

**AccentBox: High-Fidelity
Zero-Shot Accent Generation**

B237820

8,952 words

Master of Science
Speech and Language Processing
School of Philosophy, Psychology and Language Sciences
University of Edinburgh
2024

Abstract

While recent Zero-Shot Text-to-Speech (ZS-TTS) models have achieved high naturalness and speaker similarity, they fall short in accent fidelity and control. To address this issue, we propose zero-shot accent generation that unifies Foreign Accent Conversion (FAC), accented TTS, and ZS-TTS, with a novel two-stage pipeline. In the first stage, we achieve state-of-the-art (SOTA) on Accent Identification (AID) with 0.56 f1 score on unseen speakers. In the second stage, we condition ZS-TTS system on the pretrained speaker-agnostic accent embeddings extracted by the AID model. The proposed system achieves higher accent fidelity on inherent/cross accent generation, and enables unseen accent generation.

Acknowledgements

1. This project has been generously funded by the following organizations/individuals:

- **MSc Speech and Language Processing (SLP) Program**, School of Philosophy, Psychology, and Language Sciences, the University of Edinburgh: The program, under the direction of Prof. Simon King, Dr. Catherine Lai, and Dr. Korin Richmond, provided funding for the computing costs associated with the first stage accent identification (GenAID) and covered the majority of the listening test expenses in the second stage zero-shot accent generation (AccentBox). The funding for Cirrus computing and listening tests expenditure were crucial for this entire project.
- **Zhibo Su**, Transsion Holdings: My ex-colleague and close friend covered the remaining listening test costs for the second stage AccentBox. His inspiration and endorsement have been invaluable in the pursuit of this project.
- **Huiqing Jin**: My beloved mother supported the computing costs for the second stage AccentBox. Her unwavering love and financial support for my privately borrowed servers used for uninterrupted extensive experiments facilitated the project's success.

2. This project has benefited tremendously through the discussion with:

- **Weekly Supervision & Discussion**: Dr. Korin Richmond, Siqi Sun, Spencer Jenson, Noe Berger.
- **MSc SLP and School of Informatics classmates**: Linus Foo, Rowley Adams, Xinye Ma, Mingyue Jian.
- **Friends in TTS Industry**: Zhibo Su (Transsion Holdings), Yijie Zhang (iFLY-TEK).

3. We do live in a different age of speech technology, so here is to these pioneers who make speech technology research easily reproducible and openly-available. Their work greatly inspire and facilitate my research (apologies for abusing the acknowledgment section to include references and links):

- **SpeechBrain¹** (Ravanelli et al., 2021, 2024)

¹<https://github.com/speechbrain/speechbrain>

- coqui-TTS² and YourTTS³ (Casanova et al., 2022)
- Parler-TTS⁴ and the marvellous work by Lyth and King (2024)
- big tech companies X, Y, Z, ... (don't want to name any of you but you surely change the engineering field of speech)
- ... and numerous other pioneers.

4. Lastly and most importantly, words cannot describe how much I appreciate the chance of learning and growing in the MSc SLP program.

- **Prof. Simon King, Dr. Catherine Lai, and Dr. Korin Richmond** You have successfully cultivated another year of SLP students - I am sure we will all thrive upon graduation, and none of the accomplishments or growth would be possible not for your guidance, support, wisdom, and experience along the way.
- **The Centre for Speech Technology Research (CSTR)** All the research talks, group paper readings/discussions, and other activities have revolutionised my understanding of speech science and technology. I might have appeared as a reckless engineering person at first, but I believe I am picking up the art of academic research into speech science - and thank you all members of CSTR for teaching me this. It is worth mentioning the works from here that inspire me to self-propose this very dissertation: UNISYN⁵ (Fitt, 2000), VCTK (Yamagishi et al., 2012), EDACC (Sanabria et al., 2023), and Lyth and King (2024).
- **Dr. Korin Richmond** As both my advisor and supervisor, thank you from the bottom of my heart for all the help, support, and guidance along this year's journey! The late night emails for supervision/ideas, the short-notice meeting for study advice, the debate over zero-shot LLM-based TTS and so many other things - thank you for bearing with all of these!

²<https://github.com/coqui-ai/TTS>

³<https://github.com/Edresson/YourTTS>

⁴<https://github.com/huggingface/parler-tts>

⁵<https://www.cstr.ed.ac.uk/projects/unisyn>

Table of Contents

1	Introduction	1
1.1	Motivation: Accent Matters in ZS-TTS	1
1.2	Related Work: FAC, Accented TTS, and ZS-TTS	2
1.3	Task Definition: Zero-shot Accent Generation	3
1.4	Research Gap: Speaker-Accent Entanglement in AID and ZS-TTS	4
1.5	Contributions: GenAID & AccentBox	5
1.6	Roadmap	6
2	GenAID: Generalisable Accent Identification across Speakers	7
2.1	Overview	7
2.2	Data	7
2.2.1	Existing Datasets	7
2.2.2	Data Selection & Processing	8
2.2.3	Limitations	10
2.3	Problem Identification	12
2.3.1	Hypothesised Problems	12
2.3.2	Reproduction of Baselines	12
2.3.3	Results & Analysis of Baselines	13
2.3.4	Research Question	16
2.4	Methods	16
2.4.1	Overview	16
2.4.2	Validation on Unseen Speakers	16
2.4.3	Weighted Sampling	17
2.4.4	Data Augmentation by Perturbation	18
2.4.5	Information Bottleneck	18
2.4.6	Adversarial Training	19

2.5	Experiments	20
2.5.1	Systems	20
2.5.2	Configurations	21
2.5.3	Evaluation	22
2.6	Results & Analysis	23
2.6.1	Overall	23
2.6.2	Reduced Overfitting by Validation on Unseen Speakers	25
2.6.3	More Balanced Predictions by Weighted Sampling	25
2.6.4	Improved Generalisation by Data Augmentation	25
2.6.5	Effective Disentanglement by Information Bottleneck for XLSR-based AID	27
2.6.6	Better Speaker Disentanglement by Adversarial Training	29
2.6.7	Implications for Self-Supervised Learning (SSL) Models	30
2.6.8	Implications for Accent Similarity	30
2.7	Conclusions & Future Work	31
3	AccentBox: High-Fidelity Zero-Shot Accent Generation	33
3.1	Overview	33
3.2	Data	34
3.2.1	Pretraining: LibriTTS-R	34
3.2.2	Finetuning & Inference: VCTK	34
3.2.3	Stimuli for Listening Tests: <i>Comma Gets a Cure</i>	35
3.2.4	Limitations	35
3.3	Problem Identification	36
3.3.1	Accent Mismatch/Hallucination in ZS-TTS	36
3.3.2	Research Question	36
3.4	Methods	37
3.4.1	Training: Conditioning on GenAID Embedding	37
3.4.2	Inference: Inherent/Cross/Unseen Accent Generation	38
3.5	Experiments	39
3.5.1	Systems	39
3.5.2	Configurations	40
3.5.3	Objective Evaluation	40
3.5.4	Subjective Evaluation	41
3.6	Results & Analysis	41

3.6.1	Overview	41
3.6.2	Inherent Accent Generation	43
3.6.3	Cross Accent Generation	44
3.6.4	Problems of ZS-TTS	44
3.7	Conclusions	45
4	General Discussion	47
4.1	How should we define accents in data collection?	47
4.2	Is full speaker-accent disentanglement desirable and achievable?	47
4.3	Are utterance-level AID and ZS-TTS ill-defined tasks?	48
5	Conclusions & Future Work	49
	Bibliography	50
A	GenAID: Cleaning Accent Labels	59
B	AccentBox: Evaluation Materials	60
B.1	Stimuli for Listening Tests: <i>Comma Gets a Cure</i>	60
B.2	Reference Speech for Listening Tests	61
B.3	Questionnaires for Listening Tests	61

Chapter 1

Introduction

1.1 Motivation: Accent Matters in ZS-TTS

Recent advances in neural TTS systems have made it possible to generate speech, that is indistinguishable from human recordings, for English single-speaker sentence-level TTS, e.g. NaturalSpeech (Tan et al., 2024). Over the past few months, more recent advances in ZS-TTS systems have enabled speech generation of any unseen speaker’s voice in a 3-second audio clip, that is on-par quality with human recordings, e.g. NaturalSpeech 3 (Ju et al., 2024) and VALL-E 2 (Chen et al., 2024). Despite these achievements, most ZS-TTS systems have focused on replicating speakers’ voices (Jia et al., 2018) while largely ignoring accent variation. These systems are typically trained on mostly American English data without accent conditioning or control. Such disregard for accents and biased training leads to poor accent fidelity and no control over accents in the generated speech (Wang et al., 2023a).

Generating speech of high accent fidelity is crucial in TTS, especially for those communicating in a lingua franca like English or French. For native speakers (L1), having their accents accurately represented preserves their linguistic identity, which is integral to their personal and regional identity (Rosina, 1997). For non-native speakers (L2), accurate and controllable accent generation is crucial for addressing accent discrimination. TTS systems that can reproduce L2 speakers’ accent in media can alleviate their pressure to conform to native accents, allowing L2 speakers to retain their linguistic identity (Gluszek and Dovidio, 2010). Additionally, L2 speakers can benefit from a more personalized and effective language learning through TTS systems in Computer-Aided Pronunciation Training (CAPT), where L2 accents are converted to native-like accents while preserving speaker identity (Felps et al., 2009; Agarwal and

Chakraborty, 2019).

Motivated by the poor accent generation in ZS-TTS as well as the social and moral imperative for inclusive speech technology, we take an initiative to address accent-related issues in ZS-TTS. Generating accented speech in a zero-shot manner has broad and promising applications in personalised virtual assistants (Pal et al., 2019), movie dubbing (Spiteri Miggiani, 2021), CAPT (Felps et al., 2009; Agarwal and Chakraborty, 2019), and etc. Our goal is to promote the inclusivity of ZS-TTS for speakers of non-major accents.

1.2 Related Work: FAC, Accented TTS, and ZS-TTS

Previous studies on generating accented speech can be categorised into three areas, with a comparison of these tasks with our proposed task presented in Table 1.1.

1) Foreign Accent Conversion (FAC) Accent conversion is a speech-to-speech task that takes source speech from a target speaker as input, and converts the L2 accent in the source speech to a target L1 accent. Reference-free FAC proposed by Liu et al. (2020); Zhao et al. (2021) removes the need for an additional reference speech with the target L1 accent and the same content as source speech. Zero-shot FAC proposed by Quamer et al. (2022); Ding et al. (2022); Jia et al. (2023) enables FAC for unseen L2 speakers in the source speech. Despite these recent studies, reference-free zero-shot FAC is still limited by the inability to generate any given text and generalise to unseen accents or accent pairs.

2) Multi-Accent/Accented TTS Accented TTS takes target text, accent ID, and speaker ID as input, aiming to generate accented speech with high naturalness and accent fidelity. To leverage expert linguistic knowledge, some prior work on accented TTS focus on building a multi-accent front-end, such as Black et al. (1998), Fitt (2000), Sun and Richmond (2024), or leveraging the built multi-accent front-end to assist accented TTS training, such as Zhou et al. (2024a), Ma et al. (2024). These approaches are not generalisable to most languages where a high-quality multi-accent front-end is hard-to-obtain due to its labor intensive nature. To model the acoustic details of accent, different accent modelling techniques have been proposed, including Variational Auto-Encoder (VAE) (Melechovsky et al., 2023), Diffusion (Deja et al., 2023), phoneme- and utterance-level representation learning (Zhou et al., 2024b; Liu et al., 2023, 2024). To disentangle speaker and accent information which are intrinsically

intertwined in training data, multiple strategies have been used, including adversarial training of classifying speakers (Badlani et al., 2023b; Zhou et al., 2024b), data augmentation (Badlani et al., 2023a,b), and bottleneck (Ma et al., 2024). Despite these studies, accented TTS remains limited by its inability to generate speech for unseen speakers and unseen accents.

3) Zero-shot TTS (ZS-TTS) ZS-TTS generates speech using the voice in a speech prompt (i.e. reference speech) and target text as input. Jia et al. (2018) propose to condition the TTS on speaker embeddings obtained by a pretrained speaker verification model. Casanova et al. (2022) achieve great success in ZS-TTS, using the same idea with the combination of generative modelling. In more recent works, some treat TTS as a conditional Language Modelling (LM) task with the entire reference speech as context to generate target speech, leveraging audio codecs (Défossez et al., 2023; Zeghidour et al., 2022) and speech Large Language Modelling (LLM) (Wang et al., 2023a; Kharitonov et al., 2023; Le et al., 2023; Chen et al., 2024); others continue to treat TTS as a simple-to-complex distribution task, similarly using audio codecs, but relying on Diffusion instead to model the target speech conditioned on reference speech codec representation (Shen et al., 2024; Ju et al., 2024). Despite different zero-shot generation approaches, none of these studies adequately addresses accent generation, with some acknowledging poor ZS-TTS performance for accented speakers (Wang et al., 2023a).

1.3 Task Definition: Zero-shot Accent Generation

Task	Accent Generation Abilities		
	Any given text?	Any given speaker?	Any given accent?
Foreign Accent Conversion (FAC)	No.	Yes.	Only seen/trained accent pairs.
Multi-Accent/ Accented TTS	Yes.	Only seen speakers.	Only seen accents.
Zero-Shot TTS	Yes.	Yes.	No.
Zero-Shot Accent Generation	Yes.	Yes.	Yes.

Table 1.1: Different tasks proposed for generating accented speech.

We propose a new task: zero-shot accent generation. Compared with FAC, our task aims at text-to-speech generation rather than speech-to-speech mapping, which cannot generate speech from any given text or for unseen accents. Compared with accented TTS, we aim at extending the model’s ability to generate speech for unseen speakers/accents. Compared with ZS-TTS, we aim at controllable accent generation with high fidelity. Our task, zero-shot accent generation refers to generating any speech content in any given voice and accent from one audio clip, unifying the capabilities of all three tasks mentioned above.

1.4 Research Gap: Speaker-Accent Entanglement in AID and ZS-TTS

Speaker-accent entanglement is a pervasive problem in various speech technologies, including Automatic Speech Recognition (ASR) (Wang et al., 2023b), Accent Identification (AID) (Shi et al., 2021; Li et al., 2023), and TTS (Ding et al., 2022; Quamer et al., 2022; Badlani et al., 2023b,a; Zhou et al., 2024b; Ma et al., 2024). In an ideal, though unrealistic, situation, a speech dataset should include utterances from the same speaker in different accents. However, most speakers cannot consistently produce a wide range of accents. Such speaker-accent entanglement leads to limitations in both AID and ZS-TTS models, which jointly form the foundation of our proposed system.

In AID, the AESRC2020 benchmark (Shi et al., 2021) has been a standard, offering data specifically collected for accented ASR. However, this data is limited by: 1) its lack of representativeness for speech data in most languages, as it is perfectly balanced (20 hours per accent, 8 accents in total), 2) the fact that it is no longer openly available, and 3) unclear speaker composition. A more recent benchmark, CommonAccent (Zuluaga-Gomez et al., 2023), uses a subset of Common Voice (Ardila et al., 2020), which is more representative of general-purpose speech data. However, it also lacks clarity regarding speaker composition in its training/validation/testing sets. Our examination of the processing scripts reveals an overlap of speakers across these sets. The extent to which speaker-accent entanglement impacts AID performance remains largely unexplored, particularly when no effort is made to separate unseen speakers for testing.

In ZS-TTS, the closest to our work are Zhang et al. (2023a,b) and Lyth and King (2024). Zhang et al. (2023a,b) adapt a pretrained Tacotron 2-based (Shen et al., 2018)

ZS-TTS, with accent ID as input and AID as auxiliary training objective, to perform zero-shot generation for seen accents. However, their work is limited by: 1) the inability to generate unseen accents, 2) the use of limited TTS data for learning accent embeddings, 3) the reliance on pre-collected accent labels in TTS data, and 4) a lack of disentanglement between accent and speaker. Lyth and King (2024) train an AID to pseudo-label the data and then use pseudo-generated text descriptions of the speech to control different attributes (incl. accent) in text-guided ZS-TTS. However, their work is: 1) close-sourced, with no accent generation in its open-source reproduction, Parler-TTS¹, 2) unclear about how the AID is trained, susceptible to speaker-accent entanglement, 3) disregarding the continuous nature of accents with pseudo-labelled discrete accent labels as TTS input condition, and 4) unable to disentangle and separately control speaker and accent in speech generation.

To overcome the above limitations, we first propose to obtain pretrained accent embeddings from an improved AID model with speaker-accent disentanglement, termed generalisable accent identification across speakers (GenAID). This approach offers several benefits: 1) leveraging more non-TTS data to cover more speakers and accents, 2) constructing a robust accent space for even unseen accents, 3) treating accents as continuous with varying embeddings across different utterances and speakers of the same accent label, and 4) achieving greater generalisability across speakers, as the name GenAID suggests.

