Phonetic Shapes:
An Interactive, Sonic Guest Book

Mary Pietrowicz
University of Illinois at Urbana-Champaign, Department of Computer Science
Urbana, IL 61801 USA
mpietro2@illinois.edu

Karrie Karahalios
University of Illinois at Urbana-Champaign, Department of Computer Science
Urbana, IL 61801 USA
kkarahal@illinois.edu

CHI 2012, May 5–10, 2012, Austin, TX, USA.
ACM 978-1-4503-1016-1/12/05.

Abstract
Sound exists only in the moment, and cannot be referenced or searched, except in the mutable memories of people. Because of this, it has been an overlooked modality and social information channel, particularly where nonverbal communication cues and identity are concerned. Yet, it encapsulates a person's identity as effectively as a fingerprint or signature. Sound is even more useful when combined with other modalities, like the visual and gestural. In order to use the modality of sound effectively, however, we need tools that simultaneously analyze, persist, and present the important information in sound. What if you could capture identity and meaning in sound, and give it additional affordances, that go beyond those of written communication? In this paper, we explore the voice as identity and as the carrier of nonverbal information in the context of a sonic guest book.

Author Keywords
Visualization; social identity; audio; signal processing; machine learning; phonetics; guest book; persistence

ACM Classification Keywords
H.5.5 [Information Interfaces and Presentation]: Sound and Music Computing – Signal analysis; H.5.m [Information Interfaces and Presentation]: Misc.;

Introduction

Sound exists only in the moment. It is here, and gone. It has all of the properties of an idea in a preliterate society described by Walter Ong [9]. Unless it is recorded, or its qualitative features saved, it exists only in the memories of those who heard it, and is easily changed or forgotten. In addition, speech sound conveys much beyond the content of the words. Affect, emphasis, humor, and sarcasm are just a few of the nonverbal components in speech. Perhaps the most overlooked quality of speech, hidden in plain sight, is identity. No two people speak in exactly the same way. A person's speech is his sonic fingerprint.

We seek to explore making the properties and identities of spoken sound persistent, in ways which are easy to interpret, explore, and replay. An ideal application for exploring these properties in a social context is the guest book. Guest books are all about identity, so how would we present our identities and personal messages sonically, for posterity? Even a simple sound recording, though limited, has meaning. How I wish I could hear all the voices of those who signed my wedding guest book, because some of those voices are silent now.

Related Works

Significant work in speech visualization, speech and phoneme recognition, social identity, and guest book applications has been done. The Visualizing Remote Voice Conversations project [11] visualized conversation and included content analysis, simple sonic feature analysis, and basic persistence. The three visualizations in [11] recognized words in the conversation and displayed their frequency of use, displayed pitch and volume of sound for each user, and displayed turn-taking in the conversation with the corresponding volume of the speakers at each turn. This project had just begun to explore the qualitative features of speech that convey meaning and establish speaker identity. Visualizations in [7] used kinetic typography to show affect. Other typical speech visualizations include waveforms or spectrograms. While not incorrect, these tools show primarily time and frequency data, and do a poor job of conveying inflection, emphasis, rate, emotion, gender, or any other qualitative features that give unique character to both speaker and the message.

Projects from the speech therapy community have also explored voice visualization. The Phonetics Flash Animation Project [13] displays the mechanisms that produce each of the American English phonemes by showing the movement of the lips, teeth, tongue, jaw, pharynx, epiglottis, and vocal folds. These visuals are accompanied by written instructions for producing each phoneme. The project also provides similar instructions for producing each kind of phoneme articulation and voicing. It does a very good job connecting phonemes with the mechanisms for producing them, and the visuals are easy to understand. This project, however, is focused on providing instruction for producing language, and does not analyze or visualize live or recorded sound. The Spoken Impact and VocSyl Projects [3,4], however, take the opposite approach, analyze live or recorded sound, and use a variety of visualizations to help children with autism vocalize. VocSyl and ReadN'Karaoke visualized prosodic cues of pitch, amplitude, and duration in speech[4,12].

Much research has been done on speech and phoneme recognition, too many references to list here. Notably, the CMU Sphinx project [1] provides an open source

toolkit for speech recognition, and it can be used for phoneme recognition. The Sphinx implementation, however, was designed for "whole" speech recognition, which relies on the contextual constraints of the language. Pure phonemes have fewer contextual cues, so Sphinx used in this mode may be slow and have high error rates, per warning by the developers. The Sphinx research (like most other phoneme recognition projects) is also focused on interactive speech recognition, not on speech visualization.

