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 expressing their intentions, especially those who have syllable-timed languages as their native one. We design a system for non-native English speakers to visualize intonations, loudness, word count and gaps in speech in a visual format. The resulting 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 (October 2021), 3 pages. https://doi.org/10.1145/3447526.3472057

1 INTRODUCTION

English has become common language the world over for communicating with people. It has become an essential skill for non-native speakers to find job opportunities and speak effectively among their peers and in general among larger audiences.

English, however, is a stress-timed language([22], p. 72-79; [8], p. 10; [14], p. 51-62), which means that people tend to speak English with certain para-linguistic patterns that are essential in parsing the knowledge structure for finding the intentions behind people's words([3, 11]; [24], p. 81; [16], p. 283). 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 a natural way [13]([3, 11];[24], p. 81; [16], p. 283).

Currently, people just try and listen to audio and understand the various subtleties in speech through recognition by ear[1]. However, their perception of these subtleties can increase if they also visualize these audio cues visually through added layers on the English orthographical system of writing.

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.

Therefore, we design a system where we extract these para-linguistic layers of sound data and embed them over written speech for better understanding of English language among non-native English speakers.

2 RELATED WORK

2.1 Imitation-based systems

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

CALL (Computer Assisted Language Learning) [7] uses imitation-based teaching techniques for users to develop listening and speaking skills. My English Tutor (MyET) [20] also uses imitation-based teaching techniques to teach various accents of different teachers and thereby improve their pronunciation. SLION [21] uses imitation based teaching combined with automatic speech recognition for a karaoke app. However, our system is intended toward combining an imitation-based system along with a visual representation of para-linguistics in textual writing.

2.2 Systems teaching through para-linguistics

Here we explore the various works which have represented para-linguistic parameters of speech as systems to improve speech communications.

Applications like Rhema [26], Logue [12], AwareMe [5], Aging and Engaging [2], ROCSpeak [29], RoboCOP [27], VoiceCoach [28], MACH [15], and Automated Social Skills Trainer [25] provide feedback on word count [6, 12, 15, 17, 25–28], loudness or intensity [2, 6, 26, 28, 29], gaps or pauses [15, 25, 28], and pitch [6, 15, 25, 27]. However, all these applications do not provide feedbacks on the above mentioned parameters through the system of visual representation of words. We want to design a system that provides all these information 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 ([24], p. 207). Here, we represent the various ways in which pitch has been represented historically.

There has been no consensus yet for the development of a representation system of pitch([18], 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 ([23]), 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 ([19], 1967) used the 10th harmonics of audio pitch on narrowband spectrograms by highlighting their fundamental frequency. Crystal ([10]) 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 ([9]). All these systems used two layers for representation of words and pitch, which could be problematic for reading.

Bolinger ([4]) used sentences with stressed syllables and words written up or down relative to each other to represent pitch curves along with their displacements. Ladefoged ([18]) 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.

ACKNOWLEDGMENTS

REFERENCES

[1] 2019. *British English Pronunciation.* https://englishpronunciationroadmap.com

[2] 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

[3] 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.

[4] Dwight Bolinger and Dwight Le Merton Bolinger. 1986. *Intonation and its parts: Melody in spoken English.* Stanford University Press.

[5] 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.

[6] 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

[7] 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.*

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

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

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

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

[12] 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.

[13] 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.

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

[15] 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.

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

[17] 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.

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

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

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

[21] 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.

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

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

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.