SONIC SHAPES: VISUALIZING VOCAL EXPRESSION

Mary Pietrowicz
University of Illinois,
Dept. of Computer Science,
201 N. Goodwin Ave., Urbana, IL
mpietro2@illinois.edu

Karrie Karahalios
University of Illinois,
Dept. of Computer Science,
201 N. Goodwin Ave., Urbana, IL
kkarahal@illinois.edu

ABSTRACT

Sound has been an overlooked modality in visualization. Why? Because it is ephemeral. We experience it as it happens, often in community with others. Then, the sound is gone. Furthermore, sound in human communication is multidimensional and includes the semantic meaning of words, the meaning of expressive verbal gestures (paralingual and prosodic components), the nonvocal gestures, and relational gestures. Even though we are able to record sound and play it back, we typically focus more on the semantic meaning of words. In this paper, we describe a voice analytics toolkit and present visualizations that focus on the relational and expressive verbal gestures in speech. By making the overlooked channels of human communication visible and persistent, we make it possible to see beneath the surface of our words. This insight will potentially enable the development of new applications for speech therapy, the quantization and visualization of vocal trends common to speakers with medical conditions such as autism spectrum disorder (ASD), the characterization and visualization of communication patterns common in different kinds of relationships and cultures, and the development of new kinds of creative, multimodal works.

1. INTRODUCTION

Spoken language, or more generally, communication, exists only in the moment, and cannot be referenced, searched, or analyzed directly. It etches a reflection of itself into the mutable memories of people, and fades away in time. Attempts to retrieve it may fail entirely, produce partial recall, or produce a distorted, inaccurate representation of the original. Spoken communication has all of the properties of ideas in the preliterate society described by Walter Ong [1]. Without literacy, ideas are limited to what one can recall cognitively, and what one can perceive at any given moment in the world. Unless communications are recorded, they are easily distorted and forgotten, and are not available for reference in the world. Furthermore, even when communications are recorded via audio and/or video recording, they are not captured in a form that supports easy analysis of the whole, because in order to reference them, one has to play or search through the sequence once again. The experience is moment-by-moment.

If we look at spoken communication more closely, its streaming, temporal nature is even more apparent. Spoken communication has a semantic verbal channel (the meaning of the words which are uttered), an expressive verbal channel (the paralingual and prosodic features of language), the nonvocal (gestures, eye focus, body posture, etc.), and the relational (the manner in which two more individuals connect and reflect).

Considerable work has been done in interpreting and recording the semantic channel. Sound recording, indexing, and playback are common, everyday events. Live speech recognition tools are available, and speech recognition continues to be an active research topic [2]. Nonvocal communication, particularly gesture recognition, is a current research topic [3] as well. Work on the other two channels, however, lags.

We focus on the analysis and visualization of the expressive verbal and relational channels in this paper. We are particularly interested in the paralingual (pitch, rate, volume, quality, etc.), phonetic (sound content), and prosodic (rhythm, emphasis, and intonation) qualities of voice. The paralingual cues convey emotion, emphasis, humor, and sarcasm. They tell the listener how one is feeling and what he thinks about what is being said. For example, unless someone is a tax accountant, the phrase "I love doing taxes" is likely to be a sarcastic statement. Imagine a slow speaking rate, sharply-articulated, extra emphasis on the word "love," the word "love" spoken at an exaggerated higher pitch, and the end of the phrase dropping in pitch. These cues can be either conscious or unconscious to the speaker. Most listeners, however, perceive these cues very well, and they will hear the paralingual information before the semantic meaning of the words. Paralingual cues outside the norms can also signal medical conditions; many people with ASD, for example, have incorrect or unusual prosody and flat intonation.

We also focus on the interactive, social, vocal elements in a dyadic conversation. These elements represent the relationship between individuals in the conversation, and are apparent in the turn-taking behavior, use of silence, interruptions, modulation of pitch and amplitude, and entrainment. We analyze and visualize this subtle and vital information, to see beyond the words of spoken language.

We believe that vocal visualization, particularly of the expressive verbal and relational channels, may be a useful tool in behavioral analysis, particularly of children with ASD. Many of the visualizations in this paper show discourse from actual screening sessions for autism. We also believe that this kind of analysis and visualization will be even more powerful when speech is analyzed and visualized in conjunction with other signals, such as video and electrodermal activity (skin conductivity measurements). These combined visual analytics may lead to improved diagnostic techniques for ASD, which could lead to earlier intervention and improved prognosis for these children. We also intend to explore the use of these visualizations in the context of speech therapy. People with

ASD often have varying degrees of speech difficulty, and especially have trouble with vocal inflection and other prosodic elements. Comparing a target speech pattern with an actual utterance both visually and aurally could help in the perception and performance of speech. Furthermore, we intend to explore the use of our visual analytics in behavioral feedback. Our tools could be used as a conversational mirror, to highlight paralingual and relational elements in human interaction. Often, people are not aware of their behavior in group interactions, the influences that they have, and the manner in which they are influenced by others. Children with ASD have extreme difficulty picking up on conversational cues, and a conversational analytics tool and social mirror could help make social cues visible to them. Finally, we believe that these techniques have applications in interactive art, and will enable artists to extend the impact of their works by combining visual, aural, and other modalities.

