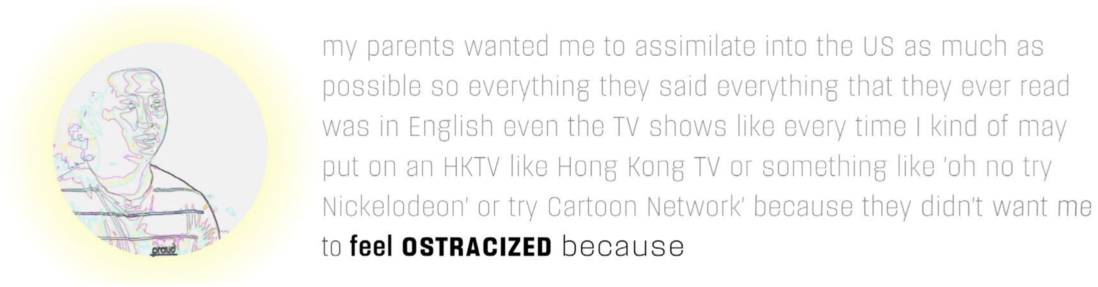
PARALINGUISTIC HYPERTEXT: Visualizing Conversation Through Expressive Digital Text

WONKI KANG, MEGAN PRAKASH, Massachusetts Institute of Technology, USA



my parents wanted me to assimilate into the US as much as possible so everything they said everything that they ever read was in English even the TV shows like every time I kind of may put on an HKTV like Hong Kong TV or something like 'oh no try Nickelodeon' or try Cartoon Network' because they didn't want me to **feel OSTRACIZED** because

Fig 1. Animated text view with corresponding video clip.

CCS CONCEPTS • **Human-centered computing; Data Visualization**

Additional Keywords and Phrases: ambiguous interface

ACM Reference Format: [...]

1 INTRODUCTION

Text medium is seemingly flat and linear, but this was not the case before the typewriter. Handwritings and calligraphies often express the writers' personalities and identities. Mechanical writings and digital texts, however, tend to flatten the complex dimensions into one single stream of letters. Through this project, we want to introduce rich and expressive affordances of writings in digital typefaces by adding high dimensional data associated with the text, namely paralinguistic cues, such as prosody, pitch, volume, intonation, or hesitancy. Our data consists of recorded speech which contains a conversation between immigrant parents and first-generation (US) Americans. We aim for an interactive and expressive visualization that encodes what happens throughout the conversation, inviting the viewers to make meaningful comparisons and associations, and to easily grasp the themes and sentiments even without watching or listening to the video.

The project is twofold. First, we unpack the contents of the conversation and lay out our findings in a web environment based on their themes. These include family, belonging, cultural heritage, as well as the speakers, whether they are immigrants or first-generation Americans. Second, we present "hypertext," an augmented form of digital text, using variable fonts that allow continuous interpolation of font styles — thickness, obliqueness, width, size, or even the design of the embellishments. The text is presented in real time, in sync with the audio that is being played alongside.

2 BACKGROUND

2.1 Representing Speech

The musical quality of speech and how to represent it has been studied for centuries in various disciplines including linguistics, phonology, and music theory. Throughout the long-lived history of musical notation, composers have come up with symbols and marks to communicate subtle intonations for vocal music. Joshua Steele (1700-1791) focused on the musical quality of speech — or linguistic quality of music. He proposes a musical model of speech notation (Fig 2) that can incorporate both melody and rhythm. [2]

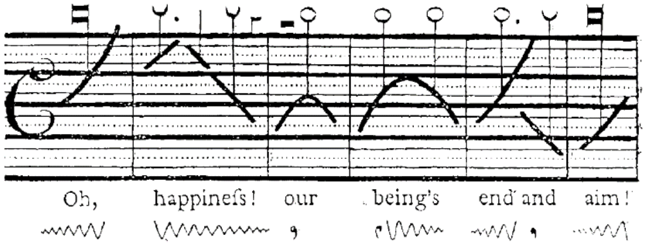


Fig 2. Notation of Spoken Words from *An Essay Towards Establishing the Melody and Measure of Speech to be Expressed and Perpetuated by Peculiar Symbols* by Joshua Steele, 1775, p.13.

Although the purpose of the notational system is essentially to deliver the composer's intentions to performers effectively, experiments of the musicians to constantly push the boundaries of what is considered "musical sound" gave rise to expressive graphics beyond the conventional, Western Art Music staff notation of the time.

2.2 Computer Input

There has been a significant body of research that focuses on extracting behavioral biometrics through digital input devices of computers. The works are in large part aiming for identification, classification, and authentication of the users. For example, Bergadano et al. [2] analyzed keyboard stroke dynamics to identify users, while Banerjee et al. [3] used the keystroke patterns to detect deceptive writings. Meanwhile, Jonathan Zong designed a novel typeface named *Biometric Sans,* which allows a variable stretch that responds to the typing speed of users. Zong describes his work as "the re-introduction of the hand in digital writing." [4]

2.3 Our Take

The challenge of communicating paralinguistic cues through visual media is due to its difficulties in quantifying them. While a number of machine learning tools are seemingly smart enough to be able to extract complex sentiments through analyzing speech, it should be noted that it is not our aim to label or classify the identities and sentiments of the speakers. Rather, we use computational tools to gently extract biometric cues of the speech to help us mediate between the speakers and the viewers in a more expressive way.

