Multisensory Musical Entertainment Systems

Zhiying Zhou, Adrian David Cheok, Wei Liu, Xiangdong Chen, and Farzam Farbiz
National University of Singapore

Xubo Yang
Shanghai Jiaotong University

Michael Haller
Upper Austria University of Applied Sciences

Two musical entertainment systems—the Bubble Bumble and the Magic Music Desk—let users create and play back music in 3D physical environments through augmented reality interfaces. During the process of creating and playing back music, users encounter visual–speech, visual–tactile, and tactile-auditory experiences.

Humans typically explore the world through their senses—sight, hearing, touch, smell, taste, and balance. The modalities corresponding to these senses are visual, auditory, tactile, olfactory, gustatory, and vestibular.[1] The different sensory modalities[2] that humans use aren't processed in isolation; instead, humans integrate them, which results in a multisensory, rich experience of the world. This process of transcending individual sensations into experience is *perception*.[3]

Aided by multimedia technologies, humans can experience the physical and the virtual world (virtual reality), or even experience both worlds simultaneously (mixed reality). Multimodal[4] systems are specifically designed to offer multisensory experiences. In this research area, *perception* describes the communication from a machine to a human;[5] more explicitly, it describes the process of transcending disparate sets of data into a unified experience.

As a result of our research into multisensory perception, we developed two unique musical entertainment systems—Bubble Bumble and Magic Music Desk (MMD)—that provide users with multisensory experiences in creating and playing back music in a tangible 3D and physical environment. (See the "Related Work" sidebar for background.) We applied visual, auditory, and tactile sensory modalities to create a system that could express a wide range of human–musical artifacts. During the process of creating and playing back music, users experience a unique sensing experience of 3D–visual–speech, 3D–visual–sound—music and nonspeech sound—and tactile-auditory perceptions. Users also can experience visual–tactile sensation by using their hands to cause a corresponding system action. Beyond the multisensory experience, our multimedia systems also emphasize tangible human–computer interactions and communications within a physical environment.

In this article, rather than focus on the application's design, we examine the systems' multimodal perceptions. We describe each system and how it works, and we present the results of a formal user study we conducted.

Multisensory approach

We developed two main demonstrations of our research. In Bubble Bumble, users collaborate to create music in a chaotic, nonlinear manner by physically capturing and bursting augmented virtual bubbles, which contain short musical segments of instrumental performance floating in the air, and dropping them down onto a virtual playback timeline, which resembles the traditional music staff. In MMD, users manipulate the virtual visual metaphor of musical resources in the physical environment by natural, intuitive interactions of hands and speech commands. We call our technique "what you say is what you see" (WYSIWYS), which turns a user's speech into a visual image. When users say "guitar," for example, a 3D character (which, in English, would be the word "guitar" but, say, might in Italian be "chitarra") appears to fly out of their mouth, dropping down to the table. When the character (or word depending on the language) hits the table, it turns into a 3D model of a guitar. The sound coming from the speaker's mouth turns into a virtual musical instrument that drops down onto the desk. Users' hand movements directly manipulate the virtual instruments.

Bubble Bumble

Our first development effort in multisensory multimedia is Bubble Bumble, a novel system

1070-986X/04/$20.00 © 2004 IEEE

Published by the IEEE Computer Society

Related Work

The advanced development of sensors and computer multimedia technologies has prompted researchers to design new musical interfaces.[1] These interfaces, which inevitably involve multiple modalities, have applied multimedia technologies to enhance the player's performance and experience. The design space multimodal environment[2] is regarded as a sort of extension of augmented reality (AR) environments. The augmentations are taking place in multiple modalities—vision, gesture, speech, music, sound, and touch.

Other researchers have focused on designing augmented multimodal musical entertainment interfaces.[3,4] Although most of these interfaces dealt with vision, music, and gesture modalities, few of them examined the speech, 3D sound, and tactile modalities. One of our research aims is to design musical entertainment interfaces that encompass all these modalities.

Although speech recognition has been applied to some industrial applications, few researchers have introduced speech in the design of musical entertainment interfaces. Designing a good musical entertainment interface that incorporates speech to enrich the user's multisensory experience is still a challenge. Our systems, Bubble Bumble and Magic Music Desk (MMD), do apply speech modality and provide multimodal perceptions by visualizing speech as a 3D virtual character or 3D virtual objects.

Three-dimensional sound is helpful to the immersiveness of virtual reality environments. The effectiveness of 3D sound in AR environments, however, will vary slightly because the interface is no longer fully immersive. Consequently, we've applied a 3D sound modality in our systems and investigated its effectiveness, as we explain in the main text. Unlike musical interfaces[4,5] in which the soundscape is not affected by the user's interactions, we intend our systems to provide a user-centric musical entertainment experience. We also investigate how 3D sound interacts with other modalities.

Tangible interaction is a major trend in human–computer interface development. The idea is to communicate and control the digital information by manipulating the physical artifacts. It encourages user collaboration and cooperation by allowing direct manipulation of digital information. Paradiso et al.'s work[5] applied passive resonant, magnetically coupled tags on trinkets, such as a small plastic cube or toy, to serve as musical controllers. With these trinkets, each tagged object acquires a set of complex musical properties when in proximity to the reader. Poupyrev et al.[4] used physical cards for the same purpose. Our system differs in that we applied more modalities and used hand gesture recognition to enhance tangible interaction.

Expressive gesture is commonly used in musical interfaces. Most systems recognize the specific gestures and use the result as a musical controller[4,5] or as a reference for visual augmentation.[6] We took a different approach and applied expressive gesture via the hand gesture modality in MMD, which allows gesture–vision, gesture–3D sound perceptions, and gesture–tactile actions. Simple hand gestures can directly manipulate virtual objects, which thus retains the advantages of tangible interaction.

Most of the previously mentioned musical interfaces try to modify musical effects by mapping the musical properties, such as tone and length, with the values retrieved from vision, gesture, and touch modalities. Instead of trying to design these mappings, our Bubble Bumble system is a composition tool that uses small musical pieces to compose music, and MMD is a scene-managing tool that deals with multiple instruments and players on a stage. Both systems give users more of an entertainment through integrated multiple modalities and integrated user interactions that provide collaborative multisensory experiences.

References

1. J. Paradiso, "Electronic Music Interfaces: New Ways to Play," *IEEE Spectrum*, vol. 34, Dec. 1997, pp. 18-30.
2. A. Camurri and P. Ferrentino, "Interactive Environments for Music and Multimedia," *Multimedia Systems*, vol. 7, no. 1, 1999, pp. 32-47.
3. M. Lyons, M. Haehnel, and N. Tetsutani, "The Mouthesizer: A Facial Gesture Musical Interface," *Proc. Siggraph 2001: Conf. Abstracts and Applications,* ACM Press, 2001, p. 230.
4. I. Poupyrev et al., "Augmented Groove: Collaborative Jamming in Augmented Reality," *Proc. Siggraph 2000: Conf. Abstracts and Applications,* ACM Press, 2000, p. 77.
5. J. Paradiso, K. Hsiao, and A. Benbasat, "Tangible Music Interfaces Using Passive Magnetic Tags," workshop report of the ACM CHI 2001 conference's Special Workshop on New Interfaces for Musical Expression, 2001; http://www.media.mit.edu/reserv/sweptRF.html and http://www.media.mit.edu/resenv/pubs/papers/tags-chi01-workshop.pdf.
6. A. Francis et al., "Emonator," http://web.media.mit.edu/~pauln/research/emonic/interfaces/emonator/emonator.html.

that combines augmented reality graphics with music to enhance the impression of reality and feeling of 3D music, in conjunction with intuitive user interaction. Bubble Bumble can be an interactive audiovisual system or, alternatively, an AR sound-visualization system that lets users compose music collaboratively in a multimodal entertainment environment.

