De-Speakerizing Accented ASR: Measuring and Mitigating Speaker Entanglement for Fair, Reliable Recognition

Jiaen Sun

School of Computing, National University of Singapore
sunjiaen@u.nus.edu

Abstract

This research statement proposes to measure and mitigate speaker entanglement, where accent features inadvertently encode who is speaking in accented automatic speech recognition (ASR). We argue that entanglement inflates scores under lenient split for the same speaker and worsens fairness gaps across accents, and we outline a parameter-efficient mitigation that combines adversarial de-speakerization with safe conditioning. The plan is grounded in established results in accented ASR, domain-adversarial learning, and parameter-efficient fine-tuning; it is feasible with public datasets and a frozen Whisper backbone, and can potentially guide low-resource data collection.

Introduction

Speaker entanglement is the phenomenon where an accent representation memorizes speaker identity cues, so models look good on seen talkers but fail on unseen speakers and unseen accents. We study how speaker entanglement degrades accented ASR and how to remove it in a way that is practical. ASR converts speech audio to text and enables captions, meeting notes, and language learning tools. Despite strong average performance from modern systems like Whisper, large audits report persistent word error rate (WER) gaps across accents both in English and in other languages, raising concerns for accessibility in diverse classrooms and international workplaces (DiChristofano et al. 2023; Bandodkar et al. 2024; Besdouri, Zribi, and Belguith 2024). This project proposes to: (1) measure entanglement rigorously; (2) mitigate it with adversarial de-speakerization and safe conditioning; (3) translate the findings into data-collection guidelines for low- vs. high-resource languages.

Background

Explicit accent information helps ASR, but many pipelines risk entanglement or weak generalization. Early end-to-end work augments ASR models with multi-task training and accent embedding in the context of end-to-end ASR, reporting sizable WER improvements over conventional baselines (Viglino, Motlicek, and Cernak 2019). Subsequent methods pair self-supervised learning (SSL) encoders with accent-dependent features, showing further gains on the Accented English Speech Recognition Challenge (AESRC-2020) (Deng, Cao, and Ma 2021; Shi et al. 2021). More recent systems move from one-hot IDs to continuous accent spaces and adapters, highlighting the value of continuous conditioning but rarely stripping speaker from those spaces (Qian, Gong, and Huang 2022). At the same time, large audits of research and commercial ASR consistently document accent disparities, indicating that conditioning alone has not solved equity (DiChristofano et al. 2023; Bandodkar et al. 2024).

Adversarial invariance is a principled way to remove nuisance factors such as speaker from learned features. domain-adversarial neural network (DANN) uses a gradient reversal layer (GRL) to make a representation useful for the main task but uninformative for a protected attribute, and has been demonstrated in computer vision, robust speech recognition, and speaker-invariant training (SIT) (Ganin et al. 2016; Shinohara 2016; Meng et al. 2018).

Open benchmarks and robust backbones make this project feasible and reproducible. AESRC-2020 supports seen / unseen studies; Common Voice and CommonAccent provide broader accent metadata, though caution is still needed regarding self-reported labels and split hygiene. (Ardila et al. 2019; Zuluaga-Gomez et al. 2023; Shi et al. 2021). Whisper offers a strong frozen backbone so we can focus on representation design rather than training from scratch (Radford et al. 2023).

Prior Work by the Applicant

I am currently working under project "Development of language-learning web application to support Human-Computer Interaction", where computer-assisted pronunciation training (CAPT) is applied to assist Mandarin learning for local medical staff and students in Singapore. While evaluating Mandarin audio from participants, some speakers' clips produced text in other languages, which is a behavior consistent with hallucination-like outputs documented for Whisper. We treat this as a representation issue rather than a single-system bug, and it motivates us to measure entanglement before attempting mitigation.

ID	Accent cue	GRL	Into ASR	Purpose
B0	none	–	none	Baseline WER and fairness
B1	continuous embedding	No	none	Full continuous representation de-speakerization
B2	continuous embedding	Yes	none	De-speakerized representation making effect of GRL
B3	continuous embedding	Yes	bounded FiLM	Full method: clean cue with conditioning

Table 1: Experiments for de-speakerization evaluation in **E1**

Setting	H (hours)	S (speakers)	M (min/spk)
Low-resource A	20	20	60
Low-resource B	20	80	15
Low-resource C	20	200	6
High-resource A	200	200	60
High-resource B	200	400	30

Table 2: Corpus design for **E2**.

Approach

Our approach quantifies speaker entanglement using diagnostic probes and then reduces it through adversarial de-speakerization, with optional bounded FiLM conditioning of a frozen backbone. We further assess downstream value for low-resource ASR by designing and analyzing a complementary data-collection protocol.

