END-TO-END AUTOMATIC SPEECH RECOGNITION FOR LOW-RESOURCE LANGUAGES

A thesis submitted in partial fulfillment for the degree of

Doctor of Philosophy

in

Computer Science

at the

School of Mathematical and Computational Sciences

Massey University

Auckland, New Zealand

SATWINDER SINGH

July 2023

Satwinder Singh: *End-to-End Automatic Speech Recognition for Low-Resource Languages*
Doctor of Philosophy ©July 2023

SUPERVISORS:

Professor Ruili Wang

Dr. Feng Hou

LOCATION:

Auckland, New Zealand

TIME FRAME:

July 2023

Ohana means family.

Family means nobody gets left behind or forgotten.

— Lilo & Stitch

Dedicated to Mum and Dad

FOREVER

ABSTRACT

Automatic speech recognition (ASR) for low-resource languages presents numerous challenges due to the lack of various crucial linguistic resources including annotated speech corpus, lexicon, and raw language text. In this thesis, we propose different approaches to improve fundamental frequency estimation and speech recognition for low-resource languages.

Firstly, we propose DeepFo, a new deep learning technique for fundamental frequency (F_0) estimation. Existing models have limited learning capabilities due to using a shallow receptive field. Our DeepFo extends the receptive field by using dilated convolutional blocks. Additionally, we enhance training efficiency and speed by incorporating residual blocks with residual connections. We achieve state-of-the-art results with DeepFo, even using 77.4% fewer network parameters.

Secondly, we introduce a new meta-learning framework for low-resource speech recognition that improves on the previous model-agnostic meta-learning (MAML) approach. Our framework addresses issues of MAML such as training instabilities and slower convergence by using a multi-step loss (MSL). MSL calculates losses at each step of MAML's inner loop and combines them using a weighted importance vector, which prioritizes the loss at the last step.

Thirdly, we propose an end-to-end ASR approach for low-resource languages that exploit the synthesized datasets along with real speech datasets. We evaluate our approach on the low-resource Punjabi language, which is widely spoken across the globe by millions of speakers, however, still lacks annotated speech datasets. Our empirical results show that our synthesized datasets (Google-synth and CMU-synth) can significantly improve the accuracy of our ASR model.

Lastly, we introduce a self-training approach, also known as pseudo-labeling approach, to enhance the performance of low-resource speech recognition. While most self-training research has centered on high-resource languages such as English, our work is focused on the low-resource Punjabi language. To weed out the low-quality pseudo-labels, we employ length normalized confidence score.

Overall, our experimental evaluation validates the efficacy of our proposed approaches and shows that they outperform existing baseline approaches for F_0 estimation and low-resource speech recognition.

PUBLICATIONS

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

- **Satwinder Singh**, Ruili Wang, and Yuanhang Qiu, "DeepFo: End-to-end fundamental frequency estimation for music and speech signals", *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, pp. 61-65, 2021, URL: https://doi.org/10.1109/ICASSP39728.2021.9414050, (CORE rank B).

- **Satwinder Singh**, Ruili Wang, and Feng Hou, "Improved Meta Learning for Low Resource Speech Recognition" *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, pp. 4798-4802, 2022, URL: https://doi.org/10.1109/ICASSP43922.2022.9746899, (CORE rank B).

- **Satwinder Singh**, Ruili Wang, Feng Hou, and Zhizhong Ma, "Enhancing End-to-End Automatic Speech Recognition for Low-Resource Punjabi Language Using Synthesized Datasets", *Computer Speech & Language*, in revision.

- **Satwinder Singh**, Ruili Wang, and Feng Hou, "A Novel Self-training Approach for Low-resource Speech Recognition", *International Speech Communication Association (INTERSPEECH)*, Accepted, (CORE rank A).

- **Satwinder Singh**, Ruili Wang, and Feng Hou, "Real and Synthetic Punjabi Speech Datasets for Automatic Speech Recognition", *Data in Brief*, in submission.

- Zhihan Wang, Feng Hou, Yuanhang Qiu, Zhizhong Ma, **Satwinder Singh**, and Ruili Wang, "CyclicAugment: Speech Data Random Augmentation with Cosine Annealing Scheduler for Automatic Speech Recognition", *In Proceedings of the International Speech Communication Association (INTERSPEECH)*, Incheon, Korea, pp. 3859-3863, 2022, URL: https://doi.org/10.21437/Interspeech.2022-526, (CORE rank A).

- Zhizhong Ma, **Satwinder Singh**, Yuanhang Qiu, Feng Hou, Ruili Wang, Christopher Bullen, and Joanna Ting Wai Chu, "Automatic speech-based smoking status identification", *In Intelligent Computing. SAI 2022. Lecture Notes in Networks and Systems*, Volume 508, pp. 193-203, Springer, Cham, 2022, URL: https://doi.org/10.1007/978-3-031-10467-1_11.

- Yuanhang Qiu, Ruili Wang, Feng Hou, **Satwinder Singh**, Zhizhong Ma, and Xiaoyun Jia. "Adversarial multi-task learning with inverse mapping for speech enhancement", *Applied Soft Computing*, Volume 120, 2022, 108568, URL: https://doi.org/10.1016/j.asoc.2022.108568.

- Yuanhang Qiu, Ruili Wang, **Satwinder Singh**, Zhizhong Ma and Feng Hou, "Self-Supervised Learning Based Phone-Fortified Speech Enhancement", *In Proceedings of the International Speech Communication Association (INTERSPEECH)*, Brno, Czechia, pp. 211-215, 2021, URL: https://doi.org/10.21437/Interspeech.2021-734, (CORE rank A).

- Zhizhong Ma, Chris Bullen, Joanna Ting Wai Chu, Ruili Wang, Yingchun Wang, and **Satwinder Singh**, "Towards the objective speech assessment of smoking status based on voice features: a review of the literature." *Journal of Voice*, Volume 37, Issue 2, 2021, URL: https://doi.org/10.1016/j.jvoice.2020.12.014.

ACKNOWLEDGEMENTS

I would like to acknowledge many individuals who have contributed towards the completion of PhD studies. First and foremost, I extend my sincere gratitude to my main supervisor, Professor Ruili Wang, for his unwavering support, encouragement, and guidance throughout my research journey. Your expertise, valuable insights, and constructive criticism were instrumental in shaping my ideas and improving my work. Professor Wang, you have given me great freedom to work on various research problems and also provided an excellent collaborative environment to carry out my research.

I would also like to express my thanks to my co-supervisor, Dr. Feng Hou, for his expert feedback during every stage of my PhD journey. You have been always available whenever I needed any expert advice and suggestions.

I am also very thankful to the staff and faculty members of the School of Mathematical and Computational Sciences, Massey University, especially, Ms. Annette Warbrooke and Ms. Sue Di Leo for their kind support and help during my PhD. Also, I would like to pay my thanks to the former Head of the School, Professor Dianne Brunton, for being kind to me and helping me during the initial period of my studies. I am grateful to have the opportunity to do my PhD at Massey University. I am so grateful to Massey University for providing me with various scholarships during my PhD.

I also want to acknowledge my friends and colleagues who have provided me with their support and encouragement throughout my studies. Their motivation and inspiration have helped me to keep going during tough times.

Finally, I am forever in the debt to my Mum, Rajinder Kaur, and Dad, Harbhajan Singh, for their love, support, and encouragement throughout my PhD. Also, how can I forget to mention my two lovely sisters, Ritu and Harpreet, for just being there whenever I needed

someone to talk to. My family is the pillar of my success, and I could not have done this without them.

On the last note, I again express my heartfelt thanks to everyone who has contributed to my research journey, and I hope my research work will contribute positively towards the advancement of knowledge in my field.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

This chapter provides an overview of this thesis. We introduce the background and previous studies on low-resource languages in Section 1.1 and Section 1.2. We explain our motivations in Section 1.3, where we discuss issues presented by existing approaches. Further, in Section 1.5, we summarize the main contributions of this thesis. Lastly, the outline of our thesis is listed in Section 1.6.

1.1 LOW-RESOURCE LANGUAGES

Automatic speech recognition (ASR) technology has revolutionized the way we interact with computers, making it easier and more natural for users to communicate. The state-of-the-art end-to-end (E2E) ASR systems, have shown excellent performance on various mainstream languages. However, the performance of these systems is often limited by the availability of annotated data (audio and corresponding text pairs) and other linguistic resources (i.e., language text and lexicon) in the target language. This is particularly true for low-resource languages, which lack a sufficient amount of annotated datasets to build an accurate ASR system. The limited availability of annotated data is a significant challenge for low-resource languages, given that only about 100 out of approximately 7,000 languages spoken worldwide have established ASR systems [1].

Before proceeding further, we first need to establish what are low-resource languages. In the field of natural language processing, specifically in the context of ASR, low-resource languages are those that lack the necessary resources to build a mature ASR system. These resources include annotated speech datasets, a vast collection of raw text, and reliable pronunciation dictionaries [2].

Annotated speech datasets play a crucial role in the development of ASR systems. The annotated datasets consist of audio recordings with corresponding transcriptions of the spoken words and phrases, which provides a reference for ASR models to learn from. However, creating annotated speech datasets is an expensive and time-consuming process that requires skillful transcribers and diverse speakers of the target language [3]. On average, transcribing an hour of spontaneous speech can take up to 6 hours and can cost around $90-$150 per hour [4]. In low-resource settings, where there is limited or no access to large annotated datasets, the ASR models produce noisy transcriptions, which can affect the overall effectiveness of the system.

A vast collection of raw text is also an essential resource for building strong language models for ASR systems. Language models use these collections to predict the probability of a given word or phrase in a particular language. However, low-resource languages often lack sufficient unpaired text, making it challenging to train reliable language models [5]. Besides, some of the low-resource languages do not even have a standard writing system, which presents unique challenges for producing and documenting linguistic data [6]. For example, the Pirahã language, which is spoken by the Pirahã people residing in the Brazilian Amazon rainforest, features a sophisticated tonal grammar, but lacks a conventional writing system [7].

Further, pronunciation dictionaries are another crucial resource for developing ASR systems [8]. These dictionaries map the phoneme sequences to their corresponding words, enabling the system to recognize and transcribe spoken words accurately. However, developing a pronunciation dictionary for a language requires expert linguists to curate a

phonetic dictionary for every possible word in the language. This process can take years and may still be subject to human errors and pronunciation variations.

1.2 ASR SYSTEMS FOR LOW-RESOURCE LANGUAGES

The research question driving this thesis is: How can we improve speech recognition performance in low-resource languages? This question is motivated by several factors. Firstly, low-resource languages represent a significant portion of the world's population and culture, and improved speech recognition technology can play a crucial role in bridging the digital divide and empowering these communities. Secondly, the advancements in speech recognition technology have enabled a wide range of applications, such as speech-based interfaces for mobile devices, accessibility technology for people with disabilities, and automatic translation systems. Improved speech recognition performance in low-resource languages can help extend these benefits to a wider audience.

The development of ASR systems for low-resource languages has been a very active area of research for the past several years. Over time, many effective approaches and research directions have been followed to solve the issues associated with low-resource languages. One of these approaches solves the scarcity of data through data augmentation techniques. Several studies have proposed standard data augmentation techniques, such as speed perturbation [9], pitch perturbation, and noise-based augmentations [10], [11]. Further, some augmentations were based on audio transformation including time and frequency masking and time warping [12]. Apart from this, text-to-speech (TTS) synthesis has also been used as an augmentation technique for low-resource languages [13]–[18].

Further, most other studies involve training an ASR model on multiple languages at once in multilingual and/or cross-lingual fashion [19]–[22]. Another approach is transfer learning, where the model is first pre-trained on high-resource languages and then fine-tuned on the target low-resource languages [23]–[25]. Recently, some studies also focused on meta-learning for low-resource languages [26], [27]. Furthermore, recent advancements

in self-supervised learning based wav2vec models have shown significant improvements in low-resource scenarios [28]–[31]. These models leverage the huge amounts of unlabeled speech data to pre-train the model to learn powerful shared feature representations across multiple languages. Besides, self-training approaches also exploit available unlabeled audio dataset in a very simple but effective way and have shown significant improvement in low-resource settings [32]–[36]. Overall, these advancements have made ASR systems more effective for low-resource languages.

1.3 MOTIVATIONS

Despite significant progress in ASR for low-resource language, there are still many challenges that need to be addressed.

- **Data-driven F_0 estimation:** F_0 estimation is important in various speech processing and music information retrieval applications, especially in the case of speech recognition of tonal languages [37]. F_0 estimation algorithms can effectively extract useful tonal features in the case of tonal languages such as Punjabi and Mandarin. Existing F_0 estimation approaches are mostly based on hand-designed statistical algorithms [38]–[42]. Very few of them are based on data-driven approaches (machine/deep learning), however, they still have limited learning capabilities due to their shallow receptive field [43].

- **Meta learning for low-resource ASR:** Meta-learning or *learning to learn*, is not a nascent concept but has gained increasing attention in recent times. Particularly in the realm of deep learning, meta-learning has gained prominence owing to its expansive range of applications and benefits. Meta-learning facilitates faster generalization to diverse tasks with limited examples and steps. Meta-learning has been employed in diverse research domains, including several computer vision tasks [44]–[47], natural language processing [48]–[50], and more recently, automatic speech recognition [27], [51], [52]. For speech recognition, model-agnostic based meta-learning (MAML)

[44], has shown encouraging results as compared to multi-task learning [27]. However, the existing MAML based approaches for ASR suffer from issues in terms of accuracy and inconsistent training behavior.

- **Availability of the annotated datasets:** There exists an ample number of annotated datasets for high-resource languages such as English, Mandarin, German, Spanish, etc. However, annotated datasets are very scarce for most low-resource languages, which presents a fundamental challenge in building ASR systems for such languages. Further, producing accurate annotated speech data could be expensive in terms of time and money.

- **Self-training for low-resource languages:** Most of the previous and recent work in self-training/ pseudo-labeling is based on and demonstrated on high-resource languages such as English [33]–[35], [53], [54]. However, most of the languages of the world are low-resource and the ASR system designed for them performs poorly when compared to those with high-resource languages. For example, very little work has been done for end-to-end self-training for the Punjabi language.

1.4 RESEARCH OBJECTIVE

In this thesis, we propose deep learning based approaches for speech processing tasks especially for F_0 estimation and automatic speech recognition for low-resource languages. Our research objectives are listed as follows:

- **Objective 1:** To address the limitations of current F_0 estimation approaches, this objective aims to develop an end-to-end deep learning-based approach called DeepF0. The objectives are twofold: (1) to design a model using dilated convolutional neural networks to capture longer context and improve accuracy, and (2) to optimize the network parameters to reduce memory and computational requirements while enhancing the model's receptive field.

- **Objective 2:** To enhance the meta-learning approach for low-resource automatic speech recognition by addressing the challenges of inconsistent training and slow convergence observed in existing meta-learning models. The research aims to propose and implement a novel multi-step loss function as a solution to these issues. The effectiveness of the enhanced meta-learning model will be empirically evaluated and compared to existing approaches using relevant performance metrics.

- **Objective 3:** To address the challenges of data scarcity in low-resource automatic speech recognition, this objective aims to develop an innovative approach for creating and utilizing both real speech and synthetic speech datasets. The objective is to design a systematic methodology to generate synthetic speech data representative of the target language. Subsequently, the efficacy of these datasets will be rigorously tested and evaluated using self-supervised crosslingual models. The research seeks to demonstrate how leveraging both real and synthetic speech data can significantly enhance the accuracy and robustness of low-resource automatic speech recognition systems.

- **Objective 4:** To enhance the performance of ASR systems in low-resource settings, this research aims to develop an innovative self-training approach. The primary goal is to leverage large unlabeled speech data in the target language to improve the accuracy and robustness of low-resource ASR systems. The research will involve designing a novel self-training framework with confidence based filtration method to iteratively improve the ASR models using the unlabeled data. The resulting self-training approach is expected to demonstrate substantial performance gains and contribute to the advancement of ASR technology in challenging low-resource scenarios.

1.5 MAIN CONTRIBUTIONS

Throughout this thesis, we will focus on the following two sub-tasks of speech processing: F_0 estimation (Chapter 2), automatic speech recognition for low-resource languages (Chapter 3, 4 and 5). The contributions in each of the aforementioned chapters are summarized as follow:

- Develop a data-driven approach called DeepFo for pitch estimation of music and speech signals. This approach could be used to extract pitch features for tonal languages for speech recognition purposes. DeepFo achieves state-of-the-art results as compared to existing baselines. This work is published as [55].

- Develop an improved meta-learning approach for automatic speech recognition of low-resource languages. We introduce a multi-step loss function to the MAML approach, which results in better performance in terms of character error rate and stable training behavior compared with the MAML approach. This work is published as [26].

- To tackle the issue of data scarcity, we produce two synthesized datasets (i.e., Google-synth and CMU-synth) and one real speech dataset called Punjabi Speech. We validate the effectiveness of synthesized datasets along with real speech datasets by developing a self-supervised framework for low-resource Punjabi language. The empirical results show significant error rate reduction when training self-supervised models with real speech and synthesized speech data. The Punjabi Speech [56], Google-synth [57], and CMU-synth [58] datasets have been made publicly available online, contributing valuable resources to the research community.

- To further improve ASR for low-resource language, we develop a simple but effective self-training approach. Our approach employs length normalized confidence based filtering method to sieve out low-quality pseudo-labels. Further, for self-training purposes, we produce an unlabeled dataset, which is compiled using audiobooks

data available on YouTube. The proposed approach reports excellent results on various datasets for the Punjabi language.

1.6 THESIS OVERVIEW

The remainder of this thesis is organized as follows:

- Chapter 2 describes our proposed pitch/fundamental frequency estimation approach.
- Chapter 3 discusses our proposed meta-learning approach for automatic speech recognition of low-resource languages.
- Chapter 4 presents our proposed self-training framework for low-resource speech recognition. The chapter lists the effectiveness of synthetic datasets in improving performance in low-resource settings.
- Chapter 5 propose a self-training approach to further improve speech recognition for low-resource languages.
- Chapter 6 summarizes the key findings of the thesis and discusses future work and directions.

Note that references related to each chapter are listed at the end of each chapter.

REFERENCES

[1] Kristin Precoda, "Non-mainstream languages and speech recognition: Some challenges," *CALICO Journal*, vol. 21, no. 2, pp. 229–243, 2004.

[2] Ekapol Chuangsuwanich, "Multilingual techniques for low resource automatic speech recognition," Ph.D. dissertation, MIT, Cambridge, United States, 2016.

[3] Karan Malhotra, Shubham Bansal, and Sriram Ganapathy, "Active Learning Methods for Low Resource End-to-End Speech Recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2019, pp. 2215–2219.

[4] Maxine Eskenazi, Gina-Anne Levow, Helen Meng, Gabriel Parent, and David Suendermann, *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. John Wiley & Sons, 2013.

[5] Ping Xu and Pascale Fung, "Cross-lingual language modeling for low-resource speech recognition," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 6, pp. 1134–1144, 2013.

[6] Gary F. Simons and Charles D. Fennig, "Ethnologue: Languages of the world," *Dallas, TX: SIL International*, 2017.

[7] DanielL Everett, "Cultural constraints on grammar and cognition in pirahã: Another look at the design features of human language," *Current anthropology*, vol. 46, no. 4, pp. 621–646, 2005.

[8] Chuandong Xie, Wu Guo, Guoping Hu, and Junhua Liu, "Web data selection based on word embedding for low-resource speech recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2016, pp. 1340–1344.

[9] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2015, pp. 3586–3589.

[10] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 5220–5224.

[11] Navdeep Jaitly and Geoffrey E Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013, p. 21.

[12] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2019, pp. 2613–2617.

[13] Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney, "Generating synthetic audio data for attention-based speech recognition systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7069–7073.

[14] Aleksandr Laptev, Roman Korostik, Aleksey Svischev, Andrei Andrusenko, Ivan Medennikov, and Sergey Rybin, "You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation," in *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, IEEE, 2020, pp. 439–444.

[15] Taniya Hasija, Virender Kadyan, and Kalpna Guleria, "Out Domain Data Augmentation on Punjabi Children Speech Recognition using Tacotron," in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1950, 2021, p. 012 044.

[16] Jason Li, Ravi Gadde, Boris Ginsburg, and Vitaly Lavrukhin, "Training neural speech recognition systems with synthetic speech augmentation," 2018. arXiv: 1811.00707.

[17] Andros TTjandra, Sakriani Sakti, and Satoshi Nakamura, "Listening while speaking: Speech chain by deep learning," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2017, pp. 301–308.

[18] Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Gary Wang, and Pedro Moreno, "Injecting text in self-supervised speech pretraining," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, pp. 251–258.

[19] Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, *et al.*, "MUCS 2021: Multilingual and code-switching ASR challenges for low resource Indian languages," in *Proc. International Speech Communication Association (Interspeech)*, International Speech Communication Association, 2021.

[20] Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert, "Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 4751–4755.

[21] Shiyu Zhou, Shuang Xu, and Bo Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," 2018. arXiv: 1806.05059.

[22] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2013, pp. 7304–7308.

[23] Changhan Wang, Juan Pino, and Jiatao Gu, "Improving Cross-Lingual Transfer Learning for End-to-End Speech Recognition with Speech Translation," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 4731–4735.

[24] Vikas Joshi, Rui Zhao, Rupesh R Mehta, Kshitiz Kumar, and Jinyu Li, "Transfer Learning Approaches for Streaming End-to-End Speech Recognition System," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 2152–2156.

[25] Xia Mao and Yulv Zhang, "Time-Delay Recurrent Neural Network for Cross-Lingual Speech Recognition," in *Recent Developments in Intelligent Computing, Communication and Devices*, Springer, 2019, pp. 341–348.

[26] Satwinder Singh, Ruili Wang, and Feng Hou, "Improved Meta Learning for Low Resource Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 4798–4802.

[27] Jui-Yang Hsu, Yuan-Jui Chen, and Hung-yi Lee, "Meta learning for end-to-end low-resource speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7844–7848.

[28] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "Wav2vec: Unsupervised pre-training for speech recognition," 2019. arXiv: 1904.05862.

[29] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[30] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2021, pp. 2426–2430.

[31] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," 2021. arXiv: 2111.09296.

[32] Jacob Kahn, Ann Lee, and Awni Hannun, "Self-training for end-to-end speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7084–7088.

[33] Daniel S. Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V. Le, "Improved Noisy Student Training for Automatic Speech Recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 2817–2821.

[34] Qiantong Xu, Tatiana Likhomanenko, Jacob Kahn, Awni Hannun, Gabriel Synnaeve, and Ronan Collobert, "Iterative pseudo-labeling for speech recognition," 2020. arXiv: 2005.09267.

[35] Yosuke Higuchi, Niko Moritz, Jonathan Le Roux, and Takaaki Hori, "Momentum pseudo-labeling for semi-supervised speech recognition," 2021. arXiv: 2106.08922.

[36] Zezhong Jin, Dading Zhong, Xiao Song, Zhaoyi Liu, Naipeng Ye, and Qingcheng Zeng, "Filter and evolve: Progressive pseudo label refining for semi-supervised automatic speech recognition," 2022. arXiv: 2210.16318.

[37] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, *et al.*, "The Kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, 2011.

[38] David Talkin and W Bastiaan Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[39] Alain De Cheveigné and Hideki Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[40] Matthias Mauch and Simon Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 659–663.

[41] Arturo Camacho and John G Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[42] Sira Gonzalez and Mike Brookes, "PEFAC-A pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.

[43] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello, "CREPE: A convolutional representation for pitch estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 161–165.

[44] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," in *International Conference on Machine Learning (ICML)*, 2017, pp. 11261135.

[45] Sachin Ravi and Hugo Larochelle, "Optimization as a model for few-shot learning," in *International Conference on Learning Representations (ICLR)*, 2017.

[46] Aniwat Phaphuangwittayakul, Yi Guo, and Fangli Ying, "Fast adaptive meta-learning for few-shot image generation," *IEEE Transactions on Multimedia*, 2021.

[47] Alex Nichol, Joshua Achiam, and John Schulman, "On first-order meta-learning algorithms," 2018. arXiv: 1803.02999.

[48] Jiatao Gu, Yong Wang, Yun Chen, Victor OK Li, and Kyunghyun Cho, "Meta-learning for low-resource neural machine translation," in *Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3622–3631.

[49] Kun Qian and Zhou Yu, "Domain adaptive dialog generation via meta learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2639–2649.

[50] Abiola Obamuyide and Andreas Vlachos, "Meta-learning improves lifelong relation extraction," in *Proc. Workshop on Representation Learning for NLP*, Association for Computational Linguistics, 2019, pp. 224–229.

[51] Florian Lux and Ngoc Thang Vu, "Meta-learning for improving rare word recognition in end-to-end ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 5974–5978.

[52] Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, Peng Xu, and Pascale Fung, "Learning fast adaptation on cross-accented speech

recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 1276–1280.