System functionalities and multisensory experiences

Bubble Bumble supports two users in composing music simultaneously. To compose a song, users manipulate wands to capture the bubbles floating in the air, collaboratively burst the bubble and release the music or vocal piece contained in it, and drop the musical symbol object onto

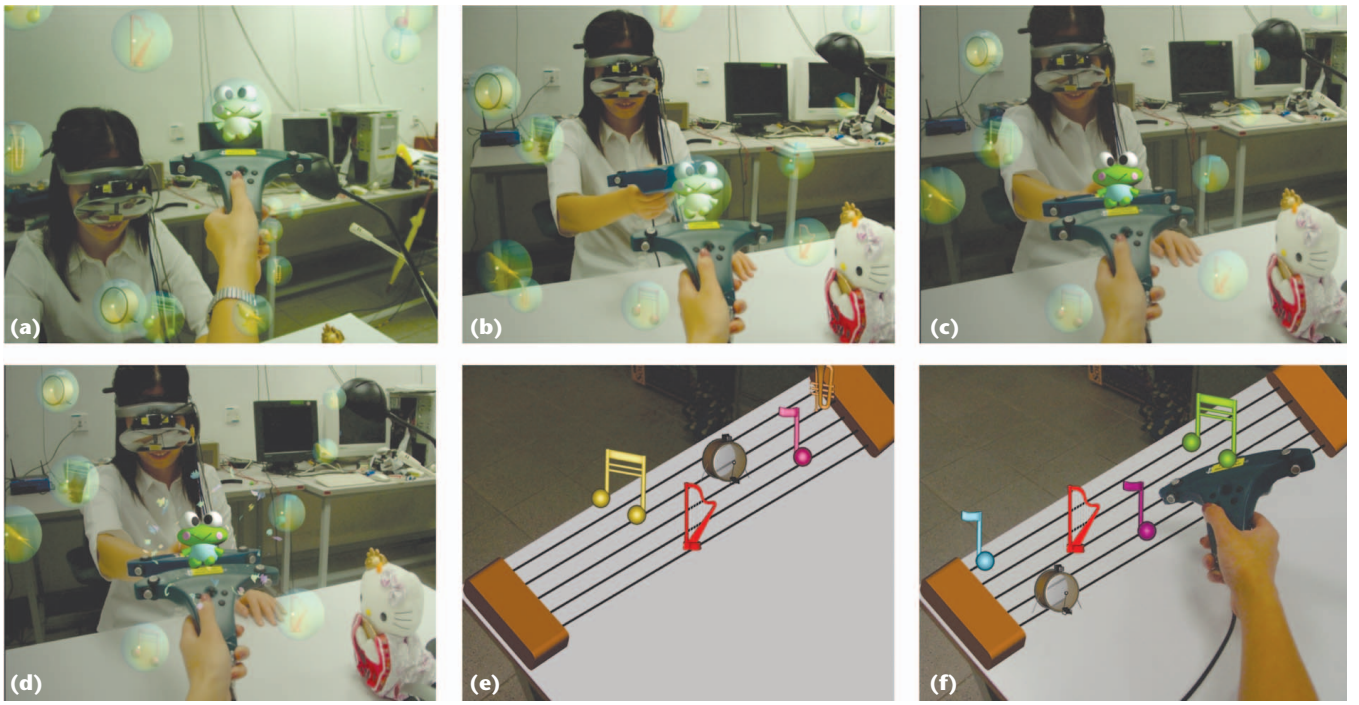

Figure 1. (a) The first user (back to camera) catches the bubble by the end of the wand. The little creature that seems to be at the end of the wand is the Japanese cartoon character "Korropi" and represents a set of notes. In (b) and (c), the first user collaboratively bursts the bubble with the second user; and (d) together, users release the musical pieces. (e) The virtual "timeline" (staff) is rolling leftward as time passes. When the first note gets to the edge, it will be played back automatically. (f) The first user drops the corresponding musical symbol object onto the rolling virtual timeline to add it into the playback queue.

the rolling virtual timeline, which is augmented at the edge of the desk (as Figure 1 shows). The virtual musical instruments correspond to the symbol objects: The bubbles in the air viewed by users have different 3D instruments inside, such as a guitar, trumpet, and so forth. When users catch them and place them onto the timeline, the Bubble Bumble system plays back short segments of music, corresponding to the appropriate instruments, when the object reaches the edge of the timeline. With this approach, users create music by temporally playing back different, brief musical pieces of different instruments.

Users can also create voice bubbles by singing or speaking into the microphone. The size of the bubble is proportional to the length of recording time controlled by a button on the wand. Our system synthesizes virtual 3D music pieces as if they are emanated from the floating bubbles. When users capture and move the bubble with the wand, the wand moves the virtual sound sources accordingly, as if the 3D sound were tangible.

Bubble Bumble's unique nature is, first, that it provides a multimodal and tangible method of creating music nonlinearly—users create music from many bubbles floating randomly in the air. Second, our system encourages cooperation between users in the process of creating music. Two users must collaborate to form a song, and ideally they interact by incorporating the modes of sound, vision, touch, music, voice, and motion. Table 1 summarizes the user experiences and perceptions, system modalities involved, and interaction features.

Discussion of multisensory experiences

Bubble Bumble encourages users' physical interactions and cooperation; however, with respect to multisensory design, speech and touch modalities function less intuitively and directly, because speech doesn't play an important role in the human–computer perception; it serves just as data. The visual–touch perception is indirect because the system translates the pressing of buttons into corresponding actions that can lead to specific visual feedback. In other words, the system's interpretation—not the user's tactile actions—is what determines visual feedback. We significantly improved both features in our later system, Magic Music Desk.

Table 1. Summary of Bubble Bumble modalities, user experiences, and features.

