Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

A Perception-Based L2 Speech Intelligibility Indicator: Leveraging a Rater's Shadowing and Sequence-to-sequence Voice Conversion

Haopeng Geng[1], *Daisuke Saito*[1], *Nobuaki Minematsu*[1]

[1]Graduate School of Engineering, The University of Tokyo, Japan
{kevingenghaopeng, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

Abstract

Evaluating L2 speech intelligibility is crucial for effective computer-assisted language learning (CALL). Conventional ASR-based methods often focus on native-likeness, which may fail to capture the actual intelligibility perceived by human listeners. In contrast, our work introduces a novel, perception-based L2 speech intelligibility indicator that leverages a native rater's shadowing data within a sequence-to-sequence (seq2seq) voice conversion framework. By integrating an alignment mechanism and acoustic feature reconstruction, our approach simulates the auditory perception of native listeners, identifying segments in L2 speech that are likely to cause comprehension difficulties. Both objective and subjective evaluations indicate that our method aligns more closely with native judgments than traditional ASR-based metrics, offering a promising new direction for CALL systems in a global, multilingual contexts.
Index Terms: computer assisted pronunciation training, speech shadowing, speech intelligibility, voice conversion

1. Introduction

Non-native speakers are often encouraged to emulate native-like pronunciation; however, achieving a native-like accent can be time-consuming and may not efficiently improve speech intelligibility. Many studies indicate that mispronunciations or accents do not necessarily result in unintelligible speech. For example, [1] found that only 16% of word-level unintelligibility in an English spoken test can be attributed to obvious mispronunciations, and experimental evaluations in [2] show that even non-native accents, such as Swedish and German-accented English, yield higher recognition accuracy using OpenAI's Whisper compared to British-accented English.

According to [3], approximately 1.14 billion non-native speakers account for 75% of the global English-speaking community. However, mainstream Automatic Speech Recognition (ASR) systems are typically built using native-accented corpora or the most prevalent accents, which can lead to biased performance that often results in higher error rates and misrecognition on non-native or less common accents. Additionally, recent ASR systems are specifically designed to predict—or even overspeculate on—what speakers actually said, which conflicts with the goals of computer-assisted language learning (CALL) systems [4, 5]. From an educational perspective, language learners require feedback on whether their accents hinder communication with speakers from diverse backgrounds rather than simply striving for native-like pronunciation. Therefore, it is essential to develop methods that capture not only native-like pronunciation but also the actual intelligibility as perceived by real human listeners.

How can we effectively identify unintelligible parts in L2 speech? One promising method is to measure native (L1) listeners' responses by asking them to shadow L2 utterances—where an L1 speaker immediately repeats the L2 speech with minimal delay. This approach, as explored in previous research [6–8], has proven effective at revealing where listening breakdowns occur, highlighting problematic segments in the L2 speech, and providing valuable feedback for learners.

Inspired by this approach, our study aims to utilize the L1 rater's shadowing process to build an unintelligibility indicator based on the rater's perception. Our contributions are as follows:

1. We propose a customized speech unintelligibility indicator that leverages native rater's shadowing data. Unlike conventional methods that focus solely on native-like pronunciation, our approach aims to capture the perceptual cues that native listeners actually use to assess intelligibility[1].
2. Our method incorporates a state-of-the-art sequence-to-sequence (seq2seq) voice conversion framework, utilizing its alignment mechanism and acoustic feature reconstruction module to simulate the cognitive process of native listeners in identifying unintelligible segments.
3. We introduce a multi-task learning strategy that jointly optimizes speech reconstruction and disfluency detection using auxiliary loss functions. Experimental evaluations demonstrate that our approach aligns more closely with native raters' perceptions than mainstream ASR, revealing the potential for delivering personalized feedback in CALL systems.

