Intuitive data representation techniques for representing para-linguistic speech data

AKASH CHAUDHARY, University of California, Santa Cruz, USA
ALEX PANG, University of California, Santa Cruz, USA

ESL(English as second language) speakers have difficulty in expressing their intentions, especially those who have syllable-timed languages as their native language. This reduces scope for effective communication, drastically impacting the quality of lives of ESL speakers, and further having a direct effect on job opportunities. We design visualizations for non-native English speakers to understand intonations, intensity, gaps and extended syllables in speech. The resulting textual representation system can help in better perception and understanding of speech modulation, thereby improving communication skills of people.

CCS Concepts: • **Human-centered computing** → *Empirical studies in HCI*; *Usability testing*; **Sound-based input / output**; **Auditory feedback**; *Natural language interfaces.*

Additional Key Words and Phrases: Learning application; Context-based learning; Stress-timed language; Intonations; Communicative expressions

ACM Reference Format:
Akash Chaudhary and Alex Pang. 2021. Intuitive data representation techniques for representing para-linguistic speech data. 1, 1 (December 2021), 12 pages. https://doi.org/10.1145/3447526.3472057

1 INTRODUCTION

English has become the common language world over for people to communicate with each other. Due to the rampant globalization in the last decade, not only has English become important for people to enjoy better job opportunities, but it also has implications on the their personal lives, like making new friends and being effective public speakers. More than 272 million migrate every year in search of job opportunites, many of whom come from countries that predominantly speak syllable-timed languages, like India[3, 27]. Languages are either syllable-timed or stress-timed, and English falls towards the stress-timed language model spectrum[11, 17, 25]. This means that English is spoken in packets or chunks of stress or intonations, containing relevant information about the context in which it is spoken and inducing intentionality of the speaker in speech[5, 14, 28].

This is a problem for non-native English speakers, specifically people with a syllable-timed native language, as they tend to speak in regular patterns of syllabic duration, making them sound monotonous, and hence, incapable of expressing intentions in the natural style of English speaking [16]([5, 14];[28], p. 81; [19], p. 283).

Some apps try to help people learn intonations by making them listen to audio and understand the various subtleties in speech through auditory recognition by ear[1]. However, the perception of these subtleties can increase if they also

Authors' addresses: Akash Chaudhary, \unskip,University of California, Santa Cruz, Santa Cruz, California, USA, 95060; Alex Pang, \unskip,University of California, Santa Cruz, Santa Cruz, California, USA, 95060.

visualize these audio cues visually through added visualization layers on top of the English orthographical system of writing. Verbose[9] is the only app that helps people understand intonations by adding visualization layers on top of the English textual written system. We follow in the footsteps of the work done in this app, and add some some more paralinguistic based stress features in our visualizations like intensity, gaps and extended syllables.

We design textual visual represents that contains the following features -

- information on raw paralinguistic parameters of specch, like intonations (pitch modulation), intensity(loudness), gaps(silences) and extended syllables.
- is able to represent all the above information together in one visual representation.
- is context-based.
- enables ease of legibility.

2 RELATED WORK

In this section, we explore the works which have tried to visualize the paralinguistic parameters of speech audio.

2.1 Audio-only based systems

Here we explore works which have focused on using audio-only based teaching techniques to teach people how to speak efficiently.

CALL (Computer Assisted Language Learning) [10] uses imitation-based teaching techniques for users to develop listening and speaking skills. My English Tutor (MyET) [23] also uses imitation-based teaching techniques to teach various accents of different teachers and thereby improve their pronunciation. SLION [24] uses imitation based teaching combined with automatic speech recognition for a karaoke app. However, our system is intended toward combining an audio-only based system along with a visual representation of para-linguistics in textual writing.

2.2 Context-less visualization of paralinguistics

Here we explore the various works which have represented para-linguistic parameters of speech to improve speech communications without the use of context.

Applications like Rhema [30], Logue [15], AwareMe [7], Aging and Engaging [4], ROCSpeak [33], RoboCOP [31], VoiceCoach [32], MACH [18], and Automated Social Skills Trainer [29] provide feedback on word count [8, 15, 18, 20, 29–32], loudness or intensity [4, 8, 30, 32, 33], gaps or pauses [18, 29, 32], and pitch [8, 18, 29, 31].