We then propose to condition a pretrained YourTTS-based (Casanova et al., 2022) ZS-TTS on these pretrained accent embeddings, named AccentBox. AccentBox is capable of high-fidelity zero-shot accent generation and offers several advantages: 1) leveraging continuous, speaker-agnostic GenAID embeddings, 2) capable of generating unseen accents, 3) no reliance on pre-collected accent labels in TTS data, and 4) providing separate control over speaker and accent in speech generation.

1.5 Contributions: GenAID & AccentBox

To summarise, our contributions in GenAID and AccentBox are three-fold:

- **Problem Identification** To the best of our knowledge, we are the first to 1) verify and quantify the *speaker-accent entanglement* issue in AID data/model, and 2) highlight the *accent mismatch/hallucination* issue in ZS-TTS.

¹<https://github.com/huggingface/parler-tts>

- **Novel Insights** We introduce novel speaker-accent disentanglement with information bottleneck and adversarial training in AID. We propose the task zero-shot accent generation and set the first benchmark for such task, unifying FAC, accented TTS, and ZS-TTS.
- **SOTA Performances** We achieve SOTA results in both AID (0.56 f1 score on unseen speakers in 13-accent classification by GenAID) and zero-shot accent generation (57.4%-70.0% accent similarity preference across inherent/cross accent generation against strong baselines by AccentBox).

1.6 Roadmap

This thesis is organised as follows. Chapters 2 and 3 detail the proposed GenAID and AccentBox respectively. Discussion, conclusions, and future work are presented in Chapters 4 and 5.

Chapter 2

GenAID: Generalisable Accent Identification across Speakers

2.1 Overview

In this chapter, we introduce the first stage of our work, Generalisable Accent Identification across Speakers (GenAID). Section 2.2 lists out the problems with current speech datasets on accents, the steps we take to curate a dataset for English AID, and the justification for each step. Section 2.3 illustrates the reproduction of the SOTA baselines by Zuluaga-Gomez et al. (2023) and identifies two key limitations these baselines possess: *intrinsic speaker-accent entanglement* and *bias towards more common accents*. Motivated by the identified problems and research question, we propose several modifications, with specific methods, experimental design, and results in Section 2.4, 2.5, and 2.6 respectively. Final conclusions and future work are presented in Section 2.7.

2.2 Data

2.2.1 Existing Datasets

Ideally, AID datasets should encompass a broad and balanced range of accents, with sufficient speakers and utterances for each accent. Table 2.1 lists, to the best of our knowledge, the largest datasets available for multi-accent identification research. AESRC2020 is no longer freely available. EDACC, L2-ARCTIC, and VCTK have limited speakers for most accents, making it challenging to train models that gener-

Corpus	Type	L1/L2	Accent Labels	#Accents	#Speakers per Accent
CommonAccent (Zuluaga-Gomez et al., 2023)	read speech, for ASR	both	self-reported by speaker	16	30+~6000+
EDACC (Sanabria et al., 2023)	conversation, for ASR	both	exhaustive questionnaire	40+	1~10
AESRC2020* (Shi et al., 2021)	read speech, for ASR	both	country of origin	10	3000+
L2-ARCTIC (Zhao et al., 2018)	read speech, for TTS	L2	assigned by expert	6	4
VCTK (Yamagishi et al., 2012)	read speech, for TTS	L1	assigned by expert	11	1~33

Table 2.1: Information of multi-accent English speech corpora with accent labels.

*: Note that AESRC2020 corpus is no longer openly available.

alise well to unseen speakers. Additionally, EDACC consists of conversational speech which mismatches the read speech generation task in later stage. The only viable option, CommonAccent, is derived from Common Voice (Ardila et al., 2020) using the portion with self-reported accent labels. Despite its broad coverage of accents and speakers/utterances for each accent, CommonAccent disregards speaker information when splitting training/validation/testing sets, resulting in overlap of speakers between training and validation/testing sets. Testing the models on speakers partially seen during training is flawed, so we reprocess CommonAccent to suit our needs.

2.2.2 Data Selection & Processing

We make the following modifications to the original CommonAccent processing pipeline to derive a multi-accent speech dataset.

1) Larger-Scale Data from Latest Common Voice To obtain larger-scale and higher-quality data, we use the latest English portion of Common Voice version 17.0¹, instead of 7.0 by Zuluaga-Gomez et al. (2023).

2) Dividing Validation/Testing Sets by Seen and Unseen Speakers To evaluate the performances of AID models on both seen and unseen speakers, we create separate validation/testing sets for seen and unseen speakers. The unseen speakers sets do not

¹<https://commonvoice.mozilla.org/en/datasets>

overlap with the training set in terms of speakers.

3) Filter out Accents with Insufficient Number of Speakers To train an AID model that generalises well to unseen speakers, we exclude accent labels with insufficient speakers. Remaining accents shall have: 1) at least 10 speakers with 50 utterances each (for training data and validation/testing on seen speakers), and 2) 20 additional speakers with at least 10 utterances each (for validation/testing on unseen speakers).

4) Balancing the Number of Utterances Across Speakers To prevent biasing the AID model towards certain speakers, we allow a maximum of 30 utterances per speaker in

Accent	Training		Validation		Testing	
			Seen Spks	Unseen Spks	Seen Spks	Unseen Spks
	#Spks	#Utrr (total)	#Spks ×#Utrr	#Spks ×#Utrr	#Spks ×#Utrr	#Spks ×#Utrr
American	6,129	78,199	948×10	10×10	948×10	10×10
English	1,737	22,481	274×10	10×10	274×10	10×10
Canadian	695	10,266	165×10	10×10	165×10	10×10
Australian	472	6,952	91×10	10×10	91×10	10×10
Irish	127	1,412	21×10	10×10	21×10	10×10
Scottish	108	1,377	23×10	10×10	23×10	10×10
New Zealand	101	1,116	17×10	10×10	17×10	10×10
South Asian	1,658	16,595	178×10	10×10	178×10	10×10
Southern African	199	2,328	36×10	10×10	36×10	10×10
Hong Kong	84	735	13×10	10×10	13×10	10×10
Filipino	77	1,103	24×10	10×10	24×10	10×10
Malaysian*	57	416	3×10	10×10	3×10	10×10
Singaporean*	43	381	8×10	9×10	8×10	10×10
TOTAL	11,487	143,361	18,010	1,290	18,010	1,300

Table 2.2: Data composition of final processed training/validation/testing sets, incl. 7 L1 accents (top) and 6 L2 accents (bottom). “Spks” - speakers; “Utrr” - utterances.

*: These two accents do not strictly meet the requirements, but are included for sufficient accent classes.

Accent	Training	Validation		Testing	
		Seen	Unseen	Seen	Unseen
		Spks	Spks	Spks	Spks
American	122.4	14.36	0.16	14.44	0.16
English	35.4	4.18	0.17	4.21	0.17
Canadian	15.8	2.47	0.16	2.50	0.15
Australian	10.8	1.36	0.16	1.33	0.16
Irish	2.3	0.34	0.15	0.34	0.15
Scottish	2.2	0.36	0.18	0.36	0.17
New Zealand	1.8	0.27	0.16	0.28	0.15
South Asian	27.2	2.81	0.16	2.83	0.17
Southern African	3.8	0.58	0.16	0.58	0.17
Hong Kong	1.2	0.19	0.17	0.19	0.16
Filipino	1.7	0.36	0.16	0.37	0.17
Malaysian	0.7	0.05	0.16	0.05	0.16
Singaporean	0.6	0.13	0.15	0.11	0.16
TOTAL	225.9	27.46	2.10	27.59	2.10

Table 2.3: Duration information of final processed training/validation/testing sets.

All duration info is calculated in hour(s).

the training set. All speakers excluded from the unseen speakers validation/testing sets are included in training to improve speaker coverage of the training data.

The composition of the final processed data is shown in Table 2.2, with its duration information in Table 2.3. Accent labels are cleaned from the self-reported labels by speakers (see Appendix A for details). Most of these accent labels do not contain granular accent variety information (e.g. Received Pronunciation, Leeds accent) and are just coarse region/country-level accent information (e.g. Singaporean, Malaysian).

2.2.3 Limitations

1) English Accents Only Most existing datasets focus on English, limiting our study. We will test our approach on other languages, if sufficient data becomes available.

2) Inherent Problems of Accent Labels Labelling speech with discrete accent labels is an ill-formed task considering the continuous nature of accents (Lyth and King,

2024). Common Voice let speakers self-report their accents, and AESRC2020 uses each speaker’s country of origin – both methods lack precision in defining accents (Sanabria et al., 2023). We recognise the inherent problems of these accent labels, however, to ignore accents in speech technology research would be unacceptable. Therefore, we leave such accent definition and labelling problem for future work.

3) Limited Coverage of Accents As shown in Tables 2.2 and 2.3, the final dataset still covers a limited number of accents. This limitation stems from the lack of accent diversity in the English portion of Common Voice. Although other L2 learner speech corpora (listed in Table 2.4) exist, each focuses on a single accent and requires unique processing to be integrated into a multi-accent dataset. As such, expanding data size and coverage by other datasets listed in Tables 2.1 and 2.4 is reserved for future work.

Corpus	Accent
BELC ² (Muñoz, 2006)	Barcelona
CUHK Corpus ³ (MacWhinney, 2017)	Chinese
Corpus PAROLE ⁴ (Hilton, 2009)	French
Dresden Corpus ⁵ (Kubaneck-German, 2000)	German
Connolly Corpus ⁶ (Green and Green, 1993; Worthington, 1997)	Japanese

Table 2.4: Information of single-accent L2 English speech corpora.

4) Imbalanced Coverage of Accents As shown in Tables 2.2 and 2.3, the most scarce accent has only 43 speakers / 0.6 hours in the training set while the most common accent has 6,129 speakers / 122.4 hours. Unfortunately, this imbalance is unavoidable during data collection due to the varying number of speakers for each accent. Downsampling data of the more common accents may unnecessarily harm overall performance. Thus, we leave such label imbalance for modelling techniques to mitigate, to be introduced in Section 2.4.

²<https://slabank.talkbank.org/access/English/BELC.html>

³<https://slabank.talkbank.org/access/English/CUHK.html>

⁴<https://slabank.talkbank.org/access/English/PAROLE.html>

⁵<https://slabank.talkbank.org/access/English/Dresden.html>

⁶<https://slabank.talkbank.org/access/English/Connolly.html>

2.3 Problem Identification

2.3.1 Hypothesised Problems

The previous section on curating the English AID dataset highlights two key problems that AID models may suffer:

1) Intrinsic Speaker-Accent Entanglement Despite our efforts above to include more speakers for each accent, still each speaker has one corresponding accent, leading to intrinsic entanglement between speaker and accent. Failure to address such entanglement could cause the AID models to generalise poorly across speakers.

2) Bias Towards More Common Accents The imbalanced coverage of accents, without proper handling, may bias the AID models towards predicting more common accents to quickly achieve lower loss and higher accuracy.

2.3.2 Reproduction of Baselines

Motivated by aforementioned data issues and hypothesised problems, we reproduce the baselines by Zuluaga-Gomez et al. (2023) and test them on our newly curated dataset to verify these problems. Zuluaga-Gomez et al. (2023) finetune both ECAPA-TDNN (Desplanques et al., 2020) and XLSR (Babu et al., 2022) with an accent classification layer on top, setting the most recent benchmark on AID. Despite our best efforts to faithfully reproduce their work, there are two differences:

1) Mismatch in Dataset We could not train/validate/test on the same dataset as the authors because: 1) the overlap of speakers in their training and validation/testing sets which hides the problem we seek identification and verification; 2) the provided processing script cannot exactly reproduce the data used in training/validating/testing the reported AID models.

2) Mismatch between Code and Paper We use the open-source code⁷ by the author and adhere to its implementation, which slightly differs from the paper. Specifically, the ECAPA-TDNN-based model is initialised from a Speaker Verification model⁸ rather than a Language Identification model⁹; and the XLSR-based model uses average pooling rather than statistical pooling when pooling frame-level embeddings to an utterance-

⁷<https://github.com/JuanPZuluaga/accent-recog-slt2022/tree/main>

⁸<https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb>

⁹https://huggingface.co/speechbrain/lang-id-commonlanguage_ecapa

level embedding for classification. These deviations do not significantly affect the performance of models in our reproduction.

2.3.3 Results & Analysis of Baselines

Baseline	Reproduced Results				Reported by CommonAccent
	Seen Spks		Unseen Spks		Mixed
	f1	acc	f1	acc	acc
ECAPA-TDNN-based (#E1 baseline vs reported)	0.76	0.85	0.26	0.29	0.79*
XLSR-based (#X1 baseline vs reported)	0.95	0.96	0.40	0.43	0.95*

Table 2.5: Reproduced and reported results of AID baselines.

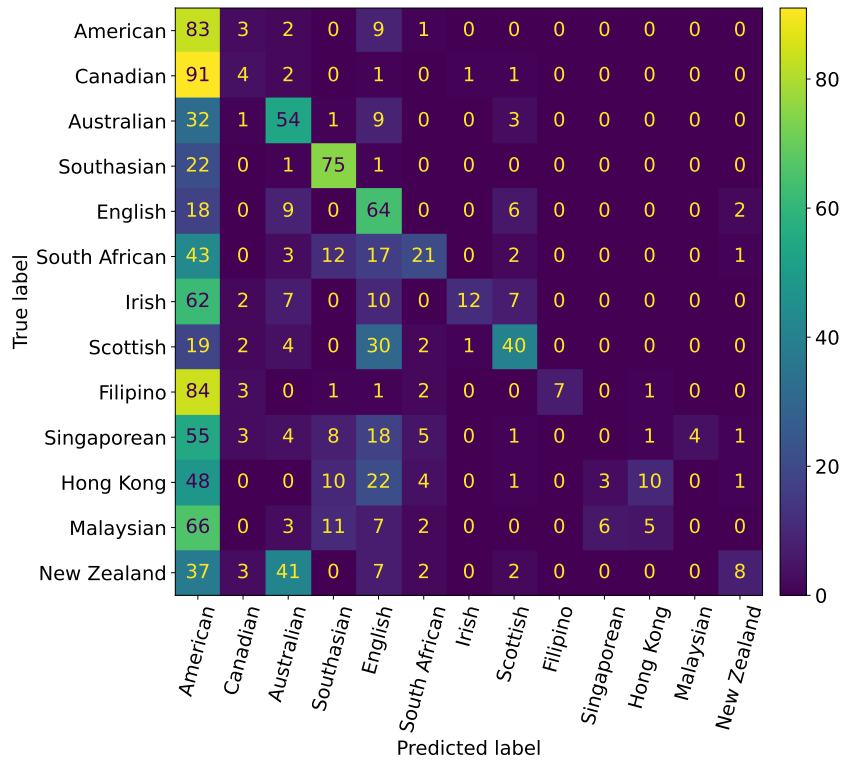
*: These numbers are taken directly from Zuluaga-Gomez et al. (2023).

“Spks” - speakers; “acc” - accuracy; “f1” - macro-average f1 score across accents.

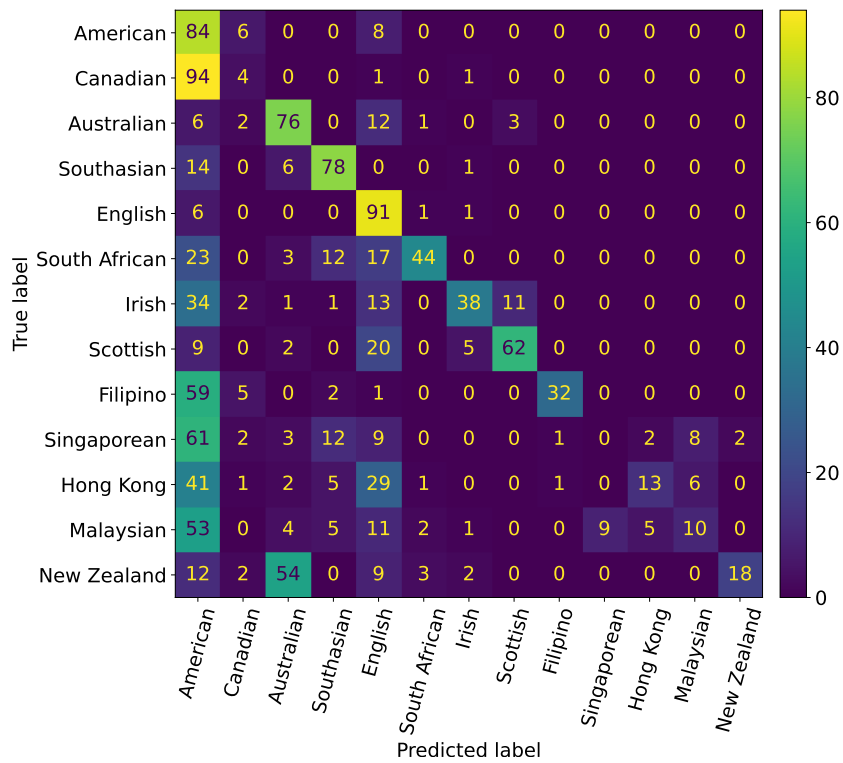
We successfully reproduce the benchmark classification performance with 0.76/0.95 macro-f1 scores on seen speakers (see Table 2.5), which resembles the mixed testing set (mostly seen speakers) used by the authors. However, when tested on unseen speakers, performance drastically drops to 0.26/0.40 macro-f1 scores. This indicates that these models have poor generalisation across speakers and are memorising the speaker-to-accent mapping rather than learning to discriminate accents from speech. The huge gap in performances between seen and unseen speakers verifies the first hypothesised problem that these AID models do not generalise well across speakers.