[53] Anton Ragni, Kate M Knill, Shakti P Rath, and Mark JF Gales, "Data augmentation for low resource languages," in *Proc. International Speech Communication Association (Interspeech)*, 2014, pp. 810–814.

[54] Tatiana Likhomanenko, Qiantong Xu, Jacob Kahn, Gabriel Synnaeve, and Ronan Collobert, "SlimIPL: Language-model-free iterative pseudo-labeling," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 741–745.

[55] Satwinder Singh, Ruili Wang, and Yuanhang Qiu, "DeepFo: End-to-end fundamental frequency estimation for music and speech signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 61–65.

[56] Satwinder Singh, Ruili Wang, and Feng Hou, "Punjabi Speech: A labeled Speech Corpus," Jul. 2023. DOI: 10.17632/sdbc8f5b77.1. [Online]. Available: https://data.mendeley.com/datasets/sdbc8f5b77/1.

[57] Satwinder, Ruili Wang, and Feng Hou, "Google-synth: A Synthesized Punjabi Speech Dataset," Jul. 2023. DOI: 10.6084/m9.figshare.23615607.v1. [Online]. Available: https://figshare.com/articles/dataset/Google-synth_A_Synthesized_Punjabi_Speech_Dataset/23615607.

[58] Satwinder Singh, Ruili Wang, and Feng Hou, "CMU-synth: A synthesized Punjabi Speech dataset," Jun. 2023. DOI: 10.6084/m9.figshare.23606697.v1. [Online]. Available: https://figshare.com/articles/dataset/_strong_CMU-synth_A_synthesized_Punjabi_Speech_dataset_strong_/23606697.

2

DEEPF0: END-TO-END FUNDAMENTAL FREQUENCY ESTIMATION

We propose a novel pitch estimation technique called DeepF0, which leverages the available annotated data to directly learns from the raw audio in a data-driven manner. F_0 *estimation is important in various speech processing and music information retrieval applications. Existing deep learning models for pitch estimations have relatively limited learning capabilities due to their shallow receptive field. The proposed model addresses this issue by extending the receptive field of a network by introducing the dilated convolutional blocks into the network. The dilation factor increases the network receptive field exponentially without increasing the parameters of the model exponentially. To make the training process more efficient and faster, DeepF0 is augmented with residual blocks with residual connections. Our empirical evaluation demonstrates that the proposed model outperforms the baselines in terms of raw pitch accuracy and raw chroma accuracy even using 77.4% fewer network parameters. We also show that our model can capture reasonably well pitch estimation even under the various levels of accompaniment noise.*

2.1 INTRODUCTION

The fundamental frequency often represented by F_0 is the lowest and predominant frequency in a complex periodic signal. It is also referred to as the pitch of the waveform

[1]. However, there is a subtle difference between F_0 and pitch since F_0 is perceived as the physical property of the audio signal, whereas pitch relates to the perceptual aspect of it. Nevertheless, outside the scope of psychoacoustics, both terms are used interchangeably in the literature [2] and in this chapter as well. Pitch estimation has been studied for almost for the last five decades because of its central importance in various domains such as speech recognition [3], speech synthesis [4], and music information retrieval [5].

There are many algorithms proposed in the past to carry out the task of pitch estimation. These algorithms can be categorized into two broad categories: digital signal processing (DSP) based approaches, and data-driven approaches. The signal processing based approaches can be further classified into the time-domain approaches (RAPT [6], YIN [7], and pYIN [8]), frequency-domain approaches (SWIPE [9] and PEFAC [10]), or hybrid (both time and frequency-domain) approaches (YAAPT [11]). Most of them use a three-stage process consisting of preprocessing of a perceived signal (usually framing and signal conditioning) followed by possible candidate search using an auto-correlation function, cross-correlation function, or cepstrum function [12]. Lastly, post-processing to track down the best possible candidates for F_0 using dynamic programming [8]. These approaches are computationally intensive, not robust in noisy environments, fail when the pitch is rapidly changing, and do not learn anything from available data [12]. On the other hand, data-driven approaches take full advantage of the available data and learn the estimation model based on the data itself. Most of the data-driven approaches are either based on traditional machine learning or deep learning approaches. Due to the availability of annotated pitch estimation datasets, and the success of deep learning models in various domains, it has become common practice to train the pitch estimation models in a data-driven manner.

Recently, numerous deep learning approaches were proposed either based on hand-crafted features or raw audio front-end. Many of the early work extracted hand-crafted features, which included constant-Q transform (CQT) [2] and spectral-domain features [13]. While extracting the acoustic features from raw audio, there is always a chance of

leaving out important features that might be crucial for pitch estimation [14]. To deal with such a situation, many researchers attempted to exploit raw waveform as the front end features in various speech-related tasks [14], [15].

In [16], a deep neural network (DNN) based pitch estimation model was proposed, which operates on raw audio. Kim et al. [12] designed a convolutional neural network (CNN) model that utilizes raw audio in the time domain and was able to outperform existing DSP based algorithms. Similarly, Dong et al. [14] proposed a convolutional residual network model to estimate pitch using raw polyphonic music. Even though these approaches perform better than DSP based algorithms, but these models still have very shallow receptive fields. Authors in [17] and [18] showed the effectiveness and applications of larger receptive in sequence modeling tasks. We intend to use that intuition in our pitch estimation task, where we can augment the network with dilation to have large memory or receptive field. We propose the DeepFo model that is based on a dilated causal temporal convolutional network (TCN). Dilation in CNN increases the receptive field exponentially, without putting a computational burden on the network in terms of the number of network parameters used [17]. To stabilize the training of deep network, we introduce residual network blocks and skip connections to the network, which can increase training efficiency, and achieve high accuracy as well [19]. The residual networks (also known as ResNet) have been successfully applied to a range of diverse areas of research [17], [20].

We evaluate our proposed model on standard datasets that include MIR-1k [21], MDB-stem-synth [22], and PTDB-TUG [23]. These datasets contain audio samples of heterogeneous timbre and characteristics. We compare our approach with state-of-the-art CREPE and SWIPE baseline algorithms, where the former is deep learning based data-driven approach and the latter is a DSP based method. Empirical evaluation demonstrates that the proposed model yields state-of-the-art results in terms of pitch accuracy. Besides, the proposed method also outperforms the baselines in the presence of a reasonable amount of background noise.

The rest of the chapter is organized as follows: Section 2.2 outlines the proposed architecture. Section 2.3 describes the experimental setup. The results are discussed in Section 2.4, followed by Section 2.5, which concludes the chapter.

2.2 PROPOSED ARCHITECTURE

The proposed approach is based on a dilated temporal convolution network. This type of architecture has been applied in the text-to-speech task (WaveNet [17]) but has not been applied in the pitch estimation task. The receptive field of the basic CNN is limited, which depends upon the linear depth of the network [18]. We can improve the receptive field by adding more convolutional layers, which will increase the receptive field linearly. However, this will put a computational burden on the network due to increased network parameters, and can also lead to a vanishing gradient problem. To address these issues dilated CNN is adopted [17]. In dilated convolutions (also referred to as convolutions with holes or atrous), we skip certain values to gain the receptive area which is usually larger than the filter size. This way we can achieve an exponentially large receptive field without even increasing the network parameters exponentially.

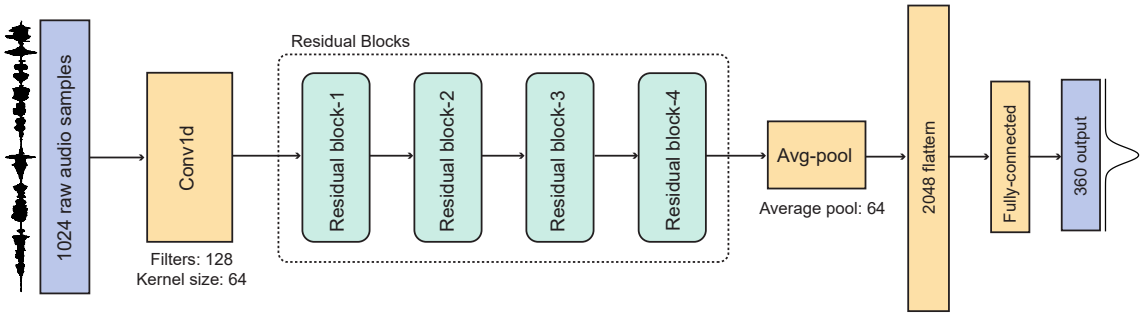

Figure 2.1: Network architecture of DeepFo.

The dilated convolutions are causal, which ensure that the current output is derived from the past outputs only and it does not look into future outputs. As we are increasing the depth of the network, which is ideal for learning robust representations, it can lead

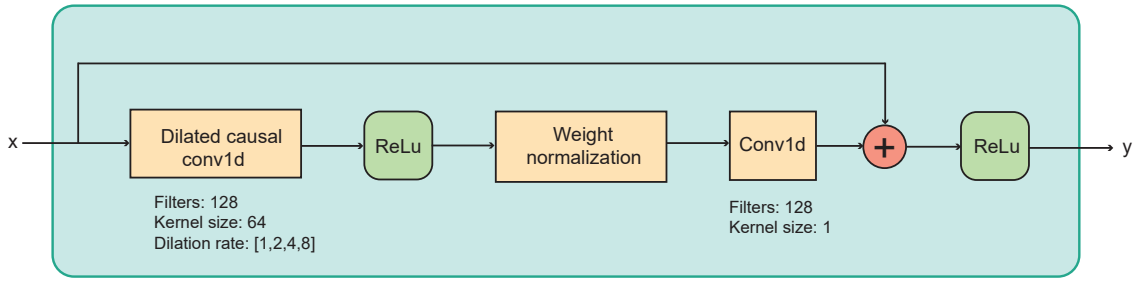

Figure 2.2: Internal view of a residual block of DeepFo.

to the classical problem of vanishing gradients. Considering this issue, we adopt residual connections that resolve the problem of gradient vanishing by making new ways to flow the gradients [17]. It also makes sure that the higher layers perform as good as the deeper layers by learning through identity mapping while training a deeper network and is expressed as the following equation:

$$y = \mathrm{ReLu}(x + \mathcal{F}(x)) \tag{2.1}$$

where x is input, and $\mathcal{F}(x)$ represents a series of transformations like convolution operations and weight normalization.

We feed our DeepFo model with an audio frame comprising 1024 samples extracted from a time-domain audio signal sampled at 16 kHz. The architecture of the model is shown in Figure 2.1. The input is passed to 1D convolution with 128 filters and 64 kernel size. The big kernel size allows it to have a wide receptive field and it also contributes to learn directly from the raw audio [14]. The output of the first convolution goes through the residual blocks. Our residual block as shown in Figure 2.2 consists of a 1D dilated causal convolution layer and a normal 1×1 convolutional layer followed by ReLu nonlinearity [24] and a weight normalization layer [25]. To achieve extended receptive fields, we use a dilation rate of $d = 1, 2, 4, 8$. In each residual block, we employ the residual/skip connections. The outcome of the last residual block is downsampled using average pooling with a pool size of 64 followed by a dense layer and sigmoid activation function. The

model uses a binary cross-entropy loss function to calculate the error between true y_i and predicted values $\hat{y}_i$. The model is optimized using the Adam optimizer [26] with a learning rate of 2e-4. Early stopping is enforced to ensure no overfitting when validation accuracy is not improving for 32 epochs. DeepFo is trained using the Nvidia Geforce RTX 2080 Ti GPU.

Following [12], the proposed model outputs a 360-dimensional vector ($c_1 - c_{360}$), which represents pitches on the logarithmic scale measured in terms of cents (a unit to measure small musical intervals). Each dimension of the output vector corresponds to the frequency bin that covers a frequency range from 32.70 Hz to 1975.5 Hz with 20 cents of intervals. The output vector estimates the Gaussian curve using the Gaussian kernel smoother [12]. Afterwards, we calculate the pitch value by taking the local weighted average of pitches closest to frequency bins having the highest peak value as shown in Equation 2.2. The resulted pitch values in cents are converted back to frequency equivalent (Hz) using Equation 2.3, where f is resulted frequency and 1200 is a single octave. f_{ref} represents the reference frequency, which is set to 10 throughout our experimentation.

$$\hat{c} = \sum_{i=m-4}^{m+4} (\hat{y}_i c_i) \Big/ \sum_{i=m-4}^{m+4} (\hat{y}_i) \qquad m = \underset{i}{\operatorname{argmax}}\, \hat{y}_i \tag{2.2}$$

$$f = f_{ref} \cdot 2^{\hat{c}/1200} \tag{2.3}$$

2.3 EXPERIMENTAL SETUP

2.3.1 *Datasets*

The proposed model is trained and evaluated on three publicly available standard datasets, namely, MIR-1k [21], MDB-stem-synth [22], and PTDB-TUG [23]. MIR-1k contains 1000 audio clips of people singing Chinese pop songs (11 males and 8 females) with pitch annotations. The right channel of the audio consists of the singing part, and the left channel holds musical accompaniment. The length of the songs is between 4 and 13 seconds,

which makes a total of 133 minutes of recordings. The MDB-stem-synth dataset contains 230 tracks resynthesized from the MedleyDB dataset [27]. It consists of 418 minutes of diverse musical instruments and singing voices. Besides that, we also use the Pitch Tracking Database provided by the Graz University of Technology (PTDB-TUG). The dataset comprises 4720 speech and laryngograph recordings of 20 native English speakers (10 males and 10 females) with a total length of 576 minutes. These three datasets have different characteristics, covering various musical instruments, singing voices, and speech signals.

2.3.2 *Methodology*

The proposed model is trained using 5-fold cross-validation with a 60/20/20 split of train, validation, and test, respectively. The split is carried out in such a way that no artist/speaker/instrument overlaps with train and test splits. To investigate the model's robustness against background noise, we trained and evaluated the model on a dataset corrupted with musical accompaniment noise. We add a musical accompaniment noise to the original clean audio for the MIR-1k dataset at different signal-to-noise ratio (SNR) levels of 20dB, 10dB, and 0dB.

2.3.3 *Baselines*

We compare the proposed model with baseline models that include CREPE [12] and SWIPE [9]. The DSP based SWIPE algorithm estimates pitch by template matching with the spectrum of the sawtooth waveform. On the other hand, CREPE is a data-driven deep learning model, which is based on CNN and trained using multiple datasets (MIR-1k [21], MDB-stem-synth [22], MedlyDB [27], RWC-Synth [8], Nsynth [28], and Bach10 [29]). We use the full version of the CREPE model (22.2M parameters) without Viterbi smoothing provided by the authors. Both the chosen models directly operate on the raw waveform in the time domain.

2.3.4 *Evaluation Metrics*

To evaluate and compare the performance of the proposed model with the baselines, we use evaluation metrics defined in [30]. Raw Pitch Accuracy (RPA) measures the percentage of audio frames where the frequency estimate is accurate within the threshold value, which is 50 cents in our case. Raw Chroma Accuracy (RCA) also measures the percentage of audio frames where the frequency estimate is correct. However, the octave errors are ignored and mapped onto a single octave. Note that both RPA and RCA ignore voicing errors.

2.4 RESULTS AND DISCUSSION

2.4.1 *Pitch Accuracy*

The proposed model is compared with state-of-the-art models that include CREPE [12] and SWIPE [9]. The results are depicted in Table 2.1. Our proposed model outperforms the baseline models in terms of raw pitch/chroma accuracies on all three datasets on clean audio. In terms of raw pitch accuracy, the DeepFo model shows 1.33% and 9.29% of relative improvement compared with CREPE and SWIPE models on the MIR-1k dataset, respectively. A similar trend can be seen in raw chroma accuracy where the proposed model outperforms both the baseline models with no added noise across all the datasets. On the MDB-stem-synth dataset, DeepFo achieves near-perfect pitch estimation with 98.38% RPA and 98.44% RCA in comparison with its baselines. We also evaluate our model on an additional dataset (PTDB-TUG), which has heterogeneous timbre (speaking voices) as compared to MIR-1k (singing voices) and MDB-stem-synth (musical instruments). Our model significantly performs better on PTDB-TUG in contrast to CREPE and SWIPE. On the PTDB-TUG dataset, CREPE performs worst of all the models. This could be attributed to the fact the model provided by the authors was not trained on the PTDB-TUG dataset.

Table 2.1: Average raw pitch accuracy and raw chroma accuracy and their standard deviation (±) tested on three different test datasets.

Model	**#Params**	**Metrics**	**Datasets**		
			MIR-1k	**MDB-stem-synth**	**PTDB-TUG**
SWIPE	-	RPA (%)	88.73 ±5.43	92.84 ±9.59	87.74 ±7.17
		RCA (%)	89.24 ±5.28	93.83 ±7.69	88.93 ±6.12
CREPE	22.2M	RPA (%)	96.51 ±3.23	97.22 ±4.12	78.18 ±10.07
		RCA (%)	96.84 ±2.56	97.55 ±3.43	79.81 ±9.39
DeepFo	5M	RPA (%)	**97.82 ± 3.34**	**98.38 ± 2.97**	**93.14 ± 3.32**
		RCA (%)	**98.28 ± 1.94**	**98.44 ± 2.87**	**93.47 ± 3.41**

Table 2.2: Average raw pitch accuracy and raw chroma accuracy and their standard deviation (±) on the MIR-1k dataset with added noise on various levels of SNR.

Model	**Metrics**	**Noise Profile**			
		Clean	**20dB**	**10dB**	**0dB**
SWIPE	RPA (%)	88.73 ± 5.43	84.45 ± 5.64	59.78 ± 11.58	32.04 ± 11.84
	RCA (%)	89.24 ± 5.28	85.31 ± 5.19	62.85 ± 11.07	37.31 ± 12.93
CREPE	RPA (%)	96.51 ± 3.23	96.49 ± 3.32	**95.11 ± 4.65**	**84.92 ± 10.70**
	RCA (%)	96.84 ± 2.56	96.96 ± 2.63	96.18 ± 3.35	**87.85 ± 8.82**
DeepFo	RPA (%)	**97.82 ± 3.34**	**97.39 ± 3.76**	94.77± 6.03	79.52± 14.0
	RCA (%)	**98.28 ± 1.94**	**98.09 ± 2.10**	**96.35 ± 3.72**	84.37 ± 10.71

DeepFo also shows more stability as it demonstrates consistently lower variance in pitch accuracy next to its baselines.

2.4.2 *Performance in Noisy Conditions*

Ideally, even in noisy environments, a model can still perform reasonably well. We put our proposed model into such testing scenarios by contaminating the signals with musical accompaniments at various levels of SNR on the MIR-1k dataset and results are presented in Table 2.2. In general, our proposed method achieves higher RPA and RCA as compared with the baselines under 10dB and 20dB noise. However, CREPE performs better when SNR is as low as 0dB. On the other hand, the performance of the SWIPE model is worst under 10dB and 0dB noise. Overall, we can say that DeepFo achieves better performance

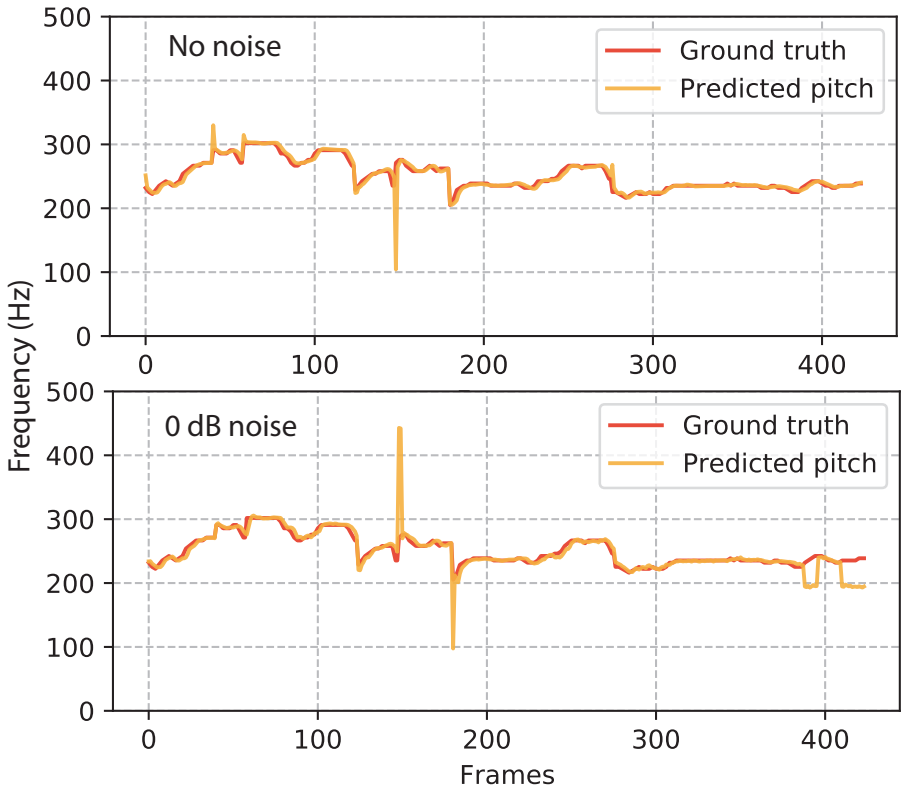

Figure 2.3: The estimated pitch trajectories of DeepFo in comparison with ground truth under clean (top) and 0dB noise (bottom). Under a noise scenario, DeepFo produces near perfect pitch estimation, while under noise there are few errors.

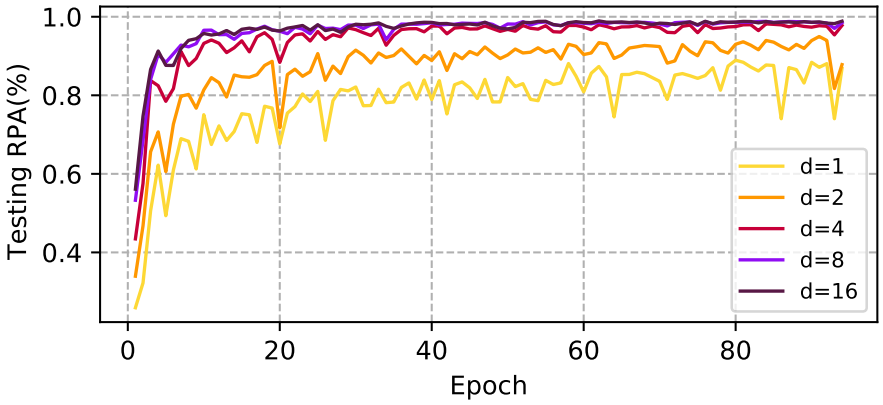

Figure 2.4: Evaluation results of the proposed model with different dilation rates on the MDB-stem-synth dataset. Dilation rate $d = 8$ shows the best results.

under low to moderate noise scenarios and CREPE works better under moderate to high noise scenarios.

2.4.3 *Model Analysis*

Our proposed model is more efficient in terms of the number of parameters used for training, which is around 5 million as compared with the CREPE model, which uses 22.2 million parameters. Thus, our DeepF0 model with 77.4% fewer parameters can still outperform the CREPE model. Further, we analyze the role of the receptive field in the task of pitch estimation. We observe that a larger receptive field indeed improves the overall performance of the model. Our experiments are with dilation rate $d = 1, 2, 4, 8, 16$ on the MDB-stem-synth dataset. The raw pitch accuracies are depicted in Figure 2.4. DeepF0 with $d = 1$, which is basically a standard CNN, only achieves about 86.40% of RPA and 86.55% of RCA. Further $d = 2, 4$ improve the performance and able to achieve similar results as the CREPE model. However, the results are not very stable in terms of variance, which ranges from ± 6.85 to ± 18.11 on dilation rate 1 to 4. We find that $d = 8$ gives the

Table 2.3: Evaluation results of the ablation study of our DeepFo model. Without residual connections accuracy of the model decrease. With the dropout layer included in the residual blocks, the performance more or less remains the same.

Models	**Metrics**	**Dataset**
		MIR-1k
DeepFo baseline	RPA(%)	97.82±3.34
	RCA(%)	98.28±1.94
w/o residual connections	RPA(%)	97.54±3.61
	RCA(%)	97.89±2.42
w/ dropout	RPA(%)	97.83±3.28
	RCA(%)	98.24±2.18

best results in terms of RPA (98.38% ±2.97), RCA (98.44% ±2.87), and variance. DeepFo does not show any performance improvements beyond the dilation rate of $d = 8$.

We also analyze some of the design choices that we made while constructing the architecture of DeepFo and the results of our ablation study are presented in Table 2.3. We find that residual connections are making a significant difference when it comes to stabilizing and speeding up the training process. Not only model converges fast, but it also helps in achieving higher performance with low variance. Besides this, we do not use dropout layers throughout our network as we observe that these layers seem redundant in the presence of the weight normalization layer and have almost zero effect on the final results.

