Emotion Recognition from Speech Signals by Mel-Spectrogram and a CNN-RNN

Roneel V. Sharan*, Cecilia Mascolo†, Björn W. Schuller‡
*School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, United Kingdom
†Department of Computer Science and Technology, University of Cambridge, Cambridge CB3 0FD, United Kingdom
‡Department of Computing, Imperial College London, London WC1E 6BT, United Kingdom
Email: roneel.sharan@essex.ac.uk, cm542@cam.ac.uk, bjoern.schuller@imperial.ac.uk

Abstract—**Speech emotion recognition (SER) in health applications can offer several benefits by providing insights into the emotional well-being of individuals. In this work, we propose a method for SER using time-frequency representation of the speech signals and neural networks. In particular, we divide the speech signals into overlapping segments and transform each segment into a Mel-spectrogram. The Mel-spectrogram forms the input to YAMNet, a pretrained convolutional neural network for audio classification, which learns spectral characteristics within each Mel-spectrogram. In addition, we utilize a long short-term memory network, a type of recurrent neural network, to learn the temporal dependencies between the sequence of Mel-spectrograms in each speech signal. The proposed method is evaluated on angry, happy, and sad emotion types and the neutral expression on two SER datasets, achieving an average accuracy of 0.711 and 0.780. These results are a relative improvement over baseline methods and demonstrate the potential of our method in detecting emotion states using speech signals.**

Index Terms—**Convolutional neural network, Mel-spectrogram, recurrent neural network, speech emotion recognition**

I. Introduction

Emotion recognition technology involves the identification of human emotions through analysis of physiological signals. While various physiological signals can be utilized for this purpose, emotion recognition using speech signals presents advantages over other modalities in its natural and non-intrusive nature [1], capturing the rich emotional nuances conveyed through intonation, pitch, and rhythm. In addition, speech is a universally accessible and ubiquitous form of communication, making it highly applicable across diverse cultural contexts.

Speech emotion recognition (SER) presents several benefits in health applications. It can provide valuable insights into individuals' emotional well-being through the analysis of vocal cues. This technology enables the monitoring of mental health conditions by detecting changes in speech patterns indicative of mood shifts or emotional distress. In remote patient monitoring and telehealth, SER can enhance virtual consultations by offering an additional layer of emotional information. The early detection of cognitive decline, stress management through voice analysis, and personalized therapeutic interventions can contribute to more effective healthcare practices. As such, integrating SER into healthcare not only enhances the understanding of emotional states but also promotes a more holistic and patient-centered approach to medical care [2].

Current methods in SER are largely based on a combination of various feature engineering techniques to capture different speech characteristics, such as time and frequency descriptors, cepstral features, and deep learning features [2], [3]. These features are then classified using conventional machine learning methods, such as random forest and support vector machine [2], [3]. Today, most studies [4], [5] incorporate deep learning methods in SER. However, most publicly available SER datasets are small in size and deep learning methods can be susceptible to overfitting when trained on small datasets without effective preventative measures.

In this work, we propose a method for SER using a time-frequency image representation of the segmented speech audio signals and a combination of different neural network classifiers, including a pretrained network. For the time-frequency representation, we use the Mel-spectrogram over the conventional spectrogram representation used in an earlier study [4]. The Mel-spectrogram makes use of the Mel-scale [6] which resembles the way humans perceive sound. It provides finer resolution at lower frequencies which are important for speech intelligibility [7]. We use a pretrained convolutional neural network (CNN) for audio classification to learn the spectral characteristics from the Mel-spectrograms of the speech signals. In addition, a recurrent neural network (RNN) is utilized to learn the relationship between successive Mel-spectrograms from the same speech utterance. The combined network [8] is evaluated on two datasets of validated emotions.

II. Materials and Methods

The steps in SER proposed in this work are illustrated in Fig. 1 and explained in the following subsections.

A. Dataset

Our work makes use of two datasets: the Ryerson audio-visual database of emotional speech and song (RAVDESS) [9] and the database of elicited mood in speech (DEMoS) [10].