We further analyse the predictions and find they are heavily biased towards more common accents, verifying the second hypothesised problem. Figure 2.1 show the confusion matrices of the reproduced systems on the unseen speakers testing set. Both systems exhibit severe biases, predicting a vast majority of utterances as having an American accent.

Finally, we visualise the embeddings learned by XLSR-based baseline using t-SNE (van der Maaten and Hinton, 2008), as shown in Figure 2.2. On the seen speakers testing set, we randomly select 10 speakers if the accent contains more speakers, to balance between accents of various data size and avoid biasing the t-SNE model. We success-

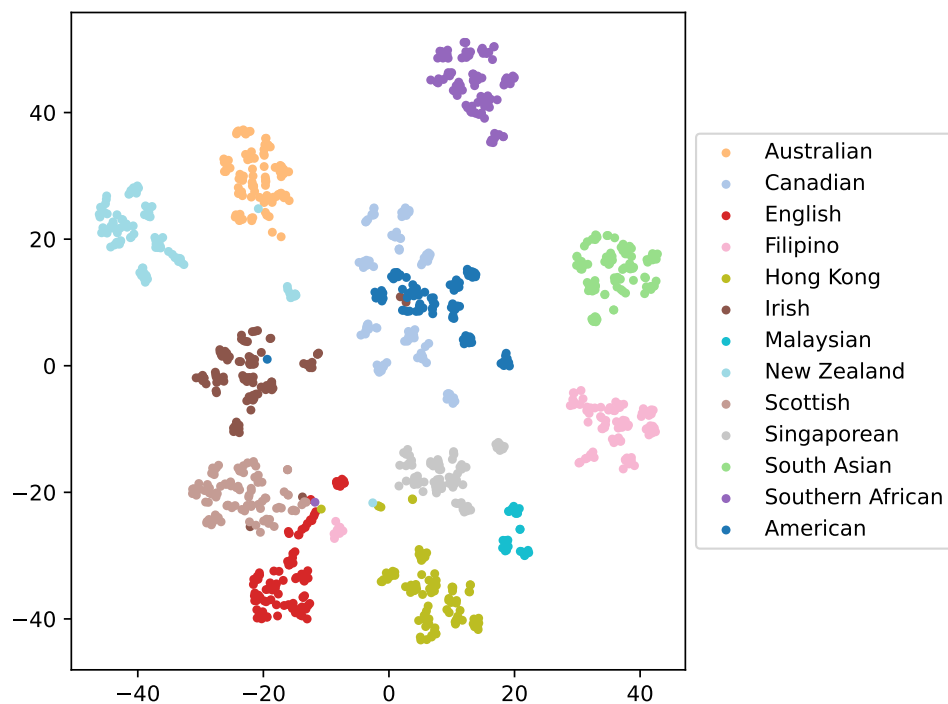


(a) ECAPA-TDNN-based baseline, #E1 baseline.

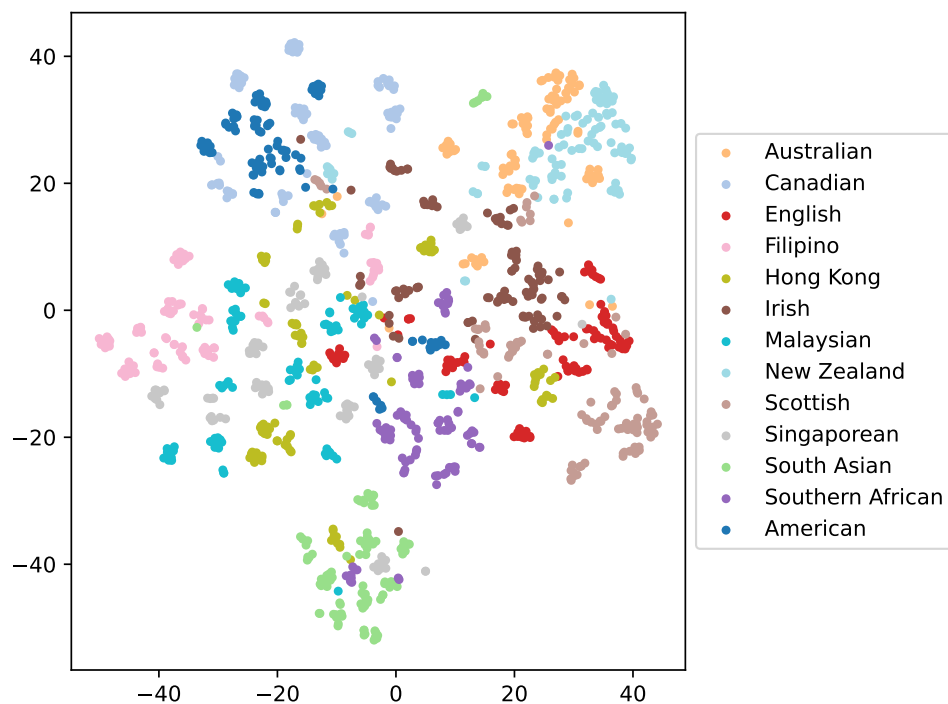


(b) XLSR-based baseline, #X1 baseline.

Figure 2.1: Confusion matrices of two reproduced baselines, showing biased predictions - most predicted labels are “American”.



(a) XLSR-based, #X1 baseline, on seen speakers.



(b) XLSR-based, #X1 baseline, on unseen speakers.

Figure 2.2: T-SNE visualisation of embeddings by XLSR-based baseline

#X1 baseline on seen and unseen speakers,

showing entanglement of speakers and accents.

fully reproduce the clearly separable accent clusters on seen speakers (no fault in our reproduction); however, the visualisation on unseen speakers reveals the baseline’s actual accent discrimination ability in real scenarios and exhibits strong speaker-accent entanglement (with each tiny cluster comprising utterances from the same speaker).

2.3.4 Research Question

The problems identified and verified in the previous sections highlight the need for a better AID model. Current benchmark approaches are flawed, and directly applying the learned representations from these AID models to the later stage, i.e. zero-shot accent generation, would lead to disastrous error propagation - the ZS-TTS models would take a problematic accent embedding extracted from one audio clip by an unseen speakers as input.

In this first stage, the formal research question we ask is: *How can we extract accent embeddings that are more discriminative of accents and less influenced by other speech factors (e.g. speaker, channel, content, etc.)?* Addressing the problem of biased prediction could improve the discriminative ability of accents, while tackling the poor generalization across speakers could enhance both accent discrimination and the removal of other speech factors in the learned embeddings.

2.4 Methods

2.4.1 Overview

In this section, five improvement techniques are proposed for a more generalisable AID across speakers. Sections 2.4.2 and 2.4.3 present two training modifications. Section 2.4.4 details the data augmentation. Sections 2.4.5 and 2.4.6 introduce the information bottleneck and adversarial training added to the model architectures, for better speaker-accent disentanglement. The final proposed systems are shown in Figure 2.3. We experiment all five techniques accumulatively with both ECAPA-TDNN-based and XLSR-based baselines (#E1 and #X1), reproduced in Section 2.3.2.

2.4.2 Validation on Unseen Speakers

Selecting the best checkpoint based on a validation set containing seen speakers can lead to overfitting. As training progresses, the model may show increasing accuracy

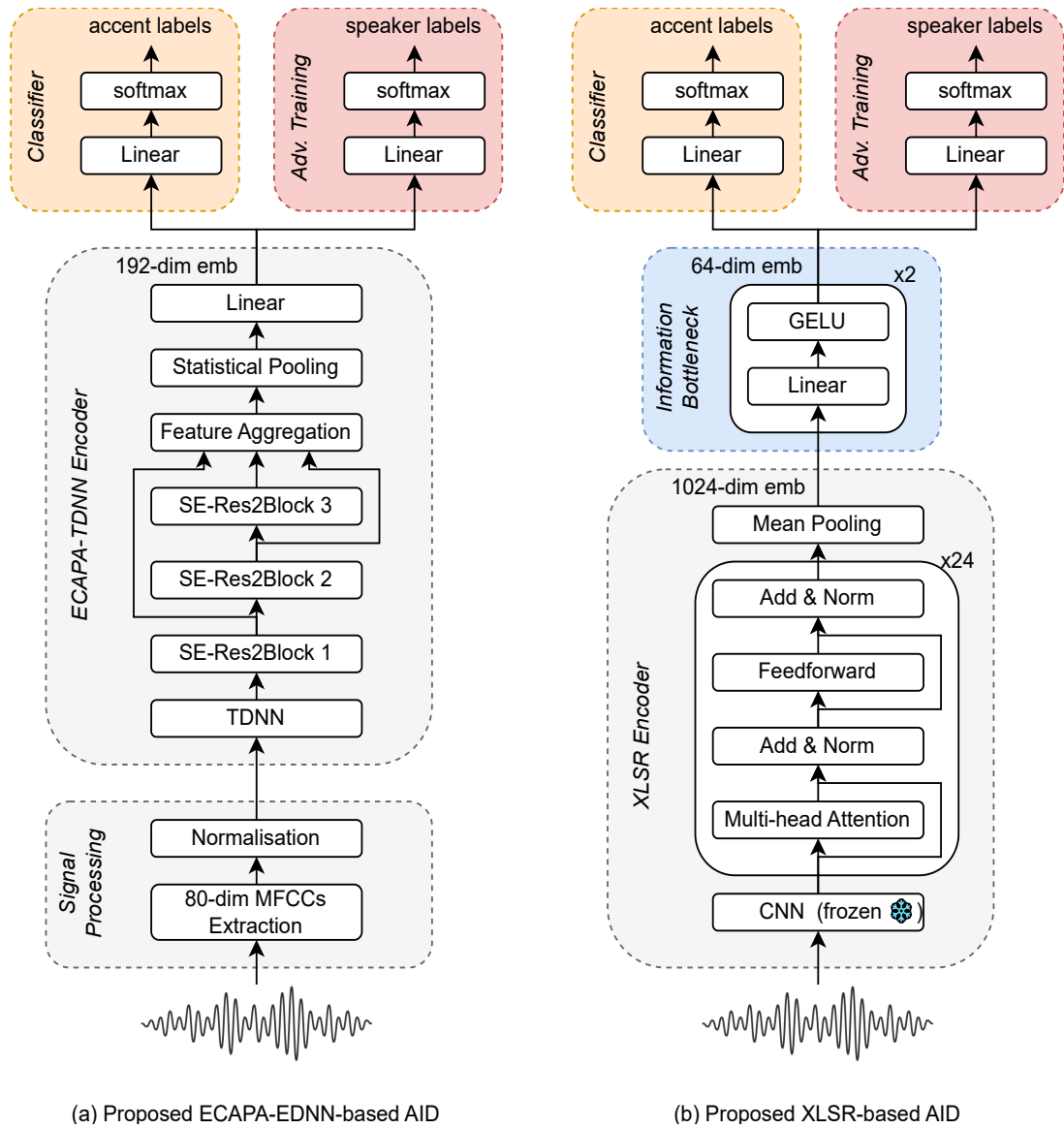


Figure 2.3: Model architectures of proposed AID systems (GenAID).

“Adv.” - Adversarial; “dim” - dimension; “emb” - embedding.

on the validation set by memorising the speaker-to-accent mapping without learning to discriminate accents. Therefore, we validate the models on unseen speakers. The resulted early stopping during training prevents the model from overfitting on seen speakers and sacrificing the generalisation ability across speakers.

2.4.3 Weighted Sampling

Both data composition (see Tables 2.2 and 2.3) and baseline model predictions (see Figure 2.1) demonstrate the severe class imbalance problem facing AID models. One

common solution is to use a weighted Cross Entropy loss where less common labels receives higher weights when calculating the loss (Ho and Wookey, 2020). However, some accent labels are so scarce that they appear in less than 0.3% of the batches. Updating model parameters at such few steps are ineffective, even with a higher loss incurred. Therefore, we apply weighted sampling instead, to ensure equal probability of sampling each accent’s data in each batch (Ling and Li, 1998). The sampling weights are the inverse frequency of each accent in the data. We hypothesise that weighted sampling would effectively mitigate biased prediction.

2.4.4 Data Augmentation by Perturbation

The large multi-accent dataset obtained in Section 2.2 is still limited in covering various speech factors (e.g. recording device, recording environment, speaking rate, etc.). This limitation hinders the performance of AID models and confounds learned accent embeddings with other speech factors. We augment the data by conducting both speed (Ko et al., 2015) and noise perturbation (Ko et al., 2017) to improve the generalisation of AID models across various scenarios. For each utterance, we generate a perturbed version by: 1) randomly changing the speech rate to one of $\{\times 0.95, \times 1, \times 1.05\}$, and 2) adding noise and reverberation with a random Signal-to-Noise Ratio (SNR) of 0-15, using the OpenRIR¹⁰ dataset.

2.4.5 Information Bottleneck

Despite the training modifications and data augmentation above, the learned embeddings still contain unnecessary information that interferes with the AID task, especially for the XLSR-based AID model where much information is learned during XLSR pre-training. Let the input speech signal be x and the ECAPA-TDNN or XLSR Encoder shown in Figure 2.3 be $\text{Encoder}(\cdot)$, the learned embedding h is obtained by:

$$h = \text{Encoder}(x). \quad (2.1)$$

h is then passed to a classifier that outputs the normalised probability over all accent labels $p(y_{acc})$:

$$p(y_{acc}) = \sigma(\text{Linear}(h)), \quad (2.2)$$

where σ is the softmax activation. Inspired by Qian et al. (2019, 2020), we construct an information bottleneck, denoted $\text{Bottleneck}(\cdot)$, that projects the encoder output h into

¹⁰<http://www.openslr.org/28>

a low-dimensional space h' which contains less information and may better disentangle various speech factors, expressed as:

$$h' = \text{Bottleneck}(h), \quad (2.3)$$

$$p(y_{acc}) = \sigma(\text{Linear}(h')), \quad (2.4)$$

where $|h'| < |h|$, i.e. the size of embedding h' is smaller than that of embedding h . The $\text{Bottleneck}(\cdot)$ we adopt is a two-layer Multi-Layer Perceptron (MLP) with GELU activation (Hendrycks and Gimpel, 2016).

2.4.6 Adversarial Training

Gradient Reversal Layer (GRL), introduced by Ganin et al. (2016) for domain adaptation, is widely used for information disentanglement and removal. GRL works by passing the target embedding (e.g. accent embedding) to an auxiliary classifier which learns to discriminate the information to be removed (e.g. speaker identity). Then, during backpropagation, the gradients are reversed (i.e. gradient ascent rather than descent), encouraging the target embedding to be invariant to the information being removed (e.g. speaker-agnostic). The total loss \mathcal{L} is expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{acc_clf}} - \alpha \cdot \mathcal{L}_{\text{spk_clf}} \quad (2.5)$$

where $\mathcal{L}_{\text{acc_clf}}$ denotes the loss for accent classification, $\mathcal{L}_{\text{spk_clf}}$ denotes the loss for speaker classification, and α denotes a positive hyperparameter to balance the scale of losses. However, GRL can be challenging when applied to disentangling speaker and accent in AID, due to *unstable training* and *problematic training objective*.

- The model is encouraged to minimise the probability of the true speaker label y_{spk}^* given the input signal x , expressed as:

$$\arg \min_{\theta} P(y_{spk} = y_{spk}^* | x; \theta), \quad (2.6)$$

where θ denotes the learnable parameters of the model. For any initial parameters θ_0 not trained on the speaker verification/identification task, $P(y_{spk} = y_{spk}^* | x; \theta_0)$ is typically extremely low, especially with numerous speaker labels $|y_{spk}| > 10,000$. This low probability incurs a large negative $\mathcal{L}_{\text{spk_clf}}$, even at the initial stage of training, causing *unstable training*.

