A. Filighera, L. Bongard, T. Steuer and T. Tregel, "Towards A Vocalization Feedback Pipeline for Language Learners," 2022 International Conference on Advanced Learning Technologies (ICALT), 2022, pp. 248-252, doi: 10.1109/ICALT55010.2022.00081.

Towards A Vocalization Feedback Pipeline for Language Learners

Anna Filighera*
Multimedia Communications Lab
Technical University of Darmstadt
Germany
ORCID: 0000-0001-5519-9959

Leonard Bongard*
Department of Computer Science
Technical University of Darmstadt
Gemarny
leonard.bongard@stud.tu-darmstadt.de

Tim Steuer
Multimedia Communications Lab
Technical University of Darmstadt
Germany
ORCID: 0000-0002-3141-712X

Thomas Tregel
Multimedia Communications Lab
Technical University of Darmstadt
Germany
ORCID: 0000-0003-0715-3889

***Abstract*—Practice is essential for language learning. This is true for writing as well as speaking. However, in contrast to writing, it can be challenging to offer students sufficient time to practice speaking while receiving corrective feedback from a teacher. Considering the importance of corrective feedback for language mastery, automatic feedback systems could provide vital assistance through additional supervised speaking exercise. For this reason, this paper proposes an end-to-end feedback generation pipeline to correct grammar errors in unconstrained speech. The approach consists of four steps, converting raw speech files into transcripts with automatic speech recognition, removing disfluency, correcting grammatical errors, and preparing the feedback for presentation to the user. An explorative analysis of the pipeline with English language learners indicates that out-of-the-box automatic speech recognition models degrade in performance when used by language learners. However, training the model with only 15 minutes of learners' speech decreases the word error rate almost by half.**

***Index Terms*—education, computer-assisted language learning, feedback, natural language processing**

I. Introduction

Speaking is an integral aspect of language learning. However, it can be challenging to ensure that students receive enough speaking time in classrooms with high student-teacher ratios. Since having access to corrective feedback on one's speech is essential to learning a language correctly [1], computer-assisted language learning (CALL) has risen in popularity to fill the gap. Here, many systems utilize automatic speech recognition (ASR) systems to provide users feedback on their speech.

However, while most ASR models perform well for native speakers, the performance degrades for second language learners where mispronunciation, accents and grammatical mistakes are common. Currently, most CALL systems mitigate the performance loss by constraining the user's speech to expected sentences, for example, by providing sentences to read aloud or giving specific prompts. Knowing what the user should be expressing simplifies the detection of deviations from the expected speech. However, being limited to specific lessons can be less motivating than freely choosing topics of interest or speaking with friends [2].

Therefore, this work proposes an end-to-end feedback pipeline for correcting grammar errors in spontaneous speech. Considering the success of Transformer-based models on many natural language processing tasks [3]–[5], we hypothesize that recently developed Transformer models perform well enough for non-native speakers to sustain an error correction pipeline. The pipeline's impact on the users' learning motivation, user experience and the quality of the speech's transcription are explored via a user study with English language learners.

II. Related Work

The idea of using automatic speech recognition systems to facilitate language learning was previously mainly discussed in the context of correcting learners' pronunciation [6]–[8] or measuring fluency [9], [10]. However, a few approaches have also addressed the need for the correction of grammatical errors [11]. Wang, Waple and Kawahara, for example, proposed a system that prompted Japanese language learners to speak or type target sentences based on visual prompts [12]. The system would then compare the learner's input to error patterns typically made by language learners based on manually defined features. Many other approaches, such as Duolingo [13], also improve speech recognition performance by constraining possible utterances through prompts [14].

Closer to our approach, Lu, Gales and Wang experiment with various components of a grammar error correction pipeline for spontaneous speech [15]. Their work mainly focuses on testing various ways to deal with disfluency in speech when correcting grammar instead of providing end-to-end feedback to the user. Similarly, [16]–[18] focus on

This research is funded by the Bundesministerium für Bildung und Forschung in the project: Software Campus 2.0 (ZN 01—S17050), Micro-project: DA-VBB.

