Performance Gestures of Musicians: What Structural and Emotional Information do they Convey?

Bradley Vines[1], Marcelo M. Wanderley[2], Regina Nuzzo[1], Daniel Levitin[1,2], Carol Krumhansl[3]

[1] Department of Psychology, McGill University
Montreal, Qc, Canada
[2] Faculty of Music, McGill University
Montreal, Qc, Canada
[3] Department of Psychology, Cornell University
Ithaca, NY, USA

Abstract. This paper analyzes how expressive gestures of a professional clarinetist contribute to the perception of structural and emotional information in musical performance. The thirty musically trained subjects saw, heard, or both saw and heard the performance. All subjects made the same judgments including a real-time judgment of phrasing, which targeted the experience of structure, and a real-time judgment of tension, which targeted emotional experience.
New statistical techniques in the field of Functional Data Analysis [13] can examine data collected from continuous processes and then explore the hidden structures of the data as they change over time.
Three main findings add to our knowledge of gesture and movement in music: 1) The visual component carries much of the same structural information as the audio. 2) Gestures elongate the sense of phrasing before an important transition and certain gestures cue the beginning of a new phrase. 3) The importance of the visual information to the experience of tension changes with loudness and note density. When loudness and note density are relatively low, the effect of removing the visual component is to decrease the experience of tension.

1 Introduction

The visual experience of a musical performance can play an important role in our appreciation and perception of music. Many of us spend considerable time and money to see live performances rather than simply listening to CDs , which offer repeatable listening and higher sound quality. The research presented in this paper investigates what the visual information in a musical performance contributes to the overall experience. We investigated the relations between modalities (auditory and visual) in conveying structural and emotional information.

Past research has shown that the movements of musicians have a close relationship with the piece being performed. Delalande [2] investigated the movements of Glen Gould and found that they were not haphazard, but reflected structural characteristics of the music. Gould's behavior at the piano changed upon moving into the studio, showing that, consciously or unconsciously, the movements of musicians can be influenced by the audience, whether through feedback or communicative intention of the performer.

Davidson [1] found that the expressive intentions of musical performers are carried most accurately by their movements. She used a point-light display to present subjects with recorded performances of classical musicians. The performers were instructed to perform in one of three ways: in a deadpan manner (without expressivity), in a standard manner (as if performing to a public audience), and in an exaggerated manner (with exaggerated expressivity). Subjects rated each performer's intended level of expressivity. Those subjects who only saw the video identified the different levels most accurately. Subjects who were denied the visual information performed more poorly at deciphering. This study shows that not only is the visual aspect of a performance not superfluous, it carries information about expressive intention, with greater resolution than the auditory component.

It has been shown that, in a ballet performance, the visual modality can convey much of the same structural and emotional information as the auditory modality. Krumhansl and Schenck [9] used a choreographed work by Ballanchine to investigate the relations between modalities for ballet. Some of their subjects only saw the dance, with sound removed. The other two groups either only heard the sound or both heard the sound and saw the dance. Subjects made four different judgments in real time: identifying section ends, identifying new ideas, a continuous judgment of tension, and a continuous judgments of the amount of emotion. The judgments were similar across modalities, especially in the coincidence of section ends, new ideas and regions of tension. Krumhansl and Schenck showed that stimulation of completely different modalities may lead to similar experiences and they opened the way for using continuous judgments in similar investigations.

Research conducted by Wanderley [16] showed that the movements of clarinetists are replicable across performances and that they are not essential to the physical execution of a piece. Wanderley used an Optotrak device to track locations on clarinetists' bodies and instruments over time, while they performed. Even though the performers were for the most part unaware of their movements, they repeated them with strong consistency from one performance of the same piece to another. Like Davidson, Wanderley instructed his musicians to use particular manners of performance. One instruction was to remain as still as possible. All four clarinetists were able to execute the pieces with accuracy, even in the absence of expressive movements. This shows that some of the movements are not essential to the physical execution of the piece, and that they may have other functions.

Wanderley's research is being used to improve the realism of synthesized musical instruments [14] [15]. By taking into account the changing relations between a virtual microphone and an electronic instrument, it is possible to model spectral fluctuations that are present in natural acoustic performances.

The research presented in this paper employed the multi-modal approach [1] [9] and real-time judgments [9] to investigate a musical performance. Relations between the audio and visual modalities are explored in terms of the structural and emotional information they convey. A continuous judgment of tension was used to observe the experience of emotion and a continuous judgment of phrasing targeted the experience of structure.

