ONLINE ACTIVITIES FOR MUSIC INFORMATION AND ACOUSTICS EDUCATION AND PSYCHOACOUSTIC DATA COLLECTION

Travis M. Doll **Ray V. Migneco** **Youngmoo E. Kim**
Drexel University, Electrical & Computer Engineering
{tmd47,rm443,ykim}@drexel.edu

ABSTRACT

Online collaborative activities provide a powerful platform for the collection of psychoacoustic data on the perception of audio and music from a very large numbers of subjects. Furthermore, these activities can be designed to simultaneously educate users about aspects of music information and acoustics, particularly for younger students in grades K-12. We have created prototype interactive activities illustrating aspects of two different sound and acoustics concepts: musical instrument timbre and the cocktail party problem (sound source isolation within mixtures) that also provide a method of collecting perceptual data related to these problems with a range of parameter variation that is difficult to achieve for large subject populations using traditional psychoacoustic evaluation. We present preliminary data from a pilot study where middle school students were engaged with the two activities to demonstrate the potential benefits as an education and data collection platform.

1 INTRODUCTION

Recently, a number of web-based games have been developed for the purpose of large-scale data labeling [1]. Similar activities can be used to collect psychoacoustic data from a large number of users, which is difficult to obtain using traditional evaluations. We present two such activities that explore the perception of audio (instrument timbre and the cocktail party problem), with the additional aim of educating users, particularly K-12 students, about aspects of music and acoustics. These web-based interfaces are designed as game activities with minimal complexity so they can be easily used by students without previous training. To maintain widespread accessibility, the activities require only internet access through a web browser and run independently of external applications. We believe this platform will enable us to collect a very large number of samples exploring myriad parameter variations to better define perceptual boundaries and quantitative models of perceived features.

2 BACKGROUND

Relatively little research has been conducted on human performance in the identification of musical instruments after timbral modifications. Saldanha and Corso demonstrated that the highest performance is achieved when the test tone consists of the initial transient and a short steady-state segment and the lowest performance occurs using test tones consisting of only the steady-state component and the ending transient [2]. Iverson examined the dynamic attributes of timbre by evaluating the effects of the transient in sound similarity tests. While results of instrument identification experiments depended on the presence of the initial transient in the sound, the results of similarity tests using tones with and without the initial transient suggest that the initial transient is not required to imply similarity. This research suggests that similarity judgments may be attributed to acoustic properties of an instrument's sound other than the transient component [3]. Martin demonstrated that humans can identify an instrument's family with greater accuracy than the instrument itself [4]. Other studies on instrument identification suggest that musically inclined test subjects perform better than non-musically inclined subjects and in particular, subjects with orchestra experience perform better than those without [5].

The performance of automatic speaker and speech recognition algorithms often use human performance as a benchmark, but it is difficult to obtain a large human subject population for comparisons. Atal [6] provides a summary of human speaker recognition evaluations before 1976 in which no more than 20 human listeners are employed for each evaluation. Stifelman [7] performed a speech recognition evaluation that simulated the environment of the cocktail party problem, testing the listening comprehension and target monitoring ability of 3 pilot and 12 test subjects. Lippmann [8] provides a summary of several human vs. machine speech recognition comparisons, all of which distinctly show humans outperforming machines with a limited number of test subjects. In the area of speaker recognition, Schmidt-Nielsen [9] conducted an evaluation in which 65 human listeners were tested with speech data from the 1998 NIST automatic speaker recognition evaluations. The experiment was administered in the same manner as the automated sys-

tems to create a direct comparison, and the results show that humans perform at the level of the best automated systems and exceed the performance of typical algorithms.

