ArabEmoNet: A Lightweight Hybrid 2D CNN-BiLSTM Model with Attention for Robust Arabic Speech Emotion Recognition

Ali Abouzeid[1*], **Bilal Elbouardi**[1*], **Mohamed Maged**[1*], **Shady Shehata**[2]
[1]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2]University of Waterloo, Waterloo, ON, Canada
{ali.abouzeid, Bilal.ElBouardi, Mohamed.Elsetohy}@mbzuai.ac.ae
shady.shehata@uwaterloo.ca

Abstract

Speech emotion recognition is vital for human-computer interaction, particularly for low-resource languages like Arabic, which face challenges due to limited data and research. We introduce ArabEmoNet, a lightweight architecture designed to overcome these limitations and deliver state-of-the-art performance. Unlike previous systems relying on discrete MFCC features and 1D convolutions, which miss nuanced spectro-temporal patterns, ArabEmoNet uses Mel spectrograms processed through 2D convolutions, preserving critical emotional cues often lost in traditional methods. While recent models favor large-scale architectures with millions of parameters, ArabEmoNet achieves superior results with just 1 million parameters, which is 90 times smaller than HuBERT base and 74 times smaller than Whisper. This efficiency makes it ideal for resource-constrained environments. ArabEmoNet advances Arabic speech emotion recognition, offering exceptional performance and accessibility for real-world applications.

1 Introduction

Speech Emotion Recognition (SER) is essential for improving human-computer interaction, particularly in linguistically diverse contexts like Arabic speech. The complexity of detecting emotions from speech arises from variations in prosody, phonetics, and speaker expression. Over time, SER has evolved from statistical approaches to deep learning, significantly enhancing recognition accuracy.

Early SER systems relied on handcrafted acoustic features (e.g., pitch, energy, and MFCCs) processed using classical machine learning models like Support Vector Machines (SVMs) and Gaussian Mixture Models (GMMs) (Lieskovska et al., 2021). While effective, these methods struggled with cross-dataset generalization, particularly in Arabic speech, which exhibits rich phonetic and prosodic diversity. Deep learning mitigated these limitations by enabling automatic feature extraction, with CNNs capturing localized spectro-temporal patterns and LSTMs modeling sequential dependencies (Fayek et al., 2017). However, many Arabic SER systems still rely on MFCCs and 1D convolutions, which fail to capture essential spectral-temporal structures for robust emotion recognition.

Transformer-based models (Vaswani et al., 2017) introduced attention mechanisms to dynamically focus on emotionally salient speech segments (Mirsamadi et al., 2017). While effective in modeling long-range dependencies and parallelizing computations across emotional speech sequences, their high computational complexity ($O(n^2)$ for self-attention) and substantial memory requirements render them impractical for resource-constrained environments. To address these constraints, we propose ArabEmoNet, a lightweight architecture leveraging Mel spectrograms with 2D convolutions, effectively capturing both fine-grained spectral features and global contextual relationships (Kurpukdee et al., 2017).

Our model achieves competitive accuracy with just 0.97M parameters, making it significantly more efficient than HuBERT (Hsu et al., 2021) and Whisper (Radford et al., 2022) while maintaining state-of-the-art performance. Additionally, we augmented the data by integrating SpecAugment (Park et al., 2019) and Additive White Gaussian Noise (AWGN), which enhances the robustness of our model (Huh et al., 2024).

Experiments on KSUEmotions (Meftah et al., 2021) and KEDAS (Belhadj et al., 2022) datasets confirm that ArabEmoNet surpasses prior architectures while maintaining efficiency, marking a significant step forward in Arabic SER.

