BETTER: An automatic feedBack systEm for supporTing emoTional spEech tRaining

Adam Wynn[0000−0002−1631−2151] and Jingyun Wang[0000−0001−9325−1789]

Durham University, Durham, DH1 3LE, United Kingdom
adam.t.wynn@durham.ac.uk

Abstract. Feedback is a crucial process in education because it helps learners identify their weaknesses whilst motivating them to continue to learn. Existing systems often only provide a score or rating with basic explanations. Although some systems provide detailed feedback, they require manual input from teachers. This paper proposes a real-time feedback visualisation system (called BETTER) for supporting emotional speech training which uses a visual dashboard to provide the learner with immediate feedback in the form of written, audio, and visual feedback. The AI-based feedback system utilises pitch tracking, transcriptions, and audio modifications in addition to one-dimensional convolutional neural networks (CNNs) to categorise speech into emotional states. A preliminary experiment was conducted involving a speech expert and 8 non-native speakers to assess their cognitive load, technology acceptance, and satisfaction while using the system.

Keywords: Automatic visualisation feedback · Emotion recognition · Speech prosody · Emotional speech training · System Evaluation

1 Introduction

In education, feedback is an important process as it enables students to identify misconceptions and reflect on their learning progress whilst reducing instructor effort [3]. The primary methods and techniques used to generate automatic feedback include comparisons with the desired solution, interactive dashboards and visualisations, and text generated by natural language processing [3]. Due to advances in machine learning and audio feature extraction techniques, new feedback mechanisms involving emotion detection, virtual worlds feedback, game-mediated feedback, and oral feedback have been proposed [2].

It can be challenging to successfully convey emotion in a public speaking environment. Therefore, this paper proposes providing automatic feedback for emotional speech training and seeks to address the following research questions: *1) Does providing automatic feedback help learners improve their speech?*; *2) What are the learner perception (including cognitive load, technology acceptance, and satisfaction for the system) differences between training one and multiple sentences?* ; *3) What are the learner perception differences between real-time feedback addressing a specific discrete emotion and addressing positive and negative emotions?*

To explore the above questions, BETTER, an AI-based feedBack systEm for supporTing emoTional spEech tRaining, which gives real-time automatic feedback via a visualisation dashboard is proposed. Our main contributions are supporting speech training by providing detailed feedback on a visual dashboard including not only the transcription and pitch information but also emotional information, and expanding the system based on the feedback of participants so that for the first time, visual strategies related to pitch changes and modified audio are provided automatically to assist the learners in conveying their speech with more accurate emotions.

2 Related Work

Although there are some systems that include techniques for acquiring features from audio signals, they usually only focus on single features at a time such as pronunciation, pitch, and emotion and most of these systems do not have feedback for the learner as their focus and focus more on how to abstract the features accurately. For example, Sztaho, et al. [14] aimed to improve the intonation and rhythm of children with hearing loss using a visual feedback system displaying the learner's pitch and the correct pitch in an easy-to-understand way using a line graph. Aucouturier, et al. [1] and Rachman, et al. [13] provide audio feedback by modifying speech signals in real-time by applying audio effects including pitch shifting, inflection and vibrato, and by using high-pass and low-pass filters. However, these effects can made speech sound unnatural and in [1], not all participants detected any manipulation. Zhou, et al. [16] use sequence-to-sequence encoders with attention to reconstruct speech which conveys a combination of emotions, however, the challenges of unnatural emotional expressions remain.

Detecting emotion involving classifying speech into emotional states is one research direction. Using Convolutional Neural Networks (CNNs), an accuracy of 71% was achieved by Franti, et al. and it was identified that someone who was speaking faster which a higher and wider pitch range were more likely to be experiencing emotions of fear, anger of joy, whereas emotion such as sadness generate slow and lower pitched speech. The framework by Issa, et al. [8] using CNNs obtains 71.61% for the RAVDESS dataset [11] with 8 classes. Using a reinforcement learning network to keep track of emotional changes in every utterance throughout the speech is another research direction. Huang et al. [7] propose a method where a sliding window would be used to collect emotional information using Mel-frequency cepstral coefficients (MFCCs), which approximate the human auditory system's response. Using this information, an agent would decide which emotion to classify the sentence as.

