A Comparison Between Convolutional and Transformer Architectures for Speech Emotion Recognition

Shreyah Iyer*†, Cornelius Glackin†, Nigel Cannings†, Vito Veneziano*, Yi Sun*
*School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield, UK
†Intelligent Voice Ltd, London, UK
Email: y.2.sun@herts.ac.uk, neil.glackin@intelligentvoice.com

Abstract—**Creating speech emotion recognition models comparable to the capability of how humans recognise emotions is a long-standing challenge in the field of speech technology with many potential commercial applications. As transformer-based architectures have recently become the state-of-the-art for many natural language processing related applications, this paper investigates their suitability for acoustic emotion recognition and compares them to the well-known AlexNet convolutional approach. This comparison is made using several publicly available speech emotion corpora. Experimental results demonstrate the efficacy of the different architectural approaches for particular emotions. The results show that the transformer-based models outperform their convolutional counterparts yielding F1-scores in the range [70.33%, 75.76%]. This paper further provides insights via dimensionality reduction analysis of output layer activations in both architectures and reveals significantly improved clustering in transformer-based models whilst highlighting the nuances with regard to the separability of different emotion classes.**

Index Terms—**speech emotion recognition, transformers, wav2vec2, convolutional neural networks, alexnet, transfer learning, mel spectrograms**

I. Introduction

In recent years, increasing attention has been paid to the problem of predicting emotions from speech. Interest in this issue has been driven by everyday human social interactions, where it is essential to be able to understand and appropriately respond to other people's emotions correctly. Research on the relationship between emotions and stress [1] indicates that the key to managing stress lies with identifying the underlying negative emotions. Recognising emotions from speech is also of interest to areas such as criminology, banking, and insurance, where detection of emotions can aid in appropriate corrective action in cases of crime and fraud.

While recognizing emotions from speech is an easy task for humans, computers still have a long way to go before emotion recognition becomes a form of artificial intelligence. The biggest impediment in using speech is that there is no single discrete speech feature that directly reflects the speaker's emotions [2]. An added challenge in the engineering side, as evidenced in past research, is the problem of having limited repositories of commercially available training data resulting in low prediction accuracies.

In this paper, a transformer-based architecture, namely Wav2Vec [3], which is a model pre-trained on approximately $53,000$ hours of unlabeled data, is compared with the well known AlexNet Convolutional Neural Network (CNN) on three publicly available speech emotion datasets. To the best of the authors' knowledge, the performance of Wav2Vec on these datasets in the existing literature was measured using either the accuracy rate or the recall, making the comparison of different models difficult. This study has produced performance measurements using recall, precision, and F1-score on those datasets. Furthermore, cross-validation has been applied to ensure a robust result. These results can be used as the benchmark for models trained with these datasets. In addition, Principal Component Analysis (PCA) is applied on the output layer activations of the deep learning architectures used in this study to gain insights into how the model architectures classify emotions.

II. Related Research

Human emotions can be detected from various channels, such as speech [4], body language [5], facial expressions [6], and text [7]. The most obvious channel for emotion recognition is through speech. Just like studying body language and text to process user sentiment, speech signals can encompass a wealth of information related to various emotional characteristics [8]. They can be used to infer different facets of human behaviour irrespective of language, ethnicity, and other distinguishing factors.

Researchers have attempted a plethora of techniques with varying degrees of success using machine learning and deep learning to solve the problem of emotion recognition from speech signals. The subsections below highlight how previous research has guided the design decisions and research pathway for the work presented in this paper.

A. Emotion Classes

One of the common attempts has been to group emotions into discrete labels such as happy, sad, disgust, fear, surprise, anger, neutral, and so on. This makes the task amenable to classification algorithms that provide accurate results with little to no ambiguity between emotion classes. An alternate theory is to consider a continuous emotional space to describe the emotions with respect to valence (positive emotions and negative emotions) and their arousal (high intensity and low intensity) [9]. Although primitive emotions such as happy or sad tend to fit well in this emotional continuum, it becomes hard to distinguish emotions such as anger and fear. It is equally possible that some other emotions (such as surprise) may lie outside this continuous spectrum, which is usually subjective and open to interpretation. Having a continuous spectrum to classify emotions also means that the range of discrete emotions contained within a single spectrum is limited, and often involves a lot of ambiguity [10]. Based on these factors, the research proposed in this paper works with emotions in the discrete space.

B. Features

Research on Speech Emotion Recognition (SER) has seen the use of a range of handcrafted features over the years. The features extracted from a speech signal can be classified as two different types, namely the prosodic (for example, fundamental frequency, pitch, intonations, rhythm) and the spectral (for example, linear predictive coding, log frequency power coefficients) [11] feature types.

Prosodic features are those features that can be perceived by humans and have been known to contain distinctive properties of emotional content in the context of SER [12]. However, it has also been observed that these features may not be able to distinguish angry and happy utterances accurately since they have similar trends in fundamental frequency and speaking rate [13].