*equal contribution

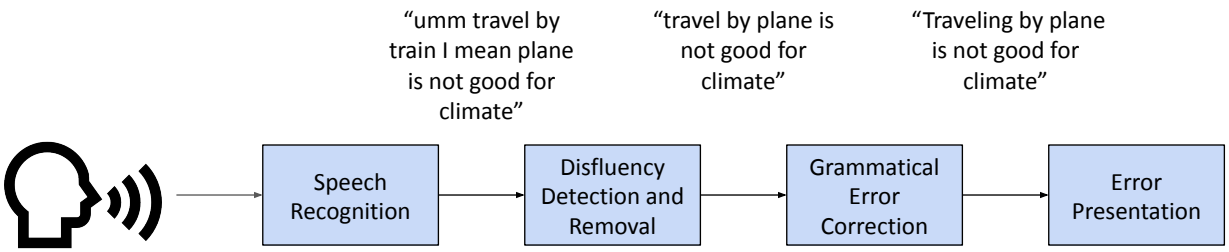

Fig. 1. Schematic depiction of the spoken error correction pipeline.

adapting ASR models to non-native speakers that tend to speak more disfluently and accented while [19] focus on improving grammar error correction models.

III. Spoken Error Correction Pipeline

The proposed pipeline consists of four modularly exchangeable components. A schematic overview can be found in Fig. 1. After recording the user's speech with a microphone, the audio files are transcribed by an ASR model. We selected the wav2vec-large-960h[1] as ASR component for two reasons. In contrast to cloud models, it offers multiple pre-trained checkpoints that can be downloaded and fine-tuned with additional data. Moreover, it does not require post-processing to produce state-of-the-art results. Many approaches post-process using language models to automatically correct mistakes made during the transcription. However, using a language model to correct transcription mistakes would most likely automatically correct the user's grammatical errors, which would defeat the pipeline's purpose.

One of the major differences between spoken and written language is disfluency. Speakers may use filler words, such as “um” or “er”, repeat words or phrases or correct themselves. As disfluency is common in spontaneous speech, especially during language learning, it is crucial to train models to handle it properly. Since the state-of-the-art grammar error correction systems are mainly trained on written texts, this work follows Lu et al.'s [15] approach of using a dedicated disfluency detection model [20] to remove disfluent utterances from the transcription. As we do not aim to provide feedback on the speaker's fluency, it makes sense to remove disfluency altogether so that the pipeline's data is more similar to the grammar error corrections model's training data.

The resulting disfluency-free transcriptions are then passed to the grammar error correction component. In this step, the text is transformed into grammatically correct text. We chose the Gector model by Grammarly [21] for this purpose due to its state-of-the-art performance, speed and generalizability. Since the model is trained to correct written texts, modifications must be made. For example, we ignore punctuation and spelling mistakes made by the ASR model.

Finally, the errors are presented to the user to facilitate learning. For this purpose, it is essential that made corrections are easily visible and explained. Since the grammar error correction model does not provide any explanation of the errors corrected, we employ the grammatical Error Annotation Toolkit (Errant) [22]. It compares the in- and output of Gector and classifies the changes made. The pipeline then highlights the error, the correction and the error type provided by Errant. Each error type is associated with a short description, which is displayed to the user. As can be seen in Fig. 2, the pipeline also displays a history of past errors made.

[1]https://huggingface.co/facebook/wav2vec2-large-960h

IV. Experiments

The main challenge for evaluating the feedback generation pipeline lies in the fact that there are no publicly available end-to-end datasets [11]. Some datasets contain transcriptions for audio files but without disfluency and error correction annotations. Often, the speakers recorded are also very proficient in the spoken language and, thus, make few mistakes in general. Other datasets contain error correction annotations but no corresponding audio recordings. Therefore, the pipeline is evaluated end-to-end with a user study. Additionally, we investigate personalizing the speech recognition component to improve its performance. In total, we aim to address the following questions:

- How motivating is it to use the feedback generation pipeline for language learning?
- How is the user experience when using the pipeline?
- What could be possible avenues to improve the pipeline in the future?

