

Flowchase: a Mobile Application for Pronunciation Training

Noé Tits¹, Zoé Broisson¹

¹Flowchase, Belgium

noe@flowchase.app, zoe@flowchase.app

Abstract

In this paper, we present a solution for providing personalized and instant feedback to English learners through a mobile application, called Flowchase, that is connected to a speech technology able to segment and analyze speech segmental and supra-segmental features. The speech processing pipeline receives linguistic information corresponding to an utterance to analyze along with a speech sample. After validation of the speech sample, a joint forced-alignment and phonetic recognition is performed thanks to a combination of machine learning models based on speech representation learning that provides necessary information for designing a feedback on a series of segmental and supra-segmental pronunciation aspects.

Index Terms: pronunciation training, language learning, speech analysis, machine learning, transfer learning, human-computer interaction

1. Introduction

In the field of Computer-Assisted Language Learning (CALL), there are today still very few solutions focusing on oral skills, and specifically to pronunciation. Computer-Assisted Pronunciation Training (CAPT) is an important research discipline, but there is a lack of concrete applications, although explicit focus on pronunciation, when combined with the use of technologies, has a significant impact on L2 learners pronunciation [1, 2]. A reason of this situation is the gap of complexity between developing feedback on written, reading or listening skills compared to spoken skills. Indeed for the first three skill sets, implementing simple heuristics based on multiple answer exercises, or matching a user answer to a gold standard is straightforward. On the contrary, providing feedback on spoken skills is not. A speech technology tailored to analyzing segmental and supra-segmental patterns is necessary.

The techniques of mispronunciation errors have been close to the findings of speech recognition area, from HMM-GMM [3], to DNN-HMM [4] and more recently, transformers [5]. Indeed the tasks share a strong common characteristics, which is extracting information from audio a representation of human speech, be it text or phonetics.

Transfer Learning [6] is today a widely used technique in Deep Learning for leveraging models trained on related tasks for which there exist abundant datasets towards task for which few data exist. This principle has been applied successfully for speech technology application [7] with few available data such as speech recognition for low resource languages, emotion recognition in speech [8], emotional or expressive speech synthesis [9, 10] or voice conversion [11], and also to pronunciation assessment [12]. A specific form of Transfer Learning that was shown very efficient is self-supervised learning where a model

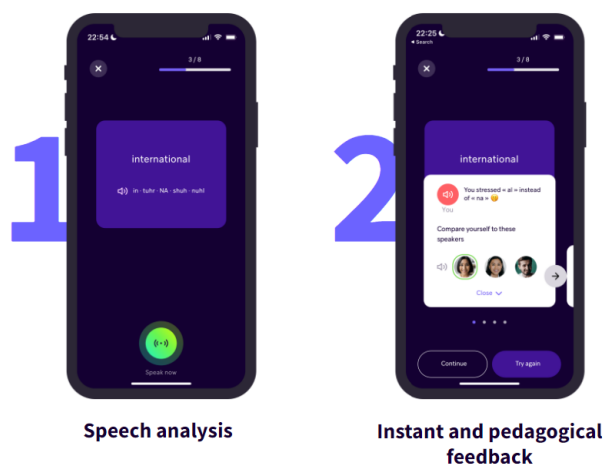


Figure 1: Sequence of steps in a Speaking Activity

is trained to learn representations of input data without the need for explicit supervision.

In this paper, we present a complete system able to provide a pronunciation training based on a speech technology built on top of a wav2vec2 [13] model adapted for mispronunciation detection, integrated in a mobile application. Although the application contains a mix of tutorials, listening activities and speaking activities, we focus here on the description of the speaking activities that involves the speech processing pipeline for analyzing English learners' pronunciation and providing feedback.

2. System

Figure 1 describes the main steps of the user experience inside a speaking exercise of a learning program. First, the exercise data is shown to the user. Specifically, it shows an English utterance that the user is expected to say, with a pronunciation guide to help him understand how it has to be pronounced. The pronunciation of the sentence can also be heard thanks to a set of different actor recordings with different English variations. On this screen, the user can record himself. Then the audio recording is sent to the speech technology backend along with the exercise information in order to perform segmentation and analysis of the speech sample. From this analysis, a number of information are extracted depending on the pronunciation aspect analyzed.

In the second screen, feedback cards are shown to the user in order to communicate him the result of the analysis, and advice in order to improve.