2. Research background

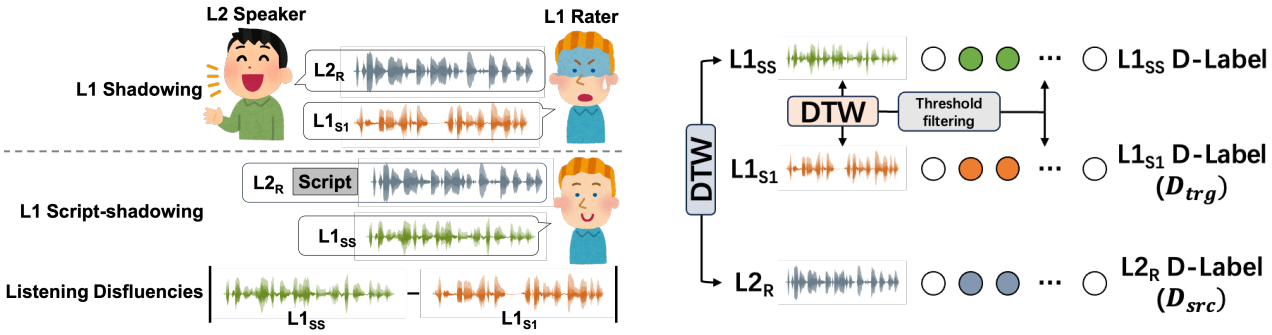
2.1. L1 rater's shadowing

Assessing L2 speech intelligibility requires fine-grained annotation, yet this process is challenging due to the need for expert-level phonetic knowledge for phoneme-level labeling. To address this challenge, [7] proposed a two-stage reverse shadowing approach to identify unintelligible segments in L2 utterances[2]. As illustrated in Fig 1a, an L1 speaker first shadows a given L2 utterance alone ($L1_{S1}$), during which listening breakdowns manifest as stuttering or inarticulate speech. Subsequently, the L1 speaker performs script-shadowing ($L1_{SS}$), where the L2 script is displayed while the corresponding L2 utterance is played acoustically. By comparing $L1_{S1}$ and $L1_{SS}$

[1]Samples are available at https://secondtonumb.github.io/publication_demo/IS_2025/

[2]In second language acquisition (SLA), shadowing usually involves learners repeating L1 speech to improve their listening skills. In our study, however, L1 raters shadow L2 speech—a process we refer to as reverse shadowing.

10.21437/Interspeech.2025-615



(a) *L1 rater shadowing for L2 speech intelligibility assessment.* (b) *Mapping disfluency label from a rater's shadowing to L2 speech.*

Figure 1: *An overview of the L1 rater's shadowing technique for immediate L2 speech intelligibility assessment. The proposed approach is designed to identify unintelligible segments in the L2 speaker's read-aloud utterances ($L2_R$).*

using dynamic time warping (DTW), this method effectively captures sequential disfluency patterns that indicate listening breakdowns. Prior research has demonstrated that phonetic posteriorgram (PPG)-based DTW alignment between $L1_{S1}$ and $L1_{SS}$ can effectively yield unintelligibility annotations at the word, syllable, and phoneme levels [8, 9].

Fig 1b illustrates the process for obtaining frame-wise disfluency labels on $L2_R$ using the two-stage reverse shadowing approach. First, DTW is applied between the first shadowing speech $L1_{S1}$ and the script-shadowing speech $L1_{SS}$ to capture listening breakdowns in $L1_{S1}$. Then, by thresholding the edit distance between $L1_{S1}$ and $L1_{SS}$, a disfluency label (**D**-Label) can be derived for both $L1_{SS}$ and $L1_{S1}$. Finally, another DTW alignment between $L1_{SS}$ and $L2_R$ is used to map the **D**-Label onto $L2_R$, thereby identifying the unintelligible frames.