Applications like TRACI (Teacher Ranging Across the Computer Interface) talks, Caroline in the City, CNN Interactive English, The Syracuse English Comprehensive Learning Series, Tell Me More Pro, and Encarta Interactive English Learning, use role-play dependent activities and task-based dialogues to improve user's speaking capabilities [10].

However, all these applications do not provide feedback on intonations, which are certain patterns of pitch modulation used to represent user intentionality through the use of relevant context. We want to design a system that provides context-based information of these intonations, gaps, intensity and extended syllables through the written system of English.

2.3 Visual Representation of Pitch

Human perception is the most sensitive to changes in pitch than to changes in other para-linguistic speech parameters ([28], p. 207). Here, we represent the various ways in which pitch has been represented historically.

Intuitive data representation techniques for representing para-linguistic speech data

There has been no consensus yet for the development of a representation system of pitch([21], despite there being a universal agreement for the representation human speech sounds as presented by International Phonetic Alphabet (IPA). An early representation system presented by James Rush (Figure 1) ([26]), portrayed the musical pitch scale for representing pitch along with the transcription of spoken text written over it to represent the position of pitch. Lieberman (Figure 2) ([22], 1967) used the 10th harmonics of audio pitch on narrowband spectrograms by highlighting their fundamental frequency. Crystal (Figure 3) ([13]) used curved lines as presented by icons to encode the meanings of intonations. He further utilized "large and small dots, capitalization, arrows, dashes, and two kinds of accent marks (grave and acute), along with curved lines placed in a vertical space" to represent the various different intonations ([12]). All these systems used two layers for representation of words and pitch, which could be problematic for reading.

Fig. 1. Notation given by James Rush.

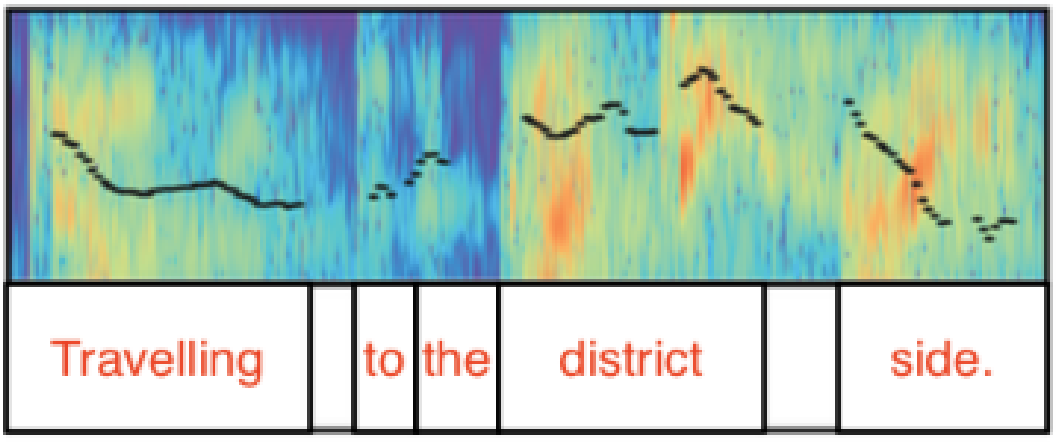


Fig. 2. Notation given by Lieberman while tracing the 10th harmonic of written speech.

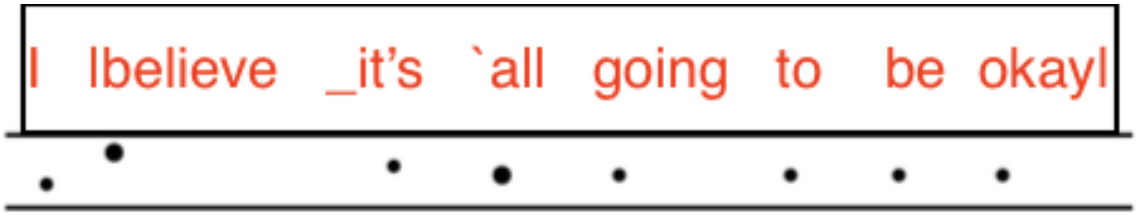


Fig. 3. Notation given by Crystal.

