Assessing the Influence of Voice Conversion on Comprehension and Engagement in Educational Lecture Recordings

Abstract

This study examines the impact of replacing foreign-accented speech with voice-converted native speech in lecture recordings. The experiment dataset consisted of parts of three speech corpora: L2-Arctic (non-native), LibriTTS (native), and CVSS-T (synthetic audio created from the S2S translation model). Through the transcription prototype, we received 210 audio file transcription results. Assessed transcription accuracy through Word Error Rate (WER) using Levenshtein distance and tracked the number of replays required. Results suggest that while synthetic speech enhances engagement, it sometimes suffers from comprehension issues similar to foreign-accented speech, potentially affecting learner interest.

Keywords: voice conversion, word error rate, comprehensibility, speech corpus, Levenshtein distance

1 Introduction

Since COVID-19, online platforms have become the primary alternative to face-to-face learning. However, according to a survey conducted at a Saudi university, 59% of the students report difficulties in understanding online lectures, emphasizing the need for improving the delivery of such lectures [1]. One promising solution is voice conversion technology, transforming the utterances of the source speaker to resemble the target speaker [2]. As 69% of Saudi university students also report their preference for recorded lectures over live sessions [1], making the application of voice conversion to lecture more viable. By applying voice conversion technology, we can enhance the comprehensibility of lectures delivered by foreign instructors, addressing both the challenge of comprehensibility and the preference for recorded content. Although this approach can potentially improve the comprehensibility of lecture audio, the issue of motivation becomes the primary concern for home learners [3]. This paper will explore how synthetic conversion speeches impact the attentiveness and willingness of language learners to face challenges compared to less comprehensible foreign-accented lectures.

2 Methods

2.1 Speech corpora

- **L2-Arctic:** Non-native English speech corpus, commonly used and designed for voice conversion tasks. The corpus was recorded in a background noise-free environment [4].

- **LibriTTS:** Native English speech corpus designed for text-to-speech task. It originates from the LibriSpeech corpus, including segmented utterances after screening out all noisy audio recordings [5].

- **CVSS-T S2S translation corpus:** Synthetic English translation corpus deemed to have high speech similarity with LibriTTS unseen speaker. Although synthetic speech preserves the vocal features of the source speakers, it is less natural and more intelligible. The corpus originates from the Common Voice speech corpus and the CoVoST 2 speech-to-text translation corpus. As the CoVoST 2 corpus contains background noises, the CVSS-T corpus likewise contains background noises. However, his minimal noise is robust enough for the speaker's speech clarity [6].

2.2 Speech-to-Speech (S2S) translation model:

The main benefit of the speech-to-speech translation model is that it preserves para-linguistic information from the source speaker, such as their emotion and prosody. The encoder from Figure [7] extracts the source speaker's linguistic and acoustic elements using mel-spectrograms. The attention module processes the extracted information, enabling the linguistic decoder to predict the translated speech's phoneme sequence. Finally, based on this sequence, synthetic speech maintaining the original speaker's vocal characteristics while translating the content into the target language is generated [7]. As translation and voice conversion models both use encoder-decoder architecture with aims to emulate the target speaker's acoustic characteristics [8], the CVSS-T translation corpus is viable for the experiment purposes in Section 3.

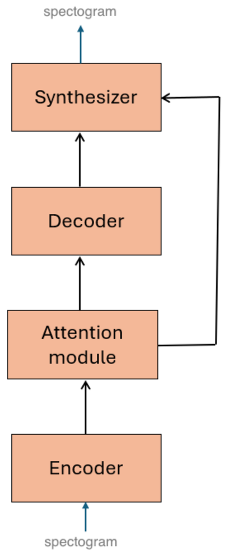


Figure 1: S2S model encoder-decoder

2.3 Levenshtein Distance:

A spelling error typically involves three types of editing operations: insertion, deletion, and substitution. One common way to quantify such spelling errors between two strings is the word error rate (WER). Phonetic comparison tasks, on the other hand, use phonetic Word Error Rate (pWER) instead. This version of the WER formula calculates accuracy by evaluating phonemes rather than entire words. The formula for pWER is defined as:

$$\text{pWER} = \frac{S + D + I}{N}$$

- where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of phonemes in the reference string.

The pWER function uses Levenshtein distance as a cost function to quantify the phonetic similarity between the two strings by measuring the minimum amount of these primary edits required to transform one string to the other.

