Effect of Speech Modifications on Wav2vec2 Models for Children Speech Recognition

Semester-II Progress Presentation

by

Abhijit Sinha

Supervisor : Dr. Hemant Kumar Kathania

Department of Electronics and Communication Engineering,
National Institute of Technology Sikkim

August 2, 2024

Contents

Introduction

- Automatic Speech Recognition (ASR) technology has advanced significantly, improving performance for adult speech.
- But they struggle with childrens due to Data scarcity and the high variability in children's speech.
- Children's speech varies significantly across different age groups.
- The study aims to evaluate the effectiveness of speech modification techniques in enhancing the performance of Wav2Vec2 models when recognizing children's speech.

Literature Survey

Table 1: Literature Survey on related work.

Authors	Year	Title	Database	Method	Performance
Baevski et al.	2020	wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations	Librispeech	self-supervised learning of representations from raw audio data	1.8/3.3 WER on the clean/other test sets of Librispeech dataseet
Jain et al.	2023	A Wav2vec2-Based Experimental Study on Self-Supervised Learning Methods to Improve Child Speech Recognition	MyST,PFSTAR and the CMU Kids dataset	Fine-Tuned wav2vec2 models for children speech recognition	7.42 on the MyST dataset, 2.91 on the PFSTAR dataset and 12.47 on the CMU KIDS dataset
Jain et al.	2023	Adaptation of Whisper models to child speech recognition	MyST,PFSTAR and the CMU Kids dataset	Fine-Tuned Whisper models for children speech recognition	12.22 on the MyST dataset, 2.98 on the PFSTAR dataset and 15.08 on the CMU KIDS dataset
Barcovschi et al.	2023	A comparative analysis between Conformer-Transducer, Whisper, and wav2vec2 for improving child speech recognition	MyST,PFSTAR and the CMU Kids dataset	Adapted Conformer-transducer models to child speech	13.61 on the MyST dataset, 4.3 on the PFSTAR dataset% and 21.21 on the CMU KIDS dataset
Abion et al.	2023	Comparison of Data Augmentation Techniques on Filipino ASR for Childrenâs Speech	Filipino Children’s Speech Corpus	Spectral warping, vocal tract length perturbation, spectrogram augmentation and MaskCycleGAN-VC	43.55% relative improvement with respect to the baseline system.

Objective

- To analyse the impact of Wav2Vec2 models of different sizes and training data volumes.
- To evaluate the impact of speech modification techniques, including pitch, speaking rate, and formant modification on Wav2vec2 performance for children's speech.

Experimental Setup

- **Datasets**:
  - PF-STAR: British English children's speech corpus.
  - CMU Kids: American English children's speech corpus.
- **Models**: Six distinct Wav2Vec2 models, including base and large English models, and multilingual models from Facebook's Massive Multilingual Speech (MMS) project.
- **Speech Modification Methods**: Pitch, speaking rate, and formant modifications applied to children's speech to normalize it towards adult speech characteristics.

Dataset Description

- **PF-STAR**:
  - British English children's speech corpus.
  - Age range: 4-13 years.
  - Total duration: 9.4 hours.
  - Training set: 8.3 hours from 122 speakers.
  - Test set: 1.1 hours from 60 speakers (28 females).
- **CMU Kids**:
  - American English children's speech corpus.
  - Age range: 6-11 years.
  - Total duration: 9 hours.
  - Contributions from 24 male and 52 female speakers.
  - Total of 5180 utterances.

Wav2Vec2 Overview

- Wav2Vec2 is a state-of-the-art, self-supervised learning model for speech recognition.
- Consists of a convolutional feature encoder and a Transformer-based context network.
- Pre-trained on unlabeled audio to learn speech representations.

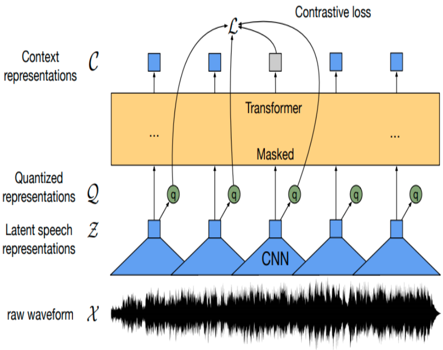

Fig : Wav2vec2 Architecture

- Advantages of Wav2Vec2 include its high accuracy, flexibility in adapting to various accents and speaking styles, and cost efficiency due to its self-supervised pre-training approach.

