Multimodal Speech and Audio User Interfaces for K-12 Outreach

Mark Hasegawa-Johnson, Camille Goudeseune, Jennifer Cole, Hank Kaczmarski, Heejin Kim,
Sarah King, Timothy Mahrt, Jui-Ting Huang, Xiaodan Zhuang, Kai-Hsiang Lin,
Harsh Vardhan Sharma, Zhen Li, and Thomas S. Huang
University of Illinois at Urbana-Champaign
{jhasegaw,cog,jscole,kacmarsk,hkim17,sborys,tmahrt2,jhuang29,xzhuang2,klin21,hsharma,zhenli3,t-huang1}@illinois.edu

***Abstract*—Elementary school children have short attention spans. This paper describes three multimodal speech and audio user interfaces that captured and held the attention of a few dozen elementary-school and high-school children during the course of a two-day university open house. The *Speech Recognition Game* demonstrated an isolated word recognizer with a rapidly-won game, in which children were challenged to get ten words in a row correctly recognized. The *Audio Easter Egg Hunt* demonstrated our *timeliner* multimedia analytics platform with a faster-than-real-time search through orchestral music for audio anomalies (cuckoo clocks, motorcycles, etc). Finally, at the *Intonation Station*, children had to pick the pitch contour that would help a friendly troll to successfully hunt dragons in the city of Champaign. Results suggest that competition, collaboration, and other forms of social interaction may motivate children more than prizes.**

***Index Terms*—Multimedia analytics, spoken language user interface, prosody, intonation**

I. Introduction

The Beckman Institute Open House [14] is a biennial outreach event at the University of Illinois, held in conjunction with the annual College of Engineering Open House [15], at which laboratories demonstrate their research in a style that is educational and entertaining for children and adults from the surrounding community. The Engineering Open House is more than a century old; the first Department of Physics Open House in 1906 was followed by an Electrical Engineering Open House in 1907, which attracted more than 1600 visitors (much to the surprise of its organizers). The open house is usually held on a Friday and Saturday. Visitors on Friday are usually elementary school classes; visitors on Saturday are usually families. The open house aims to generate excitement and to encourage children to pursue careers in the fields of science, technology, engineering and mathematics. The goal of the demonstration systems described in this paper was to educate children and to generate excitement specifically about the science, technology, engineering and mathematics of speech, audio, language and learning information environments.

Speech signal analysis has been used in educational software for at least twenty-five years [5], [7], [8], [10], [13], [18], and is now a necessary user interface technology for a wide variety of tutoring systems and computer games. There is therefore a considerable literature on the use of speech technology to teach children, but there is less consensus on the best methods to be used to teach children *about* speech technology. The demonstration systems described in this article were created without reference to the state of the art, because we are not sure that any state of the art has yet been defined in this field. Systems were therefore designed to incorporate our best understanding of standard educational principles, e.g., as expressed by [3]:

1) *Communicate high expectations*: Even in the chaos of a university open house, it is essential that children understand what they are supposed to be learning from each exhibit.
2) *Encourage active learning*: In the chaos of a university open house, a child will only pay attention to a task in which she is the protagonist.
3) *Give prompt feedback*: Children love games, because a game tells the child immediately whether or not she has succeeded in the assigned task.

This article describes three open house exhibits designed to teach children about the Science, Technology, Engineering and Mathematics of Speech, Audio, Language and Learning Information Environments. The *Speech Recognition Game* was exhibited at the 2009 Beckman Open House, and built on our previous attempts to use a speech recognizer to teach speech technology. The *Intonation Station*, exhibited in 2011, taught the pragmatics of prosody to elementary and high school students using an interactive dragon-hunting game created entirely in Microsoft Powerpoint. The *Audio Easter Egg Hunt*, also exhibited in 2011, demonstrated the utility of time-frequency signal processing by challenging children to find audio anomalies (cuckoo clocks, mooing cows, motorcycles) in a two-hour classical music recording.

