A Pilot Study of Applying Sequence-to-Sequence Voice Conversion to Evaluate the Intelligibility of L2 Speech Using a Native Speaker's Shadowings

Haopeng Geng, Daisuke Saito, Nobuaki Minematsu
Graduate School of Engineering, The University of Tokyo
E-mail: {kevingenghaopeng, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

***Abstract*—Utterances by L2 speakers can be unintelligible due to mispronunciation and improper prosody. In computer-aided language learning systems, textual feedback is often provided using a speech recognition engine. However, an ideal form of feedback for L2 speakers should be so fine-grained that it enables them to detect and diagnose unintelligible parts of L2 speakers' utterances. Inspired by language teachers who correct students' pronunciation through a voice-to-voice process, this pilot study utilizes a unique semi-parallel dataset composed of non-native speakers' (L2) reading aloud, shadowing of native speakers (L1) and their script-shadowing utterances. We explore the technical possibility of replicating the process of an L1 speaker's shadowing L2 speech using Voice Conversion techniques, to create a virtual shadower system. Experimental results demonstrate the feasibility of the VC system in simulating L1's shadowing behavior. The output of the virtual shadower system shows a reasonable similarity to the real L1 shadowing utterances in both linguistic and acoustic aspects[1].**

I. Introduction

Feedback from native speakers or language teachers is essential in second language acquisition (SLA), as it can improve the communicative skills of L2 speakers. With the assistance of Large Language Models (LLM) agents, it has become easier for L2 speakers to have their writing proof-read with detailed suggestions. However, L2 speakers still face challenges in receiving word-by-word feedback on their speaking performance, especially when educational resources are limited.

In such cases, Computer-Aided Language Learning (CALL) has proven to be effective. For example, an Automatic Speech Recognition (ASR) engine can generate transcripts of an L2 speaker's utterances, helping them understand how the system processes their speech and identify any unintelligible words. Recent advances in ASR, leveraging LLM and sophisticated feature embedding, have achieved promising performance, with Word Error Rates (WER) below 10% for Global Englishes [1]–[3].

Nevertheless, previous research indicates that relying solely on ASR for speech intelligibility evaluation is not suitable for providing feedback to L2 speakers [4]. Textual feedback alone is often insufficient for L2 speakers to identify unintelligible segments, especially in spontaneous situations, as the system has no access to the ground truth reference which is considered to exist only in the speaker's mind. Moreover, recent ASR systems are specifically designed to predict, or even over-speculate, what speakers actually said, which is contrary to the purpose of CALL systems. [5], [6]. Besides, research indicates that L2 speakers often overestimate their speech intelligibility [7], [8] and replaying their utterances is generally less effective than anticipated in helping them recognize unintelligible parts of their speech. In the worst case, replaying L2 speakers' speech may fossilize their pronunciation.

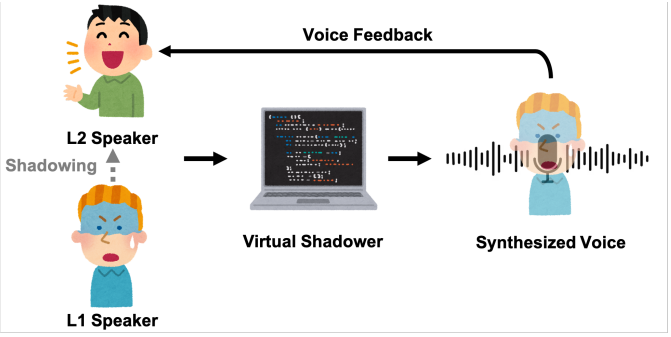

Fig. 1: The concept of the virtual shadower, which simulates the shadowing behaviors of an L1 rater hearing a given L2 speech for the first time. Stuttering or inarticulate production of speech may occur due to listening disfluencies.