The main contributions of this paper can be summarized as follows:

*Equal contribution

Proceedings of The Third Arabic Natural Language Processing Conference, pages 211–218

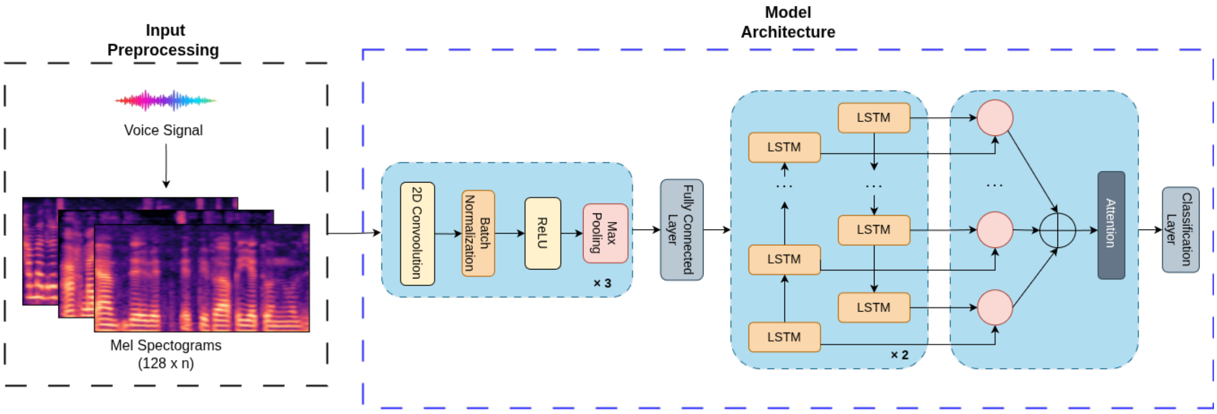

Figure 1: ArabEmoNet:2D CNN-Attention and BiLSTM Model Architecture.

- We propose ArabEmoNet: a novel lightweight hybrid architecture combining 2D Convolutional Neural Networks (CNN) with Bidirectional Long Short-Term Memory (BiLSTM) and an attention mechanism
- ArabEmoNet (1M parameters) achieves superior results with just 1 million parameters—90 times smaller than HuBERT base (95M parameters) and 74 times smaller than Whisper (74M parameters).
- We demonstrate ArabEmoNet's superior performance by achieving state-of-the-art results on the KSUEmotion and KEDAS datasets, surpassing previous benchmark models.

2 Related Work

Speech Emotion Recognition (SER) has been an active area of research for decades. Traditional approaches often relied on statistical evaluations of handcrafted speech features like pitch, energy, and spectral coefficients, combined with classifiers such as Support Vector Machines (SVMs) or Hidden Markov Models (HMMs) (Nwe et al., 2003; Schuller et al., 2011). Although these methods provided foundational insights, they often struggled to generalize across different datasets, speakers, and languages, motivating the shift towards feature learning with deep neural networks (Jahangir et al., 2021).

The advent of deep learning has established hybrid architectures combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) as a standard approach in SER (Sainath et al., 2015; Trigeorgis et al., 2016). In this paradigm, exemplified by recent studies from (Khan et al., 2024) and (Mishra et al., 2024), CNNs extract local features which are then modeled over time by an RNN. A key limitation in these models, however, is the common use of 1D convolutions, which process spectral and temporal information separately, potentially limiting their ability to capture intertwined spectro-temporal patterns.

To enhance the performance of these hybrid models, researchers have incorporated additional mechanisms. Attention mechanisms, introduced by (Bahdanau et al., 2015) and popularized by (Vaswani et al., 2017), have shown significant promise by allowing models to focus on the most salient segments of a speech utterance. A prior study by (Hifny and Ali, 2019b) successfully integrated an attention mechanism with a CNN and BiLSTM for an efficient Arabic SER system. While achieving state-of-the-art results on the KSUEmotions dataset (Hifny and Ali, 2019a), their approach was based on 13-feature Mel Frequency Cepstral Coefficients (MFCCs) and 1D convolutions, which may restrict the richness of the learned features.

Other works have explored more complex architectural variations to better exploit feature representations. For example, (Poorna et al., 2025) introduced a parallel model that processes Mel spectrograms through a CNN with a Time-Frequency Attention mechanism, while simultaneously feeding MFCC features to an attention-based BiLSTM. The learned features from these separate streams are then fused for final classification. While innovative, such parallel models can introduce significant complexity and may not fully exploit the intertwined nature of spectral and temporal patterns that exist within a single, rich input representation.