3 DEVELOPED ACTIVITIES

3.1 Timbre Game

The Timbre Game is an online activity designed to illustrate the effects of particular acoustic modifications on musical instrument timbre and to collect evaluation data on the perception of timbre. The user interface for the Timbre Game was developed in Java and is accessed by downloading applets within a client's web browser, requiring no additional software. The Timbre Game has two user interfaces, labeled "Tone Bender" and "Tone Listener". The objective of Tone Bender is primarily educational in that a player is allowed to modify time and frequency characteristics of musical sounds and immediately hear how those sounds are affected. Upon modification, the player submits the resulting sound they have created to a server database. In Tone Listener, other players will listen to the sounds whose timbre has been modified by a player in the prior component. Players will then attempt to determine the original instrument source from the modified timbre. Points are awarded to both players (modifier and listener) if the listening player enters the sound source's identity correctly. A more detailed description of each component of the activity follows.

3.1.1 Tone Bender

The Tone Bender game involves time and frequency analysis of a single instrument sound and provides a visual interface which allows a player to modify the sound's timbre. The sounds available for analysis are 44.1 kHz recordings of various musical instruments, each producing a single note with a duration of less than five seconds.

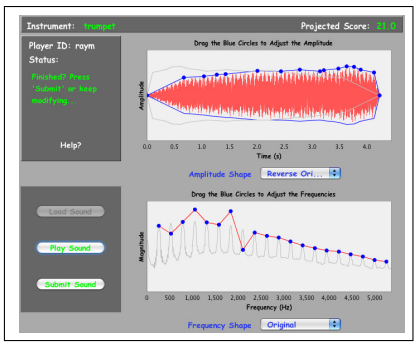

Figure 1. The Tone Bender interface

A player starts a session in Tone Bender by requesting a sound file from the server database. The sound is then analyzed in both the time and frequency domains to generate control points suitable for modifying the sound's timbre. In the time domain analysis, a heuristic method is employed to approximate the sound's amplitude envelope by picking amplitude peaks within small time intervals of the sound wave. These peaks are used as the initial amplitude envelope control points.

In the frequency domain, the sound is analyzed via an efficient Short-Time Fourier Transform (STFT) implementation optimized for Java, using 45 msec Hanning windows with 50 percent overlap. The STFTs are used to generate a time-averaged spectrum of the sound. Linear prediction is employed to establish a threshold, which is used to extract twenty of the most prominent spectral peaks from the time-averaged spectrum. These spectral peaks are used as the initial control points for the harmonic weights that the player will modify.

The visual representation of the sound's timbre is displayed to the player in two separate 'XY' plots in the Java applet as shown in Figure 1. In the amplitude plot, the sound wave is shown with the extracted amplitude control points. The player is allowed to manipulate the shape of the amplitude envelope as they wish by clicking and dragging the control points within the plot window. In the frequency plot, the time-averaged spectrum of the sound wave is shown with the spectral control points. The player is allowed to move the control points vertically so that they are only adjusting the harmonic weights of the sound without affecting pitch. After modifying the spectral envelope, the sound wave is resynthesized using additive sinusoidal synthesis and redrawn in the time plot.

After each modification in Tone Bender, the player is presented with a potential score based on the difference between the original sound and the altered sound, calculated using the signal-to-noise ratio (SNR):

$$SNR = 10 \log_{10} \left| \sum_{n=0}^{N} \frac{p[n]^2}{(p[n] - \widehat{p}[n])^2} \right| \qquad (1)$$

This score is intended to reflect the potential difficulty in correctly identifying the original instrument from the modified sound where $p[n]$ and $\widehat{p}[n]$ are the original and modified sounds, respectively. The resulting difficulty score ranges from 1-25, where 1 corresponds to a high SNR (little change) and 25 represents a low SNR (significant change). The player has the incentive to modify the timbre of the sound as greatly as possible while still maintaining the identity of the original instrument. They will be awarded points based on the difficulty score of their modifications only if a listener can correctly guess the original instrument. This encourages the player to be creative with their timbre adjustments yet still produce recognizable musical sounds.

When a player is finished altering the timbre of a specific instrument, they submit their information, including user ID and modified time and spectral envelopes, to the server

database, which collects all the players' modified sounds. The player can then load another sound from the database to continue with timbre modification.

