Speaker Identification Using Machine Learning Techniques

Matthew Wilde
22135229

BIRMINGHAM CITY University

Introduction

This project compares traditional machine learning (ML) techniques and how they are applicable for a speaker identification task.
The 3 ML techniques covered are: K-Nearest-Neighbour, Support Vector Machines, and Convolutional Neural Network.

Dataset

The dataset consisted of 5 speakers, each with 5 recorded speech files (25 samples total). Due to the small dataset size, segmentation was applied to create more samples. Voice Activity Detection (VAD) isolated voiced regions, which were then split into 1-second segments, ensuring they exceeded a loudness threshold to avoid silent or low-energy samples disrupting model training.

Speaker	Samples Before Segmentation	Samples After Segmentation
BG	5	23
GS	5	41
HW	5	27
MW	5	41
SM	5	51
Overall	**25**	**183**

Feature Extraction for SVM & KNN

Both SVM and KNN use features extracted from the segmented audio samples. A range of features were selected to capture varying aspects of speech, including timbre, energy, pitch, and spectral characteristics. For each segment, the mean and standard deviation of every feature was calculated and concatenated to form the final feature vector, providing a compact summary of the audio content.

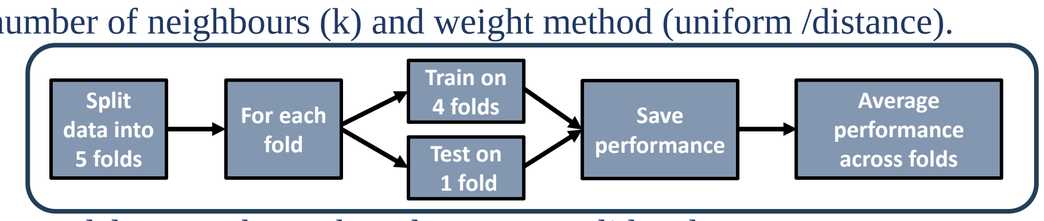
Feature	Purpose
MFCCs	Capture speech timbre characteristics.
ΔMFCC, Δ²MFCC	Capture how timbre changes over time.
Zero-Crossing Rate	Measure signal noisiness and voicing activity.
Spectral Centroid	Indicate brightness of the sound.
Spectral Bandwidth	Describe spread/width of spectral energy.
Spectral Rolloff	Identify brightness and high-frequency content.
Spectral Flatness	Distinguish tonal vs. noise-like sounds.
Spectral Contrast	Capture variations between spectral peaks/valleys.
Chroma Features	Capture pitch and harmonic content.
RMS Energy	Represent signal loudness.

Dimensionality Reduction

Principal Component Analysis was applied to the extracted feature set to reduce dimensionality while preserving the majority of the original information. PCA transformed the feature matrix from 183×128 to 183×54 dimensions whilst retaining 95% of total variance. This allowed the models to train more efficiently, reducing computational load and minimising overfitting without significant loss of descriptive power.