The continuous tension judgment, pioneered by Nielsen [12], has been shown to be correlated with continuous judgments of dominant emotions and a variety of physiological measures [8]. Krumhansl and Schenck [9] found that subjects' judgments of tension are similar to their judgments of the amount of emotion. The tension judgment is sensitive to changes in a performance [3], and it depends on a wide variety of structural, harmonic, and rhythmic features in the music [12] [10] [3] [7] [9]. This measure has proven consistent across age groups, musical skill levels, and familiarity with the stimulus music [4] [5] [6] [7]. It is an informative and consistent measure.

The continuous phrasing judgment was exploratory, in that it had never been used in the same way before. However, similar real-time judgments have been used with success [7] [8]. In these studies, beginnings and endings of musical ideas were marked as the performances were presented. In the present investigation, the shape of the phrase was included as well.

Phrasing is a structural feature. A musical phrase is analogous to a complete idea in speech. A judgment of phrasing captures the sense of form, as opposed to content.

We used new statistical techniques in Functional Data Analysis [13] to reveal changes in the effect of having either audio or visual information removed from a subject's experience. Traditional statistics have limitations when they are used with data that are sampled over time. Traditional methods produce one-dimensional descriptions that are incapable of revealing changes over time. For example, correlations and regression models return a single number to summarize the relation between entire data sets. In the experiment presented here, judgments were performed over a time span of 80 seconds, which encompassed a total of 800 data points per subject. Subjects were responding to a musical piece with great variation in dynamics, expressive movements, and harmony. It would have been an oversimplification to reduce all of the information gathered to a one-dimensional description of the relations across treatment groups. Traditional statistical techniques also ignore correlations between judgments over time. That is, they assume that all observations are independent, and that the measurement at second three has nothing to do with the measurement at second four. This would be a false assumption for the data collected in this experiment.

Functional Data Analysis, however, yields solutions that are themselves functions of time. These new statistical techniques treat data as mathematical functions. So, it is possible to ask questions like "When during the performance does removing the visual component have a strong impact on the tension judgment, as compared to the natural experience?" We only used these techniques with the tension data, due to challenges in interpreting the phrasing data in functional terms.

This paper is part of a larger investigation involving multiple performers and multiple manners of execution. We will concentrate on a standard performance of one clarinetists here.

2 Method

2.1 Stimuli:

The stimuli consisted of a performance by a professional clarinetist of Stravinsky's second piece for solo clarinet. The performer played the piece as if presenting to a

public audience. The recording was originally created for an investigation mentioned above [16].

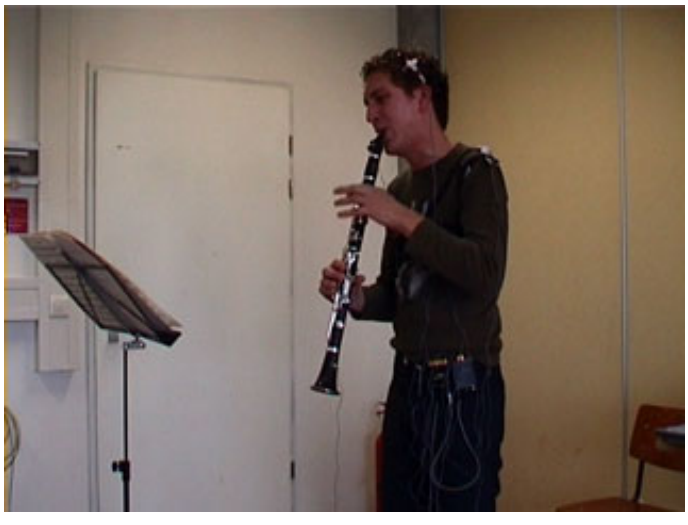

Fig. 1. A screen shot of the stimuli video used in the experiments, from [16].

We chose the Stravinsky piece for three reasons: 1) It is an unaccompanied piece, so the performer's movements and sound were not influenced by another instrument. 2) The piece is rhythmically free; it has no underlying pulse or meter. It has been shown that performers tend to entrain their movements to the underlying pulse when there is a consistent meter [16]. In the Stravinsky piece, the musicians were free to move expressively and idiosyncratically. 3) This music is standard repertoire for advanced clarinetists across cultures. This makes replication and cross-cultural comparisons accessible for future work.