The RAVDESS database features individuals articulating two lexically-matched statements in a neutral North American accent. The data is accessible in multiple formats but we focus on the speech data. While the dataset encompasses seven emotions along with an additional neutral expression, in this work, consistent with [11], we use the improvised dataset of the following four classes: *angry*, *happy*, *sad*, and *neutral*.

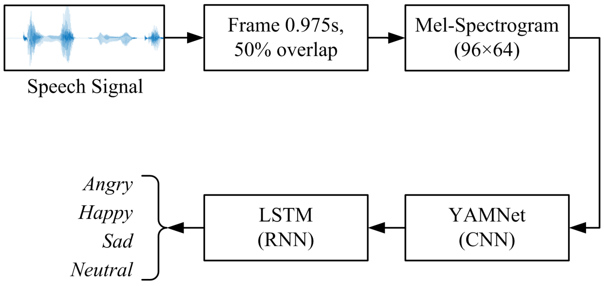

Fig. 1. Illustration of the steps in our speech emotion recognition.

TABLE I
OVERVIEW OF THE SPEECH DATASET USED IN THIS WORK

Description		**RAVDESS**	**DEMoS**
Number of subjects		24	65
Male:Female		12:12	42:23
Age (years)		26.0±3.8	23.7±4.3
Duration of speech signals (seconds)		3.70±0.34	2.61±1.13
Number of samples in each class	Angry	192	246
	Happy	192	167
	Sad	192	422
	Neutral	96	332

Each participant repeated each statement twice, delivering them at both normal and strong emotional intensities for all emotions, except neutral which is at neutral expression only. Consequently, there are eight recorded samples for each of the three emotion types and four recorded samples for the neutral expression from each participant.

While the RAVDESS dataset speech is in English, the DEMoS dataset is in Italian. The DEMoS dataset also consists of seven emotions and the neutral expression but our focus is once again on the four improvised classes, as on the RAVDESS dataset. Unlike the RAVDESS dataset, where the emotions are expressed by actors, a variety of mood induction procedures are used in the DEMoS dataset followed by a perception test, making it more authentic.

The speech utterances in both datasets are recorded at a rate of 48 kHz and 16-bit. Table I presents an overview of the two datasets. In addition, speech waveforms for angry, happy, sad, and neutral emotion types from the RAVDESS dataset are depicted in Fig. 2(a)–(d), respectively. These speech waveforms are from the same subject who spoke the same statement at the same emotional intensity in all four instances.

B. Mel-Spectrogram

The speech signals are converted into Mel-spectrogram time-frequency representation. This image-like representation forms input to the pretrained CNN. The pretrained CNN takes as input Mel-spectrogram of size 96×64, where 96 is the number of frames and 64 is the number of mel bands, from audio segments of length 0.975 s sampled at 16 kHz.

As such, we downsample the speech signals to 16 kHz and segment each speech signal into length of 0.975 s with 50% overlap between adjacent segments. Short-time Fourier transform (STFT) is computed for each segmented speech signal. This is done by sliding an analysis window (Hann window) of length of 25 ms with an overlap of 15 ms between adjacent windows. Discrete Fourier transform (DFT) of the windowed signal results in the spectrum $X_m(k)$, corresponding to the k^{th} frequency in the m^{th} frame.

The DFT values are then grouped into critical bands and subjected to weighting through triangular weighting functions [12]. Using the Mel-filters $r = 1, 2, ..., R$, where $R = 64$, the Mel-spectrum of the m^{th} frame is computed as [12]

$$MF_m(r) = \frac{1}{A_r} \sum_{k=L_r}^{U_r} |V_r(k) X_m(k)|^2, \qquad (1)$$

where the weight function for the r^{th} filter $V_r(k)$ has a DFT index range of L_r to U_r, and the normalizing filter for the r^{th} mel-filter is given as

$$A_r = \sum_{k=L_r}^{U_r} |V_r(k)|^2. \qquad (2)$$

The Mel-spectrogram representation of the speech waveforms of Fig. 2(a)–(d) are illustrated in 2(e)–(h), respectively.