2.2. Seq2seq voice conversion with self supervised speech representations

Conventional voice conversion (VC) seeks to modify non-/paralinguistic features while preserving the linguistic content of the input speech. With the advent of end-to-end architectures, recent seq2seq VC models have achieved robust mappings between sophisticated source and target features. For example, [10] introduced a transformer-based VC model capable of sequentially mapping speech with variable durations. Meanwhile, limitations of autoregressive models, such as inaccurate duration prediction and repetitive artifacts, was addressed by proposing a non-autoregressive model using a conformer structure in [11]. More recent approaches combine Monotonic Alignment Search (MAS) with joint vocoder training [12, 13], which enhances the precision of duration modeling and improves prosody synthesis by reducing alignment errors.

Advances in self-supervised learning have further enhanced VC by improving embedding features. Self-supervised speech representations (S3Rs) such as Wav2vec 2.0 [14], HuBERT [15], and WavLM [16] have set new benchmarks across diverse speech tasks [17]. Consequently, self-supervised speech representation-based VC (S3R-VC) has demonstrated notable gains in both linguistic accuracy and acoustic quality, while also proving effective for processing unintelligible speech such as electrolaryngeal speech enhancement [18, 19]. Moreover, the combination of improved architectures and advanced embeddings has enabled novel mappings between source and target speech—such as shadowing-like speech generation [20]—even when the linguistic content of the two voices is not fully parallel.

In the following sections, we describe how we adapt these recent advances in seq2seq VC, particularly the enhanced alignment module, for both speech reconstruction and disfluency detection.

3. Proposed Methods

In this section, we present two approaches for evaluating L2 speech intelligibility based on seq2seq VC. The first approach leverages alignment failure to mimic rater perception breakdowns, thereby identifying segments of unintelligibility. The second approach employs a multi-task learning strategy in a Seq2Seq VC model, where additional loss functions guide the model to detect unintelligibility.

3.1. Using alignment breakdowns to replicate rater's perception breakdowns

In parallel TTS and VC research, the MAS technique mentioned in Sec. 2.2 has been widely adopted [21, 22]. In this process, a soft alignment matrix $\mathcal{A}_{\text{soft}} \in \mathbb{R}^{T_{\text{src}} \times T_{\text{trg}}}$ is generated based on the pairwise similarity between the source and target features. The Viterbi algorithm is then used to compute the hard alignment path between, denoted by $\mathcal{A}_{\text{hard}}$. The alignment is optimized by minimizing the discrepancy between $\mathcal{A}_{\text{soft}}$ and $\mathcal{A}_{\text{hard}}$ using KL-divergence. Specifically, the binary loss is defined as

$$\mathcal{L}_{\text{bin}} = \mathcal{A}_{\text{hard}} \odot \log \mathcal{A}_{\text{soft}} \tag{1}$$

where $\odot$ denotes the Hadamard product and the final alignment loss is combined with the forward sum loss:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{ForwardSum}} + \mathcal{L}_{\text{bin}} \tag{2}$$

This idea is based on the assumption that the source and target are parallel. However, in our scenario, shadowing, a speech-to-speech process through perception, often results in speech from the rater that does not perfectly match the speech from the learner. Consequently, when aligning these semi-parallel data, segments with noticeable differences are expected to misalign. This observation motivates us to treat the alignment module's failure segments as unintelligible segments in $L2_R$ speech. Given the converged non-positive matrix $\log \mathcal{A}_{\text{soft}}$, the frame-wise alignment focus rate s_i can be calculated for each source frame i ($1 \leq i \leq T_{\text{src}}$) as follows:

$$s_i = \mathbb{I}\Big\{ \max_{j \in \{1,\ldots,T_{\text{trg}}\}} \log \mathcal{A}_{\text{soft}}^{(i,j)} < \tau \Big\} \tag{3}$$

where τ is a predefined threshold, and $\mathbb{I}\{\cdot\}$ is the indicator function that returns 1 if the condition is true, indicating that frame i contains segments that lead to perception failure during the rater's shadowing process.

