TE KUNENGA KI PŪREHUROA | MASSEY UNIVERSITY
UNIVERSITY OF NEW ZEALAND

Deep Learning for Low-Resource Speech Recognition

A thesis presented in partial fulfilment of the
requirements for the degree of

Doctor of Philosophy
in
Computer Science

School of Mathematical and Computational Sciences,
Massey University, Albany, Auckland,
New Zealand

Zhihan Wang

May 2026

It is no nation we inhabit, but a language.
Make no mistake; our native tongue is our true fatherland.

–Emil Cioran, Romanian Philosopher

Abstract

Low-resource Automatic Speech Recognition (ASR) focuses on transcribing speech signals into text for low-resource languages, which suffer from significantly limited linguistic resources (e.g., corpora, dictionaries). Many of these low-resource languages are also endangered languages. Developing deep learning-based low-resource ASR systems presents significant challenges due to the scarcity of transcribed speech data and the lack of diverse and representative speakers. These limitations hinder the generalization to out-of-domain speech, making advancements in this area critical. In this thesis, we address these challenges through four research directions: data augmentation, continual learning, text-to-speech (TTS) synthesis, and synthesized data selection.

Firstly, we propose CyclicAugment, a novel approach designed to enhance generalization in low-resource ASR systems. By dynamically transforming speech data using a cosine annealing scheduler, CyclicAugment assists escape local optima during training, improving model robustness and performance.

Secondly, we introduce Randomly Layer-wise Tuning (CLRL-Tuning), an innovative continual learning method for pre-trained ASR models. CLRL-Tuning incrementally fine-tunes randomly selected layers of the model when trained on sequentially collected datasets. This approach effectively mitigates catastrophic forgetting and enhances generalization in low-resource ASR systems.

Thirdly, we develop Zero-Voice, a zero-shot TTS model comprising an Acoustic Feature Encoder, an Acoustic Feature Refiner, and a Waveform Vocoder. Trained on 27 hours of Te Reo Māori speech data, an official and endangered language of New Zealand. Zero-Voice synthesizes high-fidelity audio from reference speech samples. This enables the generation of additional Te Reo Māori speech data, significantly enhancing ASR performance for the language.

Lastly, to address the domain mismatch between true and synthesized Te Reo Māori data generated by the Zero-Voice model, we propose a score-distribution-matching data selection approach. By aligning the score distribution of synthesized data with the prior distribution of true data, this method effectively filters synthesized data, optimizing ASR performance for Te Reo Māori and advancing the state-of-the-art for low-resource ASR.

Through these contributions, our research addresses the key challenges in low-resource

ASR research. By integrating these innovations, we achieve state-of-the-art performance in Te Reo Māori ASR.

Acknowledgements

I am profoundly grateful to the many individuals whose support, encouragement, and expertise have been invaluable throughout my PhD study.

First and foremost, I extend my deepest appreciation to my main supervisor, Professor Ruili Wang. His unwavering guidance, insightful feedback, and constant encouragement have been instrumental in shaping my research. His expertise and constructive critique have refined my ideas and strengthened my work. I am especially grateful for the intellectual freedom he provided, allowing me to explore diverse research directions in a collaborative and enriching environment.

I also sincerely thank my co-supervisor, Dr. Feng Hou, as well as my colleagues Dr Satwinder Singh, Dr Zhizhong Ma, Dr Yuanghang Qiu, and Dr Yi Wang for their patience, expert feedback, and invaluable advice throughout our collaboration. Their support and insightful suggestions have been deeply appreciated.

I gratefully acknowledge the support of the Massey University Vice-Chancellor's Doctoral Scholarship, which has significantly contributed to my study and research.

Finally, I am grateful to the staff and faculty of the School of Mathematical and Computational Sciences at Massey University for their support and assistance throughout my PhD journey. It has been a privilege to pursue my doctoral studies at Massey University.

Publications

The following research papers have been published in or submitted to International Journals and Conferences during my PhD study:

1. **Zhihan Wang**, Feng Hou, Yuanhang Qiu, Zhizhong Ma, Satwinder Singh, Ruili Wang. (2022) CyclicAugment: Speech Data Random Augmentation with Cosine Annealing Scheduler for Automatic Speech Recognition. Proc. Interspeech 2022, pp. 3859-3863, doi:10.21437/Interspeech.2022-526, (**CORE rank A**).

2. **Zhihan Wang**, Feng Hou, Wang, Ruili Wang (2023) CLRL-Tuning: A Novel Continual Learning Approach for Automatic Speech Recognition. Proc. Interspeech 2023, pp. 1279-1283, doi:10.21437/Interspeech.2023-503, (**CORE rank A**).

3. **Zhihan Wang**, Feng Hou, Wang, Ruili Wang: Zero-Voice: A Low-Resource Approach For Zero-shot Text-to-Speech. (**Submitted to Computer Speech and Language Journal** (**CORE rank A**)).

4. **Zhihan Wang**, Feng Hou, Wang, Ruili Wang: Synthesized Data Selection via Score Distribution Matching for Te Reo Māori Automatic Speech Recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2026, pp. 18172-18176, doi:10.1109/ICASSP55912.2026.11461415. (**CORE rank B**)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides an overview of this thesis. We introduce the background and previous studies on low-resource speech recognition in Section 1.1. Then, the motivation of this research is presented in Section 1.2, where we discuss issues presented by existing approaches. The research objectives are presented in 1.3. Finally, the organization of this thesis is presented in Section 1.4.

1.1 Overview of Low-Resource Automatic Speech Recognition

Deep learning-based Automatic Speech Recognition (ASR) systems have achieved significant advancements for high-resource languages, such as Chinese [1], English [2], and Japanese [3], enabling successful real-world applications across various domains, including speech and language therapy [4], and language learning [5–7]. These systems have attained remarkable precision, driven by access to large-scale labeled speech corpora encompassing diverse speakers. However, replicating this success for low-resource languages remains a formidable challenge [8, 9].

Compared to high-resource languages, low-resource languages including indigenous languages (e.g., Te Reo Māori, Fijian, and Irish) and vernacular languages (e.g., Singlish and Malay), often lack sufficient labeled speech data and have significantly fewer speakers [10]. Currently, 3,170 languages are classified as endangered, placing them at high risk of extinction [11]. This alarming statistic underscores the urgency for preservation and revitalization efforts to protect these invaluable cultural and lin-

guistic treasures. Speech processing technologies, including low-resource ASR, have emerged as vital tools in supporting the revitalization and preservation of these languages [12].

Recent approaches for developing low-resource ASR rely on semi-supervised learning methods (e.g., Cross-lingual wav2vec [8]) and multi-task multi-lingual learning methods (e.g., Whisper [9]), which leverage large-scale unlabeled speech data to pre-train an transformer-based ASR model [13]. These pre-trained models encapsulate extensive multi-lingual and speaker-specific information, that creating a robust meta prior. Fine-tuning these models with small amounts of labeled speech data enables them to achieve reasonable accuracy, while fine-tuning with additional labeled data can further enhance their performance.

During fine-tuning to improve low-resource ASR systems, data augmentation approaches have been employed to introduce variability to the input speech, enhancing model generalization and robustness [14–16]. Additionally, text-to-speech (TTS) systems have also been investigated to synthesize supplementary speech data for low-resource languages [17]. Moreover, Zero-shot TTS models [18–21] have emerged as a promising approach, aiming to generate speech in any speaker's voice using only a reference speech sample. These models enable the creation of synthetic speech data to diversify speakers' speech style within low-resource datasets. However, their impact on ASR performance remains suboptimal, as current data augmentation approaches [14–16] and low-resource zero-shot TTS system [17] show limited capability to generate synthetic speech data that captures diverse speakers' speech styles.

Note that synthesized speech generated by TTS systems often exhibits imperfections, such as unnatural prosody or artifacts [22, 23]. A significant challenge arises from the domain mismatch between true and synthesized speech data, which negatively affects the performance of low-resource ASR systems. This issue has been explored in research on children's speech ASR, where adapter-based approaches have been utilized to mitigate domain mismatches [24–27]. Despite these efforts, the gap persists, restricting the effectiveness of synthesized data in enhancing ASR accuracy for low-resource languages.

Additionally, low-resource ASR systems encounter practical constraints in data collection. Low-resource speech data are usually collected intermittently due to the challenges in organizing the recording procedure, such as coordinating recording equip-

ment and recruiting diverse speaker candidates. This incremental data collection process introduces challenges, including catastrophic forgetting during continuous model training, where the model struggles to retain knowledge previously learned from data while adapting to new data. These constraints further complicate the development of robust low-resource ASR systems [28].

In summary, while existing literature on low-resource ASR has laid a strong foundation, current methods often yield suboptimal performance due to research gaps in the aforementioned challenges. Hereby, we investigate to bridge these gaps in this thesis, introducing innovative approaches to enhance low-resource ASR systems through advanced data augmentation techniques, low-resource zero-shot TTS synthesis, strategies for addressing domain mismatch, and methods to mitigate catastrophic forgetting in continual learning. These contributions aim to significantly improve the accuracy and generalization of ASR systems for low-resource languages.

1.2 Motivations for this research

Despite notable advancements in ASR for low-resource languages, several critical challenges remain as summarized below:

- **Lack of labeled speech data, insufficient diversity in speakers, and imbalanced speech volume among speakers**. Fine-tuning pre-trained ASR models with small datasets from low-resource languages can achieve acceptable accuracy [29, 30]. However, high-resource ASR systems achieve human-level performance mainly due to the availability of large-scale labeled speech data and the inclusion of diverse and representative speakers in the training datasets [31, 32]. The scarcity of such resources for low-resource languages limits the potential for achieving similar levels of accuracy and generalization. Moreover, Low-resource datasets often exhibit an imbalanced distribution of speech volume per speaker, with a small number of speakers contributing significantly more recordings than the rest [10]. This long-tail distribution in the dataset can hinder the generalization capability of the ASR system, as it may overfit to the dominant speakers while underperforming on speech from underrepresented speakers [33].

- **Domain mismatch between true and synthesized speech data**. To ad-

dress the lack of diverse and representative speakers, zero-shot text-to-speech (TTS) systems are often employed to generate synthetic speech data [18, 19, 21]. These systems can generate unlimited labeled speech data with diverse characteristics using text input and reference speech samples. However, the synthesized data exhibit domain mismatches with true speech data due to imperfections in the generation process, which results in suboptimal performance when developing low-resource ASR systems [22–27].

- **Catastrophic forgetting during continual learning of ASR model**. The collection and labeling of speech data for low-resource languages are often conducted in an intermittent and incremental manner. This gradual data acquisition process increases computational costs and presents challenges for training ASR systems from scratch. While continual training can mitigate some of these issues by allowing models to integrate new data over time, catastrophic forgetting remains a major obstacle [28]. This phenomenon occurs when a model forgets previously learned knowledge while adapting to new data, thereby compromising its long-term accuracy and generalization across domains. Effectively alleviating catastrophic forgetting is essential to ensure that ASR systems remain robust and reliable as new data is incrementally introduced during training [34, 35] .

1.3 Research Objectives

- **Objective 1** aims to address the challenges posed by limited labeled speech data, insufficient speaker diversity, and imbalanced speech volumes among speakers in low-resource datasets. By focusing on investigating data augmentation approaches and low-resource zero-shot TTS systems, this objective seeks to generate high-quality synthetic speech data that closely resemble true speech data, thereby enhancing ASR systems for low-resource languages.

- **Objective 2** is to address the domain mismatch between true and synthesized speech data generated by zero-shot TTS systems, that exploring methods to filter and align synthesized data with the characteristics of true speech.

- **Objective 3** aims to address the challenge of catastrophic forgetting in low-

resource ASR systems caused by incrementally collected speech data, which involves exploring continual learning strategies to preserve model performance.

1.4 Contributions

Throughout this thesis, we propose four research directives to investigate the aforementioned challenges of low-resource ASR: speech data augmentation (Chapter 2, Objective 1), catastrophic forgetting in continual learning (Chapter 3, Objective 3), low-resource zero-shot TTS (Chapter 4, Objective 1), and domain mismatch for true and synthesized data (Chapter 5, Objective 2). The contributions in each of these chapters are summarized as follows:

(i) CyclicAugment: Speech Data Random Augmentation with Cosine Annealing Scheduler for Automatic Speech Recognition

- We propose CyclicAugment, a novel data augmentation approach that dynamically adjusts the magnitude of augmentation policies using a cosine annealing scheduler. This approach generates more diversified augmentation policies, assisting the model to escape local optima more effectively than static augmentation policies with consistent magnitudes for enhancing an ASR system.
- We demonstrate that dynamically configuring the magnitude of augmentation policies provides effective variability in the data augmentation process. Empirical evaluations show that CyclicAugment significantly outperforms baseline approaches, including SpecAugment [14] and RandAugment [36], by enhancing ASR model performance during training.

(ii) CLRL-Tuning: Continual Learning by Randomly Layer-wise Tuning of a Pre-trained Model for Automatic Speech Recognition

- We propose a direct Continual Learning approach, referred to as Randomly Layer-wise Tuning (CLRL-Tuning), for fine-tuning pre-trained ASR models. Using wav2vec 2.0 [29] as the base model, CLRL-Tuning tackles the randomness of incrementally collected datasets by selectively updating the parameters of a randomly chosen encoder layer of the model during each training epoch. This allows the model to quickly adapt to a new task with minimal additional training, while still leveraging the knowledge and capabilities of the pre-trained

model.

- Experimental evaluations demonstrate that CLRL-Tuning outperforms four strong baseline methods, including Knowledge Distillation [34] and Gradient Episodic Memory [35], achieving significant improvements in average word error rate (WER). By selectively updating layers during training, CLRL-Tuning effectively mitigates catastrophic forgetting and enhances model adaptability to incrementally collected datasets.

(iii) Zero-Voice: A Low-Resource Approach For Zero-shot Text-to-Speech

- We propose Zero-Voice, a novel end-to-end zero-shot text-to-speech (TTS) model based on a hierarchical framework comprising a Source Filter Network, an Acoustic Feature Encoder, a diffusion-based Acoustic Feature Refiner, and a diffusion-based Waveform Vocoder. This architecture eliminates the need for latent representations, thereby avoiding intermediate information compression and loss. It facilitates the effective transfer of speech style features from reference inputs to synthesized speech while addressing the mismatch between acoustic representations (i.e., mel-spectrograms) and waveform outputs during both training and inference. At the core of Zero-Voice is a novel Source Filter Network that enables unsupervised decoupling of prosodic components—pitch, rhythm, and timbre—from reference speech. A key innovation is the introduction of a random cut-and-connect data augmentation operation applied to reference speech, which enhances pattern diversity during training. This augmentation operation significantly improves the model's zero-shot generalization, especially in low-resource settings.

- Experimental results demonstrate that Zero-Voice exhibits strong generalization capabilities, effectively resembling reference speech samples even when trained on minimal amounts of transcribed data (e.g., 5–8 hours).

- We propose low-resource ASR development for Te Reo Māori, an official and endangered language of New Zealand, we have legitimately collected and labeled 27 hours of Māori read speech data. Zero-Voice is trained on this dataset to flexibly synthesize additional high-quality Māori speech data, addressing challenges of the lack of labeled data, insufficient speaker diversity, and imbalanced speech volumes across speakers within the true dataset. The combination of true and

synthesized data is then used to fine-tune pre-trained ASR models for Te Reo Māori, achieving improved performance on the publicly available FLEURS Te Reo Māori (mi_nz) test set [10]: Word Error Rate (WER) of 22.9% and a Character Error Rate (CER) of 11.2%. These results significantly outperform the current benchmark, which reported 37.9% WER and 13.7% CER on the Māori (`nz_mi`) test set of the Google Fleurs dataset.

(iv) Synthesized Data Selection via Score Distribution Matching for Te Reo Māori Automatic Speech Recognition

- We propose a novel score-distribution-matching data selection approach to address the domain mismatches between synthesized and true data. This approach filters synthesized data by aligning its score distribution with the prior distribution of true data, ensuring closer resemblance and improved compatibility for the synthesized dataset.

- We use our proposed data selection approach to optimize the performance of the Te Reo Māori ASR system proposed in the Zero-Voice study [21] . Specifically, we filter the synthetic speech samples generated by the Zero-Voice model and fine-tune pre-trained ASR models with a combination of true and selected synthetic data to improve recognition accuracy. Experimental results demonstrate significant improvements in ASR accuracy on the FLEURS (mi_nz) Te Reo Māori test set [10], reducing the WER to 21.4% and CER to 9.9%

1.5 Organization of this thesis

Literature reviews of the low-resource speech recognition methods are presented in each chapter corresponding to the proposed four novel approaches.

The rest of this thesis is organized as follows:

Chapter 2 presents the proposed CyclicAugment, based on our work published in the conference of *Interspeech* 2022 (2022), titled " CyclicAugment: Speech Data Random Augmentation with Cosine Annealing Scheduler for Automatic Speech Recognition" [37].

Chapter 3 presents CLRL-Tuning approach, which is based on our work published in the conference of *Interspeech* 2023 (2023), titled "CLRL-Tuning: A Novel Continual

Learning Approach for Automatic Speech Recognition" [38].

Chapter 4 presents the Zero-Voice approach, derived from our work currently submitted to the Journal *Computer Speech and Language*, titled "Zero-Voice: A Low-Resource Approach for Zero-Shot Text-to-Speech" [21].

Chapter 5 presents the Score-distribution-matching approach, which is based on our work published the conference of *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, titled "Synthesized Data Selection via Score Distribution Matching for Te reo Māori Automatic Speech Recognition" [39].

Chapter 6 concludes this thesis and discusses future work and directions.

Note that Statements of Contributions are inserted at the beginning of each relevant chapter, references related to each chapter are listed at the end of each chapter.

References

[1] T. Qian, J. Liu, and M. T. Johnson. Efficient embedded speech recognition for very large vocabulary mandarin car-navigation systems. *IEEE Transactions on Consumer Electronics*, 55(3):1496–1500, 2009.

[2] C. Chiu, T. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, and et al. State-of-the-art speech recognition with sequence-to-sequence models. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[3] K. Ohtsuki, T. Matsuoka, T. Mori, K. Yoshida, Y. Taguchi, S. Furui, and K. Shirai. Japanese large-vocabulary continuous speech recognition using a newspaper corpus and broadcast news. *Speech Communication*, 28(2):155–166, 1999.

[4] O. Saz, S.-C. Yin, E. Lleida, R. Rose, W.R. Rodríguez, and C. Vaquero. Tools and technologies for computer-aided speech and language therapy. *Speech Communication*, 51(10):948–967, 2009.

[5] R. Wang and J. Lu. Investigation of the golden speaker for a language learner from the imitation preference perspective by voice modification. *Speech Communication*, 53(2):175–184, 2011.

[6] T. Kawahara, H. Wang, and C. Waple. Computer-assisted language learning

system based on dynamic question generation and error prediction for automatic speech recognition. *Speech Communication*, 51(10):995–1005, 2009.

[7] J .Lu, R. Wang, L. C. De Silva, Y. Gao, and J. Liu. CASTLE: a computer-assisted stress teaching and learning environment for learners of english as a second language. In *Proc. Interspeech*, pages 606–609, 2010.

[8] S. Ruder, A. Søgaard, and I. Vulić. Unsupervised cross-lingual representation learning. In Preslav Nakov and Alexis Palmer, editors, *Proceedings of Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38. Association for Computational Linguistics, 2019.

[9] A. Radford, J. Wook Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, volume 202, pages 28492–28518, 2023.

[10] A. Conneau, M. Ma, S. Khanuja, Yu Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *IEEE Spoken Language Technology Workshop (SLT)*, 2022.

[11] E. M. David, G. F. Simons, and C. D. Fennig, editors. *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, twenty-seventh edition, 2024.

[12] V. Edwin, A. W. Yeo, S. F. Juan, and B. Chin. Employing information and communication technologies in the revitalisation and maintenance of indigenous languages of sarawak. In *Proc. International Conference On Minority And Majority: Language, Culture And Identity (ICMM)*, 2010.

[13] A. Vaswani, N. Shazeerand, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[14] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech*, pages 2613–2617, 2019.

[15] X. Song, Z. Wu, Y. Huang, Dan Su, and Helen Meng. Specswap: A simple data augmentation method for end-to-end speech recognition. In *Proc. Interspeech*, pages 581–585, 2020.

[16] E. Kharitonov, M. Rivire, L. Wolf G. Synnaeve, P. Mazar, M. Douze, and E. Dupoux. Data augmenting contrastive learning of speech representations in the time domain. In *Proc.IEEE Spoken Language Technology Workshop (SLT)*, pages 215–222, 2021.

[17] E. Casanova, C. Shulby, A. Korolev, A. Candido Junior, A. da Silva Soares, S. Aluísio, and M. Antonelli Ponti. Asr data augmentation in low-resource settings using cross-lingual multi-speaker tts and cross-lingual voice conversion. In *Interspeech*, pages 1244–1248, 2023.

[18] E. Casanova, J. Weber, C. Shulby, A. Júnior, E. Gölge, and M. A. Ponti. Yourtts: Towards zero-shot multispeaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning (ICML)*, 2022.

[19] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber. Xtts: a massively multilingual zero-shot text-to-speech model. In *Interspeech*, pages 4978–4982, 2024.

[20] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In *Interspeech 2024*, pages 4978–4982, 2024.

[21] Z. Wang, F. Hou, and R. Wang. Zero-voice: A low-resource approach for zero-shot text-to-speech. In *TechRxiv*, 2025.

[22] W. Wang, Z. Zhou, Y. Lu, H. Wang, C. Du, and Y. Qian. Towards data selection on tts data for children's speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6888–6892, 2021.

[23] T. Rolland and A. Abad. Improved children's automatic speech recognition combining adapters and synthetic data augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12757–12761, 2024.

[24] T. Rolland and A. Abad. Exploring adapters with conformers for children's automatic speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12747–12751, 2024.

[25] T. Rolland and A. Abad. Introduction to partial fine-tuning: A comprehensive evaluation of end-to-end children's automatic speech recognition adaptation. In *Interspeech*, pages 5178–5182, 2024.

[26] T. Rolland and A. Abad. Shared-adapters: A novel transformer-based parameter efficient transfer learning approach for children's automatic speech recognition. In *Interspeech*, pages 2370–2374, 2024.

[27] T. Rolland and A. Abad. Towards improved automatic speech recognition for children. In *IberSPEECH*, pages 251–255, 2024.

[28] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motiva-*

tion, volume 24, pages 109–165, 1989.

[29] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NIPS)*, volume 33, pages 12449–12460, 2020.

[30] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W. Hsu, A. Conneau, and M. Auli. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[32] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Oriental COCOSDA 2017*, 2017.

[33] Z. Yang, M. Ding, T. Huang, Y. Cen, J. Song, B. Xu, Y. Dong, and J. Tang. Does negative sampling matter? a review with insights into its theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 5692–5711, 2024.

[34] Z. Li and D. Hoiem. Learning without forgetting. In *Proc. European Conference on Computer Vision (ECCV)*, pages 614–629, 2016.

[35] D. Lopez-Paz and M. A. Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pages 6467–6476, 2017.

[36] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18613–18624, 2020.

[37] Z. Wang, F. Hou, Y. Qiu, Z. Ma, S. Singh, and R. Wang. Cyclicaugment: Speech data random augmentation with cosine annealing scheduler for automatic speech recognition. In *Interspeech 2022*, pages 3859–3863, 2022.

[38] Z. Wang, F. Hou, and R. Wang. Clrl-tuning: A novel continual learning approach for automatic speech recognition. In *Interspeech 2023*, pages 1279–1283, 2023.

[39] Z. Wang, F. Hou, and R. Wang. Synthesized data selection via score distribution matching for te reo māori automatic speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 18172–18176, 2026.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Zhihan Wang
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work:	Chapter 2

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Zhihan Wang, Feng Hou, Yuanhang Qiu, Zhizhong Ma, Satwinder Singh, Ruili Wang. (2022) CyclicAugment: Speech Data Random Augmentation with Cosine Annealing Scheduler for Automatic Speech Recognition. Proc. Interspeech 2022, 3859-3863.

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Zhihan Digitally signed by Zhihan Date: 2025.02.03 08:21:07 +13'00'
Date:	03-Feb-2025
Primary Supervisor's Signature:	Professor Ruili Wang Digitally signed by Professor Ruili Wang DN: cn=Professor Ruili Wang, c=NZ, o=Massey University, ou=SMCS, email=ruili.wang@massey.ac.nz Date: 2025.08.06 11:37:54 +12'00'
Date:	

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

Chapter 2

CyclicAugment: Speech Data Random Augmentation with Cosine Annealing Scheduler for Automatic Speech Recognition

Recent speech data augmentation approaches use static augmentation operations or policies with consistency magnitude scaling. However, few work is done to explore the influence of the dynamic magnitude of augmentation policies. In this chapter, we propose a novel speech data augmentation approach, CyclicAugment, to generate more diversified augmentation policies by dynamically configuring the magnitude of augmentation policies with a cosine annealing scheduler. We also propose additional augmentation operations to enlarge the diversity of augmentation policies. Motivated by learning rate warm restart and cyclical learning rates, we hypothesize that using dynamically configured magnitude for augmentation policies can also help escape local optima more efficiently than static augmentation policies with consistency magnitude scaling. Experimental results demonstrate that our approach is effective for escaping local optima. Our approach achieves 12%-35% relative improvement in word error rate (WER) over SpecAugment and RandAugment on the LibriSpeech 960h dataset, and achieves state-of-the-art result 7.1% in phoneme error rate (PER) on the TIMIT 5h dataset.