Building on these insights, our work addresses the limitations of prior approaches. We propose a unified, sequential architecture that diverges from the parallel processing of (Poorna et al., 2025) and the 1D convolutional layers used by (Mishra et al., 2024), (Khan et al., 2024), and (Hifny and Ali, 2019b). By employing **2D convolutions** directly on **Log-Mel spectrograms**, our model is designed to more effectively capture the critical spectro-temporal dependencies in a single processing stream. This architectural choice, combined with modern data augmentation techniques to enhance generalization, aims to provide a more robust and effective solution for SER.

3 Proposed Approach

In this work, we introduce ArabEmoNet, a dedicated 2D NN-Attention and BiLSTM framework optimized for Arabic Speech Emotion Recognition. Our model processes Log-Mel spectrograms to effectively capture the multifaceted nature of emotional speech through three complementary components: 2D convolutional layers that identify emotion-specific spectral patterns, bidirectional LSTMs that model the temporal evolution of emotional cues, and an attention mechanism that highlights emotionally salient segments within utterances. This integrated approach addresses the unique challenges of recognizing Arabic emotional expressions while maintaining a lightweight, efficient architecture. Figure 1 illustrates our complete model design.

3.1 Input Prepossessing

For our classification model, raw audio signals are transformed into Log-Mel spectrograms. This process involves computing the Mel spectrogram using a Fast Fourier Transform (FFT) window length of 2048 samples and a hop length of 256 samples. We generate 128 Mel bands across a frequency range from 80 Hz to 7600 Hz . A Hann window is applied to each frame to minimize spectral leakage. Subsequently, the resulting Mel spectrogram is converted to a logarithmic scale (decibels), referenced to the maximum power, to optimize the dynamic range for neural network processing.

3.2 Data Augmentation

To improve the generalization ability of the model and mitigate overfitting, we incorporate Gaussian noise augmentation during training. This technique simulates variations in the input data and leads to a more robust model. Optimization is performed using the Adam optimizer, which adapts learning rates for each parameter based on the first and second moments of the gradients. Additionally, we utilize batch normalization and early stopping based on validation loss to further stabilize the training process and prevent overfitting.

3.3 Feature Extraction via Convolutional Layers

The initial stage of the model employs a series of convolutional layers to extract high-level representations from the input Mel spectrograms. These layers are responsible for detecting local time-frequency patterns that are crucial for emotion discrimination. Mathematically, the feature maps $\mathbf{F}_l$ at layer l are computed as:

$$\mathbf{F}_l = \sigma\left(\text{Conv2D}(\mathbf{F}_{l-1}, \mathbf{W}_l, \text{padding} = p_l) + b_l\right)$$

where $\mathbf{F}_{l-1}$ represents the input to the current layer (with the initial input being the spectrogram $\mathbf{S}$), $\mathbf{W}_l$ and b_l denote the learnable weights and biases, respectively, p_l is the specified padding, and σ is the ReLU activation function. It is important to note that we employ 2D CNNs rather than 1D CNNs because Mel spectrograms provide a two-dimensional (time-frequency) representation. This allows the model to capture both temporal and spectral dependencies more effectively. The use of multiple convolutional layers, combined with max-pooling and dropout, enhances the network's ability to learn robust, hierarchical feature representations while mitigating overfitting. Following the convolutional layers, the extracted features are passed through a fully connected layer before being passed to the next stage.

3.4 Temporal Modeling with Bidirectional LSTM

After the convolutional layers, the network integrates a Bidirectional LSTM to model the temporal structure and contextual dependencies across time frames. By processing the sequential output in both forward and backward directions, the BiLSTM effectively captures transitions between emotional states, ensuring a more nuanced understanding of temporal variations in speech. The hidden state at time step t is given by:

$$\mathbf{h}_t = \left[\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t\right],$$

where $\overrightarrow{\mathbf{h}}_t$ and $\overleftarrow{\mathbf{h}}_t$ denote the forward and backward hidden states, respectively. This bidirectional processing is particularly important for SER tasks, as emotions in speech often evolve gradually rather than appearing in isolation. Capturing the transitions between emotional states allows the model to account for contextual cues, such as shifts in pitch, intensity, and rhythm, which are crucial for accurately interpreting emotional expressions over time.

