

# Hybrid Time Distributed CNN-transformer for Speech Emotion Recognition

Anwer Slimi<sup>1,2</sup><sup>a</sup>, Henri Nicolas<sup>1</sup><sup>b</sup> and Mounir Zrigui<sup>2</sup><sup>c</sup>

<sup>1</sup>Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France

<sup>2</sup>University of Monastir, RLANTIS Laboratory LR 18ES15, Monastir, Tunisia


**Keywords:** Deep Networks, Speech Emotion Recognition, Time Distributed Layers, Transformers.


**Abstract:** Due to the success of transformers in recent years, a growing number of researchers are using them in a variety of disciplines. Due to the attention mechanism, this revolutionary architecture was able to overcome some of the limitations associated with classic deep learning models. Nonetheless, despite their profitable structures, transformers have drawbacks. We introduce a novel hybrid architecture for Speech Emotion Recognition (SER) systems in this article that combines the benefits of transformers and other deep learning models.


## 1 INTRODUCTION

Communication, which may be defined as the center of our daily existence, is one of the things that binds human beings together. It is not only a means of transmitting words; rather, it is a method of delivering a large amount of information. This process may be initiated in a variety of ways, including speech, which is often considered as the quickest and most natural mode of communication (Kanth and Saraswathi, 2014). Emotions provide significance to interpersonal relationships, aid in our understanding of one another, and play a fundamental part in all social phenomena, which is why they must be thoroughly explored. Human-computer interaction has made significant strides in recent years to satisfy users' demands and responsibilities. From this vantage point, it would be ideal if computers could automatically recognize human emotions, therefore facilitating communication on both sides. The primary architecture of an emotion recognition system (Fig. 1) is to analyze an audio file to extract a collection of features, and then to detect emotions using a classifier. Thus, developing a resilient system with a high degree of accuracy requires two critical components: the most appropriate approach for feature extraction and the most performant

classification algorithm. Numerous feature extraction algorithms have been used for emotion classification in recent years, including the Gray-Level Co-Occurrence Matrix (Lima and Sajin, 2007), the Geneva Minimalistic Acoustic Parameter Set (Latif et al., 2018), the Mel-scaled spectrogram, the Chromagram, the Spectral contrast feature, and the Tonnetz representation (Issa et al., 2020). Thus, developing the optimal feature extraction method continues to be a significant challenge, since Feature Extraction tries to minimize the number of features in a dataset by generating new ones from existing ones (and then discarding the original features). These newly condensed features should then be capable of summarizing the majority of the information included in the original set of features. While the feature set is critical and determines whether or not we extract all relevant information for the emotion recognition task, the classification algorithm itself has a significant impact on the final result, as the primary reason for the model's low or high accuracy and poor or high prediction rate is the model's choice and configuration. With the advent of Deep Learning models, computer scientists are increasingly turning to Neural Networks (such as CNNs, LSTMs, and so on) to solve problems across a range of fields, owing to the fact that deep learning models outperform conventional machine learning approaches

<sup>a</sup> <https://orcid.org/0000-0003-0558-2321>

<sup>b</sup> <https://orcid.org/0000-0003-2179-4965>

<sup>c</sup> <https://orcid.org/0000-0002-4199-8925>

(Najafabadi et al., 2015). Transformers have taken over the world of NLP in recent years. The Transformer architecture makes extensive use of Attention to significantly improve the performance of deep learning translation models. It was introduced in (Vaswani et al., 2017) and was rapidly adopted as the standard design for the majority of text data applications. Transformers were created to address the difficulty of sequence transduction, also known as neural machine translation. That is, any task that converts an input sequence to an output sequence is included. This includes voice recognition and text-to-speech conversions, among other things. Along with transformers, Vision Transformers have been proposed for image classification.

Despite their usefulness, transformers have considerable drawbacks, that we will discuss in section 3.2.

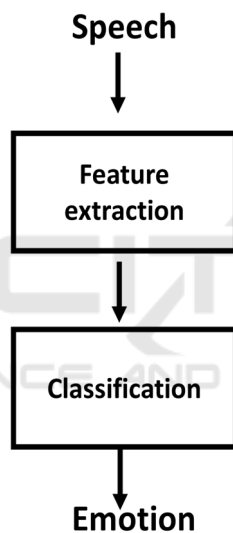


Figure 1: The general scheme of SER system.

The term "hybrid models" refers to the combination of different models to process a single input. They have recently shown excellent performance in the SER domain (Slimi et al., 2022) as well as in many other domains (Mahmoud and Zrigui, 2021; Ons et al., 2021; Bellagha and Mounir, 2020). In this paper, we propose a novel hybrid Speech Emotion Recognition system based on Time-Distributed CNNs and Transformers (Fig. 5.)