2.1 Introduction

Data augmentation has been effective to improve the generalization of Automatic Speech Recognition (ASR) models, especially by generating synthetic speech data for low resource languages [1]. Typically, the raw waveform was manipulated directly to synthesise new data by injecting noisy audio signals [2], applying speed perturbation [3] or an acoustic room simulator [4] or normalising vocal tract length [5], and so on. Since the mel-spectrogram of speech data is more effective for representing emotion, pitch, and accent than raw waveform [6], augmentation approaches that directly operate on mel-spectrogram of speech data were proposed. These approaches, such as SpecAugment [7], SpecAugment++[8], SpecSwap [9], SpecMix [10] and MixSpeech [11], treat mel-spectrogram as image and apply image data manipulation operations [12, 13]. However, these approaches use statically mixed augmentation operations (e.g., frequency masking, time masking and time warp) and cannot automatically optimize the augmentation policy.

In image processing domain, several image data augmentation approaches [14–17] were proposed to automatically find the optimal data augmentation policy based on proxy tasks. This can overcome the performance limitation caused by cumbersome empirical exploration. However, these policy search based data augmentation approaches are computationally expensive, and lack of interpretation whether the optimal augmentation policies found by the proxy tasks were also the optimal ones for the neural networks being trained. RandAugment [18] was proposed to reduce computational cost by randomly selecting image augmentation operations from a much smaller search space and only searching the strength of all the augmentation operations, i.e., the magnitude of a policy. With a much simplified algorithm, RandAugment even achieves marginal improvements over the above approaches that are computationally expensive. In the studies of SCADA (Stochastic, Consistent and Adversarial Data Augmentation) [19] for ASR, RandAugment was applied to the mel-spectrogram of speech data, and was shown to be slightly more effective than the SpecAugment approach.

Notably, the aforementioned speech and image data augmentation approaches explore augmentation operations or policies for reducing overfitting in the training of deep neural models. However, these approaches use a constant magnitude for augmentation policies; few work is done to explore the influence of the dynamic magnitude of

augmentation policies.

In this chapter, we propose a novel speech data augmentation approach, CyclicAugment, to generate more diversified augmentation policies by dynamically configuring the magnitude of augmentation policies with a cosine annealing scheduler. Motivated by learning rate warm restart of stochastic gradient descent [20] and cyclical learning rates [21], we hypothesize that training deep neural networks with dynamically configured magnitude for augmentation policies can also help escape local optima more efficiently than static augmentation policies with consistency magnitude scaling. We configure the search space of audio data augmentation operations similarly to RandAugment, but we propose additional augmentation operations to enlarge the diversity of augmentation policies. We evaluate our approach by training three popular end-to-end ASR models with our CyclicAugment approach. Experimental results demonstrate 12%-35% relative improvement in word error rate (WER) over SpecAugment and RandAugment on the LibriSpeech 960h dataset, and achieves state-of-the-art result 7.1% in phoneme error rate (PER) on the TIMIT 5h dataset. The performance gain found periodically in accordance with the cosine graph indicate that our approach is effective for escaping local optima when training an ASR model.

2.2 Speech Data Augmentation Operations

In this section, we present the definition of the individual data augmentation operations on the spectral-domain and the time-domain of speech data for our proposed CyclicAugment approach.

2.2.1 Spectral-Domain Augmentation Operations

We adopt three augmentation operations (i.e., frequency masking, time masking and time warping) proposed by SpecAugment [7]. In addition, to introduce a more variety of mel-spectrogram variants, we design two augmentation operations for the mel-spectrogram: time shifting and loudness amplifying, which are modified from two augmentation approaches for raw waveform data [22]. The definition of each augmentation operation on the spectral-domain is described in detail as below.

Frequency masking is applied to f consecutive input frequency channels where $[f_0, f_0 + f)$ are masked, $f_0 \in [0, h - f]$ where h is the number of input frequency

channels, and $f \in [0,\ h \times V]$, where V is used to adjust the frequency masking size. The masked portion of the mel-spectrogram are padded with 0 or the minimum value of the input data.

Time masking is applied to t consecutive input time steps where $[t_0, t_0 + t)$ are masked, $t_0 \in [0, w - t]$ and w is the number of time steps for the input mel-spectrogram, and $t \in [0,\ w \times V]$, where V is used for adjusting time masking size. The masked portion of the mel-spectrogram are padded with 0 or the minimum value of the input data.

Time warping is to warp the input mel-spectrogram on the time axis by t percent from the original point, where $t \in [0, V]$, and V is the time warping parameter which increases the speed of a voice. As the augmented input data on the time axis is shorter than the original input data after warping, the gap is filled with 0 or the minimum value of the input data.

Time shifting is to shift the input mel-spectrogram on the time axis by t percent from the original point, where t $\in [-V, V]$, and V is the time shifting parameter which forwards or backwards a speech with a random percentage. The gap left from the mel-spectrogram's forward and backward shifting is filled with 0 or the minimum value of the input data.

Loudness amplifying is applied to t consecutive input time steps at $[t_0, t_0 + t)$, t_0 $\in [0, w - t)$ where w is the number of time steps for the input mel-spectrogram, and $t \in [0, w \times 0.15]$. The loudness of speech at $[t_0, t_0 + t)$ is amplified by λ, $\lambda \in (0,\ V]$ where V is the extent of the loudness amplifying .

2.2.2 Time-Domain Augmentation Operations

For the raw waveform of speech data, we select 4 single augmentation operations of WaveAugment [23] developed in a public library for the time-domain speech data augmentation: pitch modification (*pitch*), additive noise (*add*), reverberation (*reverb*), and time drop (*tdrop*). We define their augmentation operations in detail as below.

Pitch modification is applied to the entire input time steps. The change in the pitch is uniformly sampled by p, and $p \in [-V, +V]$, where V is used to control pitch change.

Reverberation is applied to the entire input time steps with uniformly sampled room-size of r, $r \in [-V, +V]$,where V is the reverberation strength.

Additive noise is applied to the entire input time steps by adding a Gaussian white noise with scale of w, $w \in [-V, +V]$, where V is to control the noise scale.

Time drop is applied to t consecutive input time steps where $[t_0, t_0 + t)$ are masked with 0 values, $t_0 \in [0,\ w - t]$ and w is the number of time steps for the waveform of the input data, and $t \in [0,\ w \times V]$, where V is to adjust the time drop span.

2.3 Cyclic Augmentation Approach

In this section, we present the design of our random augmentation search space and introduce our approach of dynamically configuring the augmentation magnitude (i.e., representing the magnitude scaling) with a cosine annealing scheduler. We use the magnitude to control the scale of augmentation policies applied to the speech data.

2.3.1 Speech Data Random Augmentation with Dynamic Magnitude

Based on the individual augmentation operations defined in Section 2.3.1, we configure their unified parameters V as shown in Table 2.1 for the scale of each augmentation operation for the speech data.

Table 2.1: Parameters (V) for the spectral-domain and the time-domain augmentation operations.

	Operations	***V***
Spectral-Domain	Frequency masking	0.15
	Time masking	0.2
	Time warping	20
	Time shifting	5
	Loudness Amplifying	1.0
Time-Domain	Pitch modification	150
	Reverberation	50
	Additive noise	20
	Time drop	0.2

We merge the spectral-domain augmentation operations or the time-domain augmentation operations as shown in Table 2.1 in a search space (S), based on the form of

speech data for training an ASR model (i.e., either mel-spectrogram or raw waveform).

Algorithm 1 presents the process of the speech data random augmentation. The on-the-fly speech data is transformed by a number of (N) randomly selected augmentation operations from the search space (S), and then augmented with a unified magnitude (M) during the training of ASR models. Algorithm 1 is a similar approach to RandAugment [18] applied to speech data for enlarging the diversity of speech data variants further.

Algorithm 1: Pseudo-code for speech data random augmentation

input: search space of augmentation operations S, number of operations N, magnitude of the CyclicAugment M, input speech data x
for $n \leftarrow 1$ *to* N **do**
  $s \leftarrow$ randomly select one operation from S
  $m \leftarrow M\times$ augmentation parameter V of s
  $y \leftarrow$ apply s to x with augmentation parameter m
end
return y

2.3.2 Configuring Augmentation Magnitude Variable M

For the magnitude variable M representing magnitude scaling in Algorithm 1, we dynamically configure it with a cosine annealing scheduler as follows.

$$M = \alpha \left(\cos 2\pi \frac{T_{cur}}{T_{period}} + 1 \right) \tag{2.1}$$

where T_{cur} is the current training epoch value; T_{period} is the period of cosine function, and α is a scalar value. Figure 2.1 shows an example of the magnitude M arranged by a cosine annealing scheduler with $T_{period} = 4$ and $\alpha = 0.75$, where M varies between 0 and 1.5 during each period (i.e., 4 epochs) of the training of ASR models.

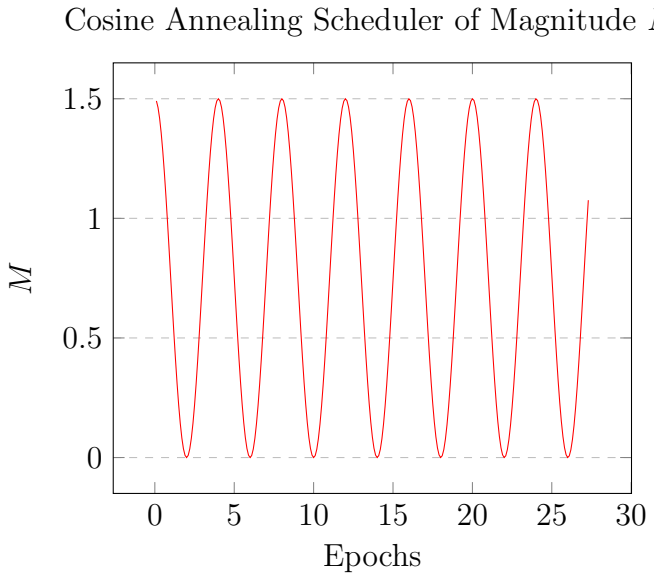

Figure 2.1: Example of variable M tuned with a cosine annealing scheduler during the training of ASR models.

2.4 Experiments and Results

2.4.1 LibriSpeech 960h Speech Recognition for Supervised Learning

2.4.1.1 Models

We use two end-to-end ASR models, the LAS (Listen, Attend, and Spell) [24] and the Conformer [25] for supervised learning setting. We use a vocabulary size of 16000 Word Piece Model [26] to tokenize the text transcripts of the LibriSpeech 960h [27] dataset for training the two models.

For the LAS model, the input data is fed into a 6 convolutional layers of very deep Convolution Neural Network (VGG) [28] to extract the input features. The LAS encoder with 4 layers of bi-directional LSTMs of size 1024, is used to process the output of VGG and generate attention vectors. The attention vectors are then passed into the LAS decoder consisting of 2 layers of bi-directional LSTMs of size 1024 to generate the text predictions.

For the Conformer model, we use the medium size architecture [25] as follows. The input data is processed by convolution kernel with size of 32 for feature extraction.

The Conformer encoder consists of 16 layers of conformer blocks with dimension size 256 and 4 attention heads. Its decoder has one layer of LSTM with cell size 640 for generating the text predictions.

For both models, we use a beam size of 8 for beam search to obtain the final transcripts of the text predictions. To get further performance improvements, we incorporate a 3-gram ARPA language model of LibriSpeech 960h dataset [27] by shallow fusion with the two ASR models.

2.4.1.2 Spectral-Domain Augmentation Policies

We extract 80-dimension filter banks as the input mel-spectrogram data for the two ASR models. To demonstrate the performance of our approach, we compare our CyclicAugment approach with two baselines: SpecAugment [7] and RandAugment [18]. Their policy settings are as follows: (i) For SpecAugment, we apply each input mel-spectrogram with time masking, frequency masking and time warping from the spectral-domain operations in Table 1; (ii) We use RandAugment approach by transforming each input mel-spectrogram with 3 randomly selected individual spectral-domain augmentation operations in Table 1; (iii) We use our CyclicAugment approach with $N = 3$, and a cosine annealing scheduler (i.e., $T_{period} = 4$, $\alpha = 2.0$), that we randomly select three spectral-domain operations in Table 1 and configure their magnitude with a period of 4 epochs cosine annealing scheduler.

All the other hyperparameters of the above three augmentation policies are fixed without applying additional tuning to inspect the effectiveness of the different augmentation policies for the training of ASR models.

2.4.1.3 Experiment Results

Both the LAS model and the Conformer model are trained from scratch on the LibriSpeech 960h datasets with the two baselines and our approach (i.e., SpecAugment, RandAugment and our CyclicAugment). Both models are trained for 200 epochs using the Adam optimizer [29] with a learning rate of 1e-4.

The WER results in Table 2.2 show that ASR models trained with our CyclicAugment approach achieves 12%-35% relative improvements compared with that trained with SpecAugment and RandAugment. In addition, the RandAugment approach has marginal improvement (i.e., less than 5% relative improvement) over the SpecAug-

ment approach. This aligns with the results in comparison with SCADA [19] approach.

Table 2.2: WER (%) of LAS and Conformer trained with the three augmentation policies assessed on LibriSpeech 960h test sets.

Model + Policy	**No LM**		**With LM**	
	Clean	**Other**	**Clean**	**Other**
LAS [24] + SpecAug [7]	8.1	16.2	6.7	14.1
LAS [24] + RandAug [18]	7.9	16.0	6.2	14.0
LAS [24] + CyclicAug (ours)	**6.9**	**14.3**	**5.1**	**12.0**
Conformer [25] + SpecAug [7]	7.3	12.2	4.7	9.2
Conformer [25] + RandAug [18]	7.0	11.2	4.3	8.8
Conformer [25] + CyclicAug (ours)	**5.5**	**9.3**	**3.6**	**6.2**

2.4.2 TIMIT 5h Phoneme Recognition for Semi-Supervised Learning

2.4.2.1 Model

Semi-supervised learning pre-trains a model with large quantities of unlabeled data and then fine-tunes it with labeled data. For our semi-supervised learning setting, we fine-tune the pre-trained wav2vec2.0 [30] ASR models on TIMIT 5h phoneme recognition corpus [31]. Specifically, we use two LARGE size wav2vec2.0 models: LARGE LS-960 (pre-trained on LibriSpeech 960h [27]) and LARGE LV-60K (pre-trained on LibriVox of 60,000 hr unlabled speech data [32]).

2.4.2.2 Time-Domain Augmentation Policies

Both aforementioned wav2vec2.0 models are pre-trained with raw audio data rather than mel-spectrogram form. Thus, we fine-tune them with the time-domain augmentation operations in Table 2.1. When fine-tuning the wav2vec2.0 models, we apply three augmentation policies to the raw waveform speech data in comparison with their performance: (i) the Time drop operation in Table 1 implemented in the original wav2vec2.0 paper [30], (ii) the RandAugment [18] approach with 2 randomly selected time-domain operations in Table 1, (iii) our CyclicAugment with parameters set to $N = 2$, and the magnitude M is scheduled by a cosine annealing scheduler (i.e., $T_{period} = 4$, $\alpha = 1.0$).

Both pre-trained models are fine-tuned for 100 epochs with Adam [29] in a learning

rate of 3e-5. For the TIMIT dataset, we use 90% TRAIN set for fine-tuning the pre-trained models, 10% TRAIN set for validation, and the TEST set for phoneme predictions with a merged 39 phonemes configuration [33].

Table 2.3: PER (%) achieved by fine-tuning LARGE LS-960 and LARGE LV-60K with three augmentation policies assessed on TIMIT test set.

Model + Policy	**TEST SET**
LARGE LS-960 [30] + Time drop (original)	8.3
LARGE LS-960 [30] + Time drop	8.4
LARGE LS-960 [30] + RandAug [18]	8.2
LARGE LS-960 [30] + CyclicAug (ours)	**7.7**
LARGE LV-60K [30] + Time drop	8.0
LARGE LV-60K [30] + RandAug [18]	7.7
LARGE LV-60K [30] + CyclicAug (ours)	**7.1**

2.4.2.3 Experiment Results

In Table 2.3, our proposed CyclicAugment approach has 8-12% relative gains measured in PER over the Time drop approach and the RandAugment approach. Further, our CyclicAugment approach achieves the state-of-the-art performance (i.e., 7.7% PER for the LARGE LS-960 model and 7.1% PER for the LARGE LV-60K model) compared with the 8.3% PER reported in the wav2vec2.0 paper [30] for the TIMIT dataset.

2.4.3 Optimal Model Checkpoints with CyclicAugment

In this sub-section, we interpret the performance gains achieved by using our CyclicAugment approach to train an ASR model. Figure 2.2 shows CyclicAugment magnitude M, training loss and validation PER % evaluated on the TIMIT dataset [31] for our semi-supervised learning setting. Curves of different colours denote the three approaches used to fine-tune the LARGE LV-60K wav2vec2.0 pre-trained model: Time drop, RandAugment [18] and our CyclicAugment approach. Combining the training loss curves and validation PER % curves, we can see that training ASR models with the Time drop approach tends to be overfitting with its weaker magnitude, and training with the RandAugment approach delays overfitting with a stronger magnitude as seen in the performance gains over the Time drop approach. While for the training loss curve of using our CyclicAugment approach, a distinct pattern is the periodical

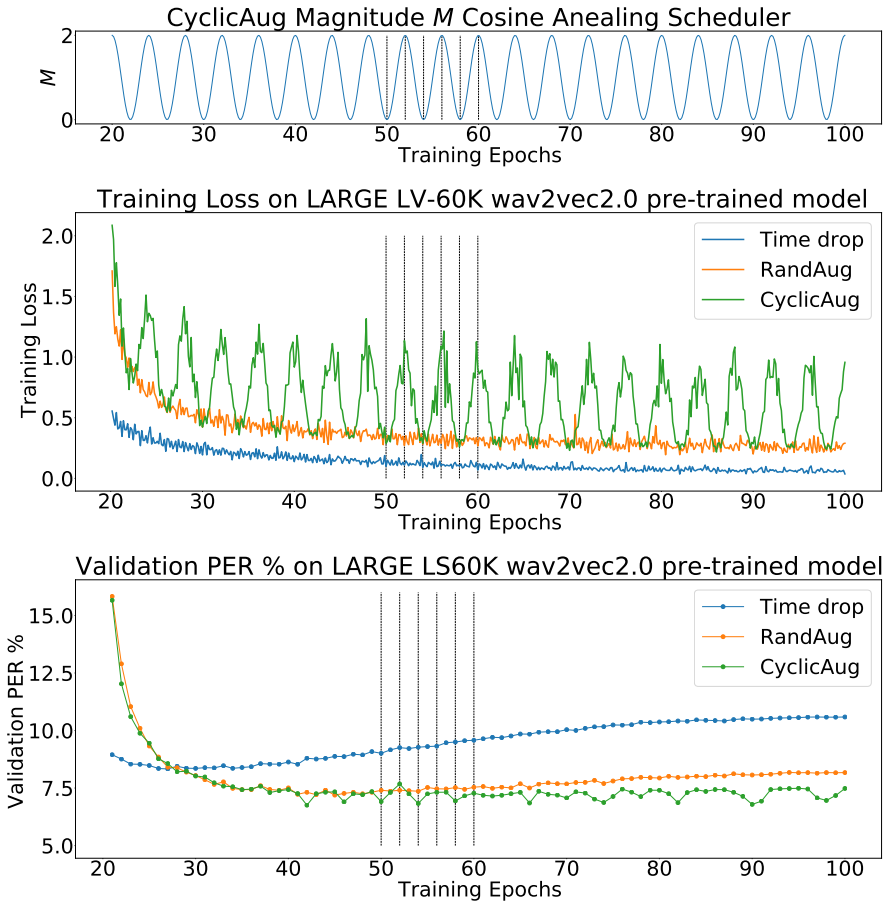

Figure 2.2: CyclicAugment magnitude M, training loss curve and validation PER % of using the three approaches to fine-tune the LARGE LV-60K wav2vec2.0 pre-trained model on TIMIT dataset.

high and low training loss synchronized with the periodical magnitude of augmentation operations (i.e., the cosine annealing scheduler of the magnitude M). Our proposed dynamic magnitude of augmentation policies works in a mechanism similar to the learning rate warm restart of stochastic gradient descent [20] and cyclical learning rate [21], which is effective for escaping local optima. Thus, training an ASR model with our CyclicAugment approach can visit multiple local optima and have higher chance to converge to the global optima. Thus, good validation results can be found frequently at the troughs of the cosine annealing scheduler of CyclicAugment, and the changes around the crests indicate that the model is likely to escape the local optimal and start moving to the next one.

Recent speech data augmentation approaches (e.g., SpecAugment [7], SpecSwap [9],

WaveAugment [23]) focus on exploring augmentation operations or policies with consistency magnitude mainly to reduce overfitting. Our proposed approach dynamically configure the magnitude of augmentation policies. It delays overfitting, and also can assist in escaping local optima during the model training with longer training epochs. Moreover, our CyclicAugment approach can be used jointly with adaptive learning rate optimizers (e.g., Adam [29] and Adadelta [34]) for training an ASR model. Training ASR models with our approach can converge quickly towards local optima with the adaptive learning rate optimizers, and periodically escape local optima and converge to a global optimum.

2.5 Conclusions

In this chapter, we propose a computationally cheap yet effective data augmentation approach, CyclicAugment, to generate diversified data augmentation policies for ASR. CyclicAugment uses a cosine annealing scheduler to dynamically configure the magnitude of augmentation policies. Experiment results show that our CyclicAugment approach is effective for both mel-spectrogram and raw waveform speech data in supervised learning and semi-supervised learning settings. Our proposed CyclicAugment approach can improve generalization, and also can help escape local optima in the training of ASR models.

We will try the slanted triangular learning rates [35] approach to tune the magnitude of augmentation policies. We will also test our CyclicAugment approach on other speech processing technologies, such as speech enhancement [36] and smoker identification [37].

References

[1] A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales. Data augmentation for low resource languages. In *Proc. Interspeech 2014*, pages 810–814, 2014.

[2] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv 1412.5567*, 2014.

[3] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur. Audio augmentation for speech recognition. In *Proc. Interspeech*, pages 3586–3589, 2015.

[4] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. N. Sainath, and M. Bac-

chiani. Generation of Large-Scale Simulated Utterances in Virtual Rooms to Train Deep-Neural Networks for Far-Field Speech Recognition in Google Home. In *Proc. Interspeech*, pages 379–383, 2017.

[5] N. Jaitly and G. E. Hinton. Vocal tract length perturbation (vtlp) improves speech recognition. In *Proc. International Conference on Machine Learning (ICML)*, volume 28, 2013.

[6] E. Sejdić, I. Djurović, and J. Jiang. Time–frequency feature representation using energy concentration: An overview of recent advances. *Digital Signal Processing*, 19(1):153–183, 2009.

[7] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech*, pages 2613–2617, 2019.

[8] H. Wang, Y. Zou, and W. Wang. SpecAugment++: A Hidden Space Data Augmentation Method for Acoustic Scene Classification. In *Proc. Interspeech*, pages 551–555, 2021.

[9] X. Song, Z. Wu, Y. Huang, D. Su, and H. Meng. Specswap: A simple data augmentation method for end-to-end speech recognition. In *Proc. Interspeech*, pages 581–585, 2020.

[10] G. Kim, D. K. Han, and H. Ko. SpecMix : A Mixed Sample Data Augmentation Method for Training with Time-Frequency Domain Features. In *Proc. Interspeech*, pages 546–550, 2021.

[11] L. Meng, J. Xu, X. Tan, J. Wang, T. Qin, and B. Xu. Mixspeech: Data augmentation for low-resource automatic speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7008–7012, 2021.

[12] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.

[13] S. Yun, D. Han, S. Chun, S. Oh, Y. Yoo, and J. Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019.

[14] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. Learning transferable architectures for scalable image recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8697–8710, 2018.

[15] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le. Autoaugment:

Learning augmentation strategies from data. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 113–123, 2019.

[16] D. Ho, E. Liang, I. Stoica, P. Abbeel, and X. Chen. Population based augmentation: Efficient learning of augmentation policy schedules. In *Proc. International Conference on Machine Learning (ICML)*, 2019.

[17] S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim. Fast autoaugment. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6665–6675, 2019.

[18] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18613–18624, 2020.

[19] G. Wang, A. Rosenberg, Z.Chen, Y. Zhang, B. Ramabhadran, and P. Moreno. Scada: Stochastic, consistent and adversarial data augmentation to improve asr. In *Proc. Interspeech*, 2020.