How can we expose those unintelligible parts in L2 speech? Native speaker shadowing, wherein an L1 speaker repeats an L2 utterance with as short a delay as possible while listening, is proposed as an advanced method for evaluating L2 speech intelligibility [9]–[11]. In this approach, the L1 speaker immediately repeats what they hear in the given L2 speech using their own accent. Any stuttering or delivery of unrelated words, known as shadowing disfluencies, indicates where listening breakdowns occur. Consequently, shadowing by the L1 speaker highlights the problematic parts of the L2 speech, enabling L2 speakers to know where their speech may be unclear or difficult to understand. However, it is impractical to provide a real shadower for every L2 speaker. In this study, as shown in Figure 1, we aim to develop a virtual shadower system that simulates the process of an L1 speaker shadowing L2 speech

[1]Audio samples are available at: https://secondtonumb.github.io/publication_demo/APSIPA_2024/index.html

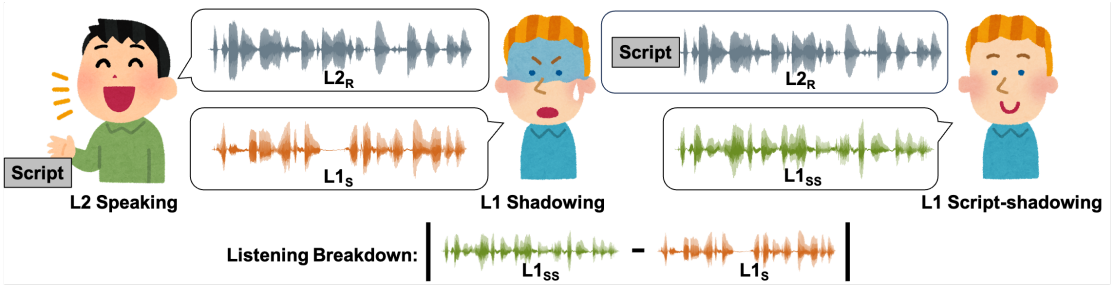

Fig. 2: The proposed shadowing technique aims to identify unintelligible parts in an L2 speaker's reading aloud utterances ($L2_R$). In this approach, $L1_{S1}$ represents a native speaker's initial shadowing, while $L1_{SS}$ denotes the script-shadowing by the native speaker. By calculating the distance between $L1_{S1}$ and $L1_{SS}$, it is possible to pinpoint the native speaker's listening breakdowns, which correspond to the unintelligible parts in the $L2_R$ as well.

(L1-shadowing-L2), thereby offering more comprehensible and constructive feedback for L2 speakers.

Our research contributions are as follows:

- We examine the similarity between an L1 shadower's behavior and Voice Conversion (VC) alignment, utilizing this similarity to construct a virtual L1-shadowing-L2 system, which simulates an L1 speaker's shadowing L2 speech. To the best of our knowledge, this is the first work to apply VC to the L2 intelligibility task.
- We examine the feasibility of using L1 shadowing speech data, a form of semi-parallel data, to develop a VC model. Experimental evaluations on both linguistic and acoustic aspects demonstrate reasonable feasibility in constructing an L1-shadowing-L2 system with Seq2Seq VC.

II. Research Background

A. Native speaker shadowing

Fine-grained annotation of the intelligibility of L2 speech is such a challenging task that it requires specialist knowledge in phonetics for phoneme-level labeling. To address this problem, [10] proposed a two-stage reverse form [2] of shadowing to identify unintelligible parts in L2 utterances. As shown in Figure 2, an L1 speaker first shadows a given L2 utterance alone ($L1_{S1}$). During this process, unintelligible parts result in stuttering or inarticulate production. Following $L1_{S1}$, the L1 speaker performs script-shadowing ($L1_{SS}$), where the L2 script is presented visually along with the L2 utterance. By comparing $L1_{S1}$ and $L1_{SS}$ using Dynamic Time Warping (DTW), we obtain sequential data on broken articulation—shadowing disfluencies that indicate listening disfluencies—based on the distance between the two.