3 METHODS

3.1 Raw Data

Our raw data is extracted from a 25-minute-long video of a conversation between immigrant parents and first-generation Americans sharing their experiences living in the United States. The conversation is generally calm and informal but has complex sentiments expressed from both sides.

In a broad scope, we analyze the contents by looking at the themes that are shared throughout the speech like cultural heritage, sense of belonging, and family. In a narrower scope, we focus on word by word, how some words are emphasized while others are not.

3.2 Data Processing

Again, the recurring themes include family, heritage, belonging, etc., and the speakers refer to these themes using longer sets of words like “my kids,” “Venezuela,” and “It's been so difficult to teach [my kids] Spanish.” These themes compose to form higher-order sentiments when you collect all the times that speakers refer to them and compare what they say.

For example, both immigrant parents and first-generation American-born individuals will refer to the theme of belonging by mentioning their origin cultures and American culture. Using “origin culture” as a nucleus, we can collect all the ways that immigrants and first-gens express their relationship to their origin culture. Then, we use “American culture” as a nucleus to see how both groups of people relate to America. Lastly, we visualize those phrases around their respective nuclei. The semiotics definitions are based on the logic as the following:

- Phrases contain semantics and paralanguage.
- Semantics include the particular words that the speaker uses to refer to common concepts, and how they relate to those concepts.
- A common concept forms a nucleus for collections of phrases.
- Nucleus concepts can be composed to form higher-level themes.

In the meantime, as the data is time-based, it is crucial to segment and process the data based on the time signatures. We use Google speech-to-text API to extract the body of texts and the timestamp of each word within the text.

The integrated database is structured like the following:

```
{ …
   {“id”: 3,
    "full_text": "my parents grew up in Cultural Revolution China…",
    "speaker": "firstgen",
    "timestamp": [{"word": "my", "weight": 1, "startTime": 0.0, "endTime": 0.2},
                  {"word": "parents", "weight": 3, "startTime": 0.2, "endTime": 0.5},
                                                                        … ]}}
… }
```

3.3 Visual Encoding

Variable fonts are typefaces where the font styles are parameterized allowing a continuous transition between styles that would otherwise require an entire array of typefaces with the styles in between. This is likely to be extremely computationally costly. The parameter space is called “axis.” We use three axes for our visual encoding: thickness, width, and slant (italics). The thickness is determined by the “weight” of each word, manually transcribed as a channel to encode the level of emphasis in the speech. The width depends on how fast each word is spoken, and the slant is conditional to the length of the pause in between the words.

4 RESULTS

The variable font was proven very effective in creating real-time text animation which requires a robust interactivity.

5 DISCUSSION

There has been a growing debate within the data visualization community over the ethical dimensions in the process of reducing the data into abstract representations: “is representing COVID-19 deaths to dots ethical?” Researchers like Michael Correll at Tableau Research argues that the data should be more anthropomorphic, actively showing that there are actual human beings behind the data. [5] However, we argue that there are alternative ways to represent data without using overt human figures while not compromising the empathetic and evocative quality to it. We intentionally de-anthropomorphized the video by filtering it to outlines in order to make our visualization more engaging to the viewers. It also allowed the viewers to focus more on the text and the audio, offering a more evocative experience than watching the original video. Also, as discussed from Section 4, the animated text takes on a certain character, or even identity, exhibiting a life-like quality.

6 FUTURE WORK

For the scope of the project, encodings of the typefaces are mostly done manually which makes it less scalable. In the future, the encodings can be more automated that will allow the scheme to be applied to larger bodies of texts and conversations. Also, variable font has an even bigger potential with customized axes in addition to existing parameters. It would be worth exploring what axes can be added that are suitable to express paralinguistic cues within text.

REFERENCES

[1] Jamie C. Kassler. 2005. Representing Speech Through Musical Notation. *Journal of Musicological Research* 24, 3–4: 227–239. https://doi.org/10.1080/01411890500233965

[2] Francesco Bergadano, Daniele Gunetti, and Claudia Picardi. 2002. User authentication through keystroke dynamics. *ACM Transactions on Information and System Security* 5, 4: 367–397. https://doi.org/10.1145/581271.581272

[3] Ritwik Banerjee, Song Feng, Jun Seok Kang, and Yejin Choi. 2014. Keystroke Patterns as Prosody in Digital Writings: A Case Study with Deceptive Reviews and Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1469–1473. https://doi.org/10.3115/v1/D14-1155

[4] Jonathan Zong. 2020. Biometric Sans and Public Display: Embodied Writing in the Age of Data. Retrieved May 19, 2021 from https://jonathanzong.com/blog/2020/05/31/biometric-sans-and-public-display-embodied-writing-in-the-age-of-data

[5] Michael Correll. 2019. Ethical Dimensions of Visualization Research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300418