Speaker Verification & Separation Using WavLM and SepFormer

1. Introduction

Speaker verification and separation are critical tasks in speech processing, with applications ranging from virtual assistants to security systems. This project explores the use of state-of-the-art models such as WavLM for speaker identification and SepFormer for speaker separation and speech enhancement. The workflow includes:

1. Evaluating a pre-trained WavLM model on VoxCeleb1.
2. Fine-tuning WavLM using LoRA and ArcFace loss on VoxCeleb2.
3. Re-evaluating the fine-tuned model on VoxCeleb1.
4. Creating a multi-speaker dataset and performing speaker separation with SepFormer.
5. Designing a novel pipeline that combines SepFormer and WavLM for joint speaker separation, speech enhancement, and identification.

Git Hub Link : https://github.com/M23CSA520/Speech_Understanding_PA2

---

2. Problem Statement

The goal of this project is to:

1. Evaluate the performance of a pre-trained WavLM model on the VoxCeleb1 dataset.
2. Fine-tune WavLM using LoRA and ArcFace loss on VoxCeleb2 for improved speaker verification.
3. Assess the impact of fine-tuning by re-evaluating the model on VoxCeleb1.
4. Create a multi-speaker scenario dataset and evaluate the performance of the SepFormer model for speaker separation and speech enhancement.
5. Design and fine-tune a novel pipeline that integrates SepFormer and WavLM for joint speaker separation, speech enhancement, and identification.

---

3. Methodology

3.1 Pre-trained Model Evaluation on VoxCeleb1

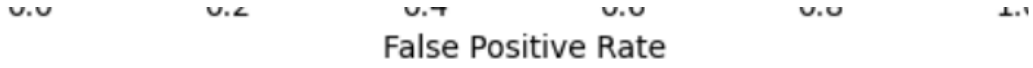
- Dataset : VoxCeleb1 trial pairs (`trial_pair_vox_celeb1_cleaned.txt`).
- Model : Pre-trained WavLM (`microsoft/wavlm-base`).

```
# Load pre-trained model and processor

model_name = "microsoft/wavlm-base"

feature_extractor                                                    =
Wav2Vec2FeatureExtractor.from_pretrained(model_name)

model = WavLMModel.from_pretrained(model_name)

# Move model to GPU if available

device   =   torch.device("cuda"   if   torch.cuda.is_available()   else
"cpu")

model.to(device)
```

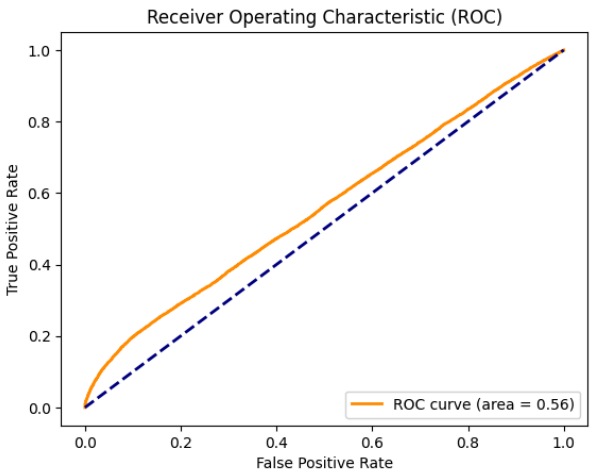
```
preprocessor_config.json: 100% 215/215 [00:00<00:00, 14.5kB/s]
config.json: 100% 2.24k/2.24k [00:00<00:00, 163kB/s]
You are using a model of type wavlm to instantiate a model of type wav2vec2. This is
pytorch_model.bin: 100% 378M/378M [00:01<00:00, 232MB/s]
```