Bimodal and Multimodal Perceptions or Actions	Modalities Involved	User Experiences	Interaction Features
Visual–speech (perception)	Vision, speech	—See the speech being stored in a virtual bubble floating in the air. —See the bubble's size is proportional to the length of voice input. —Hear the playback of the prerecorded music and voice when seeing the virtual voice object reach the edge of the timeline.	Social interactions: Physically explore the 3D space with the other user.
Visual–music (perception)	Vision, 3D music	—See the floating bubbles while hearing 3D music that is attached to the bubbles. —See the playback timeline and hear music when the object reaches the edge of the timeline. —Feel self become the focus of both the auditory experience and visual experience. This is user-specific; one user may feel sound come from the side while another feels it come from the back.	—Virtually see the 3D graphic and hear the 3D sound enable a fully immersive AR experience. —The AR environment is visual as well as auditory.
Visual–tactile (action)	Vision, touch	—See the size of the virtual bubble growing proportionally to the time of holding the voice button on the wand. —Catch a bubble in the air by moving the wand physically close to the AR bubble and pressing a button on the wand. —Burst the bubble by physically moving wand (or the hand holding the wand) to the other user's wand.	—Physically interact with virtual objects. Body movement is essential. —Tangible interaction: Physically catch and burst the virtual bubbles. —Social interactions and collaboration between users are required.
Tactile–sound (perception)	Touch, 3D sound (speech and music)	—Move the wand with 3D sound attached to it. —The 3D sound becomes tangible and is easily arranged in the 3D space.	—3D sound becomes a tangible and visible object. —Tangible interactions and collaborations between users are required.

Magic Music Desk

MMD is a multimodal musical entertainment interface that employs the principles of embodied interaction (http://www.dourish.com/embodied) and social interaction between users. MMD visualizes users' speech as both a Chinese character (or foreign-language equivalent) and musical instruments, and lets users control and arrange instruments by speech commands and hand manipulation. Similar to Bubble Bumble, MMD applies multiple modalities of vision, speech, 3D music, and touch.

By applying speech recognition technologies, MMD provides a new method to interact with virtual objects and a novel interface to visualize speech. The WYSIWYS interface visualizes both users' spoken words as 3D objects. MMD visualizes spoken words as 3D characters coming from the speaker's mouth and turns them into virtual 3D instrument objects when it drops onto the desk. We used IBM's ViaVoice Dictation Software Developer's Kit, which supports speech recognition in multiple languages to recognize the spoken words. The MMD specifically demonstrates the multilanguage interaction between English and Chinese users.

MMD bases its registration of virtual objects and virtual 3D sound sources on a vision-tracking

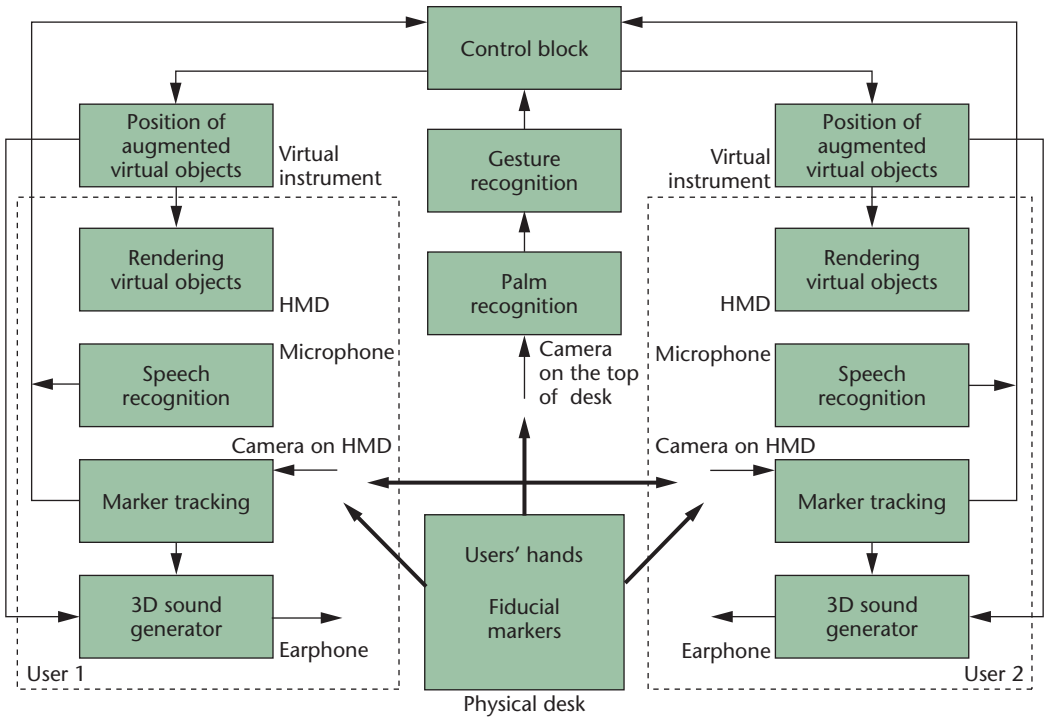

Figure 2. System architecture of the Magic Music Desk. We refer to the head-mounted displays as HMDs.

algorithm.[6] By applying 3D sound to the virtual objects, the augmented auditory environment becomes tangible; users can manipulate the virtual objects (and the 3D sound therein) by making simple gestures with their own hands. Our system represents a good combination of multiple modalities and enables users to have a fully immersive multisensory AR musical entertainment environment.

System architecture

Figure 2 shows the MMD system architecture. The user's camera, fixed on the head-mounted device (HMD), tracks the user's head movement relative to the markers. The MMD system uses the tracking result to register 3D virtual objects into the scene. The top camera recognizes hand gestures and gives the control block the hand positions relative to the markers. The speech recognition block sends the speech recognition result to the control block for rendering the characters and objects. The control block also uses the speech recognition result for controlling corresponding actions of the virtual objects, such as move, rotate, and delete. The earphone plays back the 3D sound generated by the 3D sound synthesizer block, given the 3D position information of virtual objects obtained from the marker tracking block. The HMD displays the mixed scene by augmenting virtual objects onto the real scene. The control block integrates multisensory inputs and synchronizes all other blocks.

Visual AR perception

To achieve precise registration of virtual objects, we need to calculate the transformation matrix between the marker coordinates and the camera coordinates. We used Kato and Billinghurst's ARToolkit[7] for the basic marker tracking. Figure 3 shows the coordinate systems of the cameras and the marker, where T_{C1M} and T_{C2M} denote the transformation matrix between the marker coordinates and camera 1 and camera 2 coordinates, respectively.

By calculating the transformation matrix between different cameras, we can precisely know the position and orientation of an object, whether real or virtual, in one camera given its position and orientation in the other camera. The AR environment is now a shared space for both users, which is crucial for user collaboration.

Apart from the two users' cameras, our MMD system also applied the third camera, used for hand gesture recognition mounted on the desktop. By the same transformation matrix calculation method, we can know the relationship between the maker and the camera on the desktop. The combination of these three cameras in the same coordinate system—marker coordinates—ensures precise registration of augmented graphics.

Visual–speech perception

Using speech commands to control the computer's operation is common in numerous applications.[8] Moreover, researchers have evaluated speech visualization for various purposes, such as to translate speech into graphics.[9] The MMD applies speech in both respects. To control objects, our system uses speech commands to move them and perform special operations such as zoom and rotate. Unlike Rosenberger and Neil's work[9] in which they translated speech to 2D text or graphics, we visualize speech as virtual 3D characters or objects. When viewed through the head-mounted display, the user experiences a fully immersive 3D AR environment. The MMD control block gets the speech recognition results from the speech recognition block to decide the exact position where the objects should be rendered relative to the marker coordinates.

WYSIWYS. In WYSIWYS, the MMD visualizes speech as 3D virtual characters or objects coming from the speaker's mouth. Because humans speak through the mouth, we believe it will be intuitive for MMD users to visualize speech as though it really were coming from the speaker's mouth.