Spectral features are obtained by converting the time domain speech signal into its frequency domain using Fourier transforms. Mel Frequency Cepstral Coefficients (MFCCs) [14] are one of spectral-based feature representations. MFCCs are obtained by computing the discrete cosine transform on the log scaled Mel-spectrogram, which essentially is a visual representation of an audio signal. Compared to prosodic features, spectral features can distinguish angry from happy. However, the magnitude and shift of the formants for the same emotions vary across different vowels, adding more complexity to the emotion recognition task [15].

C. Deep Learning Architectures

Research on speech emotion recognition using deep learning has shown promising results. MFCCs and features such as pitch or energy have been widely used with deep learning models in the context of speech emotion recognition [16]. Long short-term memory (LSTM) networks have also been proposed to deal with speech emotion recognition [17]. In recent years, intensive research and development activities have been carried out in the field of image processing. CNNs have always shown encouraging results on image recognition tasks. Some of the most powerful CNN-based architectures like DenseNet [18] and ResNext [19] have been applied to speech spectrogram images [20]. In [21], the authors compared the effects of Mel coefficients and spectrogram images using deep learning in speech emotion recognition, and the results showed that the spectrogram images outperformed MFCCs through the implementation of deep learning neural networks.

One major problem in SER in dealing with real-life scenarios is the poor generalisation that arises due to limited training datasets and the mismatch of the training sets and the test sets. Recent achievements in deep learning, especially in the field of natural language processing, have been employed to cope with the limitations in speech emotion recognition [22]. Transformer models, which use encoders and decoders, can learn emotional long-term temporal dependencies with the self-attention mechanism [23]. Research into this field has resulted in a library of pre-trained models that have shown to be useful for a variety of speech related tasks. One such model from this library is the Wav2Vec model [24]. Using the concept of transfer learning, these pre-trained models can be leveraged with small scale datasets to accomplish a range of tasks in speech processing.

This paper presents one such architecture using a more recent version of the Wav2Vec architecture, the Wav2Vec 2.0 model [3] to perform emotion recognition and compares the results with the ones obtained using handcrafted features as inputs to the AlexNet-based CNN model [25].

III. A Description of the Datasets

There are several datasets publicly available for SER related research. These can be classified as acted (RAVDESS [26], CREMA-D [27]), elicited (IEMOCAP [28]), and natural datasets CMU-MOSEI [29]. RAVDESS, CREMA-D and IEMOCAP, while significantly smaller than CMU-MOSEI, have well balanced emotion content available with well annotated data. The CMU-MOSEI is imbalanced in the context of emotions available and the methodology of data annotation where multiple simultaneous, and often conflicting emotion labels, obtained using crowd sourcing, adds noise and unwanted complexity to the task. Furthermore, collecting natural speech signals from the real world typically comes with barriers such as acquisition challenges (as demonstrated by CMU-MOSEI), data handling constraints, and regulatory restrictions. Hence, the research presented in this paper utilises RAVDESS (acted), IEMOCAP (elicited), and CREMA-D (acted) which contain a mixture of noisy as well as clean data whilst capturing diversity in speakers as well as emotions. Table I shows the number of emotions considered and the number of audio files used in this study.

TABLE I
Distribution of Emotions across Datasets

Emotion Class	**The Number of Utterance in Each Dataset**		
	IEMOCAP	CREMA-D	RAVDESS
Neutral	1708	1087	96
Happy	1636	1271	192
Sad	1084	1271	192
Anger	1103	1271	192
Fear	-	1271	192
Disgust	-	1271	192
Surprise	-	-	192
Calm	-	-	192