This paper's contributions are the development of a toolkit for vocal analysis, and visualizations of the expressive and relational elements of spoken communications. To put it in Ong's terms, these contributions allow us to deal with spoken language at a new level of literacy.

2. RELATED WORK

Significant work has been done in visualizing sound, from many points of view, including conversation, speech sound, music, sound collections, and general sound classes. Figure 1a-c below shows a sampling of sound visualizations from our work and the most closely related work.

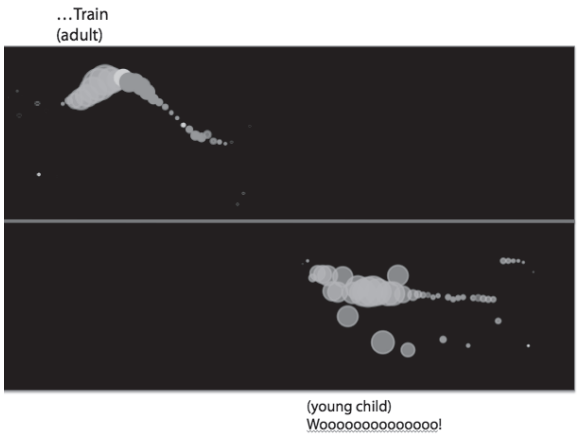

Figure 1-a: A Sonic Shapes visualization of an adult and a very young child talking about trains. The child imitates a train whistle here.

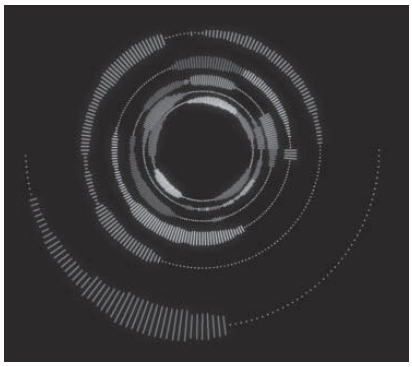

Figure 1-b: From Bergstrom and Karahalios' Conversation Clock [4,14], which shows the exchange among speakers in a conversation over time.

Figure 1-c: From Cho's Takeluma [4]. Live speech maps spoken phonemes into forces which shape a string into a moving "script".

As you can see from the samples, the visualizations focus on either conversational elements, an individual's utterances, structure in sound, or similarity and difference among sounds. Sonic Shapes seeks to visualize both the conversational dimension and the expressive elements of speech sound. More detailed discussion of work in each category of sound visualization follows.

Visualization of face-to-face (F2F) spoken conversation has many important elements. Most of this work shows when people speak, how much each person speaks, how loudly each person speaks, turn-taking, and simultaneous speaking. The Conversation Clock [5] shows the progression of a conversation over time via concentric clockwise circles. It is apparent when each speaker speaks, when there is silence, when interruptions occur, and when normal handoff from one speaker to another occurs. The relative amplitude of all speech over time is apparent, as is the relative amount of time each person speaks. Conversation Clusters [6] extracts the important words/concepts from speech, and displays them in relationship to one another. When enough related words appear, they form visual groupings, or concept clusters. Okwechime et al. [7] analyze a combination of social signals in conversation, visualize a 7-dimensional Social Dynamic Model (SDM) in the form of concentric rings around 7 axes, where the size of the nodes at the intersection of axes and rings indicate the value of the dimension for a given person involved in the conversation.

A similar body of work visualizes conversation, but focuses on text instead of live audio. Some of these visualizations focus more on the semantic channel and follow topics of conversation and the relationships among the conversation topics and the people involved in them. Angus et al. [8] create similarity matrices of concepts, comparing the concept under discussion at a given point in time with the concept under discussion at all other points in time. Periodicities and

repetitions show as distinct diagonal patterns in the visualization. Yet others show the structure of a conversation. For example, CrystalChat [9] combines the idea of a social networking graph with a structural representation of conversations. A 3D hub-and-spoke visualization shows a person's conversations with others as the spokes, and the plane of the spoke shows the conversations as strings of beads, with each bead representing a message. Darker beads indicate longer messages. The color shading on the backplane gives information about the emotional content of the message (by analysis of emoticons). Anther work by Tat [10] visualizes conversation by extracting the character from a time-series conversation and presenting it in 2D static summary visualizations. Pupyrev and Tikhonov [11] visualize online conversations with dynamic graph drawing using multidimensional scaling techniques, and show how their visualizations could help reveal temporal patterns. Hansen et al. [12] focus on threaded conversation networks. The Fugue [13] system likens conversation to a musical fugue, and focuses on the timing and paralingual elements of the conversation. The visualization stretches each person's words and cues over time, like music on a score, so that you can see who is speaking, and when. Donath [14] describes a system that emulates F2F conversation in a room, in that the user navigates though the space in order to interact with others. Only users who are in close proximity can be heard.