A. User Study Design

The first step in designing the user study was to decide on a task users should complete with the pipeline. On the one hand, the task should be comparable between all users. On the other hand, it should leave room for the users to make grammar mistakes that the pipeline can correct. For this purpose, users should verbally translate a given letter from German to English. They were not allowed to write their translation down or take notes. This ensured that all users would speak similar texts but still likely make mistakes.

The study utilized a 4-part questionnaire to collect basic demographic information, measure motivation before and after the task, and question the user's experience with the pipeline. Before the translation task, the demographic questions were answered, querying the user's age, gender, and English proficiency level. Then, we measure motivation before

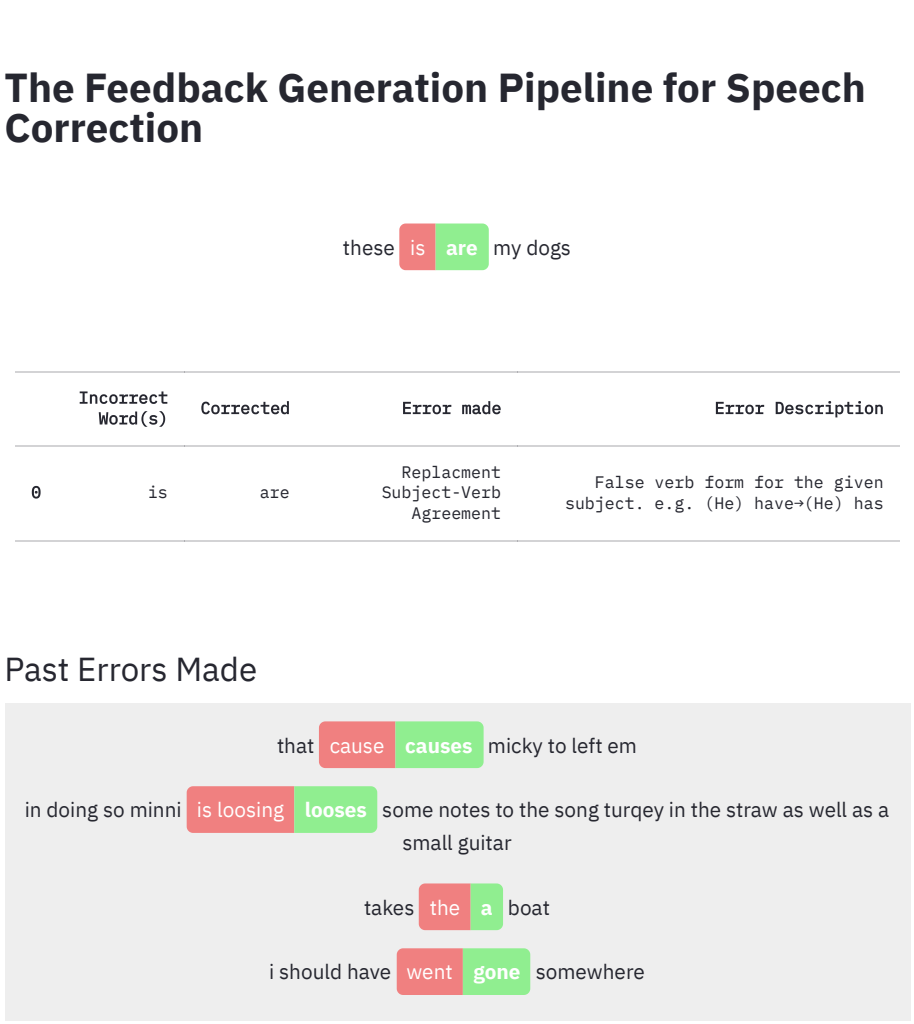

Fig. 2. Screenshot of the error presentation to the user.

and after translation based on an English learning motivation questionnaire validated by Taguchi, Magid and Papi [23]. As the questionnaire also covers aspects of language learning motivation that are not central to this study, such as writing or family influence, we selected six relevant items from the question catalogue.

The user experience items originate from an established questionnaire querying the software's attractiveness, perspicuity, efficiency, dependability, stimulation and novelty [24]. Additionally, we asked whether the pipeline had correctly transcribed the user's speech. Questions were answered on a 6-point Likert scale, with 6 being the high end of the scale.

