Adversarial Disentanglement of Speaker Identity and L2 Phonological Errors via Raw-Waveform Architectures

Amirhossein Ghanipour Amirhandeh
Department of Linguistics, Faculty of Foreign Languages
University of Isfahan, Isfahan, Iran
amirhosseinghanipour@proton.me

Abstract

Automated Mispronunciation Detection and Diagnosis (MDD) systems should assess what a learner says, not who is saying it, yet speaker identity and phonological errors overlap in acoustic space. Fixed-resolution spectral inputs such as Mel-spectrograms carry vocal tract and voice characteristics into the classifier, which can lead models to penalize anatomical differences instead of genuine phonetic errors. We study a raw-waveform architecture that combines a parameterized SincNet front-end, a strided convolutional compressor, and a Transformer encoder. An adversarial speaker branch attached through a Gradient Reversal Layer (GRL) pushes the pooled latent representation to discard speaker-specific cues. On 143,507 phone segments from four L2-ARCTIC speakers, the training cross-entropy of the phone classifier falls from 2.20 to 1.35 over five epochs, while the adversarial speaker loss stays at 1.38, close to the chance level for four speakers ($\ln 4 \approx 1.386$). t-SNE projections show phone classes forming clusters while speakers are intermixed. We discuss why speaker-invariant representations matter for fairness and learner privacy in L2 assessment, state the controls still needed to support these preliminary findings, and outline how the model could be integrated into an ICALL system.

1 Introduction and Motivation

Deep representation learning has expanded the capabilities of Intelligent Computer-Assisted Language Learning (ICALL) and automated speaking assessment (Eskenazi, 2009; Voskoboinik et al., 2023). Delivering targeted pronunciation feedback still depends on reliable Mispronunciation Detection and Diagnosis (MDD) models. A central difficulty in this setting is the acoustic variability of non-native (L2) speech, in which individual anatomy and speaking habits can easily overshadow phonological errors (Kartushina and Frauenfelder, 2014; Munro, 2022).

Standard features such as Mel-Frequency Cepstral Coefficients (MFCCs) and log-Mel spectrograms rely on fixed-window Short-Time Fourier Transforms. These transforms impose a fixed time-frequency trade-off, discard phase information, and leave anatomical traits (such as pitch and vocal tract length) entangled with phonetic content. As a result, pronunciation training tools can produce false rejections when they mistake an unfamiliar voice profile for an articulation error.

Removing speaker identity from the assessment pathway matters for two reasons. The first is *fairness*. An L2 assessment system should grade the language a learner produces, not the learner's background or physical characteristics. When a model's decisions depend on voice properties that correlate with gender, age, or first language (L1), learners from some groups receive systematically more false rejections. Comparable group-level disparities are well documented for commercial speech recognition (Koenecke et al., 2020). In a learning context such errors are not neutral: feedback that repeatedly flags correct productions undermines learners' trust in the tool and misdirects practice time (Neri et al., 2002). The second reason is *privacy*. A learner's voice is biometric data that can identify the person, and speech also reveals attributes such as gender, age, and health (Nautsch et al., 2019). Pronunciation tools that store or transmit rich speaker representations therefore create risks that are disproportionate to their pedagogical purpose. Representations that retain phonetic information but carry little speaker information reduce this exposure, although, as we discuss in Section 5, they do not by themselves guarantee anonymity (Tomashenko et al., 2020).

We address the speaker–error confound directly in the time domain. Our framework processes raw

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 58–64, November 19–20, 2026, University of Tartu Library

audio through a parameterized SincNet front-end (Ravanelli and Bengio, 2018), a bank of learnable band-pass filters. We then apply domain-adversarial training (Ganin et al., 2016) with a Gradient Reversal Layer so that the latent representation becomes uninformative about the speaker. The goal is a representation in which phonological deviations, rather than speaker-specific acoustics, drive the classifier. This paper reports a preliminary, small-scale study of this design; Section 6 lists the evaluations still required.