A third body of work focuses on the vocal quality of an individual's speech. Ueng [15] et al. provide Sammon mapping, parallel coordinates visualizations, and mappings of features onto a 2D polar coordinate plane. LocuTour's [16] commercial product targets the speech therapy market, and provides visualizations and therapy plans suitable for addressing phonological problems. Hailpern [17,18] provides comparative visualizations of an actual utterance's prosody vs. correct prosody. This visualization features pitch contour and vocal stress in a set of target words. Fell [19] visualizes syllables and infant vocalization age. Cho's Takeluma [4] translates spoken phonemes into shape and motion. It appears as though the energy of the voice shapes a flexible string into script. In time, the energy dissipates, and the string returns to its original, flat shape, until activated by the next utterance.

Music and generic sound visualizations are relevant here, too, because musical timbre is analogous to paralingual and prosodic elements of the voice. Seidenberg [20] focuses on six features (centroid, spread, skewness, flux, bark-flux, centroid-flux) and visualizes them 6 ways: 1) scatterplots, 2) slider plots, 3) histograms, 4) Lineto charts, 5) "river" plots, which show the flow of sound features over time, and 6) a similarity matrix. Foote [21] presents the self-similarity matrix which shows periodicities and repetitions, and provides a structural summary of the sound components in the music. Chan et al. [22] highlights musical voicing roles with layer braid diagrams, and structural elements and their relationships with a "theme fabric" diagram. A "collapsed theme fabric" diagram shows both theme and voice in the same diagram. Wattenberg's beautiful Arc Diagrams [23] visualize the structure and periodicity in data sequences, including sound and music. Artistic performances, such as Messa di Voce [24], have included evocative, interactive voice visualizations as well. The various sections in this piece demonstrate multimodality, and include interactivity between speakers, voice and visuals, and full-body interactions with the visuals.

Finally, work in sound collection visualization is significant, because its purpose is distinguishing and comparing sound character. The Sonic Browser [25] attempts to facilitate human browsing behavior by allowing user-defined mapping of features onto visual display components (x, y, color, shape of icon, size of icon), defining an aura (perceptual range), and plays anything that the user selects which is within the range of the aura. Elements which are close to one another sound more similar than those that are far apart on the visualization, so the user can sample the sonic character of the space by selecting the range visually and listening. Adamczyk [26] describes music genre browsing via data collected from social networks. His visualization techniques range from simple 2D to immersive 3D. Morchen [27] visualizes timbre difference over sound collections via mapping to a physical terrain model. The MEL-IRIS project [28] presents a multiple views visualization featuring chroma, and Adiloglu [29] provides a sound taxonomy, and visualizes a "spike code", which is a simplified time-frequency graph.

Many of the existing sound visualizations are suitable for engineers and domain experts, but are not easy for other users to understand. Not all users, for example, know what a centroid and bark-flux are; and not everyone knows how to read a spectrogram. Fewer still would know how to connect these representations with their voice, interpret the resulting measures, and understand why they might be important. Furthermore, the individual applications are limited in scope; none of them focus on both the expressive verbal and relational channels. We present analytics and resulting visualization techniques that are suitable for both of these channels, and that map closely to perceived qualities in the voice.

3. VISUALIZATIONS

The visualizations below show phonetic sound quality, pitch contour, breathiness, noisiness, and sound amplitude. Time progresses on the horizontal axis, and increasing pitch along the vertical y axis. Rising inflections curve upward, and falling inflections, downward. The size of the graphic shows the relative amplitude (the larger the graphic, the louder the sound). Figure 2 below shows the color mappings used in the visualizations. Obstruent, or noisy, consonants appear in blue, with the noisiest consonants having the most saturation. Sonorant consonants (the sounds which could be sung) appear in green, with the most sonorant options having the most color saturation. All vowels are pink, with the most open vowels having the strongest saturation. We chose to highlight the vowels with warm colors, because we thought that the sustained, singable quality of a vowel was more associated with warmth than the "crunchy" articulations of consonants.

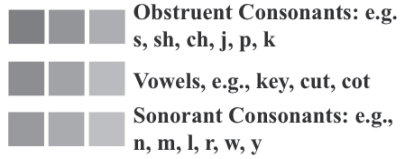

Figure 2: Current color map for phonetic quality

Figure 3 shows a child saying "no" and "now" (similar words). In the word "no," you can see the sonorant "n", the relatively loud and extended "o" sound, and an aspiration of air at the end (shows as blue, a noisy consonant). The child's voice has a slight inflection in pitch. In the word "now," you can still see the sonorant "n," but you can also see the child having some noisy air in the transition between the "n" and the "ow". "Now" is harder for a very young child to say than "no". Note that the child's voice rises while saying this word (sounds like a question), and there is a small bit of noise at the end of the word. The visualizations show the subtle sonic differences.