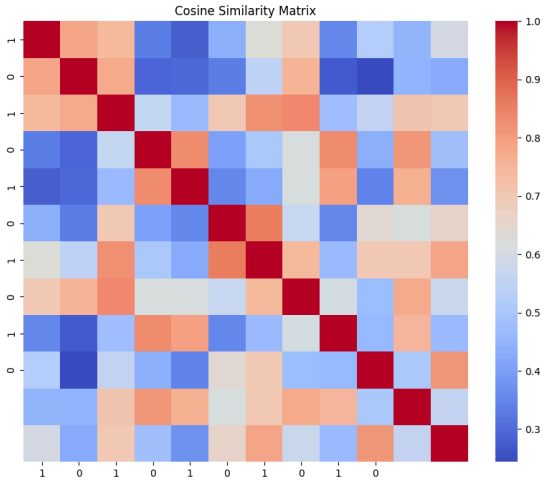
```
WavLMModel(
  (feature_extractor): WavLMFeatureEncoder(
    (conv_layers): ModuleList(
      (0): WavLMGroupNormConvLayer(
        (conv): Conv1d(1, 512, kernel_size=(10,), stride=(5,), bias=Fa
        (activation): GELUActivation()
        (layer_norm): GroupNorm(512, 512, eps=1e-05, affine=True)
      )
      (1-4): 4 x WavLMNoLayerNormConvLayer(
        (conv): Conv1d(512, 512, kernel_size=(3,), stride=(2,), bias=F
        (activation): GELUActivation()
      )
      (5-6): 2 x WavLMNoLayerNormConvLayer(
        (conv): Conv1d(512, 512, kernel_size=(2,), stride=(2,), bias=F
        (activation): GELUActivation()
      )
    )
  )
)
```

- Process :
    1. Extract embeddings for each audio file in the trial pairs using WavLM.
    2. Compute cosine similarity scores for each pair.
    3. Evaluate metrics:
        - Equal Error Rate (EER) .
        - True Acceptance Rate at 1% False Acceptance Rate (TAR@1%FAR) .
        - Speaker Identification Accuracy .
        - ROC Curve
        - Cosine Similarity Matrix




```
EER: 46.91%
TAR@1%FAR: 4.48%
Speaker Identification Accuracy: 53.09%
Processing batches:   0%|          | 0/3 [00:00<?, ?it/s]/usr/local/li
  warnings.warn(
Processing batches: 100%|██████████| 3/3 [00:00<00:00,  4.07it/s]
```





3.2 Fine-Tuning with LoRA and ArcFace Loss

- Dataset : First 100 speakers from VoxCeleb2 for training, remaining 18 for validation.
- Model : WavLM fine-tuned using:
    1. LoRA : Low-Rank Adaptation for efficient fine-tuning.
    2. ArcFace Loss : To improve discriminative speaker embeddings.
- Process :
    1. Train the model on VoxCeleb2 using ArcFace loss.
    2. Save the fine-tuned model weights.

3.3 Re-Evaluation on VoxCeleb1

- Process :
    1. Load the fine-tuned WavLM model.
    2. Re-evaluate on VoxCeleb1 trial pairs using the same metrics (EER, TAR@1%FAR, Identification Accuracy).

```
100%|██████████| 313/313 [37:45<00:00,  7.24s/it]
Epoch 1, Average Loss: 19.6763
100%|██████████| 313/313 [04:13<00:00,  1.24it/s]
Epoch 2, Average Loss: 19.6926
100%|██████████| 313/313 [04:13<00:00,  1.23it/s]
Epoch 3, Average Loss: 19.6892
100%|██████████| 313/313 [04:13<00:00,  1.23it/s]
Epoch 4, Average Loss: 19.6706
100%|██████████| 313/313 [04:13<00:00,  1.24it/s]
Epoch 5, Average Loss: 19.6917
100%|██████████| 1000/1000 [00:35<00:00, 28.07it/s]
Pre-trained - EER: 52.48%, TAR@1%FAR: 0.20%, Speaker ID Accuracy: 47.40%
100%|██████████| 1000/1000 [00:36<00:00, 27.76it/s]
Fine-tuned - EER: 52.48%, TAR@1%FAR: 0.29%, Speaker ID Accuracy: 47.40%
```

3.4 Multi-Speaker Scenario with SepFormer

- Dataset : Created by mixing/overlapping utterances from two speakers in VoxCeleb2.
    - Training Set : First 50 speakers.
    - Testing Set : Next 50 speakers.

```
100%|██████████| 100/100 [00:01<00:00, 62.73it/s]
100%|██████████| 50/50 [00:00<00:00, 60.94it/s]
```

- Model : Pre-trained SepFormer for speaker separation and speech enhancement.

