Report

DL Group

Group members:

• Miguel Marcano Da Costa Gomez (u2084335)

• Gülseren Merve Şafak (u314441)

• Ece Deniz Çevik (u809763)

• Wiktoria Agata Pawlak (u2106820)

• Hiwa Feizi (u2094203)

Data Loading and Preprocessing

Each WAV name encodes accent and gender (e.g., 3f_1021.wav → accent 3, female). Files were trimmed or zero-padded to 7 s (112 000 samples) and loaded with torchaudio. Model A (raw waveforms) used three augmentations: ±10 % time shift, 0.9–1.1× speed, and ±3 dB volume. Models B converted audio to 64-bin log Mel spectrograms (1024-pt FFT, hop 256), stored them as tensors, and indexed paths and labels in a CSV. A custom Dataset handled loading, and a stratified 80/20 split kept class balance.

Experiments

All models used a learning rate of 1 e-3, batch sizes of 64, and cross-entropy for loss function. Evaluation metrics for each model were the same such as accuracy, precision, recall, and F1 scores. Additionally, A1, B1, and B2 models used Adam optimiser while B3 used AdamW.

	Input	**Data Augmentation**	**Regularization (Dropout/BN/ Weight-decay)**
A1 (1D-CNN)	Raw 16 kHz waveform	True	0.30 / ✓ / 1 e-4
B1 (2D-CNN)	Mel-spectro gram	False	0.30 / ✓ / 3 e-6
B2 (2D-CNN +opt.)	Mel-spectro gram	True	0.30 / ✓ / 1 e-6
B3 (2D-CNN +attention)	Mel-spectro gram	True	0.40 / ✓ / 2 e-6

Table 1: Different experiment details for models

Proposed Architecture

Two stride 1D convolutions down-sample raw audio, retaining phase cues and giving a phase-aware model even though it increases parameters (Figure 1).

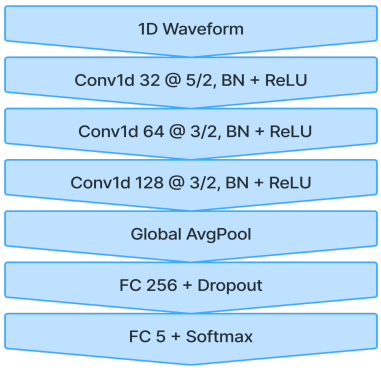

Figure 1: Raw Waveform 1D CNN - A1

Both models use 3x3 2D-CNN and 2x2 max-pool, preserving frequency detail while efficiently summarizing temporal context (Figure 2).

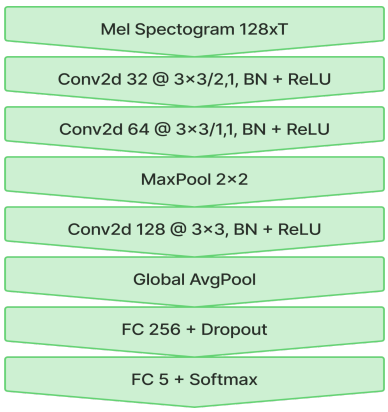

Figure 2: Spectrogram 2D CNN- B1/B2

A 1-head self-attention block is inserted after the convolution, capturing time-frequency dependencies with only small parameter overhead (Figure 3).

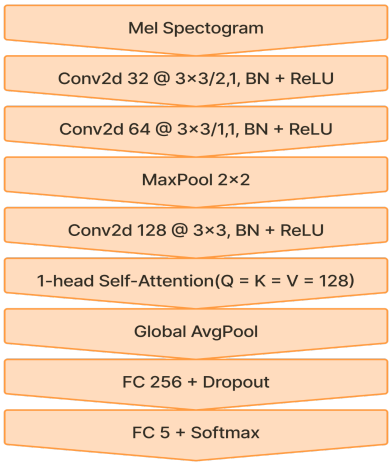

Figure 3: Spectrogram 2D CNN & Attention - B3

Results

Model A, a one dimensional convolutional network trained on raw waveforms, reached 67.3 % test accuracy after more than 60 epochs and roughly 50 to 60 minutes of training; its validation accuracy was 98.6 %, showing a 31 point generalization gap.