To render the object coming out of the user's mouth, we need to know the relative position of the mouth to the user's cameras. To the speaker's own camera viewpoint, the MMD system roughly calculates the mouth's position by adding an 8 to 10 cm translation—a term used to describe the pure movements in the x,y,z axis when no rotations are involved—from the center of the camera along its Y_C axis (as Figure 3 shows). To calculate the position of the mouth in the other user's camera viewpoint, additional calculations are required. With the marker tracking method, we can calculate the transformation matrix between camera coordinates and marker coordinates The transformation matrix can be calculated for both cameras. Then we can calculate the transformation matrix between the two camera coordinate systems. As Figure 3 shows, T_{C1C2} denotes the transformation matrix between the two cameras' coordinates, which we obtain via a simple equation: $T_{C1C2} = [T_{C2M}]^{-1} = T_{C1M}$. To obtain the speaker's mouth position relative to the other user's camera coordinates, similarly a slight calculation from the speaker's camera to her mouth is taken.

Given the position information of the speaker's mouth in the other user's camera coordinates, we can render the virtual character or object in a 3D position, which gives users the illusion that the character or object is coming from the speaker's mouth. In addition, the virtual character travels in a parabolic track after it begins to drop onto the desk. When traveling in the parabolic track, the size of the character becomes larger. The character finally turns into an instrumental object after it "splashes" onto the surface of the desk, as Figure 4 (next page) shows.

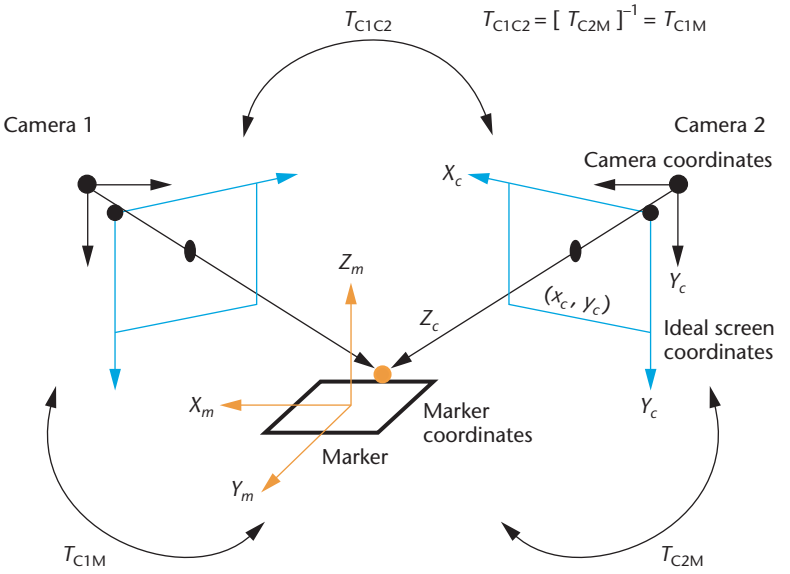

Figure 3. Coordinate systems of cameras and marker (c = center).

Speech recognition. The MMD applies the IBM ViaVoice Dictation SDK to process speech input for human–computer communication in real time. The speech engine handles a complex task, including taking the raw audio input and translating it to recognized text that our application understands. We constructed a grammar rule and a small model database as Table 2 shows. From each sentence or command, the speech recognition block tries to extract the action, object, and parameters. For example, the command "zoom tambour in" is separated into "zoom," "tambour," and "in." The MMD system determines the action (zoom), the object (tambour), and the parameters (in) and sends the results to the control block to render appropriate action of the corresponding object, as Figure 5a shows.

The index in Table 2 shows the number of actions available to users. The speech commands are composed by three words, one from each of the Action, Object, and Parameters columns. For example, the user might say "move guitar left" or "rotate tambour right." In both instances, only the three words taken from this table make up a valid speech command.

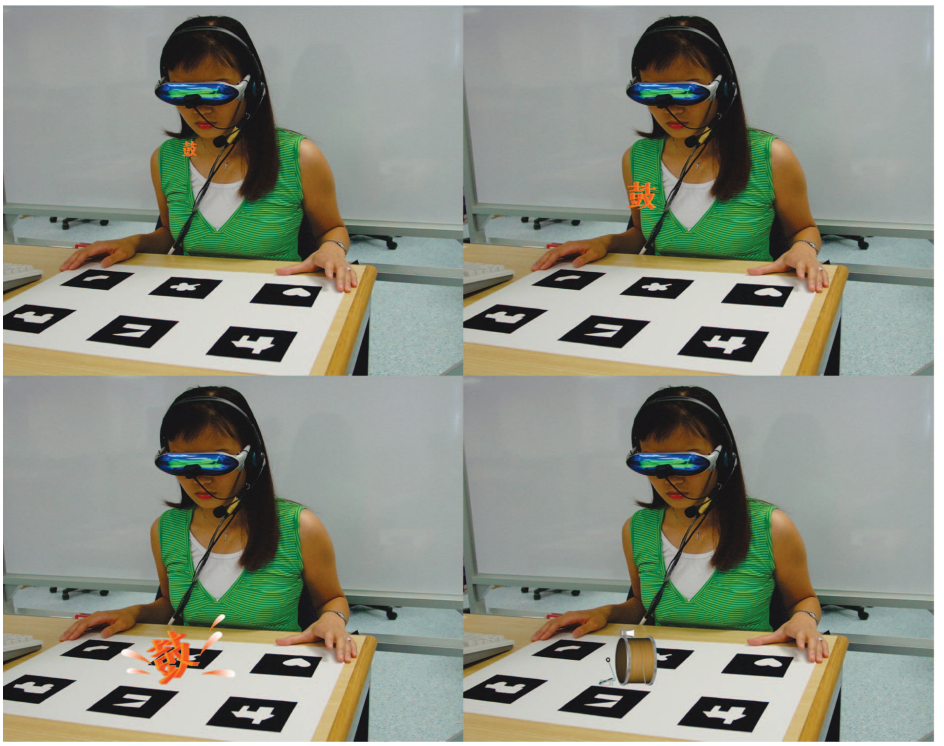

Figure 4. (a) The user is importing a virtual tambour (a type of drum) onto the desk by the speech command "input tambour now." The Magic Music Desk system interpets speech as a virtual object of the Chinese character meaning "tambour." This demonstrates a multilanguage interaction in which users can understand each other by the visualized character or object of speech. (b) The virtual Chinese character seems to come from the speaker's mouth and becomes larger as it reaches the desk. (c) A splash is shown when the character reaches the desk and (d) turns into an instrument object tambour.

Table 2. Speech commands and grammar used by the Magic Music Desk.*

Index	Action	Object	Parameters
1	input	Guitar	Now
2	delete		Now
3	play	Tambour	Now
4	stop		Now
5	move	Trumpet	Left, right, up, down
6	zoom		In, out
7	rotate	Piano	Clockwise, counterclockwise

* Grammar: Speech command = Action + Object + Parameters

Visual–gesture action. To register virtual objects onto the user's hand, we need to know the hand's positions in the marker coordinates along with the coordinates of the three cameras. The result of hand recognition will identify the hand's 2D position information, and the marker coordinates will enable the MMD to render the virtual objects onto the hand precisely.