II. The Speech Recognition Game

The *Speech Recognition Game* was developed as a testbed for automatic speech recognition (ASR) designed for people

with gross motor disability, e.g., for users with Cerebral Palsy. Word recognition accuracy (WRA) of a speaker-dependent ASR currently exceeds 99% for the most successful speakers; for example, the winner of the 2007 United States National Book Award for fiction, *The Echo Maker*, was dictated using ASR. Many adults with gross motor impairment, however, can use neither a keyboard nor ASR, because their impairment includes components of both manual disability and of *dysarthria*: reduced speech intelligibility caused by neuromotor impairment. One of our ongoing research projects seeks to develop ASR that is effective despite the distortions of dysarthria [17], [16]. The *Speech Recognition Game* was developed initially so that people with dysarthria (typically users with Cerebral Palsy) could rapidly test speaker-dependent and speaker-adaptive ASR models. In order to showcase this research, we exhibited the *Speech Recognition Game* at the 2009 open house.

A. Motivation

The *Speech Recognition Game* was not the first time that we attempted to use a speech recognizer in a Beckman Open House exhibit. In 2003, for example, we exhibited our prosody-dependent speech recognition system [2]. Word recognition accuracy, of course, was not very high in the noisy environment of an open house, but a worse limitation was the "so what?" problem: children would talk to the speech recognizer, it would print a sentence that more or less resembled what they said, and then the child would give the verbal or non-verbal response, "so what?" The *Speech Recognition Game* attempted to solve this problem by turning speech technology into a game.

B. Design

The UA-Speech corpus (http://isle.illinois.edu/UASpeech) contains recordings of thirty-six subjects, including seventeen subjects with dysarthria [9]. Each subject recorded three blocks of isolated words, with rest breaks between blocks. The core words included digits ("zero" through "nine"), letters in the international radio alphabet ("alpha, bravo, charlie, . . . "), nineteen computer commands ("command, enter, paragraph, . . . "), and the one hundred most common words in the Brown corpus of written English ("is, it, . . . ") [11]. The uncommon words were selected from children's novels digitized by Project Gutenberg (e.g., *Wizard of Oz*, *Peter Pan*) to maximize phone bigram diversity. Each subject recorded a total of 765 words, including 455 distinct words.

Talker-dependent HMM-based speech recognizers employing three configurations were developed and tested using the UA-Speech corpus: whole-word HMMs, monophone HMMs, and triphone HMMs. Whole-word HMMs used six states per word, while triphone HMMs used three states per triphone; each used two Gaussians per state. Results showed [17] that for talkers with intelligibility below 50% (less than 50% of their words correctly transcribed by human listeners), ASR generally outperforms human listeners. ASR had two important advantages over the human listeners: (1) it knew the vocabulary from which test items were being drawn, and (2) it knew the talker, in the sense that acoustic models were either speaker-dependent or speaker-adaptive. For very small vocabularies, therefore, ASR outperformed human listeners regardless of talker intelligibility, e.g., for talkers with an intelligibility of 25-75%, ASR was nevertheless able to recognize isolated digits with 90-100% word recognition accuracy. Subjects with intelligibility below 25% had much lower ASR accuracy in all tasks, but the automatic system almost always outperformed human listeners.

To test the utility of our isolated word recognition algorithms, we developed a *speech recognition game*. The player tries to raise the level of a green bar until it hits the top (ten correct recognitions). Prompt words are displayed, and the player pronounces each. Correctly recognized words raise the green bar. Incorrectly recognized words may be read again, manually replaced (by selecting from an N-best list using a button-press interface), or skipped, depending on the type of game the child has chosen to play. Three types of game were deployed: a game in which the child could respond by typing, a game in which the child had to accept the one-best ASR output, and a game in which the child could select the best output from an N-best list.

The user interface for the *speech recognition game* was programmed using Microsoft Visual C++. The back end speech recognition engine ran HVite [19] with pre-trained speaker-independent acoustic models. For the open house, two new sets of acoustic models were trained. Acoustic models for children were trained using the CMU-Kids speech corpus [6]. Acoustic models for adults were trained using TIDIGITS (and adults were therefore constrained to speak using only a ten-word vocabulary). To make the game more colorful, a set of large USB buttons (red, green, and yellow) was used to control all non-speech components of the user interface (Fig. 1).