Mel spectrogram models B, B2, and B3 performed far better, with test accuracies of 96.0, 96.7, and 96.1 % and validation scores at or near 100 % (Table 2). They converged in 19 to about 68 epochs and 25 to 40 minutes, and their validation–test gaps narrowed to 3-4 points. Gender bias was negligible: Model A scored 99.0 percent on female voices and 98.1 percent on male, while Model B2 scored 100.0 versus 99.9 %.

Confusion analysis (Table A1) shows Model A exhibits the highest confusion rate (1.42%), primarily struggling with accent discrimination between accents 2, 4, and 5. In contrast Models B and B2 misclassified only a single accent 4→5 misclassification each (0.03% error rate). This analysis confirms that mel-spectrogram representations provide substantially clearer accent distinctions compared to raw waveform processing.

Model	Input	Test Acc. (%)	Val Acc. (%)
A (1D-CNN)	Raw waveform	67.3	98.6
B (2D-CNN)	Mel-spec	96.0	100.0
B2 (2D-CNN + opt.)	**Mel-spec**	**96.7**	**100.0**
B3 (2D-CNN + var.)	Mel-spec	96.1	~100.0

Table 2: Validation and test scores for all candidate models.

Conclusions

Our experiments show that Mel spectrogram inputs greatly outperform raw waveforms: accuracy rose from 67.3 % with Model A to 96.0 % with Model B. Accent 2, often misclassified by Model A, improved by more than 30 % points in Model B2. Once any Mel spectrogram was used, extra architectural complexity brought only minor gains, as Models B, B2, and B3 all scored between 96.0 % and 96.7 %.

The selected hyperparameters—learning rate, weight decay of 1×10^{-6}, and early stopping after 16 stagnant epochs—produced stable learning. Model B2 achieved 96.7 % accuracy and a macro F1 of 0.966, with only a 3.1 % gap between training and validation, confirming effective regularization and optimization. Further progress will likely come from richer attention mechanisms, deeper networks, or pretraining on larger audio corpora, together with high quality balanced datasets and cross validation.

Appendix

Appendix A — Supporting Tables

Model	Total Errors	Error Rate	Main Confusion Pattern
A	45	1.42%	Accent 2→4 (15 errors), 2→5 (6 errors)
B	1	0.03%	Accent 4→5 (1 error)
B2	1	0.03%	Accent 4→5 (1 error)
B3	13	0.41%	Accent 5→4 (8 errors), 5→2 (3 errors)

Table A1. Classification error patterns across all models on validation data.

References

A. Gu and T. Dao, “Mamba: Linear-Time Sequence Modeling with Selective State Spaces,” arXiv.org, Dec. 01, 2023. https://arxiv.org/abs/2312.00752 A. Gu, T. Dao,

S. Ermon, A. Rudra, and C. Re, “HiPPO: Recurrent Memory with Optimal Polynomial Projections,” arXiv.org, Aug. 17, 2020. https://arxiv.org/abs/2008.07669

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A simple data augmentation method for automatic speech recognition. In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2019) (pp. 2613–2617). International Speech Communication Association. https://doi.org/10.48550/arXiv.1904.08779

Wang, X., & Aitchison, L. (2025). How to set AdamW’s weight decay as you scale model and dataset size (Version 2) [Preprint]. arXiv. https://arxiv.org/abs/2405.13698

Wang, Y., Wang, Y., Chen, Y., Zhang, S., & Yu, Z. (2020). Torchaudio: Building audio applications in PyTorch [Computer software]. PyTorch. https://pytorch.org/audio