C. Neural Networks

The Mel-spectrogram forms input to YAMNet [13], a pretrained audio classification network based on the depthwise-separable convolution architecture of MobileNet [14] and trained on the AudioSet corpus [15]. We modify the classification layer of the network to classify four classes (three emotions and neutral). In addition, for sequence-to-label classification, where a sequence of Mel-spectrograms from a speech signal are classified into a single emotion class, we add a flatten layer and two long short-term memory (LSTM) layers, with 150 hidden units in each, before the classification layer. The network parameters are optimized using adaptive moment estimation algorithm [16] with a learning rate of 0.0003. Additionally, we used a mini-batch size of 24 and trained for 50 epochs.

D. Experimental Setup and Evaluation Metrics

On the RAVDESS dataset, the performance of the proposed method is evaluated in leave-one-subject-out cross-validation whereby, in each fold, speech samples from one subject are used for testing. Of the remaining 23 subjects, 80% are used for training and 20% for validation. In addition, we perform 10-fold cross-validation on the DEMoS dataset with data from 10% of subjects for testing in each fold. Of the remaining 90% of subjects, 80% are used for training and 20% for validation. The performance is measured using *class accuracy* and *average accuracy*. Class accuracy is the proportion of correctly classified samples in a class and average accuracy is the average of the class accuracies. These metrics values range between 0 and 1 where 1 indicates an ideal value.

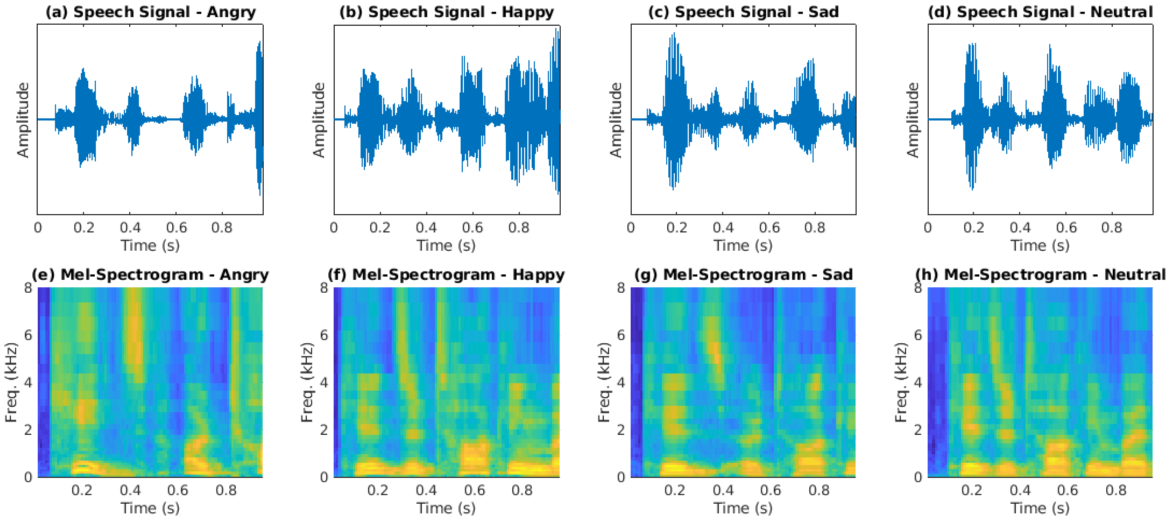

Fig. 2. Illustration of the speech waveform and Mel-spectrogram representation of angry, happy, sad, and neutral emotions from the RAVDESS dataset. The speech utterances are from the same subject who spoke the same statement and at the same emotional intensity across all four instances.

III. Experimental Results

We first present experimental results using baseline methods [3] and then using the proposed method.

A. Results Using Baseline Methods

The results in SER using baseline methods are presented in Table II. The baseline features include combination of cepstral features (Mel-frequency cepstral coefficients (MFCCs) and gammatone cepstral coefficients (GTCCs)) and feature embeddings from a pretrained deep learning network (deep learning (DL) features). These feature sets are classified using random forest (RF), support vector machine (SVM), and multilayer perceptron (MLP) classifiers. These feature extraction and classification methods are described in more detail in [3].