The music and interactive art communities have explored sound visualization, and they have often created suites of animations, which are driven by one or more features in the sound; these are common in music players. The relationship between sound and visuals in them is often unclear; and therefore, the meaning of the visualizations is also unclear. Furthermore, these tools are generic, and have not been designed or used as an integrated component of a multimedia work. The result is a visualization disconnected from the sound, instead of an integrated work where sound and visuals work together and reinforce each other. In general, this disconnection is undesirable. Myron Kruger observed and noted disconnection between sound and other modalities as a limitation in an early interactive installation (GLOWFLOW) [6]. Golan Levin et al. have explored voice visualization in several projects, particularly Messa di Voce [8], and have noted the positive effects of connecting visualization with sound at the place where the sound originates.

The MusiVerse project [16] and its derivatives began explorations of sound visualization in the context of 3D worlds and games. Interactivity in the MusiVerse projects ranged from simple mouse and keyboard to live instrumental steering of the visual world.

Social identity in an audio context has not been sufficiently addressed. Donath described identity deception in Usenet communities [5]. Toma [17] described how users present themselves in online dating communities, and the deception included in the profiles. Commercial products attempt lie detection with audio [15]. A person's voice and manner of speaking is a type of "fingerprint," and represents a potentially important, and untapped, component of virtual identity.

Guest book applications have changed very little over time. The basic function of a guest book is to capture the identity of and basic information about visitors at a physical or virtual event. Secondary functions include capturing personalized messages from visitors, capturing something of the personal essence of each visitor, and preserving memories about the event. Existing guest books as in [2] are either paper or electronic, but both have similar features. People sign, write messages, and paste pictures into either form. Some electronic forms can capture data like timestamps automatically. Neither format has significant offerings in sound. iTunes offers a video guest book that uses an iPad kiosk to record video messages, but this application doesn't do any analysis or visualization of the sound beyond the recording. [18]

Current Work

Our current work explores the development of sonic guest book applications. The project is still under development, but as of this writing, we have produced prototype visualizations of the voice, and are in the process of evaluating the visualization techniques. The

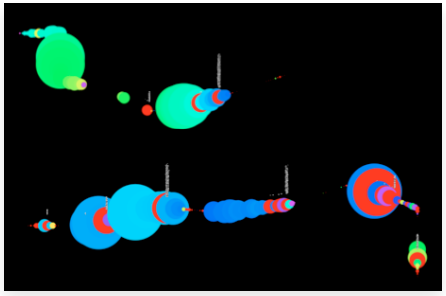

Figure 1: A first visualization prototype of the sound of the names "Mary Pietrowicz," top (210 Hz), and "Stephen Roman Pietrowicz," bottom(125 Hz), as spoken by each person. Color maps to frequency.

work breaks down into four areas: 1) audio feature exploration, 2) machine learning and phoneme recognition, 3) sound visualization, and 4) evaluation.

The first step toward working with speech and sound is reducing the input data to a form that emphasizes the most important information and excludes irrelevant details. We started with Mel Frequency Cepstral Coefficients (MFCCs), formant analysis, the amplitude contour, and relationships in the spectrum [10]. From this reduction, we inferred the fundamental frequency (or spectral centroid), pitch contour, amplitude contour, and noise profile. This data formed our basic feature set. We started with these features because MFCCs and formants are efficient phoneme recognition tools, the mechanics of MFCC analysis reduces dimensionality efficiently, and because the other components of our feature set could be used to visualize vocal inflection, dynamic range, and timbre changes directly. We also believed that these features would be useful in deriving secondary attributes and nonverbal cues in speech, such as rate and affect. An initial prototype displayed pitch contour, relative amplitude, and noise bursts in vocal signatures. The horizontal axis maps to the progression of time, and the diameter of the shape shows relative sound amplitude. The vertical axis represents the pitch contour of the sound. The amount of change in fundamental frequency steers the contour up or down from frame to frame. We also mapped color to pitch, and mapped lower voices to cool blues (typical male voice), higher voices to greens (typical female voice), and noise, which tended to be much higher frequency sounds, to bright, warm oranges and reds. Points in the sound that corresponded to high levels of noise were embellished with what appeared to be particle traces at that point in time (see Figure 1).

