

Supplementary Materials: Convert and Speak: Zero-shot Accent Conversion with Minimum Supervision

Anonymous Authors

1 APPENDIX

1.1 Examples of accent speech

Accent features affect both pronunciation units and prosody, e.g. intonation, stress and duration. Figure 1 shows the pitch contour of one example of the parallel data in which the same content is spoken with two accents. Generally, Indian-English tends to speak with more cadence while the general American-English sounds more monotonous.

1.2 In-context learning with more accent prompt types

To verify if the accent feature from the 3 seconds of accent speech will be extended through the in-context learning, we add Chinese-accent and Korean-accent as another two prompt accent types. The CommonAccent metric on Chinese-accent and Korean accent prompt is 96% and 98%, respectively, as shown in Table 1.

1.3 Effects of accent prompt length

In this section, we evaluate the effect of accent prompt length on the synthesized speech. Specifically, we build another 100 pairs of samples as Section 4.5.2. 5 general American-English speakers from LibriTTS test-clean test set and 5 Indian-English speakers from IndicTTS data set¹ are randomly selected. For each pair, the prompt utterances are cut to {3, 5, 7} seconds as testing prompts. The classification results by CommonAccent are shown in Figure 2. Based on the results, we find the accent do have effects on the prosody modeling which results in the more diverse classification results on other accent types for Indian-English accent prompt. With longer prompt length, the disturbance on the prosody becomes larger. But the effects are limited and non-directional, e.g. most of the cases are still identified as general American-English and the false cases are not all belongs to Indian-English accent type. This suggests that accent features are not effectively maintained through in-context learning.

1.4 More advantages from semantic conversion module

As previous results shown, the semantic conversion module is the key to a good accent conversion quality in the proposed framework. In this section, we explore additional advantages offered by the semantic conversion module besides phoneme conversion. We find the semantic tokens extracted from the speech provides a kind of local fine-granularity control over the prosody modeling, e.g. intonation and rhythm. As Table 2 shown, the semantic conversion module contributes to a much better accent conversion quality even on the accent speech without phoneme changes. Additionally, as a typical example shown in Figure 3, the proposed framework speaks at similar speed as the accent source while text-guided model

Liu’s lose the rhythm control completely from the source signal. We think this is a good benefit of using semantic features from the speech instead of text as the intermediate representation since the rhythm control is important for speech processing tasks.

¹<https://www.kaggle.com/datasets/tuannguyenvananh/indic-tts>

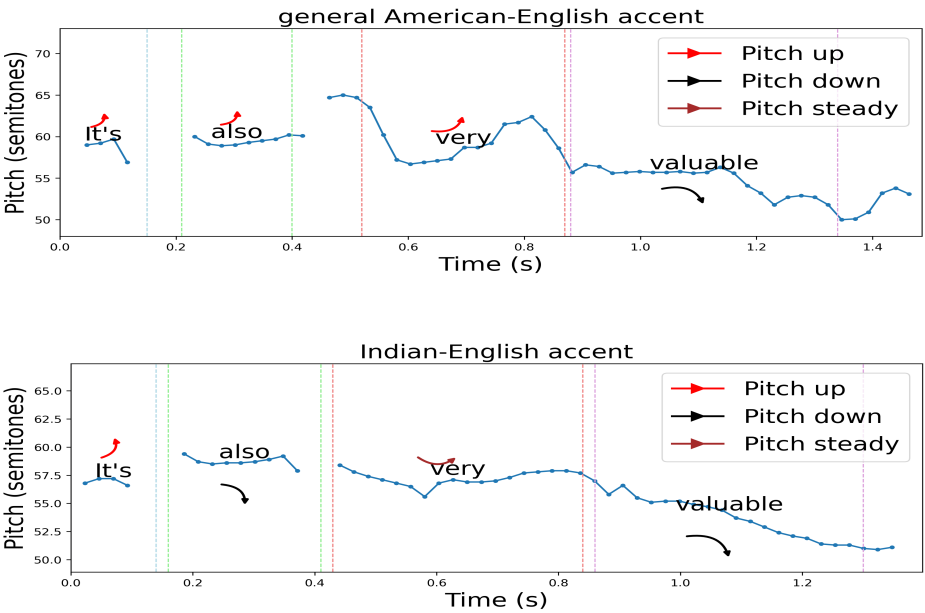


Figure 1: Pitch contour of general American-English accent compared with Indian-English accent when speaking the sentence "It's also very valuable."

Table 1: CommonAccent results on the synthesized speech with more accent prompt types: Chinese-English accent and Korean-English accent.