C. Results

At the 2009 open house, more than 200 members of the public (about 150 children, 50 adults) took turns playing the *Speech Recognition Game* (Fig. 2). Most won the game with little trouble; first-pass word recognition accuracy trended around 90%. Children generally walked up to the game, played it once or twice, were given a sticker as a prize, then walked on to the next exhibit. Children who came to the exhibit in a group, and who competed against one another, were far more likely to be engaged in game-playing than were children who approached one at a time.

Fig. 1. Hardware setup for the *Speech Recognition Game*. Red, green and yellow external buttons were used to accept all non-speech user inputs. A green bar, displayed on the computer monitor, rose with each recognized word, to peak at ten correct recognitions.

Fig. 2. At the 2009 open house, more than 200 members of the public attempted the *Speech Recognition Game*. Most won with little trouble; first pass word recognition accuracy trended around 90%.

III. Intonation Station

The goal of the *Intonation Station* was to explain the pragmatics of prosody to children and teenagers. When explaining prosody to undergraduates, we often begin with examples in which the presence vs. absence of a contrastive focus or phrase break clearly changes the meaning of the sentence. In order to teach prosody to K-12 students, the *Intonation Station* uses similar nuance shifts to change the narrative sequence of a game.

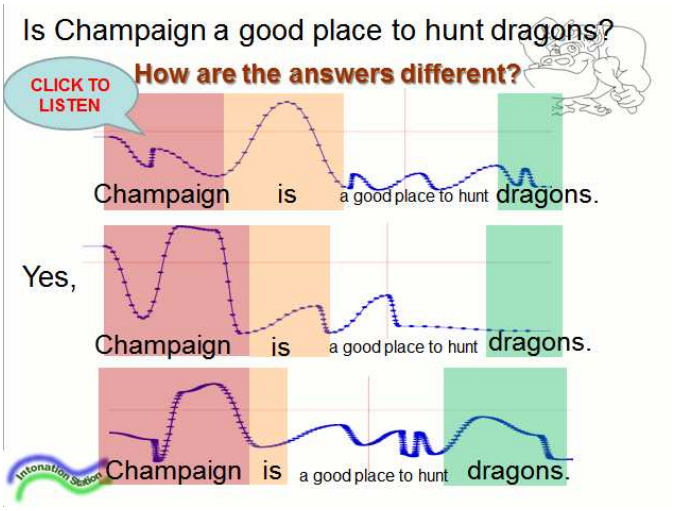

Fig. 3. The *Intonation Station* was designed to teach contrastive focus to children. A game player encounters a troll. Depending on the player's response, the troll becomes either cooperative or confused.

A. Design

When children walk up to the *Intonation Station*, they are greeted by a troll, who asks them, "Is Champaign a good place to hunt dragons?" (the Beckman Institute is situated on the border between two small cities, Champaign and Urbana). When the child clicks on an arrow key, she is taken to a screen that displays pitch contours (F0 as a function of time), transcriptions, and waveform links for three possible responses (Fig. 3). Children were invited to choose the most appropriate response. Responses differed in the location of the pitch accent:

1) Champaign IS a good place to hunt dragons.
2) CHAMPAIGN is a good place to hunt dragons.
3) Champaign is a good place to hunt DRAGONS.

All three responses contain the same words, but pragmatically, only response #1 is appropriate. Response #2 communicates that, although Champaign is a good place to hunt dragons, one should be careful to stay within the borders of the city, because outlying regions are less hospitable. Response #3 communicates that the dragons are easy to hunt, but one should beware of the gryphons and unicorns. If a child chooses to respond using response #2 or #3, the troll character becomes confused, suspicious and somewhat angry, and asks for clarification; if a child chooses to respond using response #1, then the game continues without interruption.

