Visualizing Vocal Expression

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

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

CHI 2014, Apr 26 - May 01 2014, Toronto, ON, Canada
ACM 978-1-4503-2474-8/14/04.
http://dx.doi.org/10.1145/2559206.2581331

Abstract
Sound, especially speech, is ephemeral. It is a high-speed, ordered, multichannel stream that plays for a time, and leaves shadows of its presence in our memories. When we communicate, we exchange semantic, expressive, and relational messages. Most of our communicative power lies outside semantics, yet these expressive and relational exchanges are underexplored. We and others have experimented with visual representations of speech, yet little is known about the interpretability, usability, and efficacy of the visualizations; here we focus on interpretability. We provide a system for expressive vocal analysis, new voice visualizations which map vocal parameters to different designs, and a study focusing on the interpretability of the resulting voice visualizations.

Author Keywords
Visualization; Speech; Voice; Communication

ACM Classification Keywords
H.5.2 **[Information Interfaces and presentation (e.g., HCI)]**: User Interfaces --- voice I/O

Introduction
Human speech communication is ephemeral. After the sound fades away, it lives only in fallible, limited, mutable memory. Literacy changed the human

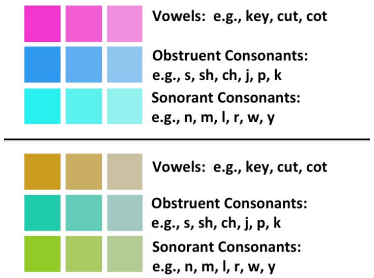

Figure 2. *Sound Quality Color Map.* This figure shows the two color maps that we compared. Brights are on the top, and muted colors on the bottom. Higher color saturation corresponds to voicing and longer duration on obstruents, higher resonance and longer duration on consonants, and vowels formed toward the back of the vocal tract.

experience, gave language a life beyond the spoken moment, and made speech persistent. People collected, absorbed, expanded, and recombined ideas, without the presence of the original speaker, in the time it took to read the words, regardless of the original speaking or writing time. Abstraction, external reference to ideas, and more complex thought processes became possible. It even changed the nature of language itself. Before literacy, speakers often grouped words and ideas into clustered phrases, like "beautiful princess," and thought in terms of these larger linguistic building blocks. Human memory is limited, and a single clustered phrase like "beautiful princess" is easier to remember than the two individual ideas "beautiful" and "princess." With literacy, these cliché clusters gradually faded out of discourse, and the "beautiful princess" could stand alone as a princess, beautiful or not [9].

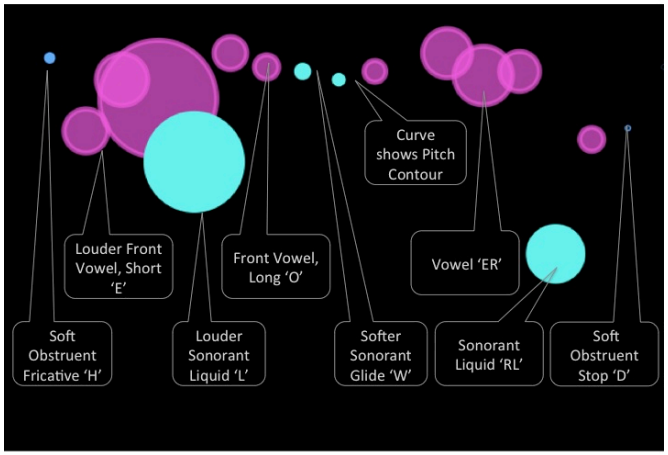

Figure 1. *Annotated Bubbles Visualization*. Visualization of "Hello World," where the bubble size at each point in time shows amplitude, the relative vertical position shows pitch, and the color hue and saturation shows phonetic quality.