- Moreover, as training progresses, minimising $\mathcal{L}_{\text{spk_clf}}$ and $P(y_{\text{spk}} = y_{\text{spk}}^* | x; \theta)$ is not meaningful for three reasons. 1) Reducing the target speaker probability $P(y_{\text{spk}} = y_{\text{spk}}^* | x; \theta)$ from 10^{-2} to 10^{-10} for instance, does not necessarily improve speaker disentanglement. 2) A model that consistently predicts all speaker labels incorrectly still indicates some learning of speaker information (an analogy is that you definitely know something about the correct answers to score 0% in a multiple-choice exam). 3) Even with extremely low $P(y_{\text{spk}} = y_{\text{spk}}^* | x; \theta)$, the model might still predict a high $P(y_{\text{spk}} = y'_{\text{spk}} | x; \theta)$ where y'_{spk} is a speaker similar to the true speaker y_{spk}^* . All of these reasons combined, point to the *problematic training objective* in GRL.

To address these limitations, we propose training the model to be maximally uncertain about speaker information, inspired by Webber et al. (2020) in their work of voice anonymisation. This is achieved using a Mean Square Error (MSE) loss \mathcal{L}_{MSE} between the predicted distribution of speaker labels $p(y_{\text{spk}} | x; \theta)$ and an even distribution across all speakers $\mathcal{U}(|y_{\text{spk}}|)$. This prevents *unstable training* by ensuring the adversarial loss is not excessively negative, and it corrects *problematic training objective* by aiming for uniform uncertainty about speakers. The total loss function is now expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{acc_clf}} + \alpha \cdot \mathcal{L}_{\text{MSE}}[p(y_{\text{spk}} | x; \theta), \mathcal{U}(|y_{\text{spk}}|)]. \quad (2.7)$$

2.5 Experiments

2.5.1 Systems

To address our research question and evaluate the proposed methods in Section 2.4, we sequentially apply these five modifications on both baselines. Modifications are retained for subsequent experiments if they result in improvement. The baselines, reproduced in Section 2.3.2, are denoted as #E1 baseline and #X1 baseline. For ECAPA-TDNN-based systems, four of the five modifications, excluding the information bottleneck, show improvement in performance and are accumulatively added as #E2 to #E5 systems. For XLSR systems, all five modifications show improvement in performance and are accumulatively added as #X2 to #X6 systems (see Table 2.7).

2.5.2 Configurations

1) Pretraining Following Zuluaga-Gomez et al. (2023), we initialise ECAPA-TDNN-based models from a Speaker Verification model¹¹, and XLSR-based models from XLSR-large¹². All model parameters are unfrozen in AID finetuning, except for the bottom CNN layers in XLSR Encoder, shown in Figure 2.3.

2) Audio Processing To be consistent with pretrained models, all waveforms are down-sampled to 16 kHz as input.

3) Classification Loss To be consistent with pretrained models, we use Additive Angular Margin (AAM) loss (Xiang et al., 2019) for ECAPA-TDNN-based models and Cross Entropy loss for XLSR-based models in accent classification, i.e. $\mathcal{L}_{\text{acc_clf}}$ in Equation 2.7.

4) Hyperparameters Tuning We use the maximum batch size available on a single GPU, 24 for ECAPA-TDNN-based and 12 for XLSR-based models. For all models, we finetune the hyperparameters shown in Table 2.6, and choose the best checkpoint based on the highest classification accuracy on unseen speakers validation set. The best ECAPA-TDNN-based system (#E5) is trained with a learning rate of 5e-5, no bottleneck, and α of 1e-2. The best XLSR-based system (#X6) is trained with a learning rate of 1e-4, bottleneck of 64 dimension, and α of 10.

Hyperparameters	Values
learning rate	1e-4, 5e-5, 2e-5, 1e-5
bottleneck dimension	1024, 192, 64, 32
α for adversarial training in ECAPA-TDNN-based AID	5e-2, 1e-2, 5e-3, 1e-3
α for adversarial training in XLSR-based AID	1, 10, 100

Table 2.6: Values of tuned hyperparameters.

5) Training Environment All models are trained on a single NVIDIA Tesla V100-SXM2-16GB GPU, with maximum 30 epochs for ECAPA-TDNN-based and 10 epochs for XLSR-based models.

¹¹<https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb>

¹²<https://huggingface.co/facebook/wav2vec2-large-xlsr-53>

2.5.3 Evaluation

1) Classification Metrics We evaluate AID performance using precision, recall, f1 score, and accuracy. For the seen speaker testing sets, the *macro-average* of precision, recall, and f1 score across all accents are reported, to avoid the influence of class imbalance. We also report the gap of f1 scores and accuracies between seen and unseen speakers to reflect the system’s generalisation ability (the smaller the gap, the better the generalisation across speakers).

2) T-SNE Visualisation To visualise the speaker and accent information captured by the AID models, we extract the latent embeddings before the final classification layer of each model, i.e. h in Equation 2.2 or h' in Equation 2.4 after bottleneck is applied, for all utterances in the unseen speaker testing set. These embeddings are then passed to t-SNE (van der Maaten and Hinton, 2008) and visualised.

3) Silhouette Coefficient for Speaker Clusters (SCSC) To quantify residual speaker information, we group embeddings of each accent label by speaker, and calculate the Silhouette coefficient (Rousseeuw, 1987) for these speaker clusters. For each data point i in speaker cluster C_I , the mean intra-cluster distance a is calculated as:

$$a = \frac{1}{|C_I| - 1} \sum_{j \in C_I, j \neq i} \text{dist}(i, j) \quad (2.8)$$

where j is another data point in the same speaker cluster C_I , and $\text{dist}(\cdot)$ represents the Euclidean distance between embeddings. The mean nearest-cluster distance b is calculated as:

$$b = \min_{J \neq I} \frac{1}{|C_J|} \sum_{k \in C_J} \text{dist}(i, k) \quad (2.9)$$

where k is a data point in a different speaker cluster C_J . The Silhouette score s of data point i is thus:

$$s = \frac{(b - a)}{\max(a, b)} \quad (2.10)$$

The Silhouette coefficient is the mean of Silhouette scores across all data points, ranging between $[-1, 1]$. A higher Silhouette coefficient indicates well-separated less overlapping clusters, while a lower Silhouette Coefficient for Speaker Clusters (SCSC) suggests that the residual speaker information in the learned embeddings is less, with more overlapping between speaker clusters.

2.6 Results & Analysis

2.6.1 Overall

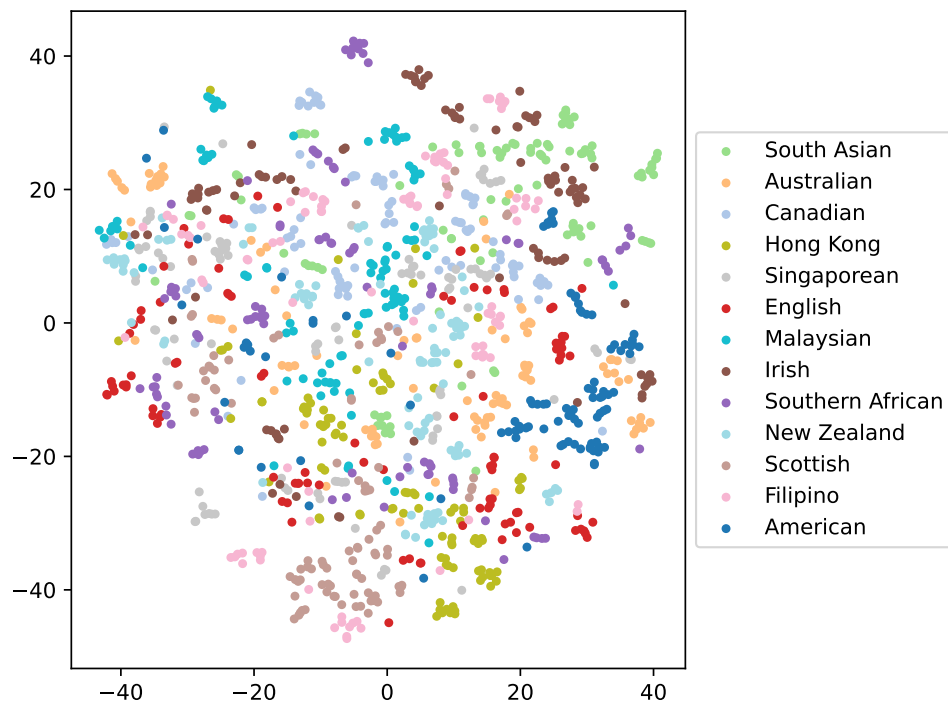
AID Systems	Seen Speakers				Unseen Speakers \uparrow				Gap \downarrow	
	prec	rec	f1	acc	prec	rec	f1	acc	f1	acc
#E1 baseline	0.85	0.72	0.76	0.85	0.46	0.29	0.26	0.29	0.50	0.56
#E2 w/ valid on unseen	0.80	0.71	0.73	0.83	0.49	0.32	0.29	0.32	0.44	0.51
#E3 w/ weighted sampler	0.72	0.91	0.80	0.81	0.46	0.43	0.41	0.43	0.39	0.38
#E4 w/ perturbation	0.53	0.88	0.61	0.62	0.48	0.50	0.47	0.50	0.14	0.12
#E5 w/ adv. training	0.56	0.88	0.63	0.58	0.49	0.50	0.47	0.50	0.16	0.08
#X1 baseline	0.97	0.94	0.95	0.96	0.56	0.43	0.40	0.43	0.55	0.53
#X2 w/ valid on unseen	0.88	0.81	0.82	0.86	0.57	0.47	0.45	0.47	0.37	0.39
#X3 w/ weighted sampler	0.75	0.87	0.77	0.58	0.56	0.47	0.46	0.47	0.31	0.11
#X4 w/ perturbation	0.78	0.90	0.81	0.63	0.60	0.50	0.48	0.50	0.33	0.13
#X5 w/ bottleneck(64dim)	0.66	0.87	0.73	0.66	0.61	0.56	0.55	0.56	0.18	0.10
#X6 w/ adv. training	0.73	0.89	0.78	0.62	0.63	0.56	0.55	0.56	0.23	0.06

Table 2.7: Accent identification results of AID systems.

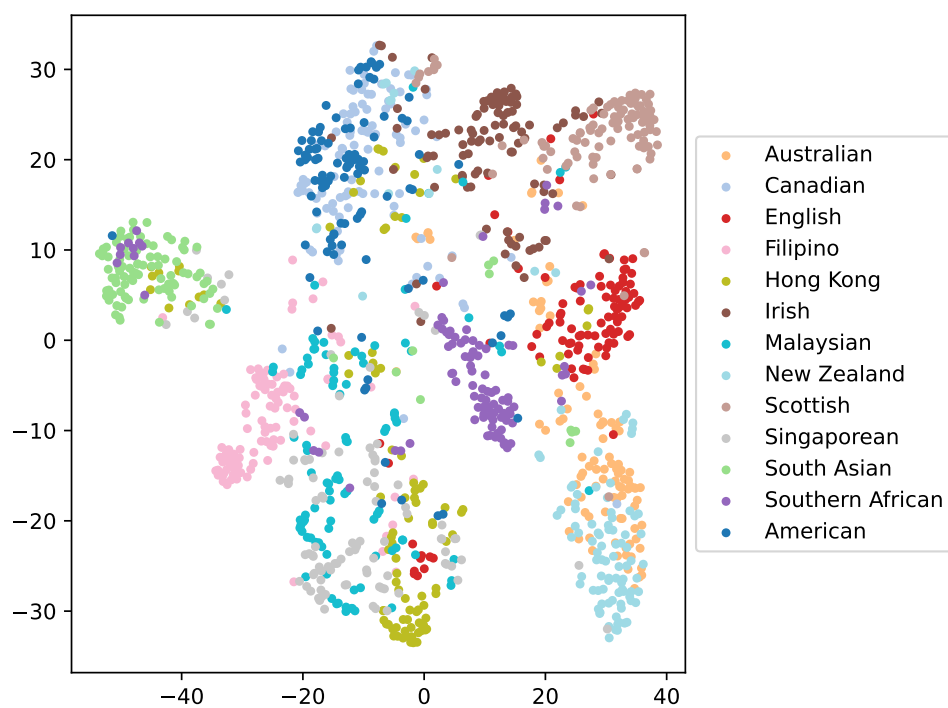
top-half: ECAPA-TDNN-based; bottom-half: XLSR-based. All “w/” changes are accumulative. “adv.” - adversarial; “prec” - precision; “rec” - recall.

Table 2.7 shows the best systems and changes from baselines. On unseen speakers which we focus on, a significant f1 score improvement is achieved: **0.21** for ECAPA-TDNN-based (#E1 vs #E5) and **0.15** for XLSR-based (#X1 vs #X6). The best system (#X6) achieves a **0.56** AID accuracy on unseen speakers, significantly better than the 0.08 random baseline. We also reduced speaker entanglement, with smaller f1 gaps between seen and unseen speakers (0.50 vs 0.16 by #E1 vs #E5 and 0.55 vs 0.23 by #X1 vs #X6). Note that high accuracy on seen speakers with a large gap to unseen speakers is not desirable. This suggests the model is memorizing speaker-accent mappings rather than learning to discriminate accents.

We visualise embeddings of the best systems on unseen speakers using t-SNE (Figure 2.4). The best XLSR-based system (#X6) shows better-separated accent clusters with and less speaker-accent entanglement compared to the XLSR baseline (#X1) in



(a) ECAPA-TDNN-based, #E5 w/ adv. training, on unseen speakers.



(b) XLSR-based, #X6 w/ adv. training, on unseen speakers.

Figure 2.4: T-SNE visualisation of embeddings by the best AID systems, showing better-separated accent clusters by the best XLSR-based system.

Figure 2.2b. By contrast, the best ECAPA-TDNN-based system, despite comparable accent classification performance, lacks well-separated accent clusters.

Weighted sampling and data augmentation by perturbation are most effective for ECAPA-TDNN-based systems, improving f1 by 0.12 and 0.06 on unseen speakers. For XLSR-based systems, the information bottleneck leads to a 0.07 f1 increase. The following sections analyse these effects, with supplementary evidence and further discussion on Self-Supervised Learning (SSL) models and accent similarity.

2.6.2 Reduced Overfitting by Validation on Unseen Speakers

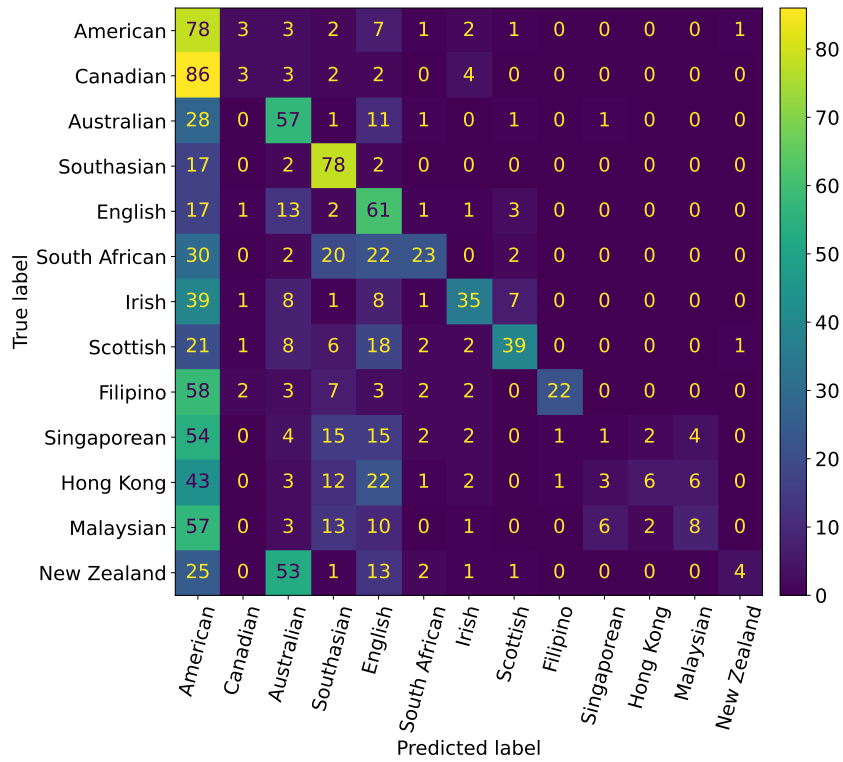
Both ECAPA-TDNN-based and XLSR-based systems exhibit modest improvements on unseen speakers after validating on unseen speakers (0.03 f1 increase by #E1 vs #E2 and 0.05 f1 increase by #X1 vs #X2). The reduced generalisation gap across speakers (0.06 f1 gap decrease by #E1 vs #E2 and 0.18 f1 gap decrease by #X1 vs #X2) confirms that this simple and classical technique effectively alleviates overfitting on seen speakers.