Bolinger (Figure 4) ([6]) used sentences with stressed syllables and words written up or down relative to each other to represent pitch curves along with their displacements. Ladefoged (Figure 5) ([21]) mixed the pitch curves obtained through pitch extraction algorithms along with linguistic speech units written over the curve. These units helped in tracking intonations embedded over the words. Although Bolinger's system used a single layer of visual representation to represent speech meaning, it was not a straight line system of representation and hence, not suitable for reading large paragraphs of speech. On similar grounds, Ladefoged's system, though continuous and more legible than the previous notation system, decreased the reading flow by not being in a straight line.

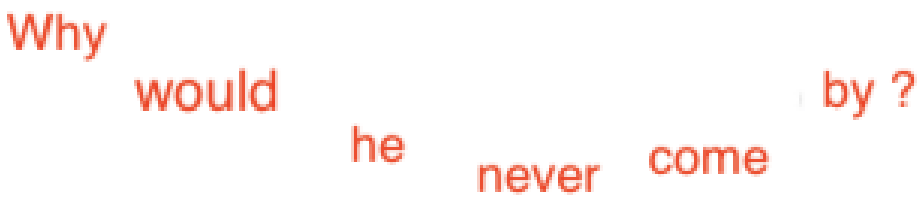


Fig. 4. Notation given by Bolinger.

Fig. 5. Notation given by Ladefoged.

2.4 Context-based visualization

Verbose (figure 6) [9] is the only app that teaches context-based intonations to non-native speakers through a visual representation system, teaching people when, how and why a particular intonation is spoken in English. However, Verbose only provides a visual representation of intonations, and not the other paralinguistic parameters of speech. We

design a system that represents intonations, intensity, gaps and extended syllables on the orthographical text-based system of English writing.

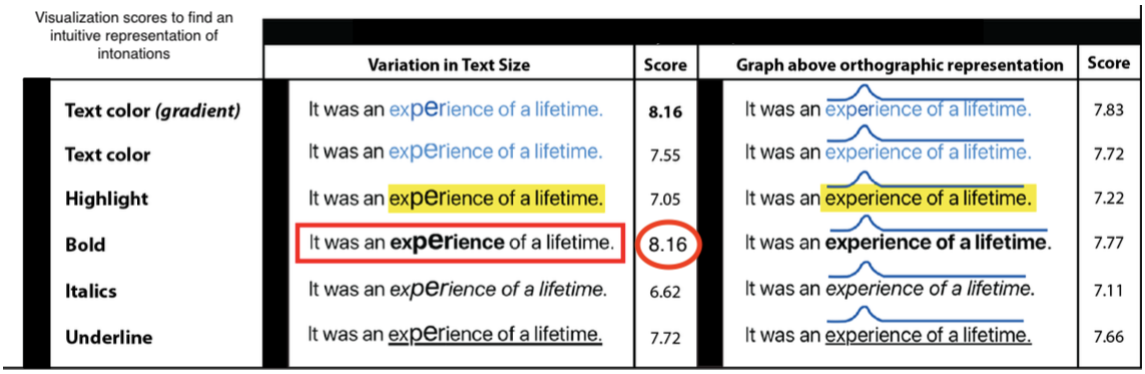


Fig. 6. Exhaustive user study performed for Verbose app that uses all options for highlighting in a standard a text editor, along with graphs and change in text size.

3 DESIGN PROPOSALS

Pitch(fundamental frequency), Intensity(amplitude), Gap(empty durations) and Extension(elongated syllables) are some of the main ways in which humans embed stress to convey intentionality in speech. These ways of embedding stress are directly in relation to the context in which a speaker speaks. The visualizations represented here display some ways to represent this context-based stress information in speech.

Following are examples of intonations, intensity, gaps and extended syllables, along with their naive sketch visualizations.

Question Intonation (QI)

Why are you standing in the corner of the room?

Accent Intonation (AI)

Is there any hotel nearby?

There is an excellent hotel near downtown.