2 Related Work

2.1 Pronunciation Assessment and MDD

Classic pronunciation scoring relies on Goodness of Pronunciation (GOP) measures derived from HMM forced alignments (Witt and Young, 2000). Later work replaced Gaussian mixture acoustic models with deep neural networks and added transfer-learning classifiers on top of GOP-style features (Hu et al., 2015). End-to-end MDD models dispense with explicit alignment-based scoring: Leung et al. (2019) recognize the phones actually produced with a CNN-RNN-CTC network and compare them with the canonical transcription, and fine-tuning self-supervised models such as wav2vec 2.0 has since become a strong approach for MDD on L2-ARCTIC (Xu et al., 2021; Peng et al., 2021). Beyond phone-level diagnosis, Transformer models jointly predict accuracy, fluency, and prosody scores at several granularities (Gong et al., 2022) on corpora such as speechocean762 (Zhang et al., 2021). Automated scoring of spontaneous speech in high-stakes tests has a longer history (Zechner et al., 2009). Within the NLP4CALL community, recent work has assessed task completion in the spontaneous speech of Finnish and Finland Swedish learners (Voskoboinik et al., 2023) and evaluated automatic pronunciation scoring against crowd-sourced corpus annotations (Hjortnaes et al., 2024). Most of these systems are trained to recognize or score speech robustly; they do not explicitly separate speaker variation from pronunciation deviations.

2.2 Speech Representations and Front-Ends

Speech processing has largely moved from handcrafted features to learned representations. Models such as wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) learn phonetic structure from raw audio with convolutional encoders followed by Transformers. They are pre-trained on tens of thousands of hours of speech, and their size makes them costly to fine-tune and deploy in lightweight CALL applications. Structured front-ends offer a compact alternative. SincNet (Ravanelli and Bengio, 2018) and LEAF (Zeghidour et al., 2021) constrain the first layer to parametric band-pass filters, which reduces the number of parameters and makes the learned filters interpretable. Such structured time-domain front-ends have so far seen little use in automated pronunciation assessment.

2.3 Speaker Variability and Adversarial Disentanglement

Kartushina and Frauenfelder (2014) showed that individual differences in L1 production relate to how accurately learners produce L2 sounds, and Munro (2022) found that the effect of elicitation method on L2 vowel production differs from speaker to speaker. Acoustic models must therefore separate general accent characteristics and individual voice properties from specific, correctable errors. Adversarial training offers a practical way to remove unwanted attributes from learned representations. In ASR, speaker-invariant training uses an adversarial speaker classifier to reduce speaker variability in senone representations (Meng et al., 2018). For pronunciation, Yang et al. (2020) added a language-adversarial branch that reduced performance gaps across L1 backgrounds; that approach targets L1-level differences and does not address individual voice characteristics. We apply the gradient reversal framework (Ganin et al., 2016) to individual speaker identity instead. Work on text representations has shown that adversarially removed attributes can often still be recovered by a classifier trained after the fact (Elazar and Goldberg, 2018), a caveat we return to in Section 5.

3 Methodology

The model operates directly on time-domain audio and combines phonetic feature extraction with an adversarial speaker penalty.

3.1 Dataset and Preprocessing

We use L2-ARCTIC (Zhao et al., 2018), a corpus of non-native English read speech with phone-level forced alignments and manually annotated substitution, deletion, and addition errors. Our experiments use four speakers: YDCK and HKK (Korean L1)

and BWC and LXC (Mandarin L1), one female and one male speaker per L1. Speaker identity in this subset is therefore confounded with both L1 and gender.

Phone boundaries and labels are read from the phone tier (`phones` or `phone`) of the TextGrid files distributed with the corpus. Silence and non-speech markers (empty labels, `sp`, `sil`, `spn`) are discarded, and every remaining phone-tier label is treated as a separate target class. This yields 143,507 segments in 126 classes. Because the forced alignments cover all of a speaker's roughly 1,130 utterances, whereas the manual error annotations cover 150 utterances per speaker (Zhao et al., 2018), most segments carry canonical phone labels rather than annotated error labels. Each segment is resampled to 16 kHz and center-padded with zeros or center-cropped to 16,000 samples (1.0 s) for batched training.

3.2 Raw-Waveform Front-End: SincNet

The first layer is a SincNet convolution (Ravanelli and Bengio, 2018) with $M = 80$ filters of 251 taps. Each filter is defined in the frequency domain as the difference of two rectangular low-pass filters:

$$G(f, f_1, f_2) = \text{rect}\left(\frac{f}{2f_2}\right) - \text{rect}\left(\frac{f}{2f_1}\right) \quad (1)$$

An inverse Fourier transform yields the discrete-time impulse response:

$$g[n, f_1, f_2] = 2f_2\,\text{sinc}(2\pi f_2 n) - 2f_1\,\text{sinc}(2\pi f_1 n), \quad (2)$$

which is truncated and smoothed with a Hamming window. Each filter has only two learnable parameters, the low cutoff f_1 and the bandwidth $f_2 - f_1$ (each constrained to at least 50 Hz), so the whole bank has $2M = 160$ parameters instead of the 80×251 weights of an unconstrained convolution. Following Ravanelli and Bengio (2018), the cutoffs are initialized uniformly on the Mel scale between 30 Hz and the Nyquist frequency. The filter outputs are batch-normalized and passed through a GELU activation.

