An AI-based Feedback Visualisation System for Speech Training

Adam T. Wynn[0000−0002−1631−2151], Jingyun Wang[0000−0001−9325−1789], Kaoru Umezawa[0000−0001−7686−097X], and Alexandra I. Cristea[0000−0002−1454−8822]

Durham University, Durham, DH1 3LE, United Kingdom
jingyun.wang@durham.ac.uk

Abstract. This paper proposes providing automatic feedback to support public speech training. For the first time, speech feedback is provided on a visual dashboard including not only the transcription and pitch information, but also emotion information. A method is proposed to perform emotion classification using state-of-the-art convolutional neural networks (CNNs). Moreover, this approach can be used for speech analysis purposes. A case study exploring pitch in Japanese speech is presented in this paper.

Keywords: CNN · automatic visualisation feedback · second language speech training · emotion recognition · speech prosody

1 Introduction

Timely feedback is important for language learning as it enables the learner to practice at their own pace [7]. Speech training applications have been used to help second language (L2) speakers identify ways to improve their speech without the requirement for manual feedback. Some systems provide pitch feedback using visualisation dashboards [13] while others provide automatic speech modification [4]. These feedback mechanisms work well with simple phrases, but don't scale well to longer speeches. Moreover, few studies have focused on supporting public speaking training. Therefore, this research is intended to support L2 English and Japanese learners in public speaking, such as speech contests. Our research question is: *Compared to prior research simply providing transcription or pitch changes as feedback, can a combination of transcription, pitch and emotional changes as feedback better support speech training?*

Determining the relationship between the pitch range of speakers from different L1 backgrounds is one research focus. For instance, in Japanese, pitch serves as the main cue to signal lexical and phrasal distinctions. Passoni, et al. [12] found that Japanese-English bilinguals had a lower mean pitch in Japanese than in English and female speakers displayed more pitch variation for different formality settings and that lower mean pitch may be due to nervousness. A method for computing the pitch of English speech is proposed by Kurniawan, et al. [8], and their experiment results suggest that an increase of pitch could be a sign of nervousness.

Emotion detection is another research direction. Using Convolutional neural networks (CNNs), Franti, et al. [2] classified speech into 6 emotional states. They identified that someone who was speaking faster with a wider pitch range was more likely to be experiencing emotions of fear, anger or joy. Kurniawan, et al. [8] captured Mel Frequency Cepstral Coefficients (MFCCs) as features from the audio signal to classify speech. MFCCs approximate human audio perception more closely, to achieve an accuracy of 92.4% using Support Vector Machines.

The main purpose of this paper is to propose an AI-based speech feedback system, which gives immediate feedback to the learner, via a visualisation dashboard. This approach is achieved by outputting the state and level of emotion identified by CNNs in each sentence, and the user can view how their pitch changes throughout the speech to detect how this might effect the emotion conveyed, along with a transcription. Moreover, multiple audio files can be uploaded by users for further comparisons and analysis. This function is illustrated by a case study exploring pitch in Japanese speech.

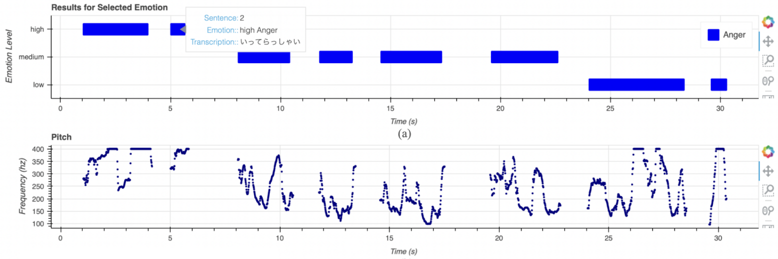
2 A Visualisation Speech Feedback System

In this research, an AI-based visualisation system which not only provides feedback for individual speech training, but also enables audio analysis, was designed and implemented. A CNN was proposed to recognise the level of emotion (low, medium or high) in each sentence which consists of 1-dimensional convolutional layers and was programmed using the Keras library [5] and TensorFlow. 40 MFCC features were used as an input, which were extracted from the data using the Librosa package [11]. 2686 speech samples from the RAVDESS [10] and CREMA-D [6] datasets were used for model training where the accuracy depends on the emotion (Anger: 82.4%, Disgust: 71.7% Fear: 79.7%, Happy: 72.5%, Sad: 70.3%). The CREPE Pitch Tracker [9] was used to identify pitch based on the fundamental frequency (f0). Readings above 400 hz or below 50 hz were removed, as they were likely erroneous measurements.

Prior to uploading audio recordings, learners need to choose their language, gender, and one emotion out of anger, disgust, fear, happiness, and sadness to focus on. The feedback provided by the system is presented visually (Fig. 1), using the Bokeh visualisation library [3]. Fig. 1(a) provides information about the emotion tracked, including three levels of intensity. The user can see how the intensity of their chosen emotion changes throughout the speech for each sentence. Fig. 1(b) shows how the pitch changes throughout the speech, which could be used to infer the relationship with emotion.

3 A Case Study – Exploring Pitch in Japanese Speech

It is proven that the mean pitch of female speakers (between 160–300 Hz) is higher than that of male speakers (between 60–180 Hz). However, few studied the pitch difference between native and non-native speakers, and also the interaction effect with gender. To study the effect of gender and whether the speaker is a



(a)

(b)

Fig. 1. Learner dashboard with visual feedback.

native Japanese or not on pitch (considering mean pitch and pitch range as two features), audio files (average duration 8 minutes) which include speeches by 2 native and 3 non-native female speakers, and 4 native and 5 non-native male speakers, consisting of 795 sentences (197 native female; 320 non-native female; 134 native male; and 144 non-native male) were uploaded to the system for analysis. The native speaker audio was collected from Toastmasters Japan [1] and non-native speaker audio was collected from Japanese speech contests.