Continuation Intonation (CI)

I am going to attend a seminar, then grab some food and head

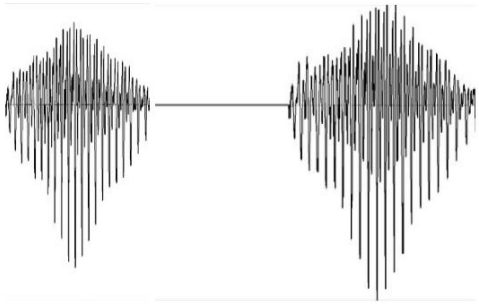
back home later.

Sentence Intonation (SI)

We will return to classes later in the semester.

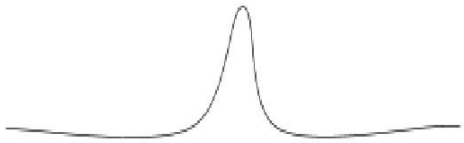
Fig. 7. The image shows four naive visual representations of intonations.

Gap



I am thinking……..of going on a trip.

Intensity



I am telling you again, I can't come.

Extended syllables

I reeaaallly have to be at home and study.

Fig. 8. The image shows three naive visual representations of gaps, intensity and extended syllables.

A summary of the paralinguistic parameters, their intentionality and the context used in their representation -

Name	Intentionality	Context in sentence
Pitch1 - Question Intonation	Open-ended intentionality	End of sentence
Pitch2 - Sentence Intonation	Close-ended intentionality	End of sentence
Pitch3 - Accent Intonation	Object-word focusing intentionality	Object-word of context
Pitch4 - Continuation Intonation	Object-word focusing, open-ended intentionality	Words in sentence denoting continui
Intensity	Object word focusing intentionality	Object-word of context
Gap	Silence intending prominence of preceding/succeeding word	Silences
Extension	Object-word focusing intentionality	Object-word of context

Table 1. Table displaying the name, intentionality and context used in sentences for visualization.

Following are the design features of visualizations that we have used -

- Relative change in size of letters is used to represent pitch change. This is done using font sizes 30 px, 50 px, 70 px and 90 px. 60 px and 80 px font sizes are used to adjust sizes of letters like a, m (etc.) which only have a lower base, as compared to letters like t, b which have an upper base as well.
- Relative change in color is used to represent intensity change. This is done using the fonts black ("000000"), dark orange ("ff8c00") and red ("ff0000).
- Dots are used to represent Relative silences.
- Letter repetition is used to represent syllable extension.

We now give examples of the visualizations for the above mentioned paralinguistic parameters. The following visualizations are implemented using fonts on HTML with the following template [2].

Fig. 9. Visual representation of pitch1.

Fig. 10. Visual representation of pitch2.

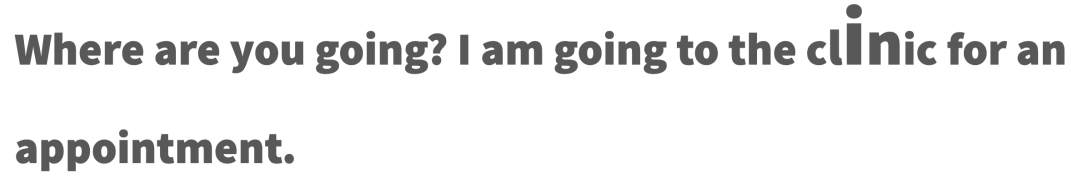


Fig. 11. Visual representation of pitch3.

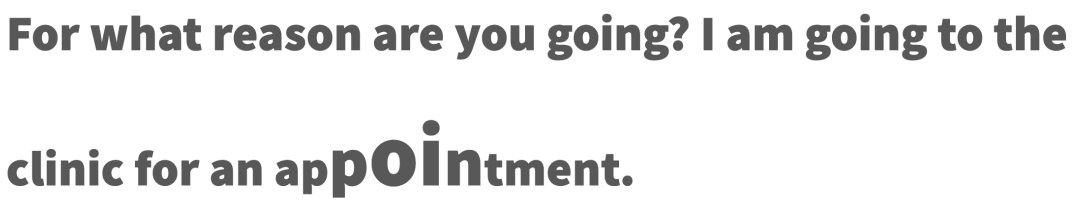


Fig. 12. Visual representation of pitch4, context1.