Previous research has shown that alignment based on phonetic posteriorgrams (PPG) performs better than acoustic features such as MFCC or Mel-Cepstrum in assessing pronunciation quality [12]. Additionally, PPG-based DTW between $L1_{S1}$ and $L1_{SS}$ can derive word, syllable, and phoneme-level annotations of intelligibility, as also shown in [11].

[2]In the field of SLA, shadowing typically involves learners repeating L1 speech to enhance their listening skills. However, in our study, learners are shadowed by L1 raters. Thus, we refer to our method as reverse shadowing.

B. Seq2Seq voice conversion

Conventional VC aims to change non-/para-linguistic features while preserving the linguistic content of input speech. However, with the advent of end-to-end architectures and self-supervised speech representations, recent studies on Seq2Seq VC enable a more robust mapping of sophisticated latent features between source and target voices. This includes tasks such as Speaking Style Conversion [13], Voice Emotion Conversion [14], and Foreign Accent Conversion [15], [16].

In [17], the authors proposed the Voice Transformer Network (VTN) that utilizes a Transformer-based Seq2Seq model and employs mel-spectrograms as input features [18]. In [19], the authors modified the input acoustic feature to linguistic representation, resulting in a significant reduction of accentedness. To address the issues of inaccurate duration prediction and repetitive artifacts caused by the auto-regressive nature of transformer models, a non-autoregressive model was proposed using a conformer structure to mitigate these problems [20], [21]. Recent works have further improved upon this by using a Monotonic Alignment Search (MAS) and joint vocoder training, outperforming non-autoregressive models in duration and prosody [22]. Furthermore, a more robust Automatic Alignment Search (AAS) method, derived from a Text-to-Speech (TTS) system, has been developed to address challenges in low-resource settings [23], [24].

In the following sections, these recent advancements in Seq2Seq VC are effectively adapted to implement our purposed L1-shadowing-L2 system.

III. Experimental Settings

A. Representing listening breakdowns with attention alignment failure

Shadowing is a voice-to-voice process in which the listener's instantaneous identification of words is crucial. However, smooth shadowing does not always occur, as listening breakdowns can happen when an L1 listener encounters unintelligible segments while shadowing L2 speech [25]. To explore the potential of VC to build virtual shadowers, Figure 3 shows our pilot experiment on Seq2Seq VC using semi-parallel data, $L2_R$ and $L1_{S1}$. We observe that the decoding path in the $L2_R$-$L1_{S1}$ VC system closely resembles the PPG-DTW editing path. This

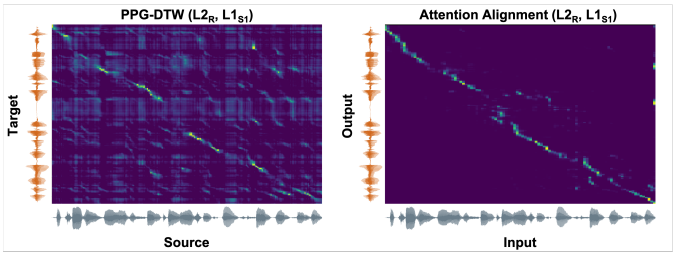

Fig. 3: The similarity of attention alignment to PPG-DTW is illustrated. The left figure shows the attention alignment observed in the inference phase of converting $L2_R$ to $L1_{S1}$, while the right figure shows the PPG-DTW path between $L2_R$ and $L1_{S1}$. Both figures exhibit a prominent diagonal contour.

inspires us to explore whether a VC system can be extended to explicitly represent L2 unintelligibility. If alignment failure (weak or faded-out attention) indicates listening breakdowns, we suggest that such a virtual shadower can be developed to resemble a real human L1 shadower.

