Report Deep Learning

Group 19

Group Members:

- Lars Heijnen (2097024)
- Lieke van Eijk (2102826)
- Aimélie Speet (2103752)
- Fleur Sülter (2103769)

Data loading and pre-processing

We worked with a dataset of single-channel, 16kHz audio files, each labelled by accent (1–5) and gender (m/f). Recordings varied in length, so we standardized all inputs to a fixed length (208 samples for waveforms, 208 timesteps for spectrograms) by cropping or zero-padding. This enabled efficient batching in PyTorch and ensured our models could process any input length at inference by dynamic padding. We chose this approach as it is standard in audio deep learning and avoids errors from variable-length inputs.

For Approach A (raw waveform), we standardized each waveform to zero mean and unit variance.
For Approach B (spectrogram), we computed log-mel spectrograms (n_fft=1024, hop_length=256, n_mels=64), then applied log-scaling. The mel scale was selected because it better reflects human auditory perception, emphasizing frequency bands relevant for speech and accents.
Gender labels were not used as model input to avoid information leakage; instead, we analyzed bias post-hoc as required by the assignment.

Experiments

We compared two main approaches and several regularization settings:

- Approach A: 1D CNNs on raw waveforms, with/without batch normalization and dropout (p=0.3, 0.5).
- Approach B: 2D CNNs on log-mel spectrograms, with/without batch normalization and dropout (p=0.3, 0.5).

All models used the same training configuration: Adam optimizer (lr=0.001), batch size 4, and cross-entropy loss, with a constant weight decay of 1e-4 applied as a default regularization technique. The weight decay is a deliberate choice, because this helps at a fundamental level to prevent overfitting by penalizing large weights in the network. Performance was evaluated using accuracy, precision, recall, and F1-score, both overall and broken down by gender and accent.

We applied data augmentation (e.g. adding noise, time-shifting, volume scaling). The amount by which each of these augmentations was applied was randomized. We introduced early stopping– continuing training only as long as validation performance improved. After comparisons, we proceeded with Approach B (2D CNNs on spectrograms), which showed the most promising results.

Proposed Architecture Visualizations

Approach A, employing 1D CNNs on raw waveforms (Figure 1), was chosen to enable the model to directly learn temporal dependencies and acoustic features inherent in the raw audio signal length capturing subtle accent variations without manual feature engineering. In contrast, Approach B utilized 2D CNNs on log-mel spectrograms (Figure 2), motivated by the effectiveness of spectrograms as visual representations of audio frequency content, allowing the 2D CNNs to exploit their spatial feature extraction capabilities to identify accent-specific patterns in the frequency domain and their evolution over time.

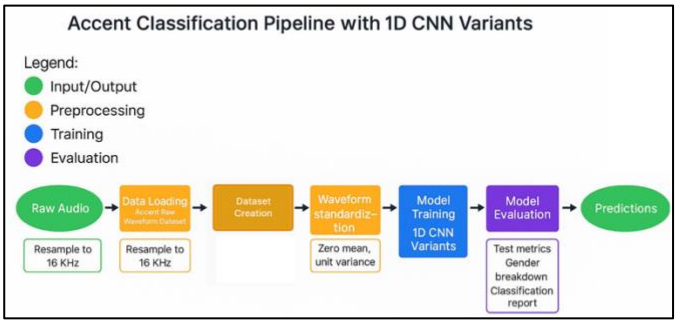


Figure 1: (Simplified) diagram of Approach A architecture.

Results

Models using Mel spectrograms (Approach B, Table 1) consistently outperformed raw waveform models (Approach A), Appendix Table 4 across all metrics. The best-performing model, CNNBaseline trained on augmented data –approach B-, demonstrated a predicted accuracy of 0.954 and F1 score of 0.952, outperforming the best waveform model CNNBaseline_BN trained on augmented data by a large margin (accuracy: 0.24, F1: 0.08).

Data augmentation proved highly effective, boosting spectrogram model accuracy by approximately 6.4% (from 0.89 to 0.954), confirming its critical role in robust model performance. Figure 3 highlights that data augmentation consistently improved accuracy across all models, confirming its value in reducing classification errors.

Overall, spectrogram-based models with augmentation provide a reliable solution for accent classification, while waveform models underperformed.