2.2 Subjects:

Thirty subjects participated in the experiment. All of them had at least five years of musical training. This criterion ensured that subjects had a developed ear and an understanding of basic terms in music, like "phrasing."

Using a Between-Subjects design, the subjects were randomly divided into three equally sized treatment groups. The first group experienced the performance in a "Natural" way, with both audio and visual information intact. The second group heard the performance with the visual component removed, and the last group saw the performance with the audio removed.

The performance was shown with digital-video quality (25 frames per second) on a G4 desktop Macintosh.

2.3 Tasks:

All subjects performed the same real-time tasks, including a continuous judgment of tension and a continuous judgment of phrasing. Both judgments were made using one track on a Peavy 1600X slider. Subjects moved the slider up and down along a track as the performance was presented. The location of the slider was sampled once every 100 milliseconds. Continuous judgments reveal a subject's mental state more accurately than retrospective judgments, which are subject to memory errors.

Tension These are the exact instructions that subjects read before performing the task:

> - Use the full range of the slider to express the TENSION you experience in the performance. Move the slider upward as the tension increases and downward as the tension decreases.

Phrasing Here are the exact instructions read by each subject:

> - Use the full range of the slider to express the PHRASING you experience in the performance. Move the slider upward as a phrase is entered and downward as a phrase is exited. The slider should be near the top in the middle of a phrase and near the bottom between phrases.

3 Results

3.1 The Tension Data:

The effects of having visual or audio information removed from a subject's experience change over time, as revealed by Functional Data Analysis.

Figure 2 shows the fitted curves for each of the treatment conditions. These functions are essentially an average of the subjects' judgments in each condition, after taking into account their unique reaction times and the range of slider that they used. The curves have been smoothed to eliminate high frequency noise in the data.

When we apply a linear model to the data functions[1], the coefficients shown in Figure 3 emerge. The dotted line shows the effect of having the audio removed and the dashed line shows the effect of having the visual component removed. When either line is above the $y = 0$ axis, the effect of being in that particular treatment group is to increase the judgment of tension in comparison to the average judgment in the Natural condition.

[1] The following linear model was used:

$$Y = U + B1(t)[if minus audio] + B2(t)[if minus video] \quad (1)$$

where U is the natural function, and $B1(t)$ and $B2(t)$ are coefficients that change with time.

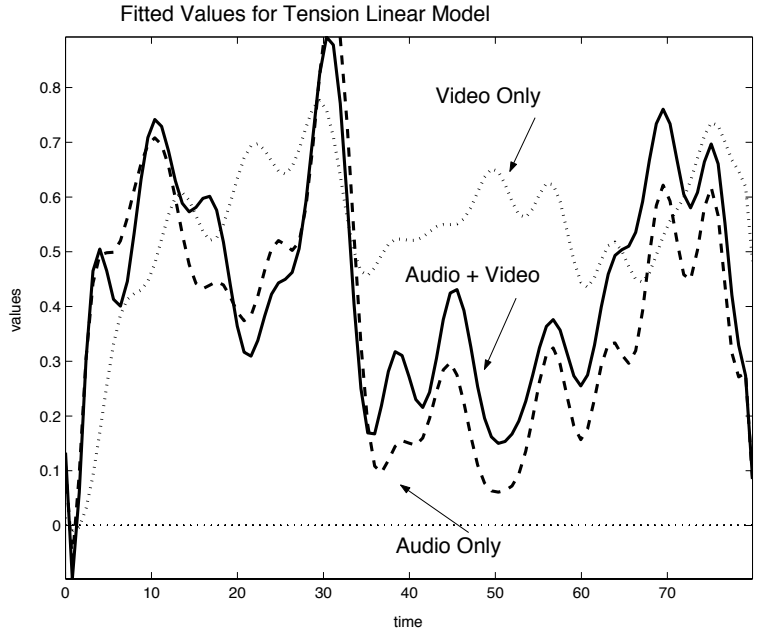

Fig. 2. The fitted curves for each treatment group after registering, scaling and smoothing. Registering takes into account varying reaction times across subjects. It is possible that two subjects, who are responding to the same physical event, might give responses at different points in time, due to variations in their reaction times. Registering, also known as *Time Warping*, takes this possibility into account by aligning major peaks and troughs in the judgments. Scaling corrects for the different spans of the slider that subjects use, and smoothing irons out the high frequency wobbles in the curves.

