MFCCFusionnet: A CNN–DNN feature-fusion model for speech emotion recognition using MFCC representations

Neslihan GÜNDOĞAN[1]
Buket İŞLER[*2]

Geliş tarihi / *Received:* 29.04.2026
Düzeltilerek geliş tarihi / *Received in revised form:* 29.04.2026
Kabul tarihi / *Accepted:* 29.06.2026
DOI: 10.17932/IAU.ABMYOD.2006.005/abmyod_v21i730010

Abstract
Speech emotion recognition has become an increasingly important research area because of its potential to enhance human–machine interaction and support the development of emotion-aware systems. However, further comparative evidence is still needed on how different deep learning architectures perform when they are trained using the same acoustic feature representation. In this study, an experimental comparison was conducted using the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset. Mel-frequency cepstral coefficients (MFCCs) were used as the sole feature extraction method, and three deep learning architectures were designed and evaluated: a Convolutional Neural Network (CNN), a Deep Neural Network (DNN), and a hybrid feature-fusion model named MFCCFusionNet. The CNN and DNN models achieved classification accuracies of 81% and 76%, respectively. The proposed MFCCFusionNet model combined the learned representations obtained from the CNN and DNN sub-models and achieved the highest accuracy among the evaluated models, with a test accuracy of 83%. The findings indicate that MFCC-based representations can be effectively processed by different deep learning architectures and that feature-level fusion may improve classification performance in speech-based emotion recognition. The study provides preliminary evidence that feature-level

[1] *İstanbul Topkapı Üniversitesi, nespo_ozkan@hotmail.com, ORCID:0000-0002-7366-3476*
[2*] *Sorumlu Yazar/ Corresponding Author:İstanbul Topkapı Üniversitesi, buketisler@topkapi.edu.tr, ORCID:0000-0002-9393-9564*

fusion of CNN- and DNN-based MFCC representations may improve classification performance in speech emotion recognition.

***Keywords**: Speech emotion recognition, MFCC, deep learning architectures*

MFCCFusionNet: MFCC temsilleri kullanılarak konuşma duygu tanıma için CNN–DNN tabanlı bir özellik birleştirme modeli

Öz

Konuşma duygu tanıma, insan–makine etkileşimini geliştirme ve duygu farkındalığına sahip akıllı sistemlerin geliştirilmesini destekleme potansiyeli nedeniyle önemli bir araştırma alanı hâline gelmiştir. Bununla birlikte, aynı akustik özellik temsili kullanılarak eğitilen farklı derin öğrenme mimarilerinin nasıl performans gösterdiğine ilişkin daha fazla karşılaştırmalı kanıta ihtiyaç duyulmaktadır. Bu çalışmada, RAVDESS veri seti kullanılarak deneysel bir karşılaştırma yapılmıştır. Mel-frekansı kepstral katsayıları (MFCC'ler) tek özellik çıkarım yöntemi olarak kullanılmış ve üç derin öğrenme mimarisi tasarlanarak değerlendirilmiştir: Evrişimli Sinir Ağı (CNN), Derin Sinir Ağı (DNN) ve MFCCFusionNet olarak adlandırılan hibrit özellik birleştirme modeli. CNN ve DNN modelleri sırasıyla %81 ve %76 sınıflandırma doğruluğu elde ederken, önerilen MFCC-Fuse modeli %83 test doğruluğu ile değerlendirilen modeller arasında en yüksek başarıyı göstermiştir. Bulgular, MFCC tabanlı temsillerin farklı derin öğrenme mimarileri tarafından etkili biçimde işlenebildiğini ve özellik düzeyinde birleştirmenin konuşma duygu tanıma performansında sınırlı fakat olumlu bir iyileşme sağlayabileceğini göstermektedir. Genel olarak çalışma, MFCC özelliklerinden öğrenilen CNN ve DNN tabanlı temsillerin birleştirilmesinin konuşma tabanlı duygu tanıma modellerinin ayırt edici kapasitesini artırabileceğine ilişkin ön kanıt sunmaktadır.

***Anahtar Kelimeler:** Konuşma duygu tanıma, MFCC, derin öğrenme mimarileri, özellik birleştirme, CNN, DNN*

Introduction

In recent years, speech emotion recognition (SER) has become an increasingly important research area within artificial intelligence and human–computer interaction. SER systems aim to automatically identify emotional states from vocal expressions and thereby support the development of more responsive, adaptive, and context-aware technologies. Such systems have potential applications in virtual assistants, customer service automation, distance education, social robotics, driver monitoring, affective computing, and digital mental health technologies. As these application areas continue to expand, the development of reliable and computationally efficient SER models has become an important research priority. The performance of SER systems largely depends on the quality of acoustic feature representation and the suitability of the classification model used to process these features. Speech signals contain complex information distributed across both temporal and frequency domains. Emotional expression may be reflected through variations in spectral structure, intensity, rhythm, articulation, and other acoustic characteristics. Therefore, feature extraction is a critical stage in SER because it transforms raw speech signals into structured representations that can be processed by machine learning or deep learning models. The main objective at this stage is to obtain compact yet informative representations that preserve emotion-related acoustic patterns without introducing excessive computational complexity.Various feature extraction techniques have been developed for speech-based emotion recognition. Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), Gammatone Frequency Cepstral Coefficients (GFCC), and RASTA-PLP have been used to represent different aspects of speech signals. LPC and PLP are effective in modelling the spectral envelope and articulatory-related characteristics of speech; however, when used alone, they may be limited in capturing the broader acoustic variations associated with emotional expression. GFCC provides a perceptually motivated representation and is often considered useful under noisy conditions, although its implementation may introduce additional computational complexity depending on the experimental setting. Despite these alternatives, Mel-frequency cepstral coefficients (MFCCs) remain one of the most widely used feature extraction methods in speech processing and SER studies because they provide a compact, computationally efficient, and perceptually informed representation of speech signals. MFCC extraction is based on the frequency sensitivity of