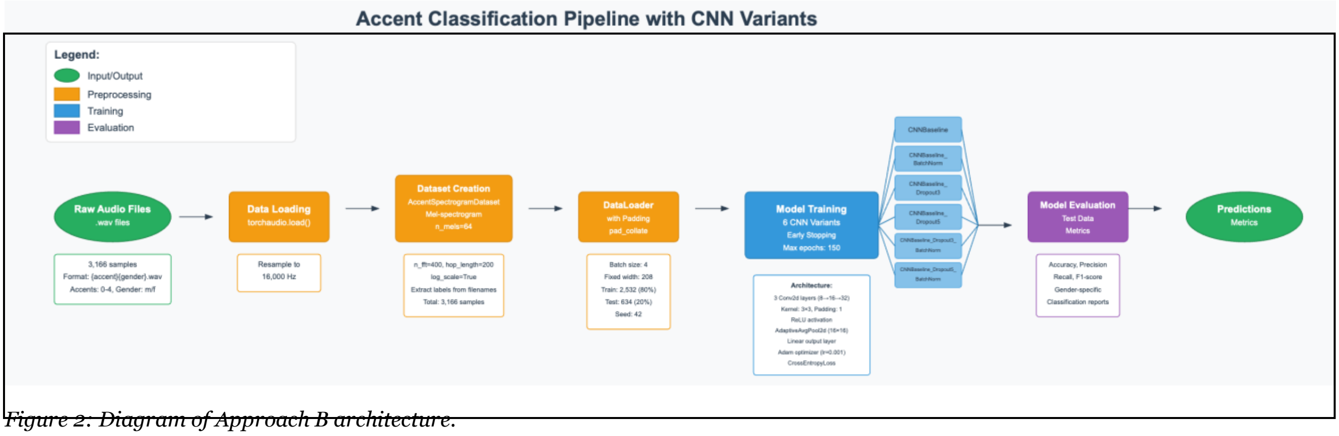
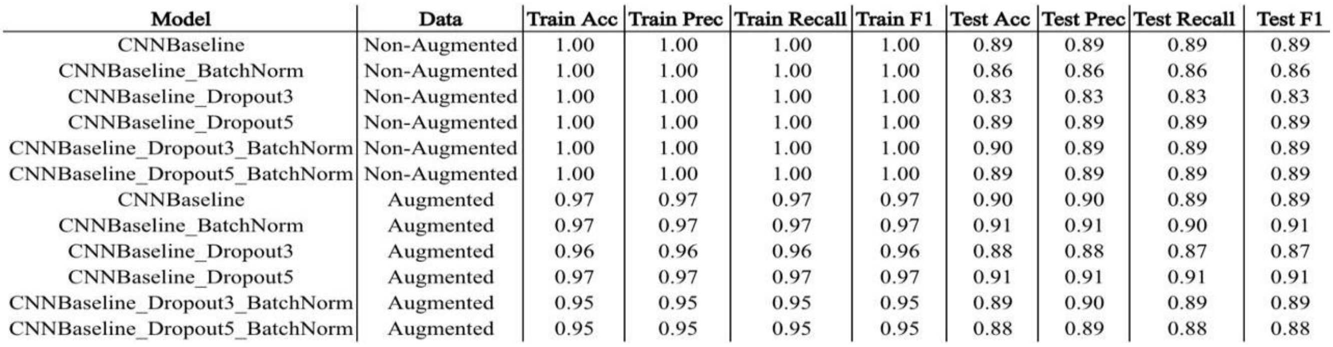


Figure 2: Diagram of Approach B architecture.



Model	Data	Train Acc	Train Prec	Train Recall	Train F1	Test Acc	Test Prec	Test Recall	Test F1
CNNBaseline	Non-Augmented	1.00	1.00	1.00	1.00	0.89	0.89	0.89	0.89
CNNBaseline_BatchNorm	Non-Augmented	1.00	1.00	1.00	1.00	0.86	0.86	0.86	0.86
CNNBaseline_Dropout3	Non-Augmented	1.00	1.00	1.00	1.00	0.83	0.83	0.83	0.83
CNNBaseline_Dropout5	Non-Augmented	1.00	1.00	1.00	1.00	0.89	0.89	0.89	0.89
CNNBaseline_Dropout3_BatchNorm	Non-Augmented	1.00	1.00	1.00	1.00	0.90	0.89	0.89	0.89
CNNBaseline_Dropout5_BatchNorm	Non-Augmented	1.00	1.00	1.00	1.00	0.89	0.89	0.89	0.89
CNNBaseline	Augmented	0.97	0.97	0.97	0.97	0.90	0.90	0.89	0.89
CNNBaseline_BatchNorm	Augmented	0.97	0.97	0.97	0.97	0.91	0.91	0.90	0.91
CNNBaseline_Dropout3	Augmented	0.96	0.96	0.96	0.96	0.88	0.88	0.87	0.87
CNNBaseline_Dropout5	Augmented	0.97	0.97	0.97	0.97	0.91	0.91	0.91	0.91
CNNBaseline_Dropout3_BatchNorm	Augmented	0.95	0.95	0.95	0.95	0.89	0.90	0.89	0.89
CNNBaseline_Dropout5_BatchNorm	Augmented	0.95	0.95	0.95	0.95	0.88	0.89	0.88	0.88

Table 1: Training and Testing Metrics of Models Trained on Augmented and on Non-Augmented Data (Approach B).