We wanted to reflect user identity in new ways, so we trained the system to recognize American English phonemes. Phonemes are the fundamental building blocks of sound in a language, and we selected the monophthongs (single-sound vowels), diphthongs (blended-sound vowels), and consonant classes shown in figures 2a-c.

	Front long	Front short	Central long	Central short	Back long	Back Short
Close	/i:/ key	/I/ kit			/u:/ cool	/ʊ/ could
Mid		/ɛ/ ken	/3/ cur	/^/ cut	/ɔ/ caught	
Open		/oe/ cat				/D/ cot

Figure 2a: Monophthong Vowels, formed with the tongue raising and lowering, with the sound focused from front to back in the mouth.

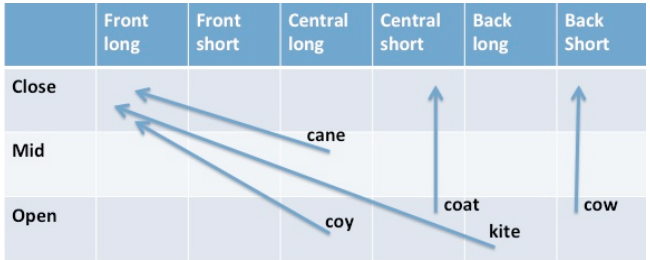

Figure 2b: Diphthong Vowels, formed by moving between two vowels and two positions in the mouth.

	Short Sustain	Long Sustain
High Noise	Africates /ch/ /nch/ /j/ Stops /p/ /b/ /t/ /d/ /k/ /g/	Fricatives /h/ /f/ /v/ /s/ /z/ /sh/ /zh/
Medium Noise	Glides /w/ /j/ /kw/ /ks/	Nasals /m/ /n/ /ng/
Low Noise		Liquids /l/ /r/

Figure 2c: Consonant Classes, by relative noise content and the degree to which they are sustained. Examples per class: /ch/ chip, /sh/ shack, /p/ pan, /j/ yes, /m/ mat, /l/ lamp

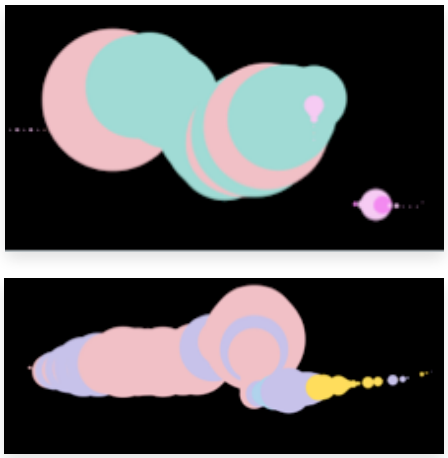

Figure 3: Visualizations of the word "kit", top, and of the word "core", bottom. Color indicates phoneme or phoneme class.

We used simple Gaussian classifiers and Hidden Markov Models (HMMs) [14], and initial results produced an average phoneme recognition accuracy of 70-75%, without any additional filtering or modeling. Closer examination of the results revealed that diphthongs were causing much of the confusion, but not in the way we expected. The diphthongs themselves were classified relatively well, but the majority of monophthongs were being misclassified as diphthongs when they were misclassified. By converting the architecture into a network of classifiers that make simple decisions, where each classifier has a high accuracy rate, we can remove much of the confusion, retain efficiency, and improve the classification rates for many of the phonemes to theoretically greater than 90%. With this architecture, three of the vowels remain difficult to classify, and include the vowel sounds in the words "cat," "cut," and "ken". We plan to explore further architectural improvements to our phoneme classifier in the future and expect that alternative time series techniques, clustering, shifting to higher levels of dimensionality, and further exploration of the feature space may be helpful.