the human auditory system. In this process, the speech signal is divided into short-time frames and transformed into the frequency domain using the short-time Fourier transform. The resulting spectrum is passed through Mel-scaled filter banks, logarithmic energy values are computed, and the discrete cosine transform (DCT) is applied to obtain cepstral coefficients. In this way, MFCCs primarily represent the short-term spectral envelope of speech and may indirectly reflect emotion-related acoustic variations embedded in the signal. Their low-dimensional structure, computational efficiency, and established use in speech-based classification tasks make them suitable for comparative deep learning experiments. In traditional SER approaches, hand-crafted acoustic features such as MFCCs are generally classified using supervised machine learning algorithms. Support Vector Machines (SVM), Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), K-Nearest Neighbour (KNN), Decision Trees, Random Forests, Naive Bayes classifiers, and Logistic Regression have been frequently used for this purpose. These methods can provide acceptable classification performance, particularly in relatively small or controlled datasets. For example, SVM can construct effective decision boundaries in high-dimensional feature spaces, whereas HMM is useful for modelling sequential characteristics of speech. However, traditional machine learning methods generally depend heavily on predefined feature representations and may have limited adaptability when speech signals vary according to speaker identity, accent, speaking style, recording condition, or emotional intensity. Consequently, their generalisation capacity may be restricted in more heterogeneous conditions [1]. To address these limitations, deep learning architectures have increasingly been adopted in SER research. Deep learning models can learn hierarchical and nonlinear representations from acoustic inputs and may capture complex patterns that are difficult to represent through traditional classifiers alone. Convolutional neural networks (CNNs), although originally developed for image processing, have been widely applied to two-dimensional acoustic representations such as spectrograms, Mel-spectrograms, and MFCC matrices. When applied to MFCC-based inputs, CNNs can learn local time–frequency patterns and generate abstract feature maps that are useful for distinguishing emotional categories [2,3]. Deep neural networks (DNNs), by contrast, process feature vectors through multilayered fully connected structures. When MFCC features are provided as input, DNNs can transform low-level acoustic representations into more

abstract feature spaces and perform classification based on nonlinear combinations of these features. Although DNNs do not explicitly model local time–frequency relationships in the same way as CNNs, they can still serve as effective baseline architectures for SER tasks when compact and informative acoustic representations are used [4]. Long short-term memory (LSTM) networks have also been widely used in SER because of their ability to model temporal dependencies in sequential speech data. LSTM architectures are particularly useful for tracking changes in emotional expression over time and for learning temporal patterns related to rhythm, continuity, and dynamic acoustic variation [5,6]. However, the present study focuses specifically on CNN and DNN architectures in order to examine whether local pattern extraction and dense feature abstraction can provide complementary representations when trained on the same MFCC-based input. Although individual deep learning architectures have produced promising results in SER, single-model approaches may capture only certain aspects of speech representation. CNNs are effective in extracting local time–frequency patterns, whereas DNNs can learn abstract relationships among acoustic features in a fully connected feature space. These differences suggest that the representations learned by CNN and DNN architectures may contain complementary information. Accordingly, hybrid architectures and feature-fusion strategies have increasingly been explored in recent SER studies as a way of combining different representational strengths within a unified classification framework [7]. Recent studies have reported promising results using transfer learning, hybrid neural architectures, speaker embeddings, and multi-feature fusion strategies. For instance, Jakubec et al. [8] used transfer learning with speaker representation models such as d-vector, x-vector, and r-vector structures. Rajapakshe et al. [9] proposed the emoDARTS framework for the joint optimisation of CNN and sequential neural network architectures. Ulgen et al. [10] examined emotional clustering in speaker embeddings, while other studies have combined MFCCs with prosodic features, CNN-based representations, or ensemble structures to improve SER performance [11–14]. These studies indicate that the integration of complementary feature representations may improve the ability of SER models to distinguish complex emotional patterns. Within this framework, the present study comparatively evaluates CNN and DNN architectures using MFCC-based features extracted from the RAVDESS dataset and proposes a hybrid feature-fusion model named MFCCFusionNet. In the proposed model, the

CNN and DNN sub-models are first trained separately on the same MFCC-based emotion recognition task. Then, the representations learned by these sub-models are reused and concatenated within a unified classification structure. Therefore, the proposed approach should be understood not as transfer learning based on an external pre-trained model, but as an intra-dataset representation reuse and feature-level fusion strategy. This distinction is important because the model aims to examine whether task-specific representations learned by different architectures from the same acoustic input can improve classification performance when integrated into a single framework. The main contribution of this study is threefold. First, it provides a controlled comparison of CNN and DNN architectures under the same MFCC-based feature extraction and classification conditions. Second, it proposes MFCCFusionNet as a CNN–DNN feature-fusion model that combines local pattern extraction and abstract feature representation within a unified structure. Third, it evaluates whether the fusion of learned representations can provide a measurable improvement over individual baseline models in speech-based emotion recognition. In this respect, the study contributes to ongoing research on compact and interpretable deep learning approaches for SER, while also providing preliminary evidence on the potential value of feature-level fusion using MFCC representations. Beyond achieving high classification accuracy, this study offers a distinctive contribution to the literature by providing a comparative analysis of how different deep learning architectures extract and leverage distinct representation information from the same acoustic feature set.

