You're The Conductor: A Realistic Interactive Conducting System for Children

Eric Lee
Media Computing Group
RWTH Aachen University
52056 Aachen, Germany
eric@cs.rwth-aachen.de

Teresa Marrin Nakra
Immersion Music
P.O. Box 230248
Boston, MA 02123 USA
teresa@immersionmusic.org

Jan Borchers
Media Computing Group
RWTH Aachen University
52056 Aachen, Germany
borchers@cs.rwth-aachen.de

ABSTRACT

This paper describes the first system designed to allow children to conduct an audio and video recording of an orchestra. No prior music experience is required to control the orchestra, and the system uses an advanced algorithm to time stretch the audio in real-time at high quality and without altering the pitch. We will discuss the requirements and challenges of designing an interface to target our particular user group (children), followed by some system implementation details. An overview of the algorithm used for audio time stretching will also be presented. We are currently using this technology to study and compare professional and non-professional conducting behavior, and its implications when designing new interfaces for multimedia. *You're the Conductor* is currently a successful exhibit at the Children's Museum in Boston, USA.

Keywords

conducting systems, interactive exhibits, gesture recognition, real-time audio stretching, design patterns

1. INTRODUCTION

While new interaction methods for digital music and other multimedia are a hot topic in today's research community, much of the work has remained out of reach for the general public. This is partly due to the fact that these interfaces are geared towards those with a musical background; unfortunately, schools continue to reduce budgets for arts and music education [2], [18]. Consequently, for many people, user interfaces for digital music remain largely limited to decades-old metaphors of play, fast-forward and rewind. We have all seen people (even ourselves!) unconsciously tapping a drum beat to a rock song or conducting to an orchestral piece from a CD; the goal of our work is to use technology to support these methods of multimedia interaction, and encourage people to more interactively explore music.

In this paper, we present the first electronic conducting system with the following features:

- Designed to be usable by children with no prior music experience.
- Displays a high-quality audio and video recording of an orchestra.
- Produces high-quality, time-stretched audio in real-time without pitch alteration using an improved phase vocoder algorithm.

You're The Conductor is one of the main exhibits in the "Making America's Music: Rhythm, Roots & Rhyme!" exhibition at the Children's Museum in Boston, USA, and has received favorable press since it opened in June 2003 [5].

2. RELATED WORK

In this section, we present some earlier conducting systems most relevant for our current discussion.

Mathews' *Radio Baton* [14] was the first system to provide an interactive conducting experience. It uses the movement of one or more batons emitting radio frequency signals above a flat receiver panel to determine conducting gestures. A MIDI file is played back synchronously with these movements. Conducting is restricted to the space above the receiver.

The *Digital Baton* measures additional parameters besides baton position, such as pressure on parts of the handle, to allow for richer expression [12]. The *Conductor's Jacket* [13] uses sixteen additional sensors to track muscle tension and respiration, translating gestures based on these inputs to musical expressions. It uses a MIDI-based synthesizer to create the resulting musical performance.

Usa's *MultiModal Conducting Simulator* [17] uses Hidden Markov Models and fuzzy logic to track gestures with a high recognition rate of 98.95–99.74%. It plays back a MIDI score, with matching tempo, dynamics, staccato/legato style, and an adjustable coupling of the orchestra to the conducting.

Ilmonen's *Virtual Orchestra*, demonstrated at CHI 2000, is one of the few systems that also feature graphical output; however, it renders the orchestra synthetically as 3D characters. Audio output is again MIDI-based [7].

Murphy, Andersen and Jensen track a real baton using a camera and computer vision [15]. A beat tracking algorithm synchronizes professional conducting gestures to an actual audio recording. The audio playback speed is adjusted in real-time using a variation of the phase vocoder algorithm; however, there is no graphical output.

Unlike the above systems, many of which focus on interpreting professional conducting styles, *You're the Conductor* is designed to provide an immersive experience for children by processing recorded audio *and* video.