[20] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[21] Leslie N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472, 2017.

[22] L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190, 1983.

[23] E. Kharitonov, M. Rivire, L. Wolf G. Synnaeve, P. Mazar, M. Douze, and E. Dupoux. Data augmenting contrastive learning of speech representations in the time domain. In *Proc.IEEE Spoken Language Technology Workshop (SLT)*, pages 215–222, 2021.

[24] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964, 2016.

[25] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040, 2020.

[26] M. Schuster and K. Nakajima. Japanese and korean voice search. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,

pages 5149–5152, 2012.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. International Conference on Learning Representations, (ICLR)*, 2015.

[29] P. K.Diederik and B. Jimmy. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[30] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12449–12460, 2020.

[31] J. Garofolo, Lori Lamel, W. Fisher, Jonathan Fiscus, D. Pallett, N. Dahlgren, and V. Zue. Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*, 1992.

[32] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. Libri-light: A benchmark for asr with limited or no supervision. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673, 2020.

[33] K.-F. Lee and H.-W. Hon. Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1641–1648, 1989.

[34] M. D. Zeiler. ADADELTA: an adaptive learning rate method. *arXiv preprint arxiv 1212.5701*, 2012.

[35] J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In *Proc. the Association for Computational Linguistics (ACL)*, pages 328–339, 2018.

[36] Y. Qiu, R. Wang, S. Singh, Z. Ma, and F. Hou. Self-supervised learning based phone-fortified speech enhancement. In *Proc. Interspeech*, pages 211–215, 2021.

[37] Z. Ma, Y. Qiu, F. Hou, R. Wang, Wai C., Joanna T., and C. Bullen. Determining the best acoustic features for smoker identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8177–8181, 2022.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Zhihan Wang
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work:	Chapter 3

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Zhihan Wang, Feng Hou, Wang, Ruili Wang (2023) CLRL-Tuning: A Novel Continual Learning Approach for Automatic Speech Recognition. Proc. Interspeech 2023, 1279-1283.

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Zhihan Digitally signed by Zhihan Date: 2025.02.03 08:41:21 +13'00'
Date:	03-Feb-2025
Primary Supervisor's Signature:	Professor Ruili Wang Digitally signed by Professor Ruili Wang DN: cn=Professor Ruili Wang, c=NZ, o=Massey University, ou=SMCS, email=ruili.wang@massey.ac.nz Date: 2025.08.06 11:38:37 +12'00'
Date:	

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

Chapter 3

CLRL-Tuning: A Novel Continual Learning Approach for Automatic Speech Recognition

We propose a novel ***C****ontinual* ***L****earning approach, which is* ***R****andomly* ***L****ayer-wise Tuning (CLRL-Tuning) of a pre-trained Automatic Speech Recognition (ASR) model. CLRL-Tuning tackles the randomness of subsequent datasets by updating the parameters of randomly selected encoder layers of the pre-trained model (such as wav2vec 2.0) for every training epoch. CLRL-Tuning is different from the previous approaches in that it neither uses previous datasets, nor expands/runs previous models. Furthermore, we perform experiments to evaluate our approach compared with four strong baselines, including* Knowledge Distillation *and* Gradient Episodic Memory*. Our approach achieves significant improvements over the baselines in average word error rate (WER) for the wav2vec 2.0 model. Additionally, we implement ablation studies for our approach by tuning one, three, six and full encoder layers of the model, and experimental results show only tuning one encoder layer of the model at each training epoch is the most effective way to mitigate catastrophic forgetting.*

3.1 Introduction

Continual/Incremental Learning is a process that trains a model sequentially on multiple incrementally collected datasets rather than trains it once on a whole combined dataset [1]. However, the differences in data distributions of datasets may induce *catastrophic forgetting* [2] of knowledge learned from previous datasets. For computer vision and automatic speech recognition (ASR), many methods have been developed to mitigate catastrophic forgetting. These methods can be categorized into three approaches: the parameter-based approach, data-based approach and regularization-based approach.

Parameter-based methods dynamically increase model capacity to maintain previous knowledge by either creating sub-networks for different tasks [3–5] or increasing the numbers of layers and/or neurons [6, 7]. Here are two examples of creating sub-networks for speech recognition: (i) Factorizing the likelihood model of an HMM-DNN-based ASR model into sub-models for different domains of datasets [8]; (ii) Inserting adapters into pre-trained ASR models for learning new languages [9]. Nevertheless, parameter-based methods will significantly increase model sizes and make it difficult to tune hyper-parameters of the models, especially when a new dataset size is unknown.

Data-based methods [10–13] replay data either from stored previous datasets or generated examples. However, the data used for the previous tasks may not be permanently stored where data storage is limited or where there are legal restrictions on data storage. (e.g., for privacy concerns [14]).

Regularization-based methods [15–19] incorporate additional regularization terms to consolidate the knowledge from a previously trained model. This approach has been applied to a multi-dialect acoustic model [20], which is continually trained with regularization-based methods (e.g, Learning without forgetting [15] and Elastic Weight Consolidation [17]). This approach has also been applied to the continual training of CTC-based ASR models [21] and pre-trained ASR models [22]. However, regularization-based methods require running previous models and can restrict the freedom of the model to learn new tasks.

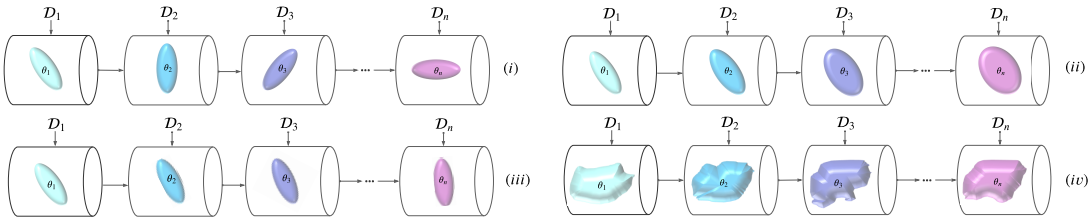

Figure 3.1: Continual learning with different methods of mitigating catastrophic forgetting illustrated in schematic parameter spaces, where the *nth* model with parameters $\boldsymbol{\theta_n}$ is trained on the *nth* dataset $\mathcal{D}_n$ sequentially. (i) Direct fine-tuning method updates the model parameters to fit a new dataset, leading to the problem of catastrophic forgetting where the model parameters diverge from the region of the previous parameters; (ii) Parameter-based methods incrementally add new parameters to the model to adapt to new dataset, but this can make it difficult to tune hyper-parameters of the model; (iii) To mitigate the issue of model parameters deviating too far from the previous parameters, Data-based methods employ the replay of previous data samples to remind the model of its prior training on previous datasets, and Regularization-based methods add regularization terms to the loss function that penalizes changes to the parameters. (iv) CLRL-Tuning makes minor adjustments to the parameter space through layer-wise tuning of a large pre-trained model. The advantage of using a pre-trained model is that it has already learned a generic and large shape of the parameter space from larges-scale of unlabeled data. Consequently, most of the pre-trained model parameters can be shared across all datasets.

3.2 Motivation

Recently, pre-trained language models with parameter-efficient prompt tuning has made significant progress, that learning lightweight task-specific embeddings while freezing parameters of a pre-trained language model [23–25]. This allows the model to quickly adapt to a new task with minimal additional training, while still leveraging the knowledge and capabilities of the pre-trained model. Intuitively, we hypotheses a large pre-trained ASR model contains prior knowledge learned from previous datasets and large-scale of unlabelled speech utterances. As a result, most of the pre-trained ASR model parameters trained on different speech datasets are likely to be reusable in solving the continual learning problem.

In this paper, we propose a **C**ontinual **L**earning approach by **R**andomly **L**ayer-wise Tuning (CLRL-Tuning) to utilize prior knowledge of a pre-trained ASR model, i.e., wav2vec 2.0. Following the principle of lightweight tuning, which is to freeze most of

the pre-trained parameters and only tune a smaller set of parameters [23–25], CLRL-Tuning updates the parameters of a randomly selected encoder layer of the model for every epoch. In Figure 3.1, we demonstrate a theoretical analysis schematically to compare our proposed CLRL-Tuning approach with the previous approaches. Our proposed CLRL-Tuning approach is different from the previous approaches in that it neither uses previous datasets, nor expands/runs previous models. We perform experiments to evaluate our approach compared with four strong baselines, including *Knowledge Distillation* [15] and *Gradient Episodic Memory* [10]. Our approach achieves significant improvements over the baselines in average word error rate (WER) for the pre-trained wav2vec 2.0 [26] model. Additionally, we implement ablation studies for our approach by tuning one, three, six, and full encoder layers of the model. Experimental results show only tuning one encoder layer of the model is the most effective way to mitigate catastrophic forgetting.

3.3 Methodology

3.3.1 Continual Learning by Randomly Layer-wise Tuning of a Pre-trainded Model (CLRL-Tuning)

Suppose a pre-trained ASR model (e.g., wav2vec 2.0 [26], HuBERT [27]) is continually trained on T datasets $\mathcal{D}_1, \mathcal{D}_2 ... \mathcal{D}_T$ with different data distributions, the model parameters trained with these datasets is denoted as $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_T$, respectively. We denote the current model as $\boldsymbol{\theta}_t$ trained on the tth dataset $\mathcal{D}_t$, and $t \in [1, \ldots, T]$. Let the transformer-based pre-trained ASR model comprises of M encoder layers (i.e., an encoder layer consists of multi-heads self-attention sub-layer and fully connected feed forward sub-layer). Before the start of every epoch to train the current dataset $\mathcal{D}_t$ on the current model $\boldsymbol{\theta_t}$, we randomly set a fixed number N encoder layers for parameters updating, and the rest of the $M - N$ encoder layers are set to be frozen. For each step of the training epoch, a batch of input speech data is passed to the current model $\boldsymbol{\theta}_t$ to generate a text prediction. We use the Connectionist Temporal Classification (CTC) [28] loss as the *objective* function to calculate the loss between the generated texts from the model $\boldsymbol{\theta}_t$ and the ground truth of the labeled text from the current dataset $\mathcal{D}_t$. This CTC loss ($\mathcal{L}_{CTC}$) is used to update the parameters of the current model $\boldsymbol{\theta}_t$.

The training framework for our proposed CLRL-Tuning approach is summarized in Algorithm 2. Our CLRL-Tuning approach trains a pre-trained ASR model on various datasets sequentially, and mitigates catastrophic forgetting by randomly selecting and tuning partial encoder layers of the model and freezing the rest of them during every training epoch. Randomly selecting and tuning partial encoder layers resembles the randomness of the future datasets distributions. Thus, our approach is more effective for mitigating catastrophic forgetting. Moreover, our approach is simple to implement and avoids the disadvantages of the parameter-based approach, the data-based approach, and the regularization-based approach, that require modifying neural network structure (or expanding neural network capacity), reusing of previous datasets (i.e., $\mathcal{D}_{t-1}$) and reusing of previous ASR models (i.e., $\boldsymbol{\theta}_{t-1}$).

Algorithm 2: CLRL-Tuning

Inputs: number of stages T, number of epochs for each stage K, number of training steps for each stage S, pre-trained model $\boldsymbol{\theta}$, number of total encoder layers M, number of encoder layers selected for tunning N, datasets $\mathcal{D} = [\mathcal{D}_1, \ldots, \mathcal{D}_T]$

$\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}$
for $t \leftarrow 1$ *to* T **do**
  $\boldsymbol{\theta}_t^0 \leftarrow \boldsymbol{\theta}_{t-1}$
  for $k \leftarrow 1$ *to* K **do**
    randomly select N indices of encoder layers
    load $\boldsymbol{\theta}_t^{k-1}$ and freeze the rest $M - N$ layers
    $\boldsymbol{\theta}_t^k \leftarrow$tune the model with S steps on $\mathcal{D}_t$
  end
  $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_t^K$
end
return $\boldsymbol{\theta}_T$

3.4 Experiments And Results

3.4.1 Datasets

We use four open source datasets, LibriSpeech-960h [29], TIMIT-5h [30], LJSpeech-26h [31], and VCTK-44h [32], denoted as $\mathcal{D}_1$,$\mathcal{D}_2$,$\mathcal{D}_3$,$\mathcal{D}_4$ respectively, for our experiments. More details about the four datasets with the configuration of their training set, test set, and validation set are shown in Table 3.1.

Table 3.1: Datasets for the continual learning experiments

Data set	Training Set	Test Set	Validation Set
LibriSpeech $\mathcal{D}_1$	train-clean-100	test-clean	dev-clean
TIMIT $\mathcal{D}_2$	90% subset of TRAIN	10% subset of TRAIN	TEST
LJSpeech $\mathcal{D}_3$	80% subset of LJSpeech	10% subset of LJSpeech	10% subset of LJSpeech
VCTK $\mathcal{D}_4$	80% subset of VCTK	10% subset of VCTK	10% subset of VCTK

3.4.2 Experimental Settings

To evaluate our proposed CLRL-Tuning approach, we adopt a commonly used transformer-based ASR model: **wav2vec 2.0** [26] pre-trained on the full LibriSpeech-960h [29] dataset. wav2vec 2.0 comprises of *twelve* encoder layers with *twelve* attention heads for each encoder layer. We extract raw waveform with 16000hz for the proposed datasets $\mathcal{D}_1$,$\mathcal{D}_2$,$\mathcal{D}_3$,$\mathcal{D}_4$. We use Adam [33] to optimise the learning rate when training the ASR model for the experiments. When decoding the prediction of the ASR model, we use the beam search with a beam size of 10.

The wav2vec 2.0 model is continually trained using three schemes: (i) **Individual** training : wav2vec 2.0 is trained on the speech datasets $\mathcal{D}_1$,$\mathcal{D}_2$,$\mathcal{D}_3$,$\mathcal{D}_4$ individually to observe the performances of the single models; (ii) **Fine-tuning**: wav2vec 2.0 is continually trained on the datasets sequentially to show catastrophic forgetting; (iii) **Mitigating Forgetting**: wav2vec 2.0 is continually trained on the datasets sequentially with the *Knowledge Distillation* [15] method, the *Gradient Episodic Memory* [10] method (two strong baselines used in the lifelong learning of end-to-end ASR studies [21]), and our proposed approach CLRL-Tuning to compare with their performances.

We continually train the pre-trained wav2vec 2.0 model with four training stages in the order of $\mathcal{D}_1$,$\mathcal{D}_2$,$\mathcal{D}_3$,$\mathcal{D}_4$, and the model is trained for 50 epochs for each stage. We use the current model trained on $\boldsymbol{\theta}_t$ to evaluate them on the test sets of the datasets $\mathcal{D}_t, \mathcal{D}_{t-1}, \ldots, \mathcal{D}_1$. For example, we evaluate the current model $\boldsymbol{\theta}_3$ on the current training dataset $\mathcal{D}_3$, and the previous datasets $\mathcal{D}_2$ and $\mathcal{D}_1$. To show the catastrophic forgetting of sequential **fine-tuning** and compare the effectiveness of **mitigating forgetting** methods, we need to start from a shared fine-tuned model on $\mathcal{D}_1$. Thus, the first model $\boldsymbol{\theta}_1$ in Algorithm 2 is fine-tuned directly on the dataset $\mathcal{D}_1$, and we apply mitigating forgetting methods to the subsequent datasets.

3.4.2.1 Knowledge Distillation (KD)

For the Knowledge Distillation [15] method, a regularization-based approach, we calculate a KL divergence [34] between the output distributions of the previous model $\boldsymbol{\theta}_{t-1}$ and the current model $\boldsymbol{\theta}_t$ as a regularization term to constrain parameters shifting from the previous model $\boldsymbol{\theta}_{t-1}$. The regularization term is summed with the CTC loss as the *objective* function for the training of the current model θ_t.

3.4.2.2 Gradient Episodic Memory (GEM)

The *Gradient Episodic Memory method* [10] is a data-based approach. To update the gradients of the current model $\boldsymbol{\theta}_t$, we use stored samples with a length of 30 minutes from the previous dataset $\mathcal{D}_{t-1}$, and select samples close to the median lengths of utterance of the dataset $\mathcal{D}_{t-1}$, that is an optimal set up for the GEM method proposed in the study of [21].

3.4.2.3 Our Approach (CLRL-Tuning)

We apply our CLRL-Tuning approach with four settings: tuning one ($N = 1$), three ($N = 3$), six ($N = 6$) and full ($N = 12$) encoder layers and freeze the rest of them to continually train wav2vec 2.0 on the datasets sequentially. As an ablation study presented in Section 3.4.4, we compare the effect of the number of encoder layers being tuned for the model.

3.4.3 Experimental Results

The results of **Individual** training, **Fine-tune** training, and **Mitigating forgetting** are summarized in Table 3.2. For a fair and reasonable comparison, we use the average word error rate (Avg WER) [8, 21] as the evaluation metric to compare our proposed *CLRL-Tuning* approach with the baselines. We present the optimal result for our approach, i.e., randomly tuning one encoder layer ($N = 1$) of wav2vec 2.0 during each epoch. Comparing the results of individual training and fine-tune training, we can see that continual training with direct fine-tuning (i.e., tuning the full ($N = 12$) encoder layers of the wav2vec 2.0 model) causes catastrophic forgetting. Using our proposed CLRL-Tuning approach can retain knowledge more effectively than using the KD method and the GEM method.

We then use the validation sets of $\mathcal{D}_1$,$\mathcal{D}_2$,$\mathcal{D}_3$,$\mathcal{D}_4$ to plot the continual learning curves of

our proposed CLRL-Tuning approach and the baselines in Figure 3.2. The curves in four different colours represent the four methods used in our experiments, and they are plotted in four windows to denote the four training stages on Librispeech $\mathcal{D}_1$, Timit $\mathcal{D}_2$, Ljspeech $\mathcal{D}_3$, and VCTK $\mathcal{D}_4$. At each training stage, the learning curves of the Fine-tuning, the KD and the GEM methods (plotted in blue, red and green in Figure 3.2, respectively) jump up when the model is trained on the new dataset, while the learning curve of our CLRL-Tuning (i.e., the black curve and $N = 1$) grows remarkably more gentle than theirs. The learning curves plotted in Figure 3.2 also indicate that our method is more effective for mitigating catastrophic forgetting than the Fine-tuning, the KD and the GEM methods, since the Fine-tuning method results in significant deviation of the model parameters from the region of the previous parameters (as demonstrated in Figure 3.1 (i)). Conversely, the KD and the GEM methods can mitigate this deviation and alleviate catastrophic forgetting. (as shown in Figure 3.1 (ii)).

3.4.4 Ablation Study for CLRL-Tuning

In this section, we inspect the effect of the number of encoder layers being tuned for the wav2vec 2.0 model on the datasets $\mathcal{D}_1$,$\mathcal{D}_2$,$\mathcal{D}_3$,$\mathcal{D}_4$ sequentially. We attempt to tune one ($N = 1$), three ($N = 3$), six ($N = 6$) and full ($N = 12$) encoder layers (i.e., full encoder layers tuning is equivalent to the Fine-tuning approach) of the model. Table 3.3 shows the less encoder layers being tuned, the more effective for the model to alleviate catastrophic forgetting. Figure 3.3 shows the learning curves on the validation sets of the datasets for the four training stages. We can see that tuning less encoder layers of the model makes the learning curves grow more gentle at each training stage. Therefore, tuning one encoder layer ($N = 1$) of the model can most effectively mitigate catastrophic forgetting.

The results for the ablation study show that a pre-trained ASR model with *good prior knowledge* can efficiently learn similar speeches patterns with only minor adjustments to its parameters, since most parameters of the model are shareable across all the datasets (as depicted in Figure 3.1 (iv)). Tuning one encoder layer ($N = 1$) of the model results in minimal changes to the parameter, while still leveraging the shared parameters for good generalization across all datasets. However, tuning a larger number of encoder layers ($N = 3, N = 6, N = 12$) results in greater deviation of the model parameters from the previous parameters, which causes catastrophic forgetting.

Table 3.2: Experiment test sets results (WER %) on various methods.

Method \ Datasets	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	*Avg*
Individual	5.8	20.1	3.1	3.2	8.1
Fine-tuning (i.e., CLRL-Tuning, $N = 12$)					
$\boldsymbol{\theta}_2$	6.8	12.0			9.4
$\boldsymbol{\theta}_3$	11.3	23.6	2.5		12.5
$\boldsymbol{\theta}_4$	16.4	27.3	11.4	2.2	14.3
Knowledge Distillation [15]					
$\boldsymbol{\theta}_2$	6.8	11.7			9.3
$\boldsymbol{\theta}_3$	8.9	16.7	2.5		9.4
$\boldsymbol{\theta}_4$	13.4	21.8	9.5	3.0	11.9
Gradient Episodic Memory [10]					
$\boldsymbol{\theta}_2$	6.7	11.8			9.3
$\boldsymbol{\theta}_3$	9.0	16.3	2.3		9.2
$\boldsymbol{\theta}_4$	13.2	20.2	8.5	3.2	11.3
Our Approach (CLRL-Tuning, $N = 1$)					
$\boldsymbol{\theta}_2$	6.9	11.7			9.3
$\boldsymbol{\theta}_3$	7.8	15.6	2.4		8.6
$\boldsymbol{\theta}_4$	9.8	16.7	6.4	2.6	8.9

Table 3.3: Test sets results (WER %) for the ablation study with N randomly selected encoder layers for CLRL-tuning.

Method \ Datasets	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	*Avg*
CLRL-Tuning ($N = 12$) (i.e.,Fine-tuning)					
$\boldsymbol{\theta}_2$	6.8	12.0			9.4
$\boldsymbol{\theta}_3$	11.3	23.6	2.5		12.5
$\boldsymbol{\theta}_4$	16.4	27.3	11.4	2.2	14.3
CLRL-Tuning ($N = 6$)					
$\boldsymbol{\theta}_2$	6.9	11.9			9.4
$\boldsymbol{\theta}_3$	9.2	21.9	2.3		11.1
$\boldsymbol{\theta}_4$	13.2	25.2	10.8	2.2	12.9
CLRL-Tuning ($N = 3$)					
$\boldsymbol{\theta}_2$	6.8	11.8			9.3
$\boldsymbol{\theta}_3$	8.8	19.3	2.2		10.1
$\boldsymbol{\theta}_4$	12.4	21.9	8.5	2.4	11.3
CLRL-Tuning ($N = 1$)					
$\boldsymbol{\theta}_2$	6.9	11.7			9.3
$\boldsymbol{\theta}_3$	7.8	15.6	2.4		8.6
$\boldsymbol{\theta}_4$	9.8	16.7	6.4	2.6	8.9

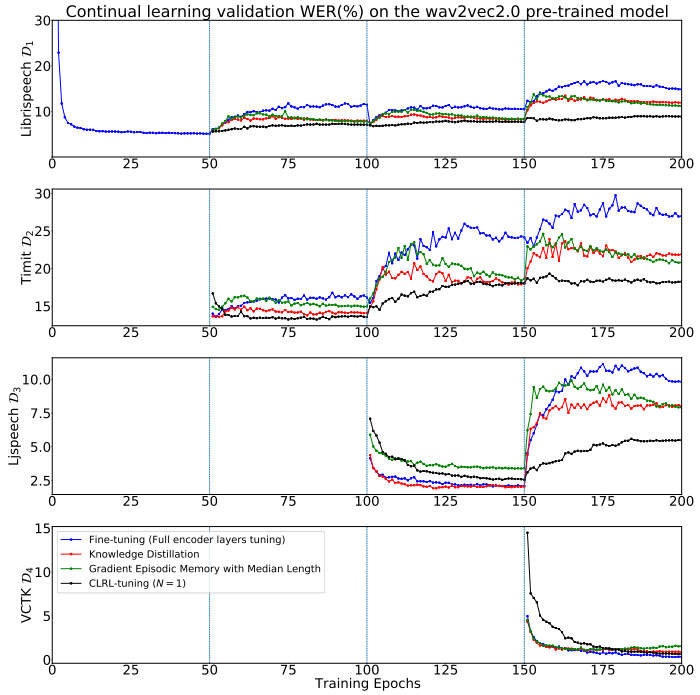

Figure 3.2: Validation sets learning curves for the wav2vec 2.0 pre-trained model with the methods of Fine-tuning, KD, GEM and our CLRL-Tuning ($N = 1$).