Fig. 13. Visual representation of pitch4, context2.

Fig. 14. Visual representation of pitch4, context3.

Fig. 15. Visual representation of intensity.

Fig. 16. isual representation of gaps.

Fig. 17. Visual representation of extended syllables.

Fig. 18. Visual representation combing intonation, intensity, gaps and extended syllables.

4 DATA

All data was recorded from only one speaker on Samsung Galaxy J6 with the distance of speaker to recording mic being 50 cm.

5 DISCUSSION

- Context can be modeled in the following four ways. “Firstly, phonetic/linguistic context, that is, what a speaker produced before or after the unit we want to analyse. Secondly, multimodal context, that is, which body posture, gestures, and facial gestures the speaker produces concomitantly, synchronously or before and after the unit we want to analyse. Thirdly, immediate situational context, that is, the overall setting (communication partners, type of communication, room characteristics, etc). Fourth, general context in time and space, that is, generally speaking, in which historic/geographic situation the communication partners are.” We have modeled the first case of context in our work. In future work, context can be changed to new and interesting cases in future cases of visualization work. Context can be narrow, concentrating on the speaker herself and her personal situation and what she has experienced in recent hours, it can be wide, including macro-sociological and political constellations, and it can simply be be narrowed down to membership of class, etc.
- Relative change can be replaced by absolute quantitative change.
- Normal written language follows Grice’s 4 cooperative maxims to communicate - "First, the maxim of quantity, where one tries to be as informative as one possibly can, and gives as much information as is needed, and no more. Second, the maxim of quality, where one tries to be truthful, and does not give information that is false or that is not supported by evidence. Third, the maxim of relation, where one tries to be relevant, and says things that are pertinent to the discussion. Fourth, the maxim of manner, when one tries to be as clear, as brief, and as orderly as one can in what one says, and where one avoids obscurity and ambiguity." If we change the maxim of manner, we get figures of speech like sarcasm and irony. It will be to interesting to visualize these figures of speech.
- We initially used Google Speech to Text for doing word transcriptions, and Parselmouth to extract pitch, intensity, gap and spectrogram information to make visualizations for any random recorded audio file. However, Google Speech to Text only provides word level information of time stamps, and hence we could not couple the letter level information with pitch to represent the visualizations. Future work could explore novel methods of visualizing randomly entered input data.

ACKNOWLEDGMENTS

REFERENCES

[1] 2019. *British English Pronunciation.* https://englishpronunciationroadmap.com
[2] 2021. *HTML Template.* Retrieved December 05, 2021 from https://html5up.net/future-imperfect
[3] 2021. *Migration WEF.* Retrieved Dec 05, 2021 from https://www.weforum.org/agenda/2020/01/iom-global-migration-report-international-migrants-2020/
[4] Mohammad Rafayet Ali, Kimberly Van Orden, Kimberly Parkhurst, Shuyang Liu, Viet-Duy Nguyen, Paul Duberstein, and M. Ehsan Hoque. 2018. Aging and Engaging: A Social Conversational Skills Training Program for Older Adults. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - IUI ’18.* ACM Press, Tokyo, Japan, 55–66. https://doi.org/10.1145/3172944.3172958
[5] Peter L Auer, Peter Auer, Elizabeth Couper-Kuhlen, Frank Müller, et al. 1999. *Language in time: The rhythm and tempo of spoken interaction.* Oxford University Press on Demand.
[6] Dwight Bolinger and Dwight Le Merton Bolinger. 1986. *Intonation and its parts: Melody in spoken English.* Stanford University Press.

[7] Mark Bubel, Ruiwen Jiang, Christine H Lee, Wen Shi, and Audrey Tse. 2016. AwareMe: addressing fear of public speech through awareness. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 68–73.