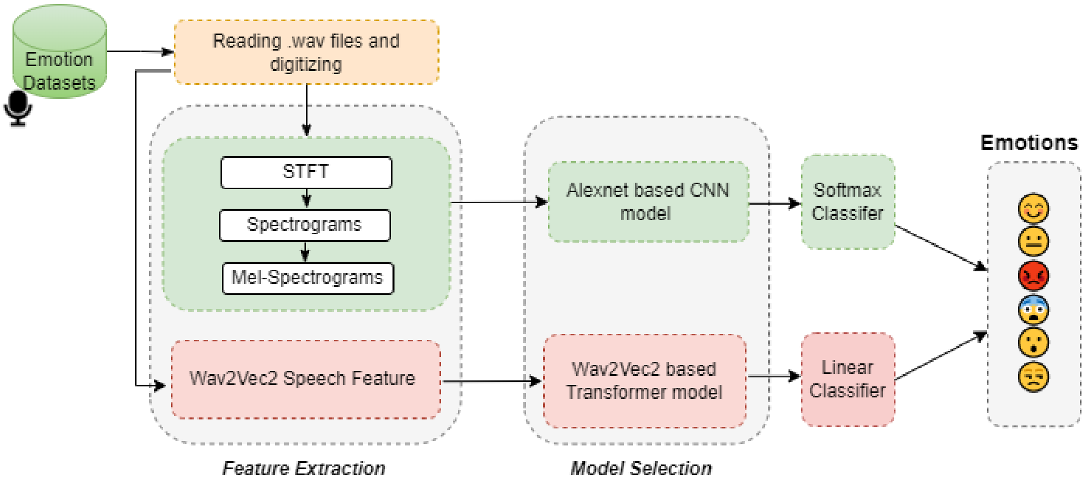

Fig. 1. Proposed methodology for speech emotion recognition. Here two architectures are proposed - an AlexNet-based CNN architecture ($M1$ in green) and a Wav2Vec2-based transformer architecture ($M2$ in red).

A. The RAVDESS dataset

RAVDESS is an audiovisual acted dataset consisting of $1,440$ utterances in a noise free environment and spoken in English with a North American accent. It consists of eight emotions *(neutral, happy, sad, fearful, angry, disgust, surprised, and calm)* enacted by 12 female and 12 male actors. The diversity in emotions as well as the gender of speakers is well balanced in this dataset. The length of the audio clips are $\approx$ 4 seconds each with predefined short sentences.

B. The CREMA-D dataset

CREMA-D is a crowd sourced audiovisual dataset consisting of $7,442$ utterances with 91 speakers (48 males and 43 females) reading 12 predefined sentences and portraying a set of six emotions *(anger, disgust, fear, happy, sad, and neutral)*. The audio files in this dataset contain added environmental noise. The length of each audio is $\approx$ 3 seconds in duration.

C. The IEMOCAP dataset

IEMOCAP contains data recorded across five sessions with ten speakers (five males and five females) spanning close to 12 hours of audiovisual data with transcripts. There are 10,093 utterances of scripted and improvised emulated scenarios representing a total of nine emotions. The average duration of an utterance is $\approx$4 seconds. However the data points present in this dataset are quite unbalanced. Hence, out of the nine available emotions, this research considers only four emotions *(anger, happy, sad, neutral)* to not only balance but to also enable effective comparison across datasets. The data labelled as *excited* was added to it to further augment the labels for *happy*.

D. Pre-processing data

Speech signals are time-variant and non-stationary signals. The audio files within the datasets used are stored in the wav format. They are first digitized using the Librosa library [30] at a sampling rate of 16 kHz. The digitized signals are trimmed to remove the leading and trailing silences. The resulting signals are zero padded to make the sizes consistent.

IV. Two Architectures Used in this Study

In this study, two model architectures are applied for SER. The first model architecture (denoted as $M1$) uses Mel-spectrograms as the feature vector with an AlexNet-based CNN [25]. The second model architecture (denoted as $M2$) is based on Wav2Vec2 [31]. Fig. 1 shows these two architectures: architecture $M1$ in green and architecture $M2$ in red. Both $M1$ and $M2$ make use of all of the above datasets.

A. Architecture $M1$ - AlexNet-based CNN

In this architecture, the pre-processed audio files (see section III-D) were transformed to the frequency domain using the Short Term Fourier Transform (STFT) with a time frame size of 256 with 50% overlap. After applying STFT, the magnitude is taken and frequencies are converted to the Mel scale, by passing the signal through several filter banks [32] to obtain Mel spectrograms. The feature extraction is performed using the Librosa library [30].

The Mel spectrograms are then passed as inputs to an AlexNet network. The AlexNet topology consists of five convolutional layers and three fully connected layers, all of which are ReLU activated [25]. The convolutional layers in the AlexNet extract essential features from the Mel spectrogram, and the fully connected layers that follow learn the data classification model parameters. The final layer, which is a softmax activated fully connected layer, is suitably modified to predict the required number of emotion classes.

B. Architecture $M2$ - Wav2Vec2-based Transformer

In this architecture, the pre-processed audio files (see section III-D) were passed onto the Wav2Vec2 model. The Wav2Vec2 model is essentially a pre-trained model for automatic speech recognition (ASR) [3]. This model learns contextualized speech representations by randomly masking feature vectors before passing them to a transformer network. This research makes use of features extracted from a pre-trained model called Wav2Vec2-Large-XLSR-53-English [33].

Architecturally, the Wav2Vec2 model has an encoder network and a context network. The encoder network transforms the digitized speech into a latent speech representation for a given number of time steps. These are then fed into the context network, which is basically a transformer, to build context representations that capture information from the entire sequence. To handle the context representations for any audio length, a mean merge strategy plan (pooling mode) is used. The output layer uses a linear classifier to map these context representations to emotions. For the experiments described in this paper, the pre-trained Wav2Vec2 model is fine-tuned with Connectionist Temporal Classification (CTC) [34].