2.1 Personal Orchestra

Personal Orchestra is a system we have previously designed for the House of Music in Vienna, Austria [4]. The user, using an infrared baton, conducts recordings of performances by the Vienna Philharmonic; the music's tempo, volume and instrument emphasis can all be controlled. The system recognizes simple up-down gestures which are synchronized to the beat of the music.

While we have built upon this earlier experience, *You're the Conductor* is different from *Personal Orchestra* in both interaction style and technical implementation. *Personal Orchestra* was targeted towards adult visitors to a music museum in Vienna, Austria, many of whom have prior music experience, and (usually) interest in music. Conversely, *You're the Conductor* was to be deployed in an environment where children aged four and up are free to roam, often unsupervised. We could not assume that these children had prior music experience – in fact, the Children's Museum hopes to expose and foster children's' interest in music through this exhibit. These factors drove some important design decisions, which will be discussed in the next section.

Technically, *You're the Conductor* improves upon *Personal Orchestra* by using a real-time, high-quality time stretching algorithm for polyphonic audio; this feature allows the system to respond instantaneously and accurately to user input. *Personal Orchestra* did not have this feature, and instead relied on pre-stretched audio tracks to implement tempo changes; consequently, the speed could only be adjusted at discrete intervals. This implementation created inherent limitations on the accuracy of the tempo following, and caused occasional audible artifacts when switching between the different audio tracks. We will discuss our time stretching algorithm in more detail later in this paper.

3. REQUIREMENTS AND DESIGN

The primary goal of this system is to engage children in an interesting way, thus encouraging them to continue exploring the world of music. Designing for children, however, imposes some interesting and challenging constraints on the physical and interaction design:

- The gesture recognition system must be robust to the possibly erratic gestures of a child, yet still work reasonably well for their parents.
- The conducting device and surrounding exhibit must be robust to use (and abuse). An expensive and relatively fragile infrared baton, such as the one used in *Personal Orchestra*, would have been unacceptable. The Children's Museum required that the baton be tethered, inexpensive, and easy to replace.
- The user interface must be carefully designed to not only attract, but keep hold of, a child's interest; because children have a very high energy level, it is often challenging to keep their attention for very long. We have personally observed children at the museum running up to an exhibit, pushing some buttons, and moving on when the system did not immediately respond in an interesting way – all in a matter of seconds.

Figure 1: *You're the Conductor* exhibit at the Children's Museum of Boston.

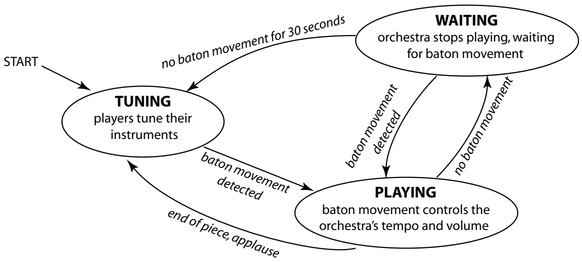

Figure 2: State machine for *You're the Conductor*.

These requirements are reflected in our design through a series of interaction design patterns for interactive exhibits, principles and guidelines based on previous experience [3]. Some of the design patterns we utilized include: *invisible hardware* for shielding the user from the complexity of the hardware required to run the exhibit; *domain appropriate devices* by selecting a baton as the natural device for controlling tempo and dynamics; and *immersive displays* to provide an interesting and engaging exhibit.

3.1 User Experience

As a user walks near the exhibit (see Fig. 1), he/she sees a large screen 2.3 meters wide, showing a movie of the Boston Pops tuning their instruments. The sound of the instruments is soft enough to be unintrusive, but audible enough to attract attention. A baton rests on a podium before the large screen; as soon as the user picks it up, the scene switches to one where the orchestra has their instruments ready to play. As the user waves the baton, the orchestra plays "Stars and Stripes Forever", matching their tempo with the speed of the user's gestures and their volume with the size of the gestures. They continue to follow the user until he/she either stops conducting and leaves, or the piece finishes, at which point the user is rewarded with applause. In either case, the system returns to tuning their instruments, waiting for the next conductor (see Fig. 2).