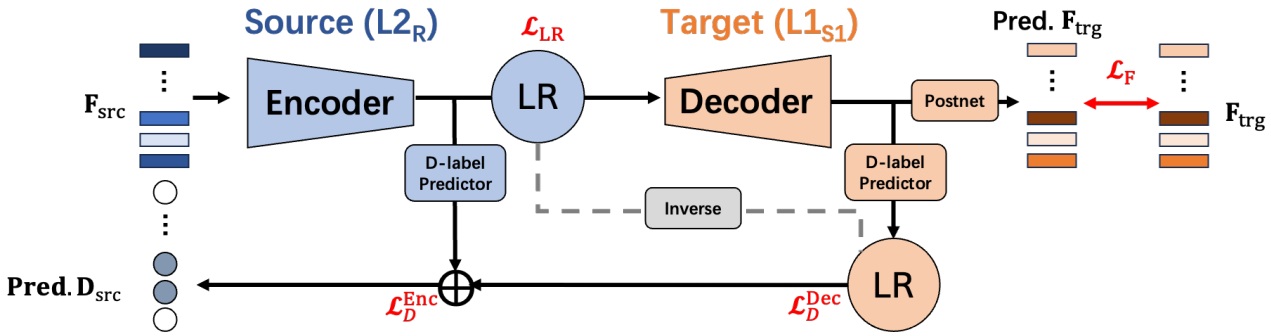


Figure 2: *Proposed multi-task learning for L2 intelligibility prediction based on seq2seq VC.*

3.2. Multi-task learning: seq2seq VC with disfluency detection

Another approach aims to convert $L2_R$ to $L1_{S1}$ while simultaneously detecting unintelligible segments in $L2_R$, is illustrated in Fig 2. In typical seq2seq VC models, let $\mathbf{F}_{\text{src}} \in \mathbb{R}^{T_{\text{src}} \times d}$ and $\mathbf{F}_{\text{trg}} \in \mathbb{R}^{T_{\text{trg}} \times d}$ denote the continuous features of the source and target utterances, respectively. The transformation process in the VC model is defined as follows:

$$\mathbf{h} = \text{Enc}(\mathbf{F}_{\text{src}}) \tag{4}$$

$$\mathbf{h}' = \text{LR}(\mathbf{h}) \tag{5}$$

$$\mathbf{z} = \text{Dec}(\mathbf{h}') \tag{6}$$

$$\mathbf{F}_{\text{trg}} = \text{Postnet}(\mathbf{z}) \tag{7}$$

where the length regulator (LR) maps the source hidden representation $\mathbf{h} \in \mathbb{R}^{T_{\text{src}} \times h}$ to the target representations $\mathbf{h}'$ and $\mathbf{z} \in \mathbb{R}^{T_{\text{trg}} \times h}$.

Directly capturing the positional information of unintelligibility from $\mathbf{F}_{\text{src}}$ is challenging because the source speech, although accented, remains fluent. To tackle this, we propose guiding the model by incorporating additional modules aimed at detecting disfluency on the target side, since the disfluency information is explicit in the rater's shadowing voice. Specifically, we integrate two disfluency label predictors (DLPs), each implemented using a 5-layer CNN, which are positioned after the encoder and decoder, respectively. By leveraging the hidden outputs from both the encoder and decoder, the model can effectively capture disfluency information. This information is then projected back to the source side via an inverse length regulation, serving as a proxy for the source's unintelligibility. The disfluency position information is computed as follows:

$$\mathbf{D}_{\text{src}} = \text{Linear}\left(\text{DLP}(\mathbf{h}) \oplus \text{LR}^{-1}(\text{DLP}'(\mathbf{z}))\right) \tag{8}$$

where LR^{-1} is the inverse length regulation from target to source. Finally, incorporating our proposed DLPs, whose loss functions $\mathcal{L}_{\text{D}}^{\text{Enc}}$ and $\mathcal{L}_{\text{D}}^{\text{Dec}}$ are implemented using the focal loss [23], we define the overall loss as:

$$\mathcal{L}_{\text{VC}} = \mathcal{L}_{\text{F}} + \mathcal{L}_{\text{LR}} + \mathcal{L}_{\text{align}} \tag{9}$$