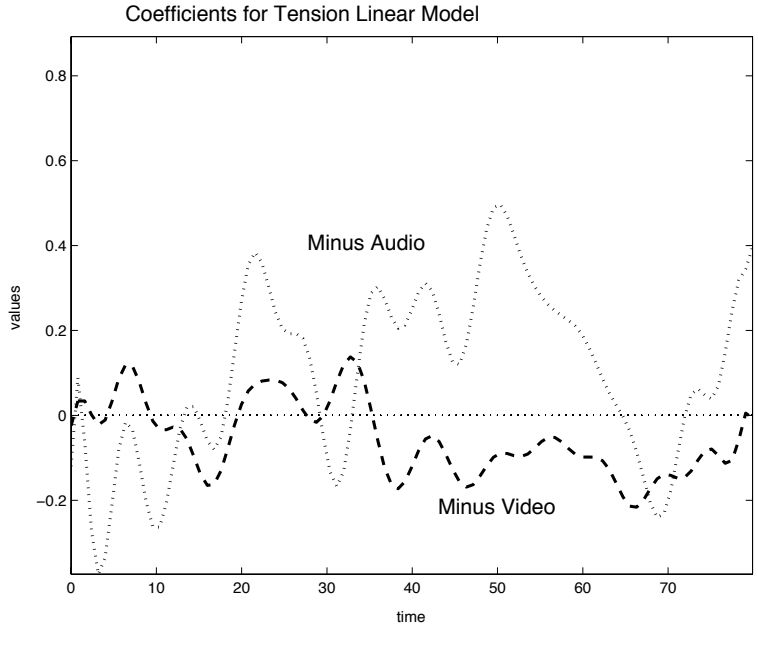

Fig. 3. A graph of the coefficients for two treatment groups over time. These are products of new statistical methods, *Functional Data Analysis*, which treat data as functions of time. The dashed line shows the effect of being in a group for which the auditory component was removed. The dotted line shows the effect of being in the group for which the visual component was removed. Values above the $y = 0$ line represent an effective increase in the judgment of tension.

The graph of the functional coefficients drew our attention to the region from about 35 to 65 seconds. During this period of time, the effect of removing audio remains strongly and consistently positive. The effect of removing the visual component is correspondingly negative for most of the region. The corresponding section in the performance encompasses the middle section in the piece. During this section, the dynamics decrease dramatically, from a mezzo forte to pianissimo, and the note density decreases as well. From this we can see that the importance of video information to the typical tension judgment is associated with loudness and note density.

3.2 Phrasing Data:

Structural Content of the Performance Gestures The raw phrasing averages for the three treatment groups, shown in Figure 4, tell a very different story from the tension averages.

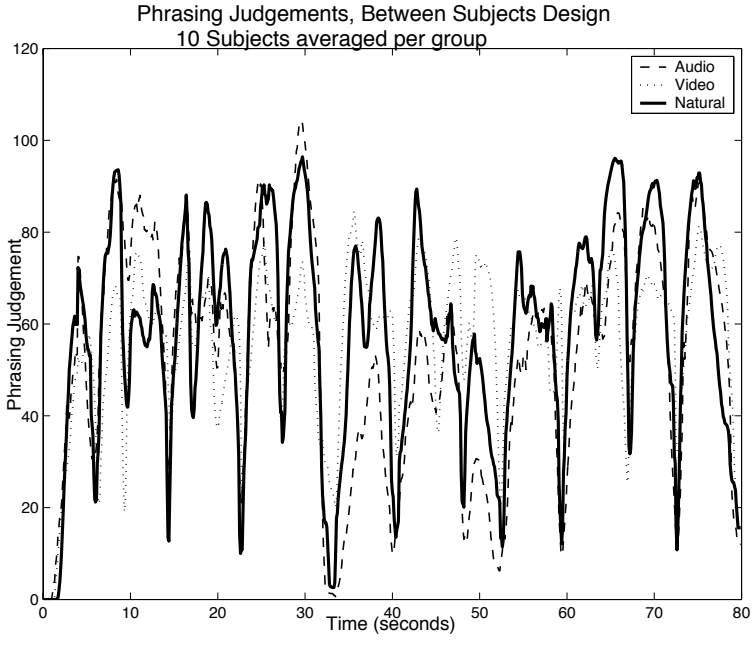

Fig. 4. Graphs of phrasing judgments. Judgments of the ten subjects from each treatment group were averaged to create these lines. The coincidence of troughs and peaks across treatment conditions shows that the visual aspect of the performance is also carrying structural information about the piece.