The Levenshtein distance d between two strings a and b is defined as:

$$d(i, j) = \min \begin{cases} max(i, j), \\ d(i-1, j) + 1, \\ d(i, j-1) + 1, \\ d(i-1, j-1) + cost(a_i, b_j). \end{cases}$$

- where $\text{cost}(a_i, b_j)$ is 0 if $a_i = b_j$, and otherwise adjusted based on the phonetic similarity between the phonemes [9].

3 Experiment

3.1 Experiment Objectives:

- **Participant goal:** Transcribe audio files to achieve the lowest possible Word Error Rate (WER) while minimizing the number of replays. As calculation utilizes phonemes, advice to phonetically transcribe the audio file to get the lower WER score is given to the participant.

- **Experimenter goal:** To assess and analyze participants' transcription performance on foreign-accented English speech and converted synthetic native speech. Examine how the audio file speech type influences participants' willingness to complete the transcription task.

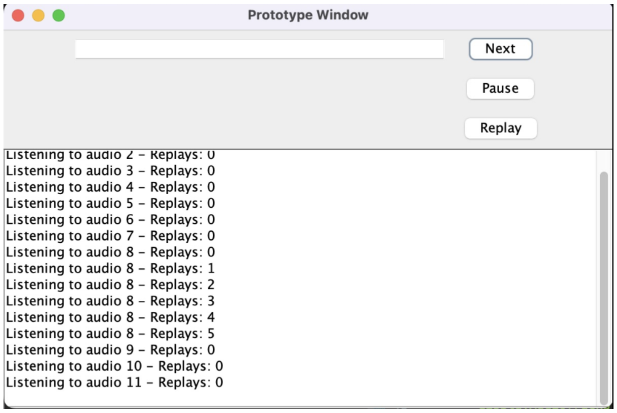
- **Experiment duration:** Depending on the number of times the participant replays the audio files, experiment completion requires approximately 10-15 minutes.

3.2 Audio dataset setup:

From the speech corpora in Section 2, we selected 500 audio files with transcriptions per category: foreign-accented English (L2-Arctic), synthetic native English (CVSS-T), and native English categories. The audio files duration is approximately 3-5 seconds, containing 14-20 spoken words with a speaking speed of 200 words per minute (WPM). Although the CVSS-T translation corpus includes a mix of background noise in its audio files, minimal background noise recordings were hand-selected for the dataset.

3.3 Prototype setup:

The prototype for the audio transcription task was developed using the Java Swing library. It randomly selects 30 categorically balanced audio files from the dataset for each session. The prototype preprocesses the participant input and the corpora transcriptions by removing capitalization and punctuation. The prototype in Figure 2 includes three buttons: next, pause, and replay. The next button submits the participant's experimental results and advances them to the following audio file. The pause button stops the current audio at the exact pressed moment and resumes when used again. The replay button restarts the current audio file from the beginning and tracks its usage for each audio file. Once the participant completes all 30 audio files, the system calculates the WER for each transcription using the LevenshteinDistance function from the Apache Commons library, comparing the participant's transcription against the corpus transcription.



```
Listening to audio 2 – Replays: 0
Listening to audio 3 – Replays: 0
Listening to audio 4 – Replays: 0
Listening to audio 5 – Replays: 0
Listening to audio 6 – Replays: 0
Listening to audio 7 – Replays: 0
Listening to audio 8 – Replays: 0
Listening to audio 8 – Replays: 1
Listening to audio 8 – Replays: 2
Listening to audio 8 – Replays: 3
Listening to audio 8 – Replays: 4
Listening to audio 8 – Replays: 5
Listening to audio 9 – Replays: 0
Listening to audio 10 – Replays: 0
Listening to audio 11 – Replays: 0
```

Figure 2: Experiment prototype window design

3.4 Experiment procedure:

1. The experimenter informs the participants of their goals.
2. The participant wears headphones to minimize environmental noise.
3. The participant begins the experiment and transcribes the audio file.
4. Step 3 is repeated for a total of 30 audio files.

4 Result

The experiment in Section 3 extracts the following numerical features from each audio transcription task: WER, replay, audio duration, transcription length, and WPM. A total of 210 audio transcription results were collected from 7 participants in the experiment.

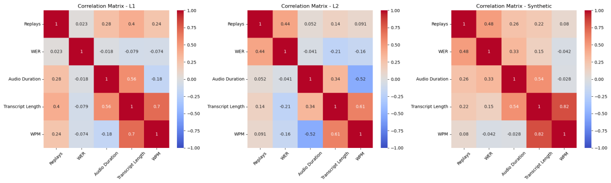


Figure 3: Correlation matrix of result features by audio type

The correlation matrix in Figure 3 shows that for all speech types, an increase in transcription length corresponds with an increase in speaking rate, and an increase in audio duration corresponds with an increase in transcription length. For L2 and synthetic speech, the primary feature WER shows a strong positive correlation with replay number. However, for L1 speech, WER did not show any significant correlation with the other features. Moreover, for L2 speech, a decrease in audio duration corresponded with an increase in the speaker's speaking rate, and for L1 speech, the number of replays increased with the audio duration.