$$\mathcal{L}_{\text{all}} = \lambda \left(\mathcal{L}_{\text{D}}^{\text{Enc}} + \mathcal{L}_{\text{D}}^{\text{Dec}}\right) + \mathcal{L}_{\text{VC}} \tag{10}$$

where $\mathcal{L}_{\text{F}}$ is the L1 loss between the predicted and target acoustic features, and $\mathcal{L}_{\text{LR}}$ is the loss for predicting the appropriate duration transformation referring [13]. The hyper-parameter λ was set to 10 to balance the training criteria. This combined loss formulation ensures balanced target feature reconstruction and D-label detection for binary unintelligibility detection.

Table 1: *Objective evaluation of the proposed L1 perception-based intelligibility indicator*

Description	Word			Frame
	F1↑	Precision↑	Recall↑	Acc. ↑
ASR-based	28.0	28.8	27.2	—
Alignment-based	27.2	23.0	**33.3**	—
Multi-task:				
$\mathcal{L}_D^{\text{Enc}}$	25.5	25.0	26.0	74.2
$\mathcal{L}_D^{\text{Dec}}$	23.5	23.0	24.0	73.3
$\mathcal{L}_D^{\text{Enc}} + \mathcal{L}_F$	28.6	26.1	31.6	74.3
$\mathcal{L}_D^{\text{Dec}} + \mathcal{L}_F$	24.5	25.5	23.6	**75.7**
$\mathcal{L}_D^{\text{Enc}} + \mathcal{L}_D^{\text{Dec}} + \mathcal{L}_F$	**30.3**	**29.2**	31.4	75.6

4. Experimental Settings

4.1. Dataset

We utilized the shadowing dataset described in [8, 9]. Reading-aloud utterances were collected from 225 Japanese learners of English ($L2_R$) with varying degrees of Japanese accents. An American male English speaker (EN-A) then shadowed and script-shadowed these recordings, producing $L1_{S1}$ and $L1_{SS}$, respectively. In total, 2,695 triplets $\{L2_R, L1_{S1}, L1_{SS}\}$ were prepared. This dataset contained 3.9 hours of valid phonation, with an average duration of 5.0 seconds. Three hundred utterances were selected as the test set, while the remaining utterances were allocated as the training and development sets.

For unintelligibility detection, we applied the method described in Sec. 2.1 to generate frame-wise labels with 10 ms hop size. We then converted the frame-level D-labels to word-level via weighted editing path concatenation, following the approach outlined in [8], since word-level degradation more closely aligns with human perception.

4.2. Feature embedding

In our preliminary experiments, we initially attempted to use HuBERT for both the source and target phases of feature mapping. However, the alignment did not converge as expected based on our observation. Consequently, based on implementations in [20], we adopted the ninth-layer output of the HuBERT-base model on the source side and a PPG-like bottleneck feature (PPG-BNF) from [24] for the target, which yielded the best alignment. The feature dimensions per frame were 768 and 144, respectively.

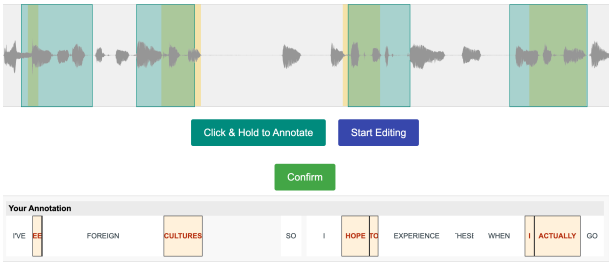


Figure 3: *Overview of annotation system for unintelligibility.*

5. Experimental Evaluation

5.1. Objective evaluation

We present our objective evaluation results in Table 1. Both frame-level and word-level metrics are considered, with the word-level evaluation computed as an F1-score based on the positional detection of unintelligible words in $L2_R$. The conventional approach—relying on an ASR system trained on librispeech dataset [25] for mispronunciation detection—serves as our baseline[3]. We first evaluated an alignment-based method (described in Sec. 3.1). Although this method achieved an improved recall rate compared to the baseline, the degradation in F1-score and precision indicates that this method tends to overpredict unintelligible segments, which means further development is required.