2.5 CONCLUSIONS

In this chapter, we propose a data-driven approach based on dilated temporal convolutional networks for the task of fundamental frequency estimation. The proposed DeepFo model operates on raw audio and outputs the pitch estimation. The experimental results

performed on three heterogeneous datasets (singing voice vs speaking voices vs musical instruments) reveal that DeepFo outperforms existing baseline models like CREPE and SWIPE. Our proposed model not only achieves better results but also uses 77.4% fewer parameters as compared with the CREPE model. Our model is also able to perform reasonably well under low to moderate noise. Further, we gain crucial insights about the large receptive field, which was not there in earlier models. We find that the length of the receptive field of the network is very significant in pitch estimation, which aids in achieving excellent results with consistently low variance. In the future, we would like to improve the noise robustness of our proposed model by introducing changes in architectural design, data augmentation, and speech enhancement techniques [31]. The performance can be further improved by post-processing the pitch estimate through temporal smoothing techniques.

In the forthcoming chapter, we delve into the domain of meta-learning for low-resource speech recognition. Acknowledging the necessity for efficient training algorithms in low-resource settings, this chapter introduces the Model-Agnostic Meta-Learning (MAML) algorithm. As we navigate the challenges associated with applying MAML to our speech recognition task, we identify areas that warrant improvement to attain more consistent training outcomes. In light of this, we propose a multi-step loss (MSL) function with the goal of stabilizing the training process and enhancing performance across diverse languages with limited data.

REFERENCES

[1] Donn Morrison, Ruili Wang, and Liyanage C De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech communication*, vol. 49, no. 2, pp. 98–112, 2007.

[2] Beat Gfeller, Christian Frank, Dominik Roblek, Matt Sharifi, Marco Tagliasacchi, and Mihajlo Velimirovic, "SPICE: Self-supervised pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[3] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *IEEE international conference on acoustics, speech and signal processing*, 2014, pp. 2494–2498.

[4] M Kiran Reddy and K Sreenivasa Rao, "Excitation modelling using epoch features for statistical parametric speech synthesis," *Computer Speech & Language*, vol. 60, p. 101 029, 2020.

[5] Sangeun Kum and Juhan Nam, "Joint detection and classification of singing voice melody using convolutional recurrent neural networks," *Applied Sciences*, vol. 9, no. 7, p. 1324, 2019.

[6] David Talkin and W Bastiaan Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[7] Alain De Cheveigné and Hideki Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[8] Matthias Mauch and Simon Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 659–663.

[9] Arturo Camacho and John G Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[10] Sira Gonzalez and Mike Brookes, "PEFAC-A pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.

[11] Kavita Kasi, "Yet another algorithm for pitch tracking:(YAAPT)," Ph.D. dissertation, Old Dominion University, 2002.

[12] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello, "CREPE: A convolutional representation for pitch estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 161–165.

[13] Kun Han and DeLiang Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 2158–2168, 2014.

[14] Mingye Dong, Jie Wu, and Jian Luan, "Vocal pitch extraction in polyphonic music using convolutional residual network," in *20th Annual Conference of the International Speech Communication Association*, 2019, pp. 2010–2014.

[15] Tara N Sainath, Ron J Weiss, Andrew Senior, Kevin W Wilson, and Oriol Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *16th Annual Conference of the International Speech Communication Association*, 2015, pp. 1–5.

[16] Prateek Verma and Ronald W Schafer, "Frequency estimation from waveforms using multi-layered neural networks.," in *Proc. International Speech Communication Association (Interspeech)*, 2016, pp. 2165–2169.

[17] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016. arXiv: 1609.03499.

[18] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018. arXiv: 1803.01271.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] Yan Tian, Xun Wang, Jiachen Wu, Ruili Wang, and Bailin Yang, "Multi-scale hierarchical residual network for dense captioning," *Journal of Artificial Intelligence Research*, vol. 64, pp. 181–196, 2019.

[21] Chao-Ling Hsu and Jyh-Shing Roger Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, 2009.

[22] Justin Salamon, Rachel M Bittner, Jordi Bonada, Juan J Bosch, Emilia Gómez Gutiérrez, and Juan Pablo Bello, "An analysis/synthesis framework for automatic fo annotation of multitrack datasets," in *18th International Society for Music Information Retrieval Conference*, 2017.

[23] Gregor Pirker, Michael Wohlmayr, Stefan Petrik, and Franz Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Proc. International Speech Communication Association (Interspeech)*, 2011, pp. 1509–1512.

[24] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *27th international conference on machine learning (ICML)*, 2010, pp. 807–814.

[25] Tim Salimans and Durk P Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Advances in neural information processing systems*, 2016, pp. 901–909.

[26] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2014. arXiv: 1412.6980.

[27] Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello, "MedleyDB: A multitrack dataset for annotation-intensive MIR research," in *International Society for Music Information Retrieval (ISMIR)*, vol. 14, 2014, pp. 155–160.

[28] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in *International Conference on Machine Learning (ICML)*, 2017, pp. 1068–1077.

[29] Zhiyao Duan, Bryan Pardo, and Changshui Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2121–2133, 2010.

[30] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel, "Mir_eval: A transparent implementation of common MIR metrics," in *15th International Society for Music Information Retrieval Conference*, 2014.

[31] Yuanhang Qiu and Ruili Wang, "Adversarial latent representation learning for speech enhancement," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 2662–2666.

3

IMPROVED META-LEARNING FOR LOW-RESOURCE SPEECH RECOGNITION

We propose a new meta-learning based framework for low-resource speech recognition that improves the previous model agnostic meta-learning (MAML) approach. The MAML is a simple yet powerful meta-learning approach. However, the MAML presents some core deficiencies such as training instabilities and slower convergence speed. To address these issues, we adopt multi-step loss (MSL). The MSL aims to calculate losses at every step of the inner loop of MAML and then combines them with a weighted importance vector. The importance vector ensures that the loss at the last step has more importance than the previous steps. Our empirical evaluation shows that MSL significantly improves the stability of the training procedure and it thus also improves the accuracy of the overall system. Our proposed system outperforms MAML based low-resource ASR system on various languages in terms of character error rates and stable training behavior.

3.1 INTRODUCTION

Modern deep learning based end-to-end (E2E) models have lately become extremely popular in the speech community [1] and have achieved a significant milestone in terms of performance. These systems have been deployed under commercial domains as they have shown consistently lower word error rates that are close to 1-2% [2]. The modern ASR sys-

tems are mostly trained in E2E fashion without requiring resources like a pronunciation dictionary and a language model as separate modules. These systems are able to achieve such a high degree of accuracy mainly because they are trained on various large vocabulary datasets. However, these E2E systems tend to perform much worse for the languages that do not have such large quantities of annotated data.

Among roughly 7000 languages spoken across the world, there are only around 100 languages that have well-established speech recognition systems [3]. The rest of the languages are considered low-resource languages because they do not have a huge amount of annotated speech data, strong pronunciation dictionaries, and a huge collection of unpaired texts. A lot of progress has been made in low-resource speech recognition, which includes efforts like transfer learning [4] and multilingual training [5]. Recently, a new paradigm, meta-learning has been explored for low-resource speech recognition [6]. Meta-learning (also known as learning to learn) is a machine learning technique, where learning is done on two levels. On one level (inner loop) model acquires task specific knowledge, whereas the second level (outer loop) facilitates task across learning [7].

Previously, Hsu et al. [6] proposed a meta-learning framework based on the MAML approach for ASR for low-resource language. The proposed framework outperformed the no-pre-training and multi-lingual training settings. Similarly, Winata et al. [8] incorporated the MAML approach for the few-shot accent adaptation task for English. The MAML approach in general is a very straightforward and powerful approach. However, it is prone to numerous problems, including unstable training and slow convergence speeds. These issues also impact the generalizability of the model. Thus, to deal with these issues, in this work, we adopt the multi-step loss [7], which is introduced to stabilize the meta-training procedure. The meta-training approach with multi-step loss calculates the inner loss after every inner step updates and later computes the weighted sum of all the inner losses.

We evaluated our proposed approach on 10 different languages present in the Common Voice v7.0 dataset. All these languages are represented in the form of a low-resource

setting where the language data ranges from 0.5 hours to 300 hours. We find that our approach indeed improves the training instabilities of the MAML approach, which in turn improves the overall accuracy of the model.

The rest of the chapter is organized as follows. We introduce the important background and related work in Section 3.2. We describe our proposed approach in Section 3.3. Following that, in Section 3.4, we discuss experimental setup including datasets and methodology. Further, Results and discussions are presented in Section 3.5, followed by conclusions in Section 3.6.

3.2 RELATED WORK

3.2.1 *Meta-Learning*

Meta-learning is not a new idea but has begun to gain attention in recent times. Recently, in the context of deep learning, meta-learning comes into the limelight due to its wide range of applications and advantages. Meta-learning helps to generalize to various tasks faster with few steps and examples. Literature suggests the application of meta-learning in two ways where the first is learning a better initialization of network parameters [9] and the second is learning a strategy or procedure for updating the parameters of the network [10], [11].

Meta-learning has been applied to a range of research domains including various computer vision tasks, natural language processing, and recently automatic speech recognition. In the computer vision area, meta-learning has been exploited for the few-shot image classification task [12], object detection [13] and video generation [14]. In the natural language processing domain, meta-learning has shown promising results in neural machine translation (NMT) for resource constraint languages [15]. Apart from this, recently researchers have tried meta-learning for speech processing tasks, such as automatic speech

recognition [6], speaker adaptation [16], [17] and recognition [18], cross-lingual [19], and cross-accent adaption [8].

3.2.2 *Low-Resource Speech Recognition*

The development of a speech recognition system for low-resource languages has been a very active research area for the past few years. The regular E2E ASR systems designed for resource-rich languages seem not to work for low-resource languages due to the lack of annotated speech data or other resources. There have been many attempts made to alleviate the scarcity of labeled speech data. These efforts include, speech data augmentation [20], transfer learning [4], multilingual [5], cross-lingual [19], and multi-task learning [6]. Recently, the unsupervised cross-lingual wave2vec 2.0 XLSR model [21] has shown a huge performance boost compared to other previous state-of-the-art models. Further, there have been recent attempts to explore a new research direction of meta-learning for low-resource languages. The idea is to extract meta parameters learned over multiple source languages and then bootstrap these learned meta parameters to fine-tune on the target languages. The whole process can be seen as learning a model that can perform fast adaptation to target languages with few epochs and data samples. As fine-tuning requires few training samples, this process of meta-learning is totally aligned with our proposed framework of ASR for low-resource languages.

3.3 PROPOSED SYSTEM

Our proposed system consists of two core components. The first is an ASR model that acquires language specific knowledge and the second is a multi-step loss based model agnostic meta-learning algorithm.

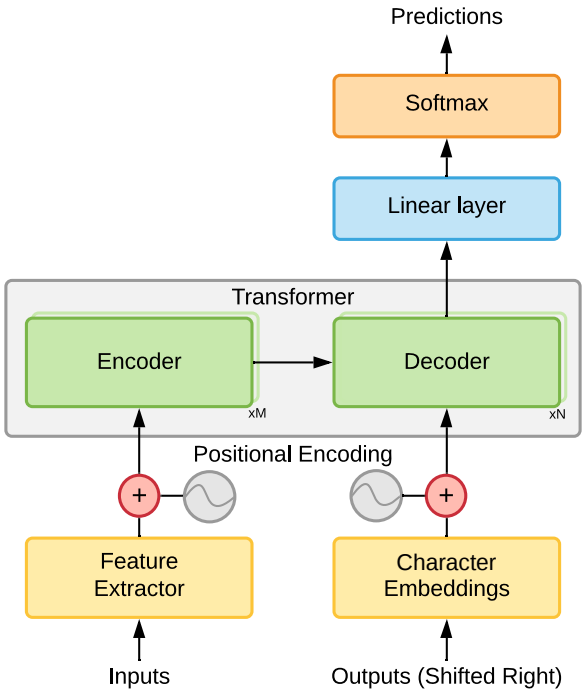

Figure 3.1: The Transformer model for ASR

3.3.1 *The ASR Model*

For our proposed system, we adopt the Transformer ASR model [22] as our language specific model. The Transformer model is a sequence-to-sequence model based on the encoder-decoder architecture. The proposed model extracts the input features using the learnable VGG based convolutional neural network (CNN) model [23]. The input embeddings produced by the feature extractor are then fed to the encoder module through the positional encoding setup. The positional encoding setup generates a vector that is served as context for the symbols. Afterward, the outputs of the encoder module are passed on to the decoder module, where a multi-head attention mechanism is employed on these encoder outputs. The attention mechanism applies masking in the decoder block to restrict the attention layer from attending to any future tokens. Finally, the output of the decoder block goes through a linear and softmax layer and generates the predictions. The entire

training process is optimized by maximizing the log probability using next step prediction based on the last output token. In the following equation, x, y_i and y'_{i-1} are the input character, next predicted character, and true label of the last character, respectively.

$$\max_{\theta} \sum_{i} \log P(y_i|x, y'_{i-1}; \theta) \tag{3.1}$$

3.3.2 *Meta-Learning Setup*

In general, our proposed meta-learning setup aims at learning the initial parameters for the model in a way that it can be quickly adapted to new languages with a fewer number of gradient descent steps. We adopt multi-step loss from MAML++ [7] procedure over standard MAML as MAML tends to have unstable training procedure. This can affect the overall speed of convergence and also has a negative impact on the accuracy of the model. Figure 3.2 shows the computational graph for both MAML and MAML with multi-step loss. We select support samples (x_S, y_S) from the training set and target samples (x_T, y_T) from the validation set of our source language set. Notably, each language within the source language set is associated with its distinct support and target sets, which are utilized for training and evaluation purposes. We start optimizing the inner loop by initializing our ASR model f with $\theta_0 = \theta$. Afterwards, the ASR model produces logits $f(x_S, \theta_{i-1})$ by using samples from training set and parameters θ_{i-1}. Here i represents i^{th} step of total N steps. In the next step, loss L^S_{i-1} is computed over true labels y_S and logits. Further, the L^S_{i-1} is utilized to update the current parameters of the model.

MAML with Multi-Step Loss (MSL MAML) represents an adaptation of the conventional MAML approach, wherein a modified loss function is introduced within the inner optimization loop. The key distinction lies in the computation and utilization of this loss to update the model's parameters during the inner loop. Instead of using θ_N parameters for computing target set loss, our approach goes on using θ_i parameters. After completing the inner loop, we obtain N target set losses as in Equation 3.2, which can be seen as a multi-step loss, where w_i is the weight of step i and specify the importance of per step

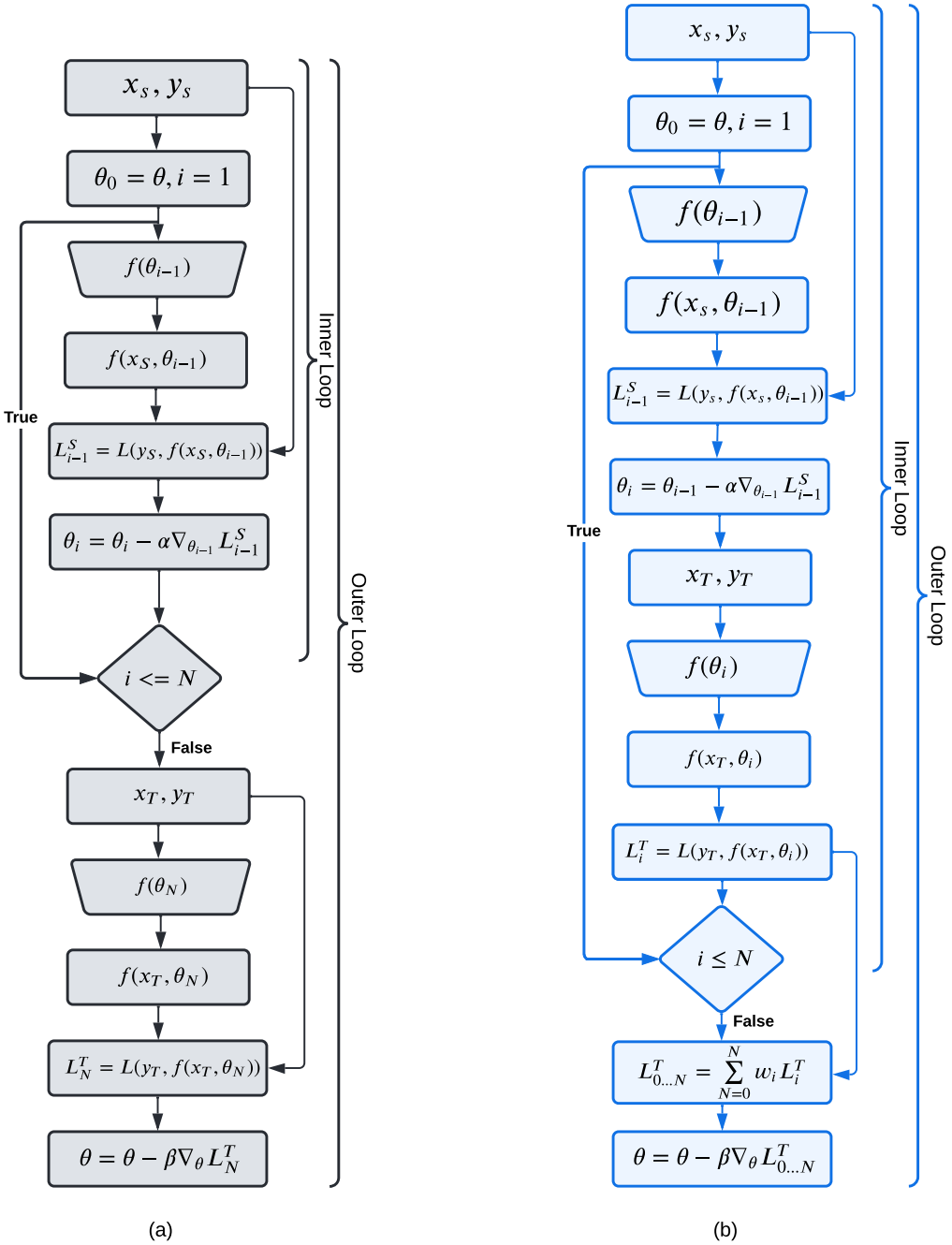

Figure 3.2: MAML (a) vs MAML with MSL (b) (adopted from [7])

target loss. Initially, all the losses have approximately the same importance, while later in training more importance is given to the losses on later steps. This way the model gradually steps towards the MAML loss, ensuring there is no issue of gradient degradation. Finally, these losses are combined together using a weighted sum of per-step losses. The combined weighted loss is then used to update the outer loop parameters θ. The advantage of calculating per step loss is reducing the gradient vanishing and exploding problem of the original MAML. Following [8] and [6], we only compute first order approximation of θ.

$$L^T_{0\ldots N} = \sum_{N=0}^{N} w_i L^T_i \tag{3.2}$$

3.4 EXPERIMENTAL SETUP

3.4.1 *Dataset*

For our experiments, we choose Common Voice dataset version 7 [24]. The Common Voice dataset is a publicly crowdsourced dataset that comprises a wide array of languages, encompassing both resource-rich and low-resource languages. It is organized with predefined train, validation, and test set splits. Consequently, for our research, we utilized the default splits provided by the dataset to ensure consistency and standardized evaluation. We select 10 low-resource languages and the description is represented in Table 3.1. Some of the languages are very low-resource having just a few hours of data. The audio from all the languages is downsampled to 16 kHz and labels are preprocessed to remove any kind of special symbols.

Table 3.1: The selected low-resource languages from the Common Voice dataset v7.0 and the total amount of speech data in terms of hours.

ID	Languages	Hours
ar	Arabic	85
as	Assames	1
hi	Hindi	8
lt	Lithuanian	16
mn	Mongolian	12
or	Odia	0.94
fa	Persian	293
pa-IN	Punjabi	1
ta	Tamil	198
ur	Urdu	0.59
	Total	615.53

3.4.2 *Methodology*

Our model receives a spectrogram as an input. These spectrogram inputs then go through a VGG based 6-layered CNN feature extractor. We use 2 encoder layers and 4 decoder layers with 8 multi-head attention layers. Our model produces input and output of dimension 512, whereas the inner layer has 2048 dimensions. We set the dropout value to 0.1 and the keys and values dimensions to 64. We multilingually pre-train our model for 100K iterations on the source language set. The primary purpose for using a source language set is to allow the model to learn from various languages during the meta-training phase. Exposing the model to multiple languages during meta-training improves its ability to generalize to new and unseen languages. We put together 3 source language sets where one set includes **fa, ar** and, **ta**. The other set has **ar, mn** and, **lt**, and the last set consists of **or, pa-IN, hi, ur,** and **as**. During the fine-tuning phase, we fine-tune the model

Table 3.2: The average experimental results in terms of character error rate (CER in %) on 5 target languages. We do not fine-tune our model on the languages that are present in the pre-train source language set. These cells are represented by a hyphen (-).

Pre-train languages	**Fine-tune**									
	Hindi		**Mongolian**		**Persian**		**Arabic**		**Tamil**	
	MAML	Our	MAML	Our	MAML	Our	MAML	Our	MAML	Our
[fa, ar, ta]	70.51	**70.47**	61.05	**60.52**	-	-	-	-	-	-
[ar, mn, lt]	71.61	**71.37**	-	-	47.96	**45.45**	-	-	40.96	**35.17**
[or, pa-IN, hi, ur, as]	-	-	62.26	**59.50**	52.42	**52.41**	**36.00**	36.09	**45.96**	46.60

on our target languages (**hi, mn, fa, ar** and **ta**) one by one for 10 epochs. The model is then evaluated on a test set of target languages using beam search with a beam size of 5.

3.5 RESULTS AND DISCUSSION

3.5.1 *Model's Accuracy Analysis*

We evaluate the performance of our proposed MSL MAML approach on 10 languages from the Common Voice dataset. Our proposed approach showcases consistent improvement in character error rates (CER in %) over the standard MAML approach. The detailed results are presented in Table 3.2. On source languages set **[fa, ar, ta]** our approach achieves 70.47% and 60.52% of CER on Hindi and Mongolian languages, respectively. Our proposed model shows around 1% improvement over standard MAML on the Mongolian language. On set **[ar, mn, lt]** our approach slightly performs better than MAML on the Hindi language. On the same set, our approach outperforms the current MAML approach with 5.23% and 14.13% of relative improvement on Persian and Tamil languages, respectively.

Further, the Mongolian language demonstrates 4.43% relative improvement over MAML on set **[or, pa-IN, hi, ur, as]**. Mostly, on this pre-train language set both MAML and our approach report similar results on Persian and Arabic languages. Interestingly, the MAML marginally outperforms our approach on the Tamil language. Overall, our approach shows consistent improvements across all the pre-train sets, where excellent performance is observed on **[ar, mn, lt]**.

3.5.2 *Training Performance Analysis*

The multi-step loss indeed stabilizes the training process of MAML as shown in Figure 3.3. The primary driver of instability in MAML is the gradient degradation problem while

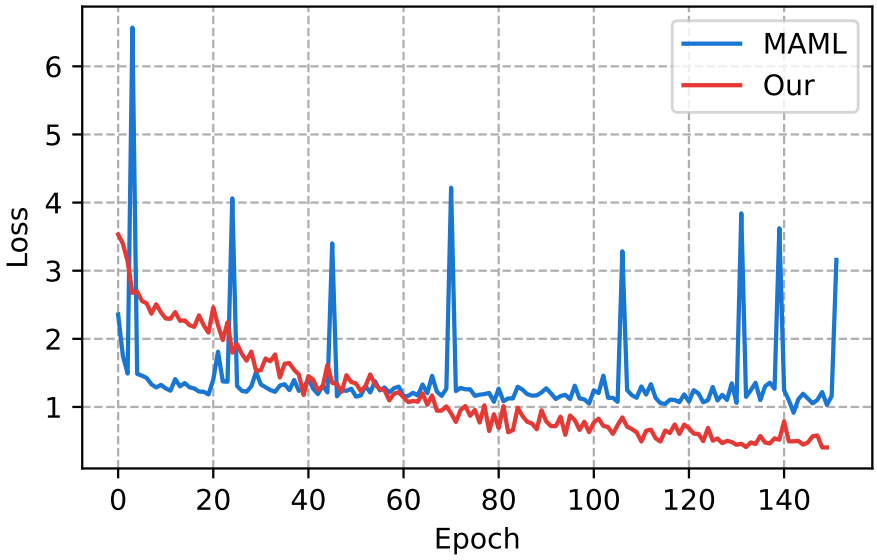

Figure 3.3: Training curve of MAML vs our approach. The training loss curve for MAML shows unstable peaks whereas our approach shows a more consistent loss curve.

training deep network [7]. Our approach resolved this issue using multi-step loss, where the model is evaluated at each step against its validation set. Further, the importance weight vector also makes sure later step loss has more importance. It also improves the convergence speed of the model as shown in Figure 3.3.