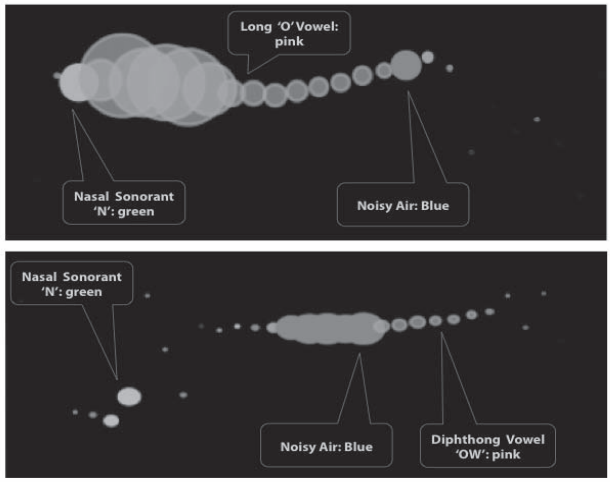

Figure 3: Top: the word "no"; Bottom: the word "now".

Figures 4 and 5 below show a child and therapist in a screening session for autism. The child is about 2 years old, and in the process of acquiring language. They are building a toy sandwich together in an unstructured interaction. The child is particularly fond of cheese, but can't say the word "cheese" very well yet. It comes out sounding like a noisy blend of "ttssseesss". The adult is talking in a slow, sing-song voice to the child, with some of the sounds elongated. The rise and fall of the adult's sing-song intonation is visible in Figure 4 in the words "Lettuce, cheese." You can also see some possible entrainment, where the two parties move together to become more similar in manner of speech. Note the similarity of the word "turkey" spoken by both parties in Figure 4.

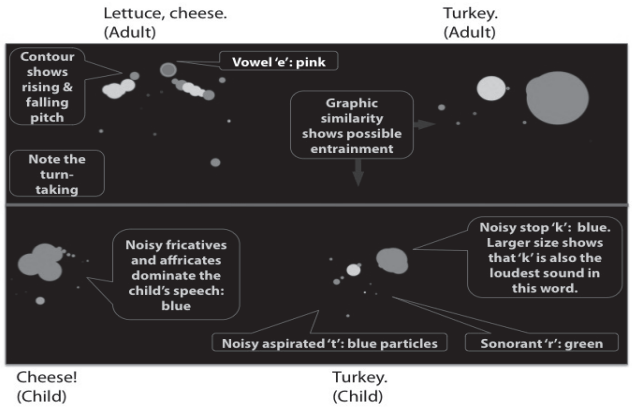

Figure 4: Building a toy sandwich, part 1

In Figure 5, the adult is trying to encourage the child to include lettuce in the sandwich, but the child, who is not fond of lettuce, will have none of it. Here, his speech becomes a very clear "NO ON!"

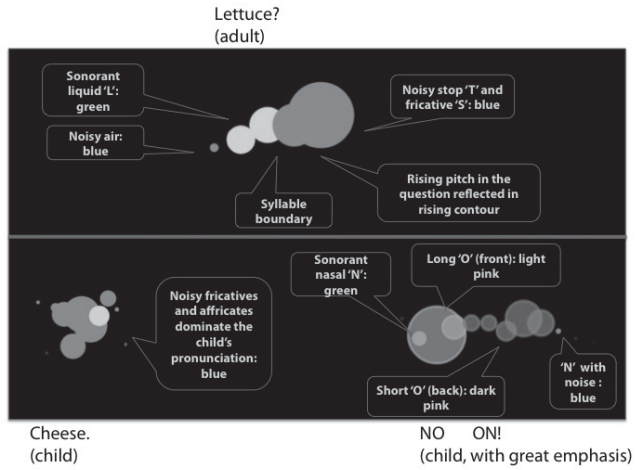

Figure 5: building a toy sandwich, part 2

Figure 6 below shows how the system could be used for speech therapy. The top panel shows the correct pronunciation of the word "lisp", while the bottom panel shows the word "lisp" spoken with a lisp. You can clearly see the difference between the correct "s" sound and the lisp, and you can see the effect that the lisp has on the preceeding vowel sound.

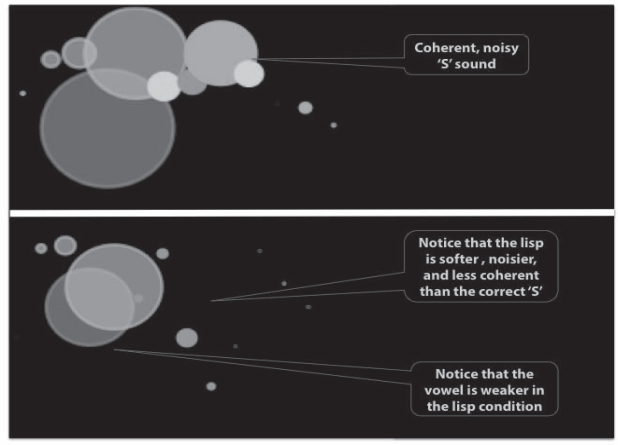

Figure 6: Top = woman saying the word "lisp" correctly; Bottom = same woman saying the word "lisp" with a lisp.