There is a striking similarity between the judgments for all three groups, in spite of the fact that the Audio-only group and the Visual-only group had no overlap in stimulus

material. Subjects in the Audio-only group had all of the visual component removed. Subjects in the Visual-only group had all the auditory component removed. Yet their judgments show that both modalities express very similar phrasing information. The magnitude of judgments varies from group to group, but the troughs and peaks align consistently. These data show that the visual component of a musical performance also conveys the structural content of a piece.

The Effect of Gestures During the Major Transition Figure 5 shows a detail of the only major transition between sections. During this transition, a fermata ends the first section and there is a pause and a breath before the new section is entered.

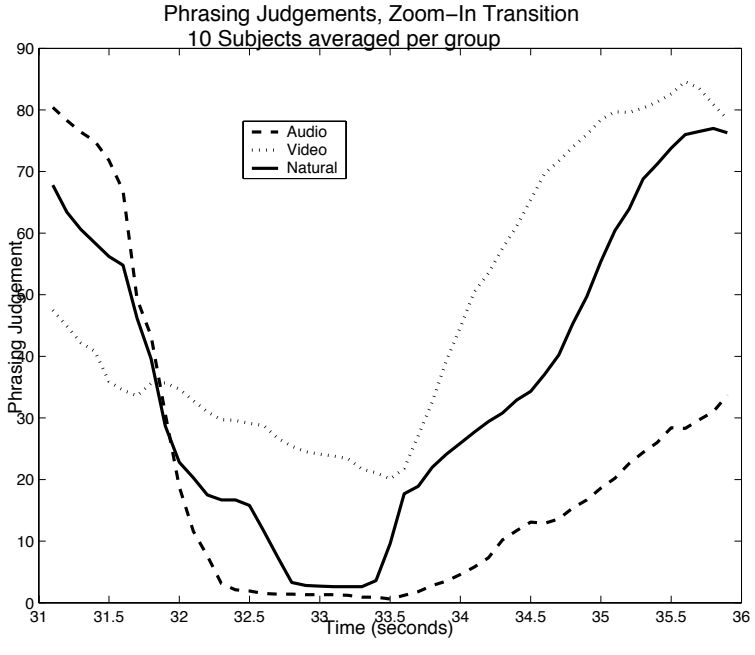

Fig. 5. A zoom-in on the transition between musical sections. The Visual-only (*dotted line*) is slowest to recognize the end of the phrase and the Audio-only (*dashed line*) group is slowest to recognize the beginning of the new phrase.

The zoomed-in region of the raw data averages shows that the Visual-only group is slow to recognize the end of the phrase and the Audio-only group is slow to recognize the beginning of the new phrase. The visual component has an important effect on the sense of phrasing in this segment.

The performer ends the fermata with a strong movement of his body. His left hand raises from the clarinet and then slowly descends during the silent rest. We hypothesize that gestural movement during the pause extends the sensation of the phrase for the

Visual-only group. These subjects did not have the benefit of hearing the note come to a clear and concise end. The sound is important for recognizing the conclusion of the phrase and the visual information elongates the experience.

Before the new phrase begins, the performer makes some movements in anticipation of the sound. He takes a deep breath, raises the clarinet and brings it back down in a swooping motion before initiating the sound at the bottom of his arc. Without these visual cues, the Audio-only group can not anticipate the onset of sound. Subjects in the Audio-only group have a sense of phrasing that lags behind the other two groups who are privy to the movement cues that anticipate the coming sound. The visual information is important for engaging the experience of a new phrase.

There are corresponding phenomena in speech. Preparation of the vocal articulators begins before the speech sound is initiated. This visual cue helps an observer to anticipate the onset of a new speech sound and gives some information about what that sound will be. Also, breathing cues are used by people engaged in conversation to help in timing their exchange [11]. In a musical situation, certain movements cue the beginning of a new phrase, including breath and body gesture.

We hypothesize that for important transitions that involve a pause in the sound, the visual information has strong effects that extend the sensation of the phrase being concluded, and cue the beginning of the new phrase with particular gestures.

4 Conclusion

We found Functional Data Analysis methods to be very useful for revealing the relations between modalities as they change over time. These statistical techniques are important tools for the fields of gesture, music, and emotion research. They reveal the hidden structures of data and how those structures change over time. Functional techniques are useful for analyzing continuous judgment data as well as measures of gestures and movements themselves (i.e. Optotrak data).