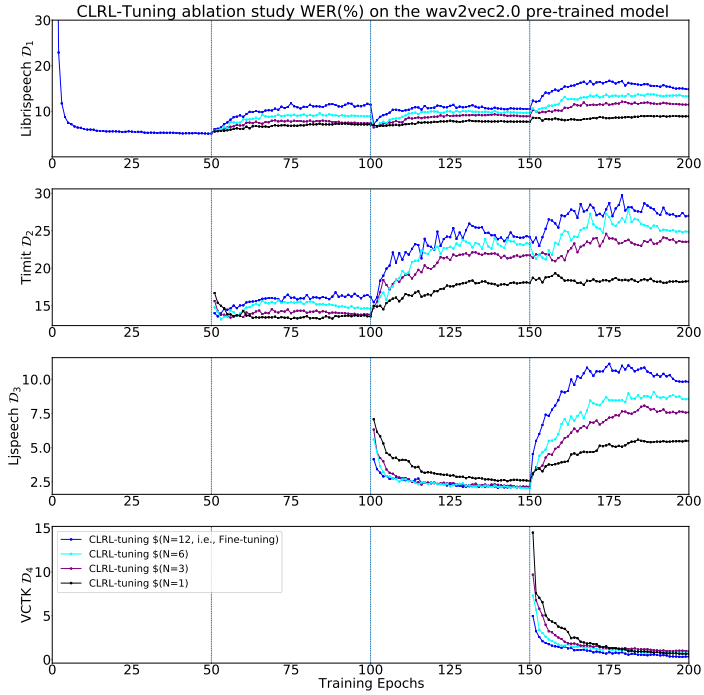

Figure 3.3: Validation sets learning curves for the wav2vec 2.0 pre-trained model with CLRL-tuning of various number of encoder layers ($N = 1, N = 3, N = 6, N = 12$).

3.5 Conclusions and Future Work

In this chapter, we propose a continual learning approach that tunes partial encoder layers of a pre-trained ASR model for effectively retaining the knowledge learned from previous datasets. Experimental results show our approach can utilize the prior knowledge of a pre-trained ASR model (i.e., trained on large amount of unlabelled data) to alleviate the catastrophic forgetting. Notably, our approach has neither memory restriction, nor extra parameter tuning, nor privacy considerations compared with previous approaches.

Future work will test CLRL-Tuning on larger wav2vec 2.0 models with more extensive unlabeled data, such as the wav2vec 2.0 model pre-trained on the LirbiVox 60K dataset [35], to examine whether larger networks and more unlabeled data can further enhance the performance of our approach. Additionally, we propose to investigate tuning more detailed sub-layers of the pre-trained ASR models, such as recent advances in speaker adaptation for the conformer transducer model [36].

References

[1] G. I. Parisia, R. Kemkerb, J. L.Part, C. Kanan, and S. Wermtera. Continual lifelong learning with neural networks: A review. In *Neural Networks*, volume 113, pages 54–71, 2019.

[2] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165, 1989.

[3] A. Mallya and S. Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7765–7773, 2018.

[4] J. Serra, D. Suris, M. Miron, and A. Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *Proc. International Conference on Machine Learning (ICML)*, pages 4548–4557, 2018.

[5] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. In *arXiv preprint arxiv 1701.08734*, 2017.

[6] R. Aljundi, P. Chakravarty, and T. Tuytelaars. Expert gate: Lifelong learning

with a network of experts. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7120–7129, 2017.

[7] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. Progressive neural networks. In *arXiv preprint arxiv 1606.04671*, 2016.

[8] S. Sadhu and H. Hermansky. Continual learning in automatic speech recognition. In *Proc. Interspeech*, pages 1246–1250, 2020.

[9] S. Kessler, B. Thomas, and S. Karout. An Adapter Based Pre-Training for Efficient and Scalable Self-Supervised Speech Representation Learning. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3179–3183, 2021.

[10] D. Lopez-Paz and M. A. Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pages 6467–6476, 2017.

[11] S.A. Rebu, A. Kolesnikov, G. Sperl, and C. H. Lampert. icarl: Incremental classifier and representation learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2001–2010, 2017.

[12] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proc. European Conference on Computer Vision (ECCV)*, pages 532–547, 2018.

[13] R. Aljundi, E. Belilovsky, T. Tuytelaars, L. Charlin, M. Caccia, M. Lin, and L. Page-Caccia. Online continual learning with maximal interfered retrieval. In *Advances in Neural Information Processing Systems (NIPS)*, pages 11849–11860, 2019.

[14] P. McClure, C. Y. Zheng, J. R. Kaczmarzyk, J. A. Lee, S. S. Ghosh, D. Nielson, P. Bandettini, and F. Pereira. Distributed weight consolidation: a brain segmentation case study. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4097–4107, 2018.

[15] Z. Li and D. Hoiem. Learning without forgetting. In *Proc. European Conference on Computer Vision (ECCV)*, pages 614–629, 2016.

[16] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin. Learning a unified classifier incrementally via rebalancing. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 831–839, 2019.

[17] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, and A. Grabska-Barwinska. Overcoming catas-

trophic forgetting in neural networks. In *Proc. National Academy of Sciences*, volume 114, 2016.

[18] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proc. European Conference on Computer Vision (ECCV)*, pages 139–154, 2018.

[19] C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner. Variational continual learning. In *arXiv preprint arXiv:1710.10628*, 2017.

[20] B. Houston and K. Kirchhoff. Continual learning for multidialect acoustic models. In *Proc. Interspeech*, pages 576–580, 2020.

[21] H. Chang, H. Lee, and L. Lee. Towards lifelong learning of end-to-end asr. In *Proc. Interspeech*, pages 2551–2555, 2021.

[22] J. Lee, C. Lee, J. Choi, J. Chang, W. Seong, and J. Lee. CTRL: Continual Representation Learning to Transfer Information of Pre-trained for WAV2VEC 2.0. In *Proc. Interspeech*, pages 3398–3402, 2022.

[23] X. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 4582–4597, 2021.

[24] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, pages 3045–3059, 2021.

[25] X. Liu, K. Ji, Y. Fu, W. Tam, Z. Du, Z. Yang, and J. Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 61–68, 2021.

[26] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NIPS)*, volume 33, pages 12449–12460, 2020.

[27] W. Hsu, Y.Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed. Hubert: How much can a bad teacher benefit asr pre-training? In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 6533–6537, 2021.

[28] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proc. International Conference on Machine Learning (ICML)*, pages 369–376, 2006.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus

based on public domain audio books. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[30] J. Garofolo, Lori Lamel, W. Fisher, Jonathan Fiscus, D. Pallett, N. Dahlgren, and V. Zue. Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*, 1992.

[31] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

[32] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. https://doi.org/10.7488/ds/2645, 2017.

[33] P. K.Diederik and B. Jimmy. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[34] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22 (1):79–86, 1951.

[35] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, and P. E. Mazaré et al. Libri-light: A benchmark for asr with limited or no supervision. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673, 2020.

[36] Y. Huang, G. Ye, J. Li, and Y. Gong. Rapid Speaker Adaptation for Conformer Transducer: Attention and Bias Are All You Need. In *Proc. Interspeech*, pages 1309–1313, 2021.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Zhihan Wang
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work:	Chapter 4

Please select one of the following three options:

○ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

◉ The manuscript is currently under review for publication – please indicate:

- The name of the journal:

Computer Speech and Language

- The percentage of the manuscript/published work that was contributed by the candidate: 80.00
- Describe the contribution that the candidate has made to the manuscript/published work:

The candidate's contributions to this manuscript include conceptualization, investigation, methodology, validation, and drafting the manuscript.

○ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Zhihan Wang Digitally signed by Zhihan Wang Date: 2026.05.17 10:50:45 +12'00'
Date:	17-May-2026
Primary Supervisor's Signature:	Prof. Ruili Wang Digitally signed by Prof. Ruili Wang DN: cn=Prof. Ruili Wang, o=Massey University, ou=School of Mathematical and Computational Sciences, email=ruili.wang@massey.ac.nz Date: 2026.05.25 10:15:27 +12'00'
Date:	25-May-2026

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

Chapter 4

Zero-Voice: A Low-Resource Approach For Zero-shot Text-to-Speech

We propose Zero-Voice, a novel end-to-end zero-shot text-to-speech (TTS) model that adopts a hierarchical framework comprising a Source Filter Network, an Acoustic Feature Encoder, a diffusion-based Acoustic Feature Refiner, and a diffusion-based Waveform Vocoder. This architecture avoids intermediate information compression and loss, facilitating effective transfer of speech style features from reference inputs to synthesized speech, while addressing the mismatch between acoustic representations, mel-spectrograms, and waveform outputs during both training and inference. At the core of Zero-Voice is a novel Source Filter Network that enables the unsupervised decoupling of prosodic components (i.e., pitch, rhythm, and timbre from reference speech). A key innovation in our approach is the random cut-and-connect data augmentation operation, which is applied to reference speech to enhance pattern diversity during training. This operation significantly improves the model's zero-shot generalization, particularly under low-resource conditions. We conduct objective experiments under low-resource settings to compare our model with recent strong zero-shot TTS baseline methods under high-resource settings (e.g., StyleTTS 2 and HierSpeech++). Experimental results demonstrate that Zero-Voice achieves comparable performance to these high-resource methods. Moreover, we legitimately collected and labeled 27 hours Te Reo Māori read speech data (i.e., an official and endangered language of New Zealand). We train the Zero-Voice model on this dataset, and use it to synthesize Te Reo Māori speech data to enhance speech recognition models for the language. This approach yields significant improvements on the Māori test set (language code: nz_mi) from the Google FLEURS dataset, outperforming baseline models by a substantial margin.

4.1 Introduction

Text-to-Speech (TTS) is a fundamental technique in speech processing that transforms written text into audio. The advent of deep neural network-based TTS synthesis models has heralded significant improvements in the speed and quality of audio synthesis. These advancements stem from the development of non-autoregressive acoustic models [1, 2], generative modeling methods [3], and neural vocoders [4, 5]. Furthermore, current diffusion models [6, 7] have demonstrated success in creating high-quality images [8], sounds [9], and specifically, speech for TTS applications [10–14].

To synthesize a single speaker's voice, deep neural TTS models typically consist of an acoustic model and a vocoder. The acoustic model produces time-frequency acoustic features based on the input text, while the vocoder generates waveforms conditioned on these acoustic features. Initially, TTS systems relied on autoregressive acoustic models (e.g., Tacotron [15] and Tacotron2 [16]) to sequentially generate acoustic features, with WaveNet [17] used as the vocoder for synthesizing waveforms. Subsequent research shifted towards generative modeling frameworks, including Generative Adversarial Networks (GANs) [18], Normalizing Flows [19], and Diffusion Models (DM) [6], which enabled parallel speech synthesis (e.g., FastSpeech [1], FastSpeech2 [2], Glow-TTS [3], and Grad-TTS [11]).

Recent investigations [20–22] have focused on synthesizing multiple speakers' voices. Additionally, numerous studies [23–36] have explored zero-shot TTS, which enables TTS systems to synthesize a voice that resembles a specific speaker voice. These approaches typically involve integrating an auxiliary embedding that encodes the vocal characteristics of a reference speech. This embedding is used as a prior information guiding the TTS process. Existing literature on zero-shot TTS can be categorized into two primary groups: (i) Cascaded TTS Models, e.g., Meta-StyleSpeech [28] and Your-TTS [29], which comprise speaker's style encoder, text encoder, duration predictor, decoder, and vocoder. These models used speaker's style encoder to incorporate features from speech prompts to achieve zero-shot TTS capability, and required only a small amount of transcribed speech and text data to synthesis high-quality audio. However, optimizing their hyperparameters is challenging due to their complex architectures, making it difficult to scale to large datasets and model sizes; (ii) Neural Codecs Language Models, e.g, Vall-E [31], SoundStorm [32], and Make-A-Voice [33],

utilized highly compressed audio codecs and transformer-decoder-only language models. By scaling up the speech training dataset to over 60,000 hours and increasing the model size to over 7 billion parameters, these models can perform in-context learning to generate discrete audio units used to transform waveforms.

Zero-shot TTS exhibits considerable potential across various domains, such as producing personalized voice and video contents. Recent zero-shot TTS models (e.g., Vall-E [31], NaturalSpeech2 [34], HierSpeech++ [35], StyleTTS 2 [36]) were proposed in high-resource scenarios, that utilizing large scale multi-speakers speech datasets and pre-trained linguistic and acoustic feature encoders to improve naturalness and verisimilitude of synthesized voices. However, the global linguistic landscape includes over 7,000 languages classified as low-resource, many of which face significant challenges in collecting sufficient transcribed speech and text data for model training [37]. While recent research has explored language transfer learning and adaptation techniques for low-resource TTS (e.g., Adaspeech 4 [27], Guided-TTS [12] and Text-Inductive Graphone-Based Language Adaptation [39]), these method still depend on collecting around 5 minuets paired speech-text paired samples from individual speakers for adaptation. As a result, they lack the flexibility required for generating speech in a zero-shot setting.

In this paper, we propose Zero-Voice, a novel zero-shot TTS approach designed specifically for low-resource settings. Zero-Voice synthesizes voice that are both natural and closely resemble a particular speaker's voice, relying on small amount of transcribed speech data for training. We adopt a conventional cascaded TTS framework for the Zero-Voice model. We employ an Acoustic Feature Encoder, a Acoustic Feature Refiner, and a Waveform Vocoder for end-to-end training and inference. The Acoustic Feature Encoder, based on the transformer architecture [38], processes input phonemes and transfers the style of speech prompts to generate corresponding acoustic features. The Acoustic Feature Refiner and a Waveform Vocoder collaboratively produce the waveform output. Specifically, we address two aspects to optimize the Zero-Voice model with low-resource settings:

- Within the Acoustic Feature Encoder, we develop a novel Source Filter Network to enhance TTS zero-shot capability by unsupervised decoupling the prosody components of the reference speaker's voice: (i) We first diversify speech prompt patterns (i.e., mel-spectrogram, pitch, and energy features extracted from the

reference speaker's waveform) during the training of Zero-Voice by applying a random cut-and-connect operation across the temporal domain of the speech prompts; (ii) Subsequently, these features are stretched or compressed along the time axis to disrupt their rhythm; The processed features are passed through the Source Filter Network, which effectively disentangles the timbre, rhythm, and pitch components of the reference speaker's voice; This network filters out irrelevant elements while preserving essential speech characteristics, thereby enhancing the accuracy and quality of style transfer.

- To enhance the quality of generated speech and improve style transfer, we implement end-to-end training and inference across the Source Filter Network, Acoustic Feature Encoder, Acoustic Feature Refiner, and Waveform Vocoder. The model jointly processes the acoustic features extracted by the Acoustic Feature Encoder. The Acoustic Feature Refiner employs a conditional score-based diffusion model [7], adapted from Grad-TTS [11], to produce refined mel-spectrograms by improving the fidelity of the acoustic features. For waveform synthesis, we integrate a second diffusion model, the DiffWave network [40] as the Waveform Vocoder, which converts the refined mel-spectrograms into high-quality speech waveforms.

Notably, Zero-Voice innovatively integrate two diffusion models (i.e., the Acoustic Feature Refiner and the Waveform Vocoder) within a hierarchical framework to effectively bridge the feature mismatch between mel-spectrograms and playable waveforms. This approach departs from previous methods, e.g., Meta-StyleSpeech [28], Your-TTS [29], HierSpeech++ [35], and StyleTTS 2 [36] , which employ Generative Adversarial Network (GAN) [18] for the vocoder, requiring additional network training. Moreover, our diffusion-based framework operates without relying on latent representations, thereby avoiding intermediate information compression and loss. This contrasts with prior diffusion-based zero-shot TTS models, e.g., StyleTTS 2 [36] and NaturalSpeech 2 [34], which introduce latent variables that can lead to information degradation, that compromising the fidelity of both style and content transfer.

To assess the performance of our proposed model, we conduct experiments on the LibriTTS (train-clean-100 partition) [41] and VCTK [43] datasets, which are considered low-resource settings. The experimental results strongly affirm the superiority of our approach over recent open-sourced baseline models. Specifically, the empiri-

cal findings in Section 4.5.4 for the Ablation Study indicate the substantial effect of reference speech feature decomposition and data augmentation, and the end-to-end training of the Source Filter Network, Acoustic Feature Encoder, Acoustic Feature Refiner and Waveform Vocoder on the quality and similarity of the synthesized voice.

In addition, TTS models have received lots of attention to synthesize speech data for improving low-resource Automatic Speech Recognition (ASR) systems [44–47]. We validate our Zero-Voice approach on Te Reo Māori language (i.e., an official and endangered language for New Zealand). We train the Zero-Voice model on our legitimately collected 27 hours of Te Reo Māori speech data and use it to synthesize additional speech for enhancing ASR systems. This approach achieves significantly improved results on the Te Reo Māori test set of the FLEURS dataset [48], outperforming existing baselines.

This research makes substantial contributions to the development of a robust zero-shot TTS model for low-resource scenarios and low-resource ASR system. The key contributions are as follows:

- We propose Zero-Voice, a novel end-to-end zero-shot text-to-speech (TTS) model that employs a hierarchical framework integrating a Source Filter Network, an Acoustic Feature Encoder, a diffusion-based Acoustic Feature Refiner, and a diffusion-based Waveform Vocoder. This architecture avoids intermediate information compression and loss, enabling effective transfer of speech style features from reference inputs to synthesized speech. It also addresses feature mismatches among acoustic representations, mel-spectrograms, and waveform outputs during both training and inference.

- In the Zero-Voice model, we design a novel Source Filter Network that enables unsupervised decoupling of the reference speaker's prosodic components. We emphasize that our proposed Random Cut-and-Connect data augmentation operation, applied to reference speech, is critical for simulating diversified speech patterns during training, thereby enhancing the zero-shot capability of TTS systems in low-resource settings.

- We demonstrate that Zero-Voice is an effective architecture for low-resource settings, achieving competitive performance in speech similarity compared to baseline models trained with high-resource data. Furthermore, experimental

results show that Zero-Voice presents strong zero-shot performance even when trained on less than 8 hours of speech data.

- With the training of a very small amount of speech data (i.e., Te Reo Māori in this paper), we demonstrate that Zero-Voice can be used to flexibly synthesize a large amount of speech data with diverse reference voice samples, significantly improving the accuracy of low-resource ASR systems.

The rest of this paper is organised as follows. Section 4.2 provides a review of related work. In Section 4.3, we present the methodology of our proposed approach. The experimental setup and results are detailed in Section 4.4 and 5.3.3, respectively. Finally, Section 5.4 concludes the paper.

4.2 Related Work

In this section, we review related work on recent zero-shot TTS methods, focusing on three main approaches: autoregressive and non-autoregressive models, normalizing flows, and diffusion models.

4.2.1 Zero-Shot TTS with Autoregressive and Non-Autoregressive Models

The advent of deep learning-based TTS commenced with autoregressive models, e.g., Tacotron 2 [16], Deep Voice 2 [49], and Deep Voice 3 [50]. These models utilized CNN or RNN-based encoder-decoder architectures to transform text into mel-spectrograms, achieving audio quality comparable to human speech. By controlling intonation and rhythm with auxiliary embedding methods, TTS models enable generation of diverse speech style [51, 52].

Other research endeavors have focused on synthesizing voices of multiple speakers [25, 49, 53] and zero-shot TTS approaches [54]. However, the autoregressive framework used in these TTS models [17], exhibits slow speech generation speeds. Thus, non-autoregressive TTS approaches have emerged to parallelize the generation of mel-spectrogram frames, i.e., Transformer TTS [55], FastSpeech [1], FastSpeech 2 [2], Parallel-WaveNet [56] and EfficientTTS 2 [57] have demonstrated success in significantly accelerating mel-spectrogram generation, surpassing autoregressive TTS

models while preserving speech synthesis quality.

Moreover, Wang et al. have developed a neural codec language model framework (Vall-E [31] for zero-shot speech synthesis. They treated zero-shot TTS as a task of conditional language modeling, that synthesis audios by performing in-context learning. However, Vall-E retains the same limitations as previous autoregressive TTS models [25, 49, 53], including slow inference speed and lack of robustness.

4.2.2 Zero-Shot TTS with Normalizing Flows

Normalizing flow-based generative models have gained significant attention for their advantages in fast TTS generation [58–60]. These models precisely estimate data likelihood using invertible transformations and can be trained to maximize this likelihood. The transformations introduced in [61, 62] have facilitated efficient density estimation and enabled fast and effective sampling.

Recently, researchers have integrated normalizing flows into TTS models (i.e., Waveglow [63], Flowavenet [64], NaturalSpeech [65] and Glow-TTS [3]) to address the slow sampling speeds inherent in earlier autoregressive TTS models [25, 49, 53]. Moreover, ensuring proper alignment between input text and the acoustic features of TTS models, as well as predicting phoneme durations are critical for generating high-fidelity audio. Jim et al. employed the Monotonic Alignment Search algorithm (MAS) [66] in Glow-TTS [3] to map input text to acoustic features, identifying the most probable hidden alignment. This method alleviates pronunciation issues encountered by the autoregressive TTS models [16, 49, 50]. Building on the Glow-TTS [3] framework, Casanova et al. proposed SC-GlowTTS [67], which incorporated a speaker-conditional module into the flow-based decoder for zero-shot TTS.

Additionally, Jim et al. developed Variance Inference Text-to-Speech (VITS [68]), which combined a variational encoder with normalizing flows and adversarial training to improve zero-shot TTS performance.In the VITS model, they incorporated a stochastic duration predictor to capture the inherent variability in speech production, enabling for the generation of speech with diverse rhythms, pitches, and pronunciations from a single text input. For zero-shot TTS applications, Casanova et al. and Lee et al. developed Your-TTS [29] and HierSpeech++ [35] models, respectively, by integrating auxiliary embeddings of a reference speaker's voice into the VITS framework.

4.2.3 Zero-Shot TTS with Diffusion Models

Recently, Diffusion Models [6–8] have emerged as the new state-of-the-art approach for deep generative models. Unlike Generative Adversarial Networks (GAN) containing both mode collapse and mode hopping issues [18], Diffusion Models consist of a forward process through the addition of Gaussian noise, and a reverse process that noise is transformed back into a sample from the target distribution for sample generation. Recent research [40, 69, 70] has explored the application of Diffusion Models [6, 7] in TTS systems. Popov et al. introduced Grad-TTS [11], which employed a score-based diffusion model [7] to generate mel-spectrograms by progressively transforming noise and acoustic features.

For zero-shot TTS using diffusion models, Kang et al. enhanced the Grad-TTS model by appending a hierarchical transformer encoder to construct a representative noise prior distribution for a reference speaker, adapting it into the score-based diffusion model (i.e., Grad-StyleSpeech [13]). Additionally, in Guided-TTS [12], speaker embedding was incorporated into the score-based diffusion model to enable zero-shot speaker adaptation. Recently, in StyleTTS 2 [36], Li et al. combined a large pre-trained speech language model, a differentiable duration predictor, style diffusion, and adversarial training to achieve synthesized voice quality comparable to human-level performance. The use of large-scale datasets has also been explored in NaturalSpeech 2 [34] , which processed a large-scale dataset comprising 44,000 hours of speech and singing data. NaturalSpeech 2 comprised a neural audio codec and residual vector quantizers to generate quantized latent vectors, which were then synthesized through a diffusion model to produce audio. The model also facilitated in-context learning within both the diffusion model and its duration-pitch predictor, further enhancing zero-shot capability.

4.3 Method

The Zero-Voice model framework, as illustrated in Figure 4.1, consists of two main components: the training procedure and the inference procedure. This section outlines a detailed overview of both procedures, focusing on the three key components of the model: the Acoustic Feature Encoder, the Acoustic Feature Refiner, and the Waveform Vocoder.

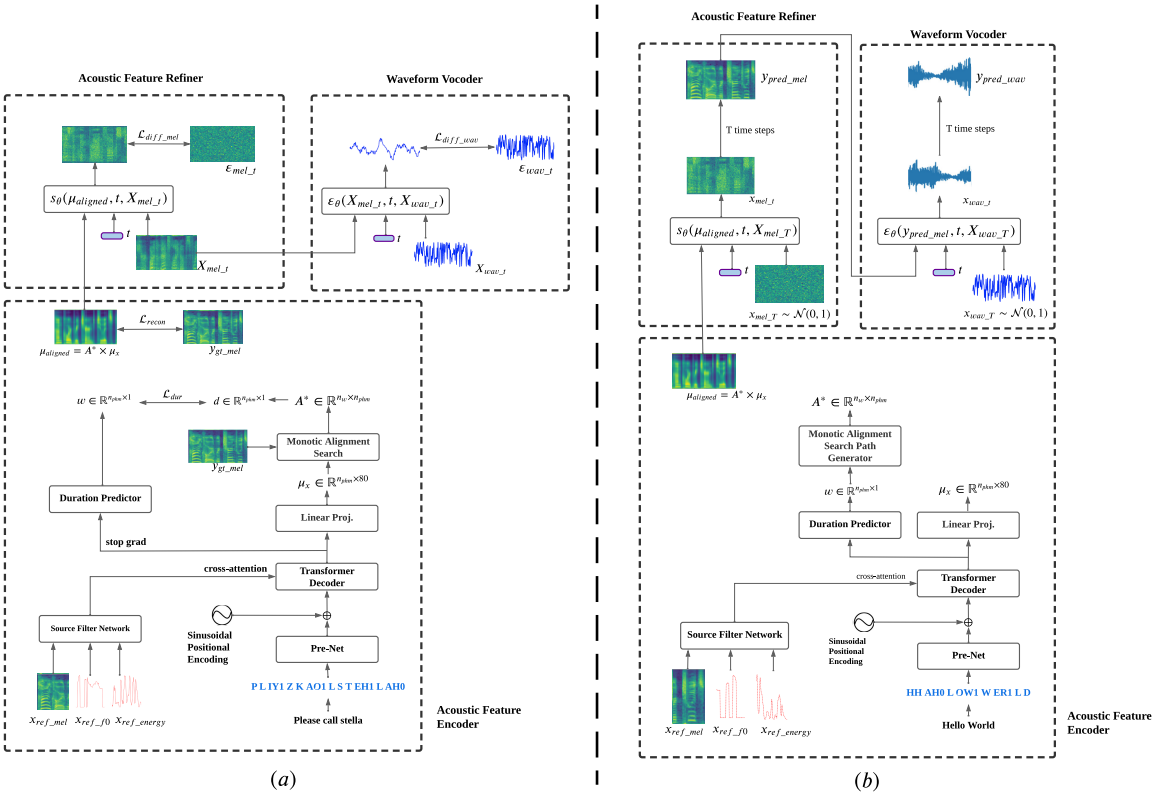

Figure 4.1: **Overall framework of Zero-Voice model:** In the Training Procedure (a), we jointly train three components: the Acoustic Feature Encoder, the Acoustic Feature Refiner, and the Waveform Vocoder. During the Inference Procedure (b), the Acoustic Feature Encoder processes speech prompts and phonemes to generate the aligned acoustic features, denoted as $\mu_{aligned}$. The Acoustic Feature Refiner then refines these features, reducing noise to produce the predicted mel-spectrogram output, y_{pred_mel}. Finally, the Waveform Vocoder converts y_{pred_mel} into the final waveform output, y_{pred_wave}. Additionally, Figure 4.2 provides detailed information on pipeline of the speech prompts processed by the Source Filter Network within the Acoustic Feature Encoder.