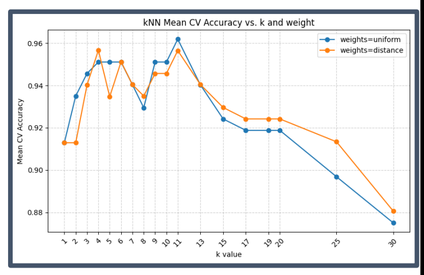
SVM & kNN Model Training

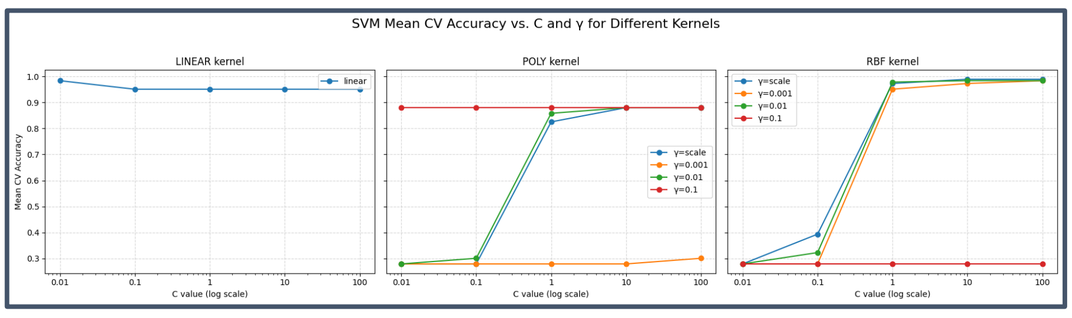
SVM and kNN classifiers were trained on PCA-reduced features using 5-fold Stratified Cross-Validation to optimise parameters.
SVM: kernel type (linear, polynomial, RBF), regularisation parameter (C), and gamma.
kNN: number of neighbours (k) and weight method (uniform /distance).



The best models were chosen based on cross-validated mean accuracy and evaluated on a separate 20% test set.

Technique	Best Parameters	F-Measure
SVM	RBF kernel, C=10, gamma=scale	0.9892
kNN	K=11, weights = uniform	0.9619



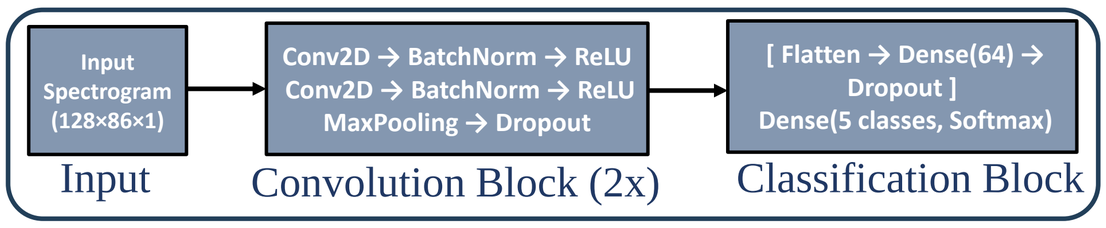


CNN Feature Extraction & Training

Mel-spectrograms were extracted and normalised from each voiced segment. The input spectrograms (128×86×1) were reshaped to fit a 2D CNN architecture. The CNN consisted of two convolutional blocks (Conv2D → BatchNorm → MaxPooling → Dropout) followed by dense classification layers. Training included:

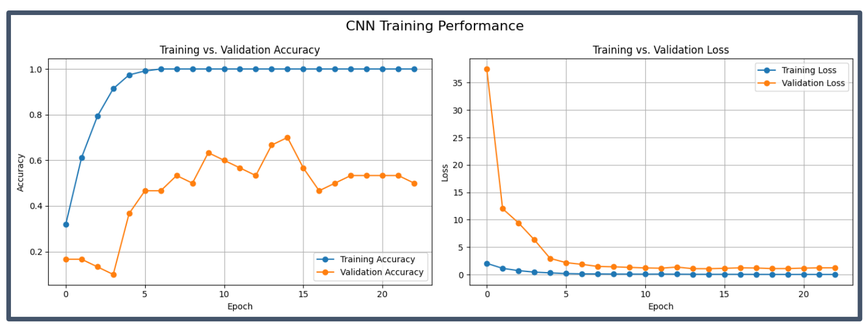
- Early Stopping to prevent overfitting.
- Model Checkpointing to save the best model.
- Validation Split (20%) during training to monitor performance.

Two convolutional blocks were used to progressively extract low/mid-level acoustic features from the mel-spectrograms before classification.



The model's training curves show rapid convergence, with training accuracy stabilising within 10 epochs. Validation accuracy plateaued at 70% at 15 epochs suggesting overfitting due to limited dataset size and diversity. Validation and training loss curves remained closely aligned, indicating minimal overfitting due to early stopping.

Early stopping was triggered once the model began to overfit, saving the best model at 15 epochs.