3.1.2 Tone Listener

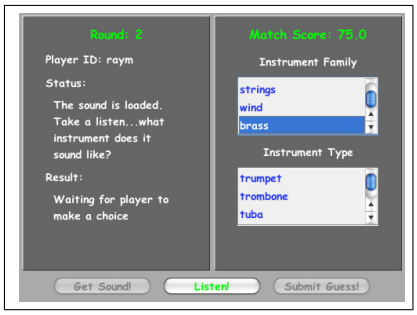

Figure 2. The Tone Listener interface

In the Tone Listener interface, a player is presented with a Java applet that allows them to listen to sounds created by other players in the Tone Bender component. A player's objective is to correctly guess the family and identity of the instrument from the modified sound with which they are presented. The modified sounds are randomly selected from the database and the sounds are re-synthesized in the applet using the amplitude and spectral envelope data. The player is allowed to listen as many times as needed before submitting their guess.

The listening player is allowed to classify the sound among three instrument families: strings, wind, and brass. The player's choice populates another list consisting of individual instruments. After the player has listened to the sound and made a selection for the family and instrument, they submit their guess. The player will receive points only if they correctly guess either the instrument family or the specific instrument. If the player correctly guesses an instrument, they receive a score proportional to the difficulty rating. If the player correctly guesses within the instrument family, they receive half of the potential point value. After each guess, the results, including the user ID, original sound information and the player response are uploaded to a server database containing all players' listening results.

3.1.3 Educational objectives

The Timbre Game is designed to educate players about sound timbre, particularly the importance of time- and frequency-domain envelopes. The interface does not require any previous background in engineering or music, and the ability to listen to timbral changes in real-time encourages the user to learn by experimentation. Additionally, the simple scoring method is designed to provide players with an indication of the degree of change imparted upon the sound, while concealing technical details, such as SNR.

3.2 Cocktail Party Game

The Cocktail Party Game is a web-based activity designed to collect data from listeners on how source locations and acoustic spaces affect identification of a speaker and the intelligibility of speech. It also provides a method of examining the perception of complex timbres, such as the dynamically varying sounds of voices. This game consists of two components: room creation and listening room simulators. The room creation component introduces the concepts of the cocktail party problem and illustrates the effects of reverberation and interfering sounds. The listening room component evaluates the ability of a listener to detect a known person's voice within a room mixture. The two components are described in further detail in the sections that follow.

3.2.1 Room Creation

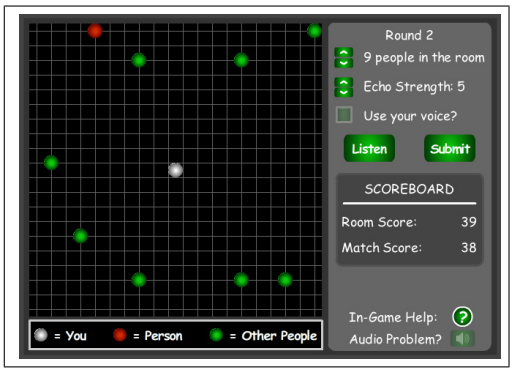

Figure 3. The Room Creation interface

In this component of the game, the player simulates a "cocktail party" situation by positioning multiple talkers, including the target person of interest, in a reverberant room, thus making it more difficult to hear the voice of the target speaker. The goal is to create a situation where the speaker of interest is obscured, but still identifiable, and more points potentially will be awarded to the player based on the "degree of difficulty" of the designed room. Initially, the game displays a room (20' x 20' x 10') containing two people: the listener and the person of interest, represented as white and red circles, respectively. The player has the option to add or remove people from the room, change the strength of the room reverberation, and alter the position of the people in the room. Audio for each of the speakers in the room is randomly drawn from the well-known TIMIT speech database [10]. The browser-based user interface communicates with the server to download the relevant speech files.

A room's potential score is based on the resulting SINR, treating the target voice as the signal and the others as interferers. These points are only added to the player's score if another player correctly determines if the speaker is in the room.