Method

Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was used as the dataset in this study. RAVDESS is an open-access and multimodal emotional expression dataset that has been widely used in speech emotion recognition research. It consists of high-quality audio and video recordings produced by 24 professional actors, including 12 female and 12 male speakers, who speak English with a North American accent. In the dataset, each actor vocalised two semantically neutral sentences, “Kids are talking by the door” and “Dogs are sitting by the door”, across eight emotional categories: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. Each emotional category was recorded at different intensity levels, except for the neutral category,

and each sentence was repeated twice. Therefore, although RAVDESS is balanced in terms of actor gender, the emotional speech subset contains a mild class imbalance across emotion categories. This imbalance mainly arises from the neutral category, as neutral expressions are not recorded with a strong-intensity condition, unlike most other emotional categories. Consequently, the neutral class contains fewer samples than the other emotion classes. Although the full audio version of RAVDESS includes both speech and song recordings, only the emotional speech subset was used in the present study. Accordingly, the analyses were conducted on 1,440 speech recordings. Song recordings were excluded from the analysis in order to ensure methodological consistency and to focus exclusively on speech-based emotion recognition. All audio files were recorded at a sampling rate of 48 kHz and a resolution of 16 bits. In the present study, the mild class imbalance in the dataset was considered during data partitioning and performance evaluation. A stratified 80:20 training–test split was used to preserve the proportional distribution of emotion categories across both subsets. This ratio was selected after preliminary empirical trials with alternative partitioning ratios, as it provided comparatively stable classification performance while retaining a sufficient number of samples for independent testing. As a result, 1,152 recordings were used for model training and 288 recordings were reserved for testing. In addition to overall accuracy, class-based precision, recall, F1-score, and confusion matrices were reported to provide a more detailed evaluation of model behaviour across both underrepresented and more frequently represented emotion classes.

Methodology

In this study, three different deep learning architectures were designed and comparatively evaluated for the task of speech emotion recognition using MFCCs. The methodological approach was structured around three main components: audio signal preprocessing, deep learning-based feature extraction, and the classification process. In the first stage, the audio samples obtained from the RAVDESS dataset were digitised and subjected to preprocessing. The audio signals were processed at a sampling rate of 48 kHz and a resolution of 16 bits, and MFCC-based features were extracted from these signals. Each audio sample was divided into segments using a 25 ms window size and a 10 ms frame shift, with each frame being represented by 64 MFCC coefficients. This procedure strengthened the representation of speech in the time–frequency domain and generated low-

dimensional, information-rich vectors suitable for model input. The MFCC features obtained after the preprocessing stage were evaluated using three different model architectures: (i) CNN, (ii) DNN, and (iii) a hybrid feature-fusion architecture developed by combining the learned representations of the CNN and DNN sub-models, referred to as MFCCFusionNet. The CNN model consisted of three consecutive convolutional blocks and a fully connected classification layer. In each block, convolution, ReLU activation, batch normalisation, and max pooling operations were applied, respectively. The feature maps obtained from the final block were flattened and transferred to a dense layer with 256 neurons, after which the final classification was performed using dropout and a softmax layer. This architecture aimed to capture the differences between emotion classes by learning local patterns from the two-dimensional representation of MFCCs. The DNN architecture consisted of three fully connected dense layers containing 512, 256, and 256 neurons, respectively. ReLU activation, batch normalisation, and dropout were applied in each layer. This structure enabled the direct classification of MFCC features as one-dimensional vectors and allowed abstract representations to be learned progressively across layers. The proposed hybrid architecture was developed by combining the learned representations obtained from the final layers of the CNN and DNN models after these models had been trained separately. The feature vectors obtained from the dense layer of the CNN and the final hidden layer of the DNN were merged in a Concatenate layer, and classification was subsequently performed through two dense layers containing 128 and 64 neurons. In this architecture, batch normalisation, ReLU activation, and dropout were applied to each dense layer to improve generalisation capability. MFCCFusionNet aimed to improve classification performance by integrating the complementary learning capacities of different architectures. All models were trained on the RAVDESS dataset using an 80:20 training–test split, which provided the highest performance following preliminary experiments conducted with various data partitioning ratios. To ensure a consistent experimental setting, the same preprocessing, feature extraction, training, and evaluation procedures were applied to all three models. The models were trained using the Adam optimisation algorithm and a multi-class classification loss function suitable for the eight emotion categories. To reduce the risk of overfitting, dropout and batch normalisation were applied within the model architectures. The final evaluation was performed on the held-out test set, which was not used

during model training. In order to provide a fair comparison among the CNN, DNN, and MFCCFusionNet architectures, the same MFCC feature configuration, data partitioning strategy, class labels, and performance metrics were retained across all experiments. Performance evaluation was conducted using class-based metrics, including accuracy, precision, recall, and F1-score. In addition, confusion matrices were generated for each model to visualise within-class and between-class prediction performance.