With the RF classifier, the average accuracy values on the MFCC+DL feature combination are slightly higher than the corresponding values on the GTCC+DL feature combination on both the datasets. However, the average accuracy values using the SVM and MLP classifiers on the GTCC+DL feature combination are slightly higher compared to the MFCC+DL feature combination. The highest average accuracy on the RAVDESS dataset is 0.564, using the GTCC+DL feature combination and SVM classifier, and 0.663 on the DEMoS dataset, once again using the GTCC+DL feature combination but with the MLP classifier.

B. Results Using Mel-Spectrogram and CNN-RNN

The SER classification results using mel-spectrogram and YAMNet-LSTM are also presented in Table II. The average accuracy values using this method are 0.711 and 0.780 on the RAVDESS and DEMoS datasets, respectively. This is a relative improvement of 26.1% and 17.6% over the highest baseline accuracy values on the RAVDESS and DEMoS datasets, respectively. The class accuracy values using the proposed method are also higher than the corresponding baseline values in all but one instance.

The highest average accuracy of 0.780 is achieved on the DEMoS dataset, where the accuracy value for angry, sad, and neutral classes are in the range of 0.754–0.880 using the proposed method. However, the accuracy value for happy emotion is 0.647, the lowest of all the classes, which is consistent with all the baseline methods. We further investigate the YAMNet-LSTM predictions on the DEMoS dataset using *t*-distributed stochastic neighbor embedding (*t*-SNE) [17]. The *t*-SNE visualization in Fig. 3 shows that the angry, sad, and neutral classes are reasonably well separated, however, the data points for happy cluster are overlapping with the other three classes. This explains the relatively lower classification accuracy for the neutral class. This is further supported by the confusion matrix in Table III where the happy class has high misclassifications with all three classes. One reason for this could be the smaller number of training data in the happy class.

IV. Discussion and Conclusion

In this work, we proposed a method for SER using Mel-spectrogram representation of the speech signals and a pretrained CNN (YAMNet) combined with a RNN (LSTM). The proposed method yielded an accuracy of 0.841, 0.647, 0.754, and 0.880 in classifying angry, happy, sad, and neutral emotion types. These values are higher than what could be achieved using various baseline methods. This highlights the usefulness of the proposed method in SER. In addition, the average accuracy on the DEMoS dataset is higher than what could be achieved on the RAVDESS dataset. The DEMoS dataset comes from more number of subjects, has more samples for all classes, except happy, and the emotions are more authentic. This potentially helps the network generalize better.

However, our work has some limitations, such as relatively small number of subjects and limited emotion types. In the future, we plan to evaluate the proposed method on a dataset

TABLE II
CLASSIFICATION ACCURACY IN SPEECH EMOTION RECOGNITION

Feature/Input	**Classifier**	**Accuracy on the RAVDESS dataset**					**Accuracy on the DEMoS dataset**				
		Angry	**Happy**	**Sad**	**Neutral**	**Average**	**Angry**	**Happy**	**Sad**	**Neutral**	**Average**
MFCC+DL	RF	0.766	0.609	0.557	0.156	0.522	0.703	0.138	0.841	0.708	0.598
MFCC+DL	SVM	0.729	0.573	0.547	0.292	0.535	0.780	0.305	0.813	0.699	0.649
MFCC+DL	MLP	0.714	0.521	0.510	0.406	0.538	0.711	0.329	0.746	0.708	0.624
GTCC+DL	RF	0.693	0.557	0.526	0.198	0.493	0.687	0.132	0.839	0.717	0.594
GTCC+DL	SVM	0.760	0.594	0.599	0.302	0.564	0.756	0.341	0.820	0.729	0.662
GTCC+DL	MLP	0.693	0.536	0.563	0.396	0.547	0.711	0.365	0.801	0.774	0.663
Mel-Spectrogram	YAMNet-LSTM	0.854	0.729	0.729	0.531	0.711	0.841	0.647	0.754	0.880	0.780

