SPEECH PROCESSING

Spoken Language Intent Detection

Introduction

Intent detection is a fundamental natural language understanding (NLU) task where in speech input is analysed to infer user's intention. This capability is essential for various AI-driven applications, such voice assistants (e.g., Siri, Alexa), automated customer support in various service domains, and assistive technologies for individuals with disabilities. Thus, robust intent detection models improve the systems' ability to comprehend user inputs and perform associated tasks. Traditional methods relied heavily on utilising the time-frequency domain and statistical approaches. However, the advent of deep learning and transformer-based language models have resulted in greater speech comprehension.

Traditionally, intent classification relies on textual inputs derived from automatic speech recognition (ASR). However, ASR errors and loss of acoustic information can degrade model performance. To overcome these limitations, this project proposes a multimodal intent detection model that integrates both textual and speech features for improved robustness and accuracy.

Supporting Literature

The authors in [2] proposed a multimodal deep learning model using both textual and acoustic data for its application in trauma-room scenarios. The data employed had almost 6424 audio-transcription pairs spanning 6 classes: Directive, Response, Question, Report, Clarification and Acknowledgement. The text-only and audio-only classifiers resulted in 57.36% and 59.64% accuracy during testing whereas the text-audio classifier exhibited 83.1% accuracy. The proposed multi-modal architecture surpassed its unimodal counterparts in accuracy.

This project has extended the use of the same architecture for its application in the banking-domain for in operating customer grievance chatbots.

Methodology

1. *Dataset:* The project makes use of the MINDS-14 dataset that comprises of speech utterances across multiple languages in the banking domain. The data columns include path, audio, transcription, English transcription, intent class and language ide. The intent label spans across 14 classes including address, ATM limit, business loans, card issues etc.
2. *Feature Extraction*

   Acoustic and textual feature extraction is carried out independently prior its fusion.

a. **Textual Features:** The transcriptions of the audio are first tokenized using the BERT architecture, which will serve as the input to the CNN that performs feature extraction on the text. The textual ConvNet embeds the tokens into a 768-dimensional space. Three parallel 1-D convolutional filters of sizes 1, 2, and 3 capture unigram, bigram, and trigram patterns; each filter's output is passed through a ReLU activation and reduced via adaptive max-pooling to a 300-dimensional vector. All three convolutional outputs are concatenated to yield a 900-dimentional representation.

b. **Acoustic Features:** Each raw audio file is processed to produce its Mel-spectrogram representation that encapsulates how the frequency changes over time mapped onto the Mel-scale enabling the model to capture patterns in pitch, tone, and intensity crucial for understanding spoken intent. To stabilize training, the resulting spectrogram is normalized by subtracting its mean and dividing by its standard deviation. The acoustic ConvNet performs a three-stage 2D convolutional network. ach stage applies a 3×3 convolution followed by ReLU and 2×2 max pooling, doubling the channel depth from 64 to 128 to 256 while halving the time–frequency resolution. The feature maps are then flattened to obtain a 1024-dimensional embedding.

c. **Feature Fusion:** Once separate embeddings are obtained, the model concatenates the 900-dimensional text vector with the 1,024-dimensional audio vector to form a single 1,924-dimensional tensor. This joint representation is passed through two successive fully connected layers, each followed by a ReLU activation: the first reduces dimensionality to 512, and the second to 256.

3. *Classification*

The final 256-dimensional vector is fed as input to the classifier, a linear transformation whose outputs will be raw logits corresponding to the intent predictions.

Experimental Results

1. *Data preprocessing*

As the project primarily aimed to perform speech intent detection for English, experimental data was curated by concatenating accented English data from American, Australian and British spoken English. There were 1809 samples which divided in a 7: 1.5 :1.5 ratio to create train, validation and test sets.