3.3 Temporal Compression and Transformer Encoder

Self-attention over raw sample sequences (16,000 points per second) would have prohibitive memory requirements, so a convolutional compressor follows the SincNet stage. It consists of two 1D convolutions (kernel sizes 7 and 5, strides 4 and 2, with 128 and 256 output channels), each followed by batch normalization and GELU, for a total downsampling factor of 8.

The compressed sequence enters a 4-layer Transformer encoder ($d_{\text{model}} = 256$, 8 attention heads, $d_{\text{ff}} = 1024$, dropout 0.1). Global average pooling over time yields a single segment-level representation $z_p \in \mathbb{R}^{256}$.

3.4 Domain-Adversarial Disentanglement

The pooled vector z_p feeds two parallel branches:

- **Phone classifier** C: a linear layer that maps z_p to one of the 126 phone-tier classes y_{phn}.
- **Speaker discriminator** D: a two-layer perceptron (256–128–4, ReLU, dropout 0.2) that predicts the speaker y_{spk}.

A Gradient Reversal Layer (GRL) (Ganin et al., 2016) sits between z_p and the speaker discriminator. In the forward pass it is the identity, and in the backward pass it multiplies incoming gradients by $-\lambda$:

$$\text{GRL}(x) = x, \qquad \frac{\partial\,\text{GRL}(x)}{\partial x} = -\lambda \mathbf{I}. \quad (3)$$

The network minimizes the sum of the two cross-entropy losses,

$$\mathcal{L} = \mathcal{L}_{\text{phn}}\big(C(z_p), y_{\text{phn}}\big) + \mathcal{L}_{\text{spk}}\big(D(\text{GRL}(z_p)), y_{\text{spk}}\big). \quad (4)$$

Because of the GRL, the discriminator parameters descend on $\mathcal{L}_{\text{spk}}$, whereas the encoder parameters θ_e receive the gradient $\nabla_{\theta_e}\mathcal{L}_{\text{phn}} - \lambda\nabla_{\theta_e}\mathcal{L}_{\text{spk}}$. The encoder is thus pushed to produce features that predict the phone class while making the speaker hard to identify. Following Ganin et al. (2016), λ is increased from 0 to 1 at every optimization step to avoid unstable gradients early in training:

$$\lambda = \frac{2}{1 + e^{-10 \cdot p}} - 1, \quad (5)$$

where $p \in [0, 1]$ is the fraction of training steps completed.

4 Experiments and Results

4.1 Training Setup

We trained the network on two NVIDIA T4 GPUs in a Distributed Data-Parallel configuration with mixed-precision (FP16) arithmetic. Training ran for 5 epochs with AdamW (learning rate 10^{-4},

Epoch	λ (start→end)	Phone	Speaker
1	$0.000 \rightarrow 0.762$	2.2001	1.4004
2	$0.762 \rightarrow 0.964$	1.7044	1.3877
3	$0.964 \rightarrow 0.995$	1.5323	1.3854
4	$0.995 \rightarrow 0.999$	1.4305	1.3821
5	$0.999 \rightarrow 1.000$	**1.3496**	**1.3827**
Uniform-guess reference		$\ln 126 \approx 4.836$	$\ln 4 \approx 1.386$

Table 1: Epoch-averaged training cross-entropy of the phone classifier and of the adversarial speaker discriminator, with the range of the GRL coefficient λ within each epoch.

weight decay 10^{-5}, batch size 128, random seed 42). All losses below are training cross-entropies averaged over each epoch; this preliminary study did not use a held-out split.

4.2 Adversarial Convergence

Table 1 summarizes the loss trajectory.