4.3.1 Zero-Voice Model Training

4.3.1.1 Acoustic Feature Encoder Training Procedure

The Acoustic Feature Encoder processes input phonemes $x \in \mathbb{R}^{n_{phm} \times 1}$ through a Pre-Net, resulting in phoneme features $x_{phm} \in \mathbb{R}^{n_{phm} \times d_{phm}}$, where n_{phm} denotes the number of phonemes, and d_{phm} represents their dimensional representation. The Source Filter Network processes the speech prompts x_{ref_mel}, x_{ref_pitch}, and x_{ref_energy} which represent the mel-spectrograms, pitch, and energy derived from the waveform, to produce $x_{ref} \in \mathbb{R}^{n_{ref} \times d_{ref}}$, where n_{ref} indicates the time steps and d_{phm} denotes dimension of the speech prompts. The phoneme features x_{phm} and speech prompt features x_{ref} are then integrated within Transformer Decoder module with sinusoidal positional encoding, resulting in a phoneme representation denoted as $\mu_x \in \mathbb{R}^{n_{phm} \times 80}$.

The Monotonic Alignment Search (MAS) algorithm in Eq. 4.1 is utilized to find the most probable monotonic alignment between the phoneme representation μ_x and the statistics of the prior distribution for ground truth mel-spectrograms y_{gt_mel} using Viterbi training [3, 66]. The duration alignment vector $A^* \in \mathbb{R}^{n_w \times n_{phm}}$ is generated from μ_x and y_{gt_mel} using Eq. 4.1, where n_w denotes learned duration length for each input phoneme within μ_x. Addionally, the stop gradient operator $sg(.)$ is applied to x_{phm} without affecting the Source Filter Network, Pre-Net, or Transformer Decoder modules that process the text and speech prompts during training. To obtain aligned acoustic features $\mu_{aligned} \in \mathbb{R}^{n_w \times 80}$, we utilize the duration information vector A^* and perform a matrix multiplication with μ_x according to Eq. 4.2.

$$A^* = MAS(sg(x_{phm}), y_{gt\_mel}) \tag{4.1}$$

$$\mu_{aligned} = A^* \times \mu_x \tag{4.2}$$

During the training of the Acoustic Feature Encoder, we optimize two loss functions: $\mathcal{L}_{recon}$ for the reconstruction loss and $\mathcal{L}_{dur}$ for the duration predictor. The reconstruction loss $\mathcal{L}_{recon}$ in Eq. 4.3 is computed as the mean squared error (MSE) between the aligned acoustic features $\mu_{aligned}$ and the mel-spectrograms of the ground truth speech y_{gt_mel} to reconstruct ground truth speech.

$$\mathcal{L}_{recon} = MSE(\mu_{aligned}, y_{gt\_mel}) \tag{4.3}$$

The loss function $\mathcal{L}_{dur}$ in Eq. 4.5 for the duration predictor is computed as the mean squared error (MSE) between w and d. Here, $w \in \mathbb{R}^{n_w \times 1}$ is the duration information vector generated from the duration predictor, which predicts the duration of each phoneme representation within μ_x, and d represents the duration label calculated from the alignment matrix A^* in the logarithmic domain as described in Eq. 4.4. This procedure aligns the phoneme representations and estimates their respective durations.

$$d = \log \sum_{j=1}^{n_{phm}} \mathbb{1}_{A^*(j)=i}, i = 1...n_{phm} \tag{4.4}$$

$$\mathcal{L}_{dur} = MSE(w, d) \tag{4.5}$$

4.3.1.2 Source Filter Network for the Reference Speaker's Voice

To enhance the model's zero-shot capabilities in low-resource settings, we develop a novel Source Filter Network designed to decouple the reference speaker's voice components. This network can be implemented either through unsupervised decomposition methods [71, 72] or by learning orthogonal subspaces in the latent space [73, 74]. In this study, we explore the unsupervised decomposition approach [71, 72] for its simplified implementation. These methods hypothesize that by breaking speech rhythm and passing the resulting features through an information bottleneck, it is possible to effectively isolate pitch, timbre, and rhythm in an unsupervised manner, thereby removing interference from irrelevant voice elements. As a result, the reference speaker's vocal attributes are efficiently disentangled, enabling more effective extraction of style features .

The pipeline of the Source Filter Network is depicted in Figure 4.2. Specifically, We propose two data augmentation operations. During the training phase of the model: (i) We convert the waveform of a reference speaker into mel-spectrograms, pitch, and energy, resulting in the speech prompts x_{ref_mel}, x_{ref_pitch}, and x_{ref_energy}; (ii) To simulate prosodic variability, we apply a random cut-and-connect operation to

the speech prompts, where time-domain features are randomly truncated and reconnected with a magnitude m_{rcac}; This approach differs from the random resampling technique employed in prior source network designs [71, 72], which introduce corruption by downsampling speech features but fail to adequately diversify reference speech patterns; In contrast, our proposed random cut-and-connect operation enriches the variability of reference speech during training, making it effective in low-resource scenarios. A comparative analysis of both augmentation techniques is presented in the Ablation Study Section 4.5.4.1; (iii) These truncated features are randomly stretched or compressed with a magnitude m_{rsc} to distort their rhythm for unsupervised filtering of speech components, producing $x_{ref\_mel} \in \mathbb{R}^{n_{ref} \times 80}$, $x_{ref\_pitch} \in \mathbb{R}^{n_{ref} \times 1}$, and $x_{ref\_energy} \in \mathbb{R}^{n_{ref} \times 1}$; (iv) The output features are aligned using three encoders (Mel-Style Encoder, Pitch Encoder, and Energy Encoder) to produce $x_{ref\_mel} \in \mathbb{R}^{n_{ref} \times d_{ref}}$, $x_{ref\_pitch} \in \mathbb{R}^{n_{ref} \times d_{ref}}$, and $x_{ref\_energy} \in \mathbb{R}^{n_{ref} \times d_{ref}}$, where d_{ref} represents the aligned dimension of these features. This process achieves the unsupervised decoupling of timbre, pitch, and rhythm features from the reference speaker's voice. Subsequently, element-wise summation of the outputs from the three encoders facilitates multimodal data fusion, and the Transformer Encoder module produces $x_{ref} \in \mathbb{R}^{n_{ref} \times d_{ref}}$, effectively preserving essential reference speech feature information.

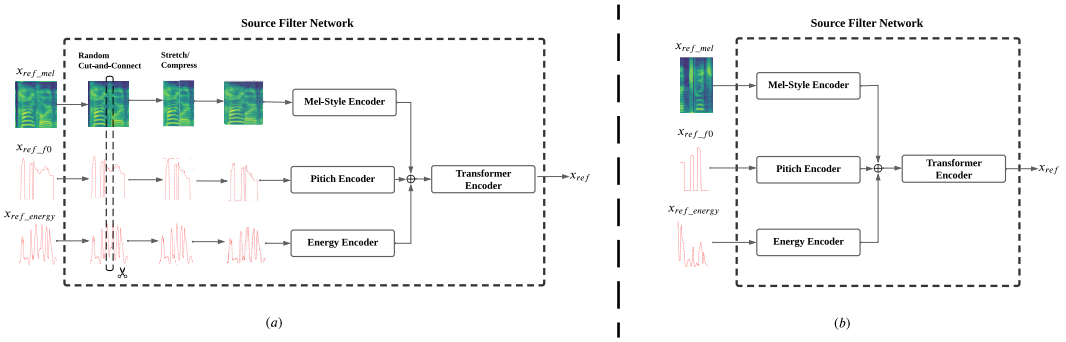

Figure 4.2: **Pipeline of Source Filter Network:** (a) During training, the speech prompts x_{ref_mel}, x_{ref_pitch}, and x_{ref_energy}, which represent the mel-spectrograms, pitch, and energy derived from the waveform, undergo a random cut-and-connect process that truncates their time-domain features. These truncated features are stretched or compressed to distort their rhythm, and are then aligned by three encoders: the Mel-Style Encoder, Pitch Encoder and Energy Encoder. Subsequently, elementwise-summation of the outpts from the three encoders, and the Transformer Encoder modules produce x_{ref}; (b) During inference, the speech prompts x_{ref_mel}, x_{ref_pitch}, and x_{ref_energy} are directly used as input for the three encoders and the Transformer Encoder to produce x_{ref}.

4.3.1.3 Acoustic Feature Refiner Training Procedure

The Acoustic Feature Refiner utilizes a score-based diffusion model to refine the aligned hidden representations $\mu_{aligned}$ produced by the Acoustic Feature Encoder. During training, forward diffusion noise is applied. To implement forward diffusion in the training of the Acoustic Feature Refiner, we integrate a property derived from Grad-TTS [11], which is based on the score-based diffusion model [7]. This integration establishes an n-dimensional stochastic process X_t that satisfies a specified Stochastic Differential Equation (SDE) depicted in Eq. 4.6. This process is designed to transform any data into Gaussian noise, assuming an indefinite time horizon T. The equation incorporates a noise scheduler β_t, a predictor μ, and a diagonal matrix $\sum$ with positive entries at t time step.

$$dX_t = \frac{1}{2}\sum^{-1}(\mu - X_t)\beta_t dt + \sqrt{\beta_t}dW_t, t \in [0, T] \tag{4.6}$$

Solving Eq. 4.6 leads us to Eq. 4.7, which computes X_{mel_t} at time step t using the aligned acoustic features $\mu_{aligned}$ and the initial condition X_{mel_0} (where X_0 is set to the ground truth mel-spectrograms y_{gt_mel}). The term dW_t represents a standard Wiener process equivalent to Gaussian noise $\mathcal{N}(0, 1)$, while a noise scheduler $\beta(t)$ regulates the intensity of noise incorporated into the $\mu_{aligned}$ at time step t.

$$\begin{aligned} X_{mel\_t} &= (I - \beta(t)^{-\frac{1}{2}})\mu_{aligned} + \beta_t^{-\frac{1}{2}} X_{mel\_0} + (I - \beta_t)^{-\frac{1}{2}} dW_t \\ \beta_t &= e^{\sum^{-1} \int_0^t \beta_s d_s}, \beta_s = \beta_{min} + (\beta_{max} - \beta_{min})s \end{aligned} \tag{4.7}$$

We optimize the tractable score estimations using the loss function $\mathcal{L}_{diff\_mel}$ in Eq. 4.8. ε_{mel_t} represents Gaussian noise $\mathcal{N}(0, 1)$ at time step t for mel-spectrogram, and s_θ utilizes $\mu_{aligned}$, t, X_{mel_t} as inputs to estimate this noise, employing a score function s_θ based on a U-net architecture [75].

$$\mathcal{L}_{diff\_mel} = \mathbb{E}_{X_{mel\_0}, t, \varepsilon_{mel\_t}} ||\varepsilon_t - \beta(t)^{-\frac{1}{2}} s_\theta(\mu_{aligned}, t, X_{mel\_t})||_2^2 \tag{4.8}$$

4.3.1.4 Waveform Vocoder Training Procedure

Recently, Chen et al. (Wavegrad 2 [76]) and Lim et al. (JETS) [77] proposed concurrent training of a vocoder (i.e. HIFI-GAN [4]), which assists the model to mitigate

the feature mismatch between acoustic features and waveforms during both training and inference stages. Motivated by this, we concurrently train a diffusion-based Waveform Vocoder alongside the Acoustic Feature Refiner to convert the predicted mel-spectrograms, y_{pred_mel}, into a waveform, y_{pred_wav}. This collaborative approach enables the generation of high-fidelity audio output based on refined acoustic features. We employ DiffWave network [40], a denoised diffusion probabilistic model that supports non-autoregressive conditional waveform generation. DiffWave transforms Gaussian noise and mel-spectrograms into a structured waveform via a Markov chain process with a predetermined number of synthesis steps. It is trained efficiently by optimizing a modified variational bound on the data likelihood. The integration of DiffWave network is simpler to train, avoiding GAN-based vocoders contain both mode collapse and posterior collapse issues [18], and training instability associated with the joint training of two networks proposed in Wavegrad 2 [76] and JETS [77].

The forward process computes X_{wav_t} at timestep t in Eq. 4.9, where X_{wav_0} represents the audio ground truth y_{gt_wav}, and ε_{wav_t} denotes Gaussian noise $\mathcal{N}(0,1)$ at timestep t for the waveform, modulated by a noise schedule β_t.

$$\begin{aligned} X_{wav\_t} &= \sqrt{\bar{\alpha}_t} X_{wav\_0} + \sqrt{1-\bar{\alpha}_t}\varepsilon_{wav\_t} \\ \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s, \alpha_t &= 1-\beta_t, \hat{\beta}_t = \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t \end{aligned} \tag{4.9}$$

During the training of the Waveform Vocoder, we optimize the loss function $\mathcal{L}_{diff\_wav}$ as described Eq. 4.10. ε_θ represents a trainable neural network which utilizes a bidirectional dilated convolution architecture. Specifically, we employ X_{mel_t} (i.e., the noisy mel-spectrograms derived from Eq. 4.7) as a conditional input for ε_θ, ensuring that the output of the Acoustic Feature Refiner y_{pred_mel} closely matches the distribution of X_{mel_t}.

$$\mathcal{L}_{diff\_wav} = \mathbb{E}_{X_{wav\_0},t,\varepsilon_t} ||\varepsilon_{wav\_t} - \varepsilon_\theta(X_{mel\_t}, t, X_{wav\_t})||_2^2 \tag{4.10}$$

Finally, we aim to minimize the total loss $\mathcal{L}_{total}$ as defined in Eq. 4.11 during the training of the Acoustic Feature Encoder, Acoustic Feature Refiner and Waveform Vocoder within our Zero-Voice model. The loss comprises the summation of $\mathcal{L}_{recon}$, $\mathcal{L}_{dur}$, $\mathcal{L}_{diff\_mel}$ and $\mathcal{L}_{diff\_wav}$.

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \mathcal{L}_{dur} + \mathcal{L}_{diff\_mel} + \mathcal{L}_{diff\_wav} \tag{4.11}$$

4.3.2 Zero-Voice Model Inference

4.3.2.1 Acoustic Feature Encoder Inference

During the inference of Acoustic Feature Encoder , the Zero-Voice model directly takes text and the speech prompts as input, generating the hidden representation μ_x. The learned duration information vector w and μ_x are used to sample the duration alignment vector A^* via the Monotonic Alignment Search Path Generator algorithm. Subsequently, the aligned hidden representation $\mu_{aligned}$ is derived through by matrix multiplication between A^* and μ_x by using Eq. 4.2.

4.3.2.2 Acoustic Feature Refiner Inference

For the inference of Acoustic Feature Refiner, reverse diffusion is used to compute the mel-spectrograms y_{pred_mel}. In our application, we develop Eq. 4.12 which is the derivation of Ordinary Differential Equation (ODE) for reverse diffusion. This ODE is to be solved in reverse, commencing from the terminal condition $X_{mel\_T} \sim \mathcal{N}(0, 1)$, taking inputs of $\mu_{aligned}$, t, and β_t, to gradually compute the mel-spectrograms X_{mel_0} (i.e., y_{pred_mel}).

$$X_{mel\_t-1} = X_{mel\_t} - \frac{1}{2}(\mu_{aligned} - X_{mel\_t} - s_\theta(\mu_{aligned}, t, X_{mel\_t}))\beta_t d_t \tag{4.12}$$

4.3.2.3 Waveform Vocoder Inference

For the inference of the Waveform Vocoder, we utilize Eq 4.13 derived from DiffWave [40], which employs Langevin dynamics sampling method. The predicted mel-spectrogram y_{pred_mel} produced by the Acoustic Feature Refiner, is used as the conditional input for the reverse processing. To generate the playable waveform data. it starts from the terminal condition $X_{wave\_T} \sim \mathcal{N}(0, 1)$, this process gradually denoises the X_{wave_T} to produce final waveform output X_{wav_0} (i.e., y_{pred_wav}).

$$X_{wav\_t-1} = \frac{1}{\sqrt{\alpha_t}}(X_{wav\_t} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(y_{pred\_mel}, t, X_{wav\_t})) + \sqrt{\hat{\beta}_t}\varepsilon_{wav\_t} \tag{4.13}$$

4.4 Experimental Settings

We conduct experiments to validate the effectiveness of the proposed Zero-Voice model on both English and Te Reo Māori. In this section, we outline the experimental settings for the datasets, the Zero-Voice model, and the corresponding baselines for both English and Te Reo Māori.

4.4.1 Datasets

The experimentation includes two English TTS datasets (i.e., LibriTTS [41] and VCTK [43]) and a Te Reo Māori dataset.

- **LibriTTS** [41] is a multi-speaker English corpus designed for TTS applications, consisting of approximately 585 hours of read English speech sampled at 24kHz. LibriTTS was derived from the original materials of the LibriSpeech [42] corpus, which comprises roughly 1000 hours of 16kHz read English speech. In our experiments, we train our Zero-Voice model on the **train-clean-100** set as the low-resource setting, which consists of only 54 hours transcribed speech data, and we use the **clean-test** set for testing.

- The **VCTK** Corpus [43] comprises 44 hours of speech data from 110 English speakers with a diverse range of accents. Each speaker was recorded reading approximately on the same 400 sentences, selected from various sources, including newspaper articles, the rainbow passage, and an elicitation paragraph from the speech accent archive. In our experiments, we partition the complete dataset into train and test sets for training and testing of the Zero-Voice model, For the test set, we adhere to the experimental settings of Your-TTS [29], selecting one representative from each accent, resulting in a total of 7 women and 4 men (i.e., speakers p225, p234, p238, p245, p248, p261, p294, p302, p326, p335, and p347).

- **Te Reo Māori**, an official language of New Zealand, has been identified as a gradually diminishing language. The 2022 New Zealand census reported that approximately 892,200 people identified as being part of the Māori ethnic group, representing 17.4% of the New Zealand population. Of these, 67.7% could speak more than a few words or phrases of Te Reo Māori, 22.1% could hold an everyday conversation in Te Reo Māori, and only 6.8% were able to speak Te Reo Māori at

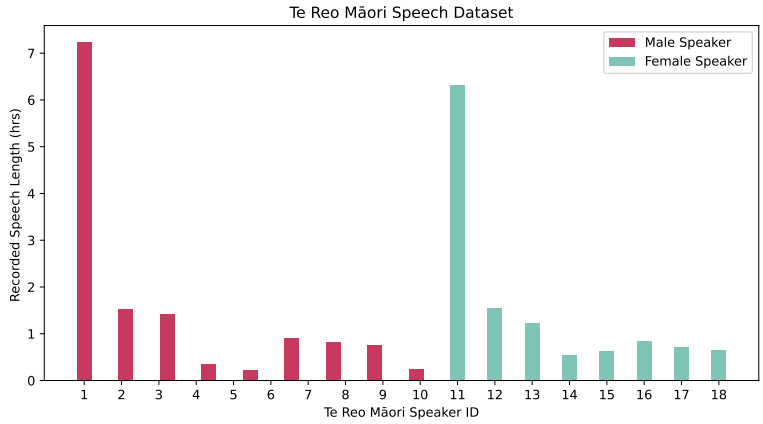

Figure 4.3: Distribution of recorded speech durations across speakers for the Te Reo Māori Speech Dataset.

least fairly well [79]. In this study, we have legitimately collected and labeled 27 hours of read speech data in Te Reo Māori with 44kHz sample rate, comprising 18 different speakers, including 10 males and 8 females native Māori speakers. The distribution of speech durations per speaker is illustrated in Figure 4.3. For our experiments, we divide the dataset into a training set and a validation set with a 90%/10% split, and use test set from the publicly available Google FLEURS Te Reo Māori subset (nz_mi) to evaluate performance in developing a low-resource ASR system.

4.4.2 Implementation Details

For the three datasets used in the experiments, we resample them to 22050 hz. We extract 80-dimensional mel-spectrograms from the the reference waveform. Pitch and Energy contours are extracted from the reference waveform with the same parameters used for computing mel-spectrograms.

For the training of the Zero-Voice model, we use Adam optimiser [80] with a learning rate of 1e-4. We train the model on the LibriTTS, VCTK and Te Reo Māori datasets using the hyperparameter settings detailed in TABLE 4.1. The notations used are as follows:

- FC-XD-YD: Fully Connected Layer with input dimension X and output dimension Y;

Table 4.1: Settings of Hyperparameters

Module	Hyperparameters
Acoustic Feature Encoder	**Source Filter Network:** Random Cut-and-Connect: $m_{rcac} = [0.2, 0.5]$ Random Stretch or Compression: $m_{rsc} = [0.5, 1.5]$ Mel-Style Encoder: FC-80D-256D → Mish → DropOut(0.1) → FC-256D-256D → Mish → DropOut(0.1) → Conv1D-256D-256D → GLU → DropOut(0.1) Conv1D-256D-256D → GLU → DropOut(0.1) Pitch Encoder: Conv1D-1D-256D → ReLU → DropOut(0.1) → Conv1D-256D-256D → ReLU → DropOut(0.1) Energy Encoder: Conv1D-1D-256D → ReLU → DropOut(0.1) → Conv1D-256D-256D → ReLU → DropOut(0.1) Transformer Encoder: [MHA-512D-8H-64D → LN → FFL-2048D] ×6 **Pre-Net:** Phoneme Embedding-512D → [Conv1D-512D-512D → LN → ReLU → DropOut(0.5)] ×3 **Transformer Decoder:** [Masked MHA-512D-8H-64D → LN → MHCA-512D-8H-64D → LN → FFL-2048D] ×6 **Linear Proj:** Conv1D-512D-80D **Duration Predictor:** [Conv1D-512D-512D → ReLU → DropOut(0.1)] ×2 → Conv1D-512D-1D
Acoustic Feature Refiner	U-net-256D-128D-64D Noise Scheduler : $\beta_{min} = 0.05$, $\beta_{max} = 20.0$
Waveform Vocoder	[DiffWave-196D] × 30 Noise Scheduler : $\beta_t = [0.0001...0.05, 50]$
Number of Parameters	176m

- Con1D-XD-YD: 1-D Convolution Layer with input dimension X and output dimension Y;
- LN: Layer Normalization;
- GLU: Gated Linear Unit;
- ReLU: Rectified Linear unit;
- MSA-XD-YH-ZD: Multi-Head Self-Attention with dimension X, Y heads, head dimension Z;
- MSCA-XH-YD-ZD: Multi-Head Cross-Attention with dimension X, Y heads, head dimension Z;
- FFL-XD: Feed Forward Layer with dimension X;

- U-net-XD-YD-ZD: U-net with Downsample Layers of dimensions X, Y, and Z, and Upsample Layers of dimensions Z, Y, and X;
- DiffWave-XD: Bidirectional Dilated Convolution Block with dimension X;
- [...] × X: module [...] with X layers.

4.4.3 English Zero-Shot TTS Baselines

To assess the effectiveness of our proposed approach under low-resource settings, we utilize five recently proposed English zero-shot TTS models as strong baselines for comparison on the two English datasets. Specifically, the pre-trained models below contain public available pre-trained model checkpoints and trained from high-resource speech corpus (i.e., conjointly trained on full LibriTTS train set, VCTK dataset, and others).