In our proposed multi-task learning framework (detailed in Session 3.2), we conducted an ablation study by incrementally incorporating the auxiliary loss functions $\mathcal{L}_{\mathrm{D}}^{\mathrm{Enc}}$, $\mathcal{L}_{\mathrm{D}}^{\mathrm{Dec}}$ and $\mathcal{L}_{\mathrm{F}}$. As shown in Table 1, predicting D-labels solely from the source features—without leveraging target speech—resulted in suboptimal performance. In particular, when all auxiliary loss functions were applied, our multi-task approach attained the highest word-level F1-score of 30.3 and the second-highest frame-level accuracy of 75.6, which is only 0.1 % lower than the best result—demonstrating nearly equivalent accuracy to the top-performing method.

5.2. Subjective evaluation

Objectively evaluating the perception-based assessment is challenging because precisely annotating the onset and offset of intelligibility corruption is inherently difficult. Thus, to examine whether our system can effectively simulate the rater's cognition, we conducted a subjective evaluation. In this evaluation, both the rater and other English speakers with diverse backgrounds were asked to annotate segments of the speech that were not immediately intelligible to them. Figure 3 illustrates our annotation system[4], which is inspired by [26]. Our participants included:

- **EN-A**: The American rater mentioned in Session 4.1[5]
- **EN-J**: A Japanese speaker learning English.
- **EN-O**: Six L2 English speakers from diverse linguistic backgrounds (excluding English and Japanese).

Participants were instructed to annotate segments of $L2_R$ that they could not identify words by holding and releasing an annotation button while the utterance was playing. After the initial real-time annotation, the participants were allowed to edit misannotated segments. The annotated frames were then used to pinpoint the corresponding words that were unintelligible to the participants. Forced alignment between the waveform and the text sequence of $L2_R$ was performed using Kaldi's WSJ implementation.

Table 2: *Precision of the proposed unintelligibility indicator compared to manual annotation*

Methods	Precision↑		
	EN-A	EN-J	EN-O
2-Stage Shadowing + DTW (Ground Truth)	45.9	40.6	31.5
ASR-based	40.5	32.8	37.6
Alignment-based	21.6	34.4	27.0
Multi-task	43.2	28.1	31.7

Table 2 presents a comparison between the predicted unintelligible segments and the participants' annotations. We evaluated the similarity between the manual annotations and the predictions of the proposed method in terms of precision rate, defined as the proportion of correctly identified unintelligible segments over all predicted segments.

Among all methods, the 2-stage shadowing approach using DTW achieved the highest precision, demonstrating its effectiveness in replicating the rater's cognition of L2 speech. Furthermore, comparing the annotations of EN-J and EN-O revealed that the PPG-DTW results more closely resemble those of the Japanese learner (EN-J). This is reasonable because Japanese learners are generally more familiar with their own accented English, whereas learners from other backgrounds (EN-O) tend to identify different words as unintelligible.

Another interesting observation is that while the ASR-based method yields consistent results among annotators EN-J and EN-O, it does not surpass the performance of our proposed multi-task method on EN-A. This discrepancy is likely because ASR focuses on common acoustic features across accents, whereas our approach captures rater-dependent features, thereby more closely aligning with the rater's judgment.

6. Conclusions and Future Work

In this study, we present a perception-based L2 speech intelligibility indicator that leverages native rater shadowing data within a sequence-to-sequence voice conversion framework. Our experiments show that this approach provides effective feedback by pinpointing unintelligible segments, offering valuable insights for CALL systems, marking the first application of such a method.