2. *Data Analysis*

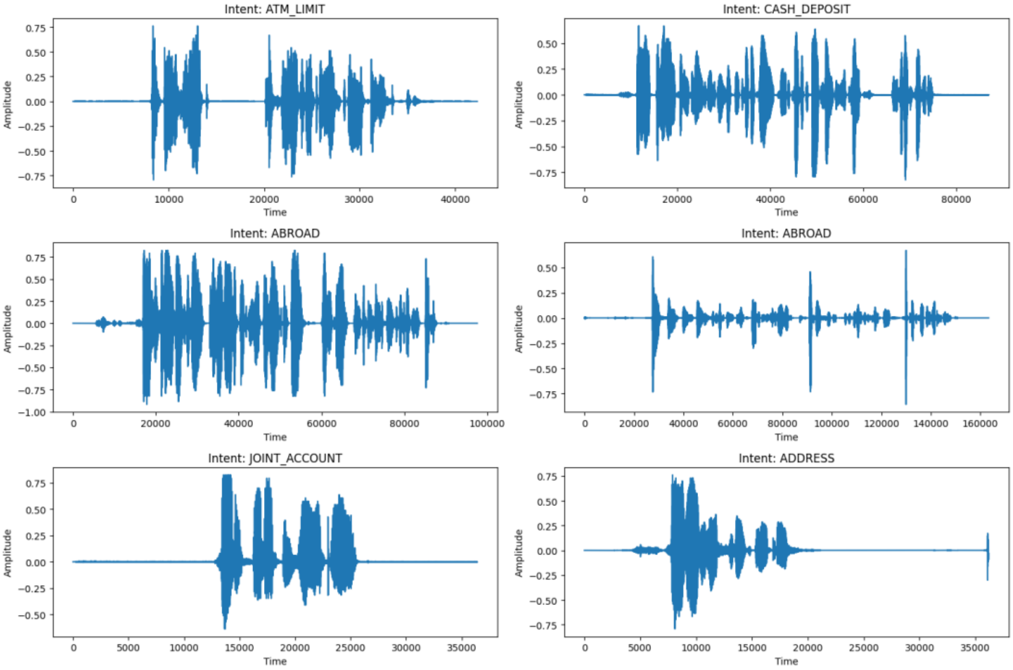


Figure 1: Amplitude-Time plots

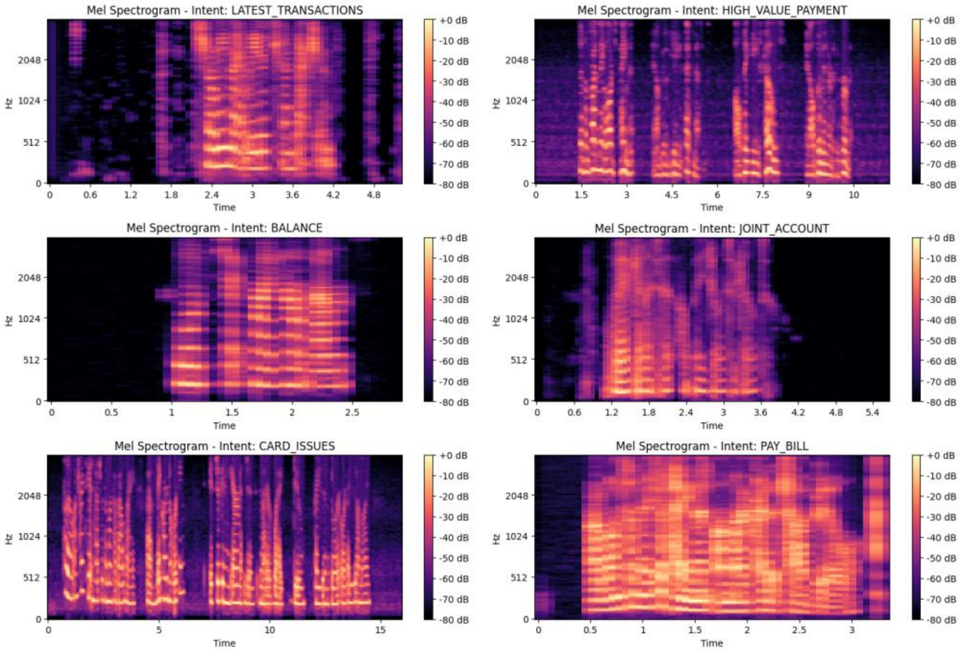


Figure 2: Mel-Spectrograms of sample audio

3. *Model Performance*

Owing to the size of the dataset and the complexity of the architecture, the model was trained for 10 epochs. Despite slight overfitting during the training phase, the model can perform extremely well on the test set with test accuracy of 94%.