A uniform guess over 126 classes gives a cross-entropy of $\ln 126 \approx 4.84$; the epoch-averaged phone loss falls from 2.20 in the first epoch to 1.35 in the fifth. The speaker loss stays between 1.38 and 1.40 throughout, essentially at the value of $\ln 4 \approx 1.386$ expected from uniform guessing over four speakers (each speaker contributes 1,130–1,131 utterances of the same ARCTIC prompts (Zhao et al., 2018), so the speaker classes are close to balanced). The jointly trained discriminator therefore cannot identify the speaker from z_p any better than chance. This is consistent with the encoder suppressing speaker information, but it is not proof of it: the speaker loss is already at chance in the first epoch, when λ is still small, and an adversarial discriminator at chance does not exclude that a separately trained probe could recover the speaker (Elazar and Goldberg, 2018). Section 6 describes the controls needed to settle this.

4.3 Latent Space and Filterbank Analysis

To inspect the representation qualitatively, we projected z_p for 2,000 segments onto two dimensions with t-SNE (perplexity 30).

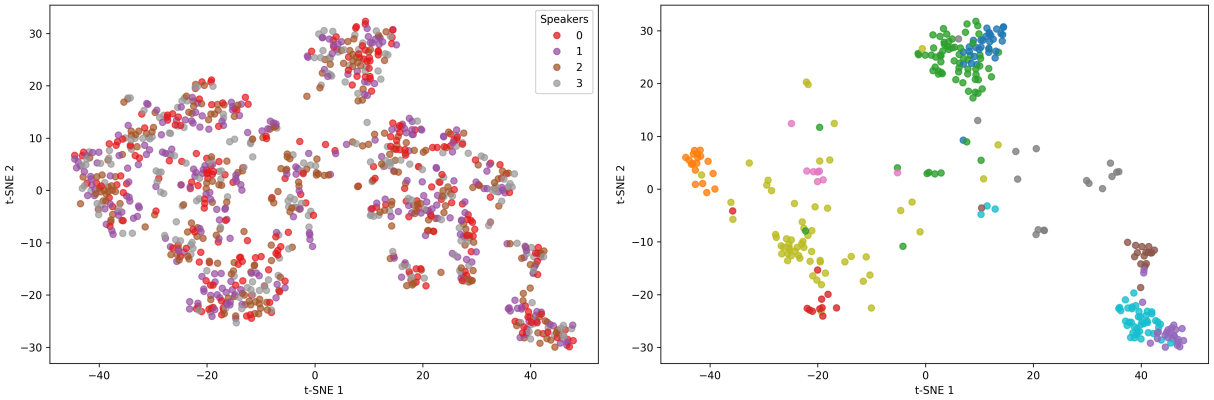
In Figure 1a the four speakers are intermixed across the whole projection and do not form speaker-specific clusters. Figure 1b colors the same points by the ten most frequent phone-tier classes, which form largely separate groups. The representation thus preserves phone-level contrasts while showing no visible speaker structure. t-SNE projections are qualitative, however, and do not replace a quantitative probe.

Figure 2 shows the SincNet cutoffs after training. The passbands still tile the full 0–8000 Hz range, with narrow bands at low frequencies, where formants are resolved, and wider bands at high frequencies, where fricative energy lies. This Mel-like layout reflects the initialization: after five epochs at a learning rate of 10^{-4} the cutoffs are visually indistinguishable from their initial values. The front-end therefore remained stable and did not collapse, but this run provides no evidence that it adapted its filters to L2 speech; longer training, or a higher learning rate for the filter parameters, would be needed to test this.

5 Discussion

Why speaker invariance matters for L2 assessment. In a pronunciation tutor, a speaker-dependent decision is a validity problem as well as a technical one. If an error detector relies partly on voice quality, pitch range, or vocal tract length, its false rejections are not randomly distributed but concentrate on learners whose voices differ from those in the training data. In our subset speaker identity is entangled with L1 and gender, so a model that encodes the speaker also encodes these group attributes. Pushing the representation towards speaker invariance is one way of building the principle that the system should assess the language produced, not the person producing it, into the model itself rather than checking for it only after deployment.

Privacy. Speaker-invariant representations also reduce how much biometric information a pronunciation service has to handle. If only z_p, rather than raw audio or speaker embeddings, is stored or sent to a remote scoring service, re-identifying a learner should become harder. Two qualifications apply. First, our evidence for invariance is that a jointly trained discriminator stays at chance; privacy claims require stronger attack models, such as speaker verification against the representation, as standardized in the VoicePrivacy evaluations (Tomashenko et al., 2020). Second, the raw audio still passes through the front-end, so the overall system protects privacy only if the audio is processed locally or discarded after encoding. We therefore see disentanglement as one component of privacy by design (Nautsch et al., 2019), not as a guarantee of anonymity.