## 2 STATE OF THE ART

The approach of Lima M. and Sajin S. (2007) consist of generating the spectrogram of the speech then

transform it to a GLCM (Gray-Level Co-Occurrence Matrix). The GLCM is a widely used technique in the domain of texture analysis. It employed basically to perform feature extraction. And in simple words, the GLCM is 2-dimensional vector that contains some derived information from the spectrogram. In their paper, this technique is used to extract texture features (standard deviation, energy, mean, and entropy). The Gray-Level Co-Occurrence Matrix is fed to an SVM (support vector machine) to classify the emotions.

Avots et al. (2019) have created their speech emotion identification system utilizing a support vector machine. They employed a sliding window to extract features, thus each speech is represented with multiple vectors where each vector has 21 features of the global signal feature and 13 Mel Frequency Cepstral Coefficients features of each frame. And to reach a single result for the complete speech, the results were amalgamated by majority vote.

Mustaqem et al. (2020) utilize a k-means approach to cluster similar frames after splitting an utterance into numerous frames. One key frame from each cluster that is closest to the centroid will be chosen and utilized to build a spectrogram. They applied the transfer learning approach for classification by deploying the pre-trained Resnet101. To recognize emotions, the output of the Convolution Neural Network (CNN) will be normalized and sent to a bidirectional Long Short-Term Memory (LSTM).

The work of Issa et al. (2020) is based on gathering together 5 different feature sets: MFCCs, Mel-scaled spectrogram, Chromagram, Spectral contrast feature and Tonnetz representation. The total number of features is 193 that were stacked together and fed to 1-Dimensional CNN.

Slimi et al. (2020) have utilized log-mel spectrograms as an input for a shallow neural network to establish that neural networks can function with limited datasets. Once the spectrograms were formed, they have shrunk them to be able to input them to the first layer of the neural network. Although their model was basic, it delivered excellent results but under one condition: a person's records should be separated into the train and the test sets, otherwise the model would not be able to generalize.

The purpose of Xia et al. (2020) is to learn salient parts for speech emotion detection. The model consists of obtaining hand-crafted low-level descriptors and separating them into equal overlapping pieces. Each segment will be submitted to a Deep Neural Network to forecast the emotion probabilities. Then an attentive temporal pooling

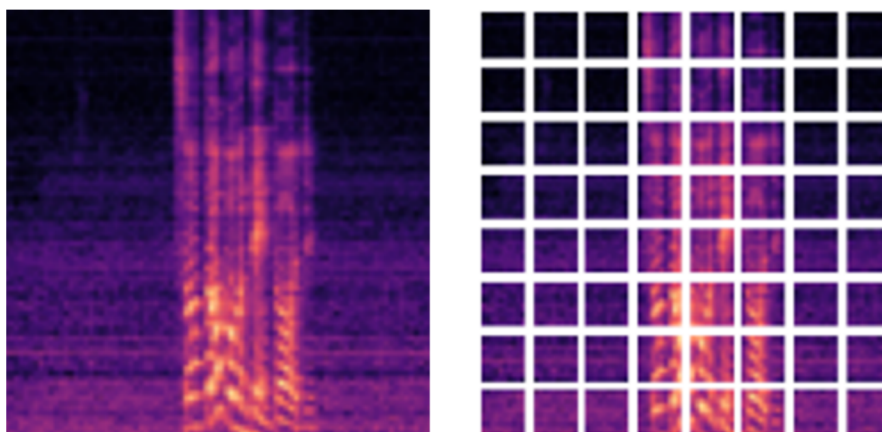


Figure 2: Log-Mel Spectrogram (left) with its corresponding patch (Right).

module made of a Gaussian Mixture Model and an auxiliary deep Neural Network, is used to learn the emotional saliency weights of each frame with self-attention.

In the work of Aouani and Ayed (2020), a vector of 42 features was derived from each signal. The employed features include Zero Crossing Rate, Harmonic to Noise Rate and Teager Energy Operator. Instead of employing the 42 features, they have opted to implement an Auto-Encoder to acquire a reduced data representation and to choose important features. The output of the auto encoder will be fed to a Support Vector Machine to identify emotions.

Because of the effectiveness of transformers, several researchers have utilized them to identify emotions. Only a few articles, such as (Tarantino et al., 2019), have used Transformers for speech emotion identification. Transformers, on the other hand, are used in multimodal systems where audio is blended with text, visual, or both.

Chen et al. (2021) employed Wav2vec speech embeddings in conjunction with Roberta text embeddings to identify emotions. They presented a key-sparse Transformer to concentrate on emotion-related information. In addition, a cascaded cross-attention block is included, which is specifically built for multimodal frameworks, to create deep interaction across multiple modalities. Similarly, Delbrouck et al. (2020) suggest two architectures that incorporate linguistic and acoustic inputs and are based on Transformers and modulation.

To identify emotion, Xie et al. (2021), suggest a transformer-based cross-modality fusion using the EmbraceNet architecture. They employed three models at the same time, each followed by a transformer block: FaceNet + RNN for image recognition, GPT for text, and WaveRNN for audio.