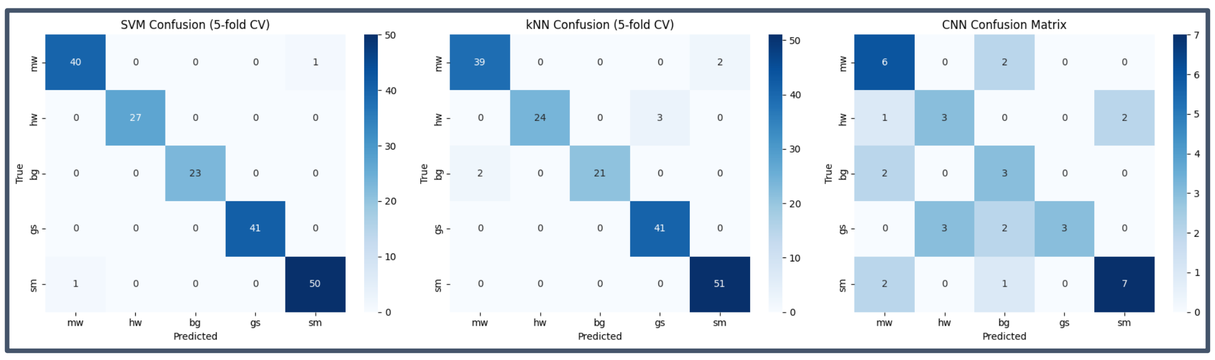
Model Evaluation & Comparison

Final model performances were evaluated using confusion matrices, classification reports, and accuracy scores on the cross-validation sets.

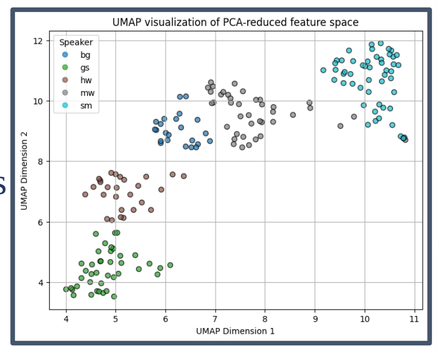
SVM achieved the highest overall test accuracy (98.9%), demonstrating exceptional generalisation on the PCA-reduced feature set.

kNN also performed strongly (96.2%), although slightly less robust than SVM on more ambiguous samples.

CNN achieved 70% accuracy, performing well at distinguishing broad speaker groups (e.g., male vs female) but struggling with precise speaker separation.



UMAP visualisation of PCA-reduced features shows distinct speaker clusters. SVM outperforms kNN due to more effective boundary modeling in complex feature spaces.



Overall, traditional ML models (SVM and kNN) outperformed the CNN, achieving test accuracies of 98.9% and 96.2% respectively compared to the CNN's 70%.These results highlight the importance of structured feature extraction and dimensionality reduction when working with small datasets, and suggest that deep learning methods like CNNs may require larger, more diverse data to fully realise their potential for speaker identification.

Classification Reports for Best Models

Classification Report for best CNN:

	precision	recall	f1-score	support
bg	0.38	0.60	0.46	5
gs	1.00	0.38	0.55	8
hw	0.50	0.50	0.50	6
mw	0.55	0.75	0.63	8
sm	0.78	0.70	0.74	10
accuracy			0.59	37
macro avg	0.64	0.58	0.58	37
weighted avg	0.68	0.59	0.60	37

5-Fold CV Classification Report for SVM:

	precision	recall	f1-score	support
mw	1.00	1.00	1.00	23
hw	1.00	1.00	1.00	41
bg	1.00	1.00	1.00	27
gs	0.98	0.98	0.98	41
sm	0.98	0.98	0.98	51
accuracy			0.99	183
macro avg	0.99	0.99	0.99	183
weighted avg	0.99	0.99	0.99	183

5-Fold CV Classification Report for kNN:

	precision	recall	f1-score	support
mw	1.00	0.60	0.75	5
hw	1.00	1.00	1.00	8
bg	1.00	1.00	1.00	6
gs	0.80	1.00	0.89	8
sm	1.00	1.00	1.00	10
accuracy			0.95	37
macro avg	0.96	0.92	0.93	37
weighted avg	0.96	0.95	0.94	37