(a) Colored by speaker

(b) Colored by phone class (ten most frequent)

Figure 1: Two-dimensional t-SNE projection of the pooled representation z_p for 2,000 segments. **(a)** Points colored by speaker (four speakers): the speakers are intermixed across the projection. **(b)** The same projection restricted to the ten most frequent phone-tier classes, each shown in a different color: segments of the same class tend to form compact clusters.

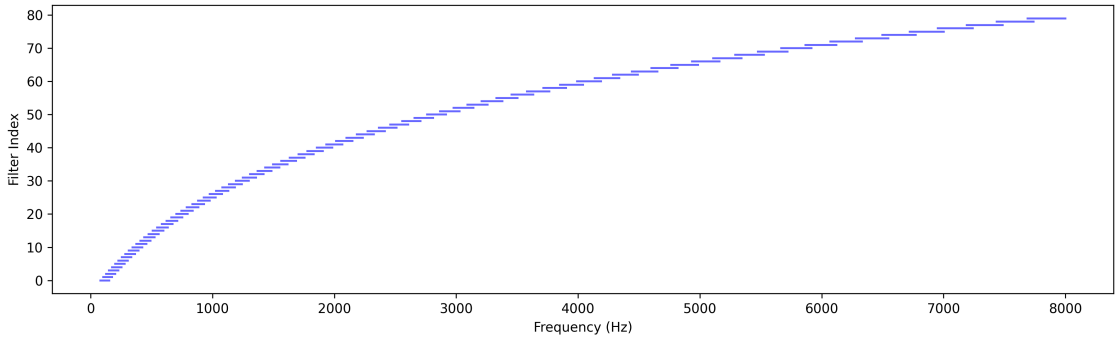


Figure 2: SincNet cutoff frequencies (f_1 to f_2, one horizontal segment per filter, sorted by f_1) after 5 epochs. The bands cover 0–8000 Hz without collapsing onto a few regions, and they remain close to their Mel-scale initialization.

Towards ICALL integration. We envision the model as the diagnosis component of a pronunciation tutor for read-aloud practice. A learner reads a prompt; the known prompt text is force-aligned to the recording; each aligned phone segment is encoded and classified; and the predicted label is compared with the canonical phone to produce feedback at the word and phone level. Since segments are scored independently and the model is small, this component could run on the learner's own device, which would also keep raw audio local. Before such integration, three questions must be answered: (i) whether the model improves detection of annotated errors over spectrogram-based and self-supervised baselines; (ii) whether false rejection rates become more uniform across speakers, L1 groups, and genders; and (iii) how learners and teachers respond to the resulting feedback in classroom or self-study use (Neri et al., 2002).

6 Limitations

This is a preliminary study, and its results should be read with the following limitations in mind. (1) All reported losses are training losses; we did not evaluate on held-out utterances or unseen speakers, and we report no error-detection metrics (precision, recall, F1) against the manual annotations. Since most training labels are canonical phones, the present model is best described as a speaker-adversarial phone classifier that would serve as the basis of an MDD system, not yet as an evaluated error detector. (2) We did not train a baseline without the adversarial branch ($\lambda = 0$), so we cannot show how much speaker information the encoder would otherwise retain; the speaker loss is near chance from the first epoch, which leaves open whether the discriminator was able to exploit speaker cues at all. (3) Speaker invariance was assessed only through the jointly trained discriminator and t-SNE; a post-hoc probe trained on frozen representations is needed (Elazar and Goldberg, 2018). (4) Only

four speakers and two L1s were used. (5) Segments are padded to 1 s, so most frames in a typical phone segment are padding, and mean pooling includes them; masking the padding may change the results.

7 Conclusion and Future Work

We presented a raw-waveform architecture that aims to separate speaker identity from phonological information in non-native speech by pairing a parameterized SincNet front-end with gradient-reversal adversarial training. On four L2-ARCTIC speakers, the phone classification loss decreased steadily while the adversarial speaker discriminator stayed at chance, and t-SNE projections showed phone-class structure without speaker clusters. These results are encouraging but preliminary.

Future work will (i) evaluate on held-out speakers from the full 24-speaker L2-ARCTIC corpus, with error-detection metrics computed against the manual annotations; (ii) compare against a non-adversarial version of the same model and against spectrogram-based and self-supervised baselines, reporting false rejection rates per speaker group; (iii) quantify residual speaker information with post-hoc probes and speaker-verification attacks; and (iv) integrate the model into an ICALL prototype for read-aloud practice and study its use with learners.