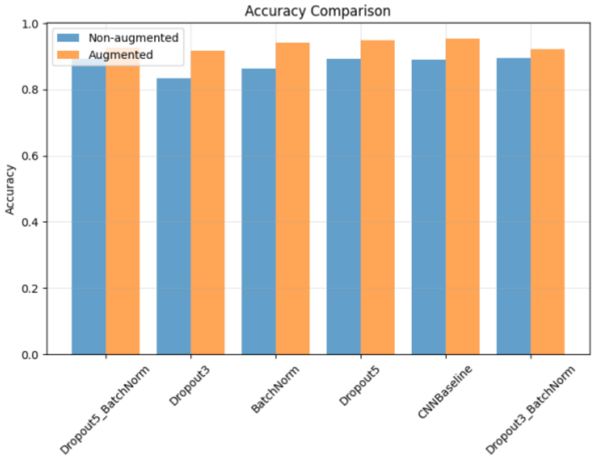
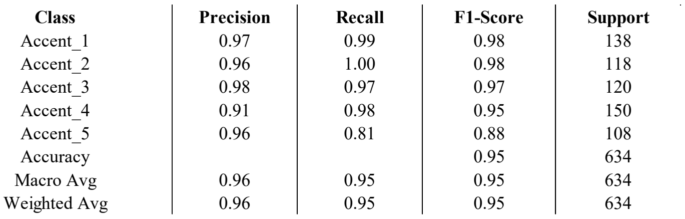


Figure 3: Depicting the impact of chosen approaches and regularization techniques on the test accuracy, comparing different models trained with and without data augmentation.

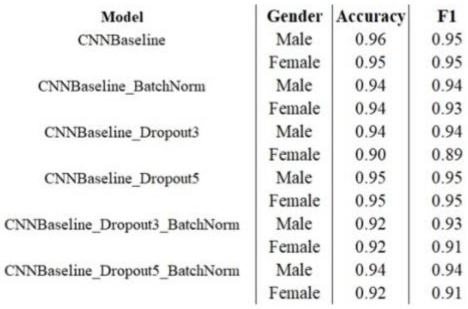
Error Analysis

As shown in Table 2, the model, CNNBaseline trained on augmented Mel spectrogram data (approach B), achieved strong overall performance but struggled most with Accent_5 (F1: 0.88, Recall: 0.81). Accent classification results for the top model were strong and balanced: Accent F1 scores: 0.88–0.98; Weighted F1: 0.954. Gender evaluation showed no bias, with male F1: 0.955 and female F1: 0.953 (Table 3).



Class	Precision	Recall	F1-Score	Support
Accent_1	0.97	0.99	0.98	138
Accent_2	0.96	1.00	0.98	118
Accent_3	0.98	0.97	0.97	120
Accent_4	0.91	0.98	0.95	150
Accent_5	0.96	0.81	0.88	108
Accuracy			0.95	634
Macro Avg	0.96	0.95	0.95	634
Weighted Avg	0.96	0.95	0.95	634

Table 2: Performance Evaluation of the CNN_Baseline Model (Best Model) for Accent Classification, Trained on Augmented Data and Evaluated on Non-Augmented Data.