Study participants were selected based on their availability and whether they spoke English as a second language. In total, 30 participants completed the task and questionnaire. Of those, 19 translated while using the pipeline and 11 acted as a control group and did the task without the pipeline. Participation in the study was voluntary and could be aborted at any time. The size difference between both groups is due to more people (N=9) declining to complete the task in the control group than in the treatment group (N=1). Most participants (N=28) spoke German as their first language and the rest spoke German fluently. Eleven of the participants identified as female, 18 as male, and one preferred not to provide a gender classification.

B. User Study Results

Fig. 3 shows the median and average rating for each of the experience items assigned by users in the treatment group. The software attributes easy, clear, creative, exciting, practical and supportive received a median score of 5. In contrast, supportiveness and efficiency have a lower median of 4.

Interestingly, efficiency and supportiveness seem to strongly correlate with how well the pipeline understood the user (Spearman's rank correlation coefficient of 0.623 and 0.515, respectively). Thus, the software may be perceived as less supportive and efficient due to mistakes made by the ASR model. Generally, the ASR model's transcription was perceived to be of moderate quality with an average score of 3.4, where 1 is "The software did not understand me correctly at all" and 6 is "Yes, the software understood correctly what I said". Especially low-proficiency speakers felt like the system did

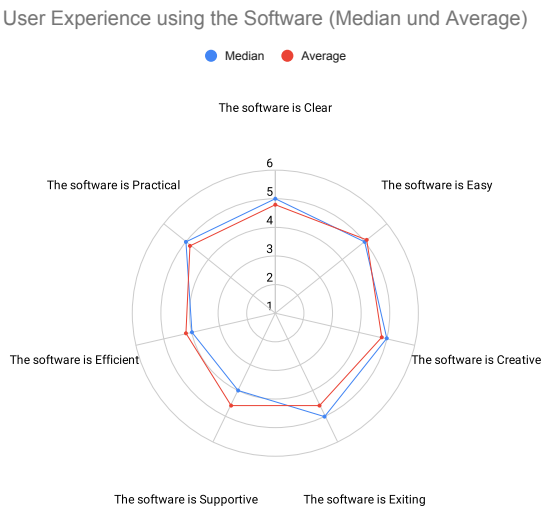

Fig. 3. Average and median user experience ratings for each category (N=19).

not understand them. The overall medium rating on the ASR quality item indicates that out-of-the-box ASR systems do not seem to perform well for English language learners with predominately German accents.

TABLE I
MEAN MOTIVATION BEFORE AND AFTER TASK COMPLETION.

Group	N	Pre		Post		Mean Δ
		Mean	SD	Mean	SD	
Control	11	23.8	3.6	24.8	4.3	1.0
Treatment	19	19.6	5.7	21.4	6.1	1.7

The results of the motivation questionnaire can be found in Table I. As can be seen, there was a higher motivation in the control group before and after the task compared to the treatment group. This could be due to less motivated individuals not completing the experiment in the control group, as discussed in the user study design. That would also explain why the standard deviation is lower in the control group. Nevertheless, on average, the motivation increased more in the treatment group (ΔM=1.7) than in the control group (ΔM=1). Thus, translating using the pipeline seems slightly more motivating than solving the task without it.

Considering that not being understood correctly can be frustrating, thus lowering motivation, a better ASR system may consolidate the observed trend.

C. Training the ASR System

The user study's results indicate that the speech recognition system did not work as intended for language learners. We hypothesize that personalizing the ASR to the user's voice and accent may mitigate the performance loss for language learners. Thus, we recorded a native German speaker reading sentences aloud for a prototypical personalization of the ASR model. Since at least 10-minutes of speech are required to train the wav2vec2 model [25], we recorded 15 minutes. We also believe this to be the upper limit of how much data could be feasibly acquired from the pipeline's users before it becomes too tedious. Similar to related work [17], [26], we observed word error rates above 20% on our dataset for common ASR models like Speech-to-Text[2], DeepSpeech[3] and wav2vec. This means that at least every fifth word is transcribed incorrectly by the system, explaining the study participants' dissatisfaction with the transcription quality of the pipeline and illustrating the need for specialized data and training.