- **Ground Truth Audios:** These are the original, unmodified audios.
- **Your-TTS [29]:** This approach is widely acknowledged in modern zero-shot TTS research. It consisted of a transformer encoder, a normalizing flow decoder, and integrated Monotonic Alignment Search as its duration predictor. In our experiments, we utilize its publicly available checkpoint [1] trained jointly on the Libri-TTS and VCTK datasets.
- **Glow-TTS [3] + FreeVC [78] :** Glow-TTS is a flow-based TTS model specifically designed for synthesizing speech in a single speaker's voice, whereas FreeVC is a source-filter speaker Voice Conversion approach. In the experiments, we generate speech using the Glow-TTS model and then adapt it to match the voice style of a reference voice using the FreeVC model. For our experiments, we employ a publicly available checkpoint [2].
- **Meta-StyleSpeech [28]:** Meta-StyleSpeech incorporated Style-Adaptive Layer Normalization (SALN) to align the text inputs and bias with the style extracted from a reference speech audio. Furthermore, a Mel-style encoder was proposed to extract the feature representations of reference speaker's voice. To enhance theadaptability of Meta-StyleSpeech to new speakers, the model includes two

[1] https://github.com/Edresson/YourTTS
[2] https://github.com/coqui-ai/TTS

discriminators trained on style prototypes and undergoes episodic training procedures. In our experiments, we employ the provided open-source checkpoint of the model[3].

- **StyleTTS 2 [36]:** StyleTTS 2 incorporated style diffusion and adversarial training into large speech language models to achieve synthesized voice quality comparable to that of human speech. The system employed large pre-trained speech language models (i.e., WavLM) as discriminators, in conjunction with a novel differentiable duration modeling method, for comprehensive end-to-end training. In our experiments, we utilize the provided open-source checkpoint of the model [4].

- **HierSpeech++ [35]:** HierSpeech++ utilized a flow-based hierarchical speech synthesis framework to markedly enhance both the robustness and expressiveness of synthetic speech. Additionally, this framework incorporated a pre-trained speech representation model to generate self-supervised speech and F0 representations, based on text representations and prosody prompts. Moreover, a highly efficient speech super-resolution framework has been implemented, which upsampled the generaed audio from 16 kHz to 48 kHz. In our experiments, we utilize the provided open-source checkpoint of the model[5].

4.4.4 English Objective Evaluation Metrics for Zero-Shot TTS

We utilize four English objective evaluation metrics, Speaker Embedding Cosine Similarity (SECS), Word Error Rate (WER), and Character Error Rate (CER) metrics, as previously proposed in [13, 29], along with the UTokyo-SaruLab Mean Opinion Score (UMOS) prediction system [83], to automatically assess speech naturalness of generated speech samples.

- **Speaker Encoder Cosine Similarity:** To assess the speech similarity between the synthesized voice and the reference speaker voice, we employ the Speaker Encoder Cosine Similarity (SECS) metric [81]. SECS calculates the cosine similarity between speaker embeddings extracted from the speaker encoder for the two audio samples, which results in a similarity score that ranges

[3] https://github.com/KevinMIN95/StyleSpeech
[4] https://github.com/yl4579/StyleTTS2
[5] https://github.com/sh-lee-prml/HierSpeechpp

from -1 to 1, with a higher value indicating a higher speech similarity. Consistent with previous research [13, 29], we compute SECS using the speaker encoder from the Resemblyzer package[6].

- **Synthesised Audio Accuracy Evaluated by OpenAI Whisper Model:** We use a pre-trained OpenAI Whisper model [82] to transcribe texts from the synthesized audios generated by various baselines and our Zero-Voice model. We then use the transcribed texts to assess their respective Word Error Rate (WER) and Character Error Rate (CER). The model checkpoint used is the large-size v3 model of OpenAI Whisper[7].

- **Speech Naturalness Evaluated by UTokyo-SaruLab Mean Opinion Score Prediction System:** For automated evaluation of speech naturalness, we employ the UTokyo-SaruLab Mean Opinion Score (UMOS) prediction system[8]. The UMOS prediction system [83] is designed to estimate MOS values of audio samples using data collected from previous Blizzard Challenges and Voice Conversion Challenges. It featured two evaluation tracks: a main track for in-domain prediction and an out-of-domain (OOD) track for scenarios with limited labeled data. In the Challenge, the UMOS system achieved the highest score on evaluation metrics for both the main and OOD tracks.

4.4.5 Te Reo Māori Baselines for Fine-Tuning Pre-Trained ASR models

For low-resource languages, a major limitation in developing robust ASR systems is the lack of diverse speaker candidates for recording sufficient and balanced speech data. As shown in Figure 4.3, which illustrates the distribution of speech durations per speaker in our Māori dataset, there is a significant imbalance: two speakers (Speakers 1 and 11) each contribute over 6 hours of recordings, while the remaining speakers provide less than 2 hours each, resulting in a long-tail distribution. This imbalance underscores common challenges in low-resource ASR, including limited speaker diversity, insufficient data volume, and uneven speaker representation, resulting in performance bias [84].

[6] https://github.com/resemble-ai/Resemblyzer
[7] https://github.com/openai/whisper
[8] https://github.com/sarulab-speech/UTMOS22

To address these issues in developing ASR for Te Reo Māori, we first train the Zero-Voice model on the available 27-hour Māori speech training set. We then use the trained Zero-Voice model to synthesize additional speech for fine-tuning pre-trained ASR models. Specifically, we utilize one voice sample (i.e., around 5-10 seconds) per speaker from our 18-speaker Māori dataset as speech prompts to generate balanced synthetic speech across speakers. To further improve speaker diversity, we incorporate voice samples from 10 additional Māori speakers sourced from out-of-domain data, including publicly available Māori read speech and contributions from private Māori speakers. This strategy expands the Te Reo Māori dataset to 360 hours of synthesized speech, substantially augmenting the original 27-hour dataset and addressing the challenges of limited speaker diversity, data scarcity, and imbalance common in low-resource ASR development.

To investigate accuracy improvements for ASR models with synthesized voice samples, we use this synthesized dataset to fine-tune two pre-trained ASR models, OpenAI Whisper Large-v3 and W2V-BERT 2.0 speech encoder. For validation and evaluation, we employ the validation set of our Te Reo Māori dataset and test set of the Google FLEURS Te Reo Māori dataset [48]:

- **OpenAI Whisper Large-v3 [82]:** A multilingual ASR model trained on 1 million hours of weakly labeled audio and an additional 4 million hours of pseudo-labeled audio generated using Whisper Large-v2. It supports over 120 languages, including Te Reo Māori, and comprises approximately 1.5 billion parameters.
- **W2V-BERT 2.0 Speech Encoder [85]:** A Conformer-based speech encoder pre-trained on 4.5 million hours of unlabeled audio spanning more than 143 languages. Notably, Te Reo Māori is not included in the training data, which limits its optimization for Māori ASR tasks in its original form. The model consists of approximately 600 million parameters.

4.5 Results and Discussions

In this section, we show the objective evaluations results by comparing Zero-Voice with the English baseline models based on the English objective evaluation metrics, and examine Te Reo Māori ASR system accuracy improvement by synthesising

speech data from the Zero-Voice model. Additionally, we demonstrate the high interoperability of the Zero-Voice model through visualizations. Further, we perform an Ablation Study on the two English datasets (i.e., LibriTTS and VCTK) to explore the influence of various components on the zero-shot TTS performance of Zero-Voice. Specifically, we examine: The impact of the data augmentation operations on speech prompts within the Source Filter Network; The effects of mel-style, pitch, and energy components on transferring the reference speaker's speech style; The importance of the Acoustic Feature Refiner and the Waveform Vocoder within the Zero-Voice to produce high-fidelity audio; The performance of Zero-Voice when trained on extremely low-resource English datasets.

4.5.1 Objective Evaluation Results

Table 4.2 presents the objective evaluation results for the SECS, UMOS, WER, and CER metrics. For the objective evaluation voice samples, we implement a many-to-many mapping approach, selecting three random voice samples from each speaker in the test sets of both the LibriTTS and VCTK datasets, ensuring synthesized voice samples cover various lengths of reference audios.

HierSpeech++ achieves the best overall performance, obtaining the highest SECS, UMOS, and WER scores on both the LibriTTS and VCTK test sets. Although the metric values of our proposed Zero-Voice model are slightly lower than those of HierSpeech++, it significantly outperforms all remaining baseline approaches on the SECS metric. Considering that HierSpeech++ is trained on high-resource data while Zero-Voice is trained under low-resource conditions, our approach demonstrates strong competitiveness and effectiveness in low-resource scenarios.

Since the evaluation requires a large number of voice samples for comparison, we have open-sourced the Zero-Voice code and pre-trained checkpoints for public review and testing. While Mean Opinion Score (MOS)-based subjective evaluations are common, conducting them with a large number of samples, similar in scale to objective evaluations, can result in unreliable responses due to participant fatigue or boredom. Thus, such subjective evaluation results may not accurately reflect the true performance of the models under test.

Table 4.2: Objective Evaluation based on SECS, WER and CER (↑ shows higher values indicate better performance and ↓ shows lower values indicate better performance)

Approaches	LibriTTS test set				VCTK test set			
	SECS ↑	UMOS ↑	WER (%) ↓	CER (%) ↓	SECS ↑	UMOS ↑	WER (%) ↓	CER (%) ↓
Your-TTS [29]	0.69	3.6	7.1	3.2	0.78	3.6	4.3	2.2
Glow-TTS [3] + FreeVC [78]	0.67	3.3	7.0	3.3	0.76	3.4	6.2	4.3
Meta-StyleSpeech [28]	0.66	3.1	8.3	5.3	0.71	3.3	7.3	5.4
StyleTTS 2 [36]	0.70	3.8	5.7	3.5	0.79	3.9	4.6	2.3
HierSpeech++ [35]	**0.79**	**3.9**	**4.6**	**3.2**	**0.87**	**3.9**	**3.8**	**2.3**
Zero-Voice (ours)	**0.77**	**3.6**	**4.7**	**3.1**	**0.83**	**3.7**	**3.2**	**2.1**
Ground Truth	0.80	4.2	3.1	1.1	0.88	4.0	3.3	1.2

4.5.2 Improved Te Reo Māori Speech Recognition Models via Synthesised Speech Data

We use the original 27 hours of Te Reo Māori speech data and gradually increase the synthesized speech data to 360 hours to fine-tune the OpenAI Whisper Large-v3 model (1.5 billion parameters) and the W2V-BERT 2.0 model (600 million parameters). Results are presented in TABLE 5.1, where we report the Word Error Rate (WER) and Character Error Rate (CER) on both the test set (FLEURS (mi_nz) test set) and the validation set.

For the Whisper Large-v3 model, which includes Te Reo Māori pre-training, the WER and CER values on the test set improve from 37.1% (WER) and 13.76% (CER) without fine-tuning, to 28.3% (WER) and 12.8% (CER) after fine-tuning with the original 27 hours of speech data. As the amount of synthesized speech data used for fine-tuning increases, the performance further improves, reaching 25.4% (WER) and 11.9% (CER) with 60 hours of synthesized data, and 22.9% (WER) and 11.2% (CER) with 360 hours of synthesized data.

For the W2V-BERT 2.0 model, which does not include Te Reo Māori pre-training, we create vocabulary tokens based on the characters of Te Reo Māori. The WER and CER values on the test set improve from 44.0% (WER) and 20.3% (CER) after fine-tuning with the original 27 hours of speech data, to 32.6% (WER) and 13.9% (CER) with 360 hours of synthesized data.

As we note the FLEURS dataset [48] is an n-way parallel speech dataset covering 102 languages, built upon the machine translation benchmark FLores-101 [86]. Each language, including Te Reo Māori, contains approximately 12 hours of speech supervision. Notably, the FLEURS dataset originates from translation tasks rather than naturally occurring Māori speech, resulting in text that reflects a more for-

Table 4.3: Speech Recognition Results Based on WER and CER with Fine-Tuning OpenAI Whisper Large-v3 and W2V-BERT 2.0 Models (↑ shows higher values indicate better performance and ↓ shows lower values indicate better performance)

Approaches	OpenAI Whisper Large-v3				W2V2.0-BERT			
	FLEURS (mi_nz) Test Set		Validation Set		FLEURS (mi_nz) Test Set		Validation Set	
	WER (%) ↓	CER (%) ↓	WER (%) ↓	CER (%) ↓	WER (%) ↓	CER (%) ↓	WER (%) ↓	CER (%) ↓
w/o fine-tuning	37.9	13.8	45.4	19.6	/	/	/	/
27hr Org data	28.3	12.8	19.1	6.9	44.0	20.3	40.7	14.3
+ 60hr Syn Data	25.4	11.9	18.0	6.6	41.3	18.3	37.9	13.8
+ 190hr Syn Data	24.1	11.5	19.2	7.2	37.2	15.6	29.1	9.4
+ 250hr Syn Data	24.6	11.3	19.9	7.1	33.6	14.2	21.7	7.4
+ 360hr Syn Data	**22.9**	**11.2**	**16.5**	**6.2**	**32.6**	**13.9**	**20.3**	**7.2**

mal, structured, and potentially anglicized form of the language. It often includes code-switching and numeric expressions, which diverge from the informal and conversational patterns of natural Māori speech. Despite these limitations, the Māori portion of the FLEURS dataset remains the only publicly available benchmark for evaluating Māori ASR. Encouragingly, our validation set results align closely with those obtained from the FLEURS test set, indicating consistent model performance across different evaluation splits.

4.5.3 Model Interoperability via Visualization

In this part, we illustrate the functionalities of the Acoustic Feature Encoder, Acoustic Feature Refiner, and Waveform Vocoder within the Zero-Voice model. As shown in Figure 4.4, the aligned acoustic features $\mu_{aligned}$ from the Acoustic Feature Encoder, the mel-spectrogram output y_{pred_mel} from the Acoustic Feature Refiner, and the ground truth mel-spectrogram y_{gt_mel} are presented alongside their corresponding waveform representations.

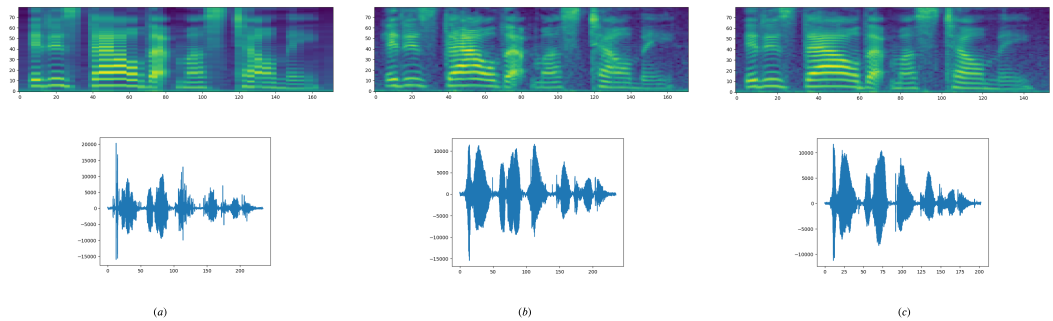

Figure 4.4: Interoperability of Zero-Voice model: (a) the output of the Acoustic Feature Encoder $\mu_{aligned}$ and its waveform; (b) the output of the Acoustic Feature Refiner y_{pred_mel} and its waveform; (c) Ground Truth mel-spectrogram y_{gt_mel} and its waveform.

4.5.3.1 No Latent State

From Figure 4.4, we observe that the intermediate representations within the Zero-Voice model modules generated by the the Acoustic Feature Encoder, Acoustic Feature Refiner, and Waveform Vocoder demonstrate high interoperability and strong alignment with the ground truth. This also indicates that the Zero-Voice framework avoids information loss and does not rely on latent compression states, unlike prior diffusion-based zero-shot TTS models, i.e., as StyleTTS 2 [36] and NaturalSpeech 2 [34].

Compared to the previous approaches using GAN-based vocoder, i.e., Meta-StyleSpeech [28], Your-TTS [29], HierSpeech++ [35], and StyleTTS 2 [36], require training additional neural networks and incorporating multiple loss terms, our diffusion-based Acoustic Feature Refiner and Waveform Vocoder each require only a single loss term during training, simplifying the optimization process while maintaining high synthesis quality. Despite this simplified training process, the proposed modules support finer granularity and richer feature decoding during inference, which contributes to enhanced synthesis fidelity and similarity in zero-shot speech generation.

4.5.3.2 The Importance of $\mu_{aligned}$ for Speaker's Voice Style Transfer

Since the visual outputs of $\mu_{aligned}$, y_{pred_mel}, and y_{gt_mel} exhibit highly similar feature patterns, we infer that the Acoustic Feature Encoder successfully integrates the reference speaker's voice style with the input text features. This effective integration is achieved as the outputs from the Source Filter Network are adequately fused within the Transformer Decoder via cross-attention mechanisms.

4.5.3.3 The Importance of y_{pred_mel} Produced by the Acoustic Feature Refiner

Comparing the output of the Acoustic Feature Refiner y_{pred_mel} and its corresponding waveform, we observe that the initial waveform derived from $\mu_{aligned}$ exhibits a subdued and less expressive voice. The Acoustic Feature Refiner transforms $\mu_{aligned}$ into a fine-grained mel-spectrogram y_{pred_mel}, which closely approximates the ground truth mel-spectrogram y_{gt_mel}. This indicates that the refiner effectively enhances the acoustic features without introducing information compression or loss. As a result, the refined mel-spectrogram y_{pred_mel} is used as a strong prior for generating the

high-fidelity waveform output y_{pred_wav}.

4.5.4 Ablation Studies

In this section, we conduct ablation studies to investigate three key aspects aimed at enhancing the Zero-Voice model: the impact of different speech prompt augmentation operations on the Source Filter Network; the contributions of mel-style, pitch, and energy features to voice style transfer in the Source Filter Network; and the improvement of synthesized voice quality through conjoint training of the Acoustic Feature Refiner and Waveform Vocoder. Furthermore, we examine the minimal dataset size required to effectively train the Zero-Voice model in low-resource settings.

4.5.4.1 Ablation: Speech Prompt Augmentation Operations in The Source Filter Network

We establish six distinct experimental settings, as detailed in TABLE 4.4, to evaluate the impact of various speech prompt augmentation strategies within the Source Filter Network on the performance metrics of the Zero-Voice model: (i) Direct training with speech prompts: Training directly on mel-spectrogram, pitch, and energy features leads to overfitting, which degrades the clarity and accuracy of synthesized speech. Although this setting yields a high SECS score, it results in significantly higher UMOS, WER, and CER values; (ii) Stretching or compressing speech prompts: Applying temporal modifications alone fails to address the overfitting issue effectively; (iii) Random cut-and-connect operation: This operation substantially reduces overfitting, resulting in high SECS scores while significantly lowering WER and CER, demonstrating its effectiveness; (iv) Combining random cut-and-connect with stretching or compressing: This hybrid approach further enhances performance across all evaluation metrics (SECS, UMOS, WER, and CER), resulting in complementary effects; (v) Random resampling operation: As used in previous Source Filter Network studies [71, 72], this operation yields inferior performance. Its limited capacity to introduce diverse reference speech patterns makes it less effective than our proposed random cut-and-connect method under the zero-shot TTS scenario; (vi) Combination of random resampling with stretching or compressing: This variant shows slight improvements over random resampling alone but still falls short of the performance achieved by the random cut-and-connect operation.

In light of these findings, the results indicate that the simple random cut-and-connect operation within the Source Filter Network yields the most significant improvement in both zero-shot performance and synthesis quality.

Table 4.4: Ablation of the Source Filter Network (Random CAC indicates the Random Cut-and-Connect operation, RR indicates the Random Resampling operation, S/C indicates the Stretch or Compress operation, ✓indicates the inclusion of the component, ↑ indicates higher values are better, and ↓ indicates lower values are better).

Zero-Voice			LibriTTS test set				VCTK test set			
Random CAC	S/C	RR	SECS ↑	UMOS ↑	WER (%) ↓	CER (%) ↓	SECS ↑	UMOS ↑	WER (%) ↓	CER (%) ↓
			0.79	1.5	84.2	74.3	**0.89**	1.5	73.6	62.1
	✓		0.78	1.2	74.6	54.6	0.84	1.4	77.2	64.3
✓			0.73	3.2	6.9	4.3	0.81	3.4	4.3	2.5
✓	✓		**0.77**	**3.5**	**4.7**	**3.1**	**0.83**	**3.7**	**3.2**	**2.1**
		✓	0.70	2.2	10.6	8.5	0.78	2.2	7.3	4.6
	✓	✓	0.71	2.5	7.6	6.6	0.79	2.4	7.2	5.3

4.5.4.2 Ablation : Mel-Style, Pitch and Energy Components in The Source Filter Network for Voice Style Transfer

Given the importance of prior voice style features in enhancing the zero-shot capability of the model, as discussed in Section 4.5.3.2, we investigate the impact of ablating the mel-style, pitch, and energy feature components in the Source Filter Network on the speech synthesized quality and similarity between the synthesized voice and the original reference speaker's voice. As shown in Table 4.5, the Source Filter Network containing mel-style, pitch, and energy components yields the best overall performance on objective evaluation metrics (i.e., SECS, UMOS, WER, and CER). Notably, the mel-style component is crucial for transferring the reference speaker's voice style and maintaining the naturalness of synthesized voice, while the pitch and energy components assist in zero-shot transformation. Without the mel-style component, overall performance significantly degrades.

Table 4.5: Ablation of Mel-Style. Pitch and Energy Components (✓indicates the inclusion of the component, ↑ indicates higher values are better and ↓ indicates lower values are better)

Zero-Voice			LibriTTS test set				VCTK test set			
Mel-Style	Pitch	Energy	SECS ↑	UMOS ↑	WER (%) ↓	CER (%) ↓	SECS ↑	UMOS ↑	WER (%) ↓	CER (%) ↓
✓	✓	✓	**0.77**	**3.6**	**4.7**	**3.1**	**0.83**	**3.7**	**3.2**	**2.1**
✓	✓		0.74	3.5	6.6	3.7	0.81	3.4	5.1	4.3
	✓	✓	0.68	3.0	10.7	8.6	0.75	3.1	9.2	7.1
✓		✓	0.74	3.4	4.6	3.6	0.80	3.5	5.2	3.3
✓			0.73	3.5	6.6	4.1	0.80	3.4	4.2	2.2
	✓		0.68	3.0	14.2	10.3	0.73	3.1	15.6	12.1
		✓	0.69	2.9	84.2	74.3	0.72	3.2	14.3	11.3

4.5.4.3 Ablation: The Acoustic Feature Refiner and Waveform Vocoder

Zero-Voice consists of the Acoustic Feature Encoder, the Acoustic Feature Refiner, and the Waveform Vocoder, which are trained in an end-to-end manner to achieve optimal performance in zero-shot TTS. In this section, we assess the effectiveness of jointly training the diffusion-based Acoustic Feature Refiner and the Waveform Vocoder. We conduct an ablation study on these components, with results presented in Table 4.6: (i) Joint training of the Acoustic Feature Refiner and Waveform Vocoder within Zero-Voice: this setting yields the best overall performance across all evaluation metrics, including SECS, UMOS, WER, and CER; (ii) Removing the Waveform Vocoder: In this setting, the model is trained to produce only mel-spectrogram features. We then utilize a pre-trained HiFi-GAN [4] [9] to convert the mel-spectrograms into waveforms. This configuration leads to a noticeable drop in performance, primarily due to the feature mismatch between the independently trained HiFi-GAN and the rest of the Zero-Voice architecture. This mismatch disrupts end-to-end coherence and limits the effectiveness of waveform reconstruction. (iii) Removing the Acoustic Feature Refiner: Excluding the Acoustic Feature Refiner from the model results in the poorest performance across all objective evaluation metrics. This highlights the critical function of refined acoustic features as intermediate representations in enhancing mel-spectrogram quality and overall speech synthesis performance as we discussed in Section 4.5.3.3.

Table 4.6: Ablation of Acoustic Feature Refiner and Waveform Vocoder (AFR indicates Acoustic Feature Refiner, WV indicates Waveform Vocoder, ✓indicates the inclusion of the component, ↑ indicates higher values are better and ↓ indicates lower values are better)

Zero-Voice		**LibriTTS test set**				**VCTK test set**			
AFR	WV	**SECS ↑**	**UMOS ↑**	**WER (%) ↓**	**CER (%) ↓**	**SECS ↑**	**UMOS ↑**	**WER (%) ↓**	**CER (%) ↓**
✓	✓	**0.77**	**3.5**	**4.7**	**3.1**	**0.83**	**3.7**	**3.2**	**2.1**
✓		0.76	2.9	10.4	7.3	0.82	2.8	11.3	8.5
	✓	0.72	2.2	14.6	9.6	0.80	2.4	12.2	10.3

4.5.4.4 Ablation: Extremely Low-resource Setting

In this section, we evaluate the performance threshold of our Zero-Voice model under extremely low-resource conditions, using significantly reduced dataset sizes and speaker counts. To simulate these settings, we randomly shuffle the LibriTTS and

[9] https://github.com/jik876/hifi-gan

VCTK training sets and extract subsets comprising one-half, one-quarter, and one-eighth of the original data. Consequently, both the dataset size and the number of speakers are substantially reduced. As shown in Table 4.7, Zero-Voice demonstrates strong generalization and robustness, even when trained on only one-eighth of the original datasets (approximately 5–8 hours of transcribed speech). While some degradation in speech naturalness and accuracy is observed under these constraints, the model still performs competitively on both English test sets.