3.6 CONCLUSIONS

In this chapter, we propose a multi-step loss based meta-learning approach for speech recognition for low-resource languages. The proposed system improves the inner loop optimization for the MAML algorithm, which results in a more stabilized training procedure. Our empirical results show that multi-step loss indeed improves the overall training procedure and also has a positive impact on the accuracy of the model. Apart from this, our model also trains faster as compared to MAML. In the future, we plan to conduct more experiments with more low-resource languages. We would extend our experiments with

different combinations of languages on the basis of their phonetic structures, geographic areas, and language family.

With insights gained from the meta-learning approach in this chapter, Chapter 4 introduces an ingenious self-supervised framework to address data scarcity in low-resource speech recognition. Recognizing the lack of annotated datasets for the Punjabi language, we present the development of three new datasets through synthesis: Google-synth, CMU-synth, and our self-recorded Punjabi Speech dataset. The chapter delves into the efficacy of these synthesized datasets by adopting pre-trained cross-lingual wav2vec models. Through empirical analysis, we aim to demonstrate the potential of synthesized data in improving error rates on real speech labeled datasets.

REFERENCES

[1] Satwinder Singh, Ruili Wang, and Yuanhang Qiu, "DeepFo: End-to-end fundamental frequency estimation for music and speech signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 61–65.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[3] Ekapol Chuangsuwanich, "Multilingual techniques for low resource automatic speech recognition," Ph.D. dissertation, MIT, Cambridge, United States, 2016.

[4] Yuan-Jui Chen, Tao Tu, Cheng chieh Yeh, and Hung-Yi Lee, "End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning," in *Proc. International Speech Communication Association (Interspeech)*, 2019, pp. 2075–2079.

[5] Shiyu Zhou, Shuang Xu, and Bo Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," 2018. arXiv: 1806.05059.

[6] Jui-Yang Hsu, Yuan-Jui Chen, and Hung-yi Lee, "Meta learning for end-to-end low-resource speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7844–7848.

[7] Antreas Antoniou, Harrison Edwards, and Amos Storkey, "How to train your MAML," in *International Conference on Learning Representations (ICLR)*, 2018.

[8] Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, Peng Xu, and Pascale Fung, "Learning fast adaptation on cross-accented speech recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 1276–1280.

[9] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel, "Meta-learning with temporal convolutions," 2017. arXiv: 1707.03141.

[10] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," in *International Conference on Machine Learning (ICML)*, 2017, pp. 11261135.

[11] Jake Snell, Kevin Swersky, and Richard Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4077–4087.

[12] Sachin Ravi and Hugo Larochelle, "Optimization as a model for few-shot learning," in *International Conference on Learning Representations (ICLR)*, 2017.

[13] Aniwat Phaphuangwittayakul, Yi Guo, and Fangli Ying, "Fast adaptive meta-learning for few-shot image generation," *IEEE Transactions on Multimedia*, 2021.

[14] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro, "Few-shot video-to-video synthesis," in *International Conference on Neural Information Processing Systems*, 2019, pp. 5013–5024.

[15] Jiatao Gu, Yong Wang, Yun Chen, Victor OK Li, and Kyunghyun Cho, "Meta-learning for low-resource neural machine translation," in *Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3622–3631.

[16] Ondej Klejch, Joachim Fainberg, and Peter Bell, "Learning to adapt: A meta-learning approach for speaker adaptation," in *Proc. International Speech Communication Association (Interspeech)*, 2018, pp. 867–871.

[17] Ondřej Klejch, Joachim Fainberg, Peter Bell, and Steve Renals, "Speaker adaptive training using model agnostic meta-learning," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 881–888.

[18] Seong Min Kye, Youngmoon Jung, Hae Beom Lee, Sung Ju Hwang, and Hoirin Kim, "Meta-Learning for Short Utterance Speaker Recognition with Imbalance Length Pairs," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 2982–2986.

[19] Wenxin Hou, Yidong Wang, Shengzhou Gao, and Takahiro Shinozaki, "Meta-adapter: Efficient cross-lingual adaptation with meta-learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7028–7032.

[20] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2019, pp. 2613–2617.

[21] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2021, pp. 2426–2430.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[23] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[24] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

4

IMPROVING SPEECH RECOGNITION WITH SYNTHESIZED DATASETS

In this chapter, we develop an end-to-end (E2E) speech recognition system for the low-resource Punjabi language. Punjabi is a widely spoken language across the globe by millions of speakers. However, there is a lack of annotated (text and speech pair) datasets for Punjabi speech recognition. To tackle the issue of scarce data, we synthesize two new datasets, namely, Google-synth and CMU-synth. We empirically test the effectiveness of the synthesized datasets on three versions of the cross-lingual wav2vec models including XLSR-53, XLS-R-300M, and XLS-R-1B. Our empirical evaluation demonstrates that our synthesized datasets and language model, the cross-lingual wav2vec models achieve significant results on real labeled datasets such as Common Voice. On average, we observe 45.98% and 59.14% of relative improvement in terms of character error rates (CER) and word error rates (WER), respectively on the Common Voice Punjabi dataset in comparison to the baseline. We also show that even 5 hours of our synthesized data and language model can significantly improve (relative improvement of 34.35% CER and 40.96% WER) the results on the Common Voice dataset. The web-based demo of our work is available at https://salp.massey.ac.nz.

4.1 INTRODUCTION

The contemporary end-to-end (E2E) automatic speech recognition (ASR) models [1] have shown promising performance compared to classical hybrid ASR systems (e.g., Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) [2], [3] and hybrid Deep Neural Network-Hidden Markov Model (DNN-HMM) [4]–[8]. These E2E models combine different components of traditional systems (acoustic, lexicon, and language model) and jointly train a single powerful model [1], [9], [10]. However, the core requirement for E2E models to work well is to train an ASR model on very large datasets, containing tens of thousands of hours of labeled speech data. This requirement is easy to fulfill for resource-rich languages such as English, German or Spanish. However, most languages of the world are low-resource languages that do not have such amount of labeled data, which presents numerous challenges to obtain decent quality transcriptions.

In the past few years, several techniques have been proposed to overcome the challenges presented by limited linguistic resources (labeled data, pronunciation lexicon, and raw language text). One of the common techniques involve multilingual/cross-lingual training of an ASR model on various languages [11]–[14]. Further, transfer learning has been used where the model is pre-trained on high-resource languages and then fine-tuned on low-resource languages [15]–[17]. Similarly, meta-learning has been explored for low-resource languages recently [18], [19]. Besides, data augmentation methods have also been employed to solve the issue of data scarcity [20]–[22]. Apart from these, more recently, self-supervised learning based wav2vec models have shown a great deal of improvement in low-resource scenarios [23]–[26]. All these advancements significantly improved the performance of ASR models for low-resource languages.

There are around 7000 languages widely spoken all across the globe. However, only as few as 100 languages have well-established ASR systems [27]. The rest of the languages are considered to be low-resource languages and the Punjabi language is also one of them. The Punjabi language belongs to an Indo-Aryan language family, spoken by roughly 110+

million native speakers across India and Pakistan. The Punjabi language is native to India and Pakistan, but Punjabi speakers are spread all across the world. Punjabi is written in two scripts Gurmukhi and Shahmukhi, where Gurmukhi is used in the Indian region and Shahmukhi in Pakistan. Punjabi is a very distinctive language in the Indo-Aryan language family as it uses different lexical tones (i.e., low, mid, and high tones). Although it is spoken by a large population, it does not have the required resources needed for the standard speech recognition recipe.

From the perspective of ASR systems, the Punjabi language is not well explored. Few attempts had been made in recent years to develop a robust ASR system for the Punjabi language [28]. However, most of them are based on a traditional pipeline, which uses complex Hidden Markov Model (HMM) based models that were not reliable in terms of accuracy and performance [29]. Apart from this, a major challenge while developing an ASR system for the Punjabi language is that there is no standard dataset available in the public domain. In most previously published papers, authors have developed their own small experimental datasets, which were not available in the public domain. Recently, Mozilla's Common Voice project released a Punjabi dataset [30]. However, our literature survey suggested that it has not been used for ASR applications so far.

Therefore, in this chapter, we propose a framework where we first generate more data using speech synthesis and later use this data in conjunction with real labeled speech datasets to fine-tune cross-lingual wav2vec 2.0 models (i.e., XLSR-53, XLS-R-300M, and XLS-R-1B). Our empirical investigation demonstrates that our synthesized datasets significantly reduce error rates on three of the real labeled speech datasets that include Common Voice, Punjabi Speech, and 50Languages datasets. At the same time, we are building a diverse Punjabi Speech dataset. The Punjabi Speech dataset has been used in this work and shows promising results.

We made the following key contributions with this work:

- To the best of our knowledge, this is the first work to integrate an E2E ASR model for the Punjabi language.

- We synthesized two new Punjabi datasets containing a combined 130K utterances and 208 hours of data.
- We also compiled the in-house self-recorded dataset called Punjabi Speech. This dataset is continuously updated with more data, which eventually be available for public use.

The rest of the chapter is organized as follows: In Section 4.2, we give an overview of the literature covering recent trends in low-resource ASR systems and also previous studies on Punjabi ASR systems. Further, in Section 4.3, we introduce our proposed framework including details about our synthesized datasets and ASR models for this work. Later, we describe the other datasets and methodology used for experimentation in Section 4.4. In Section 4.5, we present the detailed result analysis. Lastly, conclusions are outlined in Section 4.6.

4.2 RELATED WORK

4.2.1 *Low-resource Automatic Speech Recognition*

There has been a smooth transition from the traditional Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) [2], [3] and hybrid Deep Neural Network-Hidden Markov Model (DNN-HMM) [4]–[8] to pure E2E deep learning based models. So far, several studies have attempted to alleviate the issue of limited linguistic resources (i.e., annotated speech data, lexicon, and unpaired text) associated with speech recognition systems for low-resource languages. These studies include speech data augmentation [20]–[22], [31], [32], multilingual and cross-lingual systems [11]–[14], transfer learning systems [15]–[17], meta learning approaches [18], [19], and semi/unsupervised approaches [23]–[26].

Data augmentation is a very attractive choice to circumvent the scarcity of limited data in low-resource scenarios. Over the few years, many studies presented unique augmentation procedures. [33] presented very simple speed perturbation augmentation, where

additional data is obtained by changing the speech of the raw audio. While other studies focused on pitch perturbations and noise-based augmentations [34], [35]. Further, [36] proposed vocal tract length perturbation (VLTP) based augmentation. Recently, [20], proposed a SpecAugment procedure, which worked by masking frequency and time warping in log Mel spectrogram features.

Besides, some studies also focused to use text-to-speech (TTS) synthesis based augmentation [37]–[39]. Usually, the core idea is to leverage a huge collection of unpaired raw text to generate synthetic speech data using the TTS system. [40] proposed an ASR system trained by both natural and synthetic data obtained using the Tacotron-2 TTS model [41]. The resulting system yielded better WER on the Librispeech dataset. Similarly, [21], presented a speech synthesis based ASR system and demonstrated that diverse multi-speaker speech synthesis is crucial for better results. Further, [42], [43] have jointly trained ASR and TTS models in a close-loop SpeechChain system. Apart from this, [44] proposed a tts4pretrain system to leverage text to infuse useful phonetic and lexical knowledge at the pre-training stage. The further extension to this work has been presented in tts4pretain 2.0, where the inclusion of consistency regularization and contrastive loss during pretraining, facilitated the learning of robust shared representation of speech and text [45].

Previously, multilingual ASR systems were based on context-dependent DNN-HMM [14], [46]. The context-dependent approaches, for example, the shared hidden layer (SHL) network share the intermediate hidden layers across multiple source languages to facilitate multilingual learning [47]. Each language has its own dedicated softmax layer to learn language-dependent features [14]. Similarly, the SHL model has been explored using the LSTMs network and shown improvement over DNN-HMM based model [48]. Further, [13] proposed an end-to-end Transformer based multilingual ASR system for low-resource languages. The multilingual Transformer jointly learns acoustic, pronunciation, and language models in a single multi-head attention network. The proposed system outperformed the previous LSTM based SHL model [48] by an average WER reduction of 12.4%. Furthermore, [49] explored the connectionist temporal classification (CTC) model

for multilingual and cross-lingual settings. Some studies also focused on extracting bottleneck features from multiple languages and then using these features for acoustic modeling [50]–[54]. Further, [55] showed language identification information fused with acoustic features could improve the ASR system for low-resource settings.

Apart from multilingual/cross-lingual systems, some studies also exploited transfer learning for low-resource speech recognition [17]. [56] proposed the shared hidden layer (SHL) based transfer learning using language adversarial transfer learning. Adversarial learning mitigated the problem of learning unnecessary language features found in multilingual SHL models. Language identification has also been used in multilingual transfer learning [57].

Recently, meta-learning (learning to learn), a new domain of research, has been explored by some studies for low-resource speech recognition [18]. Meta-learning have shown promising results in resource constraint computer vision [58] and natural language processing [59] tasks. Recently, [18] and [19] proposed a model agnostic meta-learning based approach for low-resource speech recognition. They have shown meta-learning systems could potentially outperform multilingual systems.

In recent years, semi-supervised approaches for ASR have received a great deal of attention among many researchers [60]–[62]. The self-training approaches also known as pseudo-labeling based approaches are the most common and very impactful semi-supervised learning approaches. In self-training, the initial model (also referred to as the seed model) is obtained by training it on limited labeled data and then utilizing this seed model to generate predictions (also called pseudo labels) on large unlabeled data. The most recent studies include noisy student training [63], [64], iterative pseudo-labeling [65], [66], momentum pseudo-labeling [67], and graph-based self-training approaches [68]. Some studies also focused on entropy minimization based semi-supervised approaches [69]–[71].

Further, [23] proposed a self-supervised wav2vec model for representation learning. The core idea was to pre-train a model on raw unlabeled speech data in a self-supervised

(unsupervised) manner to learn speech representations. Afterward, the model was fine-tuned on very few hours of labeled data in a supervised fashion. The wav2vec model showed significant results even with limited hours of labeled data. The extension to this work has been presented in wav2vec 2.0 [24]. [25] adopted the wav2vec 2.0 model and proposed a cross-lingual XLSR model, which learned latent speech representation across multiple languages with shared quantized speech representation. The XLSR approach (XLSR-53) was trained on 53 languages across various datasets including multilingual Librispeech [72], Common Voice [30], and BABEL datasets [73]. The XLSR-53 model performed extremely well in terms of CER/WER and also required only a few hours of labeled data to fine-tune a model on a target language.

More recently, [26] presented a further improvement to cross-lingual representation learning using a large-scale XLS-R model. The XLS-R model was multilingually pre-trained on 128 languages as compared to 53 languages in the previous XLSR-53 model [25]. In comparison with XLSR-53, the XLS-R is also bigger in terms of the number of parameters (largest model with 2 billion parameters) used for training and also the amount of training data (half a million) used. The authors presented three variants of the XLS-R model such as XLS-R-300M (300 million parameters), XLS-R-1B (1 billion parameters), and XLS-R-2B (2 billion parameters).

4.2.2 *Existing ASR Research for the Punjabi Language*

In the past decade, many researchers attempted to develop a Punjabi ASR system. Most of the research focused on traditional statistical and hybrid approaches, which involved GMM-HMM and DNN-HMM. [28] explored the HMM based Punjabi ASR system for the isolated word recognition task. The proposed system had issues such as a limited vocabulary size (115 isolated words, 8 speakers) and it was an isolated word speech recognition system, which was not ideal for practical scenarios. Further, [74] proposed a continuous triphone HMM based ASR system for the Punjabi language. The proposed system was

trained on a speech corpus containing only 100 utterances from 9 speakers. Similarly, [29] presented a spontaneous Punjabi ASR system, however, their system was still trained on extremely limited data, which could cause generalization issues on diverse speech datasets. To tackle the limited labeled data, [75] constructed a dataset containing 5000 utterances of words recorded by 4 sets of speakers in 4 different dialects of the Punjabi language. They also used the HMM based system and presented an analysis on finding optimal parameters for HMM using genetic and differential evolution algorithms. The analysis showed that the combination of differential evolution and HMM performed better than a regular and genetic algorithm based HMM.

Further, [76] exploited the Kaldi toolkit [77] and presented a continuous Punjabi ASR system. The study demonstrated that triphone modeling units outperformed the monophone units. Also, acoustic features such as Mel-Frequency Cepstral Coefficient (MFCC) performed better than Perceptual Linear Prediction (PLP) features for the Punjabi language. Besides, some studies also focused on the development of the Punjabi ASR system for mobile devices [78], [79]. The work presented by [79] analyzed various flavors of acoustic models consisting of context-dependent and independent models, and context-dependent untied, tied, and deleted interpolation models. Their analysis demonstrated that the context-dependent untied model performs better than other models in terms of accuracy and WER. However, context-independent required less storage, which is ideal for mobile devices.

A comparative study of GMM-HMM and DNN-HMM based Punjabi ASR systems has been conducted in [80]. The study suggested that the hybrid DNN-HMM system outperformed GMM-HMM based system. Similarly, [81] various hybrid models based on the DNN-HMM architecture were compared. [82] proposed a hybrid feature extraction method for the Punjabi language. Recently, [83] presented a work in which they proposed an AutoSSR system based on the Sphinx toolkit[1]. However, their system was trained using a dated statistical toolkit and also on just 3.5 hours of data, which could potentially

1 https://cmusphinx.github.io

fail to generalize to diverse Punjabi speech samples. The rest of the recent work has been focused on the development of Punjabi ASR systems for children's speech [39], [84]–[86].

Although there are many studies presented in the past decade, the research on Punjabi ASR systems remains very limited only covering statistical and hybrid models (i.e., GMM-HMM and DNN-HMM). To our knowledge, no prior studies have examined state-of-the-art end-to-end ASR approaches for the Punjabi language. Hence, with this aim in mind, in this work, we present a new framework for the ASR system, which is based on the state-of-the-art unsupervised cross-lingual wav2vec2.0 for low-resource Punjabi language. To deal with the issue of scarcity of data, we also produced two new synthesized Punjabi speech datasets for training the ASR model.

4.3 PROPOSED FRAMEWORK

4.3.1 *The Synthesized Punjabi Datasets*

The scarcity of annotated Punjabi speech data is a major stumbling block for the development of ASR systems for the Punjabi language. Although annotated speech data is very limited for the Punjabi language but text-only data is available in large quantities. In the past, text-only data had been used to generate synthesized speech data for training an ASR model [37]–[39], [87]. The ASR model trained on these synthesized speech datasets aids in reducing word error rates by a large margin [88]. Hence, motivated by these previous studies, we synthesized our own two Punjabi speech datasets, namely, Google-synth [89] and CMU-synth [90]. For CMU-synth we adopted CMU's Clustergen model [91]. Clustergen is a statistical parametric model available under the Festival Speech Synthesis system[2]. We train our synthesis model (also known as a text-to-speech model) from scratch using the CMU INDIC Punjabi dataset[3]. The dataset consists of speech samples from a single female speaker.

2 https://www.cstr.ed.ac.uk/projects/festival
3 http://festvox.org/cmu_indic

Further, the Google-synth speech dataset is produced using Google's Cloud text-to-speech API[4]. We generate speech data that contain 4 synthetic speakers including 2 male and 2 female speakers. More details are listed in Table 4.1. To synthesize Punjabi speech, we have used Punjabi text present in the Old Newspapers dataset[5], which is a cleaned subset of the HC corpus[6]. The corpus is available freely under CC0 public domain license. The corpus consists of textual data collected from newspapers, blogs, and social media platforms in 67 languages across the world. In total, the corpus has 16,806,041 sentences. We filtered out the Punjabi sentences from the main corpus for our speech synthesis.

4.3.2 *The ASR Model*

For our proposed framework, we adopt the self-supervised pre-trained wav2vec 2.0 based crosslingual models. The wave2vec 2.0 model is a powerful speech representation learning self-supervised model as shown in Figure 4.1. The model maps the raw input speech X to latent speech representations Z using set convolutional layers. The output Z is transformed into a finite set of discrete representations Q using a quantization module. The discretized representations Q act as target representations for the self-supervised objective. The Transformer network then takes the latent speech representations Z and outputs the contextual representations C.

4.3.2.1 *Quantization and Masking*

In simple terms, quantization maps the infinite values in a continuous space to a finite set of discrete values. In wav2vec 2.0, the latent space representations Z are mapped to discrete representations using product quantization. The finite set of phonemes is represented as V entries of code words in G codebooks. The quantization module then can select the correct code words from different codebooks that represent accurate phonemes.

4 https://cloud.google.com/text-to-speech
5 https://www.kaggle.com/alvations/old-newspapers
6 https://www.kaggle.com/code/mpwolke/hc-corpora-newspapers/notebook

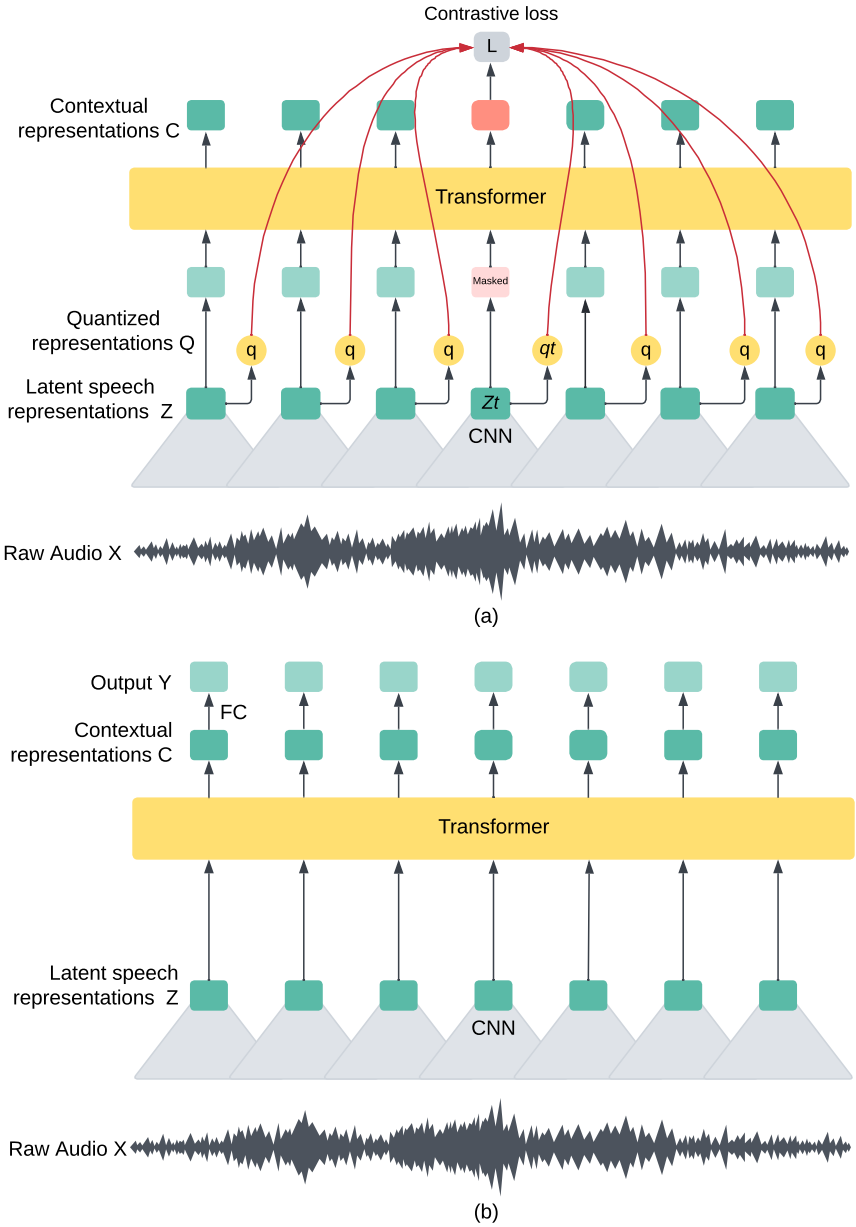

Figure 4.1: wav2vec 2.0 is pre-trained by self-supervised latent representations (a) and the ASR model is obtained by fine-tuning wav2vec on our datasets (b). For cross-lingual pre-training (e.g., XLSR-53 model) quantized representations are shared across multiple languages.

The best selected code words from different codebooks are then concatenated to acquire quantized representation using a linear transformation.

Some proportion p of latent space representations Z are randomly masked before being input to the Transformer network. The number of time steps to be masked is determined using the M parameter that allows the overlapping of multiple masks.

4.3.2.2 *Contrastive Learning*

The core idea of pre-training wav2vec 2.0 is inspired by the BERT language modeling [92], where some parts of the input (in latent space Z) to the Transformer network are masked and the goal is to predict the masked input. For speech, this is achieved using contrastive learning. Contrastive learning aims to guess whether the two different resulting transformations of the same input are still the same. In the wav2vec 2.0 case, these two transformations are quantized representations Q and contextual representations C.

