A Conceptual Framework for Motion Based Music Applications

Marcella Mandanici* Antonio Rodà† Sergio Canazza‡

Dept. of Information Engineering
University of Padova, Italy

ABSTRACT

Imaginary projections are the core of the framework for motion based music applications presented in this paper. Their design depends on the space covered by the motion tracking device, but also on the musical feature involved in the application. They can be considered a very powerful tool because they allow not only to project in the virtual environment the image of a traditional acoustic instrument, but also to express any spatially defined abstract concept. The system pipeline starts from the musical content and, through a geometrical interpretation, arrives to its projection in the physical space. Three case studies involving different motion tracking devices and different musical concepts will be analyzed. The three examined applications have been programmed and already tested by the authors. They aim respectively at musical expressive interaction (*Disembodied Voices*), tonal music knowledge (*Harmonic Walk*) and XX century music composition (*Hand Composer*).

Index Terms: H.5.2 [User Interfaces]: Auditory feedback—Interaction styles; H.5.5 [Sound and Music Computing]: Methodologies and techniques—Systems

1 INTRODUCTION

Handheld devices, virtual reality environments, touch user interfaces and smartphones play a fundamental role in the design of applications based on the so-called post-WIMP interfaces [4]. These have opened a completely new perspective on human-computer interaction because they bring in all the expressive power of reality, spatial relationships and user's long-life experience of the physical world.

More recently, a new generation of motion tracking devices like Kinect[1] and Leap Motion[2], as well as simple camera sensors, allow the user to interact with digital contents through free-handed and full body motion. This represents a further step towards the evolution of reality-based interaction styles, because now we face an interaction range much larger than before. As a matter of facts, handheld devices and smartphones represent reality-based interfaces, but they are operated in a few inches ranges and mainly through fingers' interaction. The possibility of moving into wider spaces with free limbs and body motions and with the additional option of tracking the user's position, offers even further cues and different interaction qualities. Naïve physics, body and environment awareness and skills as well as social relationships, are the main themes spotted by [4]. Now we can add also the orientation, wayfinding, route storage, spatial navigation, ego-location, kinesthesia and proprioception.

1.1 Related Work

A lot of musical applications have been developed since the date of the commercial launch of Microsoft's Kinect sensor, and later for the Leap Motion sensor, while probably millions of music apps are available for various kinds of smartphone systems.[3] The general tendency of all these applications has been to digitally reproduce both already existing music player's interfaces or traditional acoustic instrument features.[4] These last have been created in a more or less faithful way through imaginary projections of the real instrument model in the space around the performer.

In [3] a guitar model is described in a virtual scene, where subsequent blocks positioned on the guitar's neck represent the different chords, while another area at the center of the guitar body is used to trigger the sound. In [2] three more abstract models of percussion, guitar and melody playing are proposed with the aim of gathering an all virtual instruments ensemble. S. Şentürk and al. in [13] proposed a Kinect operated composing system where chord progressions, rhythm and timbre are manipulated by the performer through hand gestures. Many trials of building virtual instruments with the Leap Motion sensor are also reported in [1], where the great sensor's latency and some confusion in the detection of the different finger position seem yet to be the greatest obstacle in obtaining virtual instruments which can really fit the musical performance.

1.2 Aim of the Paper

In [3] the main difference from the real model is that in the guitar the player selects the chords through different fret finger positions, while in this digital reproduction a series of simplified chord blocks are proposed. It is clear that we are facing a trade-off solution between reality and its virtual environment projection, which can fit well from the computational point of view. Anyway, what it is important to point out is that in designing such kind of applications we always need to render some aspects of the reality by interpreting or summarizing their functionalities. And, as this rendering happens in a virtual environment, we need to give a geometrical interpretation of it through an imaginary projection. The aim of this paper is to propose and discuss a general conceptual framework for motion based music applications which has at its core the geometrical interpretations of the musical concepts we need to represent. In Section 2, the framework is presented and the main theoretical and spatial concepts upon which the motion based music applications are built are explained. Afterwards, we examine three case study of motion based application design. In Subsection 3.1 the Kinect based *Disembodied Voices* application [8], aimed at expressive musical interaction, is presented. Subsection 3.2 shows the projections

*e-mail: mandanici@dei.unipd.it
†e-mail:roda@dei.unipd.it
‡e-mail:canazza@dei.unipd.it

[1]Kinect (`http://en.wikipedia.org/wiki/Kinect`) is a motion sensing input device launched by Microsoft in the autumn of 2010. The system can interpret specific gestures using an infrared projector and camera and a special microchip to track the movements of individuals in three dimensions.