3 Methodology

3.1 BETTER Version 1.0

Both visual and basic written feedback is provided by BETTER 1.0 [15] built using the Panel [5] library. Two modes are provided: Mode 1) a positive and

negative emotions mode and Mode 2) a 5 discrete emotions mode. In mode 1, the learner is presented with a graph which shows their audio emotion and content emotion on a 5-point Likert scale (very positive, positive, neutral, negative, very negative), and their pitch. In mode 2, as shown in figure 1, the user is able to see how the intensity of the focused emotion changes throughout the speech, how long each sentence is, and how long the pauses between each sentence are. BETTER shows which emotion was most prevalent out of fear, happy, anger, sad, and neutral along with the level of the emotion detected indicated by the colour of the bar. The "Content Emotion" is also provided in green which is the emotion of the text produced by the transcript. The learner can look at the emotion identified for the content of their speech, and try to match that emotion when speaking. Finally, BETTER shows how the pitch changes throughout the speech which can be used to infer how this might affect the emotion detected. The pitch is not displayed when there is a pause between sentences to make it clearer to the user how the pitch in each individual sentence fluctuates.

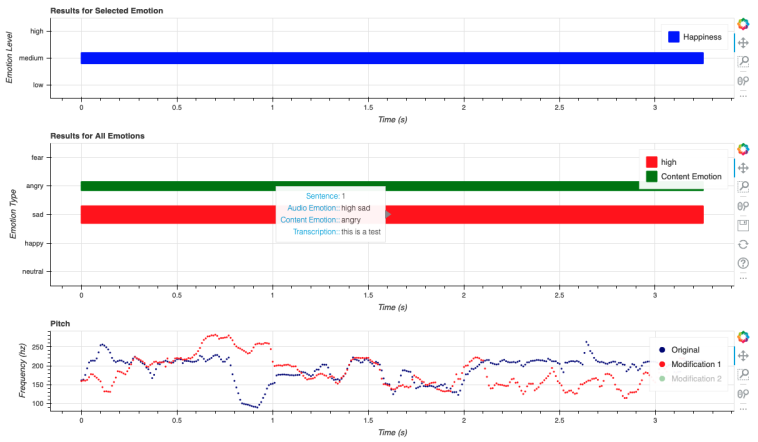

Fig. 1. Part of the system output when in Discrete Emotions Mode showing how the user's emotion and pitch changes throughout their speech.

1D CNNs are used to classify each sentence and were programmed using the Keras library [4] and TensorFlow. 40 MFCC features are extracted from the data using the Librosa package [12] and used as input into the models. To classify emotions on a 5-point Likert scale, a model was trained using batches of size 64 over 100 epochs. The model architecture consists of 4 1-dimensional convolutional layers, a batch normalisation layer, dropout layers and 3 dense layers. The model was trained using the RAVDESS [11], TESS [6], and SAVEE [9] datasets. These datasets were manually relabelled by a native speaker using the same 5-point likert scale as above to avoid inconsistencies between datasets. This model achieved a validation accuracy of 89.82% using 5-fold validation. To classify discrete emotions (anger, fear, sadness, happiness, and neutral), another model was trained with the same architecture as above, using data from the

RAVDESS [11], TESS [6], and SAVEE [9] datasets which include a total of 3876 audio recordings of English speakers. This model achieved a validation accuracy of 90.05% using 5-fold validation.

3.2 Preliminary Experiment

A preliminary experiment involving a female expert in speech training and eight male Chinese students who study in a university in the UK, was conducted using BETTER 1.0. During the experiment, participants were asked to use Mode 1 followed by Mode 2. Once the audio was analysed, the participant was presented with visual feedback showing how their emotion, pronunciation, and pitch changes (using Parselmouth [10]) throughout their speech.