2.6.3 More Balanced Predictions by Weighted Sampling

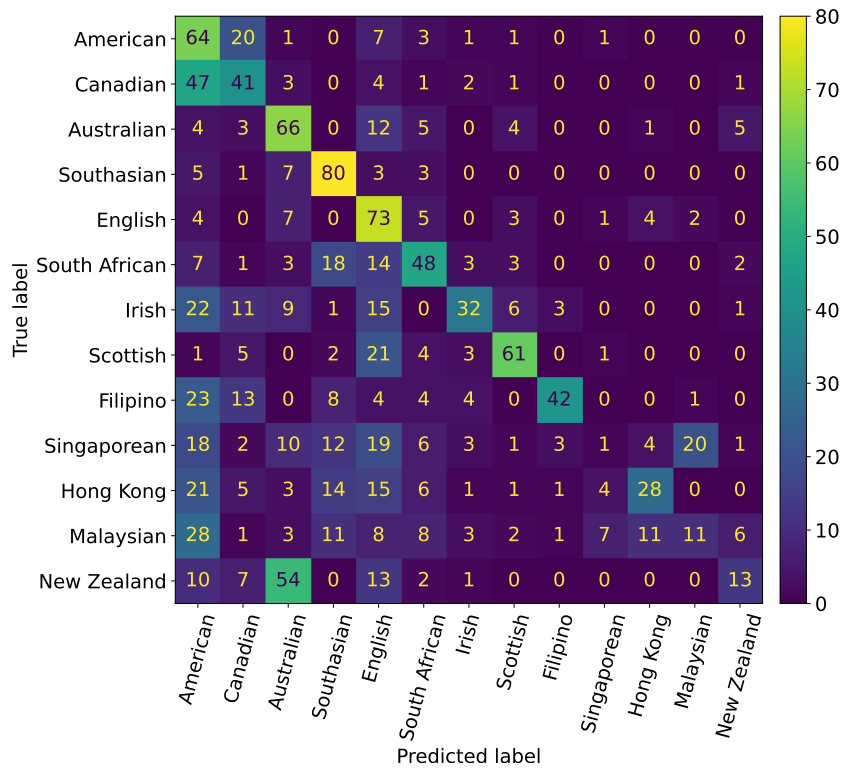
Weighted sampling significantly improves the ECAPA-TDNN-based system, with 0.12 f1 and 0.11 recall increase on unseen speakers (#E2 vs #E3). Detailed analysis, with confusion matrices shown in Figure 2.5, reveals that this improvement comes from better recall rates for scarce accent labels, reducing bias towards common accents like “American”. However, the effects on XLSR-based system is marginal, with 0.01 f1 increase on unseen speakers (#X2 vs #X3), likely due to its robustness from SSL pre-training, which exposed the model to a wide range of accents and languages.

2.6.4 Improved Generalisation by Data Augmentation

Data augmentation works effectively on ECAPA-TDNN-based system, with 0.06 f1 increase on unseen speakers and 0.25 f1 gap decrease (#E3 vs #E4), shown in Table 2.7). The improvement is likely due to the “pseudo” new speakers created by perturbation. However, the effects on XLSR-based system is less effective, with 0.02 f1 increase on unseen speakers (#X3 vs #X4), likely because SSL pretraining already provides robustness to noise and speaking rate variability, making data augmentation less impactful.



(a) #E2 w/ valid on unseen, before weighted sampling.



(b) #E3 w/ weighted sampler, after weighted sampling.

Figure 2.5: Confusion matrices of ECAPA-TDNN-based systems before and after applying weighted sampling, showing effects on debiasing predictions.

2.6.5 Effective Disentanglement by Information Bottleneck for XLSR-based AID

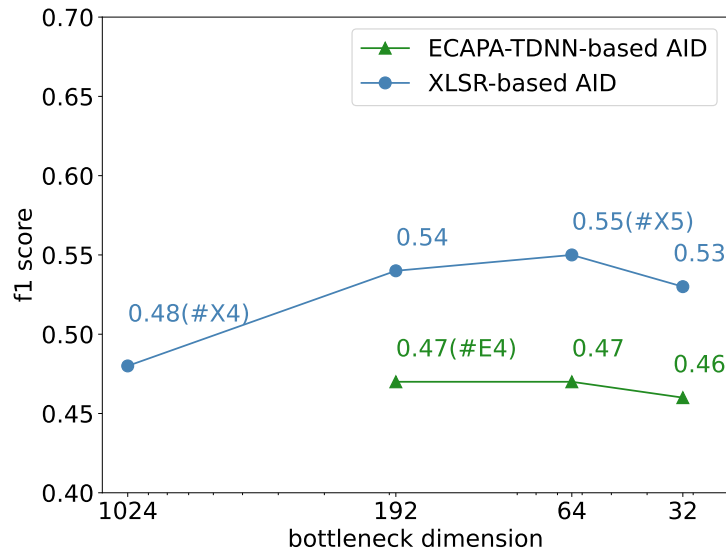


Figure 2.6: F1 scores of applying information bottleneck of different dimensions to ECAPA-TDNN-based #E4 and XLSR-based #X4, showing improvement for only XLSR-based system with 0.07 increase in f1 score at 64 dimension (#X5).

Information bottleneck does not work on ECAPA-TDNN-based system, with deteriorating f1 scores (see Figure 2.6) as the dimension of the bottleneck decreases. The ineffectiveness on ECAPA-TDNN-based system suggests that there is not much non-accent-related information that could be filtered out without harming AID performance. However, information bottleneck works very effectively on XLSR-based system, with 0.07 f1 increase at 64 dimension (see the blue line in Figure 2.6). The effects of information gap on speaker disentanglement is also shown in the reduced generalisation gap across speakers (0.15 f1 gap decrease by #X4 vs #X5 in Table 2.7). The effectiveness on XLSR-based system suggests that the bottleneck helps filter out non-accent-related information from the richer, more redundant embeddings produced by SSL pretraining.

Visualisation of all embeddings using t-SNE, shown in Figure 2.7, shows better-separated accent clusters after applying the information bottleneck. Additionally, the mean Silhouette Coefficient for Speaker Clusters (SCSC) of embeddings across all accents drops from 0.176 to 0.090 ($p\text{-value} = 6.09 \times 10^{-6}$), shown in #X4 vs #X5 in Table 2.8, demonstrating statistically significant more overlap between speaker clusters

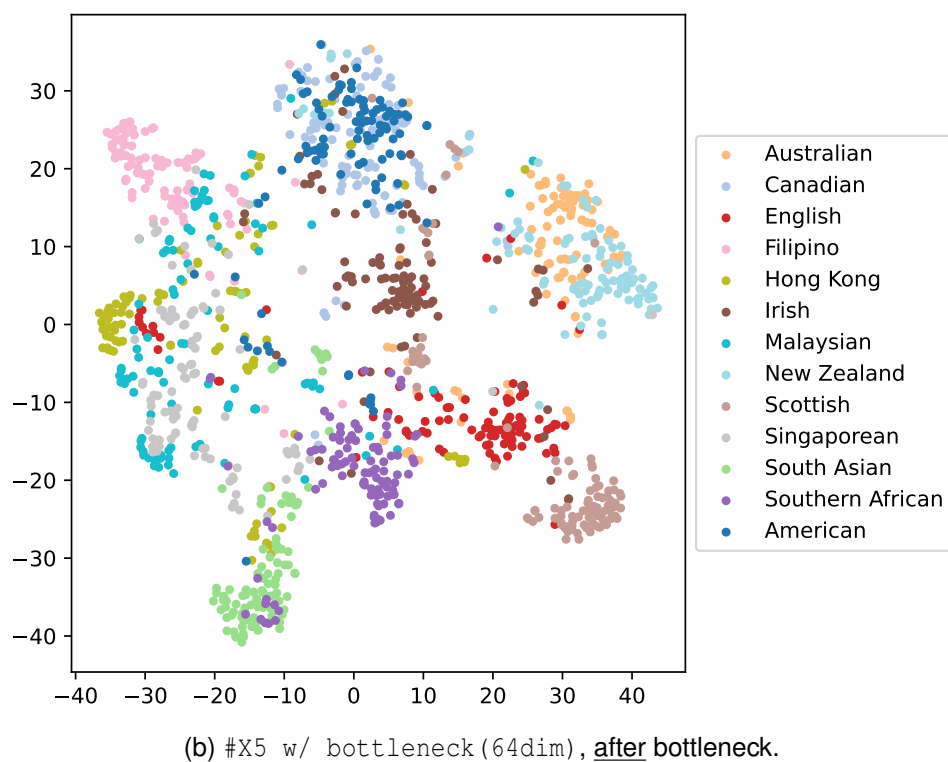
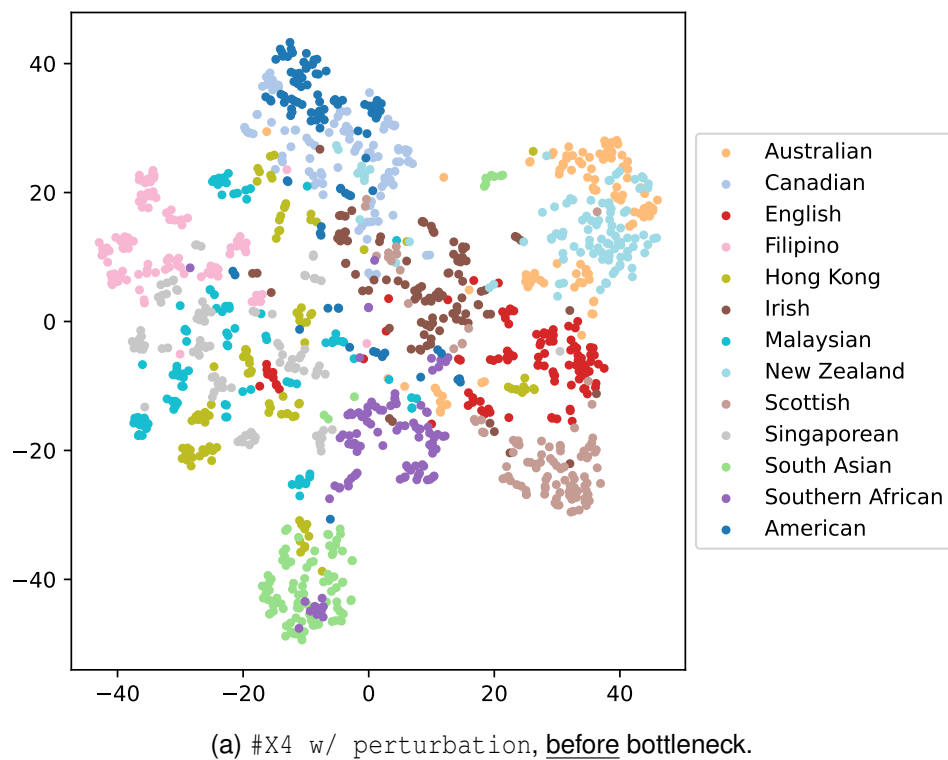


Figure 2.7: T-SNE visualisation of embeddings by XLSR-based systems on seen speakers, before and after applying information bottleneck, showing effects on separability of accent clusters.

Systems	USA	ENG	CAN	AUS	IRL	SCO	NZL
	SAS	SAF	HKG	PHL	MYS	SGP	mean ↓
#E4	0.149	0.171	0.154	0.185	0.130	0.155	0.147
	0.124	0.152	0.127	0.135	0.124	0.170	0.146
#E5	0.118	0.124	0.122	0.153	0.094	0.112	0.139
	0.098	0.116	0.085	0.104	0.106	0.144	0.114
#X4	0.170	0.160	0.175	0.148	0.125	0.152	0.097
	0.178	0.182	0.245	0.205	0.226	0.221	0.176
#X5	0.018	0.083	0.030	0.066	0.012	0.029	0.080
	0.101	0.158	0.164	0.115	0.146	0.167	0.090
#X6	0.095	0.059	0.054	0.052	-0.009	-0.018	0.064
	0.078	0.136	0.160	0.072	0.135	0.143	0.079

Table 2.8: Silhouette Coefficients for Speaker Clusters (SCSC) for each accent by different AID systems.

Effects of information bottleneck see #X4 vs #X5;

effects of adversarial training see #E4 vs #E5 and #E5 vs #E6.

USA - American, ENG - English, CAN - Canadian, AUS - Australian, IRL - Irish, SCO - Scottish, NZL - New Zealand, SAS - South Asian, SAF - South African, HKG - Hong Kong, PHL - Filipino, MYS - Malaysian, SGP - Singaporean.

in the learned accent space. Both pieces of evidence proves that there are more accent-related and less speaker-related information in the learned accent space.

2.6.6 Better Speaker Disentanglement by Adversarial Training

Adversarial training has minimal impact on accent classification results (<0.01 F1 change on unseen speakers for both #E4 vs #E5 and #X5 vs #X6). However, the SCSC results, which quantifies the residual amount of speaker information, suggest a small positive effect on speaker disentanglement, shown in Table 2.8. For ECAPA-TDNN-based systems, SCSC decreased from 0.146 to 0.114 (p-value = 1.53×10^{-7}), while for XLSR-based systems, SCSC decreased from 0.090 to 0.079 (p-value = 0.22). The limited effect with weak statistical significance on the XLSR-based system may be due to most speaker-related information already being filtered by the information bottleneck. The final 0.079 SCSC suggests heavy overlap between speaker clusters in the

learned accent space.

The limited effect could also be caused by the adversarial training scheme, which guides an auxiliary classifier to output an even distribution over all speakers $p(y_{spk}|x; \theta) \rightarrow \mathcal{U}(|y_{spk}|)$. Without a clear supervised signal, the model learns an even distribution over 10,000 undefined labels (which may not correspond to speakers). Such adversarial training is still helpful since the model is guided not to disambiguate information at such high granularity, but certainly suboptimal. We retain this adversarial training in our best systems, and leave more advanced schemes/designs for future work.

2.6.7 Implications for Self-Supervised Learning (SSL) Models

Overall, XLSR-based systems outperform ECAPA-TDNN-based systems, as shown in Table 2.7. Additionally, XLSR-based systems, aided by speech SSL pretraining, are more robust to class imbalance (analysed in Section 2.4.3) and noise/speech variability (analysed in Section 2.4.4). XLSR-based systems also possess richer information and would benefit from an information bottleneck (analysed in Section 2.4.5). This points to future research into what information these models hold and how robust they are across various speech factors like accents, speakers, channels, environments, styles, and rates.

2.6.8 Implications for Accent Similarity

T-SNE visualisation of the best system (Figure 2.4b) reveals that while most accent clusters are well-separated, some accent pairs overlap, reflecting regional proximity and accent similarity. Overlapping accents pairs include: American & Canadian, Australian & New Zealand, Singaporean & Malaysian (possibly also Hong Kong). These findings highlight the issues with self-reported, discrete accent labels, suggesting some accents may need further separation (e.g. different American accents - General American, New York, and South Carolina as in UNISYN lexicon¹³ (Fitt, 2000)) while others can be combined (e.g. arguably, Australian & New Zealand). AID models can be used to objectively measure accent similarity in speech signals by assessing cluster separation in the learned accent space, which can guide linguistic research on accents.

¹³<https://www.cstr.ed.ac.uk/projects/unisyn/>

2.7 Conclusions & Future Work

This chapter systematically addresses the challenges in AID, including dataset issues, benchmark shortcomings, and the effects of proposed modifications. Our key contributions are:

- To the best of our knowledge, we are the first to verify and quantify two critical issues in AID: 1) intrinsic speaker-accent entanglement, and 2) bias towards more common accents.
- We propose GenAID, with five effective modifications, reaching a new SOTA of 0.56 f1 score in 13-accent classification on unseen speakers.
- We propose new speaker-accent disentanglement methods, using information bottleneck and MSE-based adversarial training, and quantify the effects using proposed SCSC metric. We also raise concerns about the effectiveness of current adversarial training.

To revisit our research question: *How can we extract accent embeddings that are more discriminative of accents and less influenced by other speech factors?* Our results and analysis have addressed this question. GenAID achieves SOTA performance, generalises well across speakers, and provides better accent embeddings (which is crucial for the second stage accent generation task). In the future, we wish to focus on following four areas:

- **1) Data Size and Accent Coverage** We aim at constructing a larger dataset with broader accent coverage, incorporating other multi-accent and L2 learner speech corpora.
- **2) Quantifying other Residual/Entangled Information** Beyond speaker-accent entanglement, we would like to investigate and examine other biases and entanglement in speech, such as gender biases, and content-accent entanglement.
- **3) Explainable Accent Space** The current accent space is well-separated across most accents, but not explainable. We wish to explore how accent intensity and mixed accents are encoded in the learned accent space.
- **4) More Effective Adversarial Training** We call for research into more effective adversarial training scheme which can better remove information from

learned embeddings. We will also incorporate the disentanglement of other speech factors such as content, gender, age, etc. for AID and potentially speech forensics.