We initially planned to implement this game using Visual C++, but we realized that all necessary user interface components could be rapidly created in Powerpoint. The entire game was therefore created as a Powerpoint stack.

B. Results

Children traveling alone or with uninterested parents did not volunteer to play the *Intonation Station*, but the game was played by children traveling in groups (e.g., school trips) and by children traveling with their parents. Most children found the game too easy, and abandoned it after successfully passing the first prompt. Although it was not terribly successful as a game, it was successful as a teaching tool and as a conversation starter. Everyone who attempted the game immediately grasped the meaning of the F0 contours (either alone or with a parent's help). Those who continued the game through multiple levels usually did so because they wanted to learn more about contrastive focus. Generally, children actively participating in groups of more than one person (e.g., groups of children playing the game together and/or children playing the game together with an actively participating parent) were more likely to play past the first level than were children playing alone.

IV. Audio Easter Egg Hunt

The human ear detects anomalous audio rapidly and with high accuracy. For example, rifle magazine insertion clicks are detected with 100% accuracy at 0 dB SNR in white noise, babble, or jungle noise [1]. Unfortunately, most people are only capable of listening to one sound at a time, making the rapid browsing of large audio databases much harder than the browsing of image, video, or text data. Audio visualization tools have the potential to show a user many sounds at once, possibly allowing him or her to detect an anomaly several orders of magnitude faster than "real time." Our *Timeliner* application (Fig. 4) is an audio visualization interface that lets the user view a multi-hour recording in a single screen, and then smoothly and rapidly zoom in to regions of interest, changing from scales coarser than ten minutes per pixel to scales as small as 10μs per pixel (60000000:1, about six temporal orders of magnitude). The goal of the *Audio Easter Egg Hunt* was to exhibit our *Timeliner* application, and thereby to teach elementary and high school students about the visualization of audio using spectrograms and other time-frequency representations.

A. Design

The Timeliner application is a multi-parameter zoomable timeline, in the spirit of "non-linear editing" video editing suites. Its source datafile is a single audio recording several hours long. It displays parameters derived from the source data in the form of stacked images, synchronized over a horizontal timeline (Fig. 4). Displayed parameters include the waveform, the spectrogram [4], a spectrogram transformed to reduce the visual salience of non-anomalous events (salience-maximizing features [12]), and a plot of the output log likelihoods from a bank of supervised classifiers [21], [20].

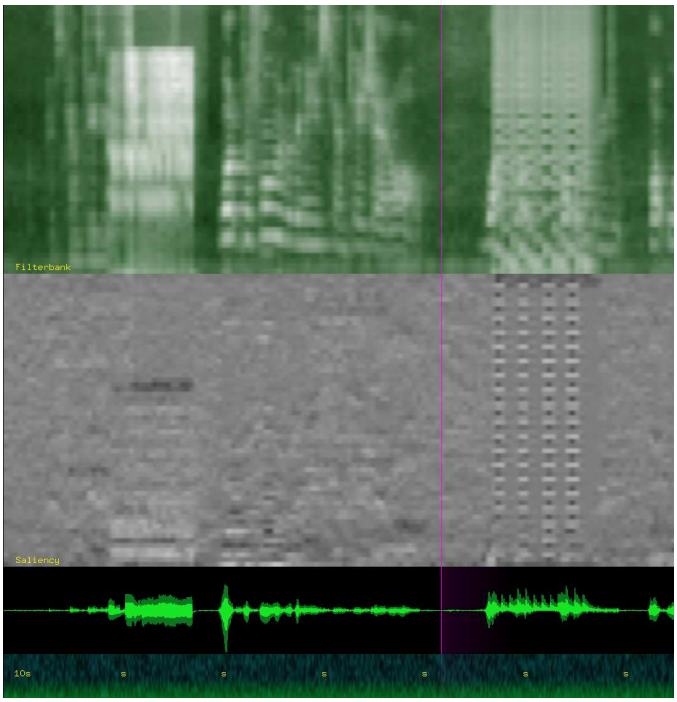

Fig. 4. The Timeliner multimedia analytic testbed, shown here, is designed to allow first responders, and other users of portable hardware, to rapidly find short anomalous audio segments buried in very long audio recordings.