3.2.2 Room Listening

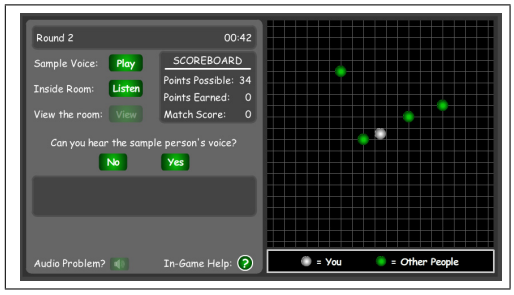

Figure 4. The Listening Room interface

In the listening component of the game, the player's goal is simply to determine if the target person's voice is in the mixture of voices in the room. The game communicates with the server to obtain parameters and sound sources for a previously created room. The game randomly determines whether or not the target voice will be present in the room. The player then listens to the mixed room audio with the option to graphically view the configuration of the people in the room. The room audio is generated in exactly the same manner as in the room creation component. Then the player decides whether or not the sample person's voice is in the room.

After the player decides if the person of interest is in the room, the response is sent to the server and the player is informed of the correct answer. The points are based on the difficulty of the room as calculated by the SINR and only awarded when the person chooses the correct answer. After the information is submitted, the player continues on to the next round until all rounds are completed for the current match.

3.2.3 Educational Objectives

The Cocktail Party Game is designed to educate and inform students about the cocktail party effect and basic room acoustics. Like the Timbre Game, the controls are designed to be simple and minimal so that players can easily experiment with sound source locations and listen to the results. The graphical layout of the game represents actual acoustic space so that players can visually correlate speaker positions with the resulting sound. The activity is intended for players without any specific training in engineering or music, which broadens its appeal.

3.3 Technical implementation details

Both the Timbre Game and Cocktail Party game employ a client-server architecture that allows distributed online game play, so that players of both games may be widely separated. The data for both activities is stored and served from a web server using a MySQL database. PHP web server scripts on the server respond to client queries with XML formatted responses from the database, such as sound and room parameters and audio file URLs.

The Timbre Game is a single Java applet, containing the interface, sound processing, and server communication code. The overall client-server architecture for the Timbre Game shown in Figure 5.

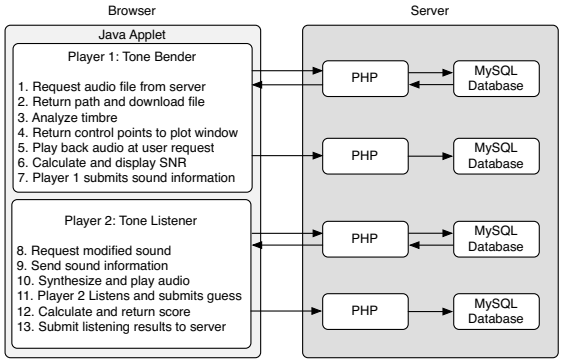

Figure 5. Diagram of Timbre Game

The user interface for the components of the Cocktail Party Game is implemented in Adobe Flash, which has rather limited audio processing and computational capabilities. Therefore, it was necessary to also use a Java helper applet for audio playback flexibility and its concurrent computation capabilities using threading, which is generally sufficient to handle the computation required for room acoustics simulation. Calculation and playback of the room audio in both components of the game is initiated via a call in Flash that sends the room configuration and speaker audio file paths to the Java applet via a JavaScript bridge. The Flash application also communicates with the server to obtain game parameters and audio file paths using Asynchronous JavaScript and XML (AJAX) calls.