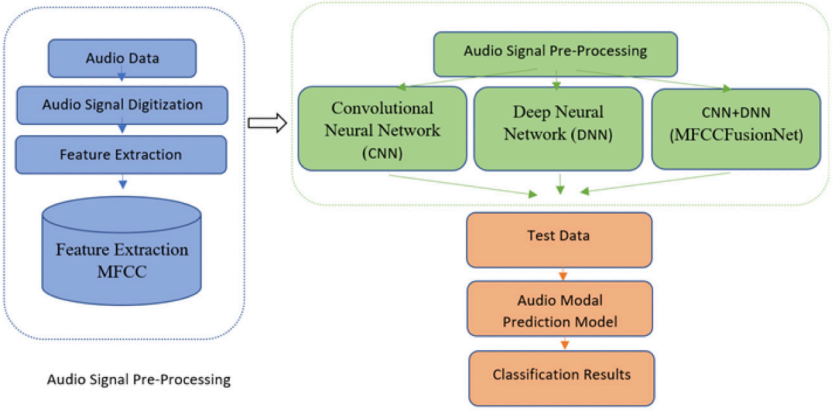

Figure 1. Deep learning emotion recognition process steps

Mel-Frequency Cepstral Coefficients (MFCC)

MFCCs are an effective feature extraction technique that enables meaningful and discriminative features to be automatically derived from audio signals. This method has been widely used in various application areas, particularly speech processing, automatic speech recognition, biometric systems, and speech-based emotion recognition. MFCC are based on the Mel scale, which models the sensitivity of the human auditory system to different frequencies. This scale transforms the frequency axis into a logarithmic structure based on the assumption that the human ear is more sensitive to lower frequencies. Thus, the audio signal can be represented in a form that is more closely aligned with human auditory perception. In the MFCC extraction process, the speech signal is divided into short-time frames to allow fixed-duration analysis. A Hamming window is then applied to each frame, and the transition from the time domain to the frequency domain is performed using the fast Fourier transform (FFT). This spectrum is filtered according to the Mel scale, and the logarithmic energy distribution is obtained. In the final stage, the discrete cosine transform (DCT) is applied

to derive the cepstral coefficients. These coefficients generate information-rich vectors that contain prosodic and acoustic characteristics of speech, such as pitch, intonation, and energy, beginning from the fundamental frequency components. One of the main advantages of MFCCs is that they provide low-dimensional features with high representational capacity. This reduces computational cost and enables classification algorithms to operate more effectively and rapidly. In addition, the relative robustness of MFCC vectors against noise increases their reliability in the analysis of audio signals. For these reasons, MFCCs have become one of the most widely preferred feature extraction methods in contemporary speech and emotion recognition systems [16].

Convolutional neural networks (CNN)

CNNs are among the deep learning architectures widely used particularly in the field of image processing, owing to their capacity to automatically learn local patterns. Features are extracted through convolution and pooling operations applied to the input data, and these representations are reconstructed at an abstract level in deeper layers. The ability to model temporal and frequency-based patterns in multidimensional data also makes CNNs an effective option for speech-based emotion recognition systems. In this context, CNNs have been shown to produce successful results on two-dimensional feature representations such as MFCC, Mel-spectrograms, and filter banks. Their ability to learn patterns containing emotional information, such as rhythm, intonation, and spectral energy, improves classification performance. In addition, parameter sharing and convolution operations reduce computational cost and mitigate the risk of overfitting. In these respects, CNN stands out in speech-based emotion recognition systems as both a powerful feature extraction mechanism and an efficient approach that provides high classification accuracy.

Deep neural networks (DNN)

DNNs are deep learning architectures capable of learning complex nonlinear relationships through multilayered fully connected structures. The connection of each neuron to all neurons in the preceding layer enables abstract representations to be extracted from high-dimensional data and improves classification performance. Particularly when trained with sufficient data, DNNs can represent the intrinsic characteristics of the data in an increasingly abstract manner across successive layers. In speech-based emotion recognition applications, DNNs operate on feature vectors

such as MFCCs and can directly learn the underlying emotional patterns in speech, thereby strengthening the distinction between classes. While architectures such as CNNs focus on spatial patterns, DNNs operate within a more general feature space and produce useful results in feature-level classification tasks. In this respect, they are widely used in speech-based emotion recognition systems and contribute to classification performance.

Hybrid architecture and feature-level fusion

This study employs a hybrid architecture strategy, which involves the integration of task-specific representations learned by separately trained CNN and DNN sub-models. Unlike conventional transfer learning, where a pre-trained model is adapted to a new domain, the approach proposed in this study focuses on intra-dataset feature fusion. By constructing a unified structure, we aim to combine the local pattern extraction capabilities of the CNN sub-model with the abstract feature representation capabilities of the DNN sub-model. This design enables the fusion of complementary acoustic information learned from the same dataset, thereby enhancing the model's ability to capture complex emotional patterns. This feature-level fusion approach ensures that the model leverages the distinct strengths of both architectures, leading to more robust emotion recognition performance without the need for external large-scale pre-trained networks.

Results

CNN analysis results

The first model architecture developed in this study was structured, as shown in Table 1, to consist of three convolutional blocks and one fully connected classification layer. Each block was designed to include a convolutional layer, a ReLU activation function, and a max-pooling layer. In this way, hierarchical learning was performed; first, low-level features were captured, and then more abstract and complex patterns were learned. The convolutional layers were responsible for extracting local patterns from the MFCC feature matrices provided as input, whereas the pooling layers reduced the spatial dimensions, thereby improving computational efficiency and reducing the risk of overfitting. By using 128, 96, and 64 filters, respectively, the features were represented at progressively higher levels of abstraction. The feature maps obtained from the convolutional blocks were flattened and transferred to a dense layer with 256 neurons. In this layer, batch normalisation and dropout at a rate of 50% were applied to improve the training process. Finally, an output layer with softmax

activation was used for the eight-class emotion recognition task. This architecture was designed to provide high classification performance by effectively learning both low-level acoustic features and emotional patterns.