This extremely simple flow of interaction satisfies our requirement for simplicity; unlike *Personal Orchestra*, there is no preamble required from the user before conducting (e.g. language or piece selection). Moreover, since the minimum age of our target user group is only four, we chose not to present the user with written instructions where it would

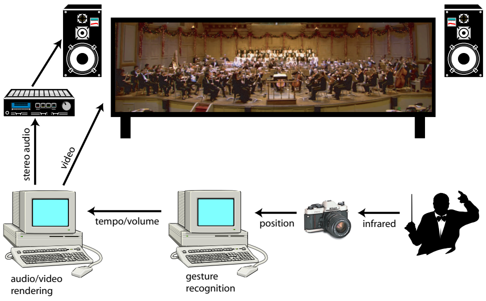

Figure 3: *You're the Conductor* system architecture.

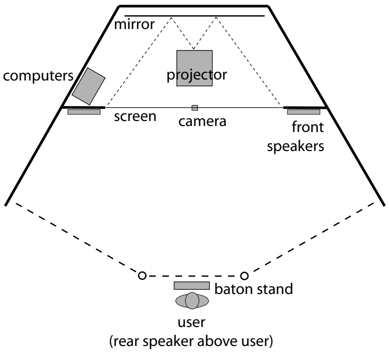

Figure 4: Floorplan of *You're the Conductor*.

interfere with their primary activity of conducting an orchestra.

One of the unique features of *Personal Orchestra* was that if the user conducted too slowly or too quickly, it would switch to a complaint sequence, pre-maturely ending the session. User observation showed that a similar "error condition" would only frustrate children: we found many of them enjoyed driving the orchestra as quickly or slowly as physically possible. It was also technically unnecessary to artificially limit the playback speed, since audio in *You're the Conductor* is stretched in real-time; in contrast, the range of playback speeds in *Personal Orchestra* is limited by the number of pre-stretched audio tracks.

3.2 System Overview

Figure 3 shows the *You're the Conductor* system architecture. The input device is an infrared baton whose signal is detected by a chip inside a camera; this signal is then passed to a host computer through an analog to digital converter. The host computer parses the gesture information and calculates the desired tempo and volume. This information is in turn passed to a second computer responsible for time stretching and playing the audio and video. Video is displayed on a large, rear-projected screen; audio speakers are placed on either side of the screen and on the ceiling directly above the user. All hardware (except the baton) is housed in a small room directly behind the screen, shielding the user from all of its intricacies, in the spirit of the *invisible hardware* design pattern (see Fig 4).

4. IMPLEMENTATION

In this section we will describe in more detail the individual components of *You're the Conductor*.

4.1 Baton Sensor System

A camera is situated just below the screen to receive the infrared illumination from the tip of the baton. The size of the user's gesture space is approximately 1.5 meters by 1.5 meters at a fixed distance of 2.2 meters from the camera, making it suitable for both adults and children. The user is directed to aim the baton toward the camera lens, and as long as the baton is angled within ±60 degrees of the camera, the infrared signal is received by the system. However, users' gestures often cause the baton to repeatedly enter/leave the camera's field of view, creating a jitter in the received signal strength that must be accounted for in the gesture recognition algorithm.

The camera houses a two-dimensional position-sensitive detector (PSD). Silicon PSDs are optoelectronic sensors that provide analog position data from light spots traveling over their sensitive surfaces. By using an infrared filter to remove the visible light frequencies (below 720 nm), we ensured that the PSD could reliably track the infrared signal from the baton.

The signals from the PSD are sent to a signal processing circuit mounted to the back of the camera, which extracts the xy position and intensity of the light spot from the moving baton. These analog output signals are then passed along to the host computer where they are converted to digital form.