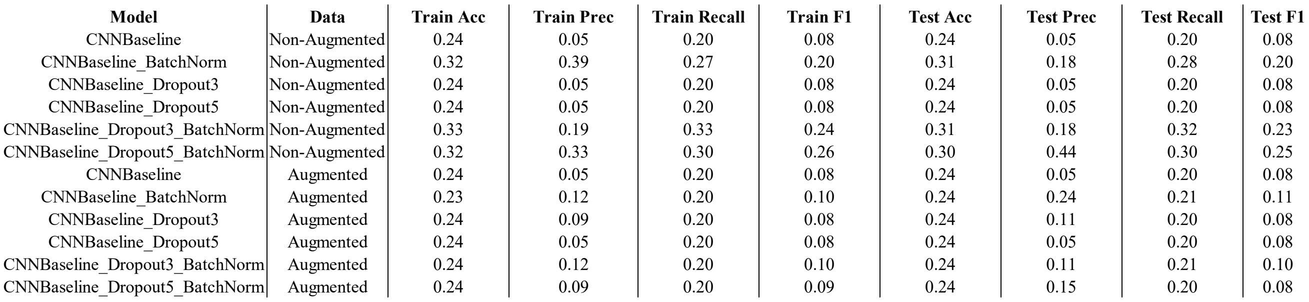
Model	Gender	Accuracy	F1
CNNBaseline	Male	0.96	0.95
	Female	0.95	0.95
CNNBaseline_BatchNorm	Male	0.94	0.94
	Female	0.94	0.93
CNNBaseline_Dropout3	Male	0.94	0.94
	Female	0.90	0.89
CNNBaseline_Dropout5	Male	0.95	0.95
	Female	0.95	0.95
CNNBaseline_Dropout3_BatchNorm	Male	0.92	0.93
	Female	0.92	0.91
CNNBaseline_Dropout5_BatchNorm	Male	0.94	0.94
	Female	0.92	0.91

Table 3: Performance Evaluation of all the models for Gender Classification, Trained on Augmented Data and Evaluated on Non-Augmented Data.

Conclusions

- **Raw Waveform Model Performance**: Models trained directly on raw waveforms consistently underperform across all metrics, even with the addition of regularization. This suggests that either more complex architectures or significantly larger datasets may be necessary to make waveform-based approaches competitive for accent classification.
- **Mel Spectrogram Superiority**: Mel spectrogram-based models outperform waveform models by a significant margin. Their superior performance is likely due to their ability to represent perceptually meaningful features of speech, making them particularly well-suited for tasks such as accent recognition.
- **Impact of Regularization**: The effect of regularization varies by input type:
  - **Waveform Models**: Batch normalization improves performance slightly (e.g., test accuracy from ~23.7% to ~30.8%), while dropout and combined methods offer no substantial gains.
  - **Spectrogram Models**: Regularization yields minor benefits without augmentation. However, with augmentation, the non-regularized CNNBaseline performs best (test accuracy = 0.954), indicating that data augmentation is a more effective form of regularization in this setting.
- **Effectiveness of Data Augmentation**: Augmenting the training data yields substantial performance gains for spectrogram models, boosting test accuracy from 0.89 to 0.95. This underscores the importance of augmentation in improving model generalization and robustness.
- **Gender Bias Evaluation**: Evaluation across male and female subsets shows nearly identical performance, with F1 scores differing by less than 0.01. This indicates that the best-performing models do not exhibit significant gender bias.
  - **Future Work:** To further improve performance, Deeper or more expressive network architectures
  - Integration of attention mechanisms
  - More sophisticated or diverse data augmentation strategies

Appendix



Model	Data	Train Acc	Train Prec	Train Recall	Train F1	Test Acc	Test Prec	Test Recall	Test F1
CNNBaseline	Non-Augmented	0.24	0.05	0.20	0.08	0.24	0.05	0.20	0.08
CNNBaseline_BatchNorm	Non-Augmented	0.32	0.39	0.27	0.20	0.31	0.18	0.28	0.20
CNNBaseline_Dropout3	Non-Augmented	0.24	0.05	0.20	0.08	0.24	0.05	0.20	0.08
CNNBaseline_Dropout5	Non-Augmented	0.24	0.05	0.20	0.08	0.24	0.05	0.20	0.08
CNNBaseline_Dropout3_BatchNorm	Non-Augmented	0.33	0.19	0.33	0.24	0.31	0.18	0.32	0.23
CNNBaseline_Dropout5_BatchNorm	Non-Augmented	0.32	0.33	0.30	0.26	0.30	0.44	0.30	0.25
CNNBaseline	Augmented	0.24	0.05	0.20	0.08	0.24	0.05	0.20	0.08
CNNBaseline_BatchNorm	Augmented	0.23	0.12	0.20	0.10	0.24	0.24	0.21	0.11
CNNBaseline_Dropout3	Augmented	0.24	0.09	0.20	0.08	0.24	0.11	0.20	0.08
CNNBaseline_Dropout5	Augmented	0.24	0.05	0.20	0.08	0.24	0.05	0.20	0.08
CNNBaseline_Dropout3_BatchNorm	Augmented	0.24	0.12	0.20	0.10	0.24	0.11	0.21	0.10
CNNBaseline_Dropout5_BatchNorm	Augmented	0.24	0.09	0.20	0.09	0.24	0.15	0.20	0.08

Table 4: Approach A, Training and Metrics of Models Trained on Non-Augmented and Augmented Data