Figure 7 shows that the system can show differences in emotion and prosody. Here, the same speaker is saying "Hello, world" in the style of a statement and a question. Note that the end of the statement is de-emphasized, but the end of the question is inflected and emphasized. The speaker continues to play with her voice in the last two panels of Figure 7, this time saying "Hello, world" with different emotions. The "bored" utterance has elongated vowels and sonorants, less inflection, is even in volume, and is less varied overall than the other utterances. The "angry" statement clearly looks angry.

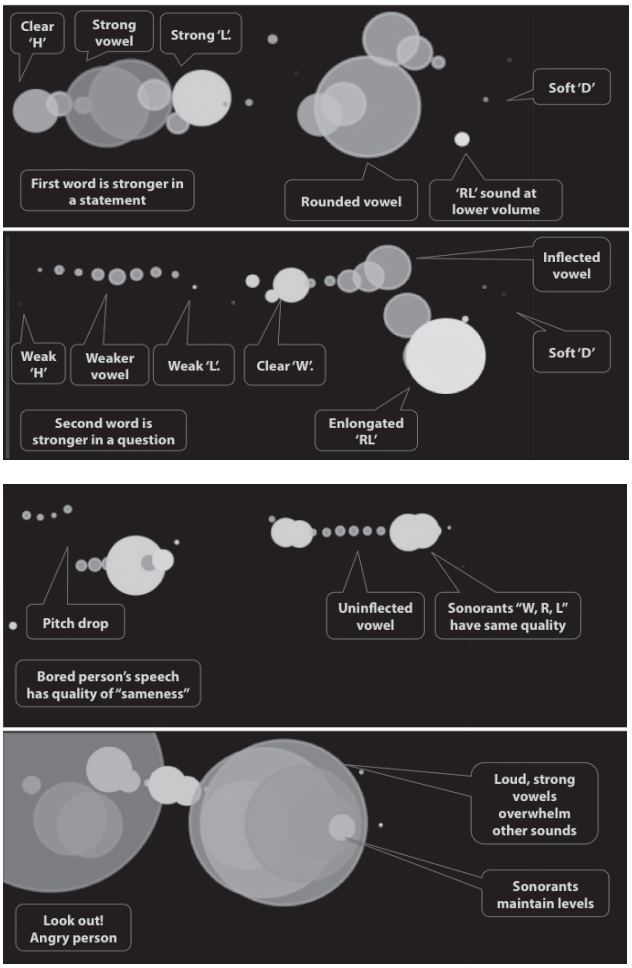

Figure 7: Top = person saying "Hello, world" as a statement; 2nd = same person saying "Hello, world" as a question; 3rd = same person saying "Hello, world" with bored emotion; Bottom = same person saying "Hello, world" with angry emotion.

Earlier versions of the software included more detailed phonetic classification in the visualizations. Figure 8 shows the color map for the earlier prototype, and Figure 9a-c show utterances by three different people (two men and one woman). We thought that showing that level of detail in the phonology might distract from our purpose of visualizing the paralingual dimension of the voice. This more phonologically rich version, however, will still be useful in speech therapy applications, in creative works for aesthetic reasons, and anytime phonological detail is desired. Also note the that in figures 9a-c, the sampling rate was 44.1KHz, and the graphic resolution is correspondingly higher than in previous figures (sampling rate 16KHz). Finally, note that the earlier version of the color map uses warm colors for consonants and cool colors for vowels (opposite of the previous approach). We thought that using warm colors wold draw the eye to the noisy, textured sounds here.

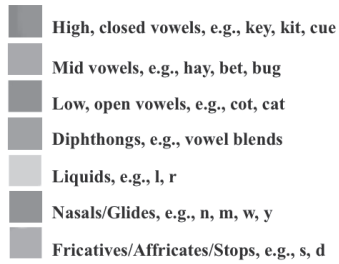

Figure 8: Color map for early prototype

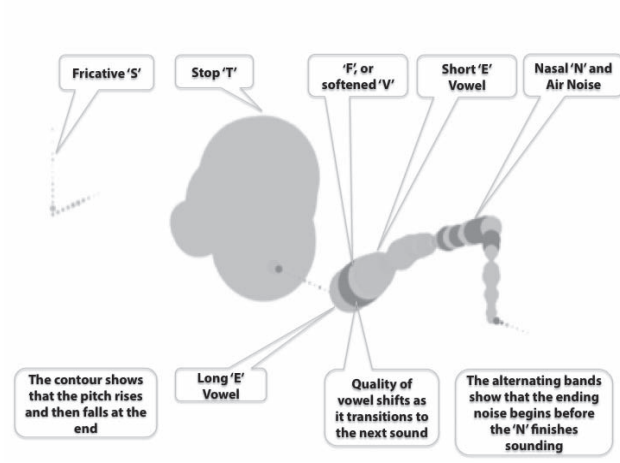

Figure 9a: An adult male saying his name, "Stephen"