C. Optimization and Hyper-parameter Tuning

In order to identify discrete emotions, categorical cross entropy is used as the loss function for architecture $M1$ and binary cross entropy for architecture $M2$. The choice of optimizer is the Adam optimizer that computes individual adaptive learning rates intelligently to speed up computation. Early stopping and model selection is used to prevent overfitting. The learning rate is set as 0.0001 and a batch size of 16 was used [35]. Architecture M1 was trained for 16 epochs with a constant learning rate schedule and the validation loss was monitored for a patience of five epochs. Architecture $M2$ used a learning rate scheduler with a warm up of 1000 steps to a peak of 0.0001 followed by an exponential decay, and was fine tuned for two epochs for the RAVDESS and CREMA-D datasets, and three epochs for the IEMOCAP dataset.

The resulting suite of models from these two architectures are extensively tested and their performance has been compared to determine the best performing model. These are detailed in the next section.

V. Experiments and Results

A. Experimental Setup

All of the experiments ensure that there is no intermixing of speakers in the training and testing datasets. It has also been ensured that all emotions are equally distributed across these subsets, both for the training set and the test set, while all the speakers (identified by their Speaker ID) are evenly apportioned across these subsets.

In RAVDESS, each training set consists of 16 speakers, while the validation and test set has four speakers each. The CREMA-D dataset is split to have a training set with 71 speakers, and a validation and test set with ten speakers each. The training set from IEMOCAP consists of four sessions, while the last session is split into the validation and test set. This ensured that all of the datasets are split with a ratio of $8 : 1 : 1$ for the training, validation, and testing set, respectively, while ensuring no mixing of speakers across sets.

For each dataset, models are trained for architectures $M1$ and $M2$ using the k-fold cross validation technique. Here, the dataset is first split into n subsets. One of these subsets is kept for testing and the remaining *(n-1)* subsets undergo *(n-1)*-fold cross validation. This leads to n different test sets and consequently n different models. To measure the performance of each model, precision, recall, and F1-score are computed from the confusion matrix. The mean and the standard deviation of each performance metric across the test sets are shown in Table II.

B. A Comparison of $M1$ and $M2$

The objective of the first experiment is to identify a suitable workflow and model architecture that derives optimal performance in the context of emotion recognition. Subsection V-B1) shows the overall results, and subsection V-B2) presents the detailed performance of each emotion.

1) Results of $M1$ vs $M2$: Table II shows the performance values across three datasets namely, RAVDESS, CREMA-D, and IEMOCAP. The first notable feature of this result is that architecture $M2$ significantly outperforms architecture $M1$ across all performance metrics on all three datasets. Model architecture $M2$ shows a mean F1-score of 77% on RAVDESS, 70.33% on CREMA-D, and 72.9% on IEMOCAP, significantly surpassing the performance metrics obtained from $M1$, which are 34.83%, 42.16% and 42.68%, respectively. This is likely because $M2$ is derived from a more robust pre-trained model indicating that pre-trained models may be the way forward to accomplish the task of speech emotion recognition.

Interestingly, $M2$ has produced relatively larger standard deviation values on RAVDESS and IEMOCAP, and smaller standard deviation values on CREMA-D when compared with $M1$. This may suggest that $M2$ generalises better when more speakers are involved in the dataset to do the fine tuning. In this case, CREMA-D involves 91 speakers, which is more than

TABLE II
Precision, Recall, and F1-score for Architectures $M1$ and $M2$ across different Datasets

Dataset	Model	Precision (%)		Recall (%)		F1 (%)	
		Mean	*Std. dev.*	*Mean*	*Std. dev.*	*Mean*	*Std. dev.*
RAVDESS	M1	33.31	2.87	37.16	2.79	34.83	2.80
	M2	**75.76**	8.66	**78.16**	7.58	**77**	8.44
CREMA-D	M1	41.66	4.19	42.16	3.8	42.16	4.48
	M2	**70.37**	3.74	**71.66**	3.14	**70.33**	3.72
IEMOCAP	M1	44.7	6.61	43.2	5.23	42.68	6.1
	M2	**73.16**	10.28	**75.6**	9.15	**72.9**	9.88

