between RMS energy and *Rx*. Therefore an RMS unstable note is directly shown. Pitch instability has not been yet explored. We have the intuition that relating the pitch instability with a light change of the color of the shape will result a meaningful feedback.

5.4. Smoothing

As the hollowness/squeakness value evolve over time, sudden jumps of this value often occur. This fact results to rapid changes of the shape, making the view of the graphic annoying. To handle this problem, we deployed a smoothing on the final shape. Every attribute cannot change more than a fixed ratio between two consecutive visual frames. However this imports a tradeoff between a satisfactory and enjoyable viewing and visualizing quick, short-time errors.

In the case of offset discarding, when a consecutive decay of RMS energy is detected, the fixed ratio between the attributes values become grater.

5.5. Transforming to the 3-Dimensional Object

The rendering system takes as input the 2-Dimensional curve. The number of drawing steps used to create this object determines the quality of the generated object. Since our module is real-time and targets to low-end machines, the performance of the implementation itself should be fast and effective.

Since we want to use lights and material features to our object, one more step is actually required, to compute the *normals* for each triangle we are going to draw. After normals calculation, our object is ready to be drawn.

We are applying color, material and lighting and also (optionally) rotation to the object. The colors used in our implementation are chosen arbitrary. However, as mentioned before, there is the persuasion that associating changes in color with instability of pitch will result a meaningful output. The object is now ready for drawing as shown in Figures 6 and 7.

6. CONCLUSION AND FURTHER WORK

In this paper, a tool has been presented that employs 3D graphics to provide real-time visualization of a beginner student performance in an educational context.

Students receive simple and intuitive visual feedback that can help them improve the quality of the sound they produce, while keeping the disruption of their practice to a minimum. An important feature of the system is that it does not only provide a measure of sound quality at a given time, but also visual feedback on different aspects of the performance integrated into a simple 3D object which offers an intuitive way of understanding *what* is wrong and, to a degree, what the student must do to correct it.

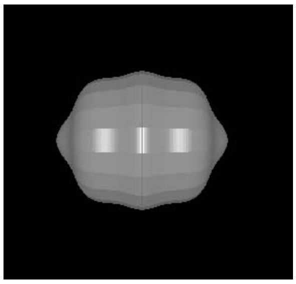

Figure 7: *The 3-Dimensional output for a hollow note*

Some first feedback from music teachers has provided clear indications that the approach is well motivated in an educational context and that if appropriately integrated into a learning setting, it may help students gain better understanding of their errors. Further feedback from music teachers, but also from students, will be collected after the visualization tool has been integrated into the overall VEMUS platform and tested with users in realistic conditions.

7. ACKNOWLEDGEMENTS

We would like to thank Ballteck for its collaboration and supply of the Real-Time Audio Recognizer tool. We also would to thank Dr. Raptis, K. Falkenberg Hansen and A. Zlatintsi for their helpful opinions during our discussions.

8. REFERENCES

[1] R. Hiraga, R.Mizaki and I. Fujishiro, “Performance Visualization: a new Challenge to Music Through Visualization”, *Proc. of the 10th ACM international conference on Multimedia*, Juan les Pince, France, Dec. 1-6, 2002, pp 239-242.

[2] R. Hiraga, F.Watanabe and I. Fujishiro, “Music Learning through Visualization”, *Proceedings of the 2nd International. Conference on Web Delivering of Music*, Darmstadt, Germany, Dec. 9-11, 2002, pp 101-108.

[3] P. McLeod and G. Wyvill, “Visualization of Musical Pitch”, *Proceedings of Computer Graphics International 2003*, Tokyo, Japan, Jul. 9-11, 2-3, p. 300-306.

[4] P. Toiviainen, “Visualization of Tonal Content with Self-Organizing Maps and Self-Similarity Matrices”,*ACM Computers in Entertainment*, Vol. 3, No. 4, Section: Music Visualization, Article 3C, pp 1-10, October 2005.

[5] S. Ferguson, A. Vande Moere, D.Cabrera. “Seeing Sound :- Real-Time Sound Visualization in Visual Feedback Loops Used for Training Musicians”, *Proceedings of the 9th International Conference on Information Visualization*, London, England, UK, Jul. 6-8, 2005, pp 97-102.

[6] www.vemus.org

[7] A. Zlatintsi, “When the Clarinet Sounds Bad – Identification Study”, M.S. Thesis in Musical Acoustics, KTH, 2006.