[2]The Leap Motion Controller (`https://www.leapmotion.com/`) is a small USB peripheral which is placed in front of the laptop. The device scans a region in the shape of an inverted pyramid centered at the devices middle point and expanding upwards for about 60 cm (2 feet).

[3]`http://www.digitaltrends.com/mobile/best-music-apps/`

[4]A Kinect virtual piano at `http://hacknmod.com/hack/diy-virtual-piano-using-kinect/`

on a flat surface of the *Harmonic Walk*, a step operated application for the accompaniment of a tonal melody [7]. At the end (Subsection 3.3) we present the *Hand Composer*[6], a tool for interactive, gesturally driven, XX century music composition.

2 The Conceptual Framework for Motion Based Music Applications

The framework we want to present and discuss is depicted in Fig. 1. The figure is subdivided into two main areas: the first is the theoretical framework area (in gray), and the second is the interaction framework area (in blue). In the picture there is a third area, containing the "Projection into a Physical Space" and the "Mapping in the Acoustical Space" modules. This is a connection area which derives from the superposition of the two previous cited areas, and which belongs to both of them.

The design process starts from the *Information* module (theoretical framework area), which contains the musical content upon which the application is based. This can be the model of a traditional acoustic instruments, as in the above mentioned example [3]. Indeed, the new generation of motion tracking devices allows much more than reproducing traditional instruments, in that they can link the real human body surroundings to the digital world, creating so the basis for true sound augmented reality environments. If this is the perspective, there is no need to reproduce already known and well experimented instrumental shapes and gestures, or, at least, this can be the first approach but it doesn't seem to be the most interesting one. As a matter of facts, the musical information can be the representation of an abstract musical concept, or new spatial arrangements of traditional instrumental features or completely new virtual instruments at all. In any case we need to have a precise representation of the musical content because, from this, we have to extract a valid *Geometrical Interpretation*. It is important to underline the word "interpretation" because in this lies the novelty and the interest of this new approach to musical production. In this phase, we need to focus the musical functionality we want to work out and we have to choose what is the main aspect to be outlined. Furthermore, we have to produce a spatial (geometrical) arrangement of this knowledge, which has to follow precise usability and learnability qualities. In the case of abstract musical concepts, a necessary link to theoretical knowledge and representations has to be developed in order to guarantee musical coherence and suitability. This aspect is particularly important when the application design is not restricted to a personal use (i.e. a composer's artistic production), but when it aims at social or communication purposes, like in educational environments. Here, the above cited human-computer interaction features must be further tested in order to assess the application's efficiency in conveying an actual musical knowledge to its users.

The green module, *Projection into a Physical Space*, is the connection module between the theoretical framework and the interaction framework area. It represents the core of the conceptual framework, because here lies the actual application's interface. The projection can be in the 3D space, employing the *x, y* coordinates plus the *z* depth data, or the spherical-polar coordinate system, centered on the torso of the user. Also 2D surfaces can be used along the transverse, sagittal or frontal plane, or along any oblique plane between two of these, or, again, on a physical floor. The projection can be marked by visual tags or by other kinds of feedback, among which, for the musical applications - of course - the auditory is the most important. Going on in the connection area, we meet the *Mapping in the Acoustical Space* module, which is responsible of the audio output of the system and, consequently, of the *Audio Feedback*. We are now in the interaction framework area where the dashed arrows, starting from the *Interaction* module, show the classical interaction loop.

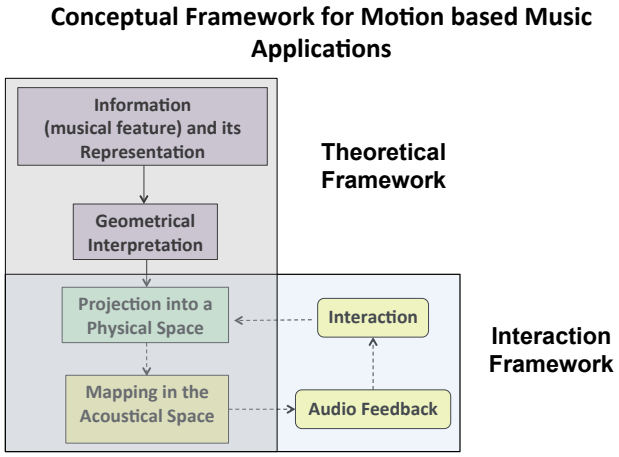

Figure 1: A conceptual framework for motion based music applications, with the theoretical framework area (in gray) and the interaction framework area (in blue). The green module (*Projection into a Physical Space*) represents the connection module among the two areas.