The approach of Hajarolasvadi and Demirel (2019) comprises dividing the speech to small frames with equal length where each frame is overlapping 50% with the previous chunk. For each frame, they extracted 88 features and its corresponding spectrogram in a way that if a speech is divided to n frames, then it will be represented with n spectrograms and 88xn matrix. For each speech, a k-means clustering algorithm is performed on the feature vector (with k=9) to select k most discriminant frames. The corresponding spectrograms of the k selected frames will be stacked together to build a 3D tensor. So finally, each speech is represented with 3D tensor which will be fed to a CNN.

Mupidi and Radfar (2021) have used the Mel-log Spectrogram as an input for their model. However, as a classifier they have not used a typical neural network, instead, they have used quaternion convolutional neural network (QCNN) as a classifier.

The work of Mustaqeem and Kwon (2020a) focuses mainly on the pre-processing phase where they used an adaptive threshold-based algorithm to remove silence, noises and irrelevant information. When it comes to their system, they generated the spectrogram from each signal and fed it to CNN to perform classification. Their preprocessing phase helped improve the accuracy by about 8% comparing to the original dataset

In (Mustaqeem and Kwon, 2020b) different blocks were used in their SER framework. They have used a ConvLSTM (combination of CNN and LSTM) for local feature learning block (LFLB), a Gated Recurrent Units (GRU) for global features learning

The model of Seo and Kim (2020) consists of training a model called VACNN (visual attention convolutional neural network) using a large dataset of log-mel spectrograms. The VACNN model is

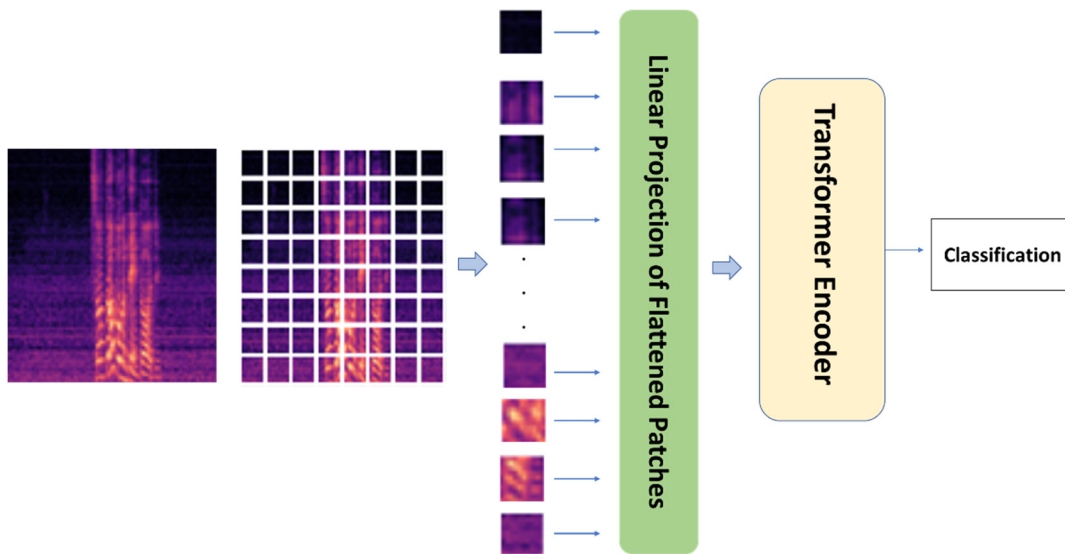


Figure 3: The architecture of a ViT (Vision Transformer).

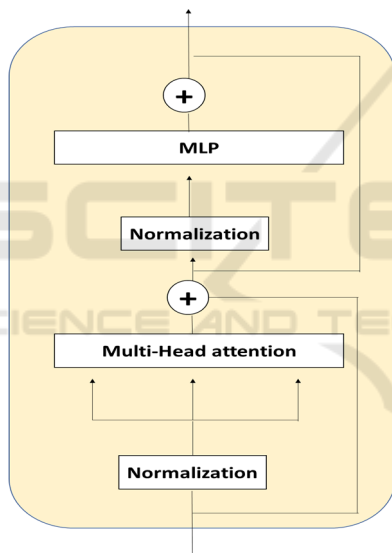


Figure 4: The architecture of a Transformer Encoder.

composed of convolution blocks along with spatial and visual attention module. Their goal is to obtain a model that can be used with smaller datasets. For a new small dataset, a bag of visual words (BOVW) is used to extract local features as an attention vector and a fine-tuned VACNN is used to extract features from log-mel spectrograms. An element-wise multiplication is applied on the outputs of the BOVW and the fine-tuned VACNN followed by an element-wise sum to sum up features. Finally, a fully-connected layer followed by a softmax layer is used to identify emotions.