B. Experimental setups

1) VC models: To ensure that the VC system's alignment failure is attributed to data mismatch rather than the model's limited alignment capabilities (as discussed in Section III-A), we evaluated two VC models in this study. Alongside the baseline Transformer-based model, VTN [26], we tested a robust alignment method, AAS, proposed in AAS-VC [23].

2) Data preparation: In this study, we utilized a reverse form of shadowing dataset mentioned in [11], [12]. With different degrees of accent, 225 Japanese English speakers' reading-aloud utterances $L2_R$ were collected. We then recruited one male native English speaker who shadowed and script-shadowed all these recordings as $L1_{S1}$ and $L1_{SS}$ respectively. Although each recording session lasted for more than 30 seconds, to avoid potential alignment failure due to lengthy input, we trimmed the original recordings sentence by sentence using forced alignment. This process resulted in 2,695 triplets of $\{L2_R, L1_{S1}, L1_{SS}\}$. Each dataset comprised an average of 3.9 hours of valid phonation, with the average duration of each sentence being 5.0 seconds. 300 utterances were selected as test sets.

3) Source-target selection: To determine the optimal setting for an L1 virtual shadower simulator, we designed three different source-target pairings:

- $L2_R$-$L1_{S1}$: This is the most straightforward approach for building a virtual shadower, as the training target is $L1_{S1}$ itself. However, it is also the most technically challenging setting, as both the linguistic context and speaker identity differ.
- $L2_R$-$L1_{SS}$: This is the typical setting for parallel-data VC. Since the source and target are intended to speak the same content, we expect this setting to reveal the pronunciation distance between L2 and L1.
- $L1_{SS}$-$L1_{S1}$: This unique training setting of the reverse shadowing dataset focuses on the listening breakdowns of the L1 shadower. We expect that this setting will reveal the disfluencies exhibited during the shadowing behavior of L1.

4) Feature embedding and pretrained models for decoding: Speaker-independent features enable stable convergence in our many-to-one VC setting. For feature embedding, we used the original PPG-like bottleneck feature extractor designed by [27]. The framewise dimension of latent feature is 144, and the features were normalized to zero mean and unit variance for embedding.

As for the PPG-to-Spec decoding, we trained a single-speaker PPG-to-Spec decoder following the implementation described in [28], where utterances were collected from the given L1 speaker.

For the vocoder, we employed Parallel-WaveGAN [29], which was trained on joint $L1_{S1}$ and $L1_{SS}$. The input Mel spectrogram has 80 frequency bins with a hop size of 256, and the target sample rate is 16 kHz.

Note that for VTN, both the PPG-to-Spec decoder and vocoder were utilized in the decoding phase. For AAS-VC, we followed the original implementation to map the source PPG to the target Mel-spectrogram, which generates the target's prosodic features better as previous study discussed [23].

IV. Experimental Evaluation

A. Metrics for evaluation

To objectively evaluate our virtual shadower, we focused on two key aspects. First, the linguistic content of the virtual shadower should closely match that of the L1 human shadower. Second, the acoustic similarity of the virtual shadower should resemble that of the L1 shadower. In this pilot study, we assessed the linguistic output of our virtual shadower only holistically. For the acoustic aspect, we calculated the segmental and prosodic similarities between the virtual shadower and the L1 shadower. Before presenting the evaluation results for these two aspects, the following section will show the ASR results for $L1_{S1}$, $L1_{SS}$, and $L2_R$, which are necessary for subsequent discussions.

B. Ablating effects of vocoding module

We first applied ASR to all testing data using a reputable and advanced ASR system[3] [31]. In this study, the ASR results reflect the expressions of the L1 shadower during the shadowing sessions (S1 and SS) and the L2 learner during the recording session (R). Inevitably, the ASR result may include errors, which are quantified using the CER/WER by comparing the results with the reference text used during the recording of $L2_R$. As shown in † of Table I, the differences in CER/WER values between $L1_{S1}$ and $L1_{SS}$ reveal the linguistic variations introduced by shadowing. The $L2_R$ utterances, characterized by a strong accent, consequently display even higher CER/WER compared to the L1 speech.