Prompt type	CommonAccent
Chinese-English accent	96%
Korean-English accent	98%

Table 2: Comparison of the accent conversion quality on accent speech without phoneme changes. CommonAccent metric shows the percentage of being predicted as general American-English.

Framework	CommonAccent
Proposed(w. semantic conversion)	95%
Generative model(w.o semantic conversion)	25%

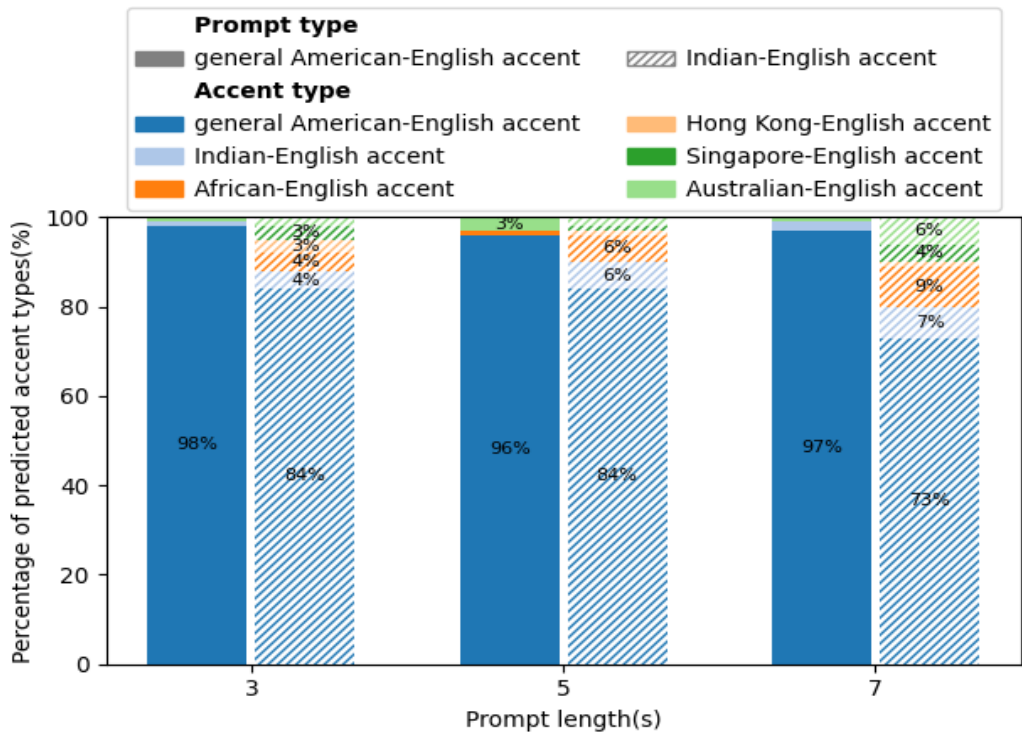


Figure 2: Accent classification distribution versus prompt length in two accent prompt types: general American-English and Indian-English accent.

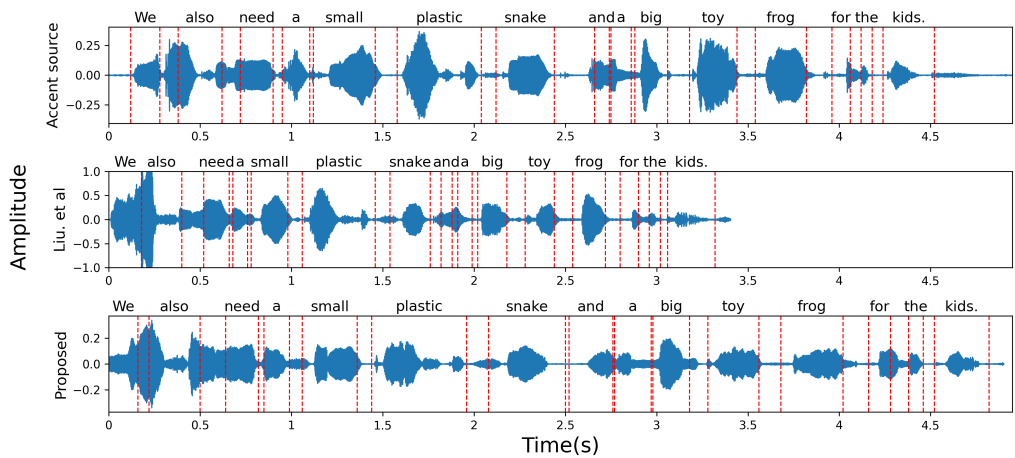


Figure 3: Comparison of the rhythm of the converted speech.