Because we assume that the user's hand is always close to the desk surface, our system can recognize hand gestures with only one camera, which avoids the problem of matching image features between different views. We achieve stable detection of the palms by extracting two kinds of features: statistical based and contour based. We developed our approach after Du and Li's methods for gesture recognition. [10]

We use two simple gestures in our system—"pick up" and "drop down"—to manipulate virtual objects. When the MMD system recognizes that the user's palm is open, as in Figure 6a, MMD recognizes this as the pick-up gesture. If the user's hand is also close enough to a virtual object, a "pick-up" event occurs and the object moves onto the user's hand. The virtual object will move by following the motion of the user's hand, assuming the user's palm stays within the marker boundaries that MMD recognizes. When the MMD system recognizes the edge of the user's hand placed vertically on the desk, as in Figure 6b, MMD interprets this as the drop-down gesture and will drop the object onto the desk. Note that we assume that the hand is near the desk vector plane in all cases, so the hand's 2D position information is sufficient for our purposes.

Visual–tactile action

As Figure 6c shows, the MMD system displays the virtual object on the user's hand; the object will move whenever the user's hand moves. Similarly, Figure 6d shows that when the user's hand turns to the side, the object drops back onto the desk. Although the hand motions don't introduce tactile devices for force feedback when the user picks up or drops down virtual objects, visual–tactile action is still obtained to some extent by the MMD's real-time updating of the virtual scene, according to the hand's movement. In this sense, the user experiences a visual–tactile action by physically using the hand to act on the virtual objects.

Visual–3D music perception

After importing the instruments onto the desk

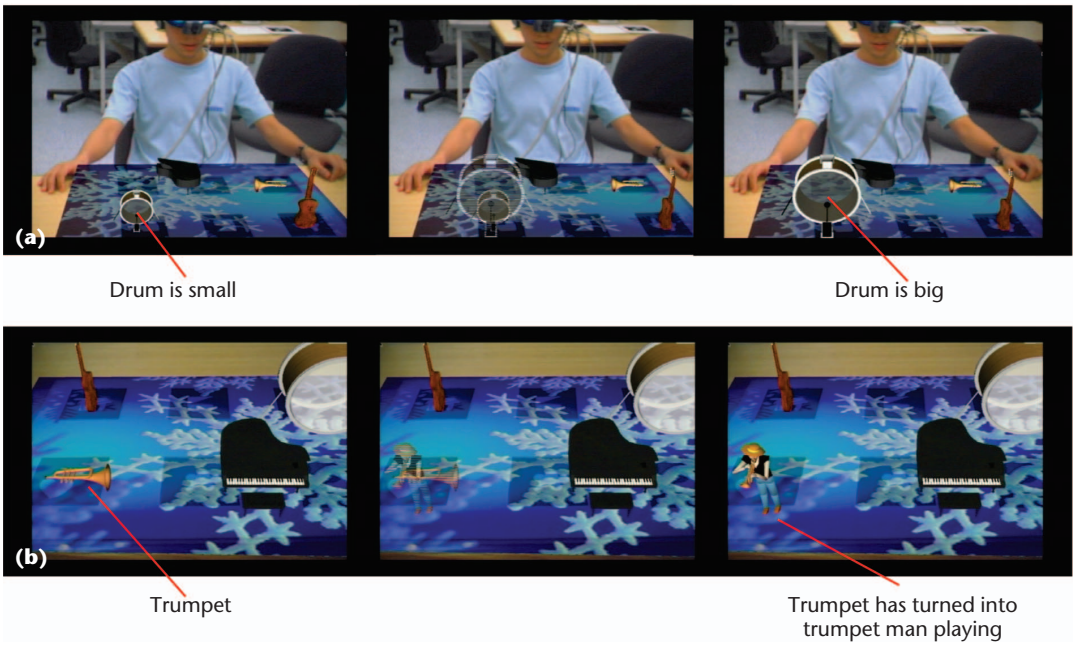

Figure 5. (a) The user enlarges the virtual tambour by saying "zoom tambour in." (b) The user's view of the speech command "play trumpet now."

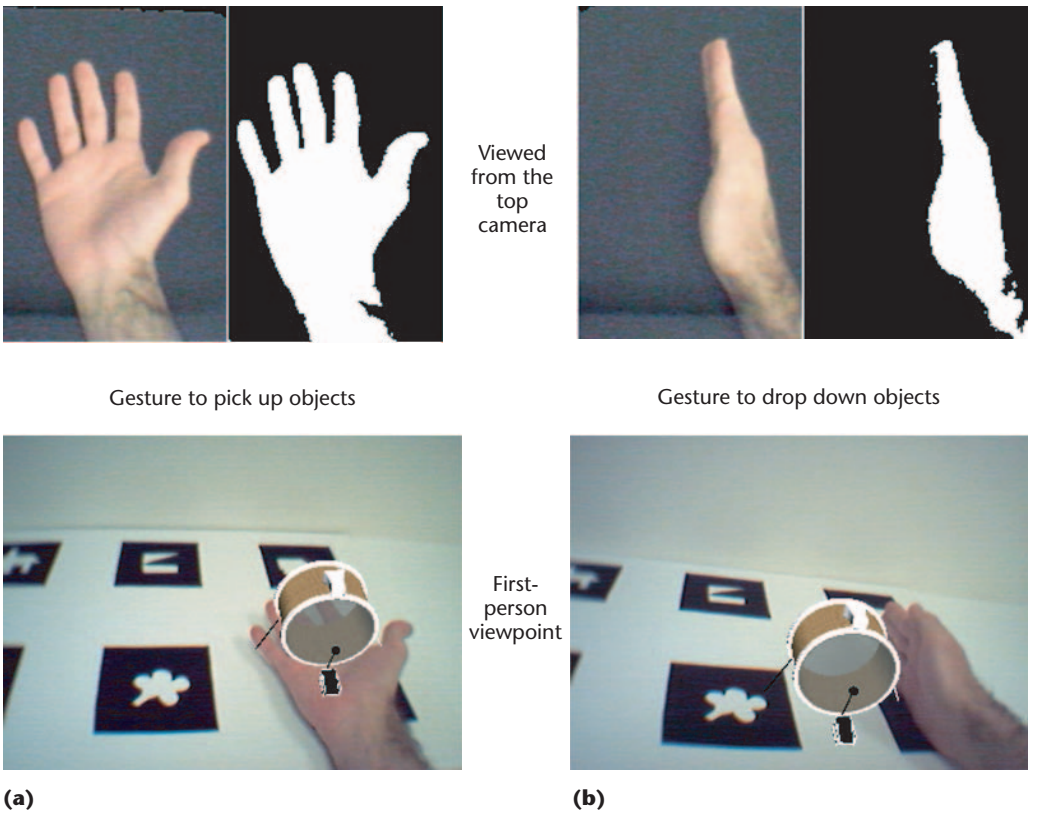

Figure 6. The Magic Music Desk recognizes two gestures to pick up and drop down virtual objects, (a) Gesture to pick up a virtual object when the palm is open and close to the object. User's hand is in the photo at the left; the black-and-white image on the right is the image after thresholding. (b) Gesture to drop down the virtual object when the edge of the user's hand is against the desk. User's hand is in the photo at the left; the black-and-white image on the right is the image after thresholding. (c) The user's view of picking up the virtual object. (d) The user's view of dropping the virtual object onto the desk.