After the speech training activity, the students were interviewed and completed a questionnaire consisting of 31 questions related to mental load, technology acceptance, and satisfaction. 62.5% and 12.5% of participants agreed and strongly agreed respectively that BETTER's feedback on their emotions was easy to understand, and 87.5% of participants found the feedback given on their emotions effective and useful. Participants had mixed opinions on whether viewing positive and negative emotions or discrete emotions was easier to understand and more useful, with 37.5% and 25% preferring each, respectively. Overall, 75% of participants found the combination of content and audio emotion helpful for improving their speech. Overall, 87.5% of participants found BETTER easy to use and navigate, and 62.5% felt confident using it to improve their speech. However, only 50% agreed that the overall feedback provided by the system was effective and useful, and 75% expressed a desire to use the system frequently. For the questions about cognitive load, participants answered on a scale from 1-7 where 1 is "very little" and 7 is "a lot". Participants rated the effort required to understand the learning activity with a mean of 3.5 (SD = 1.41). In general, participants felt more distracted when using mode 2 (Mean = 3.625, SD = 1.30) compared to mode 1 (Mean = 3.5, SD = 1.60), and more stressed or irritated when using mode 2 (Mean = 4.125, SD = 1.25) compared to mode 1 (Mean = 3, SD = 1.60).

Six of the students found the transcription and audio playback helpful for revising and fixing pronunciation mistakes. However, participants suggested that the feedback could be more detailed, such as providing an explanation of how to match the detected emotion or giving feedback on energy, speed, and other speech aspects. The expert suggested providing written feedback and a comparison with a native speaker's recorded sample sentence for easier interpretation. Participants found it difficult how to interpret the curve of the sentence's pitch and suggested an explanation of how pitch impacts the conveyed emotion.

3.3 BETTER Version 2.0

Based on the results of the preliminary study, BETTER was extended to version 2.0 and enhanced with more detailed feedback by adding written explanations of how to match the selected emotion along with a simple diagram. The explanation

consisted of how the user can adapt their pitch throughout the sentence and their rate of speech, as shown in the left part of figure 2.

Moreover, an audio playback function was added, as shown in the right part of figure 2, to provide the user with 2 modified versions of their speech, together with their original recording. The audio clips were modified by applying effects including speed changes, pitch shifting, inflection, vibrato, and filtering [1][13], and by adjusting parameters to create subtle and exaggerated versions.

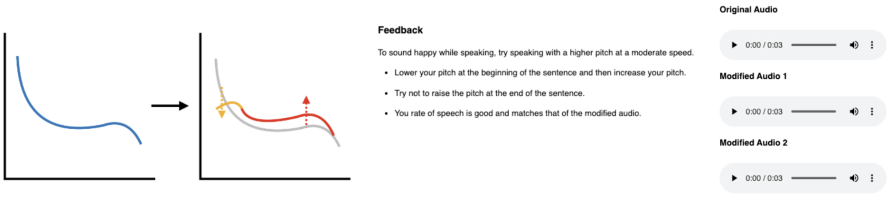

Fig. 2. Part of the system output when in Discrete Emotions Mode showing written feedback and audio modifications

A further experiment is currently being conducted to evaluate the effectiveness of BETTER 2.0. Until now, 31 students aged 18-35 consisting of 18 male and 13 female speakers have participated in a 1 hour long experiment and were required to interact with BETTER 2.0 during the experiment. It is found that most of the participants agreed that automatic feedback is effective in improving the speech of learners and thought that training one sentence at a time is slightly easier and more useful. The participants acknowledge that both modes of BETTER 2.0 are easy to use and useful for speech training and the results suggest that mode 2 is more useful despite mode 1 being easier to use. The experiment will involve more female participants to ensure a more balanced representation. The final results will be analysed in the future.