Timeliner was prototyped in the scripting language Ruby, rendering both line-based waveforms and texture-map spectrograms in OpenGL. Ruby's interactivity let us quickly evaluate many user-interface ideas. The final "two-handed input" design uses the mouse and its scrollwheel for continuous panning and zooming at a steady 60 frames per second (fps), orders of magnitude faster than a traditional timeline editor's speed. While the right hand operates the mouse, the left hand operates keys without the user needing to look at the keyboard. For example, panning is done with keys "a" and "d," zooming with "w" and "s." (This "WASD" layout for mouse-keyboard real-time games gained dominance in the mid-1990s, and was familiar to most of the children who visited our exhibit.) The thumb on the spacebar starts and pauses audio playback.

Panning and zooming through data at such high speeds uncovered a latent flaw in traditional rendering of data as waveforms or heatmaps. Because one pixel or texel represents a considerable amount of data when zoomed out, Timeliner's 60 fps zoom rate demands computational shortcuts. Unfortunately, the traditional shortcut of undersampling causes flickering powerful enough to effectively obscure data until zooming ceases. Because this defeats the whole point of Timeliner, namely skimming the data and "formulating database queries" in real time, we eliminated this flickering with a multiscale cache. This data structure, given the time interval

spanned by a texel, yields that interval's minimum, mean, and maximum data values, in time logarithmic to the length of the original dataset. (The naive linear-time approach fails utterly for long recordings.) The cache works with both scalar data (waveforms) and vector data (spectrograms, saliency maps, or anything else in our HTK-based audio parameter format [19]). The final rendering stage converts a min-mean-max triplet into a hue-saturation-value (HSV) color, through a color transfer function chosen for that kind of data.

Timeliner was demonstrated at the 2011 open house in the form of an *Audio Easter Egg Hunt*. Visitors were each given 80 seconds to find anomalous sounds in 100 minutes of classical music (orchestral, without singing). "Anomalies" were obviously anomalous when one heard them, e.g., anomalies included a cow mooing, a cuckoo clock and two types of birdcalls, a motorcycle, a "Pac-Man" video game character, and several types of spaceships and ray guns. The challenge was to find the anomalies entirely from the visual display (which is very fast), rather than listening to the audio in real time (which is very slow).

Besides its use at the open house, the *Audio Easter Egg Hunt* has also been demonstrated to groups touring the University of Illinois, including K-12 field trips, science camps, visiting researchers, delegates at university-funded academic conferences and workshops, and international goodwill tour groups. The Integrated Systems Laboratory (ISL: home of co-authors Kaczmarski and Goudeseune) is funded in part by Beckman Foundation funds, for the purpose of providing virtual reality and user interface infrastructure to other groups in the Beckman Institute; ISL demonstrates virtual reality research to 2-10 tour groups weekly. Through the University's Office of Public Engagement, the ISL frequently recruits potential students currently in their junior and senior years of high school, focusing on underrepresented groups from major metropolitan areas. First-year undergraduates at the University of Illinois have been known to cite ISL's virtual reality demonstrations (first seen when they were in elementary or high school) as a reason for their interest in science.

B. Results

Fig. 5 shows the performance of some visitors to our *Audio Easter Egg Hunt* in the Beckman open house, March 11-12, 2011. Fifteen kinds of anomalies were uniformly distributed in the music, averaging one per 25 seconds of music, each of duration 1 to 4 seconds, so a brute-force listening strategy would expect to find 3 anomalies. Fig. 5 shows that the the median visitor was able to find nine anomalies in the eighty second game (an anomaly-discovery acceleration rate of 3.0 times real time), but that a few visitors achieved much higher acceleration rates.