3.5 Attention Mechanism

To enhance the model's ability to distinguish subtle variations in emotional expressions, an attention mechanism is integrated atop the BiLSTM outputs. This mechanism computes a context vector $\mathbf{c}$ that selectively aggregates the BiLSTM hidden states, assigning higher importance to frames that carry more salient emotional cues, thereby improving emotion classification. The context vector is defined as:

$$\mathbf{c} = \sum_t \alpha_t \mathbf{h}_t, \quad \text{with} \quad \alpha_t = \frac{\exp(e_t)}{\sum_k \exp(e_k)},$$

where the attention score e_t is computed as:

$$e_t = \tanh\left(\mathbf{w}_e^\top \mathbf{h}_t + b_e\right).$$

Here, $\mathbf{w}_e$ and b_e are learnable parameters that transform the hidden states into a scalar score, and the softmax function normalizes these scores into a probability distribution over time steps. By dynamically focusing on the most emotionally informative segments of the speech signal, this mechanism enhances the model's ability to capture key variations in tone, prosody, and intensity that define different emotional states, making it more effective for Speech Emotion Recognition (SER).

3.6 Classification Layer:

Finally, the context vector is passed through one fully connected layer, culminating in an output layer that produces the logits corresponding to the target emotion classes:

$$\mathbf{o} = \mathbf{W}_o \mathbf{c} + b_o.$$

The logits are then typically passed through a softmax function during training to compute the cross-entropy loss for classification. The entire architecture is illustrated in Figure 1.

Component	Configuration
Convolutional Layers	3 stages with filters: 32, 64, 128 Kernel: 7×7, ReLU activation Max pooling: 2×2, dropout: 0.3
Fully Connected	Input: $128 \times H'$; Output: 128 ReLU activation; dropout: 0.3
BiLSTM	2 layers, 64 hidden units per direction Dropout: 0.3
Attention	Applied to 128-dim BiLSTM output
Classification	Units equal to number of emotion categories

Table 1: Model Hyperparameter Configuration

4 Experimental setup

4.1 Training Platform

Training was done on a single Nvidia RTX 4090 GPU with 24 GB of memory. The training process utilized the Adam optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. An adaptive learning rate scheduler that reduces the learning rate when a metric's improvement plateaus was incorporated to adjust the learning rate during training, and the Adam optimizer was included.

4.2 Baselines

For our baseline models, we used Whisper-base, Whisper-small, and HuBERT-base speech encoders due to their vast popularity in the speech domain. We applied two identical feed-forward sublayers, each comprising a fully connected layer followed by a ReLU activation function and a dropout layer. This feed-forward block is repeated twice. After the feed-forward modules, the output is passed to a final classification layer that maps the learned features to the desired output classes. We trained the models using Adam optimizer with learning rate 1×10^{-3} and dropout 0.5. In addition to these general speech encoders, we also compared ArabEmoNet against several dataset-specific baseline models from the literature:

- For the KSUEmotion dataset, we compared against the ResNet-based Architecture (Meftah et al., 2021) and the CNN-BLSTM-DNN Model (Hifny and Ali, 2019b).
- For the KEDAS dataset, baseline (Belhadj et al., 2022) reported in the original dataset paper.

4.3 Datasets

In this work, we utilized two Arabic emotional speech datasets: the KSUEmotions corpus and

KEDAS, both designed to advance speech emotion recognition (SER) research in Arabic, addressing the scarcity of non-English SER resources. We sampled both datasets at their native frequencies: 16kHz for KSUEmotions and 48kHz for KEDAS. To handle varying sequence lengths in the dataset, shorter sequences within a batch were padded with zeros to match the longest sequence.