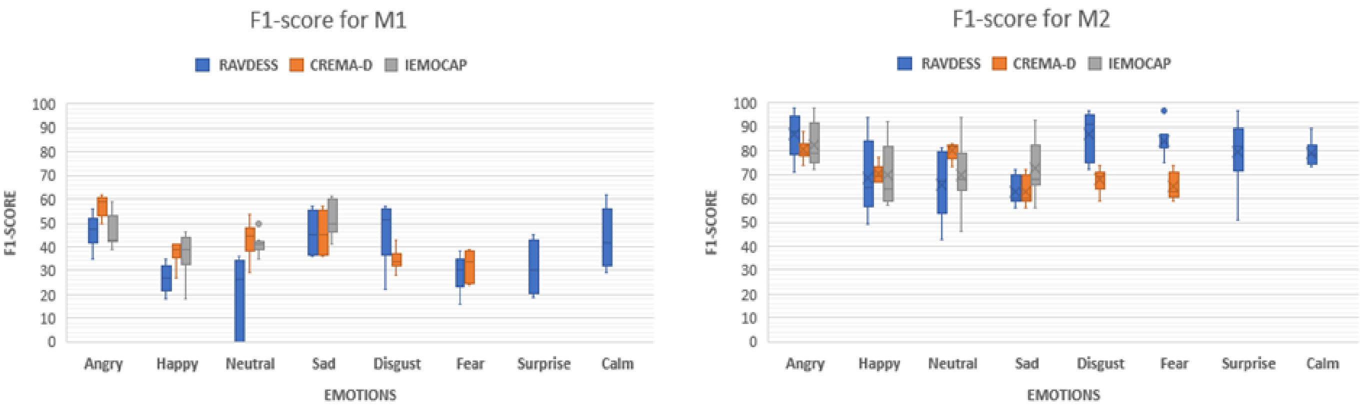

Fig. 2. Box and whisker plot of F1-scores for individual emotions

the other two datasets (24 and 10, respectively). However, this finding needs to be further validated on more datasets.

2) Results of Model Performance on Each Emotion: To observe the model performance for each emotion, the F1-scores for each of the emotions across each of the datasets are graphically depicted in the form of box plots in Fig. 2. It can be seen that,

- In general, $M2$ provides better performance over all emotions than $M1$.
- As seen in the right panel of Fig. 2, the mean values for *anger* across all three datasets are higher than other emotions, though the mean value for *disgust* and *fear* on RAVDESS, and the mean value for *neutral* on CREMA-D are equally high.
- The same architecture can have different performance on different datasets. For example, M2 provides better results for classes *anger* and *disgust* on RAVDESS, while for CREMA-D higher F1-scores are obtained for *anger* and *neutral* classes.
- Models trained using $M2$ on CREMA-D have a much smaller range of F1-scores than those trained on the other two datasets for almost all emotions, except for classes *sad* and *fear*. On the other hand, the trained M2 models tend to produce a bigger range of F1-score values on RAVDESS, especially for classes *happy* and *neutral*.

C. Principal Component Analysis of Output Layer Activations

The aim of the second experiment is to gain some insights into how the model architectures $M1$ and $M2$ classify emotions. PCA is used to visualise the output layer activations of $M1$ and $M2$, respectively. In $M1$, this layer is the softmax activated fully connected layer while in $M2$ it is the linear classifier layer.

Fig. 3 shows the PCA plots corresponding to the actual labels for models trained on the IEMOCAP dataset corresponding to model architectures $M1$ (the left panel) and $M2$ (the right panel). The total variance captured by the first two principal components is 85.89% and 86.89% for $M1$ and $M2$ respectively. The equivalent PCA plot for the predicted labels corresponding to architecture $M2$ is shown in Fig. 4. Emotional utterances that are highly correlated tend to be clustered together. These plots show the results for only one of the test sets from the IEMOCAP database to avoid cluttering and for ease of viewing, but the results have been verified to hold true for the remaining test sets as well. The confusion matrix for the $M2$ model discussed here is shown in Fig. 5.

Looking at the left panel in Fig. 3 for model architecture $M1$, it can be seen that there is significant overlap among all emotions without any clear clusters, especially for *happy* and *neutral*. This explains why the performance of the model is so poor for *happy* and *neutral* classes (see the left panel of Fig. 2). The informal listening experiments conducted by [19] further corroborate this finding. However the picture is very different and far more promising for $M2$ as seen in Fig. 3 and Fig. 4. For architecture $M2$, it is seen that the predicted emotions all have distinct clusters with only slightly overlapping decision boundaries. This suggests that transformer-based models can provide more efficient feature discrimination for speech emotion recognition, and it explains why this model outperforms the one derived from $M1$.

On looking at the right panel in Fig. 3 for actual labels corresponding to $M2$, it is seen that while most of *anger* and *sad* are well separated with clear clusters, there is quite a bit of overlap for *happy*, *neutral*, and *sad*, indicating that these emotions are more difficult to classify. Fig. 5 shows that 31 and 40 out of 161 clips labeled as *happy* have been misclassified as *neutral* and *sad*, respectively, while 73 out of 158 clips labeled as *neutral* have been misclassified as *sad*.