Pepino et al (2021) have used pre-trained wav2vec, a framework for extracting representations

from raw audio data. The extracted features, the eGeMAPS descriptors and the spectrograms are used as an input for a shallow neural network. Best result was obtained using the wav2vec features

### 3 PROPOSED MODEL

The general scheme of an emotion recognition system consists of processing an audio file to extract a set of features, then use a classifier to identify emotions.

#### 3.1 Feature Extraction

Speeches are frequently employed in a variety of disciplines of research, including speech recognition, emotion identification, gender recognition, music genre categorization, and so on. Depending on the job, we may extract various characteristics from the input signal (Trigui et al., 2015) or display it in a variety of forms such as phonemes (Terbeh et al., 2017), transcripts (Labidi et al., 2017), or spectrograms.

A spectrogram is a depiction of the spectrum of signal frequencies as they change over time. The Short Time Fourier Transform (STFT) is used to create spectrograms. It entails taking tiny frames of length  $L$  of the original signal and performing a Discrete Fourier Transform (DFT) on each frame rather than the complete signal. The Mel-scale will be utilized instead of the hertz since the Log-Mel Spectrogram has been shown to be effective in identifying emotions (Yenigalla et al., 2018).

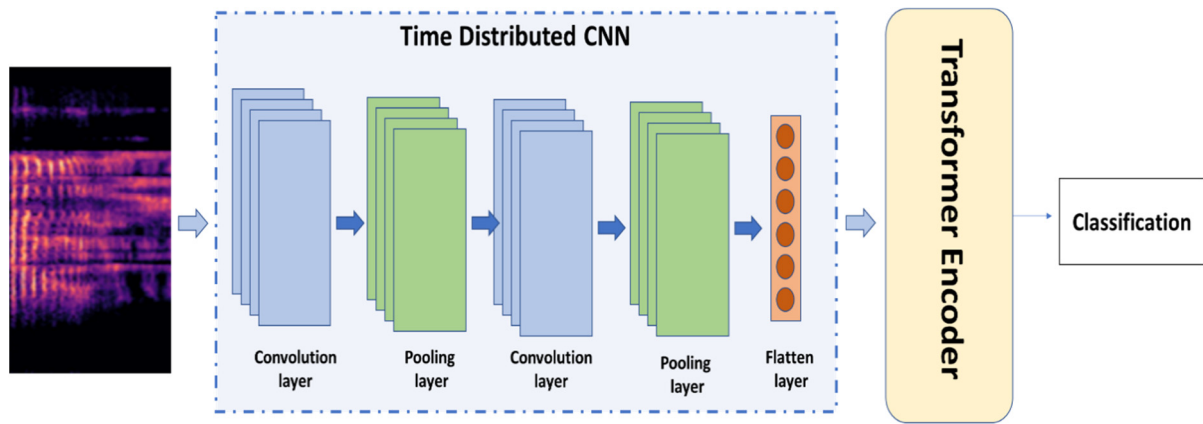


Figure 5: Proposed model.

When it comes to sound pitches, the Mel-scale is a psychoacoustic scale that may be used to distinguish between low and high pitches. Its unifying factor is the Mel. In terms of psychological sense of pitch, it relates to a subjective approximation of the psychological experience of pitch of sound (Stevens and Volkman, 1937). The following formula is used to convert from the  $f$  Hertz to the  $m$  Mel frequency range.

$$m = 2595 \log_{10} \left( 1 + \frac{f}{700} \right) \quad (1)$$

Despite the fact that the resolution of frequencies is related to both the window size and the sampling rate of the audio signal, Zhao et al (2019) and P. Yenigalla et al (2018) used the log Mel-spectrogram with the same parameters window size= 2048 and overlap= 512 on different sampling rates, whereas Zhao et al (2019) used the log Mel-spectrogram with the same parameters window size= 2048 and overlap= 512 on different sampling. Likewise, in our research, we resampled all of the data to 22050 Hz and then generated our spectrograms using the identical settings as those in (Zhao et al., 2019; Yenigalla, et al., 2018).

### 3.2 Emotion Recognition

Since the spectrograms are 2D images, a CNN can be considered the best choice as classifier since CNNs are usually used to perform a classification task on images. Most of the work described earlier in the state of the art, along with a lot of other researches, use several Convolution layers and Pooling layers stacked successively, followed by few Fully-Connected layers and an output layer with a softmax function to predict the class of the input. As a start, we have considered using several architectures, such as Inception, ResNet, VGG-19, etc, as a classification algorithm.

However, although transformers were originally designed to work with textual data, that hasn't prevented them from being used in a variety of Computer Vision tasks, including image processing, with results comparable to convolutional neural network (Dosovitskiy et al., 2021; Touvron et al., 2021). In (Dosovitskiy et al., 2021), authors proposed a Vision Transformer (ViT), in which every image is divided into patches (Fig. 2.), with each patch being passed to the transformer input layer and processed similarly to a single word embedding.