Most visitors were 8 to 12 years old, and had no prior experience with spectrograms at all, let alone with our particular implementation thereof. One 6-year old visitor was notable because, despite her age, she was able to find anomalies at a rate of more than 7 times faster than real time. More than half the visitors were female. Throughout the two days, we posted scores on a scoreboard visible to visitors. This seemed to provoke new visitors to try to do well in the game, as it gave them a benchmark by which to compare their performance to others. Still, almost no visitors tried to play the game repeatedly in order to improve their score (although the same cannot be said for some of this paper's co-authors). This may be attributed to the dozens of other interesting exhibits less than a hundred paces away.

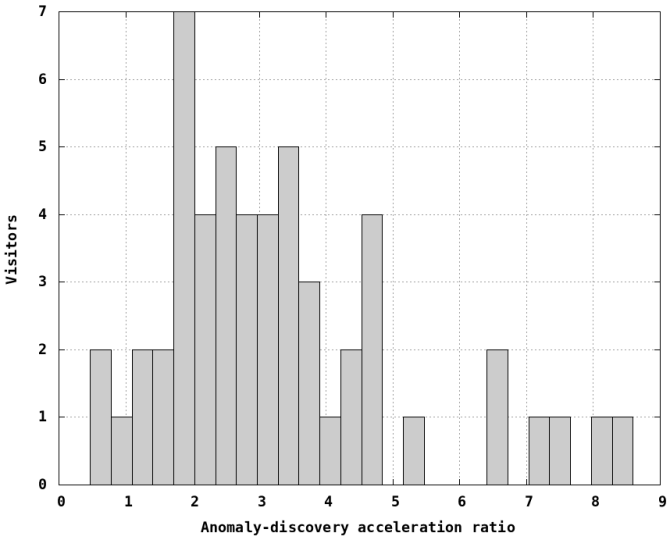

Fig. 5. Rate of discovering anomalies, as a multiple of real time, reported by visitors to the Audio Easter Egg Hunt at the Beckman Institute's 2011 Open House. Most visitors successfully exploited the spectrogram display, finding anomalies about three times faster than real-time listening. A few visitors found anomalies over seven times faster.

V. Discussion: Lessons Learned

The systems described here were all formulated as interactive games, and therefore all three were popular with children visiting the open house. There were, however, some differences among these systems.

A of the key difference is best summarized by Chickering and Gamson's second "Principle for Good Practice:" *Develop reciprocity and cooperation among students* [3]. All of the systems described in this article were one-player games. We discovered after the fact that a child's attention span is best captured, even in a one-player game, if his or her performance is measured against that of other children, even using the indirect social context provided by a high-score list (Fig. 5). In the *Speech Recognition Game*, for example, children who came to the exhibit in a group, and who competed against one another, were far more likely to be engaged in game-

playing than were children who approached one at a time. The difference between individual and cooperative play was even more starkly demonstrated by the dragon-hunting game in *Intonation Station*, perhaps because the dragon-hunting game is relatively easy to play as a group: groups of children can make jokes about the plausibility of hunting dragons in rural Illinois, about the absurdity of the troll's voice, and so on. Therefore, groups may inadvertently spend a great deal more time playing the game than a child alone. The implied lesson is that future open house demonstrations of speech, audio, language and learning information environments should be designed as multi-player games, rather than as single-player games.

VI. Acknowledgements

The *Speech Recognition Game* was supported by National Institutes of Health grant DC02717 and National Science Foundation grant 0534106. The *Intonation Station* was supported by National Science Foundation grant 0703624. The *Audio Easter Egg Hunt* was supported by National Science Foundation grant 0807329. All conclusions and findings are those of the authors and are not endorsed by their sponsors.

References

[1] Kim S. Abouchacra, Tomasz Letowski, and Timothy Mermagen. Detection and localization of magazine insertion clicks in various environmental noises. *Military Psychology*, 19(3):197–216, 2007.

[2] Ken Chen, Mark Hasegawa-Johnson, Aaron Cohen, Sarah Borys, Sung-Suk Kim, Jennifer Cole, and Jeung-Yoon Choi. Prosody dependent speech recognition on radio news. *IEEE Trans. Speech and Audio Processing*, 14(1):232–245, Jan 2006.