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☒ Generation of minor elements
☐ Translation
☒ Compression/expansion
☐ Generation from prompt

Details: Large Language Models (LLMs) were used for editing (grammar and spelling correction, flow, and style), for generating minor elements (drafting the abstract), and for compressing and expanding the author's own drafts to fit the paper format and page limits. The author designed and conducted the experiments, wrote the code, wrote the original drafts, and reviewed all AI-assisted text, retaining full editorial control and responsibility for the content and scientific claims.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21.

Maxine Eskenazi. 2009. An overview of spoken language technology for education. *Speech Communication*, 51(10):832–844.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

Yuan Gong, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass. 2022. Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment. In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7262–7266.

Nils Hjortnaes, Daniel Dakota, Sandra Kübler, and Francis Tyers. 2024. Evaluating automatic pronunciation scoring with crowd-sourced speech corpus annotations. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)*, volume 211 of *Linköping Electronic Conference Proceedings*, pages 67–77.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Wenping Hu, Yao Qian, Frank K. Soong, and Yong Wang. 2015. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, 67:154–166.

Natalia Kartushina and Ulrich H. Frauenfelder. 2014. On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology*, 5.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Wai-Kim Leung, Xunying Liu, and Helen Meng. 2019. CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8132–8136.

Zhong Meng, Jinyu Li, Zhuo Chen, Yang Zhao, Vadim Mazalov, Yifan Gong, and Biing-Hwang Juang. 2018. Speaker-invariant training via adversarial learning. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5969–5973.

Murray J. Munro. 2022. Variability in L2 vowel production: Different elicitation methods affect individual speakers differently. *Frontiers in Psychology*, 13.

Andreas Nautsch, Abelino Jiménez, Amos Treiber, Jascha Kolberg, Catherine Jasserand, Els Kindt, Héctor Delgado, Massimiliano Todisco, Mohamed Amine Hmani, Aymen Mtibaa, Mohammed Ahmed Abdelraheem, Alberto Abad, Francisco Teixeira, Driss Matrouf, Marta Gomez-Barrero, Dijana Petrovska-Delacrétaz, Gérard Chollet, Nicholas Evans, Thomas Schneider, Jean-François Bonastre, Bhiksha Raj, Isabel Trancoso, and Christoph Busch. 2019. Preserving privacy in speaker and speech characterisation. *Computer Speech & Language*, 58:441–480.

Ambra Neri, Catia Cucchiarini, Helmer Strik, and Lou Boves. 2002. The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning*, 15(5):441–467.

Linkai Peng, Kaiqi Fu, Binghuai Lin, Dengfeng Ke, and Jinsong Zhan. 2021. A study on fine-tuning wav2vec2.0 model for the task of mispronunciation detection and diagnosis. In *Proceedings of Interspeech 2021*, pages 4448–4452.

Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with SincNet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028.

Natalia Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, and Massimiliano Todisco. 2020. Introducing the VoicePrivacy initiative. In *Proceedings of Interspeech 2020*, pages 1693–1697.

Ekaterina Voskoboinik, Yaroslav Getman, Ragheb Al-Ghezi, Mikko Kurimo, and Tamás Grósz. 2023. Automated assessment of task completion in spontaneous speech for Finnish and Finland Swedish language learners. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL 2023)*, volume 197 of *Linköping Electronic Conference Proceedings*, pages 102–110.

Silke M. Witt and Steve J. Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2-3):95–108.

Xiaoshuo Xu, Yueteng Kang, Songjun Cao, Binghuai Lin, and Long Ma. 2021. Explore wav2vec 2.0 for mispronunciation detection. In *Proceedings of Interspeech 2021*, pages 4428–4432.

Longfei Yang, Kaiqi Fu, Jinsong Zhang, and Takahiro Shinozaki. 2020. Pronunciation erroneous tendency detection with language adversarial represent learning. In *Proceedings of Interspeech 2020*, pages 3042–3046.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895.

Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi. 2021. LEAF: A learnable frontend for audio classification. In *International Conference on Learning Representations (ICLR)*.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An open-source non-native English speech corpus for pronunciation assessment. In *Proceedings of Interspeech 2021*, pages 3710–3714.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-ARCTIC: A non-native English speech corpus. In *Proceedings of Interspeech 2018*, pages 2783–2787.