The usage of a Multi-head Attention mechanism module that is based on Self-attention is the secret to the transformers' success. Self-attention, also known as intra-attention, is an attention mechanism that links distinct points in a single sequence to calculate a representation of the same sequence. The Multi-head Attention technique is used numerous times in simultaneously. The outputs of the separate attention are then concatenated and linearly translated into the expected dimension.

However, despite numerous studies demonstrating that Vision Transformers outperform CNNs (Dosovitskiy et al., 2021), Transformers in general are still less powerful for local-invariant vision data (Liu et al., 2021) and vision transformers in particular are vulnerable against adversarial patches (Jindong et al., 2021). To avoid these concerns, some papers have proposed using CNN along with transformers. The use of CNN with Transformers is not novel. It's been utilized in a lot of researches. In (Karpov et al., 2020), authors used a Transformer followed by a CNN. Zhang et al. (2022) have used a CNN with Transformer in hybrid approach. Liu et al. (2021) did not use a CNN but rather some convolution filters before the Transformer.

Table 1: Accuracy of the Proposed model (RML dataset).

Work	Year	Result
Avots et al. (2019)	2019	69.30%
Xia et al. (2020)	2020	73.15%
Aouani. And Ayed (2020)	2020	74.07%
Issa et al. (2020)	2020	77.00%
<b>Ours</b>	<b>2022</b>	<b>83.88%</b>
<b>Ours (with SpecAugment)</b>	<b>2022</b>	<b>84.76%</b>

Table 2: Accuracy of the Proposed model (RML dataset).

Work	Year	Result
Hajarolasvadi and Demirel (2019)	2020	68.00%
Mustaqeem et al. (2020)	2020	71.61%
Muppidi and Radfar (2021)	2021	77.87%
Mustaqeem and Kwon (2020a)	2020	79.50%
Mustaqeem and Kwon (2020b)	2020	80.00%
<b>Ours</b>	<b>2022</b>	<b>82.72%</b>
Seo and Kim (2020)	2020	83.33%
Pepino et al. (2021)	2021	84.30%
<b>Ours (with SpecAugment)</b>	<b>2022</b>	<b>84.63%</b>

The usage of a Time Distributed CNN with a conventional transformer rather than a vision transformer, on the other hand, is novel in this research.

The transformer accepts one embedding at a time as input, whereas the Vision Transformer accepts one patch at a time. The spectrograms however are plots with special specificity where the vertical axis presents the frequency and the horizontal axis presents the time in seconds. Rather of splitting the spectrogram into patches and feeding them to a Transformer, we want to inject a chronologically ordered series of frames (time steps).

Prior to sending the frames to the transformer, we want to do per-frame convolution operations to extract key features for training the transformer. If we apply a convolution operation to each frame, each frame will have its own convolution flow, and each result will be treated as a separate input for the transformer.

However, if we train each convolution flow independently, we will encounter undesirable behaviours such as a lengthy and slow training process because we will need to train several convolution flows (one for each input frame), each convolution flow can have several different weights, resulting in different features detection that are unrelated, and some convolution flows will be unable to detect what other convolution flows can. We must ensure that the complete set of convolution flows can locate the same features. It is conceivable that certain convolutional flows discover something else, but this risk must be minimized.

A solution to this is to employ a Time Distributed layer, which enables us to apply a layer to each temporal slice of an input (i.e., perform the same operation on each timestep) and generate one output per input.

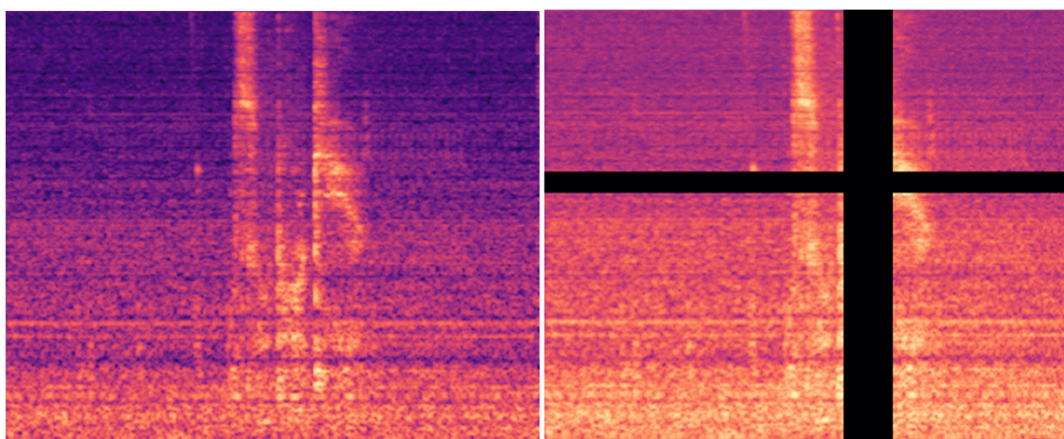


Figure 6: Example of SpecAugment (Right).