3 Three Case Studies

In this Section we present three case studies which will be discussed on the basis of the conceptual framework above described. The three cases employ three motion tracking devices and convey three different musical concepts through spatial representation and geometrical interpretation. Also the projections refer to different spatial models, as the first employs the 3D spherical-polar coordinate system, the second 2D Cartesian coordinates on a flat floor surface and the third 3D space with x, y Cartesian coordinates plus z plane for depth data.

3.1 "Disembodied Voices"

Disembodied voices[5] is an interactive environment designed for an expressive, gesture-based musical performance. The motion sensor *Kinect*[6], placed in front of the user, provides the computer with the 3D polar coordinates of the two hands. The application is designed according to the metaphor of the choir conductor: the user, through gestures, is able to run a score and to produce a real-time expressive interpretation. The software interprets the gestural data and controls articulated events to be sung and expressively performed by a virtual choir.

- **The Musical Information**. The conductor moves her/his arms and hands in the space around her/his torso and in the direction of the singers or instrumental players. Movement analysis ([9] and [10]) as well as academic teaching [12] subdivide the role of the two hands. In general the right executes musical cues while the left is devoted to iconics, metaphorics and dynamics. This model has been considered the musical information upon which to base the application's design.

[5] https://www.youtube.com/watch?v=oyf7GrMMrL8

[6] *Kinect* is a motion sensing input device launched by Microsoft in the autumn of 2010. It appears as a horizontal black bar connected to a motorized base and it consists of three devices: an RGB camera, a depth sensor and a multi-array microphone. The system can interpret specific gestures using an infrared projector and camera and a special microchip to track the movements of individuals in three dimensions.

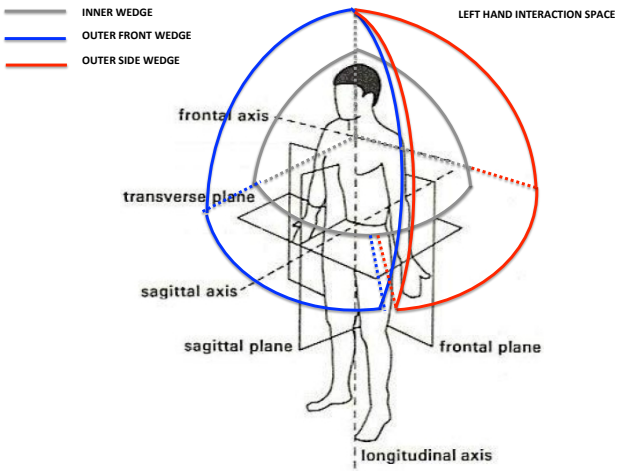

Figure 2: The left hand interactive 3D space of the *Disembodied Voices* application's user. The space is subdivided into three zones: the gray inner wedge, the blue outer front wedge and the red outer side wedge.

- **The Geometrical Interpretation**. The geometrical interpretation of the conductors interaction space is a nearly spherical wedge with the center at the basis of the neck of the conductor and the diameter corresponding approximately to the two stretched arms' length. The base of the wedge lies in the transverse plane, while the vertical ray is along the sagittal plane. Following the conductors interaction model, the wedge is subdivided in two parts, one for the right and the other for the left hand. The left hand wedge is depicted in Fig. 2, where three different wedge portions are shaped. This is a subjective interpretation of the conductor's functions, similar to that already outlined in the virtual guitar presented in [3]. The actual conductor employes hand gestures, glances, eyebrow movements and a lot of mimics to communicate her/his intentions to performers. In this computational model we have to transfer something of the conductor's communicative power into her/his left hand interaction space partition. Thus we design an inner wedge space where the hand position affects only the dynamic level of the sound. As soon as the hand enters the outer wedge beyond the dynamic level also other two digital sound processes are triggered, one in the outer front wedge and the other in the outer side wedge.

