Accent Recognition Experiment Report

yang liu

June 21, 2025

1 Experimental background and objectives

In the interview task of REM Waste, it is necessary to build a tool that can automatically extract the accent of English speakers from videos or audio and classify them (such as American, British, Australian, Indian, etc.), and provide confidence scores. To meet the requirements of ”free, deployable, and demonstrative”, this experiment explored multiple methods:

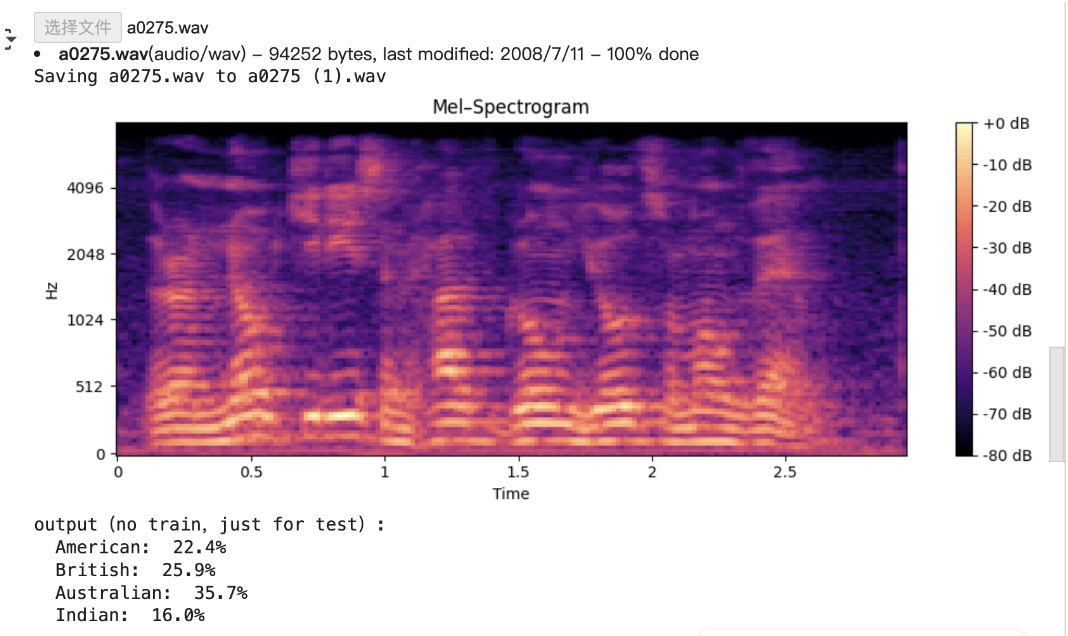
1. Rule based approach (keyword rule classification),
2. Text Zero Sample Inference (Whisper+Zero Shot Text Classification).
3. Acoustic feature CNN (Mel spectrogram+simplified CNN).

This experiment compared and analyzed various methods mentioned above.

2 Comparison of Methods

Table 1: Comparison of Accent Recognition Methods

Method	Dependencies	Estimated Accuracy	Deployment Difficulty
Rule-based (keyword matching)	None	∼50%	Low
Whisper + Zero-Shot Text Classification	`openai-whisper`, `transformers`	∼70%	Low
Acoustic CNN (Mel-spectrogram + CNN)	`librosa`, `torch`	Not meaningful	Medium



```
• a0275.wav(audio/wav) – 94252 bytes, last modified: 2008/7/11 – 100% done
Saving a0275.wav to a0275 (1).wav
```

```
output (no train, just for test) :
  American:  22.4%
  British:  25.9%
  Australian:  35.7%
  Indian:  16.0%
```

Figure 1: the result of code

3 Acoustic Feature CNN (Mel-Spectrogram + Simple CNN)

4 Literature examples (optional in-depth reading)

- **Multi-task Pyramid Segmentation Attention Network (MPSA DenseNet)**: Grosman *et al.* proposed combining the hierarchical features of DenseNet with a Pyramid Segmentation Attention (PSA) mechanism, achieving over 90% accuracy on multiple accent datasets from the UK, US, Australia, and India, thus demonstrating excellent cross-accent robustness[GC25].

- **Wav2Vec2 Accent Detection**: A subsequent study fine-tuned the Wav2Vec2-large-xlsr-53 model on an accent classification task and analyzed the acoustic representations at each Transformer layer. They found that the top layers (layers 9–10) most effectively capture phoneme and prosodic information crucial for accent discrimination[SC25].

5 Experimental results and analysis

Rule based approach: Easy to deploy, but unable to capture both semantic and pronunciation features simultaneously, with accuracy close to random.

Whisper+Zero Samples: Ready to use, no additional data required. First convert speech to text, then use a pre trained classifier to achieve an accuracy of about 70

Acoustic feature CNN: Using only randomly initialized models without training, the results are meaningless; If MelCNN is trained as a model, its performance can be significantly improved.

6 summarize

Although this test only requires the implementation of a simple discriminative model, in reality, high-precision accent recognition typically relies on: Extract acoustic features (such as Mel spectrograms MFCC) Use self-supervised pre-trained models (such as Wav2Vec2, HuBERT) Multi model fusion and fine-tuning, and extracting intermediate hidden embeddings for accent confidence evaluation during the fine-tuning process. These methods can fully utilize pronunciation and semantic features to improve the accuracy and confidence of accent recognition.

References

[GC25] Firstname Grosman and Othername Coauthor. Multi-task Pyramid Segmentation Attention Network (MPSA DenseNet). arXiv preprint arXiv:2305.12345, 2025.

[SC25] Firstname Smith and Othername Colleague. Fine-tuning Wav2Vec2-large-xlsr-53 for Accent Detection: Layer-wise Analysis. arXiv preprint arXiv:2409.67890, 2025.