4 Conclusion

The main contribution of this paper was to present a system that has been implemented from experimental results to support emotional speech training. For the first time, speech training is supported by detailed feedback provided on a visual dashboard including not only the transcription and pitch information but also emotional information with visual strategy related to pitch changes and modified audio, to support the learners in adjusting their speech.

With reference to future improvement, we plan to provide more specific personalised feedback tailored towards each user considering differences between different languages and cultures. Furthermore, we plan to carry out further experiments with more diverse participants and analyse and explore the speaker's audio collected during the experiments.

References

1. Aucouturier, J.J., Johansson, P., Hall, L., Segnini, R., Mercadié, L., Watanabe, K.: Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction. Proceedings of the National Academy of Sciences **113**(4), 948–953 (2016). https://doi.org/10.1073/pnas.1506552113
2. Bahari, A.: Computer-mediated feedback for l2 learners: Challenges versus affordances. Jouney of Computer Assisted Learning pp. 24–38 (2020). https://doi.org/10.1111/jcal.12481
3. Cavalcanti, A.P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.S., Gašević, D., Mello, R.F.: Automatic feedback in online learning environments: A systematic literature review. Computers and Education: Artificial Intelligence **2** (2021). https://doi.org/10.1016/j.caeai.2021.100027
4. Chollet, F., et al.: Keras (2015), https://github.com/fchollet/keras
5. contributors, H.: Panel. https://panel.holoviz.org (2022)
6. Dupuis, K., Pichora-Fuller, M.: Toronto emotional speech set (TESS) Collection (2010). https://doi.org/10.5683/SP2/E8H2MF
7. Huang, X., Ren, M., Han, Q., Shi, X., Nie, J., Nie, W., Liu, A.A.: Emotion detection for conversations based on reinforcement learning framework. IEEE MultiMedia **28**(2), 76–85 (2021). https://doi.org/10.1109/MMUL.2021.3065678
8. Issa, D., Fatih Demirci, M., Yazici, A.: Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control **59**, 101894 (2020). https://doi.org/10.1016/j.bspc.2020.101894
9. Jackson, F.P., Haq, S.: Surrey Audio-Visual Expressed Emotion (SAVEE) Data. http://kahlan.eps.surrey.ac.uk/savee/Introduction.html (2015)
10. Jadoul, Y., Thompson, B., de Boer, B.: Introducing Parselmouth: A Python interface to Praat. Journal of Phonetics **71**, 1–15 (2018). https://doi.org/10.1016/j.wocn.2018.07.001
11. 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
12. McFee, B., Metsai, A., McVicar, M., Balke, S., Thomé, C., Raffel, C., Zalkow, F., Malek, A., Dana, Lee, K., Nieto, O., Ellis, D., Mason, J., Battenberg, E., Seyfarth, S., Yamamoto, R., viktorandreevichmorozov, Choi, K., Moore, J., Bittner, R., Hidaka, S., Wei, Z., nullmightybofo, Hereñú, D., Stöter, F.R., Friesch, P., Weiss, A., Vollrath, M., Kim, T., Thassilo: librosa/librosa: 0.8.1rc2 (May 2021). https://doi.org/10.5281/zenodo.4792298
13. Rachman, L., Liuni, M., Arias, P., Lind, A., Johansson, P., Hall, L., Richardson, D., Watanabe, K., Dubal, S., Aucouturier, J.J.: DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech. Behav. Res. Methods **50**(1), 323–343 (Feb 2018)
14. Sztahó, D., Kiss, G., Vicsi, K.: Computer based speech prosody teaching system. Computer Speech and Language **50**, 126–140 (2018). https://doi.org/10.1016/j.csl.2017.12.010
15. Wynn, A.T., Wang, J., Umezawa, K., Cristea, A.I.: An ai-based feedback visualisation system for speech training. In: Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium. pp. 510–514. Springer International Publishing, Cham (2022)
16. Zhou, K., Sisman, B., Rana, R., Schuller, B.W., Li, H.: Speech synthesis with mixed emotions (2022). https://doi.org/10.48550/ARXIV.2208.05890