4.2 Gesture Recognition and Interpretation

The mapping scheme we decided upon was simple but effective: gesture speed is mapped to tempo, and gesture size is mapped to volume.

Given the x and y positions over time, we compute the gesture speed using $\sqrt{dx/dt + dy/dt}$. We then apply an offset value to calibrate the tempo of the music to a reasonable gesture speed; this minimizes the amount of "work" that the user needs to do to get a satisfying result. A small time-decay factor is also applied to this result; this decay exponentially reduces the tempo to zero when there is no x and y input, which causes the music to slow down imperceptibly if the baton goes out of range for a short time but wind down quickly if it stays out of range for more than one second. A number of different filters at various stages ensure that tempo changes are smooth.

Gesture size is determined by looking at the moments when the baton changes direction horizontally or vertically. The x and y positions are sampled and stored at these moments of inflection, and subtracted from their corresponding values in the opposite direction to get the width and height of the gesture. The scaled magnitude of this resulting size vector gives us the loudness value. Again, several stages of smoothing algorithms are used to account for any jitter and inconsistencies in the users' gesture patterns.

Choosing this straightforward gesture mapping scheme was motivated by our target user group of children aged four to eleven. The children we worked with during our initial stages of design had never seen a symphony orchestra, and thus we could not expect them to know the standard conducting techniques. We conducted an informal test at this time, where we asked 20 children to “conduct” to an orchestral audio/video recording. Not only did these children show a wide range in types of movements, but even when they made beat-like gestures, these beats were frequently not synchronous with the beats of the music! Thus, we decided to adopt a simple velocity-based scheme, decoupled from the “beats” that children would make.

In a later stage of implementation, we again conducted a series of informal user tests based on this algorithm. Despite the fact that our algorithm does not provide beat synchronization, we found that this did not affect their experience with the system. Most users intuitively synchronized their “ictus” (beat) gestures with the beats in the audio, irrespective of the speed of their movements.

4.3 Audio Rendering with Time Stretching

Time stretching polyphonic audio such as orchestral music is non-trivial: the simple approach of changing only the playback speed has the undesired side-effect of also changing the pitch. Current audio time stretching algorithms can be roughly categorized into three types:

Time-based: These algorithms process audio in the time domain; the earliest time stretching algorithm was granular synthesis, where audio is cut into short segments (granules) and repeated/removed to adjust the tempo [6] (see Fig. 5). These algorithms have low processing requirements, but the time stretching range is limited to a small interval around the original speed (approximately ±20%). Moreover, while they work well with structurally simple audio signals such as speech [11], there are significant artifacts for polyphonic audio signals such as an orchestra.

Frequency-based: These algorithms require processing in the frequency domain; one example is the phase vocoder [8], which can produce high quality output over a wide range of stretching factors at the expense of increased processing requirements.

Analysis-based: There are a variety of algorithms which involve more advanced analysis of an audio signal’s inherent structure. While these algorithms are capable of producing high quality output, they are often proprietary and/or require iterative processing, and thus difficult or impossible to implement in real-time. One such example is Prosoniq’s MPEX algorithm [1].

We have chosen to use a modified phase vocoder algorithm in *You’re the Conductor*.

4.3.1 The Basic Phase Vocoder Algorithm

In this section, we will briefly introduce the phase vocoder; a more rigorous description of the algorithm is given in [8]. To help understand the phase vocoder algorithm, we start with the trivial example of a sinusoid with a single frequency, f.

The naive implementation of simply changing the playback speed by a factor n has the undesired side-effect of also changing the frequency, $f' = nf$. In fact, what we really want is to increase the number of sinusoidal cycles by a factor of n, without changing the frequency.

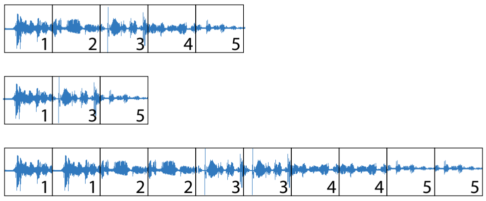

Figure 5: Time stretching using granular synthesis. The first figure shows an audio stream divided into granules. The second and third figures show how the audio could be time compressed and expanded by a factor of two, respectively.