To eliminate propagation errors potentially caused by VC decoders, we performed a framewise PPG-based VC that only converts the speaker identity based on the implementation in

[3]https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

TABLE I: CER/WER and S1-CER/WER results evaluated on original recordings, PPG-VC, VTN and AAS-VC outputs.

	Model	Training Source	Training Target	Testset	CER↓	WER↓	S1-CER ↓	S1-WER↓
†	-	-	-	$L1_{S1}$	6.83	13.71	-	-
†		-	-	$L1_{SS}$	0.85	3.60	-	-
†		-	-	$L2_R$	10.04	20.91	-	-
$*$	PPG-VC [30]	$L1_{spk}$	$L1_{spk}$	$L1_{S1}$	8.00	15.90	-	-
$*$		$L1_{spk}$	$L1_{spk}$	$L1_{SS}$	1.90	5.70	-	-
$**$		$L2_{spk}$	$L2_{spk}$	$L2_R$	21.67	38.83	-	-
$***$		$L2_{spk}$	$L1_{spk}$	$L2_R$	21.81	39.23	-	-
‡	VTN [17]	$L2_R$	$L1_{S1}$	$L2_R$	22.64	37.29	22.41	36.15
‡		$L2_R$	$L1_{SS}$	$L2_R$	**19.47**	**34.53**	19.28	33.53
‡		$L1_{SS}$	$L1_{S1}$	$L2_R$	32.21	51.24	32.08	50.56
‡	AAS-VC [23]	$L2_R$	$L1_{S1}$	$L2_R$	21.60	38.68	21.53	37.98
‡		$L2_R$	$L1_{SS}$	$L2_R$	**20.44**	**36.74**	19.98	35.97
‡		$L1_{SS}$	$L1_{S1}$	$L2_R$	26.49	46.76	26.49	46.50

TABLE II: Mel Cepstral Distortion (MCD), F0 Root Mean Square Error (F0RMSE), F0 Correlation (F0CORR), and Absolute Duration Difference (DURR) on PPG-VC, VTN and AAS-VC. PPG-VC converts multiple L2 speaker identities into the target L1 speaker to minimize the influence of speaker identity on acoustic features.

Model	Training Source	Training Target	Testset	Reference	MCD↓	F0RMSE↓	F0CORR↑	DURR↓
-	-	-	$L2_R$	$L1_{S1}$	12.84	89.65	0.084	0.337
	-	-	$L1_{SS}$	$L1_{S1}$	6.62	35.24	0.385	0.350
PPG-VC	$L2_R$	$L1_{spk}$	$L2_R$	$L1_{S1}$	8.94	47.27	0.127	0.339
	$L2_R$	$L1_{spk}$	$L2_R$	$L1_{SS}$	8.97	47.03	0.117	0.361
VTN	$L2_R$	$L1_{S1}$	$L2_R$	$L1_{S1}$	7.53	40.35	0.239	0.552
	$L2_R$	$L1_{SS}$	$L2_R$	$L1_{S1}$	7.47	39.76	0.253	0.575
AAS-VC	$L2_R$	$L1_{S1}$	$L2_R$	$L1_{S1}$	7.39	38.76	0.240	0.515
	$L2_R$	$L1_{SS}$	$L2_R$	$L1_{S1}$	7.34	37.78	0.259	0.401

[30] for analysis synthesis (ANA-SYN). The PPG decoder and vocoder described in Section III-B4 were utilized. The CER/WER of the ANA-SYN speech is presented in $*$, $**$ and $***$ of Table I. $*$ and $**$ show the VC results in which the identity of the input speaker was maintained, while $***$ shows the conversion of $L2_{spk}$ to $L1_{spk}$. The minimal variations in CER/WER for $L1_{S1}$ and $L1_{SS}$ suggest that the embedding features do not distort linguistic information. However, a larger degradation in CER/WER for ANA-SYN $L2_R$ is noted in $**$, which is considered to be due to the inevitable noises present in $L2_R$. Note that $L2_R$ recordings were collected separately in various environments using personal devices, making it difficult to ensure standardized acoustic quality. Nevertheless, the very small difference in CER/WER after speaker normalization to $L1_{spk}$, shown in $***$, indicates that these recording disturbances do not significantly affect the decoder's performance.