## 4 RESULTS

Our approach was tested using two datasets: the RML (Ryerson Multimedia Research Lab) and RAVDESS (Ryerson Audiovisual Database of Emotional Speech and Song), two of the most well-known open access datasets. The RML (Xie and Guan, 2013) dataset contains 720 audio recordings made by eight individuals. The dataset was collected in six distinct languages (English, Italian, Mandarin, Urdu, Punjabi, and Persian) and included six basic emotions: disgust, happiness, fear, anger, surprise, and sadness. We transformed each file to wav format since it is an audio-visual dataset. The RAVDESS (Livingston and Russo, 2018) is a dataset in English that was created by 24 actors (12 male and 12 female). This dataset's audio-only section has 1440 files. Eight emotions are detected in this dataset: happy, sad, disgusted, neutral, calm, angry, surprised, and fearful, although seven emotions are recognized at high intensity (angry, sad, happy, calm, disgust, fearful and surprised). Each dataset is partitioned into 80% train, 10% validation, and 10% test using stratified sampling to maintain the proportion of samples per class.

We have trained the different models simultaneously using the same data. Table 1 and Table 2 illustrate the comparison of the result produced by our proposed model (Time Distributed CNN + Transformer) on the RML and the RAVDESS datasets respectively, whereas Table 3 and Table 4 illustrate the comparison between different architectures with the RML and the RAVDESS datasets respectively.

Table 3: Accuracy of different architectures (RML dataset).

Approach	Result
ViT	65.43%
CNN+ViT	76.86%
CNN	79.30%
Time Distributed CNN + ViT	82.41%
Time Distributed CNN + Transformer	83.88%

Table 4: Accuracy of different architectures (RAVDESS dataset).

Approach	Result
ViT	62.63%
CNN+ViT	73.51%
CNN	73.59%
Time Distributed CNN + ViT	82.13%
Time Distributed CNN + Transformer	82.72%

Table 3 and Table 4 clearly show that the best result was obtained by using the Time-Distributed CNN along with the Transformer.

Table 5: Impact of the data on the system’s accuracy (RML dataset).

	CNN	ViT	Time Distributed CNN + Transformer
<b>Original dataset</b>	79.30%	65.43%	83.88%
<b>Classic augmentation techniques</b>	80.21%	65.88%	83.94%
<b>SpecAugment</b>	81.11%	57.31%	84.76%

Table 6: Impact of the data on the system’s accuracy (RAVDESS dataset).

	CNN	ViT	Time Distributed CNN + Transformer
<b>Original dataset</b>	73.59%	62.63%	82.72%
<b>Classic augmentation techniques</b>	74.20%	65.37%	83.55%
<b>SpecAugment</b>	74.41%	58.12%	84.63%

## 5 DISCUSSION AND ANALYSIS

As mentioned earlier, a lot of CNN architectures have been considered such as Inception, ResNet, VGG-19. We first trained each of the models from scratch but the results were not auspicious. This can be explained by the fact that such deep architectures require a huge amount of data (Slimi et al., 2020). The Transfer learning technique is one of the solutions that should be considered when there is no much data to work with. It has been used in previous SER systems (Mustaqeem et al., 2020; Roopa et al., 2018) but failed to attain good results, as in our work. So, finally we have decided to use less deep CNN and train it from scratch. After considering different architectures, best results were obtained using a model that is composed of three convolutional layers where each one of them is succeeded by a Max-pooling layer, followed by two Dense layers and a softmax activation function to perform classification. However, the accuracy of various algorithms varies depending on the quantity of data available (Table 5 and Table 6). This is because, in fact, deep learning models need a large amount of data to be adequately trained, otherwise they would not be able to attain high levels of accuracy.

CNN employs pixel arrays, but a Vision Transformer divides pictures into visual tokens, similar to how word embeddings are represented when using transformers to text. The vision

transformer splits a picture into fixed-size patches, embeds each one appropriately, and uses positional embedding as an input to the transformer encoder. Because there are less inductive biases, Vision transformer requires a bigger dataset to be pre-trained on. In that case, such models surpass CNNs in terms of computing efficiency and accuracy. Although Vision Transformers are so robust for image classification, the existing datasets for Speech Emotion Recognition are relatively small and that can explain the reason why using a hybrid architecture composed of a Time Distributed CNN with a simple transformer, outperforms both the CNN and the Vision Transformer. While deep networks cannot perform well in the absence of sufficient training data, it is possible to increase the effective size of existing data through the process of data augmentation, which has led to significant improvements in the performance of deep networks in many domains. Traditional augmentation techniques for speech recognition include deforming the audio waveform used for training in some way (e.g., by speeding it up or slowing it down) and adding background noise to the signal. In practice, this has the effect of making the dataset appear to be larger because multiple augmented versions of a single input are fed into the network over the period of training. It also has the effect of making the network more robust because it is forced to learn relevant features during training. Traditional ways of enhancing auditory input, on the other hand, incur significant computing costs and, in