Furthermore, the misclassified speech clips have been investigated. For example, two clips shown in coordinates of (3.2, -0.5) and (5, -1.6) respectively in the right panel of Fig. 3 have the actual label of *sad*. However, they are misclassified as *anger*, shown in the same coordinates respectively in Fig. 4. The authors have all unanimously agreed that the emotion in the clips could have been interpreted as *anger* after listening to these audio clips. It matches what the model has predicted. Therefore, the misclassifications could be caused by mislabeling due to how different listeners perceive them. In addition, some of these data points also have a mixture of

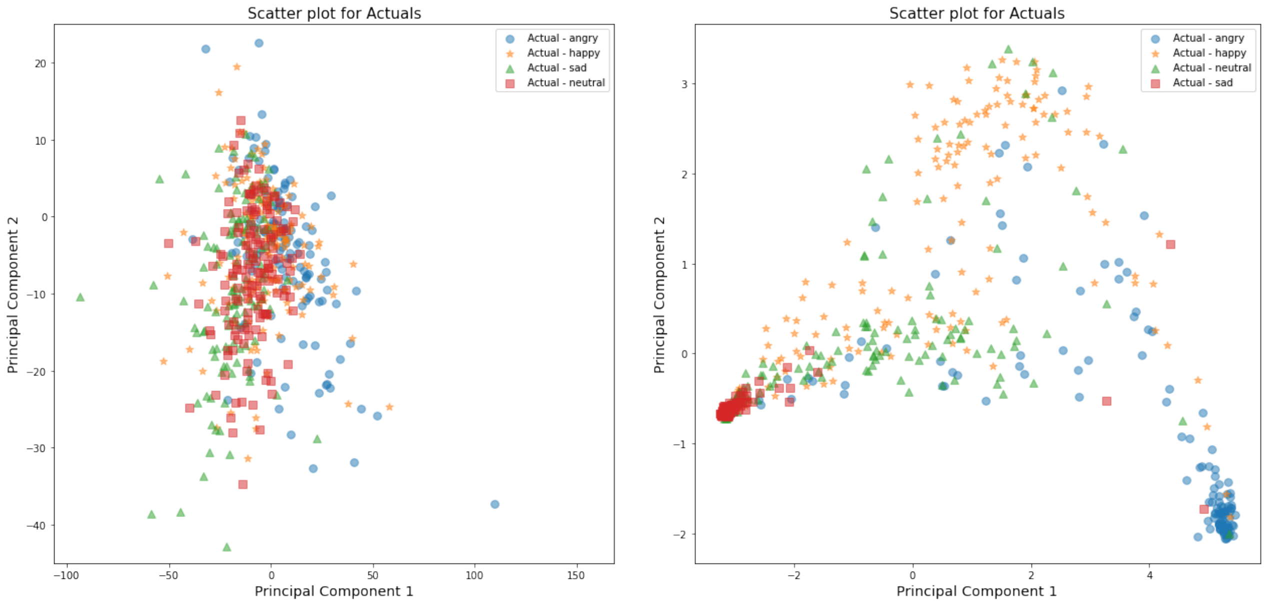

Fig. 3. PCA corresponding to actual labels derived from a model trained on the IEMOCAP database for architectures M1 (left) and M2 (right)

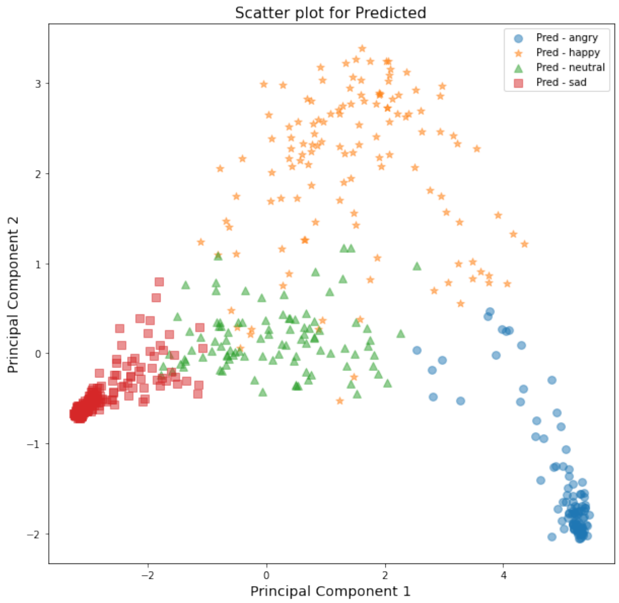

Fig. 4. PCA corresponding to predicted labels derived from a model trained on the IEMOCAP database for architecture $M2$.

emotions, such as *anger* and *sad*, but have only a single label. This also gives us insight into why *anger* can be an easily identifiable emotion in both model architectures compared to other emotions.

The PCA plots for the models built from RAVDESS and CREMA-D for $M2$ show similar clustering behaviour. In summary, it is seen that the models derived from $M2$ are able to draw piecewise decision boundaries and are a promising way to look at emotion recognition in speech for future research.