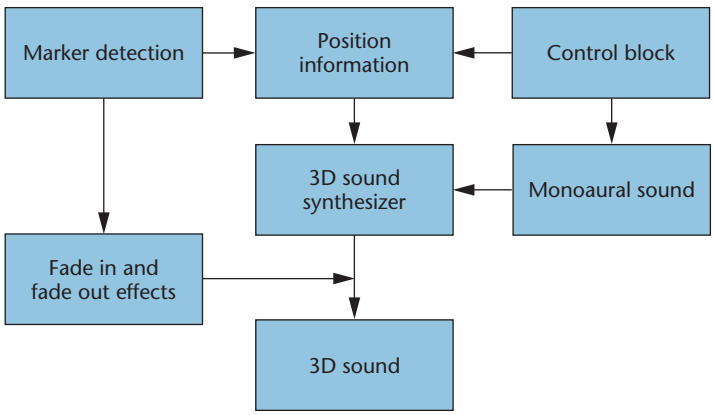

Figure 7. A flowchart depicting the components of the visual–3D sound interface.

by speech commands, the user can—by means of the speech command—play back and hear the 3D music. Figure 5b shows the user viewpoint of playing back music related to an instrument—a trumpet, in this case. By issuing a speech command "play trumpet now", MMD turns the trumpet into an animated, virtual trumpet player who starts to play the music. Meanwhile, MMD synthesizes 3D music, from the position of the virtual player, using OpenAL API (http://www.openal.org), so that the user hears it as if it were physically emanating from the trumpet player's position on the desk. Whenever users move the virtual player, either with a speech command like "move trumpet left" or with their hand to pick up and move it, the 3D sound is generated from the

Table 3. Summary of Magic Music Desk modalities, user experiences, and technical features.

Bimodal and Multimodal Perceptions or Actions	Modalities Involved	User Experiences	Interaction Features
Visual–speech (perception)	Vision, speech	—See the speech which is interpreted as a virtual character coming from the user's mouse, splash on the desk, and turn into a virtual musical instrument. —See the object manipulated visually according to the user's speech commands.	The "what you say is what you see" interface visualizes the user's speech in the form of a virtual object and in the form of a language character (speech commands can be translated into different languages). This enables a multicultural and multilanguage interaction.
Visual–nonspeech sound (perception)	Vision, 3D music	—See the four virtual players playing guitar, tambour, trumpet, and piano while hearing the 3D music as if it emanated from the instruments. —See the virtual object move while hearing the virtual sound source move, which follows the motion of the object. —Feel self become the focus of the auditory experience. This is user-specific; one user may feel sound come from the side while another feels it come from the back.	—Virtually see and hear the 3D graphic and sound enables a fully immersive AR experience. —The AR environment is visual as well as auditory.
Visual–tactile and visual–gesture (action)	Vision, touch	—Use the hand to pick up, move, and drop down the virtual objects, with two simple gestures. Although no tactile feedback occurs, the user perceives a visual–tactile sensation by seeing objects move whenever the hand moves.	—Physically interact with virtual objects. Body movement, especially of the hands, is essential. —Tangible interaction: Physically pick up, move, and drop the virtual objects. —Social interactions and collaboration between users are required.
Tactile–sound (perception)	Touch, 3D sound	—Move the objects with the hand while feeling the virtual 3D sound source move, which is attached to the object. The 3D sound becomes tangible and is easy to arrange in the 3D space.	3D sound becomes a tangible and visible object. Tangible interactions and collaborations between users are required.

character. Therefore, 3D music comes from the character the user sees. In this manner, users experience a visual–3D music perception.

Figure 7 shows a flowchart of this visual–3D sound interface. Because we fixed the camera on the HMD worn by the user, the MMD uses the position of the virtual object relative to the camera to calculate the azimuth, elevation, and distance of the virtual 3D sound source. The MMD then feeds these parameters into the 3D sound synthesizer to create virtual, 3D music. To increase the robustness of the sound system, we added fade-in and fade-out effects to avoid volume gaps when a new object is imported or moved away from the scene.

Table 3 summarizes the user experiences and perceptions, system modalities involved and interaction features in the MMD.

Experimental evaluation

To verify the effectiveness and usability of the Bubble Bumble and MMD multisensory systems, we conducted a formal user study. Our main goal was to examine user reaction and feedback on the effectiveness of multisensory integration; multimodal perceptions compared to experiences with traditional musical entertainment; the systems' physical and tangible interactions; social interactions between users; user cooperation and collaboration; and users' comparisons with other types of musical entertainment interfaces. The results suggest that by integrating multiple modalities, our systems increase the bandwidth of the human–computer communication and also integrate ubiquitous, tangible, and social user interactions for collaborative multisensory, musical entertainment experiences.

We selected 40 volunteers, 13 females and 27 males, all first-year students from the National University of Singapore with an average age of 21 years. Each was paid SGD $8 per hour for testing our systems and filling out the questionnaire. We asked the subjects to perform three sets of tasks in sequence, as follows.

- Traditional computer-music-making interface

1. Create a simple song using a traditional music game, "Music MasterWorks" (http://www.tucows.com/preview/199113.html).
2. Play the song on the computer.
3. Do steps 1 and 2 with a partner.

- Bubble Bumble

1. Create music and voice bubbles.
2. Collaborate to burst six bubbles (three for each user) captured from the floating bubbles.
3. Put the captured musical objects onto the timeline for playing back.

- Magic Music Desk

1. Import the four instrumental objects onto the desk.
2. Collaborate to arrange the positions of objects using your hands.
3. Adjust the size, position, and orientation of the objects by speech commands.
4. Play and stop the four instruments by speech commands.
5. Delete objects from the desk by speech commands.

Results: Physical interaction and tangible Interaction

We asked the subjects two questions:

- Question 1: Compare your feeling (better or worse) of being entertained by physically moving around to interact with musical resources, rather than to create, play, and hear computer music using a PC and screen. Why do you feel this way?
- Question 2: Compared with the computer-screen-based computer musical game interface, did our tangible musical entertainment systems offer a less, same, or more exciting experience than using the keyboard and mouse for interaction? Why?

As Figures 8a and 8b (next page) show, most subjects responded positively to our systems. Most reported a strong feeling of participation in our musical entertainment interface. As one subject wrote, "[I] became part of the interface due to full participation including body movements." Other reasons cited were "more realism," "more interactive," and "more involved."