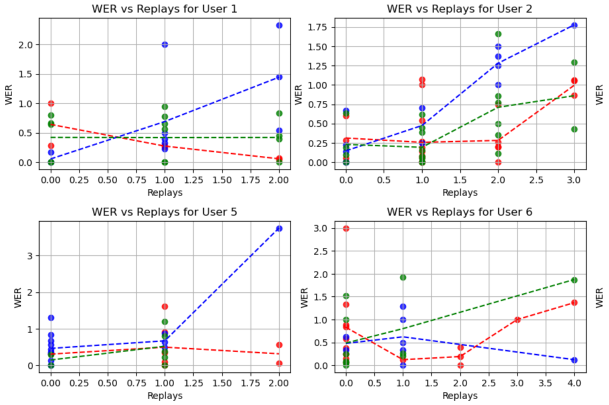


Figure 4: WER vs Replay Scatter Plot: by Participant (1,2,5,6)

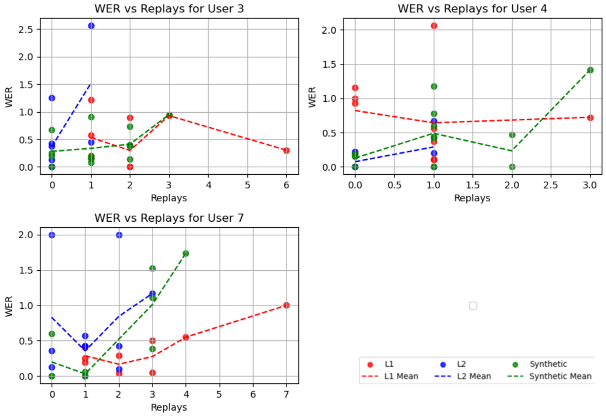


Figure 5: WER vs Replay Scatter Plot: by Participant (3,4,7)

The scatter plots in Figures 4 and 5 show the relationship between the WER of each audio file transcription and the number of replays participants required to achieve the score. The plot displays that higher WER correlates with more replays, suggesting that more difficult audio (higher WER) requires more listening attempts for proper comprehension. Individual participants show varying trends in the relationship between WER and replays. This variance could be due to differences in language abilities or familiarity with the audio types. For example, for Participants 1 and 5, L2 (foreign-accented) audio shows a steep increase in WER with more replays, indicating that these participants find L2 particularly challenging. Moreover, L1 (native) speech seems to be easiest to transcribe for most participants, as it overall has the lowest WER levels among participants. As for synthetic speech, they display a mixed pattern, with WER values scattered across different ranges. The mean line of synthetic speech sometimes overlaps with L1 and L2's mean line, hinting at the CVSS-T corpus creation procedure.

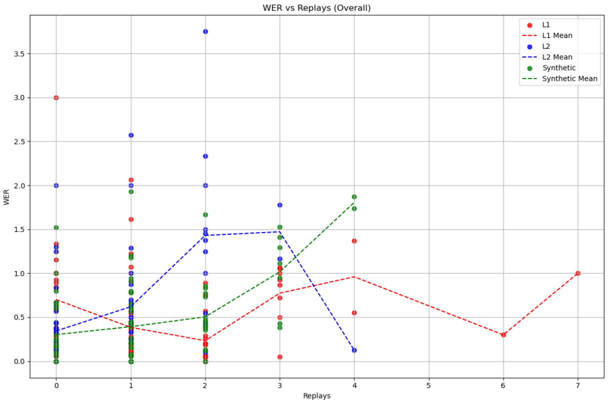


Figure 6: Overall WER vs Replay Scatter Plot

The scatter plot in Figure 6 through its WER mean lines highlights the differing correlation trends associated with each audio type. As WER lies low throughout the smaller number of replays, L1 audio is generally the easiest for participants to transcribe. On the other hand, L2 audio presents more challenges for the participants, as indicated by its higher WER and replay numbers. There is also a significant dip in trend at four replays, but as only one data point represents this pattern, it can be considered an outlier. The synthetic audio and L2 show a similar trend, but the relationship between the WER and replay count seems more positively significant for synthetic speech.