Wav2Vec2 Model Details

- Training data: LibriSpeech, Libri-Light, LibriVox, MMS datasets.

Table 2: Wav2Vec2 Model Details.

Model	Size (M)	Pretraining (h)	Finetuning (h)
Base-100h	95	960	100
Base-960h	95	960	960
Large-960-lv60	317	60K	960
Large-960-lv60-self	317	60K	960
1b-fl102	1B	491K	1.4K
1b-all	1B	491K	45K

Speech Modification Methods

- **Pitch Modification (PM):** Implemented the Real-Time Iterative Spectrogram Inversion with Look-Ahead (RTISI-LA) technique to adjust the pitch of children's speech to better match the pitch range of adults.
- **Speaking Rate Modification (SR):** Modified the speaking rate by varying the speed of the speech signal to align with the faster speaking rates typically observed in adult speech.
- **Formant Modification (FM):** Utilized a linear prediction-based method to adjust the formant frequencies of children's speech, which differ significantly from those of adults, to improve the recognition accuracy of ASR models.

Results

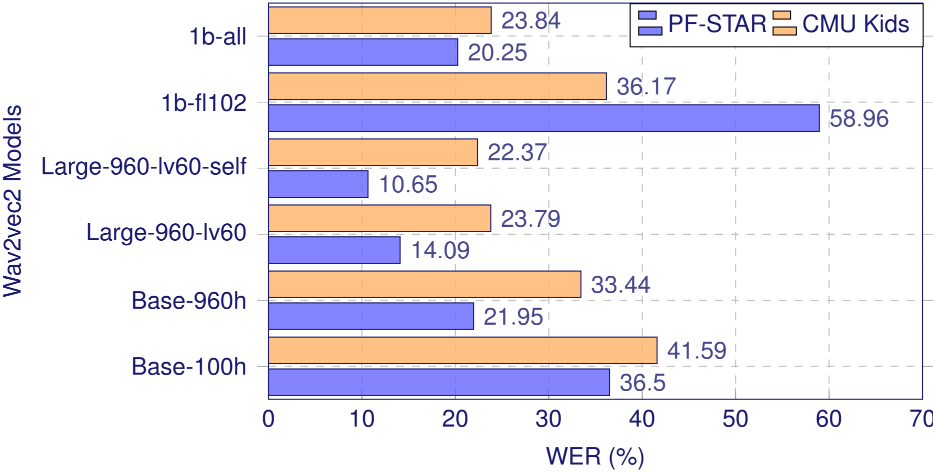

Figure 1: Performance on Unmodified Speech for PF-STAR and CMU Kids datasets

Speech Modification Impact

Table 3: WER (%) for different speech modifications on PF-STAR dataset

Model	Baseline	Pitch Modification	Speaking Rate	Formant Modification
base-100h	36.5	35.08	36.56	**32.71**
base-960h	21.95	22.5	22.74	**21.08**
large-960-lv60	14.09	15.11	15.37	**13.85**
large-960-lv60-self	10.65	11.46	11.09	**10.41**
1b-fl102	58.96	59.24	59.61	**58.57**
1b-all	**20.25**	28.21	23.84	22.90

Table 4: WER (%) for different speech modifications on CMU Kids Dataset

Model	Baseline	Pitch Modification	Speaking Rate	Formant Modification
Base-100h	41.59	42.13	42.76	**39.77**
Base-960h	33.44	34.25	35.44	**32.36**
Large-960-lv60	23.79	24.71	25.50	**24.41**
Large-960-lv60-self	**22.37**	23.20	23.80	22.63
1b-fl102	**36.17**	39.63	39.41	36.77
1b-all	**23.84**	25.65	24.70	**23.84**