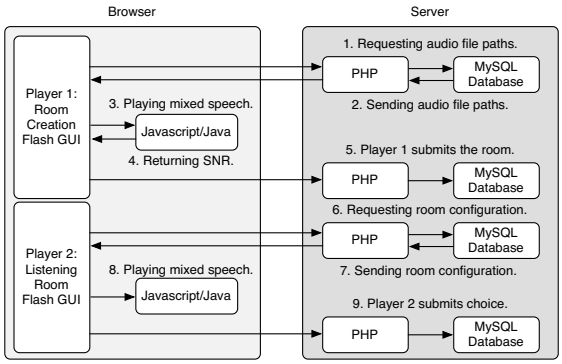

Figure 6. Diagram of Cocktail Party Game

The generation of the room audio for listening is a multi-

step process, requiring multiple components and conversions. First, the room impulse response for each source speaker location is determined based on the person's position in the room using a Java implementation of the well-known room image model [11]. Next, the resulting room impulse response is convolved with the respective speech audio using fast block convolution via the FFT in blocks of approximately 8192 samples ($\sim$ 0.5 seconds at 16 kHz sampling). An efficient Java FFT library was used to optimize the calculation speed by employing concurrent threading. In the final step, the convolved, reverberant speech from each source is combined to obtain the overall room audio for the current configuration. This audio is then played through the Java applet in the client's browser. The overall architecture of the Cocktail Party Game is given in Figure 6.

4 ACTIVITY EVALUATIONS

Evaluations of both activities were performed on a population of 56 eighth grade students attending a magnet school specializing in music performance. Activity sessions focused on the Room Simulation Game and the Timbre Game on separate days. The students were divided into groups of approximately 10 students for each of six sessions lasting 40 minutes per day. The students were switched from the creative component of the game to the listening component midway through the session to have an opportunity to both create and objectively listen to the sounds. The students played the games alone using headphones to avoid confusion in the sound creation and listening processes.

Prior to playing each component, the students were given a 2-3 minute demonstration covering the game objectives, instructions, and user controls. The students were given the opportunity to ask questions throughout the sessions.

4.1 Quantitative results

4.1.1 Timbre Game

Figure 7 provides the results of 800 listening trials from the Tone Listener game: percentage of correct identification of the instrument and instrument family vs. varying SNR levels. The plots demonstrate a very slight upward trend in percentage of correct detection with increasing SNR, with the percentage of correct family detection being greater than correct instrument detection across all SNR values. This result is expected since, in general, it is easier for a listener to identify an instrument's family. The wide variation in performance, however, is likely due to the difference between SNR as a measure of sound quality and actual perceived sound quality. It should be noted that the majority of sounds listened to were created with an SNR value under 20 dB due to players seeking to earn a high modification score by creating more difficult instrument sounds.

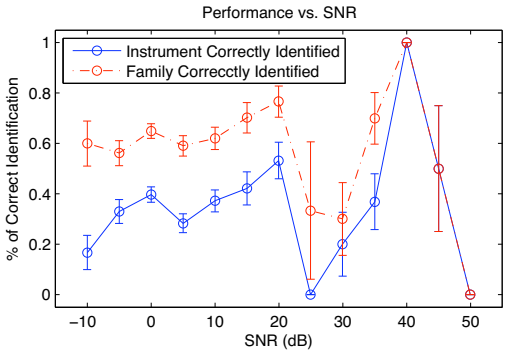

Figure 7. Timbre Game Results

4.1.2 Cocktail Party Game

The results of 817 unique listening tests were analyzed and the results are presented in Figure 8. As the figure shows, the percentage of correct detection generally increased with the SINR. This result is somewhat expected since rooms with higher SINR indicate situations where the greater energy of the target listener should make them easier to identify. This upward trend, however, was not strictly monotonic, indicating that factors other than SINR affect overall performance. The confusion matrix indicates that false negatives were far more likely than false positives, an interesting result that warrants further study.

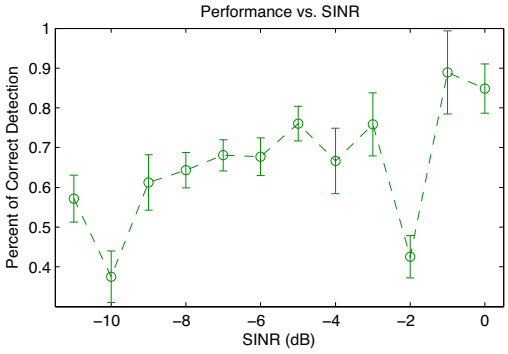

Figure 8. Cocktail Party Game Results

		Correct Answer	
		In Room	Not In Room
Player Guess	In Room	229	84
	Not In Room	215	289