Table 1. Architectural configuration and parameter specifications of the CNN model

Layer (Type)	Output Shape	Number of Parameters
Conv2D	(None, 128, 228, 128)	3328
BatchNormalization	(None, 128, 228, 128)	512
Activation	(None, 128, 228, 128)	0
MaxPooling2D	(None, 64, 114, 128)	0
Conv2D	(None, 64, 114, 96)	196704
BatchNormalization	(None, 64, 114, 96)	384
Activation	(None, 64, 114, 96)	0
MaxPooling2D	(None, 32, 57, 96)	0
Conv2D	(None, 32, 57, 64)	55360
BatchNormalization	(None, 32, 57, 64)	256
Activation	(None, 32, 57, 64)	0
MaxPooling2D	(None, 16, 28, 64)	0
Flatten	(None, 28672)	0
Dense	(None, 256)	7340288
BatchNormalization	(None, 256)	1024
Activation	(None, 256)	0
Dropout	(None, 256)	0
Dense	(None, 8)	2056

The performance of the model was evaluated on the test dataset, as presented in Table 2, with an accuracy rate of 81%. This accuracy rate indicates that the proposed CNN-based architecture achieved generally successful classification performance in the speech-based emotion recognition task. The model performance was detailed not only through the overall accuracy rate but also through the F1-score values calculated for each emotion class. According to the classification report obtained, the emotion classes for which the model achieved the highest F1-score values were “disgust” (89%), “surprised” (87%), and “calm” (86%), respectively.

Table 2. Class-based performance metrics of the CNN-based emotion recognition model

	Precision	Recall	F1-score	Support
Angry	0.78	0.78	0.78	36
Calm	0.83	0.89	0.86	44
Disgust	0.86	0.92	0.89	39
Fearful	0.89	0.71	0.79	34
Happy	0.80	0.78	0.79	41
Neutral	0.82	0.70	0.76	20
Sad	0.72	0.76	0.74	34
Surprised	0.84	0.90	0.87	40
Accuracy			0.81	288
Macro avg	0.82	0.80	0.81	288
Weighted avg	0.82	0.82	0.81	288

In addition, the classification performance of the model was analysed in greater detail through the confusion matrix presented in Figure 2. The matrix clearly shows that the model predicted the "disgust" and "surprised" classes with high accuracy. This indicates that the frequency-based features of these emotions could be easily distinguished by the model and that their likelihood of being confused with other classes was low.

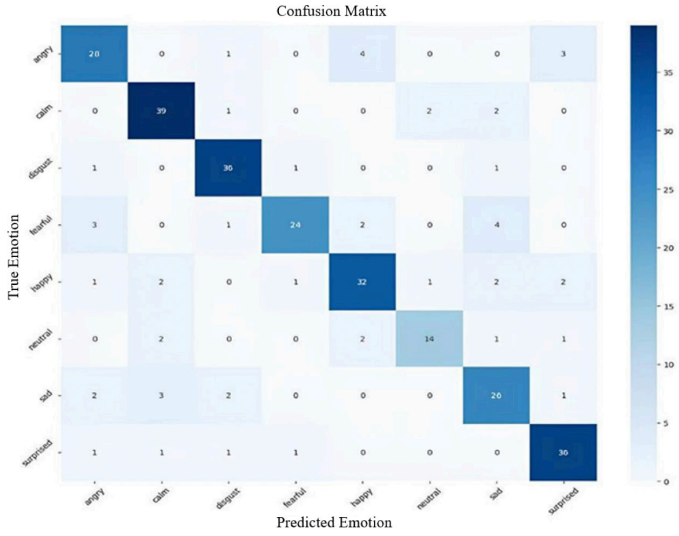

Figure 2. Confusion Matrix of the CNN Model

DNN analysis results

The second model applied in this study was the DNN-based architecture, which was constructed from three consecutive dense layers designed to process the MFCC features provided as input. These layers contained 512, 256, and 256 neurons, respectively. Each dense layer was sequentially followed by a ReLU activation function, batch normalisation, and dropout at a rate of 50%. This configuration was designed to enable the first dense layer to learn high-dimensional MFCC-based feature representations in detail, while the subsequent layers transformed these representations into more compact and abstract feature spaces.

Through the ReLU activation function applied in each dense layer, nonlinear relationships within the MFCC feature vectors were modelled. Batch normalisation contributed to the stability of the learning process, whereas dropout was applied to reduce the risk of overfitting. In particular, the 512 neurons in the first dense layer were used to support the effective processing of detailed acoustic features extracted from the MFCC matrices. The following two dense layers, each containing 256 neurons, were used to obtain lower-dimensional and more abstract representations, thereby strengthening the generalisation capability of the model. The architectural structure of the DNN model is presented in Table 3, together with the output dimensions and parameter counts of each layer.