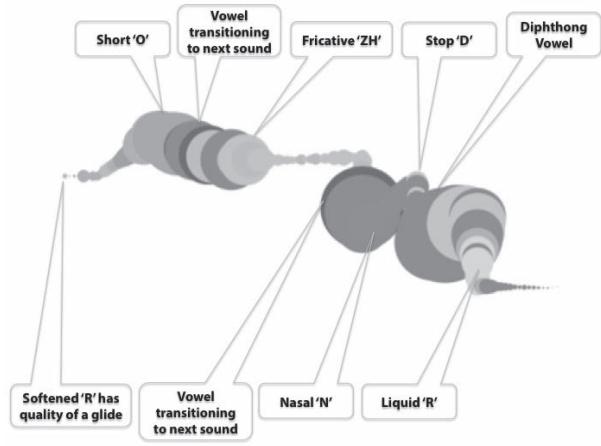

Figure 9b: another adult male, saying his name, "Rajinder"

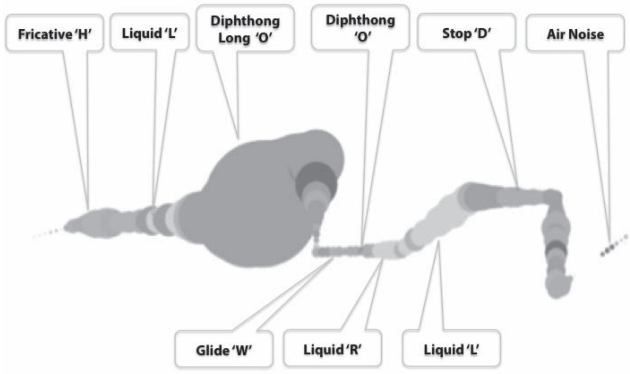

Figure 9c: An adult female saying "Hello, World"

4. SYSTEM DESCRIPTION

4.1. Preprocessing

Figure 10 below provides an overview of the system. The first step in processing speech sound was reducing the input into a form that emphasized the most important information and excluded irrelevant details. The quality of the analytic output, even the ability to generate meaningful output at all, depended on the quality of work done at this stage. This meant downsampling and filtering to improve the performance (particularly on the constrained platform of an iPad or iPhone), followed by extracting a set of audio features commonly used for representing speech [30, 31], including the Mel Frequency Cepstral Coefficients (MFCCs), pitch, normalized amplitude values, other spectral features, and the relative content of noise in the signal as determined by the ratio of harmonic to inharmonic frequencies in the sound. We also considered formant extraction, but deferred including formants in this version of the implementation for performance considerations and because the increase in accuracy of phoneme class detection was not large enough to justify the cost of the additional processing at this time.

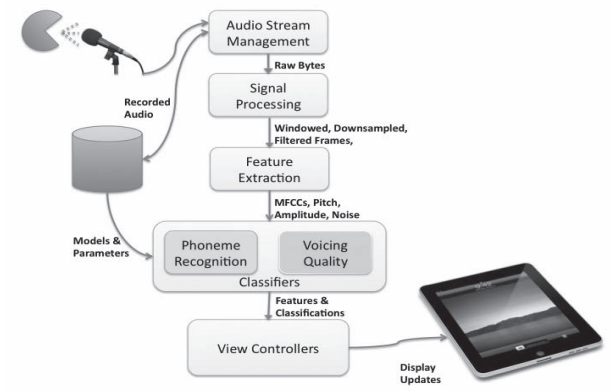

Figure 10: System Overview

We experimented with various sampling rates and found that sampling at (or downsampling to) 16KHz produced the best combination of performance and visualization quality. In general, higher sampling rates provided more information, and higher quality visualizations, but lower sampling rates provided better performance. Unless someone has a trained ear and is listening and comparing, hearing the difference in the sound for the different sampling rates is difficult. The impact on analytics was also minimal; downsampling did not introduce errors except for the occasional omission of short, softly-uttered phonemes. The largest impact of downsampling was the decrease of nuanced data available for visualization.

Additionally, processing real-time, streaming audio required framing, with a carefully-selected frame size and advance rate and overlap between signal slices. All signals analyzed in this paper used 1-second frames with a 250 msec frame advance. This meant that each 250 msec unit of sound would present itself to the system in different positions of 4 different frames, effectively taking a snapshot of the latest 1-second of audio every 250 msec. The 1-second frame size gave us optimal performance. Smaller frame sizes presented difficulty detecting longer vocal gestures, and larger frame sizes tended to overlook short sound gestures and nuance (small details got "lost" in the larger analytic scale). The advance rate was also carefully selected for optimal performance. Smaller overlap introduced misclassification errors.