4.3.2.3 *Training Objective*

The model is trained using a sum of two losses: contrastive loss and diversity loss. The contrastive loss facilitates contrastive learning, where the model learns and predicts the accurate quantized representations Q against the central time step of masked contextual representations C. On the other hand, the diversity loss is used as a regularizer to encourage the model to utilize the entire codebook entries equally.

4.3.2.4 *Fine-tuning*

The pre-trained model is then fine-tuned using a labeled dataset in a supervised fashion. To adopt a pre-trained network to ASR, the linear projection layer (fully connected (FC) layer) is added on top of contextual representations C. The fine-tuning process is optimized using the CTC loss function [93].

4.3.2.5 *Language Model and Decoding*

We use the n-gram KenLM [94] language model for decoding. We build a 5-gram language model trained on Punjabi source text from IndicCorp corpus [95]. The corpus contains a huge collection of text across 12 Indian languages including Punjabi. For the Punjabi language, the corpus is made up of 2.64 million news articles, which translates to 29.2 million sentences and 773 million tokens.

4.4 EXPERIMENTAL SETUP

4.4.1 *Datasets*

For our experiments, we use a wide range of datasets across various domains. We use 5 different datasets including the publicly available Common Voice Punjabi dataset v9, the 50Languages Punjabi dataset, our synthesized datasets, and the self-recorded custom Punjabi Speech dataset.

- **Common Voice (CV):** The Common voice datasets are the crowd-sourced collections of datasets that include data across 60 different languages [30]. We are using the Punjabi dataset (pa-IN) available in Common Voice April 2022 release version v9. The dataset contains 1 hour of validated labeled data that covers 51 speakers (76% male and 24% female).
- **Punjabi Speech (PS):** The Punjabi Speech dataset [96] is a self-recorded dataset. The dataset is recorded from audiobooks and news prompts. The dataset includes 771 utterances from 2 male speakers.
- **50Languages (50Lang):** The 50Languages dataset is provided by 50Languages Organization[7]. The provided data is divided into 100 parts based on the type of spoken

7 https://www.50languages.com/phrasebook/en/pa

Table 4.1: Statistics of the various datasets used for our experimentation. It lists the total amount of data in terms of hours, number of utterances and speakers, and average length of utterances (in seconds) in each dataset.

Dataset	#Hours	#Utterances	#Speakers	Avg. length (secs)
Common Voice	1	1210	51	4.82
Punjabi Speech	2	771	2	7.86
50Languages	3	3955	1	1.96
Google-synth	38	50K	4	2.70
CMU-synth	170	80K	1	5.87

utterance. For example, there is a category of asking for directions, which consists of all the direction related spoken utterances. However, in our setup, we do not categorize the data but instead compile the whole data into one dataset. In total the 50Languages corpus has 3955 utterances including isolated word utterances. The recordings and corresponding text labels are manually cleaned and corrected.

- **Google-synth (G-synth):** Our Google-synth speech dataset [89] is synthesized using Google's text-to-speech cloud API. The dataset is composed of 50K utterances from 4 speakers including 2 male and 2 female speakers.
- **CMU-synth (C-synth):** Our CMU-synth speech dataset [90] is synthesized using CMU's Clustergen model as explained in Section 3.1. The dataset contains 80k synthesized utterances that cover 170 hours of speech data.

4.4.2 *Methodology*

We fine-tuned the wave2vec2.0 based pre-trained models on various Punjabi datasets. These pre-trained models include XLSR-53, XLS-R-300M and XLS-R-1B. These models

Table 4.2: The pre-trained crosslingual wav2vec models used to fine-tune our Punjabi ASR system.

Model	#Hours	#Languages	#Parameters
XLSR-53	50K	53	300M
XLS-R-300M	436K	128	317M
XLS-R-1B	436K	128	965M

are pre-trained on multiple datasets including monolingual and multilingual LibriSpeech [72], BABEL [73], Common Voice [30], VoxLingua107 [97], and VoxPopuli [98] datasets. Due to the different sizes of different datasets, we use a different number of epochs for fine-tuning. For each dataset, the number of epochs are determined based on when the WER and CER of the model do not improve on the validation set. For the Common Voice dataset, we fine-tuned the model for 100 epochs with a batch size of 64. We fine-tune the pre-trained models for 150 and 50 epochs on the Punjabi Speech and 50Languages datasets, respectively. We found 30 and 10 epochs to be effective on the Google-synth and CMU-synth datasets, respectively.

When fine-tuning the pre-trained models using our synthesized datasets alongside with individual real speech labeled datasets, we used 10 epochs with a batch size of 32. We used 80%, 10%, and 10% split for training, validation, and testing sets for all the datasets, respectively. For the Common Voice dataset, we used default train, dev (validation), and test splits [30]. We optimize the fine-tuning process using Adam optimizer [99] with a learning rate of 3e-4. Further, the masking probability is set to 0.05, and the attention drop is 0.1. In our experiments, fine-tuning is done using the NVIDIA A100 GPUs.

Further, we trained a single large KenLm 5-gram language model using the IndiCorp dataset [95]. We primarily use default weights to train the KenLM language model. However, for decoding, we fine-tune the values of α and β parameters for the Punjabi language. α is a weight given to the language model for decoding an ASR transcript during shallow fusion, while the β parameter is a constant used for length adjustment during scoring.

We optimize these parameters by testing various combinations of α and β values. We find that an $\alpha = 0.7$ and a $\beta = 4.0$ yield the best results on our test sets during decoding.

4.5 EXPERIMENTAL RESULTS AND DISCUSSION

4.5.1 *Comparative Result Analysis*

Table 4.3 presents the experimental results of fine-tuning a pre-trained model on five different datasets: Common Voice, Punjabi Speech, 50Languages, Goggle-synth, and CMU-synth. The results are shown in terms of Character Error Rate (CER) and Word Error Rate (WER). The table has four sections, each one representing different fine-tuning scenarios. The baseline results are obtained by fine-tuning the models solely on real speech datasets, without incorporating any synthesized datasets or language models. The "Baseline + synth" scenario fine-tuned the models with synthesized datasets in addition to real speech datasets but without language model decoding. The "Baseline + LM" scenario fine-tuned the models with real speech datasets, but with the addition of language model decoding. Lastly, the "Baseline + synth + LM" scenario fine-tuned the models on both real and synthesized datasets, as well as incorporating a language model.

The results of the "Baseline" fine-tuning scenario demonstrate poor performance across all datasets, as evidenced by high character error rate (CER) and word error rate (WER). The incorporation of synthesized data in the "Baseline + synth" scenario resulted in improved performance, particularly in the Common Voice and Punjabi Speech datasets. The "Baseline + LM" scenario yields significant performance improvements, particularly in the 50Languages, Goggle-synth, and CMU-synth datasets. The optimal results are obtained when the model is fine-tuned with both real and synthesized datasets and language model decoding (Baseline + synth + LM scenario), resulting in the lowest CER and WER across all the datasets.

On average, the XLSR-53 model outperforms other models. On Common Voice, the XLSR-53 model achieves significant relative improvement using our synthesized datasets in terms of CER (36.68%) and WER (47.49%). Further, the addition of the language model boosts the relative improvement to 45.98% in CER and 59.14% in WER. A similar trend is seen on the Punjabi Speech dataset. Further, we find that on the 50Languages dataset, our synthesized datasets improve the WER but it is a very marginal improvement. However,

Table 4.3: Experimental results on 5 datasets. Bold font indicates the best results.

Model	**CV**		**PS**		**50Lang**		**G-synth**		**C-synth**	
	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER
Baseline										
XLSR-53	18.66	56.47	16.03	54.26	19.62	32.97	2.92	13.74	2.23	7.03
XLS-R-300M	23.94	68.74	19.82	62.95	21.80	42.09	2.86	13.61	2.23	7.02
XLS-R-1B	29.52	76.18	18.69	60.87	20.13	33.63	3.13	14.74	2.17	6.60
Baseline + synth										
XLSR-53	15.92	43.94	10.47	34.68	20.48	31.94	3.45	16.01	2.26	7.23
XLS-R-300M	16.95	45.58	10.96	34.36	19.32	30.21	3.27	15.46	2.18	6.76
XLS-R-1B	24.86	60.53	15.33	47.51	22.49	41.31	4.45	20.05	2.90	10.40
Baseline + LM										
XLSR-53	16.15	45.84	11.10	33.93	**18.45**	**26.77**	1.49	5.49	1.81	5.07
XLS-R-300M	20.66	55.31	13.28	37.46	20.47	35.36	1.44	5.34	1.82	5.08
XLS-R-1B	27.08	66.80	13.57	41.19	18.94	27.30	1.71	6.63	**1.79**	**5.00**
Baseline + synth + LM										
XLSR-53	**10.08**	**23.07**	**8.26**	**24.37**	19.66	30.74	**1.05**	**3.36**	1.81	5.08
XLS-R-300M	10.72	23.31	8.81	25.52	19.00	29.22	1.15	3.90	1.81	5.06
XLS-R-1B	14.82	29.69	9.90	28.00	20.20	30.86	1.07	3.41	1.80	5.05

the best results are reported using a language model without synthesized datasets. This could be attributed to the fact that most of the 50Language dataset, especially the test set is made up of isolated words. Our analysis reveals that most of the errors are caused by space substitution errors, where one word is recognized as two individual words. Our synthesized datasets help in circumventing some of these errors, however, they still contribute to most of the errors.

Further, we also test our synthesized speech datasets (i.e. Google-synth and CMU-synth) individually. The reported result in Table 4.3 in "Baseline + Synth" and "Baseline + synth + LM" scenarios are achieved by fine-tuning models with synthesized and real speech datasets. The best performing models manage to keep the CER and WER under 5%.

We further analyze the performance of the models by fine-tuning them on combined datasets. Our analysis shows that the models performance indeed gets better by training on combined datasets. The results are presented in Table 4.4. We find that the XLSR-53 model performs the best overall, with the lowest CER and WER on almost all the datasets.

Furthermore, as our datasets have varied lengths of utterances (see Table 4.1), we analyze the impact of different lengths of utterances on the performance. We find that in general, the utterance length does not affect the performance. However, if the training and testing set have different lengths of utterances then the resulting improvements are less significant. We reach this conclusion when we further investigate the results on the

Table 4.4: Experimental results are presented by combining all 5 datasets together. The results are decoded with a language model. Bold font indicates the best results.

Model	**CV**		**PS**		**50Lang**		**G-synth**		**C-synth**	
	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER
XLSR-53	**8.12**	**20.65**	**7.05**	**21.23**	**18.38**	**28.26**	**1.01**	**3.22**	1.36	4.85
XLS-R-300M	9.08	21.1	7.67	23.03	18.41	28.3	1.10	3.58	1.42	4.90
XLS-R-1B	12.64	27.36	8.89	27.1	19.56	29.13	1.04	3.35	**1.30**	**4.81**

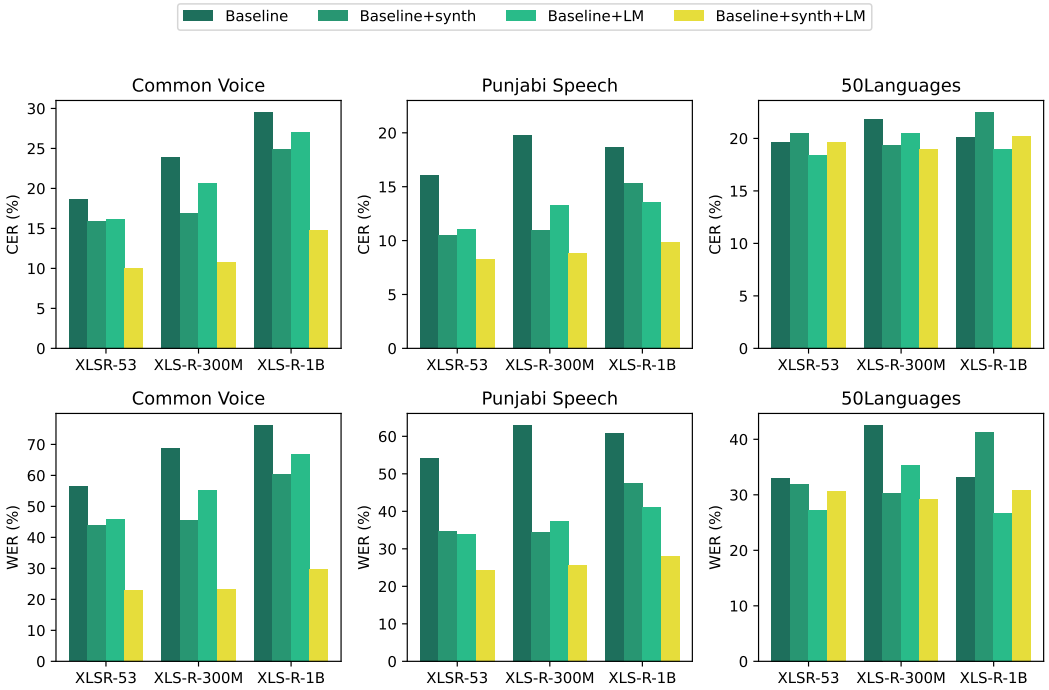

Figure 4.2: Visual representations of experimental results on three real speech datasets in terms of CER (top row) and WER (bottom row). The bar graphs clearly state that our synthesized speech datasets significantly reduce the error rates. Additional improvement is achieved by decoding the final output using the 5-gram KenLM language model.

50Languages dataset, where the training set has long utterances (average words/characters of 3.66/18.38), and the testing set consists of mostly isolated word utterances (average words/characters of 1.3/8.49). As demonstrated in Figure 4.2 for the 50Languages dataset the relative improvements are minor due to the mismatch of utterance lengths between the training and testing sets.

4.5.2 *Effectiveness of our Synthesized Punjabi Speech Datasets*

Essentially, synthesizing speech from natural text using a TTS system is equivalent to impairing the speech while keeping the text unchanged. We conjecture that the addition

Figure 4.3: Detailed error analysis on the predicted transcripts. These utterances/sentences are chosen across three real speech datasets (i.e., Common Voice, Punjabi Speech, and 50Languages). We selected XLSR-53 as our main model for analysis since it produces consistently better results than other models.

of such impaired examples can improve the robustness of an ASR model to noisy speech. We also hypothesize that the supplementary lexical content in the synthesized examples helps improve the decoding ability of an ASR model. To test our hypothesis, we conducted a details analysis of predicted utterances as presented in Figure 4.3.

Acoustically, we discover that the model trained with our synthetic datasets is more accurate in recognizing homophones, which are words that phonetically sound similar but have different spellings (e.g., "sing" and "singh"). As shown in Figure 4.3, the baseline model recognizes the word "ਸਿੱਖ" (sikkh) as "ਸਿਕ" (sik). However, the addition of our synthetic data resolves the spelling of the homophone. Additionally, both synthetic data and language model helps the model to correct homophone errors in other examples.

We further inspect the model's ability to capture nasal and stress markers in the Punjabi language. We test the model's performance on various utterances that included nasal and stress markers as presented in Figure 4.3. In the Punjabi language, nasal sounds are produced using the *Bindi* "ਅਃ" and *Tippi* "ਅੰ" (similar to the sound of the letters "n" or "m"). The stress on certain consonants is produced using *Adhak* "ਅੱ" (also called germination), which is similar to the use of double letters in English to indicate stress on certain words. Our analysis reveals that the baseline model sometimes fails to capture these nasal and stress sounds in the Punjabi language. However, the inclusion of our additional synthesized datasets assists the model to capture nasal markers, and the integration of the language model further boosts the recognition of stress markers.

We also investigate the model's performance in recognizing misspelled conjunct consonants, which are formed by combining two or more consonants. These conjunct consonants are usually added to the bottom of the consonant and written in the subscript form. In the Punjabi language, there are three conjunct consonants (i.e., ਹ, ਰ, and ਵ) as seen in examples such as "ਜੇਲ੍ਹ" (ਹ), "ਅੰਮ੍ਰਿਤ" (ਰ), and "ਸ੍ਵੈਮਾਨ" (ਵ). Our analysis indicates that the baseline model consistently missed these conjunct consonants. However, the additional exposure to synthetic data and language model to resolve these errors.

In addition, we analyze the datasets using the Speech Data Explorer toolkit [100]. We identify all utterances that consistently produced higher word error rates across different datasets. Our analysis of the Common Voice dataset shows that most of the errors are due to poor and noisy recording conditions, as well as variations in accent and speech rate. As a result, we find that the additional synthetic data aid to reduce the percentage

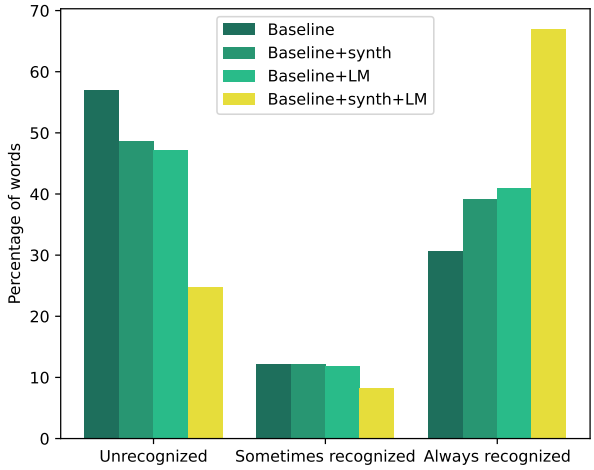

Figure 4.4: Word accuracy distribution. The word is classified as "Unrecognized" with an accuracy of 0%. The word is considered "Sometimes recognized" with an accuracy of 0% to 100%. The word is "Always recognized" with an accuracy of 100%. These results are presented on the Common Voice dataset using the XLSR-53 model.

of unrecognized words from 57.04% to 48.69%, as illustrated in Figure 4.4. Additionally, incorporating the language model further lower the rate of unrecognized words to 47.13%. Finally, when both synthetic data and a language model are used together, the number of unrecognized words drops significantly to 24.77%. Overall, the results show that using synthetic data and a language model in combination can improve the recognition rate of speech samples, thereby decreasing the proportion of unrecognized and partially recognized words.

4.5.2.1 *Resistant to Change*

We observe that transliterated utterances are the most challenging for our model to recognize. For example, as shown in Figure 4.3 left (Transliteration), the utterance is spoken as in the Hindi language but transliterated into Punjabi language making it tough to recognize. Similarly, in Figure 4.3 right (Transliteration), the sentence is a phrase from Sikh scripture (also referred to as *Sant Bhasha* translated to *language of saints*), which is

spoken just like Punjabi language but written with different spellings. Our model Baseline+synth+LM (ਏ ਮਨ ਚੰਚਲ ਵਸ ਨਾ ਆਵੇ) successfully captured the phonetic essence of the utterance but since our acoustic and language model are not exposed to *Sant Bhasha* samples, it misspelled most of the words.

Further, since, most of the 50Languages dataset is made up of isolated words, especially the test set, the model struggles with space substitution problems. For example, the word ਸਿਗਰਟਨੋਸ਼ is recognized as two isolated words ਸਿਗਰਟ ਨੋਸ਼. We attribute these errors to the pronunciation style of the speaker, where the speaker leaves some time to pronounce the whole word fluently. Although synthesized data and language model help to correct some of these mistakes, but still there are a large quantity of space substitution errors.

4.5.3 *Effectiveness of Synthesized Data in Very Limited Settings*

To further evaluate the effectiveness of our synthesized data in very limited data settings, we carried out experiments where we fine-tune the models with the Common Voice dataset and varying amounts of our synthesized data from our Google-synth and CMU-synth datasets. The results are illustrated in Figure 4.5. We find that even with less synthesized data in Google-synth models performed better as compared with the CMU-synth dataset. We report that adding more data does not always translate to better performance in the case of the CMU-synth dataset. This effect is attributed to the limited acoustic diversity in the dataset since it synthesized samples from a single female speaker. Also, the intelligibility of speech in CMU-synth is worse than Google-synth, hence comparatively worse performance compared with the Google-synth dataset.

Interestingly, the most striking result to emerge from our analysis is that the bigger XLS-R-1B model tends to perform worse as compared to relatively smaller models such as XLSR-53 and XLS-R-300M. Our analysis suggests that since the XLS-R-1B model is pre-trained on massive 128 languages and has around 1 billion parameters, the model struggles to reach an optimal solution in very low-resource scenarios. Thus, the results

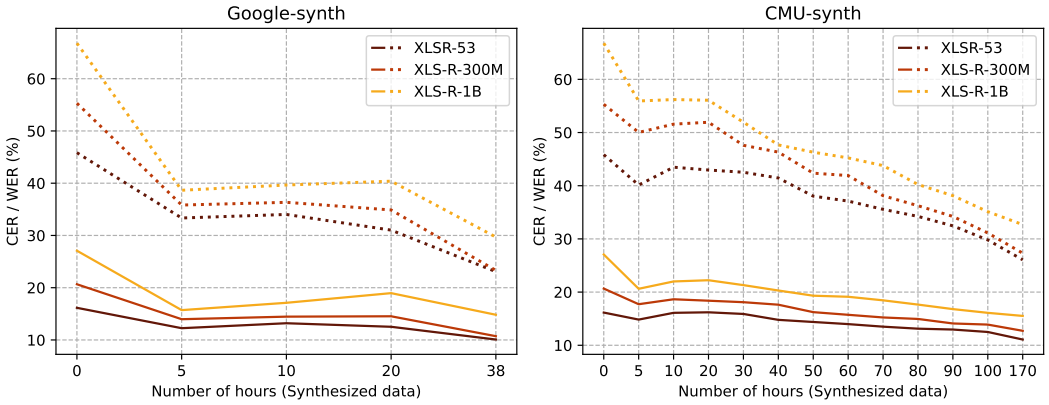

Figure 4.5: Experimental results showing the effectiveness of our synthesized data in limited data setting. The three models are fine-tuned with Common Voice, Google-synth, and CMU-synth datasets (with multiple limited hours settings). Here the solid lines represent CER and the dotted lines represent WER.

are worse than a much smaller model such as XLSR-53, which is only pre-trained on 53 languages and has relatively fewer (300 million) parameters. Besides, this could also be due to an interference problem while training multilingual systems [101]. The interference problem occurs when a model is trained on many languages and diverse styles of pronunciation among these languages negatively affect the optimization process of the model [26]. Overall, these results indicate that wav2vec 2.0 based crosslingual models fine-tuned with our synthesized datasets can aid in achieving excellent results in low-resource scenarios and can help in building robust ASR systems for such languages.

4.6 CONCLUSIONS

In this chapter, we propose a robust ASR for low-resource Punjabi language leveraging our synthesized speech datasets. We selected three variants of the pre-trained cross-lingual wav2vec 2.0 models (i.e., XLSR-53, XLS-R-300M, and XLS-R-1B) to fine-tune for Punjabi ASR. The XLSR-53 model is pre-trained on 53 different languages while other models are pre-trained with massive 128 languages in a self-supervised manner. To deal with

the scarcity of labeled Punjabi speech data, we synthesize two datasets (i.e., CMU-synth and Google-synth). The synthesized datasets are used along with the real labeled Punjabi speech datasets for fine-tuning. Our experimental results show that our synthesized datasets can improve the performance of low-resource Punjabi ASR system. We observe that on average, the XLSR-53 model mostly outperforms other models on real speech data such as the Common Voice, Punjabi Speech, and 50Languages datasets. Further, the empirical finding of this study suggests that even only a few hours of synthesized speech data can significantly improve the overall accuracy of the ASR systems.

In the future, we will conduct more experiments including investigating the different architectures for ASR, speech synthesis, and language modeling. As Punjabi is a tonal language, we will also investigate the use of pitch features alongside other features. We will use our previous work DeepFo model [102] for pitch extraction for this purpose. We will also investigate the use of our previous speech enhancement work [103] for a noise-robust ASR system for low-resource languages.

In the next chapter, we explores an impactful self-training (pseudo-labeling) approach to further enhance speech recognition for low-resource languages. By leveraging abundant unlabeled data, Chapter 5 showcases a method tailored for the low-resource languages. Our adoption of length-normalized confidence score filtering and extensive experiments showcase the potential of pseudo-labeling in improving recognition performance on various datasets.

REFERENCES

[1] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2016, pp. 4960–4964.

[2] Mark Gales, Steve Young, *et al.*, "The application of hidden markov models in speech recognition," *Foundations and Trends® in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008.

[3] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, "Revisiting hybrid and GMM-HMM system combination techniques," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2013, pp. 6744–6748.

[4] Herve Bourlard and Nelson Morgan, "Continuous speech recognition by connectionist statistical methods," *IEEE Transactions on Neural Networks*, vol. 4, no. 6, pp. 893–909, 1993.