4.3.1 KSUEmotions Dataset

The KSUEmotions corpus (Meftah et al., 2021) provides recordings from 23 native Arabic speakers (10 males, 13 females) representing diverse dialectal backgrounds from Yemen, Saudi Arabia, and Syria. The corpus was collected in two phases:

1) Phase 1: Included 20 speakers (10 males, 10 females) recording five emotions: neutral, sadness, happiness, surprise, and questioning, totaling 2 hours and 55 minutes of high-quality audio recorded in controlled environments.

2) Phase 2: Featured 14 speakers (7 males and 4 females from Phase 1, plus 3 new Yemeni females), replacing the questioning emotion with anger, contributing an additional 2 hours and 15 minutes of recordings.

4.3.2 KEDAS Dataset

The KEDAS dataset (Belhadj et al., 2022) comprises 5000 audio recording files in standard Arabic, featuring five emotional states: anger, happiness, sadness, fear, and neutrality. The recordings were collected from 500 actors within the university community, including students, professors, and staff. The dataset is based on 10 carefully selected phrases commonly used in communication, chosen through literary and scientific studies. The data collection and validation process involved 55 evaluators, including Arabic linguists, literary researchers, and clinical psychology specialists, ensuring high-quality emotional content and linguistic accuracy.

4.4 Evaluation

To evaluate our classification model's performance, we used two key metrics: Macro F1-score and Micro F1-score. Since no specific train-test split was provided for the datasets, we follow (Hifny and Ali, 2019b) and report the average of a 5-fold cross-validation with stratified splits on both datasets.

4.4.1 Macro F1-Score

The macro F1-score (Sokolova et al., 2009) calculates the unweighted mean of F1-scores for each class. It treats all classes equally, regardless of their size, making it suitable for imbalanced datasets.

4.4.2 Micro F1-Score

The micro F1-score (Sokolova et al., 2009) aggregates the contributions of all classes to compute the average metric. Instead of treating all classes equally, it is weighted by the number of instances in each class, making it more suitable for balanced datasets.

5 Results

The results presented in Table 2 demonstrate the effectiveness and efficiency of the ArabEmoNet architecture for Arabic speech emotion recognition across two distinct datasets: KSUEmotion and KEDAS.

On the KSUEmotion dataset, ArabEmoNet achieves an accuracy of 91.48%, which represents state-of-the-art performance. This significantly outperforms previously established benchmarks for this dataset, including the CNN-BLSTM-DNN model (Hifny and Ali, 2019b) and the ResNet-based architecture (Meftah et al., 2021). Furthermore, ArabEmoNet also surpasses the performance of larger, pre-trained models such as HuBERT-base (Hsu et al., 2021) and Whisper-small (Radford et al., 2022), despite its significantly smaller parameter count.

Similarly, on the KEDAS dataset, our model achieves an exceptional accuracy of 99.46%. This result substantially surpasses the original Baseline Model (Belhadj et al., 2022) and demonstrates competitive performance even when compared to highly resource-intensive pre-trained models like Whisper-small (Radford et al., 2022) and HuBERT-base (Hsu et al., 2021). Notably, ArabEmoNet achieves these superior or competitive results with significantly fewer parameters (0.97M) compared to pretrained models such as HuBERT-base (95M) and Whisper-small (74M).

6 Discussion and Analysis

6.1 CNN Kernel Size

Table 3 shows the impact of kernel size on ArabEmoNet's performance for the KSUEmotion Dataset. As the kernel size increases from 3 to 7, the model's accuracy steadily improves, peaking at 91.48% with a kernel size of 7 and a corresponding padding of 3. Beyond this point, increasing the kernel size further (to 9 and 11) leads to a decline in accuracy.