Table 3. Architectural configuration and parameter specifications of the DNN model

Layer (Type)	Output Shape	Number of Parameters
Dense	(None, 512)	7,471,616
BatchNormalization	(None, 512)	2,048
Activation	(None, 512)	0
Dropout	(None, 512)	0
Dense	(None, 256)	131,328
BatchNormalization	(None, 256)	1,024
Activation	(None, 256)	0
Dropout	(None, 256)	0
Dense	(None, 256)	65,792
BatchNormalization	(None, 256)	1,024
Activation	(None, 256)	0
Dropout	(None, 256)	0
Dense	(None, 8)	2,056

As shown in Table 4, the DNN-based model achieved an overall accuracy of 76% on the test set. The calm and angry classes stood out as the most successfully classified emotions, with F1-score values of 0.87 and 0.83, respectively. By contrast, relatively lower performance was observed for the happy and fearful classes, with F1-score values of 0.68 and 0.70, respectively. The neutral class also showed a relatively low recall value of 0.65, indicating that neutral samples were more difficult to detect correctly. When the average performance values are considered, the macro-averaged and weighted-averaged F1-scores were both approximately 0.76. This indicates that the DNN model demonstrated a generally balanced classification performance across the emotion classes, although certain acoustically similar emotions were more difficult to distinguish.

Table 4. Class-based performance metrics of the DNN-Based emotion recognition model

	Precision	Recall	F1-score	Support
Angry	0.79	0.86	0.83	36
Calm	0.85	0.89	0.87	44
Disgust	0.74	0.79	0.77	39
Fearful	0.72	0.68	0.70	34
Happy	0.74	0.63	0.68	41
Neutral	0.81	0.65	0.72	20
Sad	0.68	0.76	0.72	34
Surprised	0.78	0.78	0.78	40
Accuracy			0.76	288
Macro avg	0.76	0.76	0.76	288
Weighted avg	0.76	0.76	0.76	288

The confusion matrix presented in Figure 3 provides a more detailed view of the class-based prediction performance of the DNN model. High correct classification rates were obtained for the calm class, with 39 out of 44 samples correctly classified, and for the angry class, with 31 out of 36 samples correctly classified. Similarly, the disgust and surprised classes were classified correctly in 31 out of 39 and 31 out of 40 samples, respectively. These findings suggest that the acoustic characteristics of these emotions were relatively more distinct and discriminative for the DNN architecture. By contrast, relatively higher confusion was observed in the happy and fearful classes. The model correctly classified 26 out of 41 happy samples and 23 out of 34 fearful samples. The happy emotion

was mainly confused with fearful and surprised emotions, while smaller numbers of samples were misclassified as angry, calm, sad, and neutral. In the neutral class, 13 out of 20 samples were correctly classified, whereas the remaining samples were mostly confused with calm and happy emotions. Overall, the confusion matrix indicates that the DNN model was more successful in distinguishing emotions with clearer acoustic patterns, such as calm, angry, disgust, and surprised. However, some classification errors occurred between emotionally and acoustically overlapping categories, particularly happy, fearful, neutral, and sad.

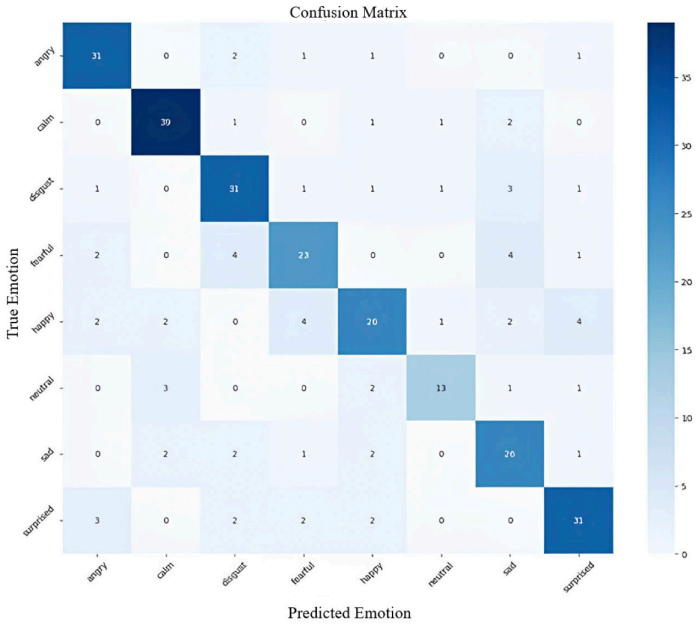

Figure 3. Confusion matrix of the DNN model

MFCCFusionnet hybrid model

The MFCCFusionNet model proposed in this study is a hybrid feature-fusion architecture developed by integrating MFCC-based representations learned by the CNN and DNN sub-models. In the first stage, the CNN and DNN architectures were trained separately on the same emotion recognition dataset in order to learn complementary acoustic representations from MFCC features. Subsequently, the trained layers of these sub-models were frozen and reused as fixed feature extractors within the hybrid architecture. In this sense, the model transfers task-specific representations learned by the individual sub-models into a unified classification structure. The proposed

model therefore does not rely on an external pre-trained network; rather, it reuses representations learned from the same dataset within an integrated feature-fusion framework. The 256-dimensional representation obtained from the dense layer of the CNN and the 256-dimensional representation obtained from the final hidden layer of the DNN were concatenated to form a 512-dimensional fused feature vector. The classification block was then constructed on this fused representation and consisted of two dense layers containing 128 and 64 neurons, respectively. Batch normalisation, ReLU activation, and dropout with a rate of 50% were applied in each dense layer to improve generalisation and reduce the risk of overfitting. This configuration enabled the complementary representations learned by the CNN and DNN sub-models to be integrated into a broader feature space. Thus, MFCCFusionNet was designed to improve emotion classification performance by combining local pattern extraction and abstract feature representation within a single architecture.