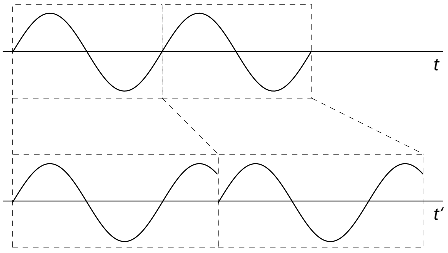

Figure 6: Time stretching a sinusoid of fixed frequency. The top figure shows the signal divided into two rectangular, non-overlapping windows. The bottom figure shows the result of time stretching these two windows without accounting for the phase.

One solution, then, is to take a window of the input signal and stretch that signal in the frequency domain by a factor of n. This has the effect of preserving the frequency whilst increasing the duration of the signal. Unfortunately, the resulting signal will be distorted unless we consider the *phase* of the signal when computing the time stretched window (see Fig. 6).

The phase vocoder operates on a similar principle, but uses overlapping and tapering windows to avoid the introduction of false high frequencies from abrupt signal changes at the window boundaries. By varying the spacing between the windows in the source and processed audio, the duration (and thus, speed) is changed. Windowing and the short time Fourier Transform are widely discussed in literature [16]; here, we will focus on the problem with phase discontinuity.

The phase vocoder algorithm estimates the phase of the signal based on the time stretch factor and the phase from the previous sample window. Since the window size is finite, computations are limited to discrete frequency intervals (*bins*), which has implications for the phase estimation calculation: there is a certain error associated with each estimate, which propagates and accumulates for each subsequent estimate. Consider an arbitrary signal with a more complex frequency spectrum – since the basic phase vocoder algorithm estimates the phase for each frequency bin independently, the signal loses phase coherence across the bins. To the human ear, this phase error sounds like a reverberation effect in the processed audio.

4.3.2 An Improved Phase Vocoder Algorithm

Laroche and Dolson describe the phase coherence problem as a loss of "vertical phase coherence" [9]. They propose a solution where the relationship between phases across multiple frequency bins is also considered when computing the phase estimation. This "rigid phase-locking" technique, which we have implemented in *You're the Conductor*, works as follows: the main frequencies of the audio signal are determined by searching for the amplitude peaks in the signal's spectrum. It is then assumed that the phase for the frequencies neighboring the peak are related to the phase of the peak, and preserved in the output signal. Their test results have shown a significant improvement in audio quality over the basic phase vocoder algorithm, which have been verified by comparing the output of our own implementations of the two phase vocoder algorithms.

Our audio is sampled at 44.1 kHz and processed with a window size of 4096 samples (93 ms) and overlap of 75%. Thus we can adjust the playback speed up to 43 times per second. This increased level of responsiveness is a significant improvement over *Personal Orchestra*, which was artificially limited to a speed change of once per second to reduce audio artifacts when switching between tracks.

4.4 Video Rendering

Unlike the audio, time stretching the video in *You're the Conductor* did not pose as large of a problem; simply changing the playback rate was sufficient. While this technique can cause unnatural movements for some types of footage, such as objects falling from the sky, there was no such problem with this system.

Unfortunately, current multimedia frameworks do not easily support the integration of a module such as the phase vocoder, which manipulates time. Thus, substantial effort was spent designing and developing a media player capable of synchronizing video played at variable speeds with the output from our phase vocoder module. The design of this player posed some interesting technical challenges, including audio-video synchronization, latency and thread management that are beyond the scope of this paper; they will be discussed in a separate publication.