Age Group Analysis

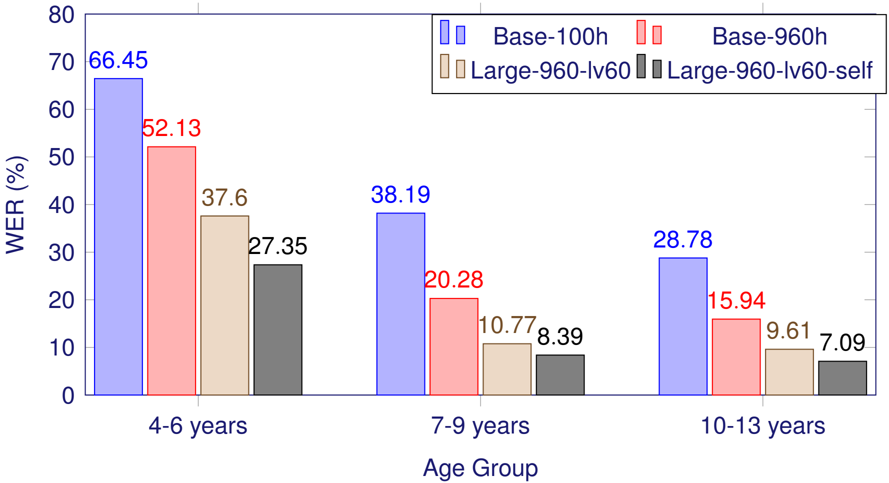

Figure 2: Comparison of model performance by age group for PF-STAR dataset

Age Group Analysis

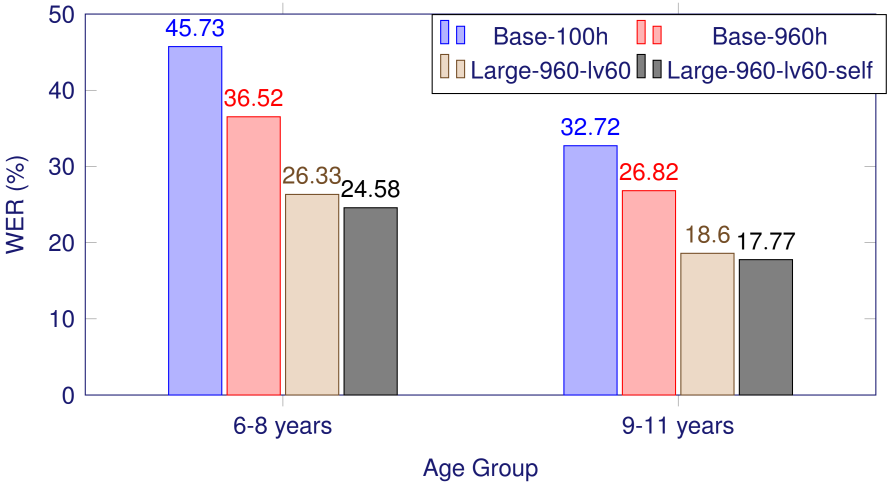

Figure 3: Comparison of model performance by age group for CMU Kids dataset

Combinations

Table 5: WER (%) for combinations of speech modifications on PF-STAR dataset

Models	Baseline	FM+PM	FM+SR	PM+SR	FM+PM+SR
base-100h	36.5	37.89	34.57	**33.05**	33.27
base-960h	21.95	23.57	22.38	**21.22**	22.15
large-960-lv60-self	10.41	11.01	10.77	**10.16**	10.49

Table 6: WER (%) for combinations of speech modifications on CMU Kids dataset

Models	Baseline	FM+PM	FM+SR	PM+SR	FM+PM+SR
base-100h	41.59	42.48	41.33	41.43	**41.31**
base-960h	**33.44**	35.21	35.00	34.24	35.18
large-960-lv60-self	23.84	23.99	23.80	**23.33**	23.57

Combinations (Age group wise)

Table 7: WER (%) for age group 4-6 of PF-STAR dataset

Models	Baseline	FM+PM	FM+SR	PM+SR	FM+PM+SR
base-100h	66.45	61.96	61.53	**58.54**	59.82
base-960h	52.13	45.94	46.15	**42.30**	42.73
large-960-lv60-self	27.35	24.57	25.00	20.29	**19.23**

Table 8: WER (%) for age group 7-9 of PF-STAR dataset

Models	Baseline	FM+PM	FM+SR	PM+SR	FM+PM+SR
base-100h	38.19	36.08	31.90	30	**29.79**
base-960h	20.28	19.75	**19.49**	**19.49**	19.65
large-960-lv60-self	8.39	8.50	8.50	8.13	**7.55**