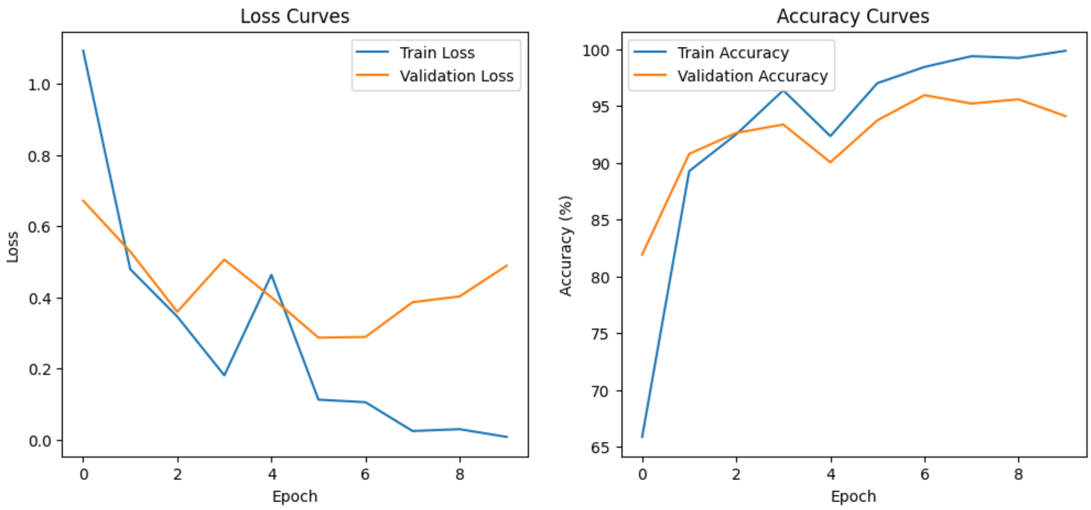
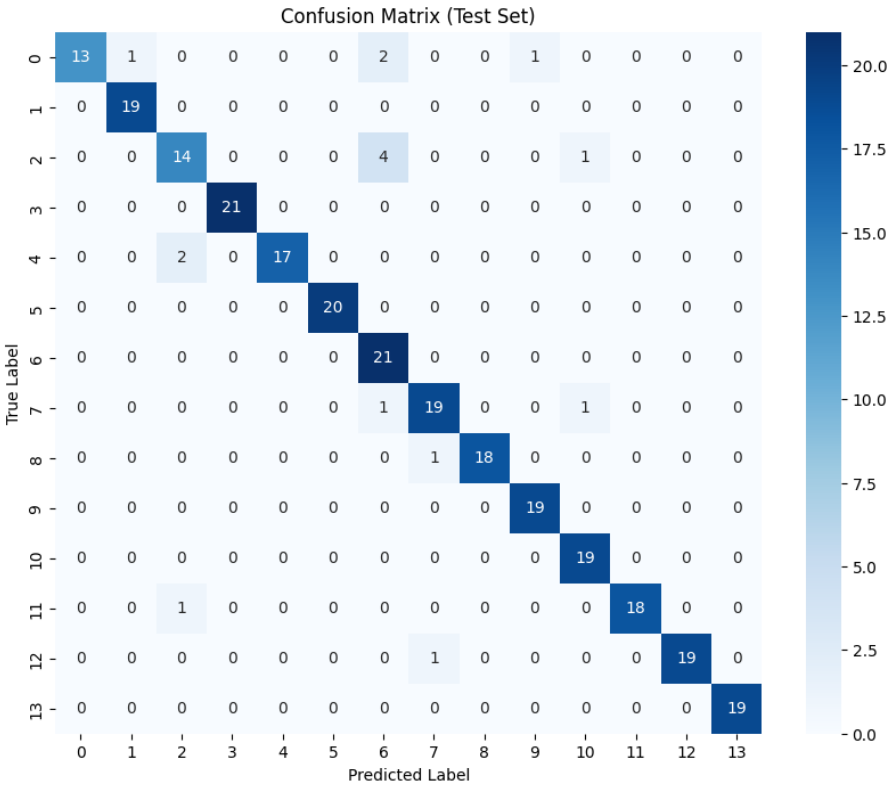


Figure 3: Visualization of Training Phase



	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	13	1	0	0	0	0	2	0	0	1	0	0	0	0
1	0	19	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	14	0	0	0	4	0	0	0	1	0	0	0
3	0	0	0	21	0	0	0	0	0	0	0	0	0	0
4	0	0	2	0	17	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	20	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	21	0	0	0	0	0	0	0
7	0	0	0	0	0	0	1	19	0	0	1	0	0	0
8	0	0	0	0	0	0	0	1	18	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	19	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	19	0	0	0
11	0	0	1	0	0	0	0	0	0	0	0	18	0	0
12	0	0	0	0	0	0	0	1	0	0	0	0	19	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	19

Figure 4: Class based distribution of the predictions

Conclusion:

The proposed project aims to improve intent detection by incorporating both textual and acoustic features during model training. This is intended such that it can have real-world impact by being able to adapt to different languages and better comprehend the intent of the speech. By bridging the gap between linguistic and acoustic understanding, the proposed approach brings voice-based AI systems closer to human-like comprehension, revolutionizing intelligent conversational interfaces.

References

[1] D. Gerz *et al.*, “Multilingual and Cross-Lingual Intent Detection from Spoken Data,” *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7468–7475, 2021, doi: https://doi.org/10.18653/v1/2021.emnlp-main.591.

[2] Y. Gu, X. Li, S. Chen, J. Zhang, and I. Marsic, “Speech Intention Classification with Multimodal Deep Learning,” *Lecture Notes in Computer Science*, pp. 260–271, 2017, doi: https://doi.org/10.1007/978-3-319-57351-9_30.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Proceedings of the 2019 Conference of the North*, 2019, doi: https://doi.org/10.18653/v1/n19-1423.

Student Details

Name: Sania Serrao

Reg No: 220962101

Subject: Speech Processing