[5] Hervé Bourlard and Nelson Morgan, "Connectionist speech recognition-a hybrid approach," Kluwer Academic Publishers, Tech. Rep., 1994.

[6] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[7] Dong Yu and Li Deng, *Automatic speech recognition*. Springer, 2016, vol. 1.

[8] Jia Pan, Cong Liu, Zhiguo Wang, Yu Hu, and Hui Jiang, "Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why DNN surpasses GMMs in acoustic modeling," in *International Symposium on Chinese Spoken Language Processing*, IEEE, 2012, pp. 301–305.

[9] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve, "Wav2Letter: an End-to-End ConvNet-based Speech Recognition System," 2016. arXiv: 1609.03193.

[10] Rohit Prabhavalkar, Kanishka Rao, Tara N Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly, "A comparison of sequence-to-sequence models for speech recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2017, pp. 939–943.

[11] Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, *et al.*, "MUCS 2021: Multilingual and code-switching ASR challenges for low resource Indian languages," in *Proc. International Speech Communication Association (Interspeech)*, International Speech Communication Association, 2021.

[12] Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert, "Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 4751–4755.

[13] Shiyu Zhou, Shuang Xu, and Bo Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," 2018. arXiv: 1806.05059.

[14] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2013, pp. 7304–7308.

[15] Changhan Wang, Juan Pino, and Jiatao Gu, "Improving Cross-Lingual Transfer Learning for End-to-End Speech Recognition with Speech Translation," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 4731–4735.

[16] Vikas Joshi, Rui Zhao, Rupesh R Mehta, Kshitiz Kumar, and Jinyu Li, "Transfer Learning Approaches for Streaming End-to-End Speech Recognition System," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 2152–2156.

[17] Xia Mao and Yulv Zhang, "Time-Delay Recurrent Neural Network for Cross-Lingual Speech Recognition," in *Recent Developments in Intelligent Computing, Communication and Devices*, Springer, 2019, pp. 341–348.

[18] Satwinder Singh, Ruili Wang, and Feng Hou, "Improved Meta Learning for Low Resource Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 4798–4802.

[19] Jui-Yang Hsu, Yuan-Jui Chen, and Hung-yi Lee, "Meta learning for end-to-end low-resource speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7844–7848.

[20] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2019, pp. 2613–2617.

[21] Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu, "Speech recognition with augmented synthesized speech," in *IEEE automatic speech recognition and understanding workshop (ASRU)*, IEEE, 2019, pp. 996–1002.

[22] Thai-Son Nguyen, Sebastian Stueker, Jan Niehues, and Alex Waibel, "Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7689–7693.

[23] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "Wav2vec: Unsupervised pre-training for speech recognition," 2019. arXiv: 1904.05862.

[24] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[25] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2021, pp. 2426–2430.

[26] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," 2021. arXiv: 2111.09296.

[27] Lucas Ondel, Pierre Godard, Laurent Besacier, Elin Larsen, Mark Hasegawa-Johnson, Odette Scharenborg, Emmanuel Dupoux, Lukas Burget, François Yvon, and Sanjeev Khudanpur, "Bayesian models for unit discovery on a very low resource language," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5939–5943.

[28] Mohit Dua, RK Aggarwal, Virender Kadyan, and Shelza Dua, "Punjabi automatic speech recognition using HTK," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 4, p. 359, 2012.

[29] Yogesh Kumar and Navdeep Singh, "A first step towards an automatic spontaneous speech recognition system for Punjabi language," *International Journal of Statistics and Reliability Engineering*, vol. 2, no. 1, pp. 81–93, 2015.

[30] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[31] Zhihan Wang, Feng Hou, Yuanhang Qiu, Zhizhong Ma, Satwinder Singh, and Wang Ruili, "CyclicAugment: Speech Data Random Augmentation with Cosine Annealing Scheduler for Automatic Speech Recognition," in *Proc. International Speech Communication Association (Interspeech)*, ISCA, 2022, pp. 3859–3863.

[32] Vladimir Bataev, Roman Korostik, Evgeny Shabalin, Vitaly Lavrukhin, and Boris Ginsburg, "Text-only domain adaptation for end-to-end asr using integrated text-to-mel-spectrogram generator," 2023. arXiv: 2302.14036.

[33] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2015, pp. 3586–3589.

[34] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 5220–5224.

[35] Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux, "Data augmenting contrastive learning of speech representations in the time domain," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021, pp. 215–222.

[36] Navdeep Jaitly and Geoffrey E Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013, p. 21.

[37] Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney, "Generating synthetic audio data for attention-based speech recognition systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7069–7073.

[38] Aleksandr Laptev, Roman Korostik, Aleksey Svischev, Andrei Andrusenko, Ivan Medennikov, and Sergey Rybin, "You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation," in *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, IEEE, 2020, pp. 439–444.

[39] Taniya Hasija, Virender Kadyan, and Kalpna Guleria, "Out Domain Data Augmentation on Punjabi Children Speech Recognition using Tacotron," in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1950, 2021, p. 012 044.

[40] Jason Li, Ravi Gadde, Boris Ginsburg, and Vitaly Lavrukhin, "Training neural speech recognition systems with synthetic speech augmentation," 2018. arXiv: 1811.00707.

[41] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 4779–4783.

[42] Andros TTjandra, Sakriani Sakti, and Satoshi Nakamura, "Listening while speaking: Speech chain by deep learning," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2017, pp. 301–308.

[43] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura, "End-to-end feedback loss in speech chain framework via straight-through estimator," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6281–6285.

[44] Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Gary Wang, and Pedro Moreno, "Injecting text in self-supervised speech pretraining," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, pp. 251–258.

[45] Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro Moreno, and Gary Wang, "Tts4pretrain 2.0: Advancing the use of text and speech in asr pretraining with consistency and contrastive losses," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7677–7681.

[46] Frank Seide, Gang Li, and Dong Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. International Speech Communication Association (Interspeech)*, 2011, pp. 437–440.

[47] Sibo Tong, Philip N Garner, and Hervé Bourlard, "Cross-lingual adaptation of a CTC-based multilingual acoustic model," *Speech Communication*, vol. 104, pp. 39–46, 2018.

[48] Shiyu Zhou, Yuanyuan Zhao, Shuang Xu, Bo Xu, *et al.*, "Multilingual Recurrent Neural Networks with Residual Learning for Low-Resource Speech Recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2017, pp. 704–708.

[49] Siddharth Dalmia, Ramon Sanabria, Florian Metze, and Alan W Black, "Sequence-based multi-lingual low resource speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 4909–4913.

[50] Jonas Gehring, Yajie Miao, Florian Metze, and Alex Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2013, pp. 3377–3381.

[51] Karel Veselỳ, Martin Karafiát, František Grézl, Miloš Janda, and Ekaterina Egorova, "The language-independent bottleneck features," in *IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2012, pp. 336–341.

[52] Ngoc Thang Vu, Florian Metze, and Tanja Schultz, "Multilingual bottle-neck features and its application for under-resourced languages," in *Spoken language technologies for under-resourced languages*, 2012.

[53] Dong Yu and Michael L Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Proc. International Speech Communication Association (Interspeech)*, 2011, pp. 237–240.

[54] Tanel Alumäe, Stavros Tsakalidis, and Richard M Schwartz, "Improved Multilingual Training of Stacked Neural Network Acoustic Models for Low Resource Languages," in *Proc. International Speech Communication Association (Interspeech)*, 2016, pp. 3883–3887.

[55] Markus Müller, Sebastian Stüker, and Alex Waibel, "Language adaptive DNNS for improved low resource speech recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2016, pp. 3878–3882.

[56] Jiangyan Yi, Jianhua Tao, Zhengqi Wen, and Ye Bai, "Language-adversarial transfer learning for low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 621–630, 2018.

[57] Ekapol Chuangsuwanich, "Multilingual techniques for low resource automatic speech recognition," Ph.D. dissertation, MIT, Cambridge, United States, 2016.

[58] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," in *International Conference on Machine Learning (ICML)*, 2017, pp. 11261135.

[59] Jiatao Gu, Yong Wang, Yun Chen, Victor OK Li, and Kyunghyun Cho, "Meta-learning for low-resource neural machine translation," in *Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3622–3631.

[60] Samuel Thomas, Michael L Seltzer, Kenneth Church, and Hynek Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2013, pp. 6704–6708.

[61] Shigeki Karita, Shinji Watanabe, Tomoharu Iwata, Atsunori Ogawa, and Marc Delcroix, "Semi-supervised end-to-end speech recognition.," in *Proc. International Speech Communication Association (Interspeech)*, 2018, pp. 2–6.

[62] Murali Karthick Baskar, Shinji Watanabe, Ramon Astudillo, Takaaki Hori, Lukáš Burget, and Jan Černockỳ, "Semi-supervised sequence-to-sequence asr using unpaired speech and text," 2019. arXiv: 1905.01152.

[63] Daniel S. Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V. Le, "Improved Noisy Student Training for Automatic Speech

Recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 2817–2821.

[64] Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," 2020. arXiv: 2010.10504.

[65] Qiantong Xu, Tatiana Likhomanenko, Jacob Kahn, Awni Hannun, Gabriel Synnaeve, and Ronan Collobert, "Iterative pseudo-labeling for speech recognition," 2020. arXiv: 2005.09267.

[66] Tatiana Likhomanenko, Qiantong Xu, Jacob Kahn, Gabriel Synnaeve, and Ronan Collobert, "SlimIPL: Language-model-free iterative pseudo-labeling," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 741–745.

[67] Yosuke Higuchi, Niko Moritz, Jonathan Le Roux, and Takaaki Hori, "Momentum pseudo-labeling for semi-supervised speech recognition," 2021. arXiv: 2106.08922.

[68] Niko Moritz, Takaaki Hori, and Jonathan Le Roux, "Semi-supervised speech recognition via graph-based temporal classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6548–6552.

[69] Dong Yu, Balakrishnan Varadarajan, Li Deng, and Alex Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech & Language*, vol. 24, no. 3, pp. 433–444, 2010.

[70] Yves Grandvalet and Yoshua Bengio, "Semi-supervised learning by entropy minimization," *Advances in neural information processing systems*, vol. 17, 2004.

[71] Vimal Manohar, Daniel Povey, and Sanjeev Khudanpur, "Semi-supervised maximum mutual information training of deep neural network acoustic models," in *Proc. International Speech Communication Association (Interspeech)*, 2015, pp. 2630–2634.

[72] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Interspeech*, 2020, pp. 2757–2761.

[73] M Harper. "IARPA babel program." Accessed: 07/11/2022. (2012), [Online]. Available: http://www.iarpa.gov/index.php/research-programs/babel.

[74] Wiqas Ghai and Navdeep Singh, "Continuous speech recognition for Punjabi language," *International Journal of Computer Applications*, vol. 72, no. 14, 2013.

[75] Virender Kadyan, Archana Mantri, and RK Aggarwal, "Refinement of HMM model parameters for Punjabi automatic speech recognition (PASR) system," *IETE Journal of Research*, vol. 64, no. 5, pp. 673–688, 2018.

[76] Jyoti Guglani and Achyuta Nand Mishra, "Continuous Punjabi speech recognition model based on Kaldi ASR toolkit," *International Journal of Speech Technology*, vol. 21, no. 2, pp. 211–216, 2018.

[77] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, *et al.*, "The Kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, 2011.

[78] Puneet Mittal and Navdeep Singh, "Speaker-independent automatic speech recognition system for mobile phone applications in Punjabi," in *International Symposium on Signal Processing and Intelligent Recognition Systems*, Springer, 2017, pp. 369–382.

[79] Puneet Mittal and Navdeep Singh, "Development and analysis of Punjabi ASR system for mobile phones under different acoustic models," *International Journal of Speech Technology*, vol. 22, no. 1, pp. 219–230, 2019.

[80] Virender Kadyan, Archana Mantri, RK Aggarwal, and Amitoj Singh, "A comparative study of deep neural network based Punjabi-ASR system," *International Journal of Speech Technology*, vol. 22, no. 1, pp. 111–119, 2019.

[81] Jyoti Guglani and Achyuta Nand Mishra, "DNN based continuous speech recognition system of Punjabi language on Kaldi toolkit," *International Journal of Speech Technology*, vol. 24, pp. 41–45, 2021.

[82] Virender Kadyan, Archana Mantri, and RK Aggarwal, "Improved filter bank on multitaper framework for robust Punjabi-ASR system," *International Journal of Speech Technology*, vol. 23, no. 1, pp. 87–100, 2020.

[83] Yogesh Kumar, Navdeep Singh, Munish Kumar, and Amitoj Singh, "AutoSSR: an efficient approach for automatic spontaneous speech recognition model for the Punjabi Language," *Soft Computing*, vol. 25, no. 2, pp. 1617–1630, 2021.

[84] Harshdeep Kaur and Virender Kadyan, "Feature Space Discriminatively Trained Punjabi Children Speech Recognition System Using Kaldi Toolkit," in *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*, 2020.

[85] Vivek Bhardwaj, Virender Kadyan, *et al.*, "Deep Neural Network Trained Punjabi Children Speech Recognition System Using Kaldi Toolkit," in *5th International Conference on Computing Communication and Automation (ICCCA)*, IEEE, 2020, pp. 374–378.

[86] Virender Kadyan, Syed Shanawazuddin, and Amitoj Singh, "Developing childrens speech recognition system for low resource punjabi language," *Applied Acoustics*, vol. 178, p. 108 002, 2021.

[87] Zhehuai Chen, Andrew Rosenberg, Yu Zhang, Gary Wang, Bhuvana Ramabhadran, and Pedro J Moreno, "Improving Speech Recognition Using GAN-Based Speech Synthesis and Contrastive Unspoken Text Selection," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 556–560.

[88] Deblin Bagchi, Shannon Wotherspoon, Zhuolin Jiang, and Prasanna Muthukumar, "Speech Synthesis as Augmentation for Low-Resource ASR," 2020. arXiv: 2012.13004.

[89] Satwinder, Ruili Wang, and Feng Hou, "Google-synth: A Synthesized Punjabi Speech Dataset," Jul. 2023. DOI: 10.6084/m9.figshare.23615607.v1. [Online]. Available: https://figshare.com/articles/dataset/Google-synth_A_Synthesized_Punjabi_Speech_Dataset/23615607.

[90] Satwinder Singh, Ruili Wang, and Feng Hou, "CMU-synth: A synthesized Punjabi Speech dataset," Jun. 2023. DOI: 10.6084/m9.figshare.23606697.v1. [Online]. Available: https://figshare.com/articles/dataset/_strong_CMU-synth_A_synthesized_Punjabi_Speech_dataset_strong_/23606697.

[91] Alan W Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *Proc. International Speech Communication Association (Interspeech)*, 2006, pp. 1762–1765.

[92] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018. arXiv: 1810.04805.

[93] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *International conference on Machine learning (ICML)*, 2006, pp. 369–376.

[94] Kenneth Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.

[95] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar, "IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4948–4961.

[96] Satwinder Singh, Ruili Wang, and Feng Hou, "Punjabi Speech: A labeled Speech Corpus," Jul. 2023. DOI: 10.17632/sdbc8f5b77.1. [Online]. Available: https://data.mendeley.com/datasets/sdbc8f5b77/1.

[97] Jörgen Valk and Tanel Alumäe, "Voxlingua107: A dataset for spoken language recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021, pp. 652–658.

[98] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 993–1003.

[99] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[100] Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg, "A toolbox for construction and analysis of speech datasets," in *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2022.

[101] Yizhou Lu, Mingkun Huang, Xinghua Qu, Pengfei Wei, and Zejun Ma, "Language adaptive cross-lingual speech representation learning with sparse sharing sub-networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6882–6886.

[102] Satwinder Singh, Ruili Wang, and Yuanhang Qiu, "DeepFo: End-to-end fundamental frequency estimation for music and speech signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 61–65.

[103] Yuanhang Qiu, Ruili Wang, Feng Hou, Satwinder Singh, Zhizhong Ma, and Xiaoyun Jia, "Adversarial multi-task learning with inverse mapping for speech enhancement," *Applied Soft Computing*, vol. 120, p. 108 568, 2022.

5

SELF-TRAINING FOR LOW-RESOURCE SPEECH RECOGNITION

In this chapter, we propose a self-training approach for automatic speech recognition (ASR) in low-resource settings. While self-training approaches have been extensively developed and evaluated for high-resource languages such as English, their applications to low-resource languages like Punjabi have been limited, despite the language being spoken by millions globally. The scarcity of annotated data has hindered the development of accurate ASR systems, especially for low-research languages (e.g., Punjabi and Māori languages). To address this issue, we propose an effective self-training approach that generates highly accurate pseudo-labels for unlabeled low-resource speech. Our experimental analysis demonstrates that our approach significantly improves word error rate, achieving a relative improvement of 14.94% compared to a baseline model across four real speech datasets. Further, our proposed approach reports the best results on the Common Voice Punjabi dataset.

5.1 INTRODUCTION

The development of end-to-end (E2E) automatic speech recognition (ASR) systems for high-resource languages has made significant strides in recent years, resulting in an excellent performance [1]. The key factors contributing to this high performance include the presence of large annotated datasets, advancements in deep learning algorithms, and easy

access to high-performance computational resources. However, the same cannot be said for ASR systems designed for low-resource languages. These systems still fall short in terms of performance when compared to their high-resource counterparts. The primary reason for this discrepancy is the lack of annotated speech datasets and their availability in the public domain. Without these resources, it is difficult for ASR systems to be trained effectively and perform at the same level as those designed for high-resource languages.

The research community has been actively exploring various ways to address the scarcity of annotated datasets for speech recognition systems [2]–[6]. The process of annotating highly accurate speech datasets is a labor-intensive task that requires a significant amount of time and resources. To overcome these challenges, many researchers have proposed self-training or pseudo-labeling (PL) based methods [7]. These methods leverage the vast amount of available unlabeled data in conjunction with a small labeled dataset to improve the performance of ASR systems. Self-training has gained increasing interest in a wide range of research areas, including computer vision [8], [9], natural language processing [10], [11], and more recently, in the field of speech recognition [2], [12]–[15]. These methods have shown promise in addressing the shortage of annotated data and improving the performance of ASR systems.

The fundamental concept behind self-training is to first train an initial seed model, either from scratch or by fine-tuning an existing pre-trained model, using a limited amount of labeled data. This seed model is then used to generate pseudo-labels for unlabeled data. The process can be repeated over multiple iterations in order to improve the quality of the generated pseudo-labels. To ensure that only highly accurate pseudo-labels are used, various filtering methods are applied to filter out any incorrect labels [12]. Additionally, a robust language model can be utilized to achieve accurate decoding of ASR transcriptions, further improving the quality of the generated pseudo-labels.

Most of the previous and recent work in self-training/ pseudo-labeling is based on and demonstrated on high-resource languages like English [2], [12]–[14], [16]. However, most of the languages of the world are low-resource and the ASR system designed for

them performs poorly when compared to those with high-resource languages. With that in mind, in this work, we focus to design a self-training approach for one of the most widely spoken languages in the world, i.e., the Punjabi language. Although the Punjabi language is spoken natively across India and Pakistan region, however, Punjabi speakers are spread across the globe, especially in countries such as Canada, the United States, the United Kingdom, Australia, and New Zealand. Despite having more than 100 million speakers, the Punjabi language still lacks mature ASR systems due to a lack of annotated datasets.

In the chapter, we propose a self-training approach to leverage the available large unlabeled audio data. We adopt a pre-trained self-supervised crosslingual wav2vec 2.0 model (XLSR-53) [17] as our seed model to generate pseudo-labels across multiple iterations. The proposed model first fine-tunes the XLSR-53 seed model on limited Punjabi datasets in a supervised fashion with language model decoding. Afterwards, the fine-tuned model is used to produce pseudo-labels for unlabeled audio data. As we are fine-tuning on limited data, it is common to get a mixed bag of pseudo-labels. To sieve out the erroneous pseudo-labels, we propose a length-normalized confidence score. The pseudo-labels with high confidence scores above the certain cut-off threshold are selected to be included into the dataset. Finally, the model is fine-tuned again from scratch using labeled and highly confident pseudo-labels. This process is repeated over multiple iterations with gradual confidence score filtration to refine the pseudo-labels. Our empirical results demonstrate that our proposed approach is able to achieve significant WER reduction compared with the baseline and also reports the best results on the Common Voice Punjabi dataset. To the best of our knowledge, this work is the first to explore the self-training approach for the Punjabi language on the mix of datasets (public, non-public, and synthesized datasets).

The rest of the chapter is structured as follows: Section 5.2 provides a review of previous studies on self-training. Section 5.3 outlines our proposed approach, including information about our seed model and self-training approach. In Section 5.4, we describe the

datasets and methodology we used for experimentation. The detailed result analysis is presented in Section 5.5, and our conclusions are outlined in Section 5.6.

5.2 RELATED WORK

Semi-supervised approaches such as self-supervision and self-training have been active areas of research for the past few years. The goal of these approaches is to leverage the large unlabeled speech data to improve the overall accuracy of the ASR systems. These approaches showed significant results for low-resource languages. Self-training is a fairly simple yet effective semi-supervised approach for utilizing unlabeled audio data. Self-training has been applied in computer vision and natural language processing tasks. However, it is recently adopted for E2E speech recognition.

Kahn et al. [7] proposed a self-training approach for the E2E ASR model. The proposed system incorporated a pseudo-label filtering method and language model decoding to improve the word error rates. Additionally, an ensemble of multiple models showed further performance improvement. Further, inspired by noisy student training (NST) from the image processing domain, Park et al. [12] adopted this method for speech recognition. The proposed approach exploited the adaptive SpecAugment [18] augmentation method. The NST approach incorporated filtering, balancing, and mixing techniques to carefully select highly accurate pseudo-labels. Furthermore, rather than selecting pseudo-labels over just one iteration, iterative pseudo-labeling (IPL) [13] refines the acoustic model over multiple iterations. IPL effectively generates the pseudo-labels for a subset of unlabeled data rather than labeling the entire unlabeled data. Also, IPL fine-tuned an existing model instead of training an entirely new model.

Recently, Higuchi et al. [14], presented momentum pseudo-labeling (MPL), which incorporated online and offline models. At different time steps, weights of multiple models are averaged using the momentum-based moving average to refine pseudo-labels. The online model (also called the teacher model), hypothesizes the pseudo-labels produced

by the offline model (also referred to as the student model) on the go. The MPL approach outperformed standard pseudo-labels and IPL approaches. The extension to this work is presented in [19], where MPL is combined with IPL. The resulting approach improved the seed model by exploiting language model knowledge using the IPL strategy. Additionally, the Conformer-based ASR architecture further enhanced the overall accuracy of the system.

While most of the self-training approaches utilized language models for decoding, the SlimIPL [16] method performed pseudo-labeling without any language model. The proposed model is built upon the IPL approach, which produced pseudo-labels over multiple iterations using a single model. SlimIPL also incorporated a dynamic cache for better and more stable model training. Furthermore, Lugosch et al. [20] presented a self-training approach to multilingually generate pseudo-labels for 60 languages. The proposed approach first pre-trained a single multilingual model followed by fine-tuning it on the target language using semi-supervision. The fine-tuned model is then used to produce pseudo-labels for that particular language, which in turn is utilized to train the final model. Overall, the multilingual pseudo-labeling demonstrated good performance and better generalization to the data from different domains.

Xu et al. [21] presented a study, which showed that self-training and pre-training can effectively be combined and are complementary to each other. Pre-training based self-supervised approaches like wav2vec [1] and its variants demonstrated excellent performance for low-resource languages. They leveraged unlabeled audio data and learn representations using contrastive learning. When combined with self-training, self-supervision showed a better generation of pseudo-labels and achieved significant results. Similarly, Jin et al. [22] exploited a self-supervised pre-trained wav2vec 2.0 model for pseudo-labeling. The proposed method filtered out the low-probability pseudo-labels and only selected the pseudo-labels with the highest probability scores. Further, self-training has also been used to improve representation learning in crosslingual settings [23]. Furthermore, Khurana et al. [24] proposed the DUST (Dropout-based Uncertainty-driven Self-Training) approach

to alleviate the issues caused by domain mismatch. The proposed approach efficiently rejected pseudo-labels that exhibit high levels of uncertainty.

Until recently, limited work has been carried out in the Punjabi language. Previous efforts relied heavily on statistical and hybrid models, using mostly non-public datasets [25]–[28]. However, in Chapter 4, we studied various E2E ASR models for the Punjabi language. The study demonstrated that E2E models trained with synthetic speech and real speech datasets could enhance the effectiveness of E2E ASR models for Punjabi.

5.3 PROPOSED APPROACH

We propose a very simple, yet effective self-training method for the Punjabi language. However, the proposed method could work for any language having a small amount of annotated data and large unlabeled audio data.