Since 15 minutes are short for training, developing and testing models, we added another hour from the LJ Speech Dataset [27]. While it contains sentences read by a native English speaker, obtaining a sufficiently sized dataset is necessary. We split 80% of the dataset off for fine-tuning, 10% for validation and another 10% for the final evaluation. We fine-tuned Facebook's pre-trained word2vec-base[4] model on our dataset following Patrick von Platen's guideline [28].

D. Results of Training the ASR System

A comparison of the fine-tuned model's performance with the out-of-the-box word2vec models' can be found in Table II. The model fine-tuned on a single hour of a mixed language proficiency corpus has almost half the word error rate of its correspondent fine-tuned on 960 hours of phone calls in the USA. It even performs better than the large model by 2.6 percentage points. This result illustrates the benefit of collecting language learner speech data for learning pipelines.

TABLE II
COMPARISON OF WORD ERROR RATE OF FINE-TUNED ASR MODEL WITH OUT-OF-THE-BOX MODELS ON OUR LANGUAGE LEARNER CORPUS.

Model	Dataset's Length	Word Error Rate
wav2vec2-base	960h	9.4%
wav2vec2-large	960h	7.5%
wav2vec2-base (ours)	1h 15 min	4.9%

V. CONCLUSION & FUTURE WORK

Correctional feedback is essential when learning to speak a language but can be expensive and time-consuming to provide. For this purpose, this work introduces an end-to-end grammar feedback pipeline for spontaneous speech. The pipeline transcribes audio input, removes disfluency, corrects grammar mistakes and then presents the feedback to the user. An explorative user study investigated the pipeline's impact on language learning motivation and user experience (N=30). We observed a slightly larger motivation increase in participants using the pipeline to complete a task compared to participants completing the same task without the pipeline. Nevertheless, considering the small sample size, a more extensive follow-up study should be conducted to statistically investigate this effect.

[2]https://cloud.google.com/speech-to-text
[3]https://github.com/mozilla/DeepSpeech
[4]https://huggingface.co/facebook/wav2vec2-base

While the user experience was satisfactory overall, the automatic transcription of the learners' speech performed less well than expected. However, personalizing the speech recognition system with only 15 minutes of recorded speech improved the word error rate almost by half compared to the out-of-the-box model. This indicates that technology may have matured sufficiently for a personalized feedback pipeline when users are willing to invest a bit of time. Still, a large-scale dataset spoken by language learners from various countries would likely improve speech recognition further.

In future work, a transcription verification step may also be added to the pipeline. This would help the user correct the ASR's mistakes and could simultaneously be used for online learning, possibly reducing the number of mistakes in future uses. Finally, while each pipeline component has been validated individually in prior work, a final end-to-end evaluation of the feedback's quality with experts would be beneficial. Besides the raw performance of the pipeline, pedagogical aspects of the provided feedback could be evaluated in such a study.

REFERENCES

[1] H. Roothooft and R. Breeze, "A comparison of EFL teachers' and students' attitudes to oral corrective feedback," *Language Awareness*, vol. 25, no. 4, pp. 318–335, 2016. [Online]. Available: https://doi.org/10.1080/09658416.2016.1235580

[2] E. A. Patall, H. Cooper, and J. C. Robinson, "The effects of choice on intrinsic motivation and related outcomes: a meta-analysis of research findings." *Psychological bulletin*, vol. 134, no. 2, pp. 270–300, 2008.

[3] L. Camus and A. Filighera, "Investigating transformers for automatic short answer grading," in *Artificial Intelligence in Education*, I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, Eds. Cham: Springer International Publishing, 2020, pp. 43–48.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[5] T. Steuer, A. Filighera, and C. Rensing, "Remember the facts? investigating answer-aware neural question generation for text comprehension," in *Artificial Intelligence in Education*, I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, Eds. Cham: Springer International Publishing, 2020, pp. 512–523.

[6] A. Neri, C. Cucchiarini, and W. Strik, "Automatic speech recognition for second language learning: How and why it actually works," in *15th International Congress of Phonetic Sciences (ICPhS-15)*, M. J. Solé, D. Recasens, and J. Romero, Eds., 2003, pp. 1157–1160.