Table 5. Configuration and parameter specifications of the MFCCFusionNet model

Layer (Type)	Output Shape	Number of Parameters
Concatenate	(None, 512)	0
Dense	(None, 128)	65,664
BatchNormalization	(None, 128)	512
Activation	(None, 128)	0
Dropout	(None, 128)	0
Dense	(None, 64)	8,256
BatchNormalization	(None, 64)	256
Activation	(None, 64)	0
Dropout	(None, 64)	0
Dense	(None, 8)	520

The class-based performance metrics presented in Table 6 show that the MFCCFusionNet-based model achieved high accuracy and balanced classification performance in the speech-based emotion recognition task. The overall accuracy of the model on the test set was 83%. In particular, emotions such as disgust (88%), angry (87%), calm (87%), and surprised (87%) achieved high F1-score values, indicating that the model demonstrated strong discriminative capability for these classes. This success can be attributed to the combination of the CNN's ability to extract local patterns and the DNN's capacity to generate abstract representations.

A notable improvement was also observed for the happy (80%) and fearful (78%) emotions compared with the previous models. However, the F1-score values for neutral (72%) and sad (74%) remained lower, suggesting that these emotions tend to overlap more with other classes in acoustic terms. The macro-averaged and weighted-averaged F1-score values of the model, both at the level of 82–83%, indicate that generally balanced and strong classification performance was achieved across all classes. These results demonstrate that MFCCFusionNet can effectively learn features representing different emotion classes and has a generalisable structure.

Table 6. Class-based performance metrics of the MFCCFusionNet-based emotion recognition

	Precision	Recall	F1-score	Support
Angry	0.89	0.86	0.87	36
Calm	0.88	0.86	0.87	44
Disgust	0.85	0.90	0.88	39
Fearful	0.83	0.74	0.78	34
Happy	0.88	0.73	0.80	41
Neutral	0.74	0.70	0.72	20
Sad	0.69	0.79	0.74	34
Surprised	0.81	0.95	0.87	40
Accuracy			0.83	288
Macro avg	0.82	0.82	0.82	288
Weighted avg	0.83	0.83	0.83	288

The confusion matrix presented in Figure 4 reflects in detail the class-based discriminative capability of the MFCCFusionNet model. When the matrix is examined, it is observed that the model predicted the angry (31/36), calm (38/44), and disgust (35/39) emotions with high accuracy. The presence of distinct acoustic patterns in these classes reduced the risk of confusion with other emotions and made a positive contribution to the overall performance of the model. By contrast, a certain degree of confusion was observed in the fearful, happy, and neutral classes. For example, the happy emotion was occasionally confused with the surprised and neutral classes. This indicates that some emotions are more difficult to distinguish because of their acoustic similarities. Overall, the enriched representation structure generated through intra-dataset representation transfer and feature fusion enabled the model to effectively distinguish different emotional patterns. Thus, MFCCFusionNet provided a solution

with high accuracy and generalisability in the speech-based emotion recognition task.

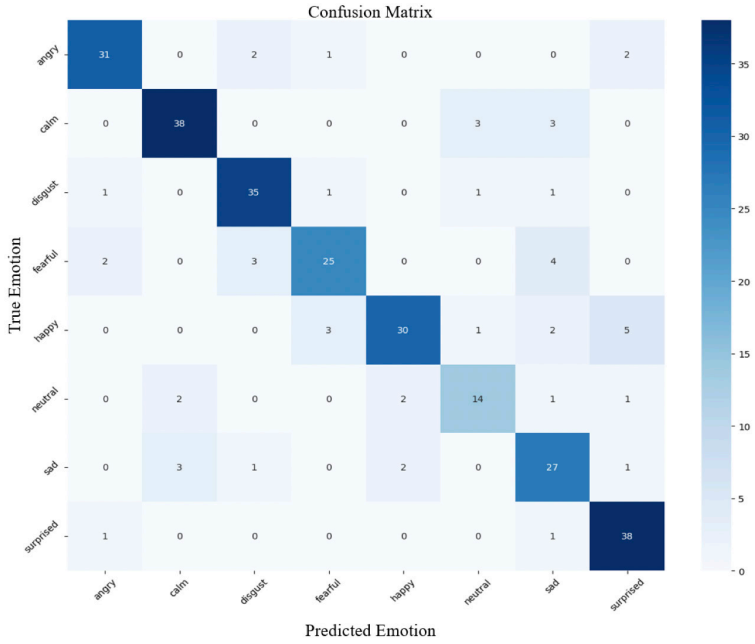

Figure 4. Confusion matrix of the MFCCFusionNet model

Discussion

In this study, three different deep learning models were developed and comparatively evaluated for the SER task using MFCC-based features extracted from the RAVDESS dataset. The findings indicate that CNN, DNN, and the proposed MFCCFusionNet model were all capable of learning discriminative representations from MFCC features. However, the hybrid feature-fusion strategy produced a more effective classification structure by integrating the complementary learning capacities of the CNN and DNN sub-models. The CNN architecture provided an effective solution for processing MFCC representations because of its ability to capture local acoustic patterns in the time–frequency domain. Since MFCC features contain compact information about the spectral and prosodic characteristics of speech, convolutional layers can identify emotion-related local variations more effectively than models based only on fully connected structures. This explains why the CNN model produced stable classification performance, particularly for emotions with more distinctive

acoustic patterns. Similar findings have also been reported in previous studies showing that CNN-based architectures can successfully process MFCC or spectrogram-based representations in SER tasks [18,19].