Based on the mean pitch and pitch range of each sentence determined by the system, a two-way MANOVA was conducted. The results indicate a significant interaction effect ($F(1, 790) = 70.66$) between gender and whether the speaker is native or not. The main effect of gender on pitch is significant ($F(1, 790) = 272.80$). Sentences by female speakers (Mean Pitch: Mean = 245.57 Hz, S.D. = 1.90; Pitch Range: Mean = 236.40Hz, S.D. = 3.08) have a significantly higher mean pitch ($F(1, 791) = 516.12$) and wider pitch range ($F(1, 791) = 49.85$) compared to those by male speakers (Mean Pitch: Mean = 174.02 Hz, S.D. = 2.51, Pitch Range: Mean = 200.27 Hz, S.D = 4.09). Also, the main effect of whether the speaker is native or not is significant ($F(1, 790) = 30.39$). Compared to sentences by non-natives (Mean = 200.94 Hz, S.D. = 2.10), those by natives (Mean = 218.65 Hz, S.D. = 2.35) have a significantly higher mean pitch ($F(1, 791) = 31.63$) and their speech (Mean = 233.13 Hz, S.D. = 2.80) has a wider pitch range ($F(1, 791) = 33.46$) in contrast to non-native speech (Mean = 203.54 Hz, S.D. = 3.81). Furthermore, the individual univariate test results (Table 1) show a significant difference between native and non-native females ($F(1, 791) = 24.044$), and between native and non-native male speakers ($F(1, 791) = 115.45$). For pitch range, there is only a significant difference ($F(1, 791) = 35.43$) between native and non-native males, and no significant difference ($F(1, 791) = 2.94$) between native and non-native female speakers.

Table 1. Individual univariate test results.

Feature	**Gender**	**Native (Hz)**	**Non-native (Hz)**	**F(1,191)**
Mean Pitch	Female	Mean = 236.37; S.D. = 2.98	Mean = 254.87; S.D. = 2.34	24.04 ($p < 0.05$)
	Male	Mean = 201.40; S.D. = 3.62	Mean = 147.01; S.D. = 3.49	115.45 ($p < 0.05$)
Pitch Range	Female	Mean = 241.68; S.D. = 4.85	Mean = 231.11; S.D. = 3.81	2.94 ($p > 0.05$)
	Male	Mean = 224.58; S.D. = 5.88	Mean = 175.96; S.D. = 5.67	35.43 ($p < 0.05$)

4 Discussion and Future Work

Despite a small number of speakers, the analysis of 795 sentences shows that non-natives have a significantly narrower pitch range, which may be due to nervousness. From a public speech training perspective, this suggests that non-native speakers should try to widen their pitch range in order to be more similar to natives. Detailed suggestions regarding their pitch and emotion could potentially help them adjust their pitch range. In summary, this case study demonstrates that our system can easily transform multiple audio files into quantitative data, which can be used in further statistical analysis for any research purpose.

In the future, speech data of more speakers will be studied to confirm this finding. Also, more detailed feedback will be provided to improve their speaking skills. In terms of emotion, we plan to train another model using Japanese emotional speech data, and design more functions to support speech training.

References

1. Toastmasters japan. https://district76.org/en/ (2021), [Accessed 6-Feb-2022]
2. Alu, D., et al.: Voice based emotion recognition with convolutional neural networks for companion robots. Romanian Journal of Information Science and Technology **20**(3), 222–240 (2018)
3. Bokeh: Bokeh. https://bokeh.org (2021), [Accessed 17-Jan-2022]
4. Bonneau, A., Colotte, V.: Automatic feedback for l2 prosody learning. Ivo Ipsic. Speech and Language Technologies, Intech pp. 55–70 (2011). https://doi.org/10.5772/20105
5. Chollet, F., et al.: Keras (2015), https://github.com/fchollet/keras, [Accessed 17-Jan-2022]
6. Cooper, D.: CREMA-D. https://github.com/CheyneyComputerScience/CREMA-D (2021), [Accessed 17-Jan-2022]
7. Golonka, E., et al.: Technologies for foreign language learning: A review of technology types and their effectiveness. COMPUTER ASSISTED LANGUAGE LEARNING **27**, 70–105 (2014). https://doi.org/10.1080/09588221.2012.700315
8. Hindra Kurniawan, Alexandr V. Maslov, M.P.: Stress detection from speech and galvanic skin response signals. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems pp. 209–214 (2013). https://doi.org/10.1109/CBMS.2013.6627790
9. Kim, J.W., et al.: Crepe: A convolutional representation for pitch estimation (2018)
10. Livingstone, S.R., Russo, F.A.: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Apr 2018). https://doi.org/10.5281/zenodo.1188976
11. McFee, B., et al.: librosa: 0.8.1rc2 (May 2021). https://doi.org/10.5281/zenodo.4792298
12. Passoni, E., et al.: Bilingualism, pitch range and social factors: preliminary results from sequential japanese-english bilinguals. Proc. 9th International Conference on Speech Prosody 2018 pp. 384–338 (2018). https://doi.org/10.21437/SpeechProsody.2018-78
13. Sztah, D., et al.: Computer based speech prosody teaching system. Computer Speech and Language **50**, 126–140 (2018). https://doi.org/10.1016/j.csl.2017.12.010