C. Linguistic similarity of virtual shadower

Since the proposed virtual shadower aims to construct potentially corrupted speech generated by the L1 shadower, the output is different from what the L2 learner actually intends, but expects to be more similar to what the shadower said. While the conventional WER is calculated as follows:

$$\text{WER} = \frac{S + I + D}{N_R}, \tag{1}$$

we introduce S1-WER to evaluate the linguistic similarity between the conversion results and $L1_{S1}$ based on their ASR results.

$$\text{S1-WER} = \frac{S + I + D}{N_{S1}}, \tag{2}$$

where S, I, and D are substitution, insertion, and deletion errors of the converted results, respectively. N_R represents the word count of L2 learners' handcrafted scripts, while N_{S1} is the word count of the recognition hypothesis of $L1_{S1}$ conducted by the mentioned ASR system. Lower S1-WER will represent higher linguistic similarity between the converted speech and the shadowing speech. S1-CER is the character-based error rate, calculated in a similar way to S1-WER.

As shown in ‡ of Table I, compared to $***$, the reduction of WER on $L2_R$-$L1_{S1}$ and $L2_R$-$L1_{SS}$ demonstrates the accent reduction capability of Seq2Seq VC. The best CER/WER are achieved using the $L2_R$-$L1_{SS}$ training pair, as the L1 shadower fluently repeated the content intended by the L2 learner by viewing the learner's script, which is a beneficial condition for VC training. For the $L2_R$-$L1_{S1}$ which is the most straightforward approach to achieve virtual shadowing

by the L1 speaker, the results are reasonable, as the semi-parallel setting brings high barriers for feature mapping from a mispronounced segment to the correct segment. In the $L1_{SS}$-$L1_{S1}$ setting, the relatively high CER/WER indicates that the model struggles to accurately capture the unintelligibility of the L2 speakers. This is likely because $L2_R$'s features are considered out-of-domain.

Regarding S1-CER/WER, although the differences are not significant, all proposed models have lower or equal S1-CER/WER values compared to CER/WER. This indicates that Seq2Seq VC systems have the potential to effectively map semi-parallel linguistic features, such as $L2_R$-$L1_{S1}$ in this study. Particularly, $L1_{SS}$ targeting models outperform all the other training settings, indicating the possibility of building a virtual shadowing system utilizing parallel data only.

D. Acoustic similarity of virtual shadower

Table II illustrates the acoustic similarity between $L1_{S1}$ and the converted speech. Compared to the $L1_{spk}$-normalized $L2_R$ conducted by PPG-VC, the better metrics in MCD and F0 demonstrate the prosody reconstruction ability of the proposed virtual shadower. In particular, AAS-VC, benefiting from its superior alignment ability, shows better prosody similarity, especially in F0 and duration.

Interestingly, although the model was not trained with $L1_{S1}$, the results showed that the output of $L1_{SS}$-targeting model closely resembles the acoustic similarity of the actual $L1_{S1}$. This indicates the potential of using parallel data directly to establish a virtual shadowing system, which supports the conclusion mentioned in Section IV-C.

V. Conclusions and future works

A. Conclusions

In this study, we first proposed the concept of a virtual shadower, which simulates the behavior of an L1 listener who instantly repeats what an L2 speaker utters while listening to it. By reflecting the L1 listener's immediate understanding, the virtual shadower aims to highlight linguistic ambiguities in L2 speech that the speaker may not notice. Experiments conducted on two Seq2Seq VC systems with various training data settings from the $L2_R$, $L1_{S1}$ and $L1_{SS}$ triplet demonstrated the potential of parallel and semi-parallel training from $L2_R$ to $L1_{S1}$/$L1_{SS}$.