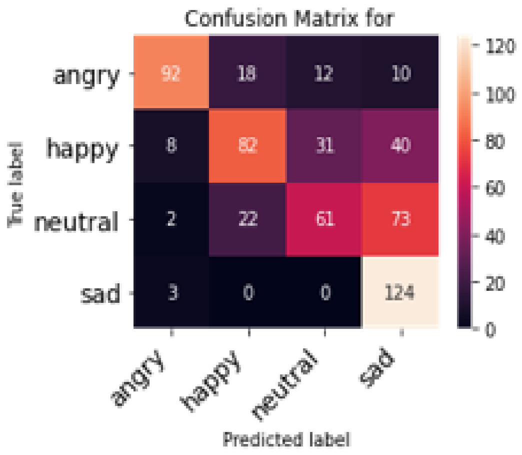

Fig. 5. Confusion matrix for model trained on the IEMOCAP database with architecture $M2$.

D. Discussion

Accurately predicting emotions from speech audio clips is extremely difficult. Table III shows a comparison of performance between the results obtained from the best performing proposed model using architecture $M2$ for each of the datasets and recent state-of-the-art literature. One can see that the performance reported through the research presented in this paper surpasses or is at par with state-of-the-art.

However, there are some caveats for the records reported in the comparison table. In [36] and [37], the experimental protocols and how the test was set up are not detailed well, making meaningful comparisons difficult. In [38], the authors used a modified Wav2Vec2 pre-trained model but only reported the recall rather than F1-score. Further, they combined *neutral* and *calm* to a single emotion for the RAVDESS model which may slightly skew the comparison results.

TABLE III
COMPARISON WITH STATE-OF-THE-ART LITERATURE

			Performance Metrics		
Dataset	**References**	**Model**	**Accuracy**	**F1**	**Recall**
IEMOCAP	Shen et al., 2020 [39]	DialogXL	-	62.4%	-
	Majumder et al., 2019 [40]	DialogueRNN	-	60.6%	-
	Kim et al.,2021 [41]	EmoBERTa	-	68.57%	-
	Padi et al., 2021 [42]	MWA-SER	-	66%	-
	Muppidi et al., [36]	QCNN	70.46%	-	-
	Pepino et al., [38]	Wav2Vec2-Pretrained	-	-	67.2%
	Proposed Methodology	Wav2Vec2-XLSR-53-English	**73.29**	**72.9**%	**75.6**%
RAVDESS	Jimenez et al., [43]	CNN-14	76.58%	-	-
	Jimenez et al., [43]	Sequentian bi-LSTM	57.08%	-	-
	Muppidi et al., [36]	QCNN	**77.87**%	-	-
	Pepino et al., [38]	Wav2Vec2-Pretrained	-	-	**84.3**%
	Proposed Methodology	Wav2Vec2-XLSR-53-English	77.05%	**75.76**%	78.16%
CREMA-D	Shukla et al., [36]	Audio-encoder	-	59.2%	-
	Ristea et al., [44]	DeepCNN Audio+Video	69.2%	-	-
	Ahmed et al., [37]	1D-CNN-LSTM-GRU	**74**%	-	-
	Proposed Methodology	Wav2Vec2-XLSR-53-English	70.77%	**70.33**%	**71.66**%

VI. CONCLUSION

This paper has compared the performance of an AlexNet-based CNN model with handcrafted features as inputs to a transformed-based model, Wav2Vec, on three speech emotion datasets.

It was found that the pre-trained Wav2Vec2 model yielded high levels of accuracy in classifying different emotions across these different datasets. Further investigation using PCA to the output layer activations in both architectures revealed that the transformer-based model is more adept at defining piecewise decision boundaries across emotion classes, thereby producing a more accurate classification of different emotions.

The research presented in this paper raises new questions and paves the path for further research to aid the evolution of SER systems. A potential extension to this research is through a multi-modal approach wherein using textual content as an added feature could boost the performance of emotion recognition. Further speech systems trained on datasets with a particular language and dialect are unlikely to perform well globally. Therefore modelling such as cultural variations, different languages, speaker speeds, pitch scales are potential avenues to explore. It also makes the role of good quality and robust data essential. Training and testing with data from varied real-world scenarios are mandatory to make this an attractive application for deployment to a larger audience.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement 823907 (MENHIR project: https://menhir-project.eu) and by an Innovate UK grant via the Knowledge Transfer Partnership programme, KTP No.11039.

REFERENCES

[1] K. Tomba, J. Dumoulin, E. Mugellini, O. Abou Khaled, and S. Hawila, "Stress detection through speech analysis.," 2018.

[2] A. Al-Talabani, H. Sellahewa, and S. Jassim, "Emotion recognition from speech: Tools and challenges," vol. 9497, 04 2015.

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[4] F. Jelinek, *Statistical methods for speech recognition*. MIT press, 1997.

[5] T. Keshari and S. Palaniswamy, "Emotion recognition using feature-level fusion of facial expressions and body gestures," in *2019 International Conference on Communication and Electronics Systems (ICCES)*, pp. 1184–1189, 2019.