Dataset	Model	Accuracy (%) ↑	Micro F1 (%)	Macro F1 (%)	Params (M)
KSUEmotion	Whisper-base (Radford et al., 2022)	78.81	76.77	78.81	74
	Hubert-base-Emotion	84.30	83.00	84.00	95
	ResNet-based Architecture (Meftah et al., 2021)	85.53	85.53	85.53	25
	Whisper-small (Radford et al., 2022)	85.98	85.96	85.98	244
	Hubert-base (Hsu et al., 2021)	87.04	<u>87.22</u>	87.04	95
	ArabEmoNet (Transformer) - **Ours**	86.66	86.66	86.66	1
	CNN-BLSTM-DNN Model (Hifny and Ali, 2019b)	<u>87.20</u>	87.20	<u>87.20</u>	-
	ArabEmoNet - **Ours**	**91.48**	**91.48**	**91.46**	1
KEDAS	Baseline Model (Belhadj et al., 2022)	75.00	75.00	75.00	-
	Whisper-base (Radford et al., 2022)	97.60	97.56	97.60	74
	Hubert-base-Emotion	98.00	97.98	98.00	95
	Hubert-base (Hsu et al., 2021)	99.35	**99.48**	**99.50**	95
	Whisper-small (Radford et al., 2022)	<u>99.40</u>	99.38	99.40	244
	ArabEmoNet - **Ours**	**99.46**	<u>99.46</u>	<u>99.42</u>	1

Table 2: Comparison of Models on KSUEmotion and KEDAS Datasets

Kernel Size	Padding	Accuracy (%)	Params (M)
11	5	89.90	1.71
9	4	91.15	1.29
7	3	**91.48**	0.97
5	2	90.08	0.71
3	1	89.71	0.55

Table 3: Impact of Changing Kernel Size for CNN Layers (KSUEmotion Dataset)

Emotion	Accuracy (%)
Neutral	93.75
Happy	88.37
Sad	95.38
Surprise	90.70
Angry	90.32
Fear	96.92

Table 4: Per-emotion results on the KSUEmotion dataset.

Larger kernels, while increasing the receptive field, may introduce too much noise or become less adept at capturing fine-grained details, leading to a dip in accuracy. Conversely, smaller kernels might not encompass enough contextual information to achieve optimal recognition. Therefore, the kernel size of 7 represents the best trade-off between performance and model complexity in this experimental setup.

6.2 Data Augmentation

To assess the contribution of data augmentation to the model's robustness and generalization, we compared the performance of our model trained with and without augmentation techniques on the KSUEmotion dataset. As shown in Table 5, employing data augmentation leads to a significant improvement in test accuracy, increasing from 89.10% to 91.48%. This improvement demonstrates the effectiveness of data augmentation in enhancing the model's generalization capabilities.

Training Strategy	Accuracy (%)
Without Augmentation	89.10
With Augmentation	**91.48**

Table 5: Impact of Data Augmentation on Model Performance (KSUEmotion Dataset)

6.3 Transformer-Based Architecture

To evaluate different architectural configurations, we performed further experiments with a CNN-Transformer model while keeping the remaining components unchanged. The Transformer-based architecture achieved an accuracy of 86.66% on the KSUEmotion dataset, as shown in Table 2, which is lower than ArabEmoNet's performance of 91.48%. This comparison suggests that the BiLSTM-based approach is more effective for Arabic dialectical speech emotion recognition tasks.

7 Conclusion

This study introduces ArabEmoNet, a lightweight yet highly effective architecture for Arabic Speech Emotion Recognition. By integrating 2D CNN layers, BiLSTM networks, and an attention mechanism with Mel spectrogram inputs, ArabEmoNet significantly advances the state-of-the-art, achieving a remarkable 4% improvement over existing models on the KSUEmotions dataset. Our results demonstrate that 2D convolutions substantially outperform traditional approaches using 1D convolutions and MFCC features, capturing richer and more nuanced acoustic patterns essential for emotion classification.

Furthermore, employing Gaussian noise augmentation successfully enhanced the model's robustness and addressed data imbalance issues, underscoring the importance of effective augmentation strategies. Comparative experiments revealed that transformer-based architectures, while powerful in other contexts, were less effective for this task, highlighting the particular suitability of BiLSTM layers in capturing temporal emotional dynamics.