5.3.1 *Pre-trained Seed Model*

For the strong baseline seed model, we adopt a self-supervised pre-trained wav2vec 2.0 based crosslingual XLSR-53 model [17]. The XLSR-53 model is pre-trained on 53 languages using a contrastive learning approach to learn audio representations in an unsupervised manner. The crosslingual wav2vec model maps raw input speech to latent speech representations, which are then transformed into a set of discrete representations using a quantization module. These discrete representations are shared across multiple languages (in this case 53 languages excluding the Punjabi language) and used as the target for the self-supervised objective. Conneau et al. [17] showed that XLSR-53 could greatly improve the character error rate (CER) and word error rate (WER) in the case of low-resource languages.

Although pre-trained XLSR-53 does not include Punjabi in the pre-training setup, the model is still able to learn powerful shared feature representation across many languages.

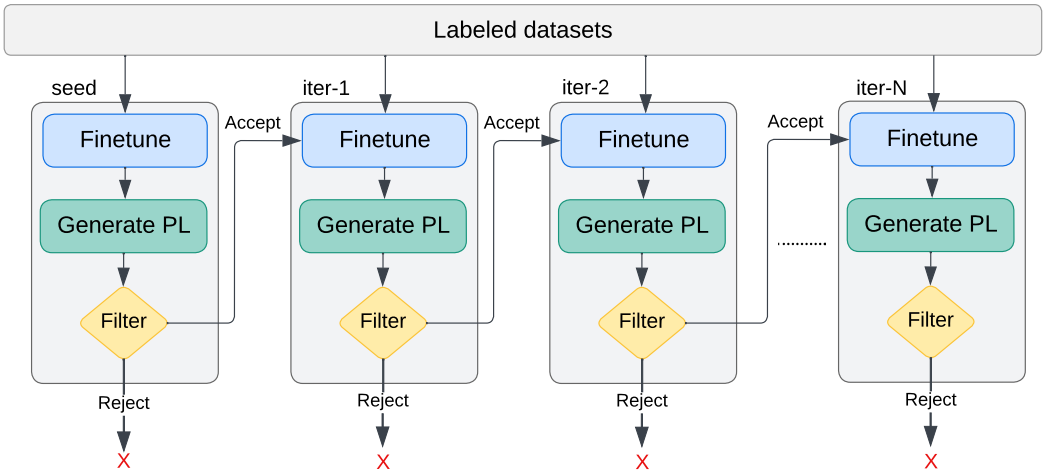

Figure 5.1: Overview of proposed self-training approach.

The crosslingual nature of the model made it easy to fine-tune the model on any new language with limited labeled data. Following that, rather than pre-training the model using the Punjabi language, we leverage the effective feature representations of the pre-trained model and only fine-tune the model on available Punjabi labeled data. This way, we keep our self-training approach simple and fast.

5.3.2 *Self-training Approach*

Our proposed self-training approach is fairly simple as shown in Figure 5.1. Initially, we fine-tune our pre-trained seed acoustic model on available limited labeled Punjabi datasets. Once the model is fine-tuned on the Punjabi language then we use this model to generate the pseudo-labels for unlabeled Punjabi data. The details of labeled and unlabeled datasets are outlined in Section 5.4.1. At this stage, we also employ the 5-gram KenLM language model [29] for decoding the seed model's outputs.

Since the seed model is only trained on limited Punjabi data, it is fairly common to produce inaccurate pseudo-labels. To tackle the erroneous pseudo-labels, we incorporated confidence based scoring strategy [7]. Our language model produces sentence-level shal-

low fusion scores while decoding. We leverage the shallow fusion score as a confidence score (CS) and normalize it to the length of the sentence. We filter pseudo-labels with different filtration thresholds of confidence scores in each iteration of self-training, i.e., [0.5, 0, -0.5, -1, -1.5, -2, -2.5]. We start with a very strict (0.5) threshold and then with each iteration of self-training, we gradually relax the filtration threshold. Afterwards, we fine-tune an entirely new model using a combination of labeled and pseudo-labeled datasets. We repeat this whole process over multiple iterations until the ASR model is not improving on the labeled test datasets.

Our approach is similar to the IPL approach [13], but with few distinctions. Firstly, we compute pseudo-labels for the entire unlabeled dataset instead of a subset of it. Although it is time-consuming to pseudo-label the entire unlabeled dataset, it would give us more high-quality filtered pseudo-labels for fine-tuning. Secondly, since the XLSR-53 model does not show any improvement over multiple continued fine-tuning phases, we find that fine-tuning a new model from a pre-trained XLSR-53 model produces the best results. Therefore, rather than continuing fine-tuning like in IPL, in every iteration, we fine-tune an entirely new model.

5.4 EXPERIMENTAL SETUP

5.4.1 *Datasets*

To train our seed model we use a range of labeled Punjabi datasets. We utilize four low-resource Punjabi real speech datasets and two synthesized datasets. For real speech datasets, we use the Shrutilipi dataset [30], Common Voice (CV) [31], Punjabi Speech [32], and 50Languages[1]. Previously, in Chapter 4, we demonstrated the effectiveness of synthetic datasets in improving the performance of the Punjabi ASR system on real speech

1 https://www.50languages.com/phrasebook/en/pa

Table 5.1: List of datasets used for our experimentations.

Datasets	**#Hours**	**#Utterances**
Real speech		
Shrutilipi	94	50K
Common Voice	1	1210
Punjabi Speech	4	2431
50Languages	3	3955
Synthesized speech		
Google-synth	38	50K
CMU-synth	170	80K
Unlabeled speech		
Audiobooks	450	322K

datasets. Thus, following that, we use two synthesized datasets i.e. Google-synth [33] and CMU-synth [34].

- **Shrutilipi**: The Shrutilipi corpus is a collection of audio and text pairs sourced from public resources such as All India Radio news bulletins, and compiled through data mining techniques. The corpus includes labeled speech data for a total of 12 Indian languages, including Punjabi.

- **Common Voice (CV)**: The Common Voice dataset is a crowd-sourced dataset available to the public, featuring speech data in numerous languages. For our purposes, we utilize the Punjabi data made available in the April 2022 release of Common Voice. The dataset comprises of a range of speech samples from a total of 39 male and 12 female speakers, providing a diverse array of recordings.

- **Punjabi Speech (PS)**: The Punjabi Speech dataset is a read speech dataset like Common Voice. The dataset consists of speech data from 2 male speakers.

- **50Languages (50Lang)**: The dataset is extracted from the 50Languages learning platform[1]. The dataset comprises speech utterances from a single male speaker.
- **Google-synth (G-synth)**: The Google-synth dataset is synthesized labeled speech data produced using Google's Cloud text-to-speech (TTS) API[2]. The dataset contains speech samples from 2 male and 2 female speakers.
- **CMU-synth (C-synth)**: The CMU-synth dataset is synthesized using CMU's Clustergen TTS model. The dataset contains speech samples from a single female speaker.
- **Audiobooks**: The Audiobooks dataset consists of speech recordings gathered from open Punjabi audiobooks featured on various YouTube channels. It is an unlabeled dataset used primarily for self-training purposes. The dataset contains audio recordings from multiple speakers and several audiobooks. To facilitate self-training, we preprocess the audio recordings, chunking them down into smaller audio segments no longer than 15 seconds.

The synthesized datasets (Google-synth and CMU-synth) are produced using Punjabi text available in the Old Newspaper corpus[3]. We ensure that there is no overlap between data used to train the ASR model, self-training, language model, and synthesized datasets.

5.4.2 *Methodology*

For our self-training approach, we adopt a pre-trained wav2vec 2.0 based crosslingual XLSR-53 model. The XLSR-53 model is pre-trained on 53 languages (50K hours) with 300 million parameters. To save time, we only fine-tune the pre-trained model on our labeled datasets instead of pre-training it from scratch. We adopt the pre-trained network by adding a fully connected layer on top of the Transformer. The fully connected layer outputs the characters using Connectionist Temporal Classification (CTC) [35]. During

2 https://cloud.google.com/text-to-speech
3 https://www.kaggle.com/datasets/alvations/old-newspapers

the fine-tuning, we freeze the weights of the feature encoder. We fine-tune a single XLSR-53 model on combined data from 6 of the labeled datasets as mentioned in Table 5.1. For all the datasets except Common Voice, we use 8:1:1 splits for train, validation, and test sets. For the Common Voice dataset, we use pre-defined data splits. We fine-tune our model for 10 epochs with a batch size of 32. Further, we optimize the training process using the Adam optimizer [36] with an initial learning rate of 3e-4, which is warmed up for the first 10% of the updates and afterward, linearly decays during the rest of the updates. All the experiments are conducted on NVIDIA A100 GPUs and results are reported in terms of word error rate (WER).

Additionally, we decode the ASR model's output with a 5-gram KenLM language model [29]. We train the Punjabi language model using the IndiCorp dataset [37]. The dataset contains 29.2 million Punjabi sentences (773 million tokens). We fine-tune α and β parameters and empirically set $\alpha = 0.7$ and $\beta = 4.0$. Here α represents the shallow fusion weight for the language model and β refers to the weight to adjust the score as per the length of the decoded sequence [38]. Further, for selecting the highly confident pseudo-labels, we adopt a normalized confidence score filtering.

5.5 EXPERIMENTAL RESULTS AND DISCUSSION

5.5.1 *Comparative Analysis*

We report the results in Table 5.2. To show the effectiveness of our 5-gram language model (LM), we demonstrate the results on the seed model with and without the language model. We find that our LM decoding further improves the seed model, consequently, we choose seed with LM as our baseline model. All of our experimental results (except seed without LM) are carried out with LM decoding. Raw PL uses all available pseudo-labels, whereas Best PL selects the best-performing pseudo-labels over multiple iterations of pseudo-labeling with different confidence thresholds. The results show that

Table 5.2: Experimental results in terms of WER (%). The best pseudo-label (PL) results are obtained by choosing the best-performing model on labeled datasets over multiple iterations of PL with different confidence thresholds.

Model	Datasets					
	Shrutilipi	**CV**	**PS**	**50Lang**	**G-synth**	**C-synth**
Seed without LM	17.23	29.88	18.88	28.08	13.39	6.82
Seed with LM (baseline)	16.87	14.04	9.84	24.16	2.93	4.96
Raw PL (no filter)	15.23	12.30	9.10	24.00	2.75	5.12
Best PL	**14.67**	**11.42**	**8.57**	**20.55**	**2.69**	**4.92**

using pseudo-labels with a filter (Best PL) outperforms Raw PL on all datasets, indicating that selecting the best pseudo-label is important to improve the performance of the ASR system. We find that iteration 5 with confidence scores (CS$\geqslant$-1.5) produces the best results as shown in Table 5.2 (Best PL) and Figure 5.2. As compared to the baseline, we find our approach achieves relative improvements of 13.04%, 18.66%, 12.91%, and 14.94% on real speech Shrutilipi, Common Voice, Punjabi Speech, and 50Languages datasets, respectively. Our PL approach demonstrates the best results on the Common Voice Punjabi dataset with a relative improvement of 50.5% over previously reported results (23.07 % vs 11.42% WER) [6].

Further, on synthetic datasets (i.e., Google-synth and CMU-synth), our PL approach also achieve better results; nevertheless, the improvements are marginal. We attribute these minimal improvements to the quality of synthetic data. We only incorporated these datasets to improve the quality of the overall ASR model, as shown in Chapter4.

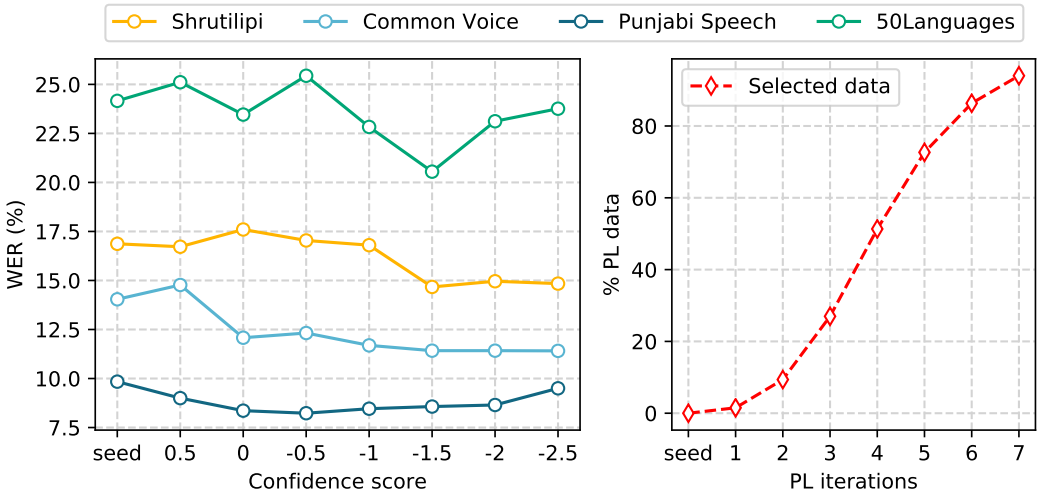

Figure 5.2: Performance on real speech datasets against various confidence score thresholds (left) and % of selected data over the PL iterations (right).

5.5.2 *Effectiveness of Gradual Filtration*

Figure 5.2 illustrates the performance of the model with different CS thresholds and the amount of PL data selected with filtration over multiple PL iterations. Initially, we start with aggressive filtering with (CS⩾0.5), which selects only 1.5% of the total unlabeled Audiobooks dataset. With this threshold, although we select high-quality PL, however, it results in marginal improvements overall (2.8% of relative improvement). As we gradually relax the confidence score threshold, our model indeed continues to produce better results over each iteration of PL. With (CS⩾0), selects 9.37% of PL, which results in 9.5% of overall relative improvement compared to baseline. Overall, as Figure 5.2 shows, (CS⩾-1.5), which selects 72.66% of PL, produces the lowest WERs across most datasets. Most consistent improvements are seen on the Common Voice and Punjabi Speech datasets. Further, our experimental analysis reveals that PL generated with (CS⩾-2 or greater) selects more than 80% of the total PL data. However, the performance of the model gradually starts to diminish as compared to the Best PL.

5.6 CONCLUSIONS

The chapter presents a self-training (pseudo-labeling) approach for automatic speech recognition for low-resource settings, specifically focusing on the Punjabi language. The proposed self-training approach generates highly accurate pseudo-labels for unlabeled Punjabi speech, resulting in a significant improvement in word error rate (WER) compared to a strong baseline. The experimental analysis demonstrates the effectiveness of our approach and highlights the potential to improve the performance of ASR systems for low-resource languages such as Punjabi.

In the forthcoming chapter, we will provide a comprehensive summary of our research findings. Additionally, this chapter will explore potential avenues for future work and directions in the field of study.

REFERENCES

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[2] Anton Ragni, Kate M Knill, Shakti P Rath, and Mark JF Gales, "Data augmentation for low resource languages," in *Proc. International Speech Communication Association (Interspeech)*, 2014, pp. 810–814.

[3] Shiyu Zhou, Shuang Xu, and Bo Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," 2018. arXiv: 1806.05059.

[4] Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu, "Applying wav2vec2. 0 to speech recognition in various low-resource languages," 2020. arXiv: 2012.12121.

[5] Satwinder Singh, Ruili Wang, and Feng Hou, "Improved Meta Learning for Low Resource Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 4798–4802.

[6] Satwinder Singh, Ruili Wang, Feng Hou, and Zhizhong Ma, "Enhancing end-to-end automatic speech recognition for low-resource Punjabi language using synthesized datasets," *Available at SSRN 4181844*,

[7] Jacob Kahn, Ann Lee, and Awni Hannun, "Self-training for end-to-end speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7084–7088.

[8] Dong-Hyun Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, 2013, p. 896.

[9] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel, "MixMatch: A holistic approach to semi-supervised learning," *Advances in neural information processing systems*, vol. 32, 2019.

[10] Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato, "Revisiting self-training for neural sequence generation," in *International Conference on Learning Representations*, 2019.

[11] Ximing Li and Bo Yang, "A pseudo label based dataless naive bayes algorithm for text classification with seed words," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1908–1917.

[12] Daniel S. Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V. Le, "Improved Noisy Student Training for Automatic Speech Recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 2817–2821.

[13] Qiantong Xu, Tatiana Likhomanenko, Jacob Kahn, Awni Hannun, Gabriel Synnaeve, and Ronan Collobert, "Iterative pseudo-labeling for speech recognition," 2020. arXiv: 2005.09267.

[14] Yosuke Higuchi, Niko Moritz, Jonathan Le Roux, and Takaaki Hori, "Momentum pseudo-labeling for semi-supervised speech recognition," 2021. arXiv: 2106.08922.

[15] Niko Moritz, Takaaki Hori, and Jonathan Le Roux, "Semi-supervised speech recognition via graph-based temporal classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6548–6552.

[16] Tatiana Likhomanenko, Qiantong Xu, Jacob Kahn, Gabriel Synnaeve, and Ronan Collobert, "SlimIPL: Language-model-free iterative pseudo-labeling," in *Proc. International Speech Communication Association (Interspeech)*, 2020, pp. 741–745.

[17] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. International Speech Communication Association (Interspeech)*, pp. 2426–2430.

[18] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. International Speech Communication Association (Interspeech)*, 2019, pp. 2613–2617.

[19] Yosuke Higuchi, Niko Moritz, Jonathan Le Roux, and Takaaki Hori, "Advancing momentum pseudo-labeling with conformer and initialization strategy," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7672–7676.

[20] Loren Lugosch, Tatiana Likhomanenko, Gabriel Synnaeve, and Ronan Collobert, "Pseudo-labeling for massively multilingual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7687–7691.

[21] Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli, "Self-training and pre-training are complementary for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 3030–3034.

[22] Zezhong Jin, Dading Zhong, Xiao Song, Zhaoyi Liu, Naipeng Ye, and Qingcheng Zeng, "Filter and evolve: Progressive pseudo label refining for semi-supervised automatic speech recognition," 2022. arXiv: 2210.16318.

[23] Zi-Qiang Zhang, Yan Song, Ming-Hui Wu, Xin Fang, and Li-Rong Dai, "XLST: Cross-lingual self-training to learn multilingual representation for low resource speech recognition," 2021. arXiv: 2103.08207.

[24] Sameer Khurana, Niko Moritz, Takaaki Hori, and Jonathan Le Roux, "Unsupervised domain adaptation for speech recognition via uncertainty driven self-training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6553–6557.

[25] Mohit Dua, RK Aggarwal, Virender Kadyan, and Shelza Dua, "Punjabi automatic speech recognition using HTK," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 4, p. 359, 2012.

[26] Jyoti Guglani and Achyuta Nand Mishra, "Continuous Punjabi speech recognition model based on Kaldi ASR toolkit," *International Journal of Speech Technology*, vol. 21, no. 2, pp. 211–216, 2018.

[27] Virender Kadyan, Archana Mantri, RK Aggarwal, and Amitoj Singh, "A comparative study of deep neural network based Punjabi-ASR system," *International Journal of Speech Technology*, vol. 22, no. 1, pp. 111–119, 2019.

[28] Yogesh Kumar, Navdeep Singh, Munish Kumar, and Amitoj Singh, "AutoSSR: an efficient approach for automatic spontaneous speech recognition model for the Punjabi Language," *Soft Computing*, vol. 25, no. 2, pp. 1617–1630, 2021.

[29] Kenneth Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.

[30] Kaushal Santosh Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra, "Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages," 2022. arXiv: 2208.12666.

[31] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[32] Satwinder Singh, Ruili Wang, and Feng Hou, "Punjabi Speech: A labeled Speech Corpus," Jul. 2023. DOI: 10.17632/sdbc8f5b77.1. [Online]. Available: https://data.mendeley.com/datasets/sdbc8f5b77/1.

[33] Satwinder, Ruili Wang, and Feng Hou, "Google-synth: A Synthesized Punjabi Speech Dataset," Jul. 2023. DOI: 10.6084/m9.figshare.23615607.v1. [Online]. Available: https://figshare.com/articles/dataset/Google-synth_A_Synthesized_Punjabi_Speech_Dataset/23615607.

[34] Satwinder Singh, Ruili Wang, and Feng Hou, "CMU-synth: A synthesized Punjabi Speech dataset," Jun. 2023. DOI: 10.6084/m9.figshare.23606697.v1. [Online]. Available: https://figshare.com/articles/dataset/_strong_CMU-synth_A_synthesized_Punjabi_Speech_dataset_strong_/23606697.

[35] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *International conference on Machine learning (ICML)*, 2006, pp. 369–376.

[36] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[37] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar, "IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4948–4961.

[38] Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 1–5828.

6

CONCLUSIONS

This chapter presents some concluding remarks about this thesis. The thesis focuses on enhancing F_0 estimation for music and speech processing applications (Chapter 2), improving meta-learning for low-resource speech recognition (Chapter 3), enhancing speech recognition for low-resource languages by utilizing our synthesized datasets (Chapter 4), and ultimately, further refining ASR through self-training for low-resource languages (Chapter 5). In this final chapter, we will recapitulate the proposed approaches and provide an outlook on the future.

6.1 RESEARCH SUMMARY

In this thesis, we have proposed different approaches to improve fundamental frequency estimation and automatic speech recognition for low-resource languages. A recap of our approaches and contributions is listed as follows.

Chapter 2 presented a novel data-driven fundamental frequency (F_0) estimation approach. Existing data-driven algorithms had limited learning capabilities due to their shallow receptive field. We addressed this issue by incorporating dilated convolutional block into the network. We also added residual blocks to make training more efficient and fast. Our empirical evaluation demonstrates significantly better results in terms of raw pitch and raw chroma accuracy. Our ablation experiments validated the effectiveness

of adding dilation and residual blocks to the model. Compared with the deep learning based baseline model CREPE [1], our approach used 77.4% fewer network parameters and still achieved state-of-the-art performance.

Chapter 3 presented a meta-learning approach for low-resource speech recognition. There has been very limited work done in the past for meta-learning for speech recognition. Existing work on meta-leaning [2] showed encouraging results. However, the baseline approach with the MAML algorithm leads to inconsistent training behavior, which could lead to subpar WERs. To deal with these issues, we proposed a framework, which employed a multi-step loss (MSL) function to stabilize the training process with the MAML algorithm. The goal of the MSL is to compute losses for each step in the inner loop of MAML and then merge them using a weighted importance vector. The vector assigns greater significance to the loss from the final step than to those from earlier steps. Our experiments demonstrate that incorporating MSL substantially enhances training stability, leading to greater accuracy in the overall system. Our proposed approach achieved better results compared with standard MAML on several languages while maintaining consistent training performance.

Chapter 4 presented a self-supervised approach to improve the accuracy of the automatic speech recognition by exploiting synthesized speech data. We tested our approach on the Punjabi language, which is spoken by more than 100 million speakers world-wild. Despite being spoken by the masses, Punjabi still lacks annotated datasets, which are essential for building robust ASR systems. To address the issue of limited data, we develop three new datasets Google-synth [3] and CMU-synth [4] - through synthesis. We also compile our self-recorded Punjabi Speech dataset [5]. We tested the efficacy of these synthesized datasets by adopting pre-trained cross-lingual wav2vec models (XLSR-53, XLS-R-300M, and XLS-R-1B). Our empirical analysis shows that our synthesized datasets helped models to improve the performance on real speech labeled datasets including Common Voice, Punjabi Speech, and 50Languages. Additionally, we demonstrated that even just a

few hours of our synthesized data could improve the error rates on real speech labeled datasets.

Chapter 5 proposed a simple yet effective self-training (pseudo-labeling) approach for improving speech recognition for low-resource languages. In low-resource scenarios, labeled data is mostly scarce, however, unlabeled data is available in large amounts. Previously, researchers utilized unlabeled data for self-training. However, most of the past methods focused on high-resource languages like English. Therefore, we present a self-training approach for the low-resource Punjabi language. To sieve out erroneous pseudo-labels, we adopted length-normalized confidence score filtering. Our experimental results validate the efficacy of the proposed approach, showing significantly improved results on various datasets.

6.2 FUTURE WORK AND DIRECTIONS

In this section, we will discuss certain directions of potential research that this thesis does not address but are crucial for the advancement of low-resource speech recognition.

6.2.1 *Expansion of target languages*

Most of the work presented in this thesis is focused on Indian languages, especially the Punjabi language. There are various other languages that need attention from the ASR perspective. For example, Te reo Māori is one of the official languages of New Zealand. However, there has been very little work done on the language.

6.2.2 *Working with different accents and dialects*

In this thesis, we do not pay much attention to various dialects, speaking styles, and accents in a language while training a model. Although our proposed approaches sig-

nificantly performed better on a variety of speech data, however, explicit knowledge of various accents and dialects distilled into the models could potentially further improve the performance [6], [7].

6.2.3 *Advance Language Models*

In our research, we employed statistical language modeling for decoding. However, recently large language model such as GPT2 [8], GPT3 [9], GPT4 [10], BERT [11], RoBERTa [12], XLNet [13], and ELECTRA [14] have shown excellent results. These models outperformed all the traditional models by a large margin, which could potentially result in additional improvements in ASR systems, especially for low-resource languages.