certain cases, need the collection of new data. A novel augmentation approach, SpecAugment, has been developed by D. S. Park et al. (2019), which applies an augmentation policy directly to the audio spectrogram rather than augmenting the input audio waveform as is normally done. This strategy is straightforward, computationally inexpensive to implement, does not need the collection of extra data, and yielded superior outcomes when compared to typical data augmentation strategies. SpecAugment alters the spectrogram by warping it in the time direction, masking blocks of consecutive frequency channels, and masking blocks of utterances in the time direction, among other things. These augmentations have been designed to assist the network in remaining resilient in the presence of time-direction deformations, partial loss of frequency information, and partial loss of tiny segments of speech from the input. Both Table 5 and Table 6 clearly demonstrate the influence of the data on the accuracy of the models. The more the amount of data we have, the higher the accuracy we get. The SpecAugment approach, however, has failed with the ViT, despite its efficacy. This failure may be explained by the fact that ViTs are vulnerable against adversarial patches.

## 6 CONCLUSION

Although transformers have shown to be effective substitutes for CNNs, there is one significant restriction that makes their implementation rather difficult: the need for huge datasets. Indeed, CNNs can learn even with a tiny quantity of data, owing mostly to the existence of inductive biases. Thus, one may argue that transformers are more capable of learning but also need more data. To maximize the benefits of both systems, these two architectures, which are based on completely opposite concepts, were merged to create something capable of using both architectures' strengths. And, although the results were pleasant and surpassed current approaches, there is still room for improvement.

## REFERENCES

- Egils Avots, Tomasz Sapinski, Maie Bachmann, Dorota Kaminska. 2019. Audiovisual emotion recognition in wild. *Mach. Vis. Appl.* 30(5): 975-985.
- Xiaohan Xia, Dongmei Jiang, Hichem Sahli. 2020. Learning Salient Segments for Speech Emotion Recognition Using Attentive Temporal Pooling. *IEEE Access* 8: 151740-151752 (2020).
- Hadhami Aouani, Yassine Ben Ayed. 2020. Speech Emotion Recognition with deep learning. *KES* 2020: 251-260.
- Dias Issa, M. Fatih Demirci, Adnan Yazici. 2020. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control.* 59: 101894.
- Anwer Slimi, Mohamed Hamroun, Mounir Zrigui and, Henri Nicolas. 2020. Emotion Recognition from Speech using Spectrograms and Shallow Neural Networks. *The 18th International Conference on Advances in Mobile Computing & Multimedia (MoMM2020)*, November 30- December 2, 2020, Chiang Mai, Thailand. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3428690.3429153>
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention Is All You Need. *CoRR abs/1706.03762* (2017)
- Maryam M Najafabadi, Flavio Villanustre, Taghi MKhoshgoftaar, Naeem Seliya, Randall Wald and EdinMuharemagic: Deep learning applications and challenges in bigdata analytics. Najafabadi et al. *Journal of Big Data* (2015) 2:1
- Lorenzo Tarantino, Philip N. Garner, Alexandros Lazaridis: Self-Attention for Speech Emotion Recognition. *INTERSPEECH* 2019: 2578-2582
- Weidong Chen, Xiaofeng Xing, Xiangmin Xu, Jichen Yang, Jianxin Pang: Key-Sparse Transformer with Cascaded Cross-Attention Block for Multimodal Speech Emotion Recognition. *CoRR abs/2106.11532* (2021)
- Jean-Benoit Delbrouck, Noé Tits, Stéphane Dupont: Modulated Fusion using Transformer for Linguistic-Acoustic Emotion Recognition. *CoRR abs/2010.02057* (2020)
- Baijun Xie, Mariia Sidulova, Chung Hyuk Park: Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Crossmodality Fusion. *Sensors* 21(14): 4913 (2021)
- Aymen Triguí, Naim Terbeh, Mohsen Maraoui, Mounir Zrigui. Statistical Approach for Spontaneous Arabic Speech Understanding Based on Stochastic Speech Recognition Module. *Research in Computing Science* 117: 143-151
- Z. Xie and L. Guan 2013. Multimodal Information Fusion of Audiovisual Emotion Recognition Using Novel Information Theoretic Tools. *International Journal of Multimedia Data Engineering and Management*, vol. 4, no. 4, pp. 1-14, 2013
- Lima Mathew, Sajin Salim. 2007. Real World Speech Emotion Recognition from Speech Spectrogram Using Gray-Level Co-Occurrence Matrix. *International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization)*
- Siddique Latif, Rajib Rana, Shahzad Younis, Junaid Qadir, Julien Epps. 2018. Transfer Learning for Improving