B. Future works

1) Generalization ability: Shadowing behavior varies among listeners with different language backgrounds. This study involved only Japanese learners of English and a single rater. Future experiments will involve participants who speak different languages and multiple raters to explore the capability and adaptability of Seq2Seq VC as more flexible and general virtual shadowers.

2) Integrating $L1_{SS}$-$L1_{S1}$ information for virtual shadowing: In this study, we acknowledge and examine the challenge of directly mapping $L2_R$-$L1_{S1}$ as they are both linguistically and acoustically different. However, the promising result with $L2_R$-$L1_{SS}$ encourage us to combine information derived from $L1_{SS}$-$L1_{S1}$. We believe the distance between $L1_{SS}$-$L1_{S1}$ has not been fully utilized yet. For instance, using a loss function derived from PPG-DTW ($L1_{S1}, L1_{SS}$) is considered a more promising and interpretable approach. Additionally, exploring the cascading of $L2_R$-$L1_{SS}$ with $L1_{SS}$-$L1_{S1}$ could yield valuable insights to convert $L2_R$ to $L1_{S1}$.

3) Self-supervised speech representations with prosodic feature considered: Although linguistically-rich embedding like PPG perform well in foreign accent conversion tasks, PPG lacks prosodic features such as intonation and rhythm, which often affect listeners' comprehension. Self-supervised speech representations have shown great potential in reconstructing unintelligible speech, such as electrolaryngeal or dysarthric speech [32]. We will explore the feasibility of using latent features from models like HuBERT [33] and WavLM [34], which are known to include both linguistic and prosodic features.

4) Pedagogical Assessment of virtual shadower: Our virtual shadower offers feedback by pinpointing segments of learners' speech that are unintelligible to listeners. Unlike conventional feedback, which is typically provided as scores or written comments, our feedback is delivered as immediate vocal response. This approach offers more realistic feedback, which we expect will enhance learners' motivation more effectively than traditional methods.

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.

[2] S. Wills, Y. Bai, C. Tejedor-García, C. Cucchiarini, and H. Strik, "Automatic Speech Recognition of Non-Native Child Speech for Language Learning Applications," in *Proc. SLATE*, 2023, 7:1–7:8.

[3] M. P. Y. Chan, J. Choe, A. Li, Y. Chen, X. Gao, and N. Holliday, "Training and typological bias in ASR performance for world Englishes," in *Proc. Interspeech*, 2022, pp. 1273–1277.

[4] A. Neri, C. Cucchiarini, and H. Strik, "Effective feedback on l2 pronunciation in asr-based call," 2001.

[5] Y. Weng, S. S. Miryala, C. Khatri, *et al.*, "Joint contextual modeling for asr correction and language understanding," in *Proc. ICASSP*, 2020, pp. 6349–6353.

[6] R. Ma, M. Qian, P. Manakul, M. Gales, and K. Knill, "Can generative large language models perform asr error correction?" *arXiv preprint arXiv:2307.04172*, 2023.

[7] T. M. Derwing and M. J. Munro, "Accent, intelligibility, and comprehensibility: Evidence from four l1s," *Studies in Second Language Acquisition*, vol. 19, no. 1, pp. 1–16, 1997.

[8] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language learning*, vol. 45, no. 1, pp. 73–97, 1995.

[9] Y. Inoue, S. Kabashima, D. Saito, N. Minematsu, K. Kanamura, and Y. Yamauchi, "A Study of Objective Measurement of Comprehensibility through Native Speakers' Shadowing of Learners' Utterances," in *Proc. Interspeech*, 2018, pp. 1651–1655.