Chapter 3

AccentBox: High-Fidelity Zero-Shot Accent Generation

3.1 Overview

This chapter introduces AccentBox, a framework for high-fidelity zero-shot accent generation, enabling speech content creation in any voice and accent from a single audio clip. Section 3.2 lists out the data used for pretraining, finetuning, and inference in various systems. Section 3.3 examines a SOTA ZS-TTS system, identifies the *accent mismatch/hallucination* problem, and formalises the research question. Motivated by the identified problem and research question, we propose a framework for accent generation and control in ZS-TTS, with specific methods, experimental design, and results in Sections 3.4, 3.5, and 3.6 respectively. Final conclusions and future work are presented in Section 3.7. Readers are highly encouraged to visit our demo page¹ where we include audio samples for accent mismatch/hallucination in current SOTA ZS-TTS (part I) and comparison between different systems and the proposed AccentBox (part IV).

¹<https://jzmzhong.github.io/AccentBox-High-Fidelity-Zero-Shot-Accent-Generation>

3.2 Data

3.2.1 Pretraining: LibriTTS-R

Derived from LibriTTS (Zen et al., 2019) using speech restoration, LibriTTS-R² (Koizumi et al., 2023) is the largest available high-quality English TTS corpus. Due to its broad and diverse coverage of speakers, we adopt the clean portion of this corpus to pretrain a ZS-TTS model. The data composition is shown in Table 3.1. Regrettably, since LibriSpeech (Panayotov et al., 2015), the basis of LibriTTS, is collected with the requirement of having accents closer to US English, most of the utterances in the dataset is North American accents. Such biased accent coverage is also verified by running GenAID #X6 system from the previous chapter on all utterances, where 75.21% of the utterances are predicted as having either “American” or “Canadian” accents.

Subset	#Utterances	#Speakers	Duration (hrs)	USA/CAN (%)*
train-clean-100	33,232	247	53.55	76.91
train-clean-360	116,462	904	190.43	74.73
TOTAL	149,694	1,151	243.98	75.21

Table 3.1: Data composition of LibriTTS-R clean portion for pretraining in ZS-TTS.

*: Percentage of utterances predicted as “American” or “Canadian” by GenAID #X6.

3.2.2 Finetuning & Inference: VCTK

As discussed in previous Section 2.2, VCTK (Yamagishi et al., 2012) and L2-ARCTIC (Zhao et al., 2018) are the two ideal corpora for TTS experiments on accent generation. Due to time and resource constraints of MSc Dissertation, we are only able to experiment on L1 accents by finetuning and inferencing on the VCTK corpus (detailed data composition shown in Table 3.2). One speaker from each accent is reserved for inference only, while the remaining speakers are used for training and validation. Note that 3 accents are not seen by GenAID, while 6 accents are hardly seen by AccentBox (having 0 to 5 speakers).

²<https://www.openslr.org/141>

Accent	GenAID	AccentBox		
	Seen?	Test Speaker	#Speakers in Train & Valid	Duration(hrs) in Train & Valid
English	Yes	p225	32	11.98
American	Yes	p294	21	8.03
Scottish	Yes	p234	18	6.70
Irish	Yes	p245	8	3.03
Canadian	Yes	p302	7	2.77
Northern Irish	<i>No</i>	p261	5	2.07
South African	Yes	p347	3	1.18
Indian	<i>No</i>	p248	2	0.69
Australian	Yes	p326	1	0.37
New Zealand	Yes	p335	0	0.00
Welsh	<i>No</i>	p253	0	0.00

Table 3.2: Data composition of VCTK for finetuning & inference in ZS-TTS.

Italic font indicates missing/scarce data.

3.2.3 Stimuli for Listening Tests: *Comma Gets a Cure*

To test the performances of different systems in terms of accent generation, we use an elicitation passage not covered in training for listening tests, known as *Comma Gets a Cure*³ (Honorof et al., 2000). This elicitation passage uses the standard lexical set words by Wells (1982), enabling examination of English pronunciation in different accents/dialects across various phonemic contexts. We split the whole passage into 23 sentences, shown in Appendix B.1.

3.2.4 Limitations

There are two major limitations with the data we use.

1) Limited and Imbalanced Coverage of Accents The accent coverage issue reoccurs in the TTS stage. During pretraining, most of the data are North American accents. During finetuning, most of the data are English, American, and Scottish accents. Other

³<https://www.dialectsarchive.com/CommaGetsACure.pdf>

than such imbalanced coverage, the number of accents covered is also limited.

2) L1 Accents Only Unfortunately, this is largely due to the time and resource constraints. We leave the training and evaluation on L2 accents for future work.

3.3 Problem Identification

3.3.1 Accent Mismatch/Hallucination in ZS-TTS

Due to the closed-source nature of most SOTA ZS-TTS systems, we choose the open-source implementation of VALL-E X⁴ (Wang et al., 2023a; Zhang et al., 2023c), to examine the accent-related issues in ZS-TTS systems.

We input a fixed reference text-speech pair to VALL-E X with different target texts. The reference text-speech pairs are the 24th utterance from different test speakers of different accents in VCTK, introduced in Section 3.2.2. The target texts are the first five stimulus in the listening tests, introduced in Section 3.2.3. We also provide the accent prediction results by GenAID. Note that GenAID is not trained on any synthesised speech and therefore the results may be inaccurate - but it does demonstrate the inconsistency of accents in the generated speech. Inference results are available at the “I. Problem Identification” part on our demo page⁵.

Since VALL-E X is trained without conditioning on any accent-specific information other than a general speaker conditioning through the reference text-speech pair, it demonstrates poor accent consistency in generation - causing severe *accent mismatch/hallucination* between the reference speech and generated speech. While we are looking at accent-related issues, we also realise that there is the *unstable generation* issue where the generated speech have skipped/repeated pronunciation (utterance 1 & 2, group I; utterance 4, group II), weird prosody (utterance 2 & 3, group II), etc.

3.3.2 Research Question

The *accent mismatch/hallucination* problem identified and verified in the previous section highlights the need for generating more controllable and higher fidelity accents in ZS-TTS. In this second stage, we formalise the research question as: *To what degree can pretrained accent embeddings help TTS control and disentangle accent informa-*

⁴<https://github.com/Plachtaa/VALL-E-X>

⁵<https://jzmzhong.github.io/AccentBox-High-Fidelity-Zero-Shot-Accent-Generation>

tion in speech generation? We seek to explore how the GenAID, built in the first stage, can facilitate accent generation and control.

3.4 Methods

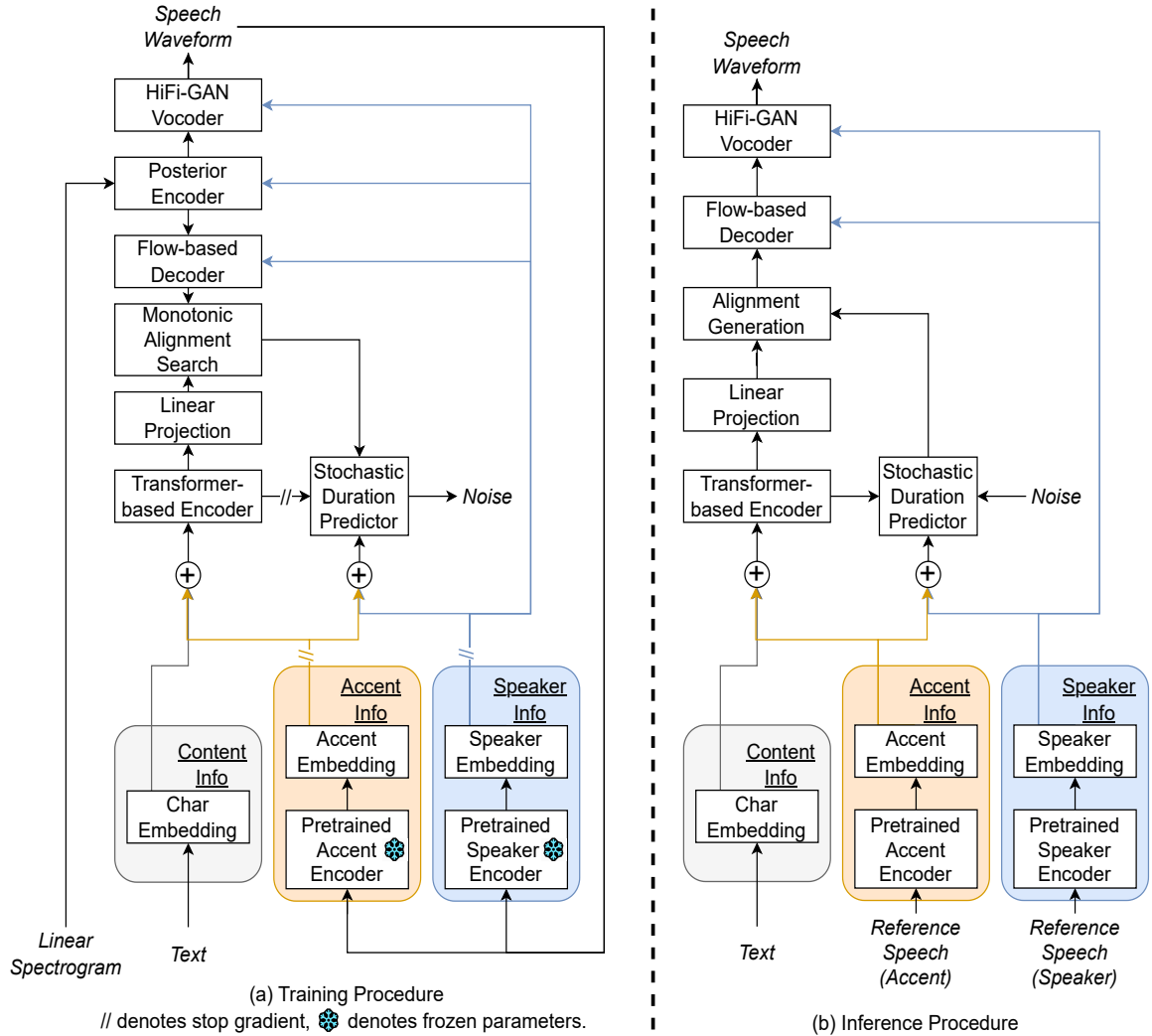


Figure 3.1: Model architecture of proposed ZS-TTS system (AccentBox).

The pretrained accent encoder (GenAID) is the same as in Figure 2.3b.

3.4.1 Training: Conditioning on GenAID Embedding

Figure 3.1 shows the model architecture for both training and inference. We build upon YourTTS (Casanova et al., 2022) instead of LLM-based ZS-TTS due to: 1) high data and computation requirements, 2) unstable generation (as verified in VALL-E X), and

3) lack of open-source models/code. Since the same text spoken by speakers of different accents exhibits distinct phonetic and prosodic variations, we condition both the Transformer-based Text Encoder and the Stochastic Duration Predictor on the accent embeddings learned by GenAID (system #X6 in previous Chapter 2). Compared with YourTTS, we replace the one-hot language embeddings in input with GenAID accent embeddings, as depicted by the pretrained accent encoder (orange block) in Figure 3.1.

Specifically, the 64-dimensional pretrained accent embedding for each utterance is linearly mapped to match the 192-dimensional character embeddings. This mapping allows the accent embedding to be added directly with the character embeddings, thereby integrating both content and accent information. The resulting combined embeddings are then passed into the Transformer-based Text Encoder and the Stochastic Duration Predictor.

To maximize the benefits of the extensive speaker coverage in the LibriTTS-R dataset and to expedite the training process across different systems, we employ a transfer learning approach. Initially, the model is pretrained on the clean portion of LibriTTS-R and subsequently finetuned on the VCTK corpus. Since LibriTTS-R lacks explicit accent labels, the accent encoder is omitted during the pretraining phase and incorporated only during the finetuning stage.

3.4.2 Inference: Inherent/Cross/Unseen Accent Generation

Table 3.3 outlines the different types of inference scenarios explored in this study, with further details provided below. It is important to note that all reference speech, both for target speaker and accent information, are from speakers not present in model’s training data, adhering to the zero-shot requirement.

Accent Generation	Target Speaker	Target Accent	Speaker-Accent Match?
Inherent	Unseen	Seen	Yes
Cross	Unseen	Seen	No
Unseen	Unseen	Unseen	Yes

Table 3.3: Different types of accent generation in AccentBox.

Inherent Accent Generation To examine the hypothesised higher accent fidelity brought by AccentBox, we use the same audio clip as reference speech for both accent and

speaker information during inference. The target accent is the inherent accent of the target speaker.

Cross Accent Generation To examine the hypothesised accent control and disentanglement brought by AccentBox, we use separate audio clips for accent and speaker information during inference. The reference speech for accent is taken from a different speaker of a different accent to the target speaker. Essentially, we are performing accent conversion by generating the target speaker’s voice in a different accent.

Unseen Accent Generation Similar to inherent accent generation, we use the same audio clip as reference speech for both accent and speaker information during inference. To explore the limits of zero-shot accent generation, we use audio clip with accents that are unseen by AccentBox. This scenario tests: 1) how generalisable the learned accent space of GenAID is, and 2) whether AccentBox can generalise synthesis to unseen accents.

3.5 Experiments

3.5.1 Systems

System	Data	Accent Info	Initialisation
VALL-E X	Unknown	N/A	inference only
Pretrained	LibriTTS-R clean	N/A	from scratch
Baseline	VCTK	N/A	from Pretrained
Accent_ID	VCTK	one-hot embedding	from Pretrained
Proposed	VCTK	GenAID embedding	from Pretrained

Table 3.4: Comparison of the training process in different ZS-TTS systems.

Table 3.4 outlines how different systems are obtained. VALL-E X is the one we have investigated in Section 3.3.1. The Pretrained system is trained on the clean portion of LibriTTS-R for 1 million steps and used for initialising the remaining three systems. The Baseline system directly finetunes on VCTK; the Accent_ID system finetunes with 4-dimensional one-hot embeddings of provided discrete accent labels in VCTK, same as the language embeddings in YourTTS; the Proposed system finetunes with

continuous 64-dimensional GenAID embeddings. All finetuned systems are trained for 200 thousand steps.

3.5.2 Configurations

1) Audio Processing To ensure high audio quality in synthesis, all waveforms are downsampled to 24 kHz as target waveform (rather than 16 kHz in original YourTTS). To be consistent with pretrained models, input waveforms to the speaker and accent encoders are still downsampled to 16 kHz.

2) Training Configurations We train all models with a batch size of 32, an initial learning rate of 0.0002, an exponentially decaying learning rate scheduler with gamma 0.999875, and the AdamW optimizer. All models are trained on a single NVIDIA Tesla V100S-PCIE-32GB GPU.

3.5.3 Objective Evaluation

1) Accent Cosine Similarity (Acc_COS) We use two AID models #X4 and #X6 from Chapter 2 to extract accent embeddings, and calculate cosine distances between reference and generated speech, avoiding biases towards Proposed which is conditioned on embeddings from AID model #X6.

2) Speaker Cosine Similarity (Spk_COS) We use Resemblyzer⁶ (Wan et al., 2018) to extract speaker embeddings of generated speech and compare them to reference speech (speaker) for cosine distance calculation.

3) Why no Word Error Rate (WER)? As verified by Sanabria et al. (2023), various SOTA ASR models have clear bias against accents and WER varies across different accents in EDACC. Despite wide usage of WER by an ASR model for evaluating ZS-TTS systems, we choose not to evaluate our systems in such way, as a high WER could indicate either unclear or more accented generation which makes ASR models harder to recognise correctly.

4) Accents for Objective Evaluation We use all 9 accents which are seen during finetuning to compare different systems. New Zealand and Welsh accents are unseen during training and therefore not able to use for evaluating Accent_ID which takes one-hot accent embedding as input condition.

⁶<https://github.com/resemble-ai/Resemblyzer>

3.5.4 Subjective Evaluation

1) Accent Similarity, Speaker Similarity, and Naturalness To holistically evaluate different aspects of generated speech, we ask listeners to compare different systems based on three metrics: i) *accent similarity* - how similar the generated speech is similar to the reference speech in terms of accent identity, ii) *speaker similarity* - how similar the generated speech is similar to the reference speech in terms of speaker identity, and iii) *naturalness* - how the generated speech sounds like human. The listening test interfaces are shown in Appendix B.3.

2) ABC Ranking and AB Preference To fully compare all systems, we conduct ABC ranking tests (Baseline vs Accent_ID vs Proposed) for inherent accent generation and AB preference tests (Accent_ID vs Proposed) for cross accent generation. The Baseline does not take any accent information as input condition and does not possess cross accent generation ability, therefore not evaluated in the later task.

3) Recruiting Listeners All listeners are recruited through Prolific⁷ with no known hearing difficulties and English as native and primary language. For different accents, we require respective listeners to be born in, spend most time in before 18, and is currently located in the respective accent region (e.g. the United States, Ireland, etc.). 10 listeners are required for each utterance.