In future work, we aim to extend ArabEmoNet's training to larger, multilingual datasets, validating its applicability and generalizability across diverse linguistic and cultural contexts. This expansion promises significant contributions toward more inclusive and effective global emotion recognition systems.

8 Limitations

A potential limitation to our architecture arises from the method used to handle variable audio lengths. To standardize the input size for model processing, the architecture employs zero-padding. Specifically, shorter audio sequences within any given batch are padded with zeros to equal the length of the longest sequence in that same batch. While this is a standard technique, it can introduce a limitation if there is significant variance in the duration of audio clips within a batch. In such cases, shorter clips will be appended with a large amount of non-informative zero values, which can lead to unnecessary computational processing and potentially impact the model's learning efficiency

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Mourad Belhadj, Ilham Bendellali, and Elalia Lakhdari. 2022. Kedas: A validated arabic speech emotion dataset. In *2022 International Symposium on iNnovative Informatics of Biskra (ISNIB)*, pages 1–6. IEEE.

Hazem M Fayek, Marcin Lech, and Luis Cavedon. 2017. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68.

Yasser Hifny and Ahmed Ali. 2019a. Efficient arabic emotion recognition using deep neural networks. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6710–6714.

Youssef Hifny and Ahmed Ali. 2019b. Efficient arabic emotion recognition using deep neural networks. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6710–6714. IEEE.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, volume 29, pages 3451–3460. IEEE.

Mina Huh, Ruchira Ray, and Corey Karnei. 2024. A comparison of speech data augmentation methods using s3prl toolkit. *Preprint*, arXiv:2303.00510.

Rashid Jahangir, Ying Wah Teh, Faiqa Hanif, and Ghulam Mujtaba. 2021. Deep learning approaches for speech emotion recognition: state of the art and research challenges. *Multimedia Tools and Applications*, 80(16):23745–23812.

Waleed Akram Khan, Hamad ul Qudous, and Asma Ahmad Farhan. 2024. Speech emotion recognition using feature fusion: a hybrid approach to deep learning. *Multimedia Tools and Applications*, 83(31):75557–75584.

Nattapong Kurpukdee, Tomoki Koriyama, Takao Kobayashi, Sawit Kasuriya, Chai Wutiwiwatchai, and Poonlap Lamsrichan. 2017. Speech emotion recognition using convolutional long short-term memory neural network and support vector machines. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1744–1749.

E. Lieskovska, M. Jakubec, R. Jarina, and M. Chmulik. 2021. A review on speech emotion recognition using deep learning and attention mechanism. *Electronics*, 10:1163.

Ali Hamid Meftah, Mustafa A. Qamhan, Yasser Seddiq, Yousef A. Alotaibi, and Sid Ahmed Selouani. 2021. King saud university emotions corpus: Construction, analysis, evaluation, and comparison. *IEEE Access*, 9:54201–54219.

Seyed-Ahmad Mirsamadi, Eslam Barsoum, and Chao Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Swami Mishra, Nehal Bhatnagar, Prakasam P, and Sureshkumar T. R. 2024. Speech emotion recognition and classification using hybrid deep cnn and bilstm model. *Multimedia Tools and Applications*, 83(13):37603–37620.

Tin Lay Nwe, Say Wei Foo, and Liyanage C. De Silva. 2003. Speech emotion recognition using hidden markov models. *Speech Communication*, 41(4):603–623.

David S Park and 1 others. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *INTERSPEECH*, pages 2613–2617.

SS Poorna, Vivek Menon, and Sundararaman Gopalan. 2025. Hybrid cnn-bilstm architecture with multiple attention mechanisms to enhance speech emotion recognition. *Biomedical Signal Processing and Control*, 100:106967.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR.

Tara N Sainath, Carolina Parada, Brian Kingsbury, and Bhuvana Ramabhadran. 2015. Convolutional, long short-term memory, fully connected deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech communication*, 53(9-10):1062–1087.

Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

George Trigeorgis, Marios A Nicolaou, Stefanos Zafeiriou, and Björn W Schuller. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Ashish Vaswani and 1 others. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.