Figure 2 details the processing steps happening in the second step explained above. The speech analysis takes as inputs

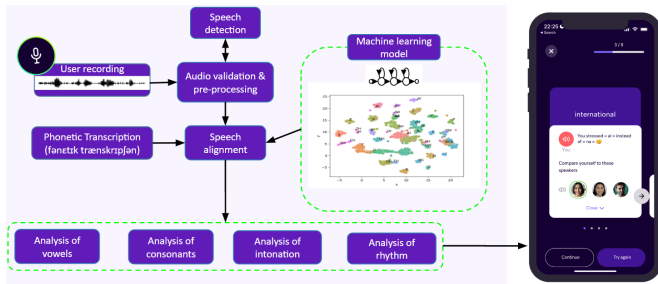


Figure 2: Description of the processing steps of a user recording towards pronunciation feedback

| phones | pred_phones_audio | GT_proba | pred_proba | start | end |
|--------|-------------------|----------|------------|-------|------|
| S | S | 1 | 1 | 0.4 | 0.54 |
| K | K | 0.60474 | 0.60474 | 0.54 | 0.8 |
| IH | IY | 0 | 0.40079 | 0.8 | 0.82 |
| L | L | 0.797643 | 0.797643 | 0.82 | 0.98 |
| Z | S | 0 | 0.096497 | 0.98 | 1.56 |

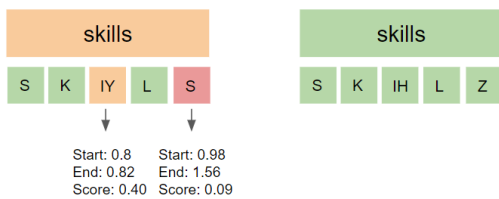


Figure 3: Pronunciation feedback on a word

exercise information such as the phonetic content and the English learner’s speech sample. The user recording has first to be validated thanks to a series of test on audio that checks is a valid speech sample, including:

- the duration of the audio is plausible to have a human speech rate compared to the expected utterance
- the speech sample contains voiced content
- the phonetic content in speech is sufficiently close from the phonetic content

If the speech sample is validated, a combination of machine learning models based on speech representation learning is used for performing a forced-alignment between the speech sample and the phonetic transcription in order to extract the start and end timings of each phoneme of the sequence. The machine learning model also analyzes the phonetic content of the audio and allows us to extract information related to set of different pronunciation aspects such as analysis of vowels or consonants, and specifically analyzing minimal pairs, as shown in Figure 3, analysis of intonation such as word stress or sentence stress, and other supra-segmental aspects like an analysis of pauses between breath groups in an utterance.

An example of analysis result on a word from a sentence is shown in Figure 3. Expected phonemes and predicted phonemes are extracted along with the start and end timings, as well as the respective posterior probabilities according to the statistical model.

3. Conclusions

In this paper, we presented Flowchase, a mobile application for personalized pronunciation training that utilizes a speech technology pipeline for analyzing English learners’ pronunciation and providing instant feedback. We employed transfer learning and self-supervised learning techniques to build a speech

technology model for detecting mispronunciations based on the wav2vec2 architecture.

The system provides feedback on both segmental and supra-segmental aspects of pronunciation. Our solution addresses the gap in current computer-assisted language learning applications, which mostly focus on written, reading, or listening skills. Flowchase provides a tool for improving oral language skills, particularly pronunciation, which is crucial for effective communication. Future work includes testing the effectiveness of the application and the speech technology pipeline in real-world settings and extending the system to support other languages.

4. Acknowledgements

This work is part of the project *REDCALL* that is partially funded by a FIRST Enterprise Docteur program from SPW Recherche¹

5. References

- [1] N. Cordier and F. Meunier, “Digital tools to improve the english pronunciation of 12 learners: a case study with flowchase.”
- [2] X. Anguera and V. Van, “English language speech assistant.” in *INTERSPEECH*, 2016, pp. 1962–1963.
- [3] S. M. Witt and S. J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.
- [4] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, “Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers,” *Speech Communication*, vol. 67, pp. 154–166, 2015.
- [5] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, “Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7262–7266.
- [6] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, “A survey on deep transfer learning,” in *International conference on artificial neural networks*. Springer, 2018, pp. 270–279.
- [7] D. Wang and T. F. Zheng, “Transfer learning for speech and language processing,” in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (AP-SIPA)*. IEEE, 2015, pp. 1225–1237.
- [8] N. Tits, K. El Haddad, and T. Dutoit, “Asr-based features for emotion recognition: A transfer learning approach,” in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*. Association for Computational Linguistics, 2018, pp. 48–52. [Online]. Available: <http://aclweb.org/anthology/W18-3307>
- [9] —, “Exploring Transfer Learning for Low Resource Emotional TTS,” in *Intelligent Systems and Applications*, Y. Bi, R. Bhatia, and S. Kapoor, Eds. Cham: Springer International Publishing, 2020, pp. 52–60.
- [10] N. Tits, F. Wang, K. E. Haddad, V. Pagel, and T. Dutoit, “Visualization and Interpretation of Latent Spaces for Controlling Expressive Speech Synthesis through Audio Analysis,” in *Proc. Interspeech 2019*, 2019, pp. 4475–4479. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2019-1426>
- [11] K. Zhou, B. Sisman, R. Liu, and H. Li, “Emotional voice conversion: Theory, databases and esd,” *Speech Communication*, vol. 137, pp. 1–18, 2022.
- [12] B. Lin and L. Wang, “Deep feature transfer learning for automatic pronunciation assessment,” in *Interspeech*, 2021, pp. 4438–4442.
- [13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

¹<https://recherche.wallonie.be/>