Using Functional Data Analysis, we have observed that the importance of the visual component to the typical tension judgment is dependent upon the loudness and the note density in sound. When the note density is low and the dynamics are soft, the effect of removing the visual component is to decrease the judgment of tension.

The visual modality was shown to carry much of the same structural information as the auditory modality, as indicated by similarities in the phrasing judgments across treatment groups. Despite the fact that the Audio-only group and the Visual-only group had no physical overlap in their stimuli, their judgments of phrasing showed precise synchrony in the onsets and conclusions of phrases.

During the strongest transition in the piece, where a pause in sound occurs, the clarinetist's gestures have a marked effect on the sense of phrasing. The group that is denied auditory information shows a lag in its recognition of the phrase ending because body movements continue during the pause. The movements elongate the sense of phrasing. There is no such ambiguity in the audio component. When the note ends, there is only silence to clearly mark the phrase end. During the major transition, gestures serve to cue the beginning of a new phrase. Without visual information, the sense of phrasing lags behind after the long pause in sound. The Audio-only group did not experience the

gestures that cue the beginning of the new phrase, including a substantial breath and a swooping motion of the clarinet.

This research augments our understanding of multi-modal relations in a musical performance and sheds light upon the important involvement of performance gestures in the perception of music.

References

[1] Davidson, J. Visual Perception of Performance Manner in the Movements of Solo Musicians. *Psychology of Music* 21: 103 - 113, 1993.

[2] Delalande, F. La Gestique de Gould. In *Glen Gould Pluriel*, pages 85-111. Louse Courteau, editrice, inc., 1988.

[3] Fredrickson, W. E. A Comparison of Perceived Musical Tension and Aesthetic Response. *Psychology of Music* 23: 81-87, 1995.

[4] Fredrickson, W. E. Elementary, Middle, and High School Student Perceptions of Tension in Music. *Journal of Research in Music Education* 45(4): 626-635, 1997.

[5] Fredrickson, W. E. Effect of Musical Performance on Perception of Tension in Gustav Holst's First Suite in E-flat. *Journal of Research in Music Education* 47(1): 44-52, 1999.

[6] Fredrickson, W. E. Perception of Tension in Music: Musicians versus Nonmusicians. *Journal of Music Therapy* 37(1): 40-50, 2000.

[7] Krumhansl, C.L. A Perceptual Analysis of Mozart's Piano Sonata K. 282: Segmentation, Tension, and Musical Ideas. *Music Perception* 13(3): 401-432, 1996.

[8] Krumhansl, C. L. An Exploratory Study of Musical Emotions and Psychophysiology. *Canadian Journal of Experimental Psychology* 51(4): 336-352, 1997.

[9] Krumhansl, C. L. S. and Schenck, D.L. Can dance reflect the structural and expressive qualities of music? A perceptual experiment on Balanchine's choreography of Mozart's Divertimento No. 15. *Musicae Scientiae* 1(Spring): 63-85, 1997.

[10] Madsen, C. K. and Fredrickson, W. E. The Experience of Musical Tension: A Replication of Nielsen's Research Using the Continuous Response Digital Interface. *Journal of Music Therapy* 30(1): 46-63, 1993.

[11] McFarland, D.H. Respiratory Markers of Conversational Interaction. *Journal of Speech, Language, and Hearing Research* 44(1): 128-143, 2001.

[12] Nielsen, F.V. *Oplevelse af misikalsk spending* (The experience of musical tension). Copenhagen: Akademisk Forlag, 1983.

[13] Ramsay, J.O. and Silverman, B.W. *Functional Data Analysis*. New York: Springer-Verlag, 1997.

[14] Wanderley, M. M. Non-Obvious Performer Gestures in Instrumental Music. In A. Braffort, R. Gherbi, S. Gibet, J. Richardson, and D. Teil (eds.) *Gesture-Based Communication in Human-Computer Interaction*. Berlin, Heidelberg: Springer Verlag, pages 37-48, 1999.

[15] Wanderley, M. M. and Depalle, P. Gesturally Controlled Digital Audio Effects. In *Proceedings of the COST-6 Conference on Digital Audio Effects (DAFx-01)*. Limerick, Ireland, pages 165-169, 2001.

[16] Wanderley, M. M. Quantitative Analysis of Non-Obvious Performer Gestures. In I. Wachsmuth and T. Sowa (eds.) *Gesture and Sign Language in Human-Computer Interaction*. Berlin, Heidelberg: Springer Verlag, pages 241-253, 2002.