[7] C. Cucchiarini and H. Strik, "Automatic speech recognition for second language pronunciation training," in *The Routledge handbook of contemporary English pronunciation*. Routledge, 2017, pp. 556–569.

[8] A. K. Elimat and A. F. AbuSeileek, "Automatic speech recognition technology as an effective means for teaching pronunciation." *JALT CALL Journal*, vol. 10, no. 1, pp. 21–47, 2014.

[9] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *The Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 989–999, 2000. [Online]. Available: https://doi.org/10.1121/1.428279

[10] A. Caines, E. Flint, and P. Buttery, "Collecting fluency corrections for spoken learner English," in *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 91–100. [Online]. Available: https://aclanthology.org/W17-5010

[11] K. Knill, M. Gales, P. Manakul, and A. Caines, "Automatic grammatical error detection of non-native spoken learner english," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8127–8131.

[12] H. Wang, C. J. Waple, and T. Kawahara, "Computer assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition," *Speech Communication*, vol. 51, no. 10, pp. 995–1005, 2009, spoken Language Technology for Education. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639309000430

[13] K. Teske, "Duolingo," *CALICO Journal*, vol. 34, no. 3, pp. 393–401, 2017. [Online]. Available: https://www.jstor.org/stable/90014704

[14] J. Van Doremalen, C. Cucchiarini, and H. Strik, "Optimizing automatic speech recognition for low-proficient non-native speakers," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–13, 2009.

[15] Y. Lu, M. J. Gales, and Y. Wang, "Spoken language 'Grammatical Error Correction'," in *Proceedings of Interspeech 2020*, 2020, pp. 3840–3844.

[16] Y. Lu, M. J. Gales, K. M. Knill, P. Manakul, L. Wang, and Y. Wang, "Impact of ASR performance on spoken grammatical error detection," in *Proceedings of Interspeech 2019*, 2019, pp. 1876–1880.

[17] T. Viglino, P. Motlicek, and M. Cernak, "End-to-end accented speech recognition." in *Proceedings of Interspeech 2019*, 2019, pp. 2140–2144.

[18] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007, intrinsic Speech Variations. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639307000404

[19] Y. Fathullah, M. Gales, and A. Malinin, "Ensemble distillation approaches for grammatical error correction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 2745–2749.

[20] P. Jamshid Lou and M. Johnson, "Improving disfluency detection by self-training a self-attentive model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 3754–3763. [Online]. Available: https://aclanthology.org/2020.acl-main.346

[21] K. Omelianchuk, V. Atrasevych, A. Chernodub, and O. Skurzhanskyi, "GECToR – grammatical error correction: Tag, not rewrite," in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Seattle, WA, USA → Online: Association for Computational Linguistics, Jul. 2020, pp. 163–170. [Online]. Available: https://aclanthology.org/2020.bea-1.16

[22] C. Bryant, M. Felice, and T. Briscoe, "Automatic annotation and evaluation of error types for grammatical error correction," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 793–805. [Online]. Available: https://aclanthology.org/P17-1074

[23] T. Taguchi, M. Magid, and M. Papi, *4. The L2 Motivational Self System among Japanese, Chinese and Iranian Learners of English: A Comparative Study*. Multilingual Matters, 2009, pp. 66–97. [Online]. Available: https://doi.org/10.21832/9781847691293-005

[24] B. Laugwitz, T. Held, and M. Schrepp, "Construction and evaluation of a user experience questionnaire," in *Symposium of the Austrian HCI and usability engineering group*. Springer, 2008, pp. 63–76.

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[26] J. Huang, O. Kuchaiev, P. O'Neill, V. Lavrukhin, J. Li, A. Flores, G. Kucsko, and B. Ginsburg, "Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition," *arXiv preprint arXiv:2005.04290*, 2020. [Online]. Available: https://arxiv.org/abs/2005.04290

[27] K. Ito and L. Johnson, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[28] P. von Platen, "Fine-tune Wav2Vec2 for English ASR with HuggingFace transformers," https://huggingface.co/blog/fine-tune-wav2vec2-english, 2021.