Unimodal Speech Emotion Recognition: Feature Extraction and Classifier Models

Zakaria Baannou, Yassine Hamzaoui, Khawla Rouchdi
International University of Rabat

***Abstract*—Speech Emotion Recognition (SER) plays a crucial role in fields like human-computer interaction, customer service, and mental health monitoring, enabling machines to detect and respond to human emotions in speech. This paper provides a comprehensive review of current SER approaches, focusing on feature extraction methods and machine learning classifiers. Specifically, we compare Deep Neural Networks (DNN) using Mel frequency spectral coefficients (MFCC) with Whisper, a transformer-based model that utilizes Mel spectrograms, and a vision transformer we finetune for our task. We analyze their performance on our dataset. The findings suggest that while all models are effective, Whisper outperforms DNNs in challenging scenarios, closely followed by ViT B/16, which provides a much better cost/performance ratio. The paper also highlights key metrics and proposes future directions for SER research.**

Speech Emotion Recognition (SER) is an emerging technology that enables machines to detect emotional states from spoken language. This capability is essential in applications such as human-computer interaction (HCI), customer service, and mental health monitoring. The challenge of SER lies in accurately identifying emotional cues in speech, as various factors—such as individual differences, linguistic variations, and environmental noise—can complicate recognition.

The audio data for SER can be categorized into two types: **spontaneous**, **acted**, and **elicited** speech [1]. Acted speech is recorded in controlled settings where speakers deliberately convey specific emotions, resulting in more consistent emotional cues. In contrast, natural speech reflects more authentic emotions, but poses greater challenges to recognition due to subtler variations in tone, pitch, and pacing.

This work evaluates three SER models, starting with a basic **DNN using MFCC Features**. The low resource approach where DNNs learn from MFCCs key acoustic features of speech that capture temporal and spectral characteristics. Next, OpenAI's **Whisper** (Whisper-small), a state-of-the-art transformer model pretrained for various speech tasks which uses MEL spectrograms to process audio. Whisper is particularly suited for noisy environments, where it outperforms traditional DNNs [2]. Finally, Google's **ViT B/16** [3], an encoder-only vision transformer that divides images into 16x16 patches and uses attention to extract features between these non-overlapping patches, globally throughout the image.

The code we worked with is available in a comprehensive Python Notebook at our Github repository [4].

; date of current version January 20, 2025

BACKGROUND

Speech datasets differ in their nature, additionally, there are also multi-modal datasets, and multilingual datasets [5]. In this work, we're using acted speech only as training data. In particular, it's the balanced TESS dataset [6] with 7 emotion labels (Angry, Disgust, Fear, Happy, Pleasant surprise, Sad, Neutral). Other works have used multi-modality to achieve greater performance, as shown in the works of Ngoc-Huynh Ho et al. (2020) [7], and Cai et al. (2021) [8]. However, we're focusing on feature extraction from monolingual speech.

EXPERIMENTS

For our experiments, we're testing the aforementioned architectures with the corresponding input. The loss function used is categorical cross-entropy.

Preprocessing went as follows; Each audio clip was either trimmed or padded to a consistent length of 3 seconds. Shorter clips were padded, while longer clips were truncated. Signal framing, or speech segmentation, is a key preprocessing step in speech emotion recognition (SER). It divides continuous speech into fixed-length sections, typically 20 to 30 milliseconds, to account for the non-stationary nature of speech. Emotions can change over time, but speech remains relatively constant over short periods.

Setup

For all of the following experiments, we've used Kaggle's free P100 GPU compute. We've also split the data as 70% for training, 15% for testing and 15% for validation.

Deep Neural Network: For the deep NN, 13 MFCC features along with their deltas were extracted and concatenated into a feature vector of size 39, for a fixed sequence length of 128, making the input a 39x128 long vector.

The NN architecture has 4 hidden layers, with sizes 512, 256, 128, 64. Throughout the hidden layers, we're using ReLU as activation function, a dropout of 30%, and a batch normalization after every layer.

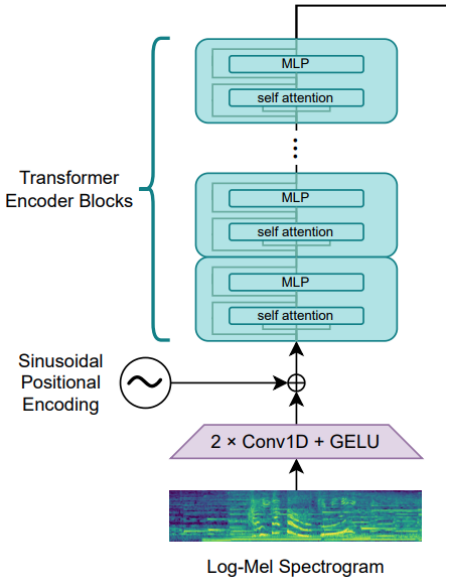

FIGURE 1. Whisper's Encoder Architecture

Whisper: We take Whisper-small's encoder and add a dense layer of size 256 and the final output layer. We freeze Whisper and only train the classification head.