- **The Projection into the Physical Space** The 3D space of the spheric wedge is mapped through the spherical-polar coordinate system, where the arm's length is represented by r (the ray), the lateral movements angle by ϕ (the azimuth) and the elevation angle by θ (the zenith). Therefore the inner wedge is determined by a threshold along the ray length, the outer zones are delimited by azimuth thresholds, while the dynamic level depends on the elevation angle.

3.2 "Harmonic Walk"

The *Harmonic Walk*[7] is an interactive physical environment designed for experiencing a novel spatial approach to musical creation. In particular, the system allows the user to get in touch with some fundamental tonal music features in a very simple and readily available way. The application's interface consists of a camera placed on the ceiling which can trace the presence of a user who walks on a flat surface within the camera's view.[8] The *Harmonic Walk*, through the body movement in space, can provide a live experience of tonal melody structure, chord progressions, melody accompaniment and improvisation. Enactive knowledge and embodied cognition allow the user to build an inner map of these musical features, which can be acted by moving on the mapped surface with a simple step.

- **The Musical Information**. As soon as a listener is presented a tonal melody, s/he first tries to interpret the sequence of notes, grouping them after a metrical and harmonic frame [11]. This produces a segmentation of the composition into different harmonic regions which, in case of one key melody, shall all belong to the same tonality. Therefore, melodic segmentation and the underlying harmonic structure are the leading features of a tonal composition.

- **The Geometrical Interpretation**. The time proceeding of the various musical units is led by the melody, whose metaphoric scheme is expressed by the so-called "source-path-goal" schema [5]. Following this metaphor and imagining the simplest motion in space a human can do - the walk - we could represent the tonal composition as a sequence of spatial blocks, where each step corresponds to the next musical unit. As far as concerns the harmonic space, we start from the classical spatial representation of the *tonnetz*[9] as the leading spatial model, and we define the first six roots of the major scale as the one tonality major melody harmonic space. Anyway, given the impossibility of the *tonnetz*'s spatial model to allow movement from one chord to the other without touching some other chord, we need to interpret the geometrical form of the six available musical chords, expanding the original *tonnetz* grid to obtain an inner empty zone, which can allow the transition to the other remaining five chords. Thus, we arrive to a six parts sliced circular ring projection where the six roots are conveniently displayed. At the end, the geometrical interpretation of the two outstanding tonal melody's features to be used in the *Harmonic Walk* application are a straight line and a circular ring.

- **The Projection into the Physical Space**. This two geometrical features are laid on the actual surface, respectively at the borders of the mapped area (the straight line) and at the centre (the circular ring), as shown in Fig. 3. The user's paths are identified through visual tags which are positioned at the center of each zone. The first is the one corresponding to the straight line and is marked with the white crosses, while the second is the circular one, marked with the black crosses. The beginning of each path is marked with an arrow. In this way the user has a visual cue of the center of the various regions and is guided through them by his inner feeling of the music structure.

[7] https://www.youtube.com/watch?v=c4ru468eqM0&list=UU1E9xCq8TWqlzessRIzUGxw

[8] The video application called *Zone Tracker*, tracks the user's position by analyzing the input images and by calculating the user blob's barycenter. It employs a series of masks to subdivide the mapped surface, allowing to detect the portion of space occupied by the user.

[9] Literally "web of tones". The *tonnetz* appeared for the first time in a harmony treatise written by Euler in 1739. In the 19th century the *tonnetz* was reviewed and discussed by the German music theorist Hugo Riemann.

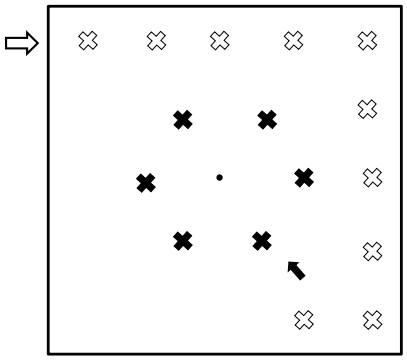

Figure 3: Visual tags of the straight and circular path of the *Harmonic Walk*. The straight path (white crosses) follows the interface perimeter and displays a 11 units musical sequence, while the circular path (black crosses), centered with respect to the application's surface, displays the six harmonic space chords distribution.

3.3 "Hand Composer"

The *Hand Composer*[10] is a gesture-driven composition system, based on the analysis of the existing relationships between music generative models and musical composition in the context of the XX century music history background. The system framework is based on a number of interactive machines performing various patterns of music composition and producing a stream of MIDI data to be compatible with a Disklavier performance.[11] Hand gestural input, captured by the Leap Motion controller[12], can control some parameters of the music composing machines, changing interactively and in real time their musical output.