Spoken language, however, includes semantics, vocal expression, nonvocal body language, and relational interaction. Literacy captures the semantics only, and in doing so, misses much of the message [7]. Yet, the impact of literacy was large. What could the impact of capturing, analyzing, visualizing, and making persistent non-semantic communication be, when these non-semantic channels carry most of the information? We believe that capturing vocal expression could impact language again, and change us socially. It could also provide enabling technology for speech therapy and new creative works. To begin exploring the possibilities, we extended our previous vocal analysis work [11], focused on vocal expression, and visualized it in multiple ways (see Figures 1-3). Finally, we evaluated user interpretation of the visualizations, user preference for visualization of their own voices, and preferences for variations of specific visual elements.

Representative Related Work Samples

Relevant work in voice visualization ranges over visualizing vocal qualities [10,2,5], music [13], generic sound [14], and sound collections [8]. Music and sound applies because timbre, dynamics, pitch, and articulation in these domains have similar functions in speech. Generic and ambient sound visualizations display timbre, too. Collection visualization is important because of the visualization of clustered groups of sound. Works from creative arts and kinetic typography [2,5,6] are engaging as are works for children and those with playful intent[1,4,10]. Scientific voice vis [3,12,14] (e.g. waveforms/spectrums) has a more traditional look and feel, along with realistic speech visualizations [12]. We recognize the importance of live and text conversation visualization, but we do not yet focus on it.

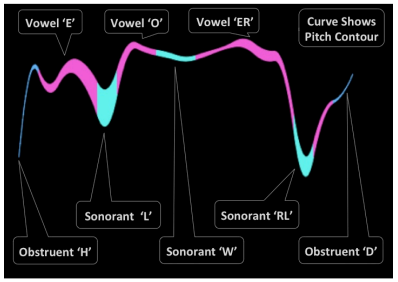

Figure 3. *Annotated Ribbons Visualization.* Visualization of "Hello World," where the ribbon thickness shows amplitude, the vertical position tracks pitch, and the color hue and saturation shows phonetic quality.

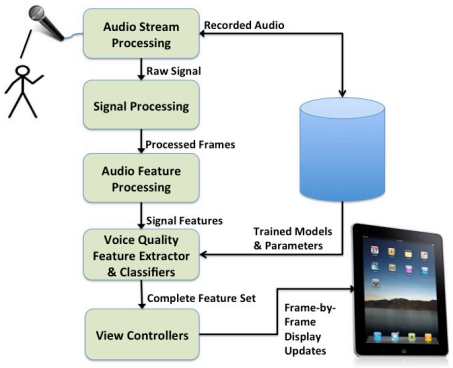

Figure 4. *System Overview.* shows the flow of real-time processing and visualization of the voice.

Visualizations

To begin our explorations, we selected a small set of voice characteristics related to human vocal expression. We wanted users to understand the visualizations, see the expressivity, think about their voices, and identify with the visualizations. We selected pitch, amplitude, phonetic quality, and noisiness. Pitch changes are prosodic markers; usually pitch rises for questions and falls for statements in English. Furthermore, people exaggerate pitch change to create emphasis on what they are saying and thinking. Imagine, for example, the difference between an astounded person saying, "Whaaaat?!" and a distracted person saying, "What." Musicians talk about pitch in terms of up and down, and we map this idea directly to the up and down position on the screen. Similarly, the size of the signal is the amplitude, and we map this idea to size on the screen.

Phonetic quality is important, not for semantics, but for understanding how a word is spoken. With just a few phonetic classes, such as vowels, sonorants, and obstruents, we can easily see if someone is stretching out vowel sounds, or emphasizing a starting or ending consonant. Imagine a person thinking "absolutely not" and saying "nnooooooo," vs. a neutral, "no." Phonetic class information is also very useful in detecting syllables. Finally, noisiness reveals sighs, deep breaths, overemphasized consonants, whispers, breathiness, and aspirations. Overemphasized consonants and strong aspirations, for example, can reveal that someone is literally, spitting mad. We map these qualitative vocal elements (which color the voice) onto the screen as color, transparency, and outline.