Table 4.7: Ablation of Datasets Size (↑ indicates higher values are better and ↓ indicates lower values are better)

Zero-Voice	**LibriTTS test set**				**VCTK test set**			
Dataset Size	**SECS ↑**	**UMOS ↑**	**WER (%) ↓**	**CER (%) ↓**	**SECS ↑**	**UMOS ↑**	**WER (%) ↓**	**CER (%) ↓**
full	**0.77**	**3.6**	**4.7**	**3.1**	**0.83**	**3.7**	**3.2**	**2.1**
1/2	0.75	3.4	6.8	3.7	0.82	3.4	4.4	2.3
1/4	0.75	3.2	6.9	4.0	0.81	3.2	5.3	3.1
1/8	0.74	3.1	8.8	4.1	0.81	3.1	5.6	3.7

4.6 Conclusion

In this chapter, we present Zero-Voice, a novel zero-shot text-to-speech (TTS) framework designed for low-resource settings. Zero-Voice integrates four core components: a Source Filter Network (to extract style features from the speech prompt), an Acoustic Feature Encoder (to fuse text and style features), an Acoustic Feature Refiner (to enhance acoustic representations), and a Waveform Vocoder (to decode refined acoustic features into waveform). The entire system is jointly optimized using two separate diffusion models within an end-to-end training and inference pipeline. This architecture enables effective style transfer from a reference speaker to synthesized speech while producing high-fidelity audio under low-resource settings.

We evaluate Zero-Voice on the LibriTTS (train-clean-100 partition) and VCTK datasets using standard objective metrics. Experimental results show that our proposed random cut-and-connect augmentation for reference speech, combined with jointly training the Acoustic Feature Refiner and Waveform Vocoder via diffusion models, leads to a robust low-resource zero-shot TTS system. Zero-Voice significantly outperforms existing strong zero-shot model baselines, including Your-TTS [29] and Meta-StyleSpeech [28], and even demonstrates competitive performance with high-resource models such as HierSpeech++ [35] and StyleTTS 2 [36].

Importantly, Zero-Voice exhibits strong generalization, even when trained on limited

speaker diversity and small datasets (as little as 5–8 hours of transcribed speech). Furthermore, we demonstrate that Zero-Voice can be used as an effective data generator to improve accuracy in low-resource ASR systems. By synthesizing large volumes of high-fidelity speech-text pairs, our method enables effective adaptation of large pre-trained ASR models to Te Reo Māori, achieving recognition accuracy on par with human performance.

Zero-Voice's hierarchical framework with diffusion models can also be extended to other types of generative models, such as flow-based models [58–60] and auto-regressive models [38], showing a promising direction for future research and exploration.

References

[1] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao and T. Liu, "FastSpeech: Fast, Robust and Controllable Text to Speech," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.

[2] Y. Ren, C. Hu, X. Tan, T. Qin, S/ Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and highquality end-to-end text to speech," in *Proc. ICLR*, 2021.

[3] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 8067-8077.

[4] J. Kong, J. Kim, and J. Bae, "Hifigan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 17022–17033

[5] K., Kundan, K. Rithesh, D. Boissiere et al., "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol 32.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020.

[7] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. ICLR*, 2021.

[8] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol 34, pp. 8780–879.

[9] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou and Dong Yu, "Diff-

sound: Discrete Diffusion Model for Text-to-Sound Generation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1720-1733, 2023.

[10] M. Jeong, H. Kim, S. Cheon, B. Choi and N. Kim, "Diff-tts: A denoising diffusion model for text-to-speech," in *Proc. Interspeech*, 2021, pp. 3605–3609.

[11] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova and M. A. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *Proc. ICML*, 2021.

[12] H. Kim, S. Kim, and S. Yoon, "Guided-tts: A diffusion model for text-to-speech via classifier guidance," in *Proc. ICML*, 2022, pp. 11119–11133.

[13] M. Kang, D. Min and S. Hwang, "Grad-StyleSpeech: Any-speaker Adaptive Text-to-Speech Synthesis with Diffusion Models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023.

[14] N. Ellinas, M. Christidou, A. Vioni, J. Sig Sung, A. Chalamandaris, P. Tsiakoulis, and P. Mastorocostas, "Controllable speech synthesis by learning discrete phoneme-level prosodic representations," in *Speech Communication*, vol. 146, 2023.

[15] Y. Wang et,al., "Tacotron: Towards End-to-End Speech Synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[16] J. Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2018, pp. 4779-4783, doi: 10.1109/ICASSP.2018.8461368.

[17] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," in *Proc. IISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza B. Xu, D. Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 139–144.

[19] D. J. Rezende and S. Mohamed "Variational inference with normalizing flows," in *Proc. ICML*, 2015, pp. 1530–1538.

[20] J. Kim, J. Kong and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, 2021, pp. 5530–5540.

[21] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao and T. Qin, "Multispeech: Multi-speaker text to speech with transformer," in *Proc. Interspeech*, 2020, pp. 4024–4028.

[22] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen,

R. Pang, I. Moreno and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4485–4495.

[23] H. -T. Luong and J. Yamagishi, "NAUTILUS: A Versatile Voice Cloning System," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2967-2981, 2020.

[24] Y. Chen, Y. M. Assael, B, Shillingford, D. Budden, S. E. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie, Ç. Gülçehre, A. Oord, O. Vinyals and N.Freitas, "Sample efficient adaptive text-to-speech," in *Proc. ICLR* 2019.

[25] S. Arik, J. Chen, K. Peng, W. Ping and Yanqi Zhou, "Neural voice cloning with a few samples," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10040–10050.

[26] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao and T. Liu, "Adaspeech: Adaptive text to speech for custom voice," in *Proc. ICLR* 2021.

[27] Y. Wu, X. Tan, B. Li, L. He, S. Zhao, R. Song, T. Qin and T. Liu, "AdaSpeech 4: Adaptive Text to Speech in Zero-Shot Scenarios," in *Proc. Interspeech*, 2022, pp.2568-2572.

[28] D. Min, D. Lee, E. Yang and S. Hwang, "Meta-stylespeech : Multi-speaker adaptive text-to-speech generation," in *Proc. ICML*, 2021.

[29] E. Casanova, J. Weber, C. Shulby, A. Júnior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multispeaker TTS and zero-shot voice conversion for everyone," in *Proc. ICML*, 2022.

[30] Z. Liu, Y. Guo and K. Yu, "DiffVoice: Text-to-Speech with Latent Diffusion" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp.1-5.

[31] S. Chen et al., "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 705-718, 2025.

[32] Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi, "Soundstorm: Efficient parallel audio generation," *arXiv preprint arXiv:2305.09636*, 2023.

[33] R. Huang, C. Zhang, Y. Wang, D. Yang, L. Liu, Z. Ye, Z. Jiang, C.Weng, Z. Zhao, and D. Yu, "Make-a-voice: Unified voice synthesis with discrete representation," *arXiv preprint arXiv:2305.19269*, 2023.

[34] K. Shen, Z. Ju, X. Tan, Y. liu, Y. Leng, L. He, T. Qin, S. Zhao and J. Bian, "NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers," *arXiv preprint arXiv:2304.09116*, 2023.

[35] S. Lee, H. Choi, S. Kim and Seong-Whan Lee, "HierSpeech++: Bridging the Gap between Semantic and Acoustic Representation of Speech by Hierarchical Variational Inference for Zero-shot Speech Synthesis," *arXiv preprint arXiv:2311.12454*, 2023.

[36] Y. Li, C. Han, V. Raghavan, G. Mischler and Nima Mesgaran, "StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, vol 36, pp.19594-19621.

[37] M. David, G. Simons and C. Fennig (eds.). 2024. Ethnologue: Languages of the World. Twenty-seventh edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, G. Llion, A. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is All you Need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5998–6008, 2017.

[39] T. Saeki, S. Maiti, X. Li, S. Watanabe, S. Takamichi and H. Saruwatari, "Text-Inductive Graphone-Based Language Adaptation for Low-Resource Speech Synthesis," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 32, pp. 1829-1844, 2024.

[40] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *Proc. ICLR*, 2020.

[41] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from librispeech for text-to-speech," in *Proc. Interspeech*, 2019, pp. 1526–1530.

[42] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2015, pp.5206-5210,

[43] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," University of Edinburgh. The Centre for Speech Technology Research (CSTR) (2019): 271-350.

[44] L. Besacier, E. Barnard, A. Karpov and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," in *Speech Communication*, vol. 56, 2014.

[45] E. Casanova, C. Shulby, A. Korolev, A. Junior, A. Soares, S. Aluísio and M. Ponti, "ASR data augmentation in low-resource settings using cross-lingual multi-speaker TTS and cross-lingual voice conversion," in *Proc. Interspeech*, 2023, pp.

1244–1248.

[46] M .Bartelds, N. San, B. McDonnell, D. Jurafsky, Dan and M. Wieling, "Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation," in *Proc. the Association for Computational Linguistics (ACL)*, 2023, vol 1: Long Papers, pp. 715-729.

[47] K. Fatehi, M. Torres and A. Kucukyilmaz, "An overview of high-resource automatic speech recognition methods and their empirical evaluation in low-resource environments," in *Speech Communication*, vol. 167, 2025.

[48] A. Conneau, M. Ma, S. Khanuja, Yu Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera and A. Bapna, "FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech," in *IEEE Spoken Language Technology Workshop (SLT)*, 2022.

[49] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp 2962–2970.

[50] W. Ping, K. Peng, A. Gibiansky, S. Arik, A. Kannan, Sharan Narang, J. Raiman and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *Proc. ICLR*, 2018.

[51] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proc. ICML*, 2018, pp 4700–4709.

[52] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, 2018, pp 5167–5176.

[53] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez-Moreno and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp 4485–4495.

[54] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 6184–6188.

[55] N. Li, S. Liu, Y. Liu, S. Zhao and M. Liu, "Neural speech synthesis with transformer network," In *Proc. AAAI*, 2019, pp. 6706–6713.

[56] V. Oord et al., "Parallel WaveNet: Fast High-Fidelity Speech Synthesis," in *Proc. ICML*, 2018, pp. 3918–3926.

[57] C. Miao, Q. Zhu, M. Chen, J. Ma, S. Wang and J. Xiao, "EfficientTTS 2: Variational End-to-End Text-to-Speech Synthesis and Voice Conversion," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1650-1661, 2024.

[58] C. Durkan, A. Bekasov, I. Murray and G. Papamakarios, "Neural spline flows," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp 7509–7520.

[59] E. Hoogeboom, R. Berg and M. Welling. "Emerging convolutions for generative normalizing flows," in *Proc. ICML*, 2019 , pp. 2771–2780.

[60] J. Serrà, S. Pascual, and C. Perales. "Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 6790–6800.

[61] L. Dinh, J. Sohl-Dickstein and S. Bengio. "Density estimation using real NVP," in *Proc. ICLR*, 2017.

[62] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10236–10245.

[63] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 3617–3621.

[64] S. Kim, S. Lee, J. Song, J. Kim and S. Yoon. "Flowavenet: A generative flow for raw audio," in *Proc. ICML*, 3370–3378, 2019

[65] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He, S. Zhao, T. Qin, F. Soong and T. Liu, "NaturalSpeech: End-to-End Text-to-Speech Synthesis With Human-Level Quality" in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 06, pp. 4234-4245, 2024.

[66] L. R Rabiner. "A tutorial on hidden markov models and selected applications in speech recognition," in *Proc. of the IEEE* 1989, vol 77(2) pp.257–286.

[67] E. Casanova, C. Shulby, E. Gölge, N. Müller, Frederico S. Oliveira, A. Junior, A. Soares, S. Aluisio and M. Ponti, "SC-GlowTTS: an Efficient Zero-Shot Multi-Speaker Text-To-Speech Model", in *Proc. Interspeech*, 2021, pp.3645-3649.

[68] J. Kim, J. Kong, and J. Son," Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech", in *Proc. ICML*, 2021.

[69] N. Chen, Y. Zhang, H. Zen, R. J.Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," in *Proc. ICLR*, 2020.

[70] S. Lee, H. Kim, C. Shin, X. Tan, C. Liu, Q. Meng, T. Qin, W. Chen, S. Yoon and Tie-Yan Liu, "PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior," in *Proc. ICLR*, 2022.

[71] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised speech decomposition via triple information bottleneck," *Proc. ICML*, 2020.

[72] Z. Luo, S. Lin, R. Liu, J. Baba, Y. Yoshikawa, and H. Ishiguro, "Decoupling speaker-independent emotions for voice conversion via source-filter networks," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 11–24, 2023.

[73] S. Sadok, S. Leglaive, L. Girin, X. Alameda-Pineda, and R. Séguier, "Learning and controlling the source-filter representation of speech with a variational autoencoder," in Speech Communication, vol. 148, 2023.

[74] R. Liu, K. Liang, D. Hu, T. Li, D. Yang and H. Li, "Noise Robust Cross-Speaker Emotion Transfer in TTS Through Knowledge Distillation and Orthogonal Constraint," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 812-827, 2025.

[75] O. Ronneberger, Philipp Fischer and F. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[76] N. Chen, Y. Zhang, H. Zen, R. Weiss, M. Norouzi, N. Dehak and W. Chan, "WaveGrad 2: Iterative Refinement for Text-to-Speech Synthesis," in *Proc. Interspeech*, 2021, pp. 2958-1796.

[77] D. Lim, S. Jung and Eesung Kim, "JETS: Jointly Training FastSpeech2 and HiFi-GAN for End to End Text to Speech," in *Proc. Interspeech*, 2021, pp. 2958-1796.

[78] J. Li, W. Tu and L. Xiao, "Freevc: Towards High-Quality Text-Free One-Shot Voice Conversion," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1-5.

[79] "Te reo Māori proficiency and support continues to grow https://www.stats.govt.nz/news/te-reo-maori-proficiency-and-support-continues-to-grow/," Stats NZ, 2022.

[80] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. ICLR* 2015.

[81] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen,and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*,

2020, pp. 6184–6188.

[82] A. Radford, J. Kim, T. Xu, G. Brockman, C. McLeavey and Ilya Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *ICML*, pp. 28492-28518, 2023.

[83] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 24521–452, doi: 10.21437/Interspeech.2022-439.

[84] Z. Yang, M. Ding, T. Huang, Y. Cen, J. Song, B. Xu, Y. Dong and J. Tang, "Does Negative Sampling Matter? a Review With Insights Into its Theory and Applications," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5692-5711, 2024.

[85] Y.A. Chung, Y. Zhang, W. Han, C. Chiu, J. Qin, R. Pang and Y. Wu, "w2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training," *in IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 244-250,

[86] N. Goyal, C. Gao, V. Chaudhary, P. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan. 2022, "The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation," *Transactions of the Association for Computational Linguistics*, 2022, vol 10 pp 522–538.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Zhihan Wang
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work:	Chapter 5

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Zhihan Wang, Feng Hou, Wang, Ruili Wang: Synthesized Data Selection via Score Distribution Matching for Te Reo Māori Automatic Speech Recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2026, pp. 18172-18176, doi: 10.1109/ICASSP55912.2026.11461415

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Zhihan Wang Digitally signed by Zhihan Wang Date: 2026.05.17 10:50:45 +12'00'
Date:	17-May-2026
Primary Supervisor's Signature:	Prof. Ruili Wang Digitally signed by Prof. Ruili Wang DN: cn=Prof. Ruili Wang, o=Massey University, ou=School of Mathematical and Computational Sciences, email=ruili.wang@massey.ac.nz Date: 2026.05.25 10:17:02 +12'00'
Date:	25-May-2026

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

Chapter 5

Synthesized Data Selection via Score Distribution Matching for Te Reo Māori Automatic Speech Recognition

Fine-tuning pre-trained Automatic Speech Recognition (ASR) models with synthesized datasets leads to suboptimal performance in low-resource ASR, due to domain mismatches between true and synthesized speech data. To address this issue, we propose a novel score-distribution-matching approach for data selection, which filters synthesized data by explicitly aligning its score distribution with the prior distribution of the true data score. We apply our proposed approach on the ASR of Te Reo Māori (the Māori langauge), an official yet endangered language of New Zealand. We legitimately collect and label 27 hours of Te Reo Māori speech data and synthesize an additional 520 hours using the recently developed Zero-Voice model, a zero-shot text-to-speech (TTS) system. Using both true and filtered synthesized data of Te Reo Māori by our proposed approach, we fine-tune the pre-trained ASR model `Whisper-large-v3`*. This results in a remarkable accuracy improvement for low-resource ASR on the FLEURS Māori (mi_nz) benchmark outperforming baseline models.*

5.1 Introduction

Data augmentation with synthesized speech, particularly through zero-shot text-to-speech (TTS) methods (e.g., [1–4] and including our recently proposed Zero-Voice approach [5]) has emerged as an effective strategy for improving low-resource automatic speech recognition (ASR) systems [6, 7]. This improvement is achieved by fine-tuning pre-trained ASR models (e.g., [8–11]) using synthesized speech data [12–15]. However, while the zero-shot TTS models generate synthesized speech with accurate textual transcriptions, the corresponding audio often exhibits a domain mismatch compared to true speech data, since synthesized speech often contains unnatural prosody or artifacts. This issue has been recently investigated in the context of English ASR [16] and children's ASR [17–21], where merely accumulating synthesized speech data for fine-tuning pre-trained ASR models achieves diminishing returns, reaching an upper bound in accuracy improvements.

As a pilot study for low-resource languages, and building on our recent work Zero-Voice [5] (a zero-shot TTS method for data augmentation to enhance Te Reo Māori ASR), we address this issue to further optimize ASR performance for Te Reo Māori (an official yet endangered language of New Zealand). We propose a synthesized data selection method based on *score-distribution-matching*, which explicitly aligns the score distributions of true and synthesized data using character error rate (CER) as the scoring metric. This alignment enables the selection of an optimal subset of synthesized data to improve ASR accuracy in low-resource settings. Experimental results show that our method outperforms baselines including: (i) Synt++ by Hu et al. [16], and (ii) the method by Wang et al. [17], both of which rely on resources developed for high-resource languages, such as pre-trained speaker embeddings or discriminator models trained on English speaker features, to be used as data selection metrics, which limits their applicability to low-resource scenarios. (iii) Adapter Double-way Fine-tuning [18], which is a adapter-based method for this issue.

5.2 Our Method

We propose to select a subset of synthesized data that closely aligns with the score feature distribution of the true dataset, leveraging the true data's distribution as a strong prior to direct the selection process. Through the selection process based on

this alignment, we aim to construct a synthesized data subset that captures a diverse range of speech samples, spanning both low- and high-quality speech examples to enhance the performance of low-resource ASR models and achieve optimal results. Notably, our method rely on minimal pre-trained resources for low-resource languages compared with [16, 17]. In addition, the previous synthesized data selection methods for Children's ASR [17–22] have not explored the impact of varying speech quality on ASR model performance. Our method addresses this gap by explicitly considering the quality variation in synthesized data and leveraging it to improve the generalization capabilities of low-resource ASR systems. The steps of the algorithm are :

(i) Firstly, we use a score estimator, $Estimator()$, to evaluate the score set S_{true} of the true dataset X_{true} and the score set S_{syn} of the synthesized data X_{syn}, as shown in Eq. 5.1. The score values are the measurement of the quality of the data.

$$\begin{aligned} S_{true} &= Estimator(X_{true}) \\ S_{syn} &= Estimator(X_{syn}) \end{aligned} \tag{5.1}$$

(ii) Next, we fit the S_{true} into an observed prior distribution function $Prior()$ as the intermediate state (i.e., it can be a Gaussian, Beta or Uniform distribution, which suits the distribution of S_{true} through observation) and estimate the prior distribution's parameters, $params$, by using the method of moments $Moments()$ [23] in Eq. 5.2.

$$params = Moments(Prior, S_{true}) \tag{5.2}$$

(iii) We implement a rejection sampling algorithm [24] to filter out the score set S_{syn} of the synthesized data that significantly deviates from the score distribution of the true dataset S_{true}. To achieve this, we first compute the maximum probability density value M of the prior distribution's *probability density function*, $Prior.PDF()$, over the scores of the true dataset S_{true} and the prior distribution parameters $params$, as shown in Eq. 5.3.

$$M = max(Prior.PDF(S_{syn}, params)) \tag{5.3}$$

(iv) We determine the acceptance criterion as defined in Eq. 5.4, to filter the subset

S'_{syn} from S_{syn}, where s_i represents a candidate sample drawn from the synthesized data score set S_{syn}.

$$P_{accept}(s_i) = \frac{Prior.PDF(s_i, params)}{M} \tag{5.4}$$

(v) We filter the score set of synthesized dataset S'_{syn}, as defined in Eq. 5.5. Data samples from the score set of synthesized data S_{syn}, of size n, are accepted or rejected using the acceptance criterion, by comparing a uniform random variable $u_i \sim Uniform(0, 1)$ with the acceptance probability $P_{accept}(s_i)$. This selection process ensures that the score distribution of S'_{syn} closely aligns with S_{true}, that effectively mitigating the domain mismatch.

$$\begin{aligned} S'_{syn} = \{s_i \in S_{syn} | u_i \leq P_{accept}(s_i), \\ s_i \sim S_{syn}, u_i \sim Uniform(0, 1), i = 1, 2...n\} \end{aligned} \tag{5.5}$$

Finally, the filtered synthesized dataset X'_{syn} is constructed based on its corresponding score set S'_{syn} for fine-tuning downstream tasks, as shown in Eq. 5.6.

$$X'_{syn} = \{(x'_i, s'_i) | s'_i \in S'_{syn}, i = 1, 2...n\} \tag{5.6}$$

5.3 Experiments

5.3.1 Experiments Set Up

(i) **Te Reo Māori Speech Dataset:** We use our collected 27 hours labeled Te Reo Māori read speech dataset featuring recordings from 18 native speakers (10 male, 8 female) [5]. Notably, the dataset is imbalanced: two speakers (Speakers 1 and 11) account for over 6 hours of recordings each, while the remaining speakers contribute less than 2 hours each, resulting in a long-tail distribution. This imbalance indicates persistent challenges in low-resource language ASR (i.e., limited speaker diversity, insufficient data volume, and uneven speaker representation).

(ii) **Te Reo Māori Synthesized Dataset:** To address the aforementioned challenges of developing low-resource ASR for Te Reo Māori, we employ our Zero-Voice

[5] model, a low-resource, zero-shot TTS framework. Compared with recent zero-shot TTS models, such as Your-TTS [1], XTTS [2], ZMM-TTS [4] and HierSpeech++ [25], which rely on cross-lingual datasets and extensive labeled training data ($>$ 1000 hours data), Zero-Voice achieves high-quality synthesized speech and demonstrates comparable zero-shot capabilities, even when trained on significantly smaller and imbalanced datasets (i.e., less than 30 hours of labeled data).

Following the experimental setup used in Zero-Voice [5], we initially train the model on a 27-hour labeled Te Reo Māori speech dataset. To enhance ASR performance, we synthesize 360 hours of labeled speech data by selecting reference samples from each of the 18 original speakers, along with 10 additional Māori speakers [5]. Based on this experimental setup, we expand the synthesized dataset to 520 hours by incorporating additional speech corpora, substantially increasing both the data volume and speaker diversity for downstream ASR model fine-tuning.

(iii) **Te Reo Māori Validation and Test Sets:** We allocate 10% of our 27-hour collected Te Reo Māori speech dataset as a validation set. For the test set, we utilize the test set of publicly available Māori subset of the FLEURS dataset [26]. All the experimental results, including WER and CER, are reported exclusively on this test set to ensure consistency and comparability across all methods.

(iv) **Pre-trained Model Fine-tuning Configuration:** In this study, we fine-tune the `Whisper-large-v3` model [10] as foundation model, a Transformer-based [27] architecture designed for multilingual speech recognition and translation (including the Māori language pre-trained on weakly supervised data). This pre-trained model leverages extensive prior knowledge acquired from diverse speakers and languages, making it a strong foundation for enhancing low-resource ASR systems. For fine-tuning the `Whisper-large-v3` model on our Te Reo Māori ASR system, we use the following hyperparameter settings: models are fine-tuned for 15 epochs with a learning rate of 1e-5, using the AdamW optimizer [28] and a cosine learning rate scheduler with warm-up steps equal to half an epoch. The optimal model checkpoint is selected for comparison based its performance on the validation set.

(v) **Our Score-distribution-matching Data Selection Method Set Up:** In the step (i) of Section 5.2, the estimator function, $Estimator()$, we utilize the `Whisper-large-v3` model to compute the Character Error Rate (CER) as the score value for both the true and synthesized speech datasets, denoted as S_{true} and S_{syn}, respectively, as de-

fined in Eq. 5.1. Beta distribution is set as the prior distribution function, $Prior()$, to calculate the parameters *params* in Eq. 5.2 in the step (ii). Then, we follow the step (iii)-(v) to construct synthesized dataset X'_{syn}. Lastly, we fine-tune the `Whisper-large-v3` model using the filtered synthesized dataset X'_{syn} exclusively, as well as using X'_{syn} in conjunction with the true dataset to evaluate our proposed method.