4.2. Analysis and Classification

Two commonly-used machine learning models were included in this implementation: a simple Gaussian classifier, and a Hidden Markov Model (HMM) [32, 33]. We used these models to detect phoneme classes and voicing quality. Training data for the system included 10-20 samples of each sound, collected at 44.1KHz, from adult men and women primarily with Midwestern dialects. One adult man, however, was from the west-coast region of the United States and had virtually unaccented speech. One notable difference in speech between the Midwestern speakers and the west-coast speaker was the pronunciation of 't' in the inner syllables of words. Midwesterners typically pronounce this sound like soft 'd' (called a 'tap' by speech and language pathologists), while the west-coast speaker used the same stopped 't' sound that they did in the beginnings and ends of words. Another difference was that the Midwestern speakers gave a stronger diphthong (blended vowel) character to long vowels. No children's voices were used to train the system, because of the unavailability of suitable corpus. The system classified children's voices almost as well as adult voices, however.

We trained the system to recognize phonemes by class. All vowels are formed by raising and lowering the tongue in the mouth (close/mid/open), and by directing the focus of the sound forward or backward (front/central/back). Figure 9 shows the monophthong vowels and their placement. Notice that some variation occurs with regional dialect.

	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 9: Monophthong Vowels

Diphthong vowels are blended, two-part sounds where the first sound "morphs" into the second. We detect the diphthong vowels heard in the words "cane," "coy," "coat," "core," "kite," and "cow".

Individual consonants are much more difficult to distinguish, but for the purposes of this work, recognizing the following English consonant classes is sufficient:

- Stops: /p/ pan, /b/ ball, /t/ tan, /d/ Dan, /k/ can, /g/ gas
- Fricatives: /h/ hat, /f/ fat, /v/ vat, /s/ sat, /z/ zap, /sh/ shack, /zh/ luge
- Nasals: /m/ mat, /n/ gnat, /ng/ sing
- Liquids: /l/ lamp, /r/ rat
- Glides: /w/ wing, /j/ yes, /kw/ quick, /ks/ box
- Affricates: /ch/ chip, /nch/ bench, /j/ jello

The phoneme detection analysis in this paper is based on training over recordings of 22 classes of isolated phonemes, outside the context of real speech. This approach worked better than using in-context phonemes (phonemes isolated from words in real speech) for training, probably because of the transitional qualities between phonemes. In real speech, a phoneme typically begins sounding before its predecessor ends. Furthermore, real speech has prosodic cues embedded in the utterances that are unrelated to a phoneme itself. The system recognized phoneme classes accurately over 70% of the time at this level of granularity, and some of the phonemes were consistently recognized over 95% of the time. At the end of this processing pipeline, the most common errors were "dropped" sounds, usually sounds uttered too quickly and softly to be detected accurately.

5. FUTURE WORK

We have demonstrated an analysis and visualization of the human voice that highlights the expressive vocal and relational channels. These visualizations can be used many ways, including applications for speech therapy, as a tool to help screen for autism, a tool to highlight the expressive content in a speaker's voice, and a means to make the ephemeral, qualitative voice persistent. We have just begun to explore the visualization possibilities. A possible next step is creating a suite of visualizations to address an expanded set of qualitative elements of speech. It would be useful to present these at different levels of detail, and include some visualizations that summarize vocal qualities. These could be released as a package and made available to iPad/iPhone developers. We might also improve the resolution of downsampled visualizations, and we are actively exploring options toward this end.

A companion to the visualization package could be an extended package for voice and sound analysis on iOS. Many other analysis techniques and representative features are possible, and a toolkit could enable the development of new applications. It would be enlightening to focus on enabling the creation of new media for artists, who could help us expand the visual repertoire.

The scope could also be expanded to include multimodal analysis and visualization. For example, we could process video with sound, or sensor data with sound. We could also incorporate the vocal semantic channel into our work.

We did not address the need for working specifically with children's voices. This need should be addressed by collecting speech samples (for training and testing) from children, and extending the analytics appropriately so that they will process children's voices efficiently.

We also did not address the mobility that iPhones and iPads provide. We could support dyadic and multi-way conversation across devices, do analysis on a server apart from the device, and display the results on yet another device (such as a public display wall).

Finally, soliciting feedback would help prioritize the next steps. Informally, feedback has been good. Users liked the visualizations, and enjoyed experimenting with their voices. Some liked the color selections, and many thought that the high resolution visuals were beautiful. A user study across the suite of possible visualizations would help focus new development

6. CONCLUSION

In this paper we showed that it is useful to visualize the expressive and relational parts of spoken communication, and that both of these channels can be addressed in the same application. We also demonstrated that these analytic and visualization techniques could enable a range of applications, such as 1) speech therapy tools, 2) the quantization and visualization of voice characteristics for typical speakers, and for speakers with medical conditions such as ASD, 3) the characterization and visualization of communication patterns found within different relationships and cultures, and 4) the creation of new kinds of multimodal creative works. And, we demonstrated that it's possible to do this in a constrained, mobile environment. This paper's contributions are 1) the development of a simple toolkit for vocal analysis on a mobile platform (iPad/iPhone), and 2) development of visualizations representing the expressive and relational elements of spoken communications, and 3) the enabling of possible new applications described above. We envision a path forward that will expand the capabilities of the analytics and visualizations, and believe that the research trajectory will extend into multimodal analysis and visualization. The challenge will be enabling a new level of literacy, while discouraging a new level of complexity.