We present two visualization styles which we call "Bubbles" and "Ribbons," selected for their opposing discrete and continuous qualities. Speech is both qualitatively discrete (with distinct phonemes, words, syllables, phrases, points of emphasis) and continuous (sounds flow one to another behind the flow of human breath), and we and show the test phrase, "Hello, world" visualized in each of these styles as an example for comparison. The "Bubbles" visualization shows the sonic quality of each frame we sampled via small circles, with time progressing to the right. The size of each bubble shows the amplitude, and the vertical position in the window shows the pitch. Higher pitches appear near the top of the screen, and lower pitches near the bottom. The sonic quality corresponds to the color (see Fig. 1,2). This visualization's strength is its clear differentiation in amplitude and the "particle" nature of the noisy consonants. This style is especially fitting for voiced, sustained noisy consonants, like the "zh" sound. The speaker is blowing air and moisture to say these sounds, which often look like particle clouds. The "Ribbons" visualization shows the voice as a continuous, sinuous, ribbon of sound. We used similar mappings where thickness shows amplitude, time moves left to right, relative pitch is higher toward the top of the screen and lower to the bottom, and color shows phonetic quality. This visualization's strength is its flowing, continuous nature, like the force of the human breath that drives the voice, and its likeness to writing, where the energy and quality of the voice controls the pen. See Figure 3.

Method

We prepared a 1.5-hour exploratory study to measure user preference and interpretation of our visualizations and selected 13 participants, both men and women. We began with a short demonstration and discussion of the system, to ensure that each participant understood

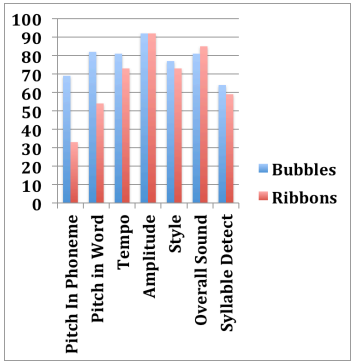

Figure 5. *Percent Accuracy in Visualization Presentation and Interpretation.* Pitch interpretation improves when heard in the context of a real word, compared to inflection on a single phoneme. The Bubbles visualization style shows higher accuracy in pitch and tempo interpretation, compared to Ribbons. In all other attributes, Bubbles and Ribbons provide similar accuracy.

the application, visualizations, and terminology we used in the study. At the end of this discussion, we gave the user a visualization key for use during the study.

Next, we ran a survey to test whether a person could interpret the different visualization styles. We asked the users to interpret our presentation of pitch inflections, time, speaking tempo, and sound amplitude by asking them to play sounds and select the correct visualization from a list of possibilities for each one. We also asked the user to distinguish a test phrase, such as "Hello, world," spoken in four ways: question, statement, monotone, and emphatic speech. Finally, we checked whether the user could interpret our color-sound mappings for sonorants, obstruents, and vowels. All participants evaluated both the Ribbons and Bubbles styles. To control for bias, we randomized the order of the questions given to the users, and the order in which the possible answers were presented. Then, we asked users about their visualization feature preference by showing them (with the corresponding audio) similar visualizations with variations in color and size. Finally, we asked the users to record their own voices by saying their names or a representative phrase. They played and viewed their recordings, and answered a questionnaire about their visualization preferences, and the degree to which they thought the visualizations represented themselves. This process emphasized open-ended questions designed to encourage critical thought, capture impressions, and solicit opinion. To close, we asked the users whether and how they would use their voice visualizations in various social media.

System Design

The system is a pipeline of sound processing and visualization that begins with a layer of essential signal processing and feature extraction. We downsampled and filtered the signal to improve performance, and reduced the signal to a meaningful representation by extracting a set of features historically used for speech. These features include the Mel Frequency Cepstral Coefficients (MFCCs), other spectral values, pitch, amplitude, and noise content. See Figure 4.