[8] Mark Bubel, Ruiwen Jiang, Christine H. Lee, Wen Shi, and Audrey Tse. 2016. AwareMe: Addressing Fear of Public Speech through Awareness. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16*. ACM Press, Santa Clara, California, USA, 68–73. https://doi.org/10.1145/2851581.2890633

[9] Akash Chaudhary, Manshul Belani, Naman Maheshwari, and Aman Parnami. 2021. Verbose: Designing a Context-based Educational System for Improving Communicative Expressions. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. 1–13.

[10] Hao-Jan H Chen. 2001. Evaluating five speech recognition programs for ESL learners. In *ITMELT 2001 Conference, Hong Kong. http://elc. polyu. edu. hk/conference/papers2001/chen. htm.*

[11] Alan Cruttenden. 1997. *Intonation* (2 ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139166973

[12] David Crystal. 1975. *The English tone of voice: essays in intonation, prosody and paralanguage.* Hodder Arnold.

[13] David Crystal. 2003. *The Cambridge Encyclopedia of the English Language.* Cambridge University Press.

[14] F Cummings and R Port. 1998. Rhythmic constraints on stress timing in English. *Journal of Phonetics* 26 (1998), 145–171.

[15] Ionut Damian, Chiew Seng Sean Tan, Tobias Baur, Johannes Schöning, Kris Luyten, and Elisabeth André. 2015. Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques. In *Proceedings of the 33rd annual ACM conference on Human factors in computing systems*. ACM, 565–574.

[16] Tanusree Das, Latika Singh, and Nandini C Singh. 2007. Rhythmic structure of Hindi and English: new insights from a computational analysis. *Progress in brain research* 168 (2007), 207–272.

[17] Rebecca M Dauer. 1983. Stress-timing and syllable-timing reanalyzed. *Journal of phonetics* (1983).

[18] Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. 2013. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 697–706.

[19] Jody Kreiman and Diana Sidtis. 2011. *Foundations of voice studies: An interdisciplinary approach to voice production and perception.* John Wiley & Sons.

[20] Kazutaka Kurihara, Masataka Goto, Jun Ogata, Yosuke Matsusaka, and Takeo Igarashi. 2007. Presentation sensei: a presentation training system using speech and image processing. In *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 358–365.

[21] Peter Ladefoged and Keith Johnson. 2006. A Course in Phonetics (5th). *Thomson Wadsworth* (2006).

[22] Philip Lieberman. 1967. Intonation, perception, and language. *MIT Research Monograph* (1967).

[23] Yi-Jing Lin and Chialin Chang. 2017. MyET and English Pedagogy. (2017).

[24] Dania Murad, Riwu Wang, Douglas Turnbull, and Ye Wang. 2018. SLIONS: A Karaoke Application to Enhance Foreign Language Learning. In *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 1679–1687.

[25] Peter Roach. 1982. On the distinction between 'stress-timed'and 'syllable-timed'languages. *Linguistic controversies* 73 (1982), 79.

[26] James Rush. 1833. The philosophy of the human voice. (1833).

[27] SR Savithri, M Jayaram, D Kedarnath, and S Goswami. 2007. Speech rhythm in Indo Aryan and Dravidian languages. In *Proceedings of the International Symposium on Frontiers of Research on speech and music*. 170–174.

[28] Bjorn Schuller and Anton Batliner. 2013. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing* (1st ed.). Wiley Publishing.

[29] Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2015. Automated social skills trainer. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 17–27.

[30] M Iftekhar Tanveer, Emy Lin, and Mohammed Ehsan Hoque. 2015. Rhema: A real-time in-situ intelligent interface to help people with public speaking. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 286–295.

[31] Ha Trinh, Reza Asadi, Darren Edge, and T Bickmore. 2017. RoboCOP: A Robotic Coach for Oral Presentations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 27.

[32] Xingbo Wang, Haipeng Zeng, Yong Wang, Aoyu Wu, Zhida Sun, Xiaojuan Ma, and Huamin Qu. 2020. VoiceCoach: Interactive Evidence-based Training for Voice Modulation Skills in Public Speaking. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[33] Ru Zhao, Vivian Li, Hugo Barbosa, Gourab Ghoshal, and Mohammed Ehsan Hoque. 2017. Semi-Automated 8 Collaborative Online Training Module for Improving Communication Skills. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 32.