7. ACKNOWLEDGMENT

Many gracious thanks to the United States National Science Foundation, which sponsored this work in part through Award CCF-1029679.

8. REFERENCES

[1] Ong, W., Orality vs Literacy. Routledge, New York, 2002.
[2] Rabiner, L. R., and Juang, B., Fundamentals of Speech Recognition, Prentice Hall, 1993
[3] Yang, M., Face Detection and Gesture Recognition for Human-Computer Interaction, Springer, 2001.
[4] Cho, Peter, Takeluma: An Exploration of Sound, Meaning, and Writing, MFA Thesis, UCLA Department of Design | Media Arts, June 2005
[5] Bergstron, T., and Karahalios, K., Conversation clock: Visualizing audio patterns in co-located groups, HICSS '07.
[6] Bergstrom, T., and Karahalios, K., Conversation Clusters: Grouping Conversation Topics through Human-Computer Dialog, Virtual Reality 2009.
[7] Okwechime, D., Ong, E., Gilbert, A., Bowden, R., Visualisation and Prediction of Conversation Interest through Mined Social Signals, Social Dynamics 2013.
[8] Angus, D., Smith, A., Wiles, J., Conceptual recurrence plots: revealing patterns in human discourse, IEEE TVCG 2012.
[9] Tat, A., Carpendale, S., CrystalChat: Visualizing Personal Chat History, HICSS '06.
[10] Tat, A., Carpendale, S., Visualising Human Dialog, Work, 2002.
[11] Pupyruv, S., and Tikhonov, A., Analyzing conversations with dynamic graph visualization, ISDA 2010.
[12] Hansen, D., Schneiderman, B., Smith, M., Visualizing threaded conversation networks: mining message boards and email lists for actionable insights, Active Media Technology 2010.
[13] Rosenberger Shankar, T., VanKleek, M., Vincente, A., Smith, B., Fugue: A Computer Mediated Conversational System that Supports Turn Negotiation, Text 2000.
[14] Donath, J., Karahalios, K., Viegas, F., Visualzing conversation, HICSS '99
[15] Ueng, S., Luo, C, Tsai, T., Chang, H., Voice Quality Assessment and Visualization, CISIS 2012
[16] Scarry-Larkin, M., Speech Visualization, LocuTour Multimedia, LocuTour, Inc.
[17] Hailpern, J., Karahalios, K., Halle, J., Creating a Spoken Impact: Encouraging Vocalization through Audio Visual Feedback in Children with ASD, CHI '09
[18] Hailpern, J., et al., VocSyl: Visualizing Syllable Production for Children with ASD and Speech Delays, Extended Abstracts of ASSETS 2010.
[19] Fell, H.J., and MacAuslan, J. Vocalization Analysis Tools, MAVEBA 2005.
[20] Siedenburg, K., An Exploration of Real-Time Visualizations of Musical Timbre, Proceedings of the 3rd international workshop on learning semantics of audio signals, 2009.
[21] Foote, J., Visualizing music and audio using self-similarity, MULTIMEDIA '99
[22] Chan, W., Qu, H., Mak, W. , Visualizing the Semantic Structure in Classical Music Works, IEEE Transactions on Visualization and Computer Graphics, Jan. 2010.
[23] Wattenberg, M., Arc Diagrams: Visualizing Structure in Strings, InfoVis 2002.
[24] Levin, G., Lieberman, Z., Blonk, J. LaBarbara, J., et al., Messa di voce project report (for performance version), http://www.tmema.org/messa, 2003.
[25] Brazil, E., Fernstrom, M., Audio Information Browsing With the Sonic Browser, Coordinated and Multiple Views in Exploratory Visualization, 2003.
[26] Adamczyk, P., Seeing Sounds: Exploring Musical Social Networks, MULTIMEDIA '04.
[27] Morchen, F., Ultsch, A., Thies, M. et al., MusicMiner: Visualizing timbre distances of music as topographical maps, ISMIR 2008.
[28] Margounakis, D., Politis, D., Mokos, K., MEL-IRIS: An Online Tool for Audio Analysis and Music Indexing, International Journal of Digital Multimedia Broadcasting, 2009
[29] Adiloglu, K., Annies, R., Wahlen, E., et al., A Graphical Representation and Dissimilarity Measure for Basic Everyday Sound Events, IEEE Transactions on Audio Speech and Language Processing 2012.
[30] McKinney, M., Breebaart, J., Features for audio and music classification, ISMIR 2003.
[31] Peeters, G., A large set of audio features for sound description (similarity and classification) in the CUIDADO project, Ircam, Analysis/Synthesis Team Technical Report, http://www.ircam.fr/
[32] Rabiner, L.R., A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, Vol. 77, No. 2, Feb 1989.
[33] Vaich, T., and Cohen, A., HMM phoneme recognition with supervised training and Viterbi algorithm. Eighteenth Convention of Electrical and Electronics Engineers in Israel, 1995.