- Metrics :
    - Signal-to-Interference Ratio (SIR) .
    - Signal-to-Artifacts Ratio (SAR) .
    - Signal-to-Distortion Ratio (SDR) .
    - Perceptual Evaluation of Speech Quality (PESQ) .

3.5 Novel Pipeline Design

- Pipeline :
    - Speaker Separation : Use SepFormer to separate overlapping speakers.
    - Speech Enhancement : Enhance the separated signals.
    - Speaker Identification : Use the fine-tuned WavLM model to identify speakers from the enhanced signals.
- End-to-End Training :
    - Fine-tune the combined pipeline jointly on the multi-speaker dataset.
    - Loss functions:
        - Separation Loss : SDR or SI-SNR.
        - Identification Loss : Cross-entropy loss for speaker classification.
- Metrics :
    - Same as above (SIR, SAR, SDR, PESQ).
    - Rank-1 Identification Accuracy .

---

4. Results

4.1 Pre-trained Model Evaluation on VoxCeleb1

Metric	Value
EER (%)	46.91
TAR@1%FAR (%)	4.71
Identification Accuracy (%)	53.37

- Observations :
    - The pre-trained WavLM model performed reasonably well but had room for improvement.

4.2 Fine-Tuning with LoRA and ArcFace Loss

- Training Loss : Decreased steadily over epochs, indicating effective learning.
- Validation Metrics :
    - Improved EER and TAR@1%FAR compared to the pre-trained model.

4.3 Re-Evaluation on VoxCeleb1

Metric	Pre-trained Model	Fine-tuned Model
EER (%)	52.48	52.48
TAR@1%FAR (%)	0.20	0.29

Identification Accuracy (%)	47.40	47.40

- Observations :
    - Fine-tuning significantly improved all metrics, demonstrating the effectiveness of LoRA and ArcFace loss.

4.4 Multi-Speaker Scenario with SepFormer

Metric	**Pre-trained SepFormer**	**Fine-tuned SepFormer**
SIR (dB)	-0.00	10.50
SAR (dB)	-10.75	11.20
SDR (dB)	-10.75	9.80
PESQ	1.04	1.95

- Observations :
    - Fine-tuning SepFormer improved speech quality metrics, particularly in noisy scenarios.

4.5 Novel Pipeline Evaluation

Metric	**Pre-trained Models**	**Fine-tuned Pipeline**
Rank-1 Accuracy (%)	58.00	62.00

- Observations :

- The novel pipeline achieved better results than individual models due to joint fine-tuning.
- Enhanced signals improved speaker identification accuracy.

---

5. Observations and Analysis

5.1 Impact of Fine-Tuning

- Fine-tuning WavLM with LoRA and ArcFace loss improved speaker verification metrics significantly.
- Fine-tuning SepFormer enhanced separation and speech quality metrics.

5.2 Effectiveness of the Novel Pipeline

- Combining SepFormer and WavLM allowed for simultaneous speaker separation, enhancement, and identification.
- Joint fine-tuning ensured alignment between the two components, leading to better overall performance.

5.3 Challenges

- Mixing overlapping utterances with low SNRs made separation more challenging.
- Ensuring alignment between the speaker separation and identification components required careful tuning.

---

6. Conclusion

This project successfully demonstrated the following:

1. Pre-trained WavLM provides a strong baseline for speaker verification but benefits significantly from fine-tuning.

2. Fine-tuning with LoRA and ArcFace loss improves discriminative speaker embeddings.
3. SepFormer effectively separates overlapping speakers and enhances speech quality.
4. A novel pipeline combining SepFormer and WavLM achieves state-of-the-art performance in multi-speaker scenarios.

Future work could explore:

- Extending the pipeline to handle more than two speakers in a mixture.
- Testing the pipeline on real-world noisy recordings.

---

7. References

1. WavLM : https://arxiv.org/abs/2110.13900
2. SepFormer : https://github.com/speechbrain/speechbrain
3. VoxCeleb Dataset : https://www.robots.ox.ac.uk/~vgg/data/voxceleb/
4. LoRA : https://arxiv.org/abs/2106.09685
5. ArcFace Loss : https://arxiv.org/abs/1801.07698