[3] Arthur W. Chickering and Zelda F. Gamson. Seven principles for good practice in undergraduate education. *Higher Education Bulletin*, pages 3–7, March 1987.

[4] Dave Cohen, Camille Goudeseune, and Mark Hasegawa-Johnson. Efficient simultaneous multi-scale computation of FFTs. Technical Report FODAVA-09-01, NSF/DHS FODAVA-Lead: Foundations of Data and Visual Analytics, 2009.

[5] Maxine Eskenazi. Detection of foreign speakers' pronunciation errors for second language training - preliminary results. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 1996.

[6] Maxine Eskenazi, Jack Mostow, , and David Graff. The CMU kids corpus. Technical Report LDC97S63, Linguistic Data Consortium, Philadelphia, 1997.

[7] Horacio Franco, Leonardo Neumeyer, María Ramos, and Harry Bratt. Automatic detection of phone-level mispronunciation for language learning. In *Proceedings of the European Speech Technology Conference (EUROSPEECH)*, 1999.

[8] W. Holmes, P. Andrews, B. Seegar, and D. Radcliff. Implementation of a voice activated computer system for teaching and communication. In *Proc. TADSEM '85: Australasian Seminar on Devices for Expressive Communication and Environmental Control*, pages 43–48. Ryde Australia: Technical Aid to the Disabled, 1985.

[9] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas Huang, Kenneth Watkin, and Simone Frame. Dysarthric speech database for universal access research. In *Proc. Interspeech*, 2008.

[10] Y. Kim, H. Franco, and L. Neumeyer. Automatic pronunciation scoring of specific phone segments for language instruction. In *Proc. Eurospeech*, pages 645–8, Rhodes, Greece, 1997.

[11] Henry Kucera and W. Nelson Francis. *Computational Analysis of Present-Day American English*. Brown University, Providence, RI, 1967.

[12] Kai-Hsiang Lin, Xiaodan Zhuang, Sarah King, Camille Goudeseune, Mark Hasegawa-Johnson, and Thomas S. Huang. Improving beyond-real-time human acoustic event detection by saliency-optimized audio visualization. In Preparation.

[13] K. Nagano and K. Ozawa. English speech training using voice conversion. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 1169–1172, 1990.

[14] University of Illinois. Beckman institute open house 2011. http://www.beckman.illinois.edu/events/boh2011.aspx. Accessed: 2011/06/15.

[15] University of Illinois. History of engineering open house. http://eoh.ec.uiuc.edu/history.php. Accessed: 2011/06/15.

[16] Harsh Vardhan Sharma and Mark Hasegawa-Johnson. State-transition interpolation and MAP adaptation for HMM-based dysarthric speech recognition. In *HLT/NAACL Workshop on Speech and Language Processing for Assistive Technology (SLPAT)*, Denver, CO, 2010.

[17] Harsh Vardhan Sharma, Mark Hasegawa-Johnson, Jon Gunderson, and Adrienne Perlman. Universal access: Speech recognition for talkers with spastic dysarthria. In *Proc. Interspeech*, volume 42862, pages 1–4, Brighton, 2009.

[18] Silke Witt and Steve Young. Language learning based on non-native speech recognition. In *Proceedings of the European Speech Technology Conference (EUROSPEECH)*, 1997.

[19] Steve Young, Gunnar Evermann, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. *The HTK Book*. Cambridge University Engineering Department, Cambridge, UK, 2002.

[20] Xiaodan Zhuang, Xi Zhou, Mark A. Hasegawa-Johnson, and Thomas S. Huang. Real-world acoustic event detection. *Pattern Recognition Letters*, 31(2):1543–1551, Sep 2010.

[21] Xiaodan Zhuang, Xi Zhou, Thomas S. Huang, and Mark Hasegawa-Johnson. Feature analysis and selection for acoustic event detection. In *Proc. ICASSP*, 2008.