5 Discussion

5.1 Impact of Speech Type on Listening Difficulty

L1 speech is consistently the easiest to comprehend, having the lowest overall WER and the fewest replays. In comparison, synthetic speech is more challenging, and L2 speech is the most difficult to understand. However, there were outliers to this pattern, with Participant 4 performing the best on L2 and worst on L1 speech transcription.

Although more challenging, the minimal difference in comprehension levels between L2 and synthetic speech suggests the result is unreliable, as the individual participant differences seen in Figures 4 and 5 might have created this difference.

5.2 Learner Engagement and Motivation

The relationship between WER and replay counts informs the learner's engagement and motivation with the transcription task. The strong positive correlation between these metrics for L2 and synthetic speech means that the experiment was stimulating enough to represent the learner's behavior for a typical learning task, unlike the relationship in L1 speech. Although this shows that learners are willing to put in more effort to comprehend challenging audios, with the additional replays, the WER score for the transcriptions was not improving, suggesting their efforts were useless. It conveys the possibility of some audio files being inherently incomprehensible.

For L2 speech, although most transcription attempts with low WER required only two replays, occasionally, some attempts display extremely high WER with high replay amounts. The result presents that the audio files are entirely incomprehensible, or after many replays, the user gives up and submits the phonetically approximated transcription. The latter option is more plausible, as we selected the dataset from reliable corpora. With a higher number of replays compared to L1, the synthetic native speech seems to be more motivating, with a stimulating yet reasonable difficulty in its comprehensibility level.

6 Limitation

The experiment was limited to only 7 participants, which potentially introduced personal biases and affected the generalizability of the results. The lack of a concrete method to cross-validate learners' attentiveness and motivation beyond performance metrics leaves some uncertainty about the accuracy of the findings. Integrating bio-metric recording devices could help address this limitation by providing data on participants' engagement levels.

Furthermore, training a voice conversion model on the L2-Arctic dataset could result in more personalized and relevant data compared to the CVSS-T corpus used for the experiment. This adjustment would likely enhance the task's alignment with the objective, potentially leading to more precise and insightful results.

7 Conclusion

Voice conversion technology can enhance the comprehensibility of foreign-accented lecture recordings to be more engaging and motivating for learners. The results reveal that native speech (L1) is generally the easiest for learners to understand, while foreign-accented speech (L2) and synthetic speech pose varying degrees of difficulty. Additionally, the study highlights that while synthetic speech provides a reasonable level of engagement, it does not improve transcription accuracy consistently, suggesting inherent challenges in some audio files. The limited participant pool and the absence of a cross-validation method present potential biases and usage limits for the experiment results. Future research should address these limitations with a larger participant pool and robust cross-validation control variable.

Reference

[1] Al-Jarf, R. (2021). EFL speaking practice in distance learning during the coronavirus pandemic 2020-2021. *International Journal of Research - GRANTHAALAYAH*, 9(7), 179-196.

[2] Felps, D., Bortfeld, H., and Gutierrez-Osuna, R. (2009). Foreign accent conversion in computer assisted pronunciation training. *Speech Communication*, 51(10), 920-932.

[3] Alkhudiry, R., & Alahdal, A. (2021). The Role of Online Learning during and Post COVID-19: A Case of Psycho-Social Study. *TESOL International Journal*, 16(1), 119-138.

[4] Zhao, G., Chukharev-Hudilainen, E., Sonsaat, S., Silpachai, A., Lucic, I., Gutierrez-Osuna, R., & Levis, J. (2018). *L2-arctic: A non-native English speech corpus*. In *Interspeech 2018*.

[5] Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., & Wu, Y. (2019). *LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech*. In *Interspeech 2019*.

[6] Jia, Y., Ramanovich, M. T., Wang, Q., & Zen, H. (2022, June). *CVSS Corpus and Massively Multilingual Speech-to-Speech Translation*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 6691-6703).

[7] Jia, Y., Ramanovich, M. T., Remez, T., & Pomerantz, R. (2022, June). Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *International Conference on Machine Learning* (pp. 10120-10134). PMLR.

[8] Lian, J., Zhang, C., & Yu, D. (2022, May). Robust disentangled variational speech representation learning for zero-shot voice conversion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6572-6576). IEEE.

[9] Gueddah, H., Yousfi, A., & Belkasmi, M. (2015, November). The filtered combination of the weighted edit distance and the Jaro-Winkler distance to improve spellchecking Arabic texts. In *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)* (pp. 1-6). IEEE.

[10] Graham, S. (2006). Listening Comprehension: The Learners' Perspective. System: *An International Journal of Educational Technology and Applied Linguistics*, 34(2), 165-182.