4) Accents for Subjective Evaluation Due to budget constraints, we are only able to conduct listening tests on two accents. We choose American and Irish accents, with different data size (8.03 and 3.03 hours respectively) in the finetuning data.

5) Statistical Testing When interpreting the subjective preferences between two systems, we set the null hypothesis to be adding accent one-hot or GenID embedding does not bring improvement (i.e. $\text{Baseline} \geq \text{Accent_ID}$, and $\text{Baseline} \geq \text{Proposed}$), and calculate the p-values which represent the chance the null hypothesis stands.

3.6 Results & Analysis

3.6.1 Overview

Table 3.5 shows the objective evaluation results of 5 systems on 9 accents that are included in the VCTK training & validation datasets, i.e. seen by the three finetuned

⁷<https://www.prolific.com>

System	Inherent Accent Generation			Cross Accent Generation		
	Acc_COS (#X4)	Acc_COS (#X6)	Spk_COS	Acc_COS (#X4)	Acc_COS (#X6)	Spk_COS
VALL-E X	0.7801	0.9077	0.8605	/	/	/
Pretrained	0.7510	0.8911	0.8413	/	/	/
Baseline	0.7232	0.8989	0.8362	/	/	/
Accent_ID	0.7837	0.9291	0.8386	0.7350	0.8985	0.8073
Proposed	0.8037	0.9336	0.8293	0.7538	0.9067	0.8100

Table 3.5: Objective evaluation results on 9 seen accents.

Acc_COS - Accent Cosine Similarity, Spk_COS - Speaker Cosine Similarity.

#X4 and #X6 are two AID systems in Chapter 2. **Bold** font indicates best results.

/: These three systems cannot conduct cross accent generation.

Comparison	Accent	Accent Similarity		Speaker Similarity		Naturalness	
		Pref. (%)	p-value	Pref. (%)	p-value	Pref. (%)	p-value
vs Baseline	US	69.1%	1.82E-04	70.0%	1.18E-03	60.0%	1.07E-02
	Irish	61.3%	1.40E-02	57.8%	9.40E-02*	33.9%	2.75E-03
vs Accent_ID	US	57.4%	8.39E-02*	62.2%	2.05E-02	56.1%	3.38E-02
	Irish	65.7%	4.91E-06	59.1%	9.30E-03	43.9%	2.56E-02

Table 3.6: Subjective evaluation results for inherent accent generation.

“Pref.” - preference rate for Proposed. *: weak statistical significance.

Comparison	Accent	Accent Similarity		Speaker Similarity		Naturalness	
		Pref. (%)	p-value	Pref. (%)	p-value	Pref. (%)	p-value
vs Accent_ID	US	70.0%	1.09E-06	45.2%	3.19E-02	65.2%	1.48E-04
	Irish	61.7%	1.33E-02	61.3%	1.14E-02	63.0%	3.10E-02

Table 3.7: Subjective evaluation results for cross accent generation.

“Pref.” - preference rate for Proposed.

Not compared against Baseline due to its inability of cross accent generation.

systems. Table 3.6 shows the subjective evaluation results for inherent accent generation by comparing the preferences among the three finetuned systems. Table 3.7 shows the subjective evaluation results for cross accent generation by comparing the preferences between the only two systems which can perform accent conversion.

For the task of unseen accent generation which is significantly more difficult, requiring TTS models to generalise to unseen accents, we include generated audios in the demo page with comparison between `Baseline` and `Proposed`. We leave more systematic evaluation of such task for future work. Note that all audios samples in the demo page are not cherry-picked - we use all stimulus for objective/subjective evaluation and arbitrarily put the first five stimulus for demonstration.

3.6.2 Inherent Accent Generation

Accent Similarity The `Proposed` system achieves higher accent similarity across both objective and subjective evaluations. In objective evaluations, regardless of which model is used to extract accent embeddings, the `Proposed` system outperforms the other systems, including the open-source `VALL-E X`, which is trained on a larger dataset with more model parameters. In subjective evaluations, the `Proposed` system consistently outperforms both the `Baseline` and `Accent_ID` systems in generating both American and Irish accents. These results collectively demonstrate that the `Proposed` system has higher accent fidelity in the task of inherent accent generation.

Speaker Similarity The `Proposed` system shows higher speaker similarity in subjective evaluations, contrasting with the lower speaker cosine similarity scores in objective evaluations. This discrepancy may arise from two reasons: 1) The speaker embeddings might be biased towards more common accents due to the training data used in the speaker verification model. 2) Listeners may not fully separate accent and speaker identities in their perception, relating higher accent similarity with higher speaker similarity. Further research is needed to develop more effective methods for evaluating speaker similarity when the generated speech include varying accents.

Naturalness The `Proposed` system demonstrates higher naturalness when generating the American accent compared to the other two systems, but it shows lower preference rates for the Irish accent. We hypothesize that this discrepancy may be due to two reasons. 1) *Difference in data size*: With only 3.03 hours of Irish accent data in the training set, the more granular accent conditioning provided by the `GenAID` embeddings likely requires a larger amount of data to accurately model diverse accents.

These continuous embeddings capture not only country-level accent labels but also more fine-grained utterance-specific accent variations. Further research is needed to explore this hypothesis, especially with more extensive accented data available during finetuning. 2) *Monotonic prosody in the reference speech*: As ZS-TTS systems are highly sensitive to reference speech with frequently unstable speech generated, it could be that the Proposed system picks up more of the monotonic prosodic pattern in the Irish reference speech - leading to overall worse naturalness.

3.6.3 Cross Accent Generation

Lower Objective Similarity The overall objective results for cross accent generation is lower than those of inherent accent generation, as shown in Table 3.5. This demonstrates that accent conversion is a more difficult task than the previous inherent accent generation.

Accent/Speaker Similarity The Proposed system shows higher accent similarity in both objective and subjective evaluations, demonstrating higher accent fidelity in accent conversion. The subjective speaker similarity results are controversial, with higher subjective similarity preference on Irish but not on American accent. We hypothesise that this could be again the listeners' perception problem - regarding the generated speech with higher accent similarity to be more distant in terms of speaker identity from the original reference speech (speaker) which is in English accent.

Naturalness The Proposed system demonstrates higher naturalness on both accents during accent conversion. We hypothesise that this could be due to the more consistent accent being generated in accent conversion. The Accent_ID system learns the accent embeddings by one-hot labels on limited TTS data, inferior to the pretrained accent embeddings in the Proposed system. Swapping one-hot accent embedding from one accent to another forces the model to generalise to unseen speaker-accent pairs using limited information from the accent embedding, resulting in inconsistent accent and unnaturalness in the utterance.

3.6.4 Problems of ZS-TTS

A lot of the above problems that zero-shot accent generation suffers are also common to ZS-TTS. There is simply too little information in one audio clip. Despite the success of ZS-TTS in engineering and industrial applications, from a speech science perspec-

tive, where the prosody comes from in the generated speech remain highly unclear - the reference speech (speaker & accent), the input text, and the stochastic noise for generative modelling modules, can all determine the generated prosody. More research is needed in explaining, controlling, and disentangling different factors in the generated speech in ZS-TTS systems. This study initiates the first step towards accent control and disentanglement in ZS-TTS, with AccentBox still suffering from numerous other entangled factors such as prosody.

3.7 Conclusions

In this chapter, we take an initial approach to leverage pretrained accent embeddings for zero-shot accent generation, achieving higher accent fidelity while maintaining speaker similarity and naturalness. Our key contributions are:

- To the best of our knowledge, we are among the first to highlight the problem of *accent mismatch/hallucination* in ZS-TTS.
- We propose a zero-shot accent generation framework and establish the first benchmark for inherent/cross/unseen accent generation, enabling the generation of any text, speaker, and accent.
- We develop AccentBox, which uses GenAID embeddings for high-fidelity zero-shot accent generation, showing superior accent similarity in both inherent and cross-accent scenarios.

To revisit our research question: *To what degree can pretrained accent embeddings help TTS control and disentangle accent information in speech generation?* Our results indicate a decent level of coarse accent control, though challenges like unstable prosody and inconsistent accent remain. Future work will focus on:

- **1) Foreign Accent Generation (FAC)** We will adapt our framework for FAC, particularly in CAPT, and comparing AccentBox's performance in L2 accent generation with existing approaches.
- **2) Accent Modelling** We seek to incorporate multi-level accent modelling (Zhou et al., 2024b) to model different levels of accent variations in speech.

- **3) Phoneme vs Character Input** We will investigate the use of a base lexicon for improved pronunciation modeling and stability in ZS-TTS, while free from reliance on an accent-specific front-end.

Chapter 4

General Discussion

In this chapter, we reflect on three challenging questions arising from the results and analyses of GenAID and AccentBox.

4.1 How should we define accents in data collection?

In both AccentBox and GenAID, accent labels based on region or country are often either too broad, like the varied American accents, or too narrow, as seen with the similar New Zealand and Australian accents. Accents are not just continuous but also multidimensional. L2 accents, in particular, don't fit on a simple scale of "accentedness" but instead occupy a complex, multidimensional space. Defining accents is inherently difficult, and this challenge complicates both accent discrimination and generation modeling. Our current answer to the question: *We don't know.*

4.2 Is full speaker-accent disentanglement desirable and achievable?

In AccentBox, evaluating speaker similarity in cross-accent generation is challenging, especially when the reference and generated speech differ in accent. From a human perception standpoint, accent is crucial for distinguishing speakers, suggesting that complete speaker-accent disentanglement may not be desirable. Similarly, in GenAID, even with minimal residual speaker information, a 0.23 f1 gap between seen and unseen speakers remains. The model may still memorize speaker-accent mappings despite attempts to disentangle them. Therefore, we question whether full disentanglement

ment is even achievable. Our tentative answer to the question: *Maybe not.*

4.3 Are utterance-level AID and ZS-TTS ill-defined tasks?

Despite broad applications for AID/ZS-TTS based on a single short utterance, our experiments with GenAID and AccentBox suggest that these might be ill-defined tasks. One utterance provides too little information, making it difficult even for experts to discern certain accents. Worse, models are forced to predict an accent without reliable confidence levels. The challenge is to train AID models that can output “unknown” when there is too little information, or ZS-TTS models that default to a neutral or standard accent when the reference speech provides insufficient accent information. We leave these issues for future work.

Chapter 5

Conclusions & Future Work

This thesis introduces zero-shot accent generation and a novel two-stage pipeline as a benchmark. In the first stage AID, we verify, quantify, and address the problems of speaker-accent entanglement and biased prediction across accents, with SOTA performance of 0.56 f1 score in 13-accent classification on unseen speakers. In the second stage zero-shot accent generation, we highlight and address the problem of accent mismatch/hallucination in ZS-TTS, with better accent fidelity in inherent/cross accent generation while enabling unseen accent generation.

In the future, we will prioritise L2 accent generation and expanding zero-shot accent generation to cover more accents. Additionally, we aim to enhance accent disentanglement and control using a factorised neural codec, as proposed in the latest SOTA ZS-TTS, NaturalSpeech 3 (Ju et al., 2024).

Bibliography

- Agarwal, C. and Chakraborty, P. (2019). A Review of Tools and Techniques for Computer Aided Pronunciation Training (CAPT) in English. *Education and Information Technologies*, 24(6):3731–3743.
- Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., and Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.
- Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2022). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.
- Badlani, R., Arora, A., Ghosh, S., Valle, R., Shih, K. J., Santos, J. F., Ginsburg, B., and Catanzaro, B. (2023a). VANI: Very-lightweight Accent-controllable TTS for Native and Non-native speakers with Identity Preservation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE.
- Badlani, R., Valle, R., Shih, K. J., Santos, J. F., Gururani, S., and Catanzaro, B. (2023b). RAD-MMM: Multilingual Multiaccented Multispeaker Text To Speech. In *Proc. INTERSPEECH 2023*, pages 626–630.
- Black, A., Taylor, P., Caley, R., and Clark, R. (1998). The FESTIVAL Speech Synthesis System.

- Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E., and Ponti, M. A. (2022). YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2709–2720. PMLR.
- Chen, S., Liu, S., Zhou, L., Liu, Y., Tan, X., Li, J., Zhao, S., Qian, Y., and Wei, F. (2024). VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers. *arXiv preprint arXiv:2406.05370*.
- Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. (2023). High Fidelity Neural Audio Compression. *Transactions on Machine Learning Research*. Featured Certification, Reproducibility Certification.
- Deja, K., Tinchev, G., Czarnowska, M., Cotescu, M., and Droppo, J. (2023). Diffusion-based Accent Modelling in Speech Synthesis. In *Proc. INTERSPEECH 2023*, pages 5516–5520.
- Desplanques, B., Thienpondt, J., and Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proc. Interspeech 2020*, pages 3830–3834.
- Ding, S., Zhao, G., and Gutierrez-Osuna, R. (2022). Accentron: Foreign Accent Conversion to Arbitrary Non-native Speakers Using Zero-shot Learning. *Computer Speech & Language*, 72:101302.
- Felps, D., Bortfeld, H., and Gutierrez-Osuna, R. (2009). Foreign Accent Conversion in Computer Assisted Pronunciation Training. *Speech communication*, 51(10):920–932.
- Fitt, S. (2000). Documentation and User Guide to UNISYN Lexicon and Post-Lexical Rules. Centre for Speech Technology Research, University of Edinburgh.
- 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.

- Gluszek, A. and Dovidio, J. F. (2010). The Way They Speak: A Social Psychological Perspective on the Stigma of Nonnative Accents in Communication. *Personality and social psychology review*, 14(2):214–237.
- Green, C. and Green, J. M. (1993). Secret Friend Journals. *TESOL journal*, 2(3):20–23.
- Hendrycks, D. and Gimpel, K. (2016). Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*.
- Hilton, H. (2009). Annotation and Analyses of Temporal Aspects of Spoken Fluency. *Calico Journal*, 26(3):644–661.
- Ho, Y. and Wookey, S. (2020). The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling. *IEEE Access*, 8:4806–4813.
- Honorof, D. N., McCullough, J., and Somerville, B. (2000). Comma Gets a Cure. *diagnostic passage*.
- Jia, D., Tian, Q., Peng, K., Li, J., Chen, Y., Ma, M., Wang, Y., and Wang, Y. (2023). Zero-Shot Accent Conversion using Pseudo Siamese Disentanglement Network. In *Proc. INTERSPEECH 2023*, pages 5476–5480.
- Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Chen, z., Nguyen, P., Pang, R., Lopez Moreno, I., and Wu, Y. (2018). Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, Y., Leng, Y., Song, K., Tang, S., et al. (2024). NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models. *arXiv preprint arXiv:2403.03100*.
- Kharitonov, E., Vincent, D., Borsos, Z., Marinier, R., Girgin, S., Pietquin, O., Sharifi, M., Tagliasacchi, M., and Zeghidour, N. (2023). Speak, Read and Prompt: High-fidelity Text-to-Speech with Minimal Supervision. *Transactions of the Association for Computational Linguistics*, 11:1703–1718.
- Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio Augmentation for Speech Recognition. In *Proc. Interspeech 2015*, pages 3586–3589.

- Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., and Khudanpur, S. (2017). A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224.
- Koizumi, Y., Zen, H., Karita, S., Ding, Y., Yatabe, K., Morioka, N., Bacchiani, M., Zhang, Y., Han, W., and Bapna, A. (2023). LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus. In *Proc. INTERSPEECH 2023*, pages 5496–5500.
- Kubanek-German, A. (2000). Early language programmes in germany. *An Early Start: Young Learners and Modern Languages in Europe and Beyond*, pages 59–70.
- Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., Williamson, M., Manohar, V., Adi, Y., Mahadeokar, J., and Hsu, W.-N. (2023). Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 14005–14034. Curran Associates, Inc.
- Li, R., Xie, Z., Xu, H., Peng, Y., Liu, H., Huang, H., and Chng, E. S. (2023). Self-supervised Learning Representation based Accent Recognition with Persistent Accent Memory. In *Proc. INTERSPEECH 2023*, pages 1968–1972.
- Ling, C. X. and Li, C. (1998). Data Mining for Direct marketing: Problems and Solutions. In *Kdd*, volume 98, pages 73–79.
- Liu, R., Sisman, B., Gao, G., and Li, H. (2024). Controllable Accented Text-to-Speech Synthesis With Fine and Coarse-Grained Intensity Rendering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2188–2201.
- Liu, R., Zuo, H., Hu, D., Gao, G., and Li, H. (2023). Explicit Intensity Control for Accented Text-to-Speech. In *Proc. INTERSPEECH 2023*, pages 22–26.
- Liu, S., Wang, D., Cao, Y., Sun, L., Wu, X., Kang, S., Wu, Z., Liu, X., Su, D., Yu, D., and Meng, H. (2020). End-To-End Accent Conversion without Using Native Utterances. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6289–6293.
- Lyth, D. and King, S. (2024). Natural Language Guidance of High-Fidelity Text-to-Speech with Synthetic Annotations. *arXiv preprint arXiv:2402.01912*.