- Speech Emotion Classification Accuracy. INTERSPEECH 2018: 257-261.
- Naim Terbeh, Mounir Zrigui. 2017. A Robust Algorithm for PathologicalSpeech Correction. PACLING 2017: 341-351.
- Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, Jithendra Vepa. 2018. Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. INTERSPEECH 2018: 3688-3692
- Mustaqeem, Muhammad Sajjad, Soonil Kwon. 2020. Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. IEEE Access 8: 79861-79875
- Nithya Roopa S., Prabhakaran M, Betty.P. 2018. Speech Emotion Recognition using Deep Learning. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-4S, November 2018.
- Mohamed Labidi, Mohsen Maraoui, Mounir Zrigui. 2017. Unsupervised Method for Im-proving Arabic Speech Recognition Systems. PACLIC 2017: 161-168
- Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Hervé Jégou.2021. Training data-efficient image transformers & distillation through atten-tion. ICML 2021: 10347-10357
- Anwer Slimi, Henri Nicolas and Mounir Zrigui. 2022. Detection of Emotion Categories' Change in Speeches. In Proceedings of the 14th International Conference on Agents and Artificial Intelligence - Volume 3, ISBN 978-989-758-547-0, ISSN 2184-433X, pages 597-604. DOI: 10.5220/0010868100003116 .
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.
- Jindong Gu, Volker Tresp, Yao Qin. 2021. Are Vision Transformers Robust to Patch Per-turbations? CoRR abs/2111.10659 (2021)
- Pavel Karpov, Guillaume Godin, Igor V. Tetko:. 2020. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. J. Cheminformatics 12(1): 17 (2020)
- Jiang Zhang, Chen Li, Ganwanming Liu, Min Min, Chong Wang, Jiyi Li, Yuting Wang, Hongmei Yan, Zhentao Zuo, Wei Huang, Huaifu Chen. 2022. A CNNtransformer hybrid approach for decoding visual neural activity into text, Computer Methods and Programs in Biomedicine, Volume 214, 2022, 106586, ISSN 0169- 2607, <https://doi.org/10.1016/j.cmpb.2021.106586>.
- Yun Liu, Guolei Sun, Yu Qiu, Le Zhang, Ajad Chhatkuli, Luc Van Gool. 2021. Transform-er in Convolutional Neural Networks. CoRR abs/2106.03180 (2021).
- Meddeb Ons., Maraoui M ohsen and Zrigui, Mounir. 2021. Personalized Smart Learning Recommendation System for Arabic Users in Smart Campus. International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 16(6), 1-21. <http://doi.org/10.4018/IJWLTT.202111101.0a9>
- Mustaqeem, Soonil Kwon. 2020a. A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. Sensors 20(1): 183
- Mustaqeem, Soonil Kwon. 2020b. CLSTM: Deep Feature-Based Speech Emotion Recognition Using the Hierarchical ConvLSTM Network. Mathematics 2020, 8, 2133. <https://doi.org/10.3390/math8122133>
- Noushin Hajarolasvadi and Hasan Demirel. 2019. 3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms. Entropy 2019, 21, 497.
- Bellagha Med Lazhar and Zrigui Mounir. 2020. Speaker Naming in TV programs Based on Speaker Role Recognition. 2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA), 1-8.
- Seo Minji, Kim Myungho. 2020. Fusing Visual Attention CNN and Bag of Visual Words for Cross-Corpus Speech Emotion Recognition. Sensors 20, no. 19: 5559. <https://doi.org/10.3390/s20195559>.
- Leonardo Pepino, Pablo Riera, Luciana Ferrer. 2021. Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings. CoRR abs/2104.03502
- Aneesh Muppidi, Martin Radfar. 2021. Speech Emotion Recognition Using Quaternion Convolutional Neural Networks. ICASSP 2021: 6309-6313
- Mr. N. Ratna Kanth and Dr. S. Saraswathi: A Survey on Speech Emotion Recognition. Advances in Computer Science and Information Technology (ACSIT) Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 1, Number 3; November, 2014 pp. 135-139.
- Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, Quoc V. Le: SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. CoRR abs/1904.08779 (2019)
- S.S. Stevens and J. Volkman: A scale for the Measurement of the Psychological Magnitude Pitch. J.A.S.A January 1937, Volume 8.
- Jianfeng Zhao, Xia Mao, Lijiang Chena: Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control 47 (2019) 312–323.
- Livingstone SR, Russo FA. 2018. 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 13(5): e0196391.
- Adnen Mahmoud and Mounir Zrigui. 2021. Hybrid Attention-based Approach for Arabic Paraphrase Detection, Applied Artificial Intelligence, 35:15, 1271-1286, DOI: 10.1080/08839514.2021.1975880