5. SOFTWARE AND HARDWARE

The baton was designed in conjunction with Superior Exhibits of Illinois, responsible for assembling the exhibit. It is a machined aluminum tube, similar in size and shape to a bicycle handle; at the tip is a high-power GaAIAs infrared LED from Opto Diode Corp. The aluminum handle acts as a heat sink for the LED, which is powered in continuous mode with 13.5 V at 160 mA.

The camera system consists of a Minolta SRT100 camera body that has been modified to house a Hamamatsu two-dimensional PSD (S5991-01). The PSD chip sits at the focal plane of a standard manual-focus lens (Vivitar 28mm f2.8), with the camera's shutter permanently left open. An analog to digital converter card from Measurement Computing provides the data to a host PC running National Instruments' Labview, which also provides the support framework for the gesture recognition software.

The media player with audio stretching runs on an Apple PowerMac system with dual 1.42 GHz G4 processors. Audio time stretching is implemented as a Core Audio module and synchronized to a video stream played back using QuickTime.

While the current system runs on two computers, this implementation was simply a matter of convenience; at the time it was difficult to find a compatible A/D converter module for the Macintosh. We are currently working on integrating the software onto a single machine.

6. EVALUATION

In addition to the user feedback we obtained throughout the design of *You're the Conductor*, we were able to observe visitors to the exhibit on a Sunday afternoon shortly after it opened to the general public. Over the period of one hour starting at 1:20 pm, there were 66 users of the system: 18% adults, 44% children supervised by an adult, 36% unsupervised children and 2% staff members. Users spent an average of 33 seconds at the exhibit, ranging from 5 seconds to 2 minutes. Only 8% of the users had trouble using the system; in all cases but one, this was corrected by a supervising adult. All users except one (98%) were able to determine that their movements controlled the orchestra. Finally, the baton underwent heavy use (and abuse) during this hour: the baton was dropped on the floor at least once, and one child even licked it like an ice cream cone!

We were able to determine two main causes for the problems users had with the system. Some users (especially small children) held the baton in such a way that it was pointing upwards, out of the camera's line of sight of the camera. Also, the small conductor's podium that was part of the exhibit design also blocked the camera's line of sight for the smaller children. In both cases, a second camera would help address these usability issues (see Future Work).

7. FUTURE WORK

We are currently exploring topics that build on our experience with designing *You're the Conductor*:

Improved input device: Our baton uses only one infrared emitter and is thus simpler and cheaper to produce than the Buchla Lightning II used in *Personal Orchestra*; however, because we use only one camera to detect the infrared signal, we have not been able to achieve the same degree of omni-direction that the Buchla provides. We are currently exploring ways of using multiple cameras to address this issue, while further reducing the cost and complexity of the baton device.

Improved tempo following: While our gesture recognition system works well for children and adults with little or no musical experience, professional conductors find the experience unsatisfying: the current system, unlike a real orchestra, does not react to their more complex and subtle movements, only to their gesture speed and size. Thus, we are currently performing studies on the movements of professional conductors and how their behavior and mental models differ from those with little or no professional experience; we hope to eventually create a new system similar to *You're the Conductor*, but with gesture recognition algorithms that adjust to the users' level of experience with conducting.

Improved time stretching algorithm: While our time stretching algorithm works well even when the audio is slowed down to one-half the original speed, audio artifacts begin to emerge below this threshold; they are especially noticeable for transient signals such as drum beats, which ideally should not be time stretched. A more advanced analysis-based algorithm, such as the one proposed by [10], could possibly address these artifacts, at the expense of more computing power.

New application areas: We believe that the ability to create high-quality, pitch-preserving time stretched audio in real-time allows us to examine unexplored areas in interfaces with music and other multimedia. For example, imagine an "instrumental karaoke" where a user could play along to a recording, but have the recording adjust to the *user's* tempo and dynamics (rather than the other way around). Or a dancer whose routine is no longer constrained by the tempo/dynamics of a particular recording – rather, the music would spontaneously react to his/her movements. This could allow a greater freedom of artistic expression currently difficult to accomplish with current systems. We hope to further explore these venues as we continue our work on innovative interactive exhibits.