- **The Musical Information**. In the XX century, particularly after the II world war, a new way to think musical composition was established. Starting from A. Schoenberg's idea of serial composition[13], many XX century composers begun to develop a new compositional approach where musical events, similar to many physical phenomena, depend on the control of parameters like density, range, frequency, distribution and so on. This way of composing allows a direct link to algorithmic composition, where the composing machines need to be clearly determined and fed by the model's parameters. We outline some of these models which we define as the author's fingerprint and which make up our model's musical information.

- **The Geometrical Interpretation**. As an example, the Ligeti's model consists of a band of pitches with various starting points, ranges and densities. The algorithmic model of a shaped composition of random events has been named "tendency mask" technique and has been theorized by G.M. Koenig in 1966.[14] It is defined by two envelopes for lower and higher boundaries, with at least another two controls for values distribution and density (probability tables, random walks, etc.). The geometrical interpretation of a band of pitches may look like the one depicted in Fig. 4.[15] Nevertheless, the Ligeti music composition machine is actually simpler than the original model.

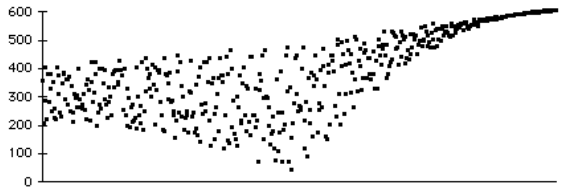

Figure 4: The tendency mask of a shaped random event.

In fact, firstly, no random event distribution information is provided by the system; secondly, the events' density depends on a preloaded metronome value which, in the present implementation, has no possibility of being interactively changed.

- **The Projection into the Physical Space**. The imaginary projection of the pitch band is the result of the hand movement in the 3D applications physical space depicted in Fig. 5. The different inclination of the hand shifting along the x axis allows the system to record a pitch minimum and maximum, which are respectively interpreted as the lower and the high boundary of the band. Moving the hand upwards to downwards causes the system to record a new boundary, shaping in this way the band's envelopes.

4 Conclusion and Further Work

In this article a conceptual framework for motion based music applications has been presented, with a particular emphasis on spatial projections. These are very strong tools, as they can make immediately available any imaginary interface useful to represent both concrete (e.g. acoustical instrument) and abstract (e.g. harmonic space) musical information. The more they are sticking to the related musical concept, the more the actual application's interface will be efficient in conveying meaning to the user. Another important point outlined from case study analysis is that geometric interpretations always imply some difference from the original model, due to computational optimization or to extended or reduced application's functionality. These differences must be carefully considered to understand the nature of the possibilities offered by imaginary projections, which, from one side, subtract something, but, from the other, add much more to the virtual environment. A last consideration is about the different nature of imaginary projections employed in the three analyzed cases. *Disembodied Voices* is an environment devoted to expressive interaction. The musical information about the conductor's movements is very rich in gestural data analysis, but these scarcely concern the hand position in the 3D space. Moreover, very often conductors employ deictic gestures, which acquire meaning de-

[10] https://www.youtube.com/watch?v=mdsn9_5Ig_A&list=UU1E9xCq8TWqlzessRIzUGxw

[11] The Disklavier is a standard acoustic piano, except that it can also employ electromechanical solenoids to move key and pedals independently of any human performer. Thus the Disklavier can be played in the traditional way, but can also be controlled by MIDI messages sent from a computer through a USB cable or stored in memory units or CDs.

[12] The Leap Motion controller is a small USB peripheral which is placed in front of the laptop. The device scans a region in the shape of an inverted pyramid centered at the devices middle point and expanding upwards for about 60 cm (2 feet).

[13] In serial composition the musical parameters are not disposed following the traditional melodic and tonal harmony frame, but rather after a series of values which rule pitches, note durations, dynamics and even timbre.