We used both a simple Gaussian classifier and Hidden Markov Model in our implementation to detect phoneme classes and voicing quality and trained the system with samples of each phoneme, provided by adult men and women with Midwestern accents, and one West-coast speaker. Our training set did not include speakers from other regional dialects in the US, other English-speaking countries, or non-native English speakers. The system recognized phonemes by class, including vowels ("singable" sounds), sonorant consonants ("singable" but noisy sounds), and obstruent consonants (noisy, "unsingable" sounds). These three sound categories were sufficient for visualizing the most important parts of phonetic quality relating to our research questions.

Results and Conclusions

The first part of our study tested whether users could understand our representations of pitch inflection, time/tempo, amplitude variation, speaking style, syllable articulation, and overall sound quality. Figure 5 summarizes the results. With the exception of pitch, the interpretive accuracy between the bubbles and ribbons visualizations was not statistically significant. Most of the results yielded around 80% accuracy, with syllable interpretation, and pitch interpretation (for ribbons) being notably lower. Why did we see differences in pitch interpretation? Closer examination of the data revealed a "closure artifact" common in

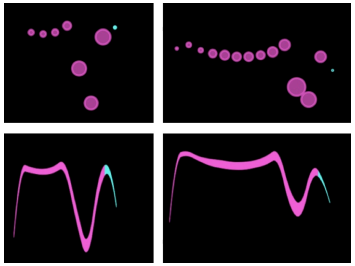

Figure 6. *Closure Artifact in Upward Inflection.* Left: Phoneme "oh" inflected up. Right: The word "hello" inflected up. Closure artifacts confuse interpretation of short utterances.

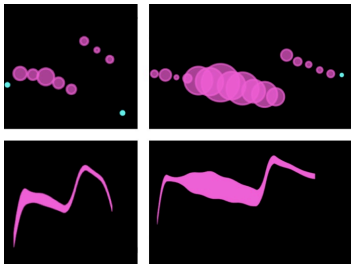

Figure 7. *Closure Artifact in Downward Inflection.* Left: Phoneme "oh" inflected down. Right: The word "hello" inflected down. Closure artifacts confuse interpretation of short utterances.

inflections at the end of utterances – a short "wobble" in pitch (see Figure 6). These wobbles are short (milliseconds), and do not confuse us aurally, but do confuse our visual interpretation when the utterance is also short. Figure 7 shows another artifact common in downward inflections at the end of utterances.

We also found that when users had to use multiple features to interpret a visualization, as in a style or overall voice quality judgment, they performed equally well using Bubbles and Ribbons. The "style" judgment asked users to match an utterance of "Hello, World" with its visualization, depending on speaking style (question, statement, monotone, or emphasis). The overall "voice quality" test asked users to listen to a word which was either strongly sonorant (singable sounds only) or strongly obstruent (many noisy sounds), and pick the correct visualization. Users performed these tests surprisingly well, even when phonetic recognition errors were present.

In the second part of the study, we asked users about color and size preference over the two visualization styles, and introduced a new color palette of muted tones for comparison against the brights. We played sonorant and consonant words, presented the variations to the user, and asked for their preferences. Most users chose colors for readability, or to match the mood and timbre of the vocal sound. A few users thought of matching color with the meaning of the utterance. In general, most users preferred brights for readability because color contrast was greater. Muted colors tended to "blend" over the continuous line in a Ribbon. Also, bright colors were easier to see in thin lines and small particles. Many users liked the use of transparency in Bubbles, and commented that it let them see layers of sound (functional and aesthetic). About half of the participants attempted to map color to the timbre of speech. They liked bight colors for loud, emphatic, strong sounds and muted colors for quiet, monotone, and mellow sounding speech. We did not map color in that way in our visualizations, nor did we ask our participants to comment on this, but we may want to experiment with this in the future, given the feedback. Most users preferred larger bubbles and thicker lines, again, because they were easy to see. Furthermore, a thicker line supported visualization of a larger dynamic range in amplitude. However, some study participants and professional designers who were external to the study thought that the thinner objects were more aesthetically pleasing. Some of the users, however, thought about mapping the size onto the dominant vocal quality they heard. Almost everyone wanted to make the visualizations of the monotone utterances smaller, whether they were quiet or not. We saw this in our examination of color, too – when people perceived a dominant quality in a sound, they wanted to see the quality reflected back into the visualization along multiple channels. For example, if a user heard a monotone utterance, and this dominated their perception, the user would expect to see this quality reflected in multiple ways on the screen. Monotones, say our study participants, should have muted colors, small size, and minimal inflections.