REFERENCES

[1] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello, "CREPE: A convolutional representation for pitch estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 161–165.

[2] Jui-Yang Hsu, Yuan-Jui Chen, and Hung-yi Lee, "Meta learning for end-to-end low-resource speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7844–7848.

[3] Satwinder, Ruili Wang, and Feng Hou, "Google-synth: A Synthesized Punjabi Speech Dataset," Jul. 2023. DOI: 10.6084/m9.figshare.23615607.v1. [Online]. Available: https://figshare.com/articles/dataset/Google-synth_A_Synthesized_Punjabi_Speech_Dataset/23615607.

[4] Satwinder Singh, Ruili Wang, and Feng Hou, "CMU-synth: A synthesized Punjabi Speech dataset," Jun. 2023. DOI: 10.6084/m9.figshare.23606697.v1. [Online].

Available: https://figshare.com/articles/dataset/_strong_CMU-synth_A_synthesized_Punjabi_Speech_dataset_strong_/23606697.

[5] Satwinder Singh, Ruili Wang, and Feng Hou, "Punjabi Speech: A labeled Speech Corpus," Jul. 2023. DOI: 10.17632/sdbc8f5b77.1. [Online]. Available: https://data.mendeley.com/datasets/sdbc8f5b77/1.

[6] Yanmin Qian, Xun Gong, and Houjun Huang, "Layer-wise fast adaptation for end-to-end multi-accent speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2842–2853, 2022.

[7] Shahram Ghorbani and John HL Hansen, "Domain expansion for end-to-end speech recognition: Applications for accent/dialect speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[10] OpenAI, "Gpt-4 technical report," Tech. Rep., 2023.

[11] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, 2019, pp. 4171–4186.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019. arXiv: 1907.11692.

[13] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[14] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," 2020. arXiv: 2003.10555.

Part I

APPENDIX

A

REAL AND SYNTHESIZED SPEECH PUNJABI DATASETS

A.1 INTRODUCTION

Speech recognition has been an active area of research for several decades, and the availability of large annotated datasets has played a crucial role in the development of robust speech recognition systems. In recent years, with the advent of deep learning and other machine learning techniques, there has been a renewed interest in creating and using large datasets to train speech recognition models.

There are many speech datasets available such as Common Voice [1], LibriSpeech [2], TIMIT [3], TED-LIUM [4], Wall Street Journal [5], Switchboard [6], Google Speech Commands datasets [7] and so on. However, most of these datasets are compiled for high-resource languages such as English. Most of the languages of the world are low-resource languages and do not have enough linguistic resources such as high-quality annotated datasets. There are approximately 7000 languages spoken worldwide, but only a small fraction of them, roughly 100, have well-established automatic speech recognition (ASR) systems [8]. The remaining languages, including Punjabi, are considered low-resource languages. Punjabi, belonging to the Indo-Aryan language family, is spoken by over 110 million native speakers in India and Pakistan, as well as throughout the world. Punjabi is unique in the Indo-Aryan language family because it uses distinct lexical tones, including low, mid, and high tones, and is written in two scripts: Gurmukhi in India and Shahmukhi

in Pakistan. Despite its large population of speakers, Punjabi lacks the quality annotated datasets to build an accurate speech recognition system.

With that in mind, we create three labeled Punjabi speech datasets namely, Punjabi Speech [9], Google-synth [10], and CMU-synth [11]. While Punjabi Speech is a real-speech dataset and the other two are synthesized speech datasets. Besides this, we also compiled an unlabeled dataset called Audiobooks, which is made up of Punjabi audiobooks.

We have demonstrated the effectiveness of our compiled datasets in Chapter 4 and 5. We show that our compiled datasets can improve the results for Punjabi Speech recognition. Our synthesized datasets, i.e., Google-synth and CMU-synth, reduce the error rates on publicly available Common Voice (refer to Chapter 4). Further, the Audiobooks dataset has shown excellent performance in self-training (refer to Chapter 5)

A.2 DATASET CREATION

For recording a Punjabi Speech and synthesizing Google-synth and CMU-synth datasets, we utilized text available in the Old Newspapers dataset[1]. This particular dataset is a carefully curated subset of the HC corpus, and it is available to the public for free under the CC0 public domain license. The corpus contains a vast amount of textual data that has been collected from a wide range of sources, including newspapers, blogs, and various social media platforms. This corpus has been designed to cover 67 different languages spoken across the world, and it comprises 16,806,041 sentences in the TSV (Tab Separated Values) file format.

As our focus is on the Punjabi language, we filtered out the Punjabi sentences from the original corpus. This leaves us with a more manageable dataset that we can work with more easily. To normalize the text, we filter out all the special symbols and numeric entries.

1 https://www.kaggle.com/alvations/old-newspapers

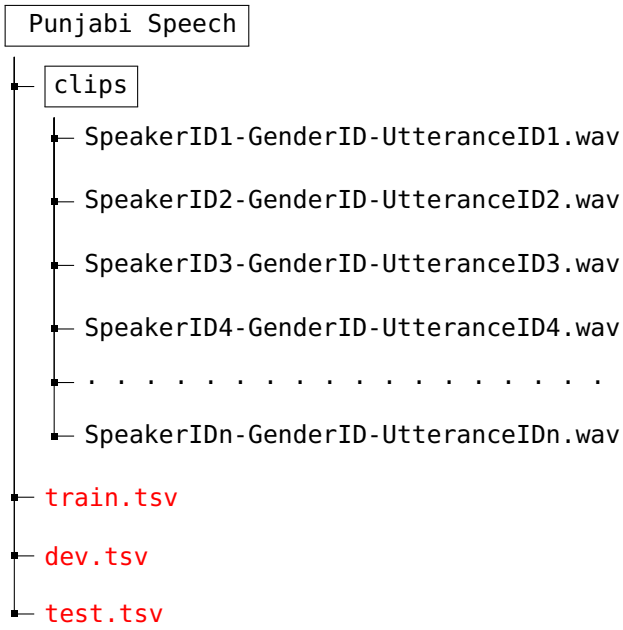

Figure A.1: Directory structure of Punjabi Speech corpus. Here box represents the directory and TSV files are the transcript files that include all the associated labels for audio files. The Google-synth and CMU-synth follow the same directory structure, except the audio file name only includes UtteranceID.

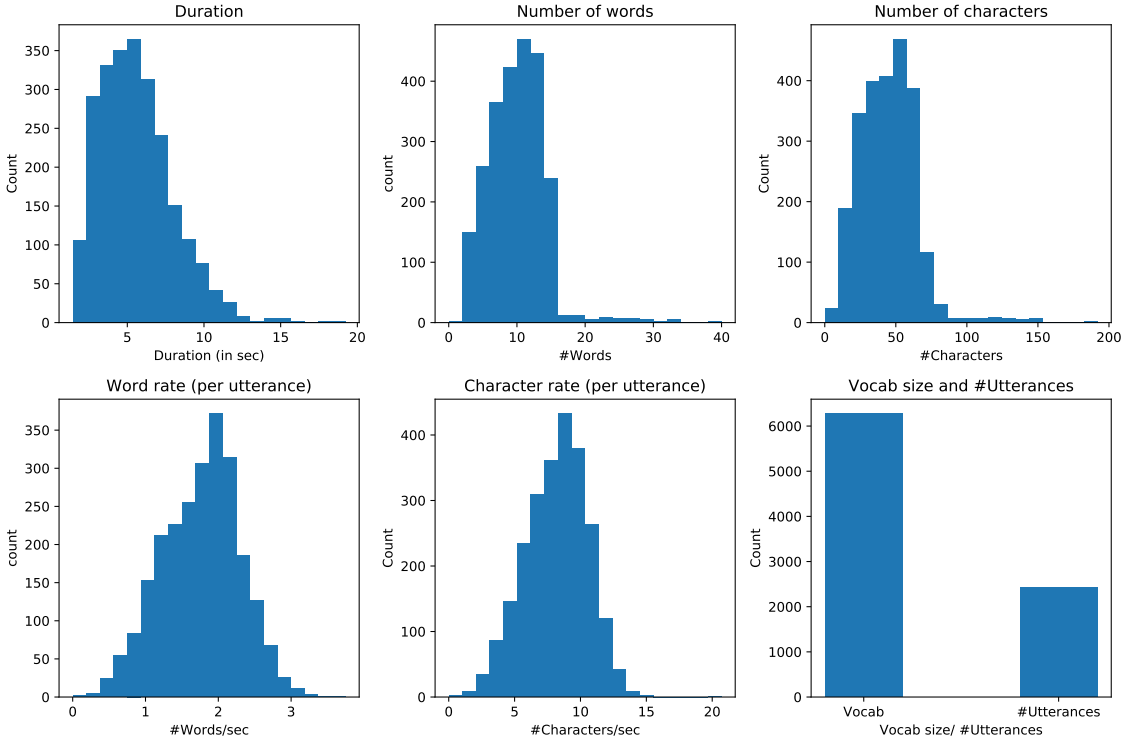

Figure A.2: Statistics of Punjabi Speech dataset

A.2.1 *Punjabi Speech*

The Punjabi Speech dataset is a read speech dataset, recorded in the studio and open environment. We record the speech samples at 16kHz in WAV file format. We keep our recording below 15 seconds to avoid memory issues while training on the GPUs. Presently, this dataset contains speech samples from two male speakers and has a total of 2429 spoken utterances making it 4 hours of data. We pre-define the data splits with 80% for training and 10% for validation and 10% for testing purposes. The Punjabi speech dataset follows a very simple structure. All the speech files are present in *clips* directory and all the transcript files (*train*, *dev*, *test*) in TSV format are present in the parent directory of the corpus as illustrated in Figure A.1.

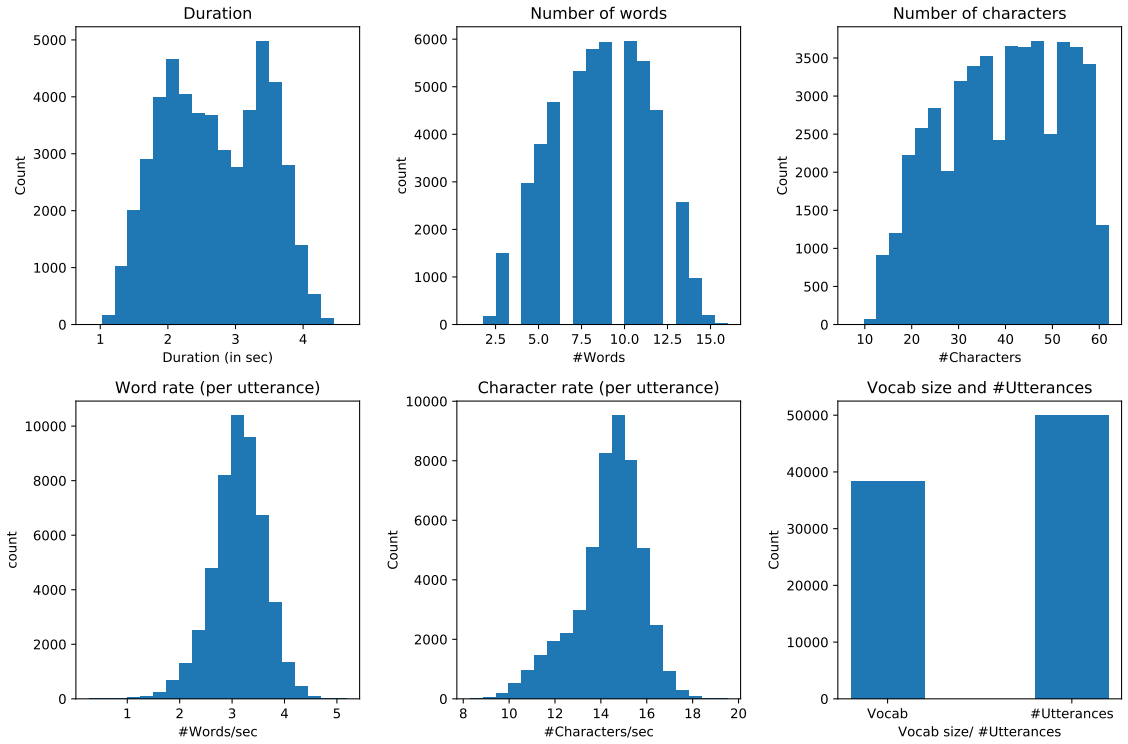

Figure A.3: Statistics of Google-synth dataset

In transcript files, each line represents a label for a single speech sample present in the clips directory. The first column in the line represents the path/name to the WAV file and the second column separated by a tab holds the actual transcript in text form. Figure A.2 demonstrates that the majority of audio recordings fall within the 2 to 15 second range, with an average duration of 5 to 7 seconds. On average, these recordings contain approximately 10 words and 45 characters and are spoken at a rate of 0.5 to 3 words per second. The Punjabi Speech dataset includes a vocabulary of 6281 words.

A.2.2 *Google-synth*

The Google-synth dataset is a synthetic dataset produced using Google's Text-to-speech (TTS) APIs. By using a TTS engine, it is possible to generate large volumes of speech

data in a relatively short amount of time. This can be especially useful for training speech recognition models, where the annotated speech data in the target language is limited, for example, in the case of Punjabi. Speech synthesis eliminates the need for human transcribers, reducing the cost and time associated with manual data collection and annotation.

Google Text-to-Speech supports a wide range of languages and voices, and users can customize the speech rate, pitch, and volume to suit their preferences. For Punjabi (*language code="pa-Guru-IN"*), Google offers TTS models in four different voices (2 male and 2 female). We carefully selected around 50K sentences from Old Newspapers dataset[2] for synthesis. With 50K utterances, we generated about 38 hours of speech. We synthesize speech at 44kHz with a similar directory style as the Punjabi Speech dataset. Figure A.3 demonstrates the dataset statistics for the Google-synth dataset. The majority of the utterances in the dataset are between 2 and 4 seconds in duration, with a total range spanning from 1 to 4 seconds. On average, each utterance contains approximately 8 words. The rate of speech is estimated to be roughly 3 words per second or 15 characters per second. The dataset encompasses a vocabulary of 38281 words.

A.2.3 *CMU-synth*

We produce a CMU-synth dataset using CMU's Clustergen TTS model [12]. Clustergen is a statistical parametric model that is incorporated within the Festival Speech Synthesis system[3]. We train the Clustergen TTS model from scratch using CMU INDIC Punjabi dataset[4]. The dataset comprises 0.4 hours of annotated speech data from a single female speaker. We generated around 80K utterances, which translates to 170 hours of speech data.

2 https://www.kaggle.com/alvations/old-newspapers
3 https://www.cstr.ed.ac.uk/projects/festival
4 http://festvox.org/cmu_indic

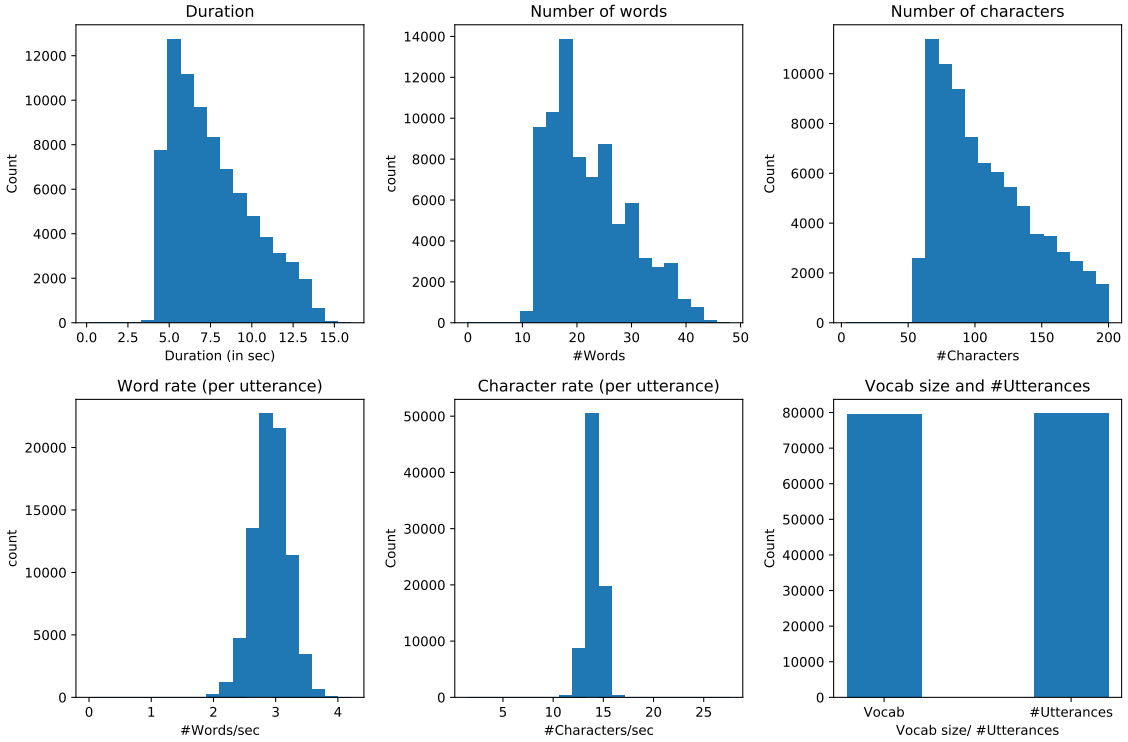

Figure A.4: Statistics of CMU-synth dataset

As shown in Figure A.4, on average, each utterance contains approximately 108/22 characters/words. The rate of speech is roughly around 3 words per second or 14 characters per second. The dataset contains a vocabulary of 79512 unique words.

A.2.4 *Audiobooks*

The Audiobooks dataset is an unlabeled Punjabi dataset comprising audio from multiple audiobooks available on YouTube. This dataset is created for pre-training or self-training purposes. To prepare this dataset we use multiple Python libraries and tools. Firstly, we manually listened to multiple audiobooks available on YouTube to examine the quality of the audio. We ensure that we only include high-quality audios in our dataset. After

shortlisting the high-quality audios, we use `pytube`[5] Python library to download audios in bulk mode.

Further, all the audiobooks are lengthy in nature, which is not practical while training deep learning models. To make them more usable, we use pre-trained voice activity detection (VAD) available under `Silero models`[6] to chunk the audio into smaller utterances. The Silero VAD model works perfectly to chunk the audio where it finds the speech part, ignoring the presence of noise and music in the audio file. Overall, we chunked around 322K utterances, which translates to 450 hours of high-quality unlabeled data.

A.3 CONCLUSIONS

We present multiple datasets to deal with data scarcity in the Punjabi language. We recorded a Punjabi Speech labeled dataset in mixed environments. Besides, we created two new synthesized datasets namely, Google-synth and CMU-synth, leveraging the available Punjabi text. We show the effectiveness of our labeled and synthesized datasets in Chapter 4. We also compiled an unlabeled Punjabi Audiobooks dataset, which could be used for self-training or pretraining self-supervised models. Our empirical evaluation presented in Chapter 5 validates the effectiveness of our Audiobooks dataset.

REFERENCES

[1] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

5 https://pytube.io/en/latest/
6 https://github.com/snakers4/silero-models

[2] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[3] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27 403, 1993.

[4] Anthony Rousseau, Paul Deléglise, and Yannick Estève, "Ted-lium: An automatic speech recognition dedicated corpus," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012, pp. 125–129.

[5] Douglas B Paul and Janet Baker, "The design for the wall street journal-based CSR corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[6] John J Godfrey, Edward C Holliman, and Jane McDaniel, "Switchboard: Telephone speech corpus for research and development," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE Computer Society, vol. 1, 1992, pp. 517–520.

[7] Pete Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018. arXiv: 1804.03209.

[8] Lucas Ondel, Pierre Godard, Laurent Besacier, Elin Larsen, Mark Hasegawa-Johnson, Odette Scharenborg, Emmanuel Dupoux, Lukas Burget, François Yvon, and Sanjeev Khudanpur, "Bayesian models for unit discovery on a very low resource language," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5939–5943.

[9] Satwinder Singh, Ruili Wang, and Feng Hou, "Punjabi Speech: A labeled Speech Corpus," Jul. 2023. DOI: 10.17632/sdbc8f5b77.1. [Online]. Available: https://data.mendeley.com/datasets/sdbc8f5b77/1.

[10] Satwinder, Ruili Wang, and Feng Hou, "Google-synth: A Synthesized Punjabi Speech Dataset," Jul. 2023. DOI: 10.6084/m9.figshare.23615607.v1. [Online]. Available: https://figshare.com/articles/dataset/Google-synth_A_Synthesized_Punjabi_Speech_Dataset/23615607.

[11] Satwinder Singh, Ruili Wang, and Feng Hou, "CMU-synth: A synthesized Punjabi Speech dataset," Jun. 2023. DOI: 10.6084/m9.figshare.23606697.v1. [Online]. Available: https://figshare.com/articles/dataset/_strong_CMU-synth_A_synthesized_Punjabi_Speech_dataset_strong_/23606697.

[12] Alan W Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *Proc. International Speech Communication Association (Interspeech)*, 2006, pp. 1762–1765.

B

STATEMENT OF CONTRIBUTION

I confirm that the *Statement of Contribution to Doctoral Thesis Containing Publications (DRC16)*, have been completed for each published article within the thesis, and are bound into the thesis and included in the electronic copy.

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:	Satwinder Singh
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:

Satwinder Singh, Ruili Wang, and Yuanhang Qiu, "DeepF0: End-to-end fundamental frequency estimation for music and speech signals", In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, pp. 61-65, 2021.

( ) 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:	Satwinder Singh Digitally signed by Satwinder Singh Date: 2023.04.12 08:58:27 +05'30'
Date:	12-Apr-2023
Primary Supervisor's Signature:	Prof Ruili Wang Digitally signed by Prof Ruili Wang DN: cn=Prof Ruili Wang, c=NZ, o=Massey University , ou=School of Natural and Computational Sciences , email=ruili.wang@massey.ac.nz Date: 2023.04.13 17:14:51 +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.

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:	Satwinder Singh
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:

Satwinder Singh, Ruili Wang, and Feng Hou, "Improved Meta Learning for Low Resource Speech Recognition" In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, pp. 4798-4802, 2022.

( ) 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:	Satwinder Singh Digitally signed by Satwinder Singh Date: 2023.04.12 08:59:19 +05'30'
Date:	12-Apr-2023
Primary Supervisor's Signature:	Prof Ruili Wang Digitally signed by Prof Ruili Wang DN: cn=Prof Ruili Wang, c=NZ, o=Massey University , ou=School of Natural and Computational Sciences , email=ruili.wang@massey.ac.nz Date: 2023.04.13 17:14:08 +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.

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:	Satwinder Singh
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:

-Proposed a end-to-end-approach for speech recognition of low-resource languages exploiting synthesized data.
-Produced 2 synthesized datasets and one real-speech labeled speech dataset

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

Candidate's Signature:	Satwinder Singh Digitally signed by Satwinder Singh Date: 2023.04.12 09:00:56 +05'30'
Date:	12-Apr-2023
Primary Supervisor's Signature:	Prof Ruili Wang Digitally signed by Prof Ruili Wang DN: cn=Prof Ruili Wang, c=NZ, o=Massey University , ou=School of Natural and Computational Sciences , email=ruili.wang@massey.ac.nz Date: 2023.04.13 17:13:01 +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.

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:	Satwinder Singh
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:

Data in Brief

- 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:

- Produced 2 synthesized datasets and one real-speech labeled speech dataset that are available in public domain
- Compiled real speech dataset, which is also available in public domain

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

Candidate's Signature:	Satwinder Singh Digitally signed by Satwinder Singh Date: 2023.07.26 18:31:36 +12'00'
Date:	26-Jul-2023
Primary Supervisor's Signature:	Prof Ruili Wang Digitally signed by Prof Ruili Wang DN: cn=Prof Ruili Wang, c=NZ, o=Massey University , ou=School of Natural and Computational Sciences , email=ruili.wang@massey.ac.nz Date: 2023.07.26 18:35:00 +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.

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:	Satwinder Singh
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:

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

- The name of the journal:

International Speech Communication Association (INTERSPEECH) 2023

- 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:

-Proposed a new self-training approach for automatic speech recognition of low-resource languages.
-Implemented and validated the effectiveness of proposed approach.

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

Candidate's Signature:	Satwinder Singh Digitally signed by Satwinder Singh Date: 2023.04.12 09:02:07 +05'30'
Date:	12-Apr-2023
Primary Supervisor's Signature:	Prof Ruili Wang Digitally signed by Prof Ruili Wang DN: cn=Prof Ruili Wang, c=NZ, o=Massey University , ou=School of Natural and Computational Sciences , email=ruili.wang@massey.ac.nz Date: 2023.04.13 17:12:25 +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.