[14] http://www.koenigproject.nl/indexe.htm

[15] The picture is taken from http://www2.ak.tu-berlin.de/~abartetzki/CMaskPaper/cmask-article.html

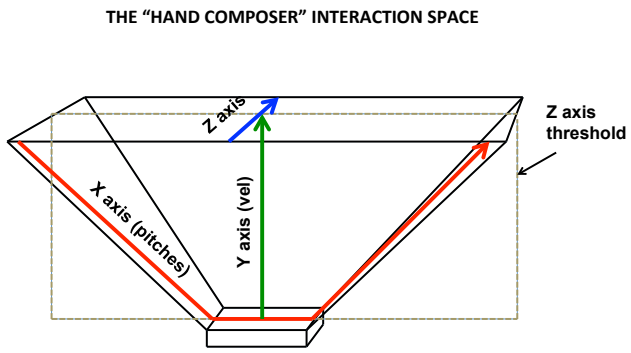

Figure 5: The 3D interaction space of the *Hand Composer* application. x,y Cartesian coordinates are employed to map the two-dimensional vertical plane, while z coordinates map the depth data.

pending of the musician they address to and which are a nonsense in a virtual environment, where all gestural and position data are interpreted by a machine. Thus, in this case, there is a wide space for extended functionalities and meaning. In the *Harmonic Walk* environment the relationship between musical information and geometrical representation is much more defined. Here, we deal with a precise music compositional structure and, moreover, with an already well established spatial representation of harmonic relationships. So, our contribution is limited to a spatial optimization of melodic and harmonic musical features. The case of *Hand Composer* is again different, as here we interact with the actual music production. The music composing machines not only require an imaginary projection of the author's "fingerprint" in the physical space, but, being subject to time proceeding, create ever changing, evolutionary events images to be constantly followed by the user.

The conceptual framework discussed in Section 2 is very general and can be applied in many sensor-controlled environments. Anyway, it is a good starting point to begin to study how the user reacts in a virtual environment where only audio feedback, with no haptic and/or visual cues, is provided.

The above described applications can be used not only for artistic production, but also as a basis for learning environments. Thus, an extensive research in the field of cognitive sciences and human-computer interaction is required to understand how these environments influence the user and how to deploy their unique features to help a more direct and effective knowledge transmission.

REFERENCES

[1] J. Han and N. Gold. Lessons learned in exploring the leap motion tm sensor for gesture-based instrument design. *Proc. NIME'14*, 2014.

[2] M.-H. Hsu, W. Kumara, T. K. Shih, and Z. Cheng. Spider king: Virtual musical instruments based on microsoft kinect. In *Awareness Science and Technology and Ubi-Media Computing (iCAST-UMEDIA), 2013 International Joint Conference on*, pages 707–713. IEEE, 2013.

[3] M. H. Hsu, T. K. Shih, and J. S. Chiang. Real-time finger tracking for virtual instruments. In *Ubi-Media Computing and Workshops (UMEDIA), 2014 7th International Conference on*, pages 133–138. IEEE, 2014.

[4] R. J. Jacob, A. Girouard, L. M. Hirshfield, M. S. Horn, O. Shaer, E. T. Solovey, and J. Zigelbaum. Reality-based interaction: a framework for post-wimp interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 201–210. ACM, 2008.

[5] G. Lakoff and M. Johnson. *Metaphors we live by*. University of Chicago Press, Chicago, 2008.

[6] M. Mandanici and S. Canazza. The "hand composer": gesture-driven music composition machines. *In Proc. of 13th Intl. Conf. on Intelligent Autonomous Systems.*, July 15-19 2014.

[7] M. Mandanici, A. Rodà, and S. Canazza. The harmonic walk: an interactive educational environment to discover musical chords. *Proceedings of ICMC-SMC Conference 2014, Athens*, 2014.

[8] M. Mandanici and S. Sapir. Disembodied voices: a kinect virtual choir conductor. *Proceedings of the 9th Sound and Music Computing Conference*, pages 271–276, 2012.

[9] T. Marrin and R. Picard. The 'conductor's jacket': A device for recording expressive musical gestures. In *Proceedings of the International Computer Music Conference*, pages 215–219. Citeseer, 1998.

[10] D. Murphy, T. H. Andersen, and K. Jensen. Conducting audio files via computer vision. In *Gesture-based communication in human-computer interaction*, pages 529–540. Springer, 2004.

[11] D.-J. Povel and E. Jansen. Harmonic factors in the perception of tonal melodies. *Music Perception*, 20(1):51–85, 2002.

[12] M. Rudolf and M. Stern. *The grammar of conducting: A comprehensive guide to baton technique and interpretation.* Schirmer Books, New York, 1994.

[13] S. Sentürk, S. W. Lee, A. Sastry, A. Daruwalla, and G. Weinberg. Crossole: A gestural interface for composition, improvisation and performance using kinect. *Proc. NIME'12*, 2012.