Although our user studies are small and not

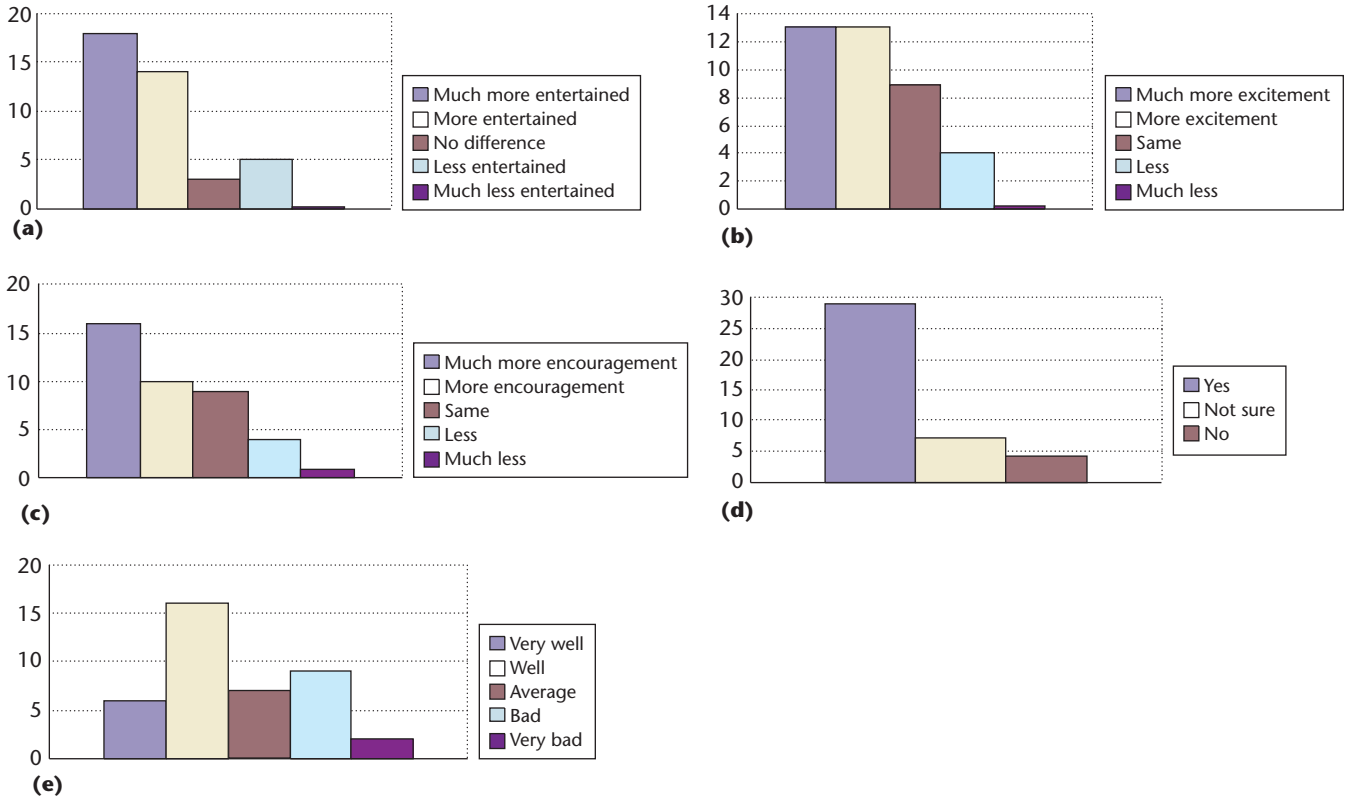

Figure 8. Quantitative results of questions. (a) Question 1, (b) results of Question 2, (c) results of Question 3, (d) results of Question 4, and (e) results of Question 7.

statistically large, these results partly prove that integrating physical body interaction with a musical interface can enhance personal entertainment. Both our systems offer many more opportunities for user involvement than those available with a computer interface experienced only by a person sitting in front of a screen.

Equally important to note are the negative comments. The subjects complained about the heavy HMD, dizzy feelings, fatigue from so much physical movement, and too many wires. These comments indicate that the hardware should be smaller, lighter, and possibly wireless to enable better entertainment experiences.

Results: Social interaction

We next asked:

- Question 3: Did the tangible AR systems encourage more or less cooperation and collaboration between users than the computer-screen-based music collaboration?

As Figure 8c shows, 16 subjects gave the positive answer of "much more encouragement" and 10 subjects claimed "more encouragement." Only four claimed "less" and one claimed "much less." The positive comments were: "Yes, very much so," "Yes, I need to burst the bubble with the help of my partner," "We were both composing the same song," and "Yes, when my partner is dealing with an object, I can choose the other." These results suggest that our musical entertainment systems encourage user collaboration.

Again, the negative comments are noteworthy. "The interactions between players are too simple." This comment implies that our systems need to be more sophisticated to provide more satisfaction.

Our fourth question was:

- Question 4: By visualizing speech in Magic Music Desk as a character and an object coming from the speaker's mouth, do you think it will prompt the interaction between users speaking different languages? Please give comments to your answer.

The result in Figure 8d suggests that most of the subjects (29 out of 40) agreed that MMD prompts multilanguage interaction. The positive comments include: "Yes, I know what my partner is saying even if I can't read the characters," and "The people speaking different languages

can see their own language character and the associated object on the desk." These results confirm the usability of our WYSIWYS interface.

Negative comments were: "Not really, there are only a few characters available," and "The speech commands are less flexible," "The accuracy of speech recognition is not satisfactory." These negative comments point out the direction of our future work—to extend both the model database and the speech command database, to apply a more flexible grammar, and to use more accurate speech recognition tools.

Results: Audiovisual experience

Next, we asked:

- Question 5: Does 3D sound improve your perception of immersiveness in the AR environment? Do you think it is more realistic to have 3D sound in audiovisual perception?

Almost all (39 out of 40) subjects claimed that 3D sound improves the immersiveness perception when sound is integrated with visual augmentation. A total of 38 subjects claimed that the experience is more realistic with 3D sound, and one subject said it was "a little better." The main reasons they provided to support their viewpoint include: "It is natural feeling and more realistic with 3D sound," "You can feel the sound from all directions which gives a more 3D feeling," "The feeling is closer to the real world and 3D sound helps me to navigate," and "I can hear and judge distance from the object producing the sound."

We then asked if 3D sound helped identify objects:

- Question 6: Does 3D sound help you to identify different spatial objects?

We applied different 3D sound—music—to different musical instruments in our systems. The responses to this question suggest that 3D sound, when applied to different spatial objects in an AR environment, can help users identify different spatial objects. A total of 36 subjects claimed that they can constantly be aware of different instrument objects simultaneously. As one subject noted, "yes, very much so. For instance, if one instrument is heard in the left earpiece of the user's earphones and the other is heard in the right earpiece, the directional information determines the positions of the different instruments."

Results: Product evaluation

In this section of our study, we asked volunteers to critique the systems in terms of product features:

- Question 7: How well do you think these will do as commercial products?

As Figure 8e shows, more than half of the subjects thought that our systems have the potential to be good commercial products. They regarded our systems as having more immersive, exciting, and entertaining interfaces than others they'd experienced. On the other hand, seven, nine, and two users responded "average", "bad," and "very bad," respectively. We partially found the answers in the next question:

- Question 8: For commercializing the Bubble Bumble and Magic Music Desk, what improvements should we make?