Table 9: WER (%) for age group 10-13 of PF-STAR dataset

Models	Baseline	FM+PM	FM+SR	PM+SR	FM+PM+SR
base-100h	28.78	32.69	30.02	**28.66**	29.88
base-960h	16.94	20.01	18.26	**16.21**	17.99
large-960-lv60-self	7.09	7.75	7.68	**7.02**	8.10

Combinations (Age group wise)

Table 10: WER (%) for age group 6-8 of CMU Kids dataset

Models	Baseline	FM+PM	FM+SR	PM+SR	FM+PM+SR
base-100h	45.73	45.76	44.82	45.13	**44.21**
base-960h	36.52	38.11	37.80	**37.23**	37.99
large-960-lv60-self	24.58	26.46	26.29	**25.49**	26.82

Table 11: WER (%) for age group 9-11 of CMU Kids dataset

Models	Baseline	FM+PM	FM+SR	PM+SR	FM+PM+SR
base-100h	32.72	35.45	33.87	**33.73**	35.01
base-960h	26.82	28.96	28.91	**27.85**	29.16
large-960-lv60-self	17.77	18.74	21.30	**18.71**	19.16

Conclusion and Future Work

- Speech modifications significantly enhance ASR performance for children's speech. Larger Wav2Vec2 models demonstrate higher robustness, likely due to extensive pretraining.
- Challenges include data scarcity, variability in children's speech, age-specific recognition difficulties, domain mismatch with modified speech, and potential benefits from fine tuning and integrating language models.
- Studied how fine-tuning pre-trained models with in-domain and out-domain data, including various speech modifications, affects childrenâs speech recognition performance.

Conclusion and Future Work

- Investigate the impact of different pre-trained features on the performance of keyword spotting systems for children.
- Develop robust models for both speech recognition and keyword detection in children's speech using features from pretrained ASR models.
- Examine how pre-trained features perform in age and gender classification and speaker verification to improve personalization and security in children's speech applications.

References

- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449-12460.
- Hsu, W.-N., Sriram, A., Baevski, A., Likhomanenko, T., Xu, Q., Pratap, V., Kahn, J., Lee, A., Collobert, R., Synnaeve, G., & Auli, M. (2021). Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. Interspeech 2021, 721-725.
- Jain, R., Barcovschi, A., Yiwere, M. Y., Bigioi, D., Corcoran, P., & Cucu, H. (2023). A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition. IEEE Access, 11, 46938-46948.
- Lee, S., Potamianos, A., & Narayanan, S. S. (1997). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. The Journal of the Acoustical Society of America, 105(3), 1455-1468.
- Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., & Auli, M. (2023). Scaling speech technology to 1,000+ languages.
- Russell, M. (2006). The pf-star British English children's speech corpus. The Speech Ark Limited.

References

- Xu, Q., Baevski, A., Likhomanenko, T., Tomasello, P., Conneau, A., Collobert, R., Synnaeve, G., & Auli, M. (2020). Self-training and pre-training are complementary for speech recognition.
- Eskenazi, M., Mostow, J., & Graff, D. (1997). The cmu kids corpus. Linguistic Data Consortium.
- Kathania, H. K., Kadiri, S. R., Alku, P., & Kurimo, M. (2022). A formant modification method for improved ASR of children's speech. Speech Communication.
- Shahnawazuddin, S., Adiga, N., Kathania, H. K., & Tarun Sai, B. (2020). Creating speaker independent ASR system through prosody modification based data augmentation. Pattern Recognition Letters.
- Zhu, X., Beauregard, G. T., & Wyse, L. L. (2007). Real-time signal estimation from modified short-time Fourier transform magnitude spectra. IEEE Transactions on Audio, Speech, and Language Processing.

Publications

1. Abhijit Sinha, Mittul Singh, Sudarsana Reddy Kadiri, Mikko Kurimo, Hemant Kumar Kathania, " **Effect of Speech Modifications on Wav2vec2 Models for Children Speech Recognition** ", accepted for publication in *IEEE International Conference on Signal Processing and Communications (SPCOM)*, 2024.

1. Vishaka Kumari, Abhijit Sinha, Hemant Kumar Kathania, " **Role of Acoustics and Prosodic Features for Children's Age Classification** ", accepted for publication in *IEEE International Conference on Signal Processing and Communications (SPCOM)*, 2024.

Thank You!