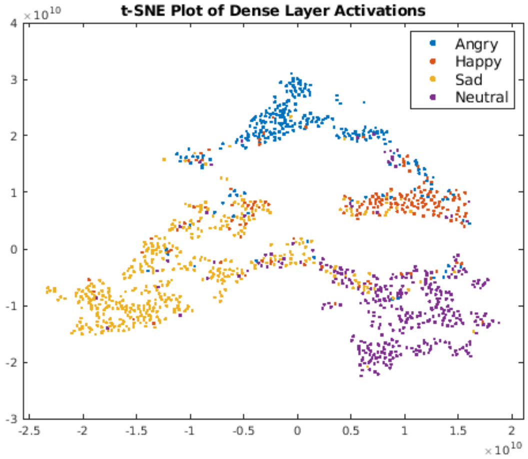

Fig. 3. *t*-SNE visualization of YAMNet-LSTM activations on DEMoS dataset.

TABLE III
CONFUSION MATRIX FOR YAMNET-LSTM ON DEMOS DATASET

		Predicted			
		Angry	Happy	Sad	Neutral
Actual	Angry	**207**	21	6	12
	Happy	21	**108**	17	21
	Sad	21	37	**318**	46
	Neutral	14	10	16	**292**

from more number of subjects which has the potential to improve the generalization of the network. We also plan to use speech utterances from other emotion types and from different settings for this purpose. In addition, in this work we evaluated only one neural network architecture. Despite yielding promising results, it is important to emphasize the increasing interest in deep learning research, where a myriad of innovative architectures are consistently being introduced. As such, in the future we plan to explore other neural network architectures for SER, including transformer models.

Subject to further independent and external validation, such SER technology can be integrated into smartphone applications and provide real-time monitoring and analysis of individuals' emotional states through their speech patterns. This can facilitate early detection of mental health conditions, enable personalized interventions for stress management and therapy, assist in remote patient monitoring for chronic illnesses, enhance medication adherence by detecting emotional barriers, and aid in early diagnosis of neurological disorders. As such, by leveraging smartphones as accessible and pervasive tools, SER has the potential to improve healthcare delivery by offering timely interventions, improving patient outcomes, and enhancing overall well-being.

REFERENCES

[1] Y. Alemu *et al.*, "Detecting clinically relevant emotional distress and functional impairment in children and adolescents: Protocol for an automated speech analysis algorithm development study," *JMIR Research Protocols*, vol. 12, 2023, Art. no. e46970.

[2] H. Aouani and Y. B. Ayed, "Speech emotion recognition with deep learning," *Procedia Computer Science*, vol. 176, pp. 251–260, 2020.

[3] R. V. Sharan, "Speech emotion recognition using gammatone cepstral coefficients and deep learning features," in *Proceedings of ICMLANT*, 2023, pp. 1–4.

[4] W. Lim *et al.*, "Speech emotion recognition using convolutional and recurrent neural networks," in *Proceedings of APSIPA*, 2016, pp. 1–4.

[5] Y. Zhao and X. Shu, "Speech emotion analysis using convolutional neural network (CNN) and gamma classifier-based error correcting output codes (ECOC)," *Scientific Reports*, vol. 13, no. 1, 2023.

[6] S. S. Stevens *et al.*, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 2005.

[7] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd ed. Piscataway, NJ: IEEE Press, 2000.

[8] R. V. Sharan *et al.*, "Detecting cough recordings in crowdsourced data using CNN-RNN," in *Proceedings of BHI*, 2022, pp. 1–4.

[9] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, 2018.

[10] E. Parada-Cabaleiro *et al.*, "DEMoS: An Italian emotional speech corpus," *Language Resources and Evaluation*, vol. 54, 2020.

[11] Z. Zhao *et al.*, "Hierarchical network with decoupled knowledge distillation for speech emotion recognition," in *Proceedings of ICASSP*, 2023, pp. 1–5.

[12] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*, 2nd ed. New Jersey: Prentice Hall, 2011.

[13] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proceedings of ICASSP*, 2017, pp. 131–135.

[14] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv:1704.04861*, 2017.

[15] J. F. Gemmeke *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proceedings of ICASSP*, 2017, pp. 776–780.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2017.

[17] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.