To prepare our data, we resample files to a sample rate of 16,000 Hz. Each clip is padded or truncated to a fixed duration of 30 seconds to match the model's input requirements. In addition, To capture local features and maintain the connection between frames, we cut the audio into segments that are overlapped. To segment the audio, we use a window size of 25ms, and a hop size of 10ms, leaving us with an overlap of 15ms. Then, our input is a Mel spectrogram with 80 bins, as the sequence length based on audio duration (30 s = 600 frames).

Vision Transformer B/16: As the ViT expects a 3-channel input, we triple the depth of the spectrogram such as the 3 channels contain the same values.

We only train the top 4 layers of the encoder, and freeze the rest.

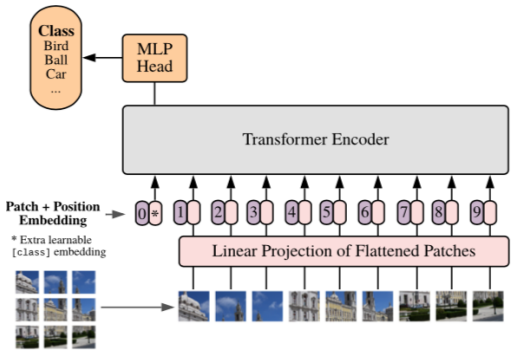

FIGURE 2. ViT's Patch + Position Embedding

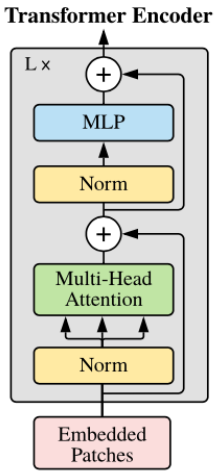

FIGURE 3. ViT's Encoder Architecture

The model is able to extract the information it needs through the attention mechanism only.

Results:

To recapitulate; Although two of our models used the MEL Spectrogram as input, we have varying input shapes as we can see in the table below.

Model	Input Shape
DNN	**(39, 128)**
Whisper-small	**(80, 600)**
ViT B/16	**(3, 224, 224)**

TABLE 1. Input shapes for the trained/fine-tuned models.

Training on 20 epochs took varying durations for each model. Accuracy and loss evolution was different even though we used Adam for all of them using the same learning rate. The figures below visualize the evolution of the mentioned metrics for the DNN, Whisper, and ViT models respectively.

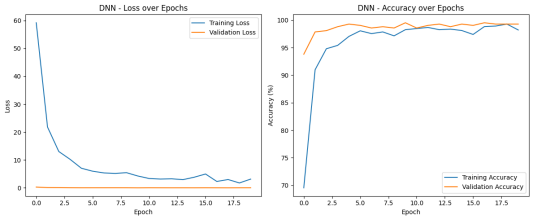

FIGURE 4. DNN Loss and Accuracy Evolution

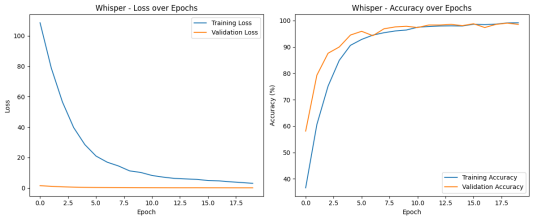

FIGURE 5. Whisper's Loss and Accuracy Evolution

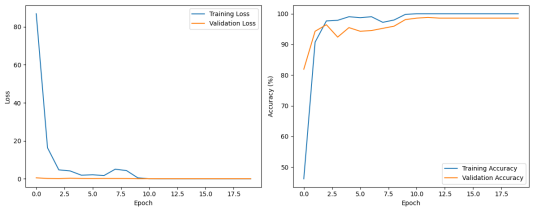

FIGURE 6. ViT's Loss and Accuracy Evolution

Testing accuracy scores were different. The table below lists both metrics.

Model	Test Accuracy (%)	Training Time (s)
DNN	97.63	538.55
Whisper-small	99.76	3400.37
ViT B/16	98.34	963.11

TABLE 2. Comparison of models: test accuracy and training time.

Due to accuracy scores being often misleading, we have the confusion matrices for the 3 models on our test data.