- Ma, L., Zhang, Y., Zhu, X., Lei, Y., Ning, Z., Zhu, P., and Xie, L. (2024). Accent-VITS: Accent Transfer for End-to-End TTS. In Jia, J., Ling, Z., Chen, X., Li, Y., and Zhang, Z., editors, *Man-Machine Speech Communication*, pages 203–214, Singapore. Springer Nature Singapore.
- MacWhinney, B. (2017). A Shared Platform for Studying Second Language Acquisition. *Language Learning*, 67(S1):254–275.
- Melechovsky, J., Mehrish, A., Herremans, D., and Sisman, B. (2023). Learning Accent Representation with Multi-Level VAE Towards Controllable Speech Synthesis. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 928–935.
- Muñoz, C. (2006). *Age and the Rate of Foreign Language Learning*, volume 19. Multilingual Matters.
- Pal, D., Arpnikanondt, C., Funilkul, S., and Varadarajan, V. (2019). User Experience with Smart Voice Assistants: The Accent Perspective. In *2019 10th international conference on computing, communication and networking technologies (ICCCNT)*, pages 1–6. IEEE.
- Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). LibriSpeech: An ASR Corpus based on Public Domain Audio Books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.
- Qian, K., Zhang, Y., Chang, S., Hasegawa-Johnson, M., and Cox, D. (2020). Unsupervised Speech Decomposition via Triple Information Bottleneck. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7836–7846. PMLR.
- Qian, K., Zhang, Y., Chang, S., Yang, X., and Hasegawa-Johnson, M. (2019). AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5210–5219. PMLR.
- Quamer, W., Das, A., Levis, J., Chukharev-Hudilainen, E., and Gutierrez-Osuna, R. (2022). Zero-Shot Foreign Accent Conversion without a Native Reference. In *Proc. Interspeech 2022*, pages 4920–4924.

- Ravanelli, M., Parcollet, T., Moumen, A., de Langen, S., Subakan, C., Plantinga, P., Wang, Y., Mousavi, P., Libera, L. D., Ploujnikov, A., Paissan, F., Borra, D., Zaiem, S., Zhao, Z., Zhang, S., Karakasidis, G., Yeh, S.-L., Champion, P., Rouhe, A., Braun, R., Mai, F., Zuluaga-Gomez, J., Mousavi, S. M., Nautsch, A., Liu, X., Sagar, S., Duret, J., Mdhaffar, S., Laperriere, G., Rouvier, M., Mori, R. D., and Esteve, Y. (2024). Open-Source Conversational AI with SpeechBrain 1.0.
- Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., and Bengio, Y. (2021). SpeechBrain: A General-Purpose Speech Toolkit. arXiv:2106.04624.
- Rosina, L.-G. (1997). *English with an Accent: Language Ideology and Discrimination in the United States*. Routledge.
- Rousseuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of computational and applied mathematics*, 20:53–65.
- Sanabria, R., Bogoychev, N., Markl, N., Carmantini, A., Klejch, O., and Bell, P. (2023). The Edinburgh International Accents of English Corpus: Towards the Democratization of English ASR. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., and Wu, Y. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.
- Shen, K., Ju, Z., Tan, X., Liu, E., Leng, Y., He, L., Qin, T., sheng zhao, and Bian, J. (2024). NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers. In *The Twelfth International Conference on Learning Representations*.
- 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 *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6918–6922.
- Spiteri Miggiani, G. (2021). Exploring Applied Strategies for English-language Dubbing.
- Sun, S. and Richmond, K. (2024). Learning Pronunciation from Other Accents via Pronunciation Knowledge Transfer. In *Proc. Interspeech 2024*.
- Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., et al. (2024). NaturalSpeech: End-to-End Text-to-Speech Synthesis with Human-Level Quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.
- Wan, L., Wang, Q., Papir, A., and Moreno, I. L. (2018). Generalized End-to-End Loss for Speaker Verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883.
- Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., et al. (2023a). Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *arXiv preprint arXiv:2301.02111*.
- Wang, Y., Gu, H., Shen, R., Li, Y., Jiang, W., and Huang, J. (2023b). Disentanglement of Speaker Identity for Accented Speech Recognition. In *2023 8th International Conference on Communication, Image and Signal Processing (CCISP)*, pages 1–6.
- Webber, J. J., Perrotin, O., and King, S. (2020). Hider-Finder-Combiner: An Adversarial Architecture for General Speech Signal Modification. In *Proc. Interspeech 2020*, pages 3206–3210.
- Wells, J. C. (1982). *Accents of English: Volume 1*, volume 1. Cambridge University Press.
- Worthington, L. (1997). Let's Not Show the Teacher: EFL Students' Secret Exchange Journals. In *Forum*, volume 35, page n3. ERIC.
- Xiang, X., Wang, S., Huang, H., Qian, Y., and Yu, K. (2019). Margin Matters: Towards More Discriminative Deep Neural Network Embeddings for Speaker Recognition.

- In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1652–1656.
- Yamagishi, J., Veaux, C., and MacDonald, K. (2012). CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). <https://datashare.ed.ac.uk/handle/10283/3443>.
- Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. (2022). SoundStream: An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.
- Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., and Wu, Y. (2019). LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proc. Interspeech 2019*, pages 1526–1530.
- Zhang, M., Zhou, X., Wu, Z., and Li, H. (2023a). Towards Zero-Shot Multi-Speaker Multi-Accent Text-to-Speech Synthesis. *IEEE Signal Processing Letters*.
- Zhang, M., Zhou, Y., Wu, Z., and Li, H. (2023b). Zero-Shot Multi-Speaker Accent TTS with Limited Accent Data. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1931–1936.
- Zhang, Z., Zhou, L., Wang, C., Chen, S., Wu, Y., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., et al. (2023c). Speak Foreign Languages with Your Own Voice: Cross-Lingual Neural Codec Language Modeling. *arXiv preprint arXiv:2303.03926*.
- Zhao, G., Ding, S., and Gutierrez-Osuna, R. (2021). Converting Foreign Accent Speech without a Reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2367–2381.
- Zhao, G., Sonsaat, S., Silpachai, A., Lucic, I., Chukharev-Hudilainen, E., Levis, J., and Gutierrez-Osuna, R. (2018). L2-ARCTIC: A Non-native English Speech Corpus. In *Proc. Interspeech 2018*, pages 2783–2787.
- Zhou, X., Zhang, M., Zhou, Y., Wu, Z., and Li, H. (2024a). Accented Text-to-Speech Synthesis With Limited Data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1699–1711.

Zhou, X., Zhang, M., Zhou, Y., Wu, Z., and Li, H. (2024b). Multi-Scale Accent Modeling with Disentangling for Multi-Speaker Multi-Accent TTS Synthesis. *arXiv preprint arXiv:2406.10844*.

Zuluaga-Gomez, J., Ahmed, S., Visockas, D., and Subakan, C. (2023). CommonAccent: Exploring Large Acoustic Pretrained Models for Accent Classification Based on Common Voice. In *Proc. INTERSPEECH 2023*, pages 5291–5295.

Appendix A

GenAID: Cleaning Accent Labels

Original Label	Cleaned Label
United States English	American
England English	English
Canadian English	Canadian
Australian English	Australian
Irish English	Irish
Scottish English	Scottish
New Zealand English	New Zealand
India and South Asia (India, Pakistan, Sri Lanka)	South Asian
South African accent, Southern African (South Africa, Zimbabwe, Namibia)	Southern African
Hong Kong English	Hong Kong
Filipino	Filipino
Malaysian English	Malaysian
Singaporean English	Singaporean

Table A.1: Mapping from original labels to cleaned labels in data processing for GenAID, incl. 7 L1 accents (top) and 6 L2 accents (bottom).

Appendix B

AccentBox: Evaluation Materials

B.1 Stimuli for Listening Tests: *Comma Gets a Cure*

The content information in inference is provided by the stimuli below taken from *Comma Gets a Cure*¹ (Honorof et al., 2000).

1. Well, here's a story for you.
2. Sarah Perry was a veterinary nurse who had been working daily at an old zoo in a deserted district of the territory.
3. So, she was very happy to start a new job at a superb private practice in North Square near the Duke Street Tower.
4. That area was much nearer for her and more to her liking.
5. Even so, on her first morning, she felt stressed.
6. She ate a bowl of porridge, checked herself in the mirror and washed her face in a hurry.
7. Then she put on a plain yellow dress and a fleece jacket, picked up her kit and headed for work.
8. When she got there, there was a woman with a goose waiting for her.
9. The woman gave Sarah an official letter from the vet.
10. The letter implied that the animal could be suffering from a rare form of foot and mouth disease, which was surprising, because normally you would only expect to see it in a dog or a goat.
11. Sarah was sentimental, so this made her feel sorry for the beautiful bird.
12. Before long, that itchy goose began to strut around the office like a lunatic, which made an unsanitary mess.

¹<https://www.dialectsarchive.com/CommaGetsACure.pdf>

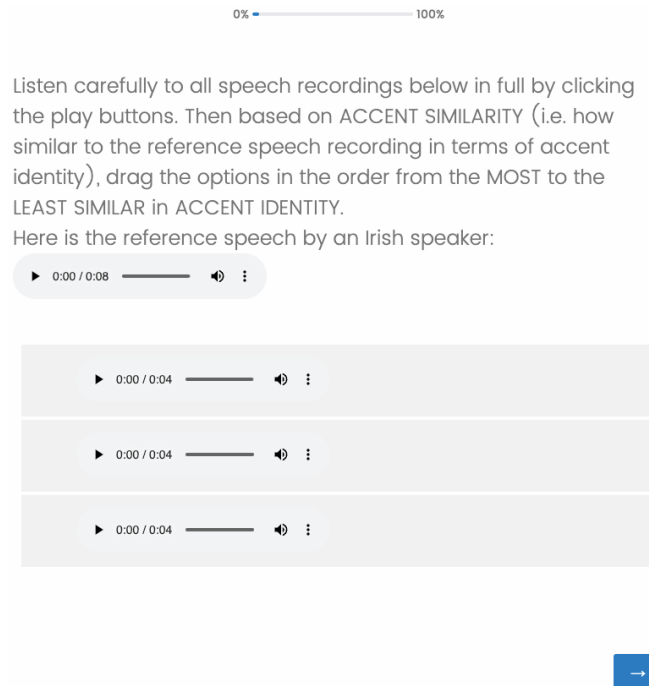
13. The goose's owner, Mary Harrison, kept calling, "Comma, Comma," which Sarah thought was an odd choice for a name.
14. Comma was strong and huge, so it would take some force to trap her, but Sarah had a different idea.
15. First she tried gently stroking the goose's lower back with her palm, then singing a tune to her.
16. Finally, she administered ether.
17. Her efforts were not futile.
18. In no time, the goose began to tire, so Sarah was able to hold on to Comma and give her a relaxing bath.
19. Once Sarah had managed to bathe the goose, she wiped her off with a cloth and laid her on her right side.
20. Then Sarah confirmed the vet's diagnosis.
21. Almost immediately, she remembered an effective treatment that required her to measure out a lot of medicine.
22. Sarah warned that this course of treatment might be expensive - either five or six times the cost of penicillin.
23. I can't imagine paying so much, but Mrs. Harrison - a millionaire lawyer - thought it was a fair price for a cure.

B.2 Reference Speech for Listening Tests

The speaker and accent information in inference is provided by the reference speeches. All reference speeches can be accessed and downloaded at: https://groups.inf.ed.ac.uk/cstr3/s2526235/listening_tests/AccentBox_v1.2/Speech_Prompts/. We take the 24th utterance for all eleven test speakers in VCTK (Yamagishi et al., 2012).

B.3 Questionnaires for Listening Tests

To evaluate accent similarity, we ask the participants to rank two or three audio samples based on their accent similarity to the reference speech (accent), shown in Figure B.1. Similarly, to evaluate speaker similarity, we ask the participants to rank two or three audio samples based on their speaker similarity to the reference speech (speaker), shown in Figure B.2. And finally, to evaluate naturalness, we ask the participants to rank two or three audio samples based on which one is more like a human would say it without reference speech, shown in Figure B.3.



0% ————— 100%

Listen carefully to all speech recordings below in full by clicking the play buttons. Then based on ACCENT SIMILARITY (i.e. how similar to the reference speech recording in terms of accent identity), drag the options in the order from the MOST to the LEAST SIMILAR in ACCENT IDENTITY.

Here is the reference speech by an Irish speaker:

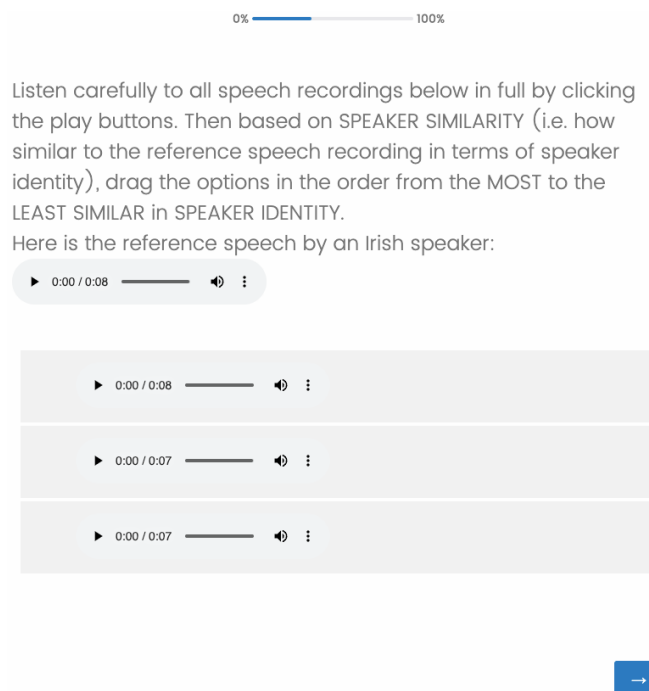
▶ 0:00 / 0:08 ————— 🔊 ⋮

▶ 0:00 / 0:04 ————— 🔊 ⋮

▶ 0:00 / 0:04 ————— 🔊 ⋮

▶ 0:00 / 0:04 ————— 🔊 ⋮

→

Figure B.1: Listening test interface for evaluating accent similarity.

0% ————— 100%

Listen carefully to all speech recordings below in full by clicking the play buttons. Then based on SPEAKER SIMILARITY (i.e. how similar to the reference speech recording in terms of speaker identity), drag the options in the order from the MOST to the LEAST SIMILAR in SPEAKER IDENTITY.

Here is the reference speech by an Irish speaker:

▶ 0:00 / 0:08 ————— 🔊 ⋮

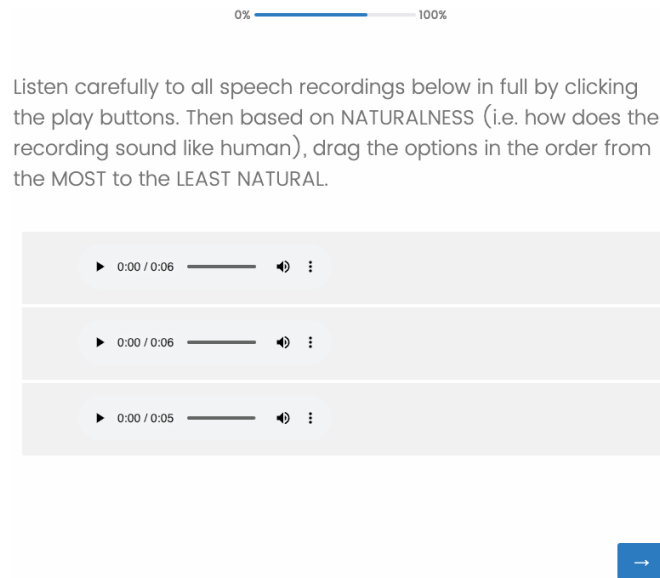
▶ 0:00 / 0:08 ————— 🔊 ⋮

▶ 0:00 / 0:07 ————— 🔊 ⋮

▶ 0:00 / 0:07 ————— 🔊 ⋮

→

Figure B.2: Listening test interface for evaluating speaker similarity.

Figure B.3: Listening test interface for evaluating naturalness.

All questionnaires are available at the following URLs, shown in Table B.1.

Generation	Accent	URL
Inherent	American	https://edinburgh.eu.qualtrics.com/jfe/form/SV_3UEtXMT2Q54EE5M
	Irish	https://edinburgh.eu.qualtrics.com/jfe/form/SV_72qDqacbRZijc6q
Cross	American	https://edinburgh.eu.qualtrics.com/jfe/form/SV_8ldW0Hscf6tDFK6
	Irish	https://edinburgh.eu.qualtrics.com/jfe/form/SV_51kCvIP9Zu89uvQ

Table B.1: Listening test URLs to various questionnaires.