5.3.2 Baselines

We propose the following baselines for fine-tuning the pre-trained model `Whisper-large-v3` .

(i) **Without Fine-tuning:** Baseline performance of the `Whisper-large-v3` model on the test set without any additional training or fine-tuning.

(ii) **True Dataset:** The 27 hours of collected and labeled Te Reo Māori speech data.

(iii) **Synthesized Dataset:** The synthesized dataset generated using the Zero-Voice model, ranging from 0 to 520 hours, derived from the true Te Reo Māori dataset.

(iv) **True + Synthesized Dataset:** A combination of the true dataset and synthesized datasets, with the volume of synthesized data varying from 0 to 520 hours.

(v) **True + Quality-based Synthesized Dataset:** We combine the true dataset with synthesized data selected from three score ranges: 0–50th percentile, 25–75th percentile, and 50–100th percentile, ranked in ascending order of scores. This method aims to investigate the impact of synthesized data quality (categorized as high, medium, and low thresholds) on ASR model performance.

(vi) **Synt++ by Hu et al.** [16] pre-train a discriminator model and (vii) the method proposed by **Wang et al.** [17]: use pre-trained speaker embeddings, as evaluation metrics to select synthesized data for improving ASR accuracy, we implement both methods with our dataset for comparison.

(viii) **Adapter Double-way Fine-tuning** [18]**:** A state-of-the-art adapter-based method derived from the recent Children's ASR studies using synthesized speech data [18–22]. In this method, adapter layers are trained on synthesized data while keeping the pre-trained model frozen. Subsequently, the adapter layers and the entire model are fine-tuned using a mix of synthesized and real data, with only synthesized

data passing through the adapters during training. For our experiments, we integrate adapters at the feed-forward layers of the `Whisper-large-v3` model to evaluate its efficacy in improving performance for Te Reo Māori ASR systems.

5.3.3 Results and Discussions

(i) **Suboptimal Performance with Synthesized Speech Data:** We present baseline results for fine-tuning the model `Whisper-large-v3` as summarized in Table 5.1. Experimental findings reveal that simply increasing the volume of synthesized Māori speech data does not consistently improve ASR accuracy. Fine-tuning the `Whisper-large-v3` model with the 27-hour true dataset supplemented by synthesized data initially achieves optimal test set results, with a WER of 22.9% and a CER of 11.2%, which represents a significant improvement over the model's baseline performance without fine-tuning (i.e., WER of 37.9% and CER of 13.8%).

However, performance begins to degrade when the synthesized data volume exceeds 360 hours. Moreover, fine-tuning the `Whisper-large-v3` model exclusively on the synthesized dataset yields only marginal improvements, underscoring the limitations of synthesized data when used independently. This outcome aligns with recent studies on synthesized Children's speech ASR [17–21], which emphasize that imperfections in synthesized data can negatively impact ASR training. While moderate volumes of synthesized data are shown to enhance generalization, excessive amounts may lead to overfitting in pre-trained ASR models.

(ii) **Baseline Methods Results:** For the quality-based data selection methods, we evaluate subsets of synthesized speech data selected based on high, medium, and low-quality thresholds. The experimental results in Table 5.1 reveal worse WER and CER values compared to the True + Synthesized Dataset baseline. This indicates that the `Whisper-large-v3` model derives limited benefit from synthesized data selected solely based on specific quality thresholds, as such selection fails to account for a balanced representation of the synthesized data [29]. The baseline methods, Synt++ [16] and the method by Wang et al. [17], perform worse than the True + Synthesized Dataset baseline, as the 27 hours of Te Reo Māori data are insufficient to train robust pre-trained speaker embeddings or discriminator models for these methods. For the Adapter double-way fine-tuning [18] method, the experimental results demonstrate a marginal improvement over the True + Synthesized Dataset baseline, achieving a

Table 5.1: WER and CER on Baseline Methods for Fine-Tuning Whisper Large-v3 Model

Methods	**FLEURS (mi_nz) Test Set**		**Validation Set**	
	WER (%) ↓	**CER (%) ↓**	**WER (%) ↓**	**CER (%) ↓**
w/o fine-tuning [10]	37.9	13.8	45.4	19.6
27hr True data	28.3	12.8	19.1	6.9
+ 60hr Syn Data	25.4	11.9	18.0	6.6
+ 190hr Syn Data	24.1	11.5	19.2	7.2
+ 250hr Syn Data	24.6	11.3	19.9	7.1
+ 360hr Syn Data [5]	**22.9**	**11.2**	**16.5**	**6.2**
+ 450hr Syn Data	23.8	11.6	17.5	6.5
+ 520hr Syn Data	24.3	11.5	17.9	6.7
60hr Syn Data	35.4	23.9	28.0	16.6
190hr Syn Data	34.1	22.5	29.2	17.2
250hr Syn Data	34.6	22.3	29.0	17.1
360hr Syn Data	**34.3**	**22.2**	**27.7**	**16.4**
450hr Syn Data	34.8	22.6	28.5	16.5
520hr Syn Data	34.5	22.5	28.9	16.7
High Quality (0-50th percentile)	23.9	11.1	18.4	6.6
Medium Quality (25-75th percentile)	23.3	10.8	19.1	6.9
Low Quality (50-100th percentile)	23.4	11.9	18.0	6.6
Synt++ [16]	**24.6**	**12.2**	**17.4**	**7.3**
Wang el al. [17]	**23.8**	**11.5**	**16.8**	**7.2**
Adapter Double-way Fine-tuning [18]	**22.6**	**11.0**	**16.4**	**6.2**

WER of 22.4% and a CER of 10.9% on the test set.

(iii) **Improving Low-Resource ASR Accuracy with Our Score-distribution-matching Data Selection Method:**

Our proposed *score-distribution-matching* data selection method greatly improves the WER and CER on both the test and validation sets, as shown in Table 5.2. Fine-tuning the `Whisper-large-v3` model with true and filtered synthesized data achieves optimal ASR accuracy with a WER of 21.4% and a CER of 9.9% on the test set. These results outperform all the basline methods. Additionally, fine-tuning the `Whisper-large-v3` model exclusively with the filtered synthesized dataset further reduces the WER to 32.9% and the CER to 21.2% on the test set, surpassing the performance of fine-tuning with 360 hours of unfiltered synthesized data alone.

Fig. 5.1 visualizes our approach. The left panel shows the score distributions of the true dataset S_{true} and the synthesized dataset S_{syn}, computed using the $Estimator()$ function. The blue and green bars clearly show the domain mismatch between S_{true} and S_{syn}. The red curve represents the estimated Beta distribution for S_{true}, with

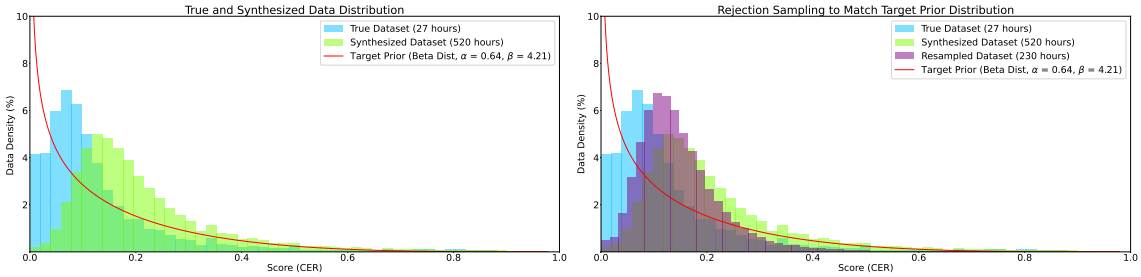

Figure 5.1: Score-distribution-matching Data Selection Method. **Left:** The score distributions of true and synthesized data, and the beta distribution fit for the true data scores. **Right:** A subset of synthesized data selected based on the beta distribution.

parameters $\alpha = 0.64$, $\beta = 4.21$, calculated using the method of moments [23].

The right panel depicts the score distribution of the selected synthesized dataset S'_{syn}, represented by the purple bars, 230 hours resampled synthesized data have been obtained through the rejection sampling algorithm. Notably, the score distribution of the filtered synthesized data S'_{syn} more closely aligns with that of the true data S_{true} than that of the synthesized data S_{syn}, effectively mitigating the domain mismatch, which underscores the effectiveness of our selection methodin improving the ASR model's performance by fine-tuning `Whisper-large-v3` with true and filtered synthesized data.

Table 5.2: WER and CER on Score-distribution-matching Data Selection Methods

Methods	**FLEURS (mi_nz) Test Set**		**Validation Set**	
	WER (%) ↓	**CER (%) ↓**	**WER (%) ↓**	**CER (%) ↓**
27hr True data	28.3	12.8	19.1	6.9
+ Score-distribution-matching	**21.4**	**9.9**	**16.1**	**6.0**
Score-distribution-matching	**32.9**	**21.2**	**26.5**	**16.2**

(iv) **CER as the Score Estimator:** As illustrated in Fig. 5.1, the score distributions of the true, synthesized, and filtered datasets exhibit a bell-shaped curve, rather than a uniform (square-shaped) distribution with equal frequencies across scores. This visual observation provides empirical support for adopting CER (computed from the pre-trained `Whisper-large-v3` base model) as a reliable and consistent metric for evaluating the quality of synthesized speech. The presence of a well-formed, interpretable distribution highlights CER's ability to effectively differentiate between varying quality levels across the speech samples. Our proposed method leverages this property by aligning the score distribution of synthesized data with that of the true

data. Furthermore, experimental results validate this rationale, demonstrating that using CER as the score estimator in our *score-distribution-matching* method leads to improved ASR performance for Te Reo Māori.

5.4 Conclusion

In this chapter, we propose a *score-distribution-matching* data selection method for fine-tuning pre-trained ASR models on the Māori language, effectively mitigating domain mismatches between synthesized and true data. Through extensive experiments on the Te Reo Māori dataset, our method demonstrates its effectiveness by significantly enhancing low-resource ASR performance over the three strong baselines. Moreover, we show that CER is a plausible estimation metric, as it clearly present normality for the true, synthesized and filtered data distributions. To further validate the robustness of our method, we plan to investigate evaluating additional prior distribution settings within the algorithm, e.g., the Normal and Gamma distributions. In future work, we also plan to apply this method to other low-resource languages.

References

[1] E. Casanova, J. Weber, C. Shulby, A. Júnior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multispeaker tts and zero-shot voice conversion for everyone," in *International Conference on Machine Learning (ICML)*, 2022.

[2] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, "Xtts: a massively multilingual zero-shot text-to-speech model," in *Interspeech*, 2024, pp. 4978–4982.

[3] W. Wang and Y. Qian, "Universal cross-lingual data generation for low resource asr," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 973–983, 2024.

[4] C. Gong, X. Wang, E. Cooper, D. Wells, L. Wang, J. Dang, K. Richmond, and J. Yamagishi, "Zmm-tts: Zero-shot multilingual and multispeaker speech synthesis conditioned on self-supervised discrete speech representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4036–4051, 2024.

[5] Z. Wang, F. Hou, and R. Wang, "Zero-voice: A low-resource approach for zero-shot text-to-speech," in *Techrxiv.173750018.83521236*, 2025.

[6] C. Edresson, S. Christopher, K. Alexander, C. Arnaldo, S. Anderson, A. Sandra, and A. Moacir, "Asr data augmentation in low-resource settings using cross-lingual multi-speaker tts and cross-lingual voice conversion," in *Interspeech*, 2023, pp. 1244–1248.

[7] J. Vásquez-Correa, H. Arzelus, J. Martin-Doñas, J. Arellano, a. Gonzalez-Docasal, and A. Álvarez, "When whisper meets tts: Domain adaptation using only synthetic speech data," in *Text, Speech, and Dialogue*. Cham: Springer Nature Switzerland, 2023, pp. 226–238.

[8] S. Ruder, A. Søgaard, and I. Vulić, "Unsupervised cross-lingual representation learning," in *Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics, 2019, pp. 31–38.

[9] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, K. S, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W. Hsu, A. Conneau, and M. Auli, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[10] A. Radford, J. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning (ICML)*, vol. 202, 2023, pp. 28 492–28 518.

[11] Y. Z. et al., "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1519–1532, 2022.

[12] Z. Lu, Y. Wang, Y. Zhang, W. Han, Z. Chen, and P. Haghani, "Unsupervised data selection via discrete speech representation for asr," in *Interspeech*, 2022, pp. 3393–3397.

[13] R. Gody and D. Harwath, "Unsupervised fine-tuning data selection for asr using self-supervised speech models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[14] Y. Chen, W. Ding, and J. Lai, "Improving noisy student training on non-target domain data for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[15] N. Lagos and I. Calapodescu, "Unsupervised multi-domain data selection for asr fine-tuning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 711–10 715.

[16] T. Hu, M. Armandpour, A. Shrivastava, J. Chang, H. Koppula, and O. Tuzel, "Synt++: Utilizing imperfect synthetic data to improve speech recognition,"

in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7682–7686.

[17] W. Wang, Z. Zhou, Y. Lu, H. Wang, C. Du, and Y. Qian, "Towards data selection on tts data for children's speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6888–6892.

[18] T. Rolland and A. Abad, "Improved children's automatic speech recognition combining adapters and synthetic data augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 757–12 761.

[19] T. Rolland and A. Abad, "Exploring adapters with conformers for children's automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 747–12 751.

[20] T. Rolland and A. Abad, "Introduction to partial fine-tuning: A comprehensive evaluation of end-to-end children's automatic speech recognition adaptation," in *Interspeech*, 2024, pp. 5178–5182.

[21] T. Rolland and A. Abad, "Shared-adapters: A novel transformer-based parameter efficient transfer learning approach for children's automatic speech recognition," in *Interspeech*, 2024, pp. 2370–2374.

[22] T. Rolland and A. Abad, "Towards improved automatic speech recognition for children," in *IberSPEECH*, 2024, pp. 251–255.

[23] K. O. Bowman and L. R. Shenton, "Estimator: Method of moments," *Encyclopedia of statistical sciences*, p. 2092–2098, 1998.

[24] G. Casella, P. Robert, and T. Martin, "Generalized accept-reject sampling schemes." *Institute of Mathematical Statistics*, p. 342–347, 2004.

[25] S. Lee, H. Choi, S. Kim, and S.-W. Lee, "Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis," in *arXiv preprint arXiv:2311.12454*, 2023.

[26] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *IEEE Spoken Language Technology Workshop (SLT)*, 2022.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, G. Llion, L. K. A. Gomez, and I. Polosukhin, "Attention is all you need," in *Neural Information Processing Systems (NIPS)*, vol. 30, 2017, pp. 5998—-6008.

[28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019.

[29] Z. Yang, M. Ding, T. Huang, Y. Cen, J. Song, B. Xu, Y. Dong, and J. Tang, "Does negative sampling matter? a review with insights into its theory and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5692–5711, 2024.

Chapter 6

Summary

This chapter offers concluding remarks on the thesis. Throughout our research, we have contributed to four research directives of low-resource Automatic Speech Recognition (ASR). In this final chapter, we will review the proposed methods in Section 6.1 and provide insights into future research directions in Section 6.2.

6.1 Research Summary

In this thesis, we have explored four research directions for low-resource automatic speech recognition (ASR), including data augmentation, mitigating catastrophic forgetting in continual learning, zero-shot text-to-speech (TTS) and addressing domain mismatch between true and synthesized speech data. Below is a summary of our methods and contributions:

Chapter 2 presents CyclicAugment, a novel data augmentation approach to enhance ASR training by diversifying augmentation policies and employing a cosine annealing scheduler to adjust augmentation magnitudes dynamically [1]. CyclicAugment enables models to escape local optima, resulting in optimal ASR accuracy. Our experimental results show that CyclicAugment outperforms two strong baselines, SpecAugment [2] and RandAugment [3], in both supervised ASR training and fine-tuning of pre-trained ASR models.

Chapter 3 presents CLRL-Tuning (Continual Learning Layer-wise Tuning), an approach to mitigate catastrophic forgetting in low-resource ASR [4]. CLRL-Tuning

selectively updates the parameters of a randomly chosen layer of a pre-trained ASR model during each training epoch, balancing adaptability and knowledge retention. Theoretical analysis demonstrates that this strategy allows the model to efficiently adapt to new tasks with minimal additional training, leveraging the prior knowledge encoded in large pre-trained ASR models. Experimental results indicate that CLRL-Tuning outperforms four strong baselines, including Knowledge Distillation [5] and Gradient Episodic Memory [6], achieving significant improvements in average Word Error Rate (WER). Notably, our CLRL-Tuning approach is established as a baseline in two recent works (i.e., low-rank adaptation (LoRA) for multilingual ASR [7] and efficient rehearsal via singular value tuning for continual learning in ASR [8]) with additional benchmarking conducted across more diverse domains.

Chapter 4 presents Zero-Voice, a novel end-to-end zero-shot text-to-speech (TTS) system designed specifically for low-resource languages [9]. Zero-Voice comprises source filter network, acoustic feature encoder, acoustic feature refiner and waveform vocoder to synthesizes high-fidelity speech directly from mel-spectrograms, closely resembling the reference speech sample. Zero-Voice only requires minimal labeled data for training of low-resource languages. We train Zero-Voice on 27 hours of Te Reo Māori speech data and use it to synthesize additional speech data. This synthesized dataset is then used to fine-tune pre-trained ASR models for Te Reo Māori. The resulting system achieve remarkable improved accuracy on the Māori (`nz_mi`) test set of the Google Fleurs dataset, an open-source benchmark [10].

Chapter 5 presents a Score-distribution-match approach to address the domain mismatch issue between true and synthesized Te Reo Māori data generated by the Zero-Voice model [11]. This approach filters synthesized data by aligning its score distribution with the prior distribution of true data, that ensures closer resemblance and improved compatibility for the synthesized dataset with the true data. Experimental results demonstrate that this approach further improves the accuracy of Te Reo Māori ASR achieved in Chapter 4 on the Māori (`nz_mi`) test set of the Google Fleurs dataset [10].

To summarize our work in this thesis, we present four research directives aimed at enhancing low-resource ASR systems, with a particular focus on Te Reo Māori (i.e., an official language of New Zealand) in Chapter 4 and 5. These contributions address key challenges in low-resource ASR, including data augmentation, continual learning,

low-resource zero-shot TTS synthesis and domain mismatch mitigation between true and synthesized datasets. Our work has been recognized in the speech processing community through publications in top-tier conferences such as *Interspeech*, *ICASSP* and is under review in prestigious journals. These contributions mark significant advancements in low-resource ASR, especially for endangered languages.

6.2 Future work and directions

In this section, we discuss three directions of future research: constructing diversified and balanced speech datasets in Section 6.2.1, developing improved low-resource zero-shot TTS System in Section 6.2.2, and multilingual fine-tuning of pre-Trained Models for low-Resource Languages in Section 6.2.3.

6.2.1 Constructing Diversified and Balanced Speech Datasets

In Chapter 4 and 5, We have made significant progress in developing a low-resource ASR system for Te Reo Māori, achieving accuracy at a human-comparable level. However, a performance gap persists in reaching above-human-level accuracy, similar to that of high-resource ASR systems. For example, state-of-the-art multilingual ASR models [12, 13] including those for English and Chinese, are trained on large-scale datasets such as LibriSpeech [14] and AIShell-1 [15], each comprising over 1,000 hours of labeled speech from more than 1,000 speakers. These datasets offer extensive speaker diversity and a balanced distribution of labeled speech across individuals, supporting robust and generalizable feature learning across a wide range of speaker characteristics.

Our Te Reo Māori dataset contains a limited number of speakers and exhibits significant imbalances in the distribution of labeled speech data across speakers, which leads to underrepresented features for some speakers and overrepresented ones for others. These limitations degrade the generalization ability of the low-resource ASR system. Moreover, the development of zero-shot TTS systems is also constrained, as underrepresented speaker data often results in suboptimal synthetic speech quality for further training of low-resource ASR systems. Therefore, addressing this issue requires the construction of a more diversified and balanced speech dataset, even with limited labeled data [16]. Ensuring representative coverage across speakers will improve ASR performance, and enhance the quality and diversity of synthetic speech

generated through zero-shot TTS approaches [9].

6.2.2 Developing Improved Low-resource Zero-Shot TTS System

Exploiting the development of more powerful zero-shot TTS systems is also a promising research direction. By leveraging recent advancements in generative modeling paradigms, such as diffusion models [17, 18] or scalable transformer-based generative approaches [19, 20], zero-shot TTS systems for low-resource languages could achieve significantly enhanced performance. These approaches can enable the synthesis of speech that more closely resembles the acoustic characteristics, prosody, and quality of the reference speech samples, ensuring higher fidelity and speaker similarity.

With such advancements, the synthetic speech could be of sufficiently high fidelity and similarity to be effectively treated as true data. This would address the imperfections in synthesized datasets and minimize the domain mismatch between true and synthesized data, as discussed in Chapter 5. Consequently, the enhanced synthesized datasets could be used as a robust foundation for training low-resource ASR models, further narrowing the performance gap between low-resource and high-resource ASR systems.

6.2.3 Multilingual Fine-Tuning of Pre-Trained Models

We demonstrate the success of combining zero-shot TTS and data selection in achieving desirable results for Te Reo Māori in Chapters 4 and 5. Furthermore, multilingual ASR training has been shown to significantly improve accuracy and overall performance, particularly in the context of low-resource languages[12, 13].

Additionally, recent multi-modal large language models (LLMs)(i.e., those capable of interpreting text, images, video, and audio) [21–23] enhanced through self-supervised fine-tuning (SFT) and reinforcement learning (RL) on large-scale multi-modal data, have shown remarkable emergent abilities and generalization across a wide range of tasks (e.g., agentic coding, navigation, and planning).

However, they lack knowledge of low-resource languages, which makes them weak at transcribing low-resource audio information [24].

Hence, we hypothesize that applying multilingual fine-tuning to pre-trained multilingual ASR models by using multiple low-resource languages (each with limited labeled data) could further improve model performance. When integrated with our proposed zero-shot TTS system (i.e., Zero-Voice [9]) and score-distribution-matching [11] data selection approach, this strategy has the potential to significantly enhance ASR performance across a broad range of low-resource languages.

References

[1] Z. Wang, F. Hou, Y. Qiu, Z. Ma, S. Singh, and R. Wang. Cyclicaugment: Speech data random augmentation with cosine annealing scheduler for automatic speech recognition. In *Interspeech 2022*, pages 3859–3863, 2022.

[2] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech*, pages 2613–2617, 2019.

[3] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems (NIPS)*, pages 18613–18624, 2020.

[4] Z. Wang, F. Hou, and R. Wang. Clrl-tuning: A novel continual learning approach for automatic speech recognition. In *Interspeech 2023*, pages 1279–1283, 2023.

[5] Z. Li and D. Hoiem. Learning without forgetting. In *Proc. European Conference on Computer Vision (ECCV)*, pages 614–629, 2016.

[6] D. Lopez-Paz and M. A. Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pages 6467–6476, 2017.

[7] C . Kwok, H . Liu, Jia Yip, S. Li, and S. Chng. A two-stage lora strategy for expanding language capabilities in multilingual asr models. *IEEE Transactions on Audio, Speech and Language Processing*, 33:2576–2590, 2025.

[8] S . Eeckt and H. Van hamme. Efficient rehearsal for continual learning in asr via singular value tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 34:978–991, 2026.

[9] Z. Wang, F. Hou, and R. Wang. Zero-voice: A low-resource approach for zero-shot text-to-speech. In *Techrxiv.173750018.83521236*, 2025.

[10] A. Conneau, M. Ma, S. Khanuja, Yu Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna. Fleurs: Few-shot learning evaluation of universal rep-

resentations of speech. In *IEEE Spoken Language Technology Workshop (SLT)*, 2022.

[11] Z. Wang, F. Hou, and R. Wang. Synthesized data selection via score distribution matching for te reo māori automatic speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 18172–18176, 2026.

[12] A. Radford, J. Wook Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, volume 202, pages 28492–28518, 2023.

[13] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W. Hsu, A. Conneau, and M. Auli. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[15] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Oriental COCOSDA 2017*, 2017.

[16] Z. Yang, M. Ding, T. Huang, Y. Cen, J. Song, B. Xu, Y. Dong, and J. Tang. Does negative sampling matter? a review with insights into its theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 5692–5711, 2024.

[17] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020.

[18] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *Proceedings of International Conference on Learning Representations*, 2021.

[19] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems (NIPS)*, 2024.

[20] J. Han, J. Liu, Y. Jiang, B. Yan, Y. Zhang, Z. Yuan, B. Peng, and X. Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *arXiv 2412.04431*, 2024.

[21] Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, 2026.

[22] Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, 2026.

[23] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

[24] X. Shi et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.