Table 1. Cocktail Party Evaluation confusion matrix

4.2 Qualitative Observations

The Tone Listener interface had the broadest appeal among the students and provided the most instant gratification. This was likely due to the simple objective of the activity, which only required them to listen to sounds and guess the original instrument producing them. Additionally, the instant feedback and scoring added a distinct competitive aspect that encouraged the students to keep playing and comparing scores with each other. In the case of Tone Bender, student reactions to the game objective were mixed. Some students appeared intimidated by or uninterested in the visual representations of instrument timbre. This behavior was evident in students repeatedly asking the demonstrators (graduate students) to explain the game or simply not participating. The students who were more engaged with the activity, however, attempted to modify many different sounds without assistance.

Similarly, the room creation component of the Cocktail Party Game raised more questions and requests for clarification from the students than the listening component. This was expected since the game requires students to be creative and to achieve some understanding of the task in order to successfully design a challenging room. The activity could be improved by altering the audio processing chain so that the room audio responds in real-time to game parameter changes (position and number of sources, etc.). Then the player would receive instant audio feedback, reducing the time to iterate a particular room design. The lack of information provided to room creators regarding the performance of their rooms when listened to by other players was also frustrating and reduced one of the motivating competitive aspects of the game. Overall, the room creation component may need to be simplified in order for middle school students to better understand the objectives of the activity.

5 FUTURE WORK

The websites for both activities will eventually be made publicly available through the web. Another improvement we believe will enhance the activities is to allow players to record their own sounds for immediate use in the games. For examples, an instrumentalist could record their own instrument to be used in the Timbre Game, and players could record their own voices for use in the Cocktail Party game. This feature would enable continuous expansion of the sound databases, keeping the game fresh for frequent users.

A relatively straightforward extension of the Cocktail Party Game would be to extend the sound sources to musical instruments, providing a method of examining the perception of individual instruments within mixtures. We are also investigating the utility of time limits for different phases of the game, in order to keep the activity moving and to increase competition. We particularly wish to pursue a detailed analysis of acquired performance data for cases that deviate from anticipated "difficulty" in terms of SNR. We plan to investigate other acoustic factors affecting listening performance as well as other metrics that may be better correlated to perceptual task performance than SNR.

6 ACKNOWLEDGEMENTS

This work is supported by NSF grants IIS-0644151 and DGE-0538476.

7 REFERENCES

[1] L. von Ahn, "Games with a purpose," *Computer*, vol. 39, no. 6, pp. 92–94, 2006.

[2] E. Saldanha and J. Corso, "Timbre cues and the identification of musical instruments," in *Journal of Acoustic Society of Ameria*, 1964, pp. 2021–2026.

[3] P. Iverson and C. L. Krumhansl, "Isolating the dynamic attributes of musical timbre," in *Journal of Acoustic Society of Ameria*, vol. 94, no. 5, 1993, pp. 2595–2603.

[4] K. Martin, "Sound-source recognition: A theory and computational model," Ph.D. dissertation, Massachusetts Institute of Technology, 1999.

[5] A. Srinivasan, D. Sullivan, and I. Fujinaga, "Recognition of isolated instrument tones by conservatory students," in *Proc. International Conference on Music Perception and Cognition*, July 2002, pp. 17–21.

[6] B. S. Atal, "Automatic recognition of speakers from their voices," vol. 64, no. 4, 1976, pp. 460–475.

[7] L. J. Stifelman, "The cocktail party effect in auditory interfaces: a study of simultaneous presentation," in *MIT Media Laboratory Technical Report*, 1994.

[8] R. Lippmann, "Speech recognition by machines and humans," in *Speech Communication*, vol. 22, no. 1, 1997, pp. 1–16.

[9] A. Schmidt-Nielsen and T. H. Crystal, "Human vs. machine speaker identification with telephone speech," in *Proc. International Conference on Spoken Language Processing*. ISCA, 1998.

[10] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, August 1990.

[11] J. Allen and D. Berkley, "Image method for efficiently simulating small room acoustics," in *Journal of Acoustic Society of Ameria*, April 1979, pp. 912–915.