A revised visualization uses color to represent the different phonemes and phoneme classes. In this second visualization, the x axis still represents the progression of time, the diameter of the visualization represents amplitude, and the bending along the vertical y axis represents relative pitch contour. The colors, however, have been changed to represent the phoneme class. Cool shades represent the vowels; the monophthong vowels formed in the mouth in close position are blue, mid position, aqua, and open position, green. The degree of saturation in the color shows where in the mouth the sound is made (front to back). Sounds with the lowest saturation are formed in the front of the mouth, and sounds with the highest saturation, in the back. Diphthongs have a special color – violet – and its saturation corresponds to the amount of change required to produce the sound. A maximum transition is maximally saturated. Consonants are represented by splashes of warm, bright colors, including a yellow-orange, a claret red, and a fuchsia. Consonants with low noise content (like "r") are yellow, and those with high noise content (like "sh") are fuchsia. The higher the nose level, the brighter and warmer the splash of color. The duration of the consonant sound ("p" vs "sh") controls the saturation. Longer durations correspond to saturated colors. We removed the particle traces from this prototype because we thought they were cliché, and we wanted to challenge ourselves to develop new approaches.

Future Work

Remaining steps to completing this project include implementation of remaining guestbook features, experimenting with alternate visualizations, improving phoneme recognition, experimenting with different platforms and venues, and user evaluation. To complete the guest book features, we will create a suite of guest book layouts, create a recording interface for the signers, create a browsing and playback feature for the users, and experiment with multiple views and alternate mappings. We plan to improve the phoneme recognition with the techniques we mentioned in the previous section, and others from the speech and hearing community. Guest books will be optionally online, searchable, and browsable. We hope to explore interaction via different media, including phone, tablet, large displays for museums and other public spaces,

remote interfaces to public spaces, waterscapes, and more. Finally, user studies are required to understand the efficacy of the different visualizations and interaction mechanisms we used.

We are already exploring the use of the analysis and visualization techniques in the speech therapy community for children who are speech delayed or who are on the autism spectrum. Beyond this, we plan to use the techniques developed for the guest book as a springboard for developing other social applications, particularly platforms for audio histories. Personal journals, musical taste profiles and visualizations, crowd journals, and contemporary histories all could make use of the enabling technologies we developed for the guest book. We are particularly interested in combining audio with other modalities such as video, motion, or galvanic skin sensors.

Acknowledgements

Thanks to Curtis Wang, who collaborated with us on the HMM model development for this project.

References

[1] The CMU Sphinx Project. http://cmusphinx.sourceforge.net/wiki/phonemerecognition

[2] Guestbook Store Products. http://www.guestbookstore.com/

[3] Hailpern, J., Karahalios, K., and Halle, J. Creating a Spoken Impact: encouraging vocalization through audio visual feedback. Proc CHI 2009.

[4] Hailpern, J. et al. VocSyl: Visualizing Syllable Production for Children with ASD and Speech Delays. Proc ASSETS '10 2010.

[5] Kollock, P., Donath, J. et al. Communities in Cyberspace. Routledge 1999.

[6] Kruger, M.K. Artificial Reality. Addison-Wesley 1983

[7] Lee, Johnny, et al. The Kinetic Typography Engine: An Extensible System for Animating Expressive Text. Proc UIST 2002.

[8] Levin, G. and Lieberman, Z. In-Situ Speech Visualization in Real-Time Interactive Installation and Performance. Proc 3rd International Symposium on Non-Photorealistic Animation and Rendering 2004.

[9] Ong, W. Orality vs Literacy. Routledge, New York, 2002.

[10] O'Shaughnessy, D. Speech communications human and machine. IEEE Press 2000.

[11] Mathur, P. and K. Karahalios, K. Visualizing Remote Voice Conversations. Proc CHI EA 2009, ACM Press (2009), 4675.

[12] Patel, Rupal, and Furr, William. ReadN'Karaoke: Visualizing Prosody in Children's Books for Expressive Oral Reading. Proc CHI 2011.

[13] The Phonetics Flash Animation Project at the University of Iowa. http://www.uiowa.edu/~acadtech/phonetics/

[14] Rabiner, L.R. A tutorial on hidden markov models and selected applications in speech recognition. Proc IEEE, Vol. 77, No. 2, Feb. 1989.

[15] The Sharper Image. http://www.sharperimage.com/

[16] Smith, B. and Garnett, G. MusiVerse. Proc Intl Computer Music Conference (ICMC) 2007.

[17] Toma, et al. Separating Fact From Fiction: An Examination of Deceptive Self-Presentation in Online Dating Profiles. Pers Soc Psychol Bull 2008.

[18] Video Guest Book on iTunes. http://itunes.apple.com/us/app/video-guestbook-for-your-wedding/id425394798?mt=8