8. SUMMARY

We have presented a system for children to conduct a realistic electronic orchestra. This system uses an improved phase vocoder algorithm for high quality audio time stretching without pitch alteration. Some of the requirements and challenges of designing an interactive music exhibit for children were discussed. The interface is specifically designed for children: it is simple, and our gesture recognition system is robust to their wide range of movements. We then presented a technical discussion of the *You're the Conductor* system architecture and its components. Our primary technical innovation is the use of an improved phase vocoder algorithm with rigid-phase locking to provide high quality audio time stretching in real-time without pitch alteration. Based on the expertise and technology we have gained from building *Personal Orchestra* and *You're the Conductor*, we hope to continue our research in designing interactive music exhibits with novel interfaces.

9. ACKNOWLEDGMENTS

We would like to thank the Children's Museum of Boston for their generous support; the Boston Symphony Orchestra and Boston Pops for providing the audio and video footage for the exhibit; Jeff Kennedy Associates for the physical exhibit design; Ron Yeh and Jahangir Nakra for their technical contributions; Julius Smith for his consultation on the phase vocoder; and Rafael Ballagas for his insightful comments on this paper.

10. REFERENCES

[1] Prosoniq MPEX - minimum perceived loss time compression/expansion. http://mpex.prosoniq.com/.

[2] School arts 'undermined by curriculum'. BBC News, June 2003. http://news.bbc.co.uk.

[3] J. Borchers. *A pattern approach to interaction design.* John Wiley & Sons, New York, 2001. http://hcipatterns.org.

[4] J. Borchers, E. Lee, W. Samminger, and M. Mühlhäuser. Personal orchestra: A real-time audio/video system for interactive conducting. *ACM Multimedia Systems Journal Special Issue on Multimedia Software Engineering*, 9(5):458–465, March 2004. Errata published in next issue.

[5] A. Deahl. Striking a chord with kids. Boston Globe, June 2003. http://www.boston.com/globe/.

[6] D. Gabor. Theory of communication. *Journal of Institution of Electrical Engineers*, pages 429–457, 1946.

[7] T. Ilmonen. The virtual orchestra performance. CHI. ACM, 2000.

[8] J. Laroche. Time and pitch scale modification of audio signals. In M. Kahrs and K. Brandenburg, editors, *Applications of DSP to Audio & Acoustics*, pages 279–309. Kluwer Academic Publishers, 1998.

[9] J. Laroche and M. Dolson. Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing*, 7(3):323–332, May 1999.

[10] S. Levine and J. O. Smith III. A sines+transients+noise audio representation for data compression and time/pitch-scale modifications. 105th Audio Engineering Society Convention, San Francisco, 1998.

[11] D. Malah. Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-27(2):121–133, April 1979.

[12] T. Marrin. Possibilies for the digital baton as a general-purpose gestural interface. CHI, pages 311–312. ACM, 1997.

[13] T. Marrin Nakra. *Inside the Conductor's Jacket: Analysis, interpretation and musical synthesis of expressive gesture.* PhD thesis, Massachusetts Institute of Technology, 2000.

[14] M. V. Mathews. *Current Directions in Computer Music Research*, chapter The Conductor Program and Mechanical Baton. MIT Press, Cambridge, 1991.

[15] D. Murphy, T. H. Andersen, and K. Jensen. Conducting audio files via computer vision. Proceedings of the Gesture Workshop, Genova, 2003.

[16] A. W. Peevers. A real time 3D signal analysis/synthesis tool based on the short time fourier transform. Master's thesis, University of California, Berkeley, 1995.

[17] U. Satoshi and M. Yasunori. A conducting recognition system on the model of musicians' process. *Journal of the Acoustical Society of Japan*, 1998.

[18] G. Sealey. Just the three R's? budget tightening, standards blamed for squeezing favorite programs striking a chord with kids. ABC News, August 2003. http://abcnews.go.com.