The final (free response) section of our study asked users about their overall impressions of each style, preferences between styles, and impressions of the visualizations of their own speech. Many of the participants liked the elegance of the Ribbons better for viewing others' speech, but then preferred Bubbles for visualizing their own speech. Some participants thought

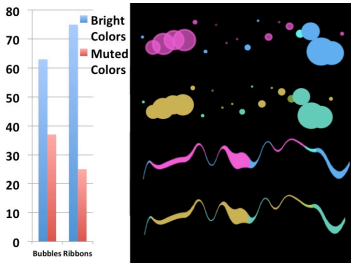

Figure 8. *Color Preference.* The word "cacophonous" shown in the two visualization styles and color palettes. The bar chart shows the strong preference for bright colors in both visualization styles, for this vocal expression.

the disconnection in Bubbles hard to follow, but others thought that Bubbles tracked their voices better in real time. Users disagreed on use of their visualizations in social media. Some liked the idea of visual signatures, but others thought that their visualizations would reveal things about their voices that they would not like to share publically, such a non-native speaker accent.

Based on designer and study feedback, our next steps will be to experiment with blended elements that leverage the strengths of both styles. The particles in the Bubbles visualization show noise quality and sound breaks better, while the Ribbons style shows pitch variation and sound connection better. We will also experiment with a wide range of new sonic and affective qualities in the voice. When a dominant quality is apparent, we will experiment by visualizing it using multiple, simultaneous techniques. It would be revealing to study how different visualizations of the same utterance affect perception. We also think that exploring the connection between identity and vocal expression – how people would choose to represent themselves and their ideas – would be revealing. For example, if we gave users the ability to customize their vocal visualizations, what would they choose, and why? Finally, given the ability to analyze and visualize real-time speech, what kinds of applications in medicine, speech therapy, and the creative arts could be enabled?

References

[1] Cavazza, M. et al. Emotional input for character-based interactive storytelling. AAMAS '09.

[2] Cho, P. Takeluma: An Explortion of Sound, Meaning, and Writing. MFA Thesis, UCLA Department of Design, Media Arts. June 2005.

[3] Fell, H.J., & MacAuslan, J. Vocalization Analysis Tools. MAVEBA '05.

[4] Hailpern, J., et al. VocSyl: Visualizing Syllable Production for Children with ASD and speech Delays, Extended Abstracts of ASSETS 2010.

[5] 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.

[6] Levin, G. and Lieberman, Z., with Ars Electronica Futurelab. Re:MARK. 2002. http://www.flong.com/projects/remark

[7] Mahrabian, A. Silent Messages. Wadsworth, 1971.

[8] Morchen, F. et al. MusicMiner: Visualizing timbre distances of music as topgraphical maps. ISMIR '09.

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

[10] Patel, R., and Furr, W. ReadN'Karaoke: visualizing prosody in children's books for expressive oral reading. CHI'11.

[11] Pietrowicz, M. and Karahalios, K. Phonetic Shapes: An Interactive Sonic Guest Book. CHI EA '12.

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

[13] Seidenburg, K. An exploration of Real-time Visualizations of Musical Timbre. Proc. 3rd intl. workshop on learning semantics of audio signals, 2009.

[14] Viegas, F.B. et al. Artifacts of the presence era: Using information visualization to create an evocative souvenir. Proc. IEEE Symposium on Information Visualization, p 105-111.

[15] Watanabe, A. et al. Speech visualization by integrating features for the hearing impaired. IEEE Transactions on Speech and Audio Processing, 2000.