To summarize, the suggested improvements are mainly in the following categories:

- Hardware—"Lighter goggles," "fewer wires," and "wireless devices are needed."
- Software—"Better computer graphics," "more complicated tasks," and "speech recognition should be more accurate" (MMD).
- User interface—"It is not so friendly—sometimes I don't know what to do. More hints are needed, for example, the function of keys can be told when I press the button [Bubble Bumble]," and "it will be great if the hand can move in full dimensions."
- System database—"More characters in different languages" (MMD), and "bigger music database is needed."

Although the wireless HMD, wireless IS900 system, and wireless microphone and earphone have been commercially available for some time, the high price is still the biggest hindrance for commercial implementation. The issues addressed in the other categories point out the direction of our future work.

Summary

We conclude from the users' feedback that Bubble Bumble and MMD provide strong feelings

of participation, providing multiperceptional experiences of entertaining realism, interaction, and immersion; that the systems encourage cooperation and collaboration between users, and prompt multicultural interaction between users speaking different languages.

From the users' negative feedback, we can conclude that our systems could stand ergonomic improvement; designed to perform more sophisticated tasks and involve users to a greater degree; and be equipped with an easier user interface design featuring better speech recognition. Therefore, in the future version of this research we are going to concentrate on these aspects to improve our system designs for the future of multimodal musical interaction.

Conclusions

The concepts of ubiquitous and tangible computing presuppose that computers are embedded in our environment, in objects, and in the background. Similarly, social computing presupposes that the real-time and real-space activities of humans as social beings receive primary importance. Consequently, in our research we've applied the theories of ubiquitous, tangible, and social computing—together with mixed reality—to construct a unique musical entertainment space offering the exciting elements of computer-facilitated musical entertainment as well as natural, physical world interactions. In our musical entertainment systems, the real-world environment is essential and intrinsic.

The research[11-13] in multimodal multimedia systems has potential for professional musicians. Currently, our systems are used only for simple musical entertainment tasks, but they can be further developed for professional musical applications by introducing well-designed modeling of musical knowledge—for example, a tool for improvisational music composition in Bubble Bumble and a 3D design/preview tool of pictures and sound effects in MMD. A Web site featuring videos of this work can be seen at http://mixedreality.nus.edu.sg. **MM**

Acknowledgment

This project is supported by the Defence Science Technology Agency, Land Transport Division, Singapore.

References

1. D. Silbernagel, *Taschenatlas der Physiologie* [Bag Atlas of Physiology], Thieme, 1979 (in German).
2. J. Coutaz, "Multimedia and Multimodal User Interfaces: A Taxonomy for Software Engineering Research Issues," *Proc. East-West Human–Computer Interaction Conf. (*HCI 92), Int'l Centre for Scientific and Technical Information, 1992, pp. 229-240.
3. P. Lindsay and D. Norman, *Human Information Processing*, Academic Press, 1977.
4. L. Nigay and J. Coutaz, "A Design Space for Multimodal Systems: Concurrent Processing and Data Fusion," *Proc. INTERCHI '93*, IOS Press, 1993, pp. 172-178.
5. J. Schomaker, L. Nijstmans, and A. Camurri, *A Taxonomy of Multimodal Interaction in The Human Information Processing System*, tech. report, Esprit Basic Research Action 8579, 1995.
6. H. Kato and M. Billinghurst, "Marker Tracking and HMD Calibration for a Video-Based Augmented Reality Conferencing System," *Proc. 2nd IEEE and ACM Int'l Workshop on Augmented Reality*, IEEE CS Press, 1999, pp. 85-94.
7. H. Kato and M. Billinghurst, "ARToolkit," publications, tutorials, source code; http://www.hitl.washington.edu/artoolkit/.
8. C. Benoit et al., "Audio-Visual and Multimodal Speech Systems," *Handbook of Standards and Resources for Spoken Language Systems*, D. Gibbon, ed., Mouton de Gruyter, 1997.
9. T. Rosenberger and R.L.M. Neil, "Prosodic Font: Translating Speech into Graphics," *Proc. CHI 99 Extended Abstracts*, ACM Press, 1999, pp. 252-253.
10. W. Du and H. Li, "Vision Based Gesture Recognition System with Single Camera," *Proc. 5th Int'l Conf. Signal Processing* (WCCC-ICSP 00), vol. 2, IEEE Press, 2000, pp. 1351-1357.
11. J. Paradiso, "Electronic Music Interfaces: New Ways to Play," *IEEE Spectrum*, vol. 34, Dec. 1997, pp. 18-30.
12. A. Camurri and P. Ferrentino, "Interactive Environments for Music and Multimedia," *Multimedia Systems*, vol. 7, no. 1, 1999, pp. 3247.
13. F. Sparacino et al., "Augmented Performance in Dance and Theater," *Proc. Int'l Dance and Technology Conf.* (IDAT 99); http://ic.media.mit.edu/icSite/icpublications/Conference/AugmentedPerformance.html.

Zhiying Zhou, a PhD candidate, is a research fellow with the Department of Electrical and Computer Engineering, National University of Singapore. His research interests include augmented and virtual reality, computer vision, speech signal processing, tangible user interfaces, and multimodal

human–computer interaction. Zhou received a B.Eng and a M.Eng in instrument science and engineering from Southeast University, Nanjing, China.

Adrian David Cheok is director of the Mixed Reality Lab at the National University of Singapore. His research interests include mixed reality, human–computer interaction, wearable computers and smart spaces, fuzzy systems, embedded systems, power electronics, and multi-modal recognition. Cheok has a B.Eng and a PhD in electrical engineering from the University of Adelaide, Australia.

Wei Liu is a research engineer in the Department of Electrical and Computer Engineering at the National University of Singapore. Her research interests include augmented and virtual reality and novel human–computer interaction systems. Liu received a B.Eng in electrical engineering from Xian Jiaotong University and an M.Eng in electrical and computer engineering from National University of Singapore.

Xiangdong Chen is a research fellow with the Department of Electrical and Computer Engineering at the National University of Singapore. Currently his research interests include neural network, speech recognition, computer vision, augmented reality, human–computer interaction, and digital system design. Chen received a B.Eng from Xi'an Jiao Tong University, and an M.Eng and a PhD from the Institute of Semiconductors, Chinese Academy of Sciences.

Farzam Farbiz is a research fellow in the Department of Electrical and Computer Engineering at the National University of Singapore and is a consultant with the Singapore Defense, Science, and Technology Agency. His research interests include augmented and mixed reality, computer vision, and human–computer interaction. Farbiz received a PhD in electronics from Amirkabir University of Technology, Tehran, in Iran.

Xubo Yang is an associate professor in the Department of Computer Science and Engineering, Shanghai Jiaotong University, China, where he also serves as vice chair of the Digital Entertainment Department in the Software School. His research interests include visualization, mixed reality, mobile computing, and human–computer interaction. Yang has a B. Eng, an M. Eng, and a PhD in computer science from Zhejiang University, HangZhou, China.

Michael Haller is an assistant professor in the Media Technology and Design Department at the Upper Austria University of Applied Sciences. His research interests include real-time computer graphics, augmented reality, virtual reality, and human–computer interaction. Haller received a master and a PhD of science from the University of Linz, Austria.

Readers may contact Adrian Cheok via email at adriancheok@nus.edu.sg.