[10] Z. Lin, R. Takashima, D. Saito, N. Minematsu, and N. Nakanishi, "Shadowability annotation with fine granularity on l2 utterances and its improvement with native listeners' script-shadowing.," in *Proc. Interspeech*, 2020, pp. 3865–3869.

[11] C. Zhu, R. Hakoda, D. Saito, N. Minematsu, N. Nakanishi, and T. Nishimura, "Multi-granularity annotation of instantaneous intelligibility of learners' utterances based on shadowing techniques," in *Proc. ASRU*, 2021, pp. 1071–1078.

[12] J. Yue, F. Shiozawa, S. Toyama, *et al.*, "Automatic Scoring of Shadowing Speech Based on DNN Posteriors and Their DTW," in *Proc. Interspeech*, 2017, pp. 1422–1426.

[13] G. Maimon and Y. Adi, "Speaking style conversion in the waveform domain using discrete self-supervised units," in *EMNLP*, H. Bouamor, J. Pino, and K. Bali, Eds., 2023, pp. 8048–8061.

[14] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *Proc. ICASSP*, 2021, pp. 920–924.

[15] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Converting foreign accent speech without a reference," *IEEE/ACM TASLP*, vol. 29, pp. 2367–2381, 2021.

[16] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent conversion using phonetic posteriorgrams," in *Proc. ICASSP*, 2018, pp. 5314–5318.

[17] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice Transformer Network: Sequence-to-Sequence Voice Conversion Using Transformer with Text-to-Speech Pretraining," in *Proc. Interspeech*, 2020, pp. 4676–4680.

[18] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 6000–6010.

[19] W.-C. Huang and T. Toda, "Evaluating Methods for Ground-Truth-Free Foreign Accent Conversion," in *Proc. APSIPA ASC*, 2023.

[20] T. Hayashi, W.-C. Huang, K. Kobayashi, and T. Toda, "Non-autoregressive sequence-to-sequence voice conversion," in *Proc. ICASSP*, 2021, pp. 7068–7072.

[21] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[22] T. Okamoto, T. Toda, and H. Kawai, "E2e-s2s-vc: End-to-end sequence-to-sequence voice conversion," in *Proc. Interspeech*, 2023, pp. 2043–2047.

[23] W.-C. Huang, K. Kobayashi, and T. Toda, "Aas-vc: On the generalization ability of automatic alignment search based non-autoregressive sequence-to-sequence voice conversion," *arXiv preprint arXiv:2309.07598*, 2023.

[24] K. J. Shih, R. Valle, R. Badlani, A. Lancucki, W. Ping, and B. Catanzaro, "Rad-tts: Parallel flow-based tts with robust alignment learning and diverse synthesis," in *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

[25] Y. Hamada, "The effectiveness of pre-and post-shadowing in improving listening comprehension skills," *The Language Teacher*, vol. 38, no. 1, pp. 3–10, 2014.

[26] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Pretraining techniques for sequence-to-sequence voice conversion," *IEEE/ACM TASLP*, vol. 29, pp. 745–755, 2021.

[27] S. Liu, Y. Cao, N. Hu, D. Su, and H. Meng, "Fastsvc: Fast cross-domain singing voice conversion with feature-wise linear modulation," in *Proc. ICME*, 2021, pp. 1–6.

[28] W.-C. Huang, S.-W. Yang, T. Hayashi, and T. Toda, "A Comparative Study of Self-Supervised Speech Representation Based Voice Conversion," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1308–1318, 2022.

[29] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, 2020, pp. 6199–6203.

[30] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE/ACM TASLP*, vol. 29, pp. 1717–1728, 2021.

[31] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[32] L. P. Violeta, D. Ma, W.-C. Huang, and T. Toda, "Pretraining and adaptation techniques for electrolaryngeal speech recognition," *IEEE/ACM TASLP*, vol. 32, pp. 2777–2789, 2024.

[33] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM TASLP*, vol. 29, pp. 3451–3460, 2021.

[34] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.