[6] V. V. Salunke and C. G. Patil, "A new approach for automatic face emotion recognition and classification based on deep networks," in *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pp. 1–5, 2017.

[7] S. Park, B. Bae, and Y. Cheong, "Emotion recognition from text stories using an emotion embedding model," in *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 579–583, 2020.

[8] R. Cowie and E. Douglas-Cowie, "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," pp. 1989 – 1992 vol.3, 11 1996.

[9] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, pp. 5–32, 04 2003.

[10] S. R. Kadiri, P. Gangamohan, V. Mittal, and B. Yegnanarayana, "Naturalistic audio-visual emotion database," in *Proceedings of the 11th International Conference on Natural Language Processing*, (Goa, India), pp. 206–213, NLP Association of India, Dec. 2014.

[11] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan, "Speech emotion recognition using both spectral and prosodic features," in *2009 International Conference on Information Engineering and Computer Science*, pp. 1–4, 2009.

[12] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Analysis of prosodic variation in speech for clinical depression," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439)*, vol. 3, pp. 2925–2928 Vol.3, 2003.

[13] G. Paidi, S. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech—a review," *Soc. Believ. Behav. Syst.*, vol. 1, pp. 205–238, 2016.

[14] N. Kamaruddin, A. W. Abdul Rahman, and N. S. Abdullah, "Speech emotion identification analysis based on different spectral feature extraction methods," in *The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M)*, pp. 1–5, 2014.

[15] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, 2021.

[16] K. Han and I. Yu, D.and Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," September 2014.

[17] X. Chen, R. Huang, X. Li, L. Xiao, M. Zhou, and L. Zhang, "A novel user emotional interaction design model using long and short-

term memory networks and deep learning," *Frontiers in Psychology*, 04 2021.

[18] S. Cheng, D. Zhang, and D. Yin, "A densenet-gru technology for chinese speech emotion recognition," in *International Conference on Frontiers of Electronics, Information and Computation Technologies*, ICFEICT 2021, (New York, NY, USA), Association for Computing Machinery, 2021.

[19] S. Padi, S. O. Sadjadi, D. Manocha, and R. D. Sriram, "Improved speech emotion recognition using transfer learning and spectrogram augmentation," 2021.

[20] S. Kadyrov, C. Turan, A. Amirzhanov, and C. Ozdemir, "Speaker recognition from spectrogram images," pp. 1–4, 04 2021.

[21] S. Demircan and H. K. Örnek, "Comparison of the effects of mfccs and spectrogram images via deep learning in emotion classification," *Traitement du Signal*, pp. 51–57, 2020.

[22] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, 2021.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[24] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *CoRR*, vol. abs/1904.05862, 2019.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84 – 90, 2012.

[26] 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, pp. 1–35, May 2018.

[27] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[28] C. Busso, M. Bulut, C.-C. Lee, E. A. Kazemzadeh, E. M. Provost, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[29] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," *CoRR*, vol. abs/1808.05561, 2018.

[39] W. Shen, J. Chen, X. Quan, and Z. Xie, "Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition," 2020.

[30] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, pp. 18–25, 2015.

[31] A. Conneau, A. Baevski, R. Collobert, A. rahman Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *ArXiv*, vol. abs/2006.13979, 2021.

[32] N. Hajarolasvadi and H. Demirel, "3d cnn-based speech emotion recognition using k-means clustering and spectrograms," *Entropy*, vol. 21, p. 479, May 2019.

[33] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *CoRR*, vol. abs/2006.13979, 2020.

[34] A. GRAVES, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. Int. Conf. on Machine Learning, 2006*, pp. 369–376, 2006.

[35] R. Poojary and A. Pai, "Comparative study of model optimization techniques in fine-tuned cnn models," in *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pp. 1–4, 2019.

[36] A. Muppidi and M. Radfar, "Speech emotion recognition using quaternion convolutional neural networks," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6309–6313, 2021.

[37] M. R. Ahmed, S. Islam, P. D, A. K. M. M. Islam, P. D, S. Shatabda, and P. D, "An ensemble 1d-cnn-lstm-gru model with data augmentation for speech emotion recognition," 2021.

[38] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," 2021.

[40] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," 2019.

[41] T. Kim and P. Vossen, "Emoberta: Speaker-aware emotion recognition in conversation with roberta," 2021.

[42] S. Padi, D. Manocha, and R. D. Sriram, "Multi-window data augmentation approach for speech emotion recognition," 2021.

[43] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, "Multimodal emotion recognition on ravdess dataset using transfer learning," *Sensors*, vol. 21, no. 22, 2021.

[44] N.-C. Ristea, L.-C. Dutu, and A. Radoi, "Emotion recognition system from speech and visual information based on convolutional neural networks," 02 2020.