While recent work of speech assessment has focused on evaluating speech quality and naturalness [27–29], speech intelligibility is still commonly measured using ASR's word error rates, which does not align well with real human's perception. Moreover, most existing methods provide holistic ratings over entire utterances, whereas intelligibility assessment requires a more fine-grained analysis. Shadowing, on the other hand, offers detailed insights into intelligibility and may lead to improved assessments that better accommodate the linguistic diversity of English speakers. Future work will extend our approach to cover a wider range of languages, accents, and diverse raters, facilitating broader multilingual accessibility.

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

[4] https://secondtonumb.github.io/DisfluencyAnnotationExp

[5] To reduce memory effects, the subjective annotations were conducted four years after the shadowing experiment and a randomly selected set of utterances were used

7. References

[1] A. Loukina, M. Lopez, K. Evanini, D. Suendermann-Oeft, A. V. Ivanov, and K. Zechner, "Pronunciation accuracy and intelligibility of non-native speech," in *Proc. Interspeech 2015*, 2015, pp. 1917–1921.

[2] C. Graham and N. Roll, "Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits," *JASA Express Letters*, vol. 4, no. 2, 2024.

[3] Ethnologue, "The most spoken languages in the world," https://www.ethnologue.com/insights/most-spoken-language/, 2023.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] J. Yue, F. Shiozawa, S. Toyama, Y. Yamauchi, K. Ito, D. Saito, and N. Minematsu, "Automatic Scoring of Shadowing Speech Based on DNN Posteriors and Their DTW," in *Proc. Interspeech*, 2017, pp. 1422–1426.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE JSTSP*, vol. 16, no. 6, pp. 1505–1518, 2022.

[17] S.-w. Yang, H.-J. Chang, Z. Huang, A. T. Liu, C.-I. Lai, H. Wu, J. Shi, X. Chang, H.-S. Tsai, W.-C. Huang, T.-h. Feng, P.-H. Chi, Y. Y. Lin, Y.-S. Chuang, T.-H. Huang, W.-C. Tseng, K. Lakhotia, S.-W. Li, A. Mohamed, S. Watanabe, and H.-y. Lee, "A large-scale evaluation of speech foundation models," *IEEE/ACM TASLP*, vol. 32, pp. 2884–2899, 2024.

[18] 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.

[19] L. P. Violeta, W.-C. Huang, D. Ma, R. Yamamoto, K. Kobayashi, and T. Toda, "Electrolaryngeal speech intelligibility enhancement through robust linguistic encoders," in *Proc. ICASSP*, 2024, pp. 10 961–10 965.

[20] H. Geng, D. Saito, and N. Minematsu, "A pilot study of applying sequence-to-sequence voice conversion to evaluate the intelligibility of l2 speech using a native speaker's shadowings," in *Proc. APSIPA ASC*, 2024, pp. 1–6.

[21] 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 *Proc. ICML*, 2021.

[22] R. Badlani, A. Łańcucki, K. J. Shih, R. Valle, W. Ping, and B. Catanzaro, "One tts alignment to rule them all," in *ICASSP 2022*, 2022, pp. 6092–6096.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 ICCV*, 2017, pp. 2999–3007.

[24] 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.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 ICASSP*, 2015, pp. 5206–5210.

[26] S. Coulange, T. Kato, S. Rossato, and M. Masperi, "Exploring impact of pausing and lexical stress patterns on l2 english comprehensibility in real time," in *Proc. Interspeech 2024*, 2024, pp. 1030–1034.

[27] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," *Proc. Interspeech*, pp. 4521–4525, 2022.

[28] W.-C. Huang, S.-W. Fu, E. Cooper, R. E. Zezario, T. Toda, H.-M. Wang, J. Yamagishi, and Y. Tsao, "The voicemos challenge 2024: Beyond speech quality prediction," in *SLT 2024*, 2024, pp. 803–810.

[29] J. Shi, H.-j. Shim, J. Tian, S. Arora, H. Wu, D. Petermann, J. Q. Yip, Y. Zhang, Y. Tang, W. Zhang *et al.*, "Versa: A versatile evaluation toolkit for speech, audio, and music," *arXiv preprint arXiv:2412.17667*, 2024.