A1: Measure entanglement. Several evaluation metrics may be applied: (1) compare performance of model training on dataset with speaker-disjoint vs. lenient splits to expose inflation; (2) train a tiny probe to predict *speaker ID* from any accent embedding, where high macro-F1 implies leakage; (3) compute equal-error rate (EER) of speaker verification directly on the embedding; and (4) compare the embedding with strong speaker representations like X-vectors and ECAPA-TDNN via computing cosine or Centered Kernel Alignment (CKA) similarity (Snyder et al. 2018; Desplanques, Thienpondt, and Demuynck 2020).

We will empirically quantify the magnitude of these effects and link them to unseen-speaker/accents WER in subsequent experiments.

A2: De-speakerized accent embedding and conditioning. We will add a small Accent Identification (AID) head on pooled encoder states to produce a continuous accent embedding $\mathbf{e}$ while a speaker classifier penalizes speaker information in $\mathbf{e}$ via GRL. We then condition the ASR with feature-wise linear modulation (FiLM) at decoder cross-attention via keys/values. This is supposed to preserve the strengths of continuous conditioning while removing voice memorization (Ganin et al. 2016; Qian, Gong, and Huang 2022).

A3: Corpus design for low- vs. high-resource languages. With total hours H, we will vary the number of speakers S (breadth) versus minutes per speaker M (depth), measure unseen-speaker WER and embedding EER, and make exploration to prioritize *breadth* (large S) in low-resource settings to avoid entanglement; add *depth* for personalization once diversity is secured.

Evaluation

We aim to answer two questions: **E1**-does our method reduce speaker entanglement and improve fairness; **E2**-how should we design a corpus for low- vs. high-resource settings.

All models use *speaker-disjoint* train/dev/test splits. We report results on AESRC-2020 for **E1**, and on Common Voice/CommonAccent for **E2**.

Table 1 lists the experiments for **E1**: a control without accent cues (B0), a continuous embedding without GRL (B1), a de-speakerized embedding with GRL (B2), and our full method with bounded FiLM (B3). Whisper stays frozen; only small heads and adapters are trained.

Table 2 defines training budgets for **E2**. We vary the number of speakers (S) and minutes per speaker (M) under fixed total hours (H), then measure unseen-speaker WER, fairness gap on a shared test set. Precise values will be finalized based on pilot runs.

Discussion

We expect de-speakerized accent embeddings to reduce fairness gaps and unseen-speaker WER. By removing speaker identity cues from the accent embeddings, conditioning should generalize better across speakers. We also hypothesize an actionable corpus-design guidance: in low-resource settings maximize the *number of speakers* per hour budget; in high-resource settings maintain broad coverage then add depth for optional personalization.

Conclusion

This project reframes accented ASR robustness as a representation problem: measure and remove speaker entanglement, then use the findings to shape better datasets in low-resource context. Grounded in adversarial invariance and continuous conditioning, the plan is realistic (with frozen backbones and small heads), feasible (accessible public data and modest compute), and valuable (improved fairness and clear corpus design rules).

References

Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Bandodkar, G.; Agarwal, S.; Sughosh, A. K.; Singh, S.; and Choi, T. 2024. "Allot?" Is "A Lot!" Towards Developing More Generalized Speech Recognition System for Accessible Communication. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23327–23334.

Besdouri, F. Z.; Zribi, I.; and Belguith, L. H. 2024. Arabic automatic speech recognition: challenges and progress. *Speech Communication*, 163: 103110.

Deng, K.; Cao, S.; and Ma, L. 2021. Improving accent identification and accented speech recognition under a framework of self-supervised learning. *arXiv preprint arXiv:2109.07349*.

Desplanques, B.; Thienpondt, J.; and Demuynck, K. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

DiChristofano, A.; Shuster, H.; Chandra, S.; and Patwari, N. 2023. Performance disparities between accents in automatic speech recognition (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 16200–16201.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59): 1–35.

Meng, Z.; Li, J.; Chen, Z.; Zhao, Y.; Mazalov, V.; Gong, Y.; and Juang, B.-H. 2018. Speaker-invariant training via adversarial learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5969–5973. IEEE.

Qian, Y.; Gong, X.; and Huang, H. 2022. Layer-wise fast adaptation for end-to-end multi-accent speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 2842–2853.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Shi, X.; Yu, F.; Lu, Y.; Liang, Y.; Feng, Q.; Wang, D.; Qian, Y.; and Xie, L. 2021. The accented english speech recognition challenge 2020: open datasets, tracks, baselines, results and methods. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6918–6922. IEEE.

Shinohara, Y. 2016. Adversarial multi-task learning of deep neural networks for robust speech recognition. In *Interspeech*, 2369–2372. San Francisco, CA, USA.

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5329–5333. IEEE.

Viglino, T.; Motlicek, P.; and Cernak, M. 2019. End-to-End Accented Speech Recognition. In *Interspeech*, 2140–2144.

Zuluaga-Gomez, J.; Ahmed, S.; Visockas, D.; and Subakan, C. 2023. Commonaccent: Exploring large acoustic pretrained models for accent classification based on common voice. *arXiv preprint arXiv:2305.18283*.