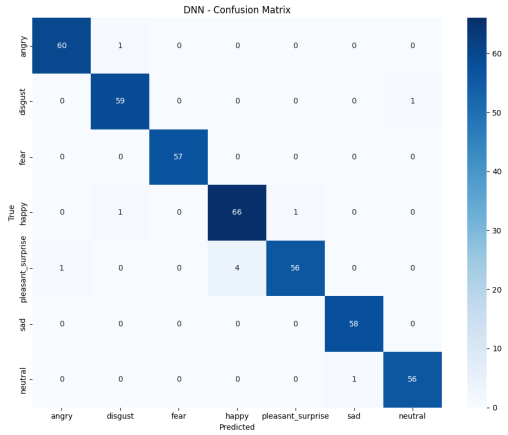

FIGURE 7. DNN's test data confusion matrix

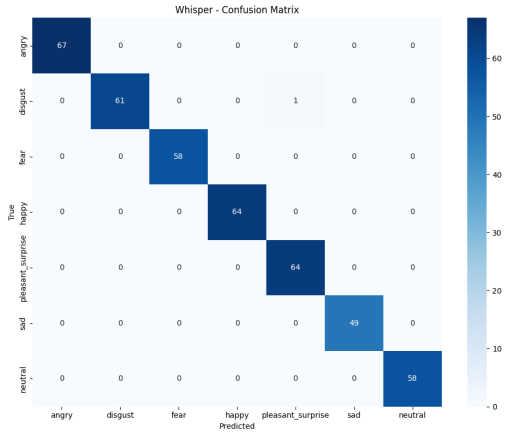

FIGURE 8. Whisper's test data confusion matrix

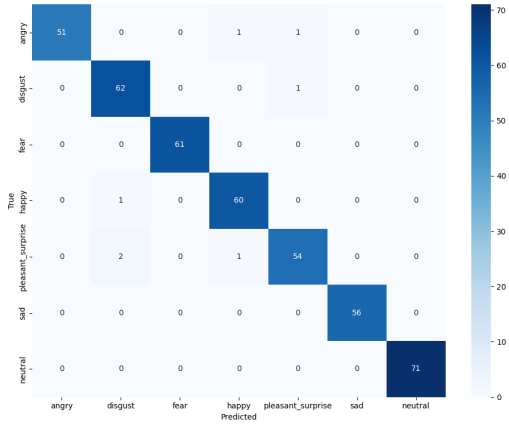

FIGURE 9. ViT's test data confusion matrix

With comparable performance, one must wonder about the cost. Below we have average inference times in seconds for our models.

Model	Inference Time (s)
DNN	0.0179
Whisper	0.9100
ViT B/16	0.0242

TABLE 3. Inference times for our trained/fine-tuned models.

It is easy to notice that Whisper is the most resource demanding model by a large margin. While the DNN and ViT need much less time for inference.

CONCLUSION

While the traditional DNN model using MFCC features provides good results and doesn't cost much, we find it doesn't generalize well to noisy audios and more nuanced speech. The transformer-based Whisper model, and the vision transformer B/16 both generalize much better. However, Whisper costs a lot more for predictions and training.

Due to the vision model's good performance, it's worth looking into it more. Further work on the fine-tuning process, or the preprocessing of the data could yield better results.

The high performance all around could be due to the dataset being too simply and not noisy at all.

Future work could explore combining these models and adopting multi-modal approaches to further enhance SER performance in diverse settings.

ACKNOWLEDGMENTS

We thank Professor Hakim Hafidi for his invaluable supervision, Professor Mehdi Zakroum for organizing this learning opportunity, and Professor Youness Moukafih along with the rest of the academic teaching staff for their dedication to quality teaching, which has greatly enhanced our learning experience and allowed us to be more independent.

We're grateful to the University of Toronto for the accessible quality dataset. We're also gratefulto Kaggle for the free compute they provide to everyone.

REFERENCES

1. T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47 795–47 814, 2021.
2. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech

recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356
3. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929
4. Z. Baannou, Y. Hamzaoui, and K. Rouchdi. (2025) Ser code repository. [Online]. Available: https://github.com/TheZakaria/SER_RnD
5. Z. Ma, M. Chen, H. Zhang, Z. Zheng, W. Chen, X. Li, J. Ye, X. Chen, and T. Hain, "Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark," 2024. [Online]. Available: https://arxiv.org/abs/2406.07162
6. M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020. [Online]. Available: https://doi.org/10.5683/SP2/E8H2MF
7. N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," *IEEE Access*, vol. 8, pp. 61 672–61 686, 2020.
8. X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning," in *Interspeech 2021*, 2021, pp. 4508–4512.

Zakaria Baannou 5th Year Big Data & AI Engineering Student at the International University of Rabat in Morocco. Loves to learn and curious about all things Machine Learning and Biology. Contact him at zakaria.baannou@uir.ac.ma .

Yassine Hamzaoui 5th Year Big Data & AI Engineering Student at the International University of Rabat in Morocco. Contact him at yassine.hamzaoui@uir.ac.ma .

Khawla Rouchdi 5th Year Big Data & AI Engineering Student at the International University of Rabat in Morocco. Contact her at khawla.rouchdi@uir.ac.ma .