The confusion matrix results further suggest that emotions with clearer acoustic boundaries were classified more successfully, whereas acoustically overlapping categories were more difficult to distinguish. This is an expected outcome in SER studies because emotions such as neutral, sad, happy, and fearful may share similar prosodic cues depending on speaker characteristics, intensity level, and the acted nature of the dataset. Therefore, the classification errors observed between these categories should not be interpreted only as a limitation of the model architecture, but also as an indication of the intrinsic complexity of emotion recognition from speech. This issue has also been emphasised in previous studies that reported similar difficulties in distinguishing acoustically close emotion classes [20]. The DNN model also demonstrated a meaningful capacity to process MFCC-based input features. Its multilayered fully connected structure enabled the extraction of abstract representations from the acoustic feature vectors. However, unlike CNN architectures, DNNs do not explicitly model local spatial relationships within the MFCC representation. This may limit their ability to capture fine-grained time–frequency patterns that are relevant for emotion classification. Nevertheless, the findings indicate that a DNN architecture can still provide a useful baseline for SER when compact and information-rich acoustic features such as MFCCs are used. This supports previous studies showing that dense neural architectures may produce competitive results in speech-related classification problems when appropriate feature representations are provided [21,22]. The comparative findings show that the individual CNN and DNN models learned different types of information from the same MFCC input. While the CNN architecture was more effective in extracting local and structural patterns, the DNN architecture contributed to the formation of more abstract feature-level representations. This difference provides the main rationale for developing the proposed MFCCFusionNet model. By combining the representations learned by both sub-models, the hybrid architecture was able to construct a more comprehensive feature space for the final classification stage. MFCCFusionNet differed from the individual models by reusing the representations learned by the separately trained CNN and DNN sub-models within a unified feature-fusion structure. In this study, transfer learning was therefore applied in the sense of

transferring task-specific representations learned from the same dataset, rather than using an external pre-trained model. This point is important because the proposed model should be understood as a lightweight intra-dataset representation transfer and feature-fusion approach. Through this structure, the model benefited from both the local pattern extraction capacity of CNN and the abstract representation learning capability of DNN. The performance of the proposed hybrid model is consistent with the general direction of recent SER research, in which hybrid and feature-fusion-based architectures have increasingly been preferred to overcome the limitations of single-model approaches. Previous studies have shown that combining different neural components, such as CNN, BiLSTM, Transformer-based structures, ensemble learning, or transfer learning mechanisms, can improve the ability of SER systems to represent complex emotional patterns [23,24]. In this respect, the present study contributes to the literature by demonstrating that a relatively simple CNN–DNN fusion structure based only on MFCC features can produce a competitive and computationally efficient model. Compared with more complex hybrid or ensemble models, MFCCFusionNet has the advantage of relying on a single acoustic feature set and a relatively straightforward model integration strategy. Some previous approaches have used additional spectral images, temporal-context layers, dimensionality reduction procedures, or external pre-trained networks to improve classification performance [24,25]. Although such methods may achieve strong results, they may also increase computational complexity and reduce practical applicability, particularly in resource-constrained environments. By contrast, the proposed model offers a simpler and more interpretable structure while still benefiting from feature-level integration.

Another important implication of the findings is that MFCC features remain useful for SER when they are combined with appropriate deep learning architectures. Although recent studies have increasingly used complex acoustic descriptors, spectrogram images, or learned embeddings, MFCCs continue to provide a compact and computationally efficient representation of speech signals. The results of this study suggest that the effectiveness of MFCC features can be improved when they are processed through complementary neural architectures and then integrated at the representation level. Despite these contributions, several limitations should be acknowledged. First, the study was conducted using only the RAVDESS dataset, which consists of acted emotional speech.

Therefore, the generalisability of the findings to spontaneous speech, real-life conversations, multilingual contexts, or noisy environments remains limited. Second, although MFCC features provide an efficient acoustic representation, the exclusive use of a single feature type may restrict the model's ability to capture all emotional cues present in speech. Future studies may therefore integrate MFCCs with prosodic, spectral, or learned embedding-based features. Third, the proposed hybrid model was evaluated under experimental conditions, and further testing is required to determine its suitability for real-time or edge-device deployment. Overall, the findings indicate that feature-level integration is a promising strategy for improving speech emotion recognition performance. The proposed MFCCFusionNet model demonstrates that the complementary strengths of CNN and DNN architectures can be effectively combined within a unified classification framework. This provides both a methodological contribution to SER research and a practical direction for developing lightweight, efficient, and interpretable emotion recognition systems.

Conclusions

In this study, three deep learning architectures were compared for speech-based emotion recognition using MFCC features extracted from the RAVDESS dataset. The CNN, DNN, and MFCCFusionNet models were evaluated under the same experimental conditions. The results showed that all three models achieved acceptable classification performance; however, the hybrid MFCCFusionNet model achieved the highest accuracy among the evaluated architectures. This finding suggests that combining representations learned by CNN and DNN sub-models through feature-level fusion may improve the discriminative capacity of speech emotion recognition systems. The study indicates that MFCC features remain useful for emotion classification when they are processed through appropriate deep learning architectures. In particular, the proposed CNN–DNN fusion structure offers a relatively simple and computationally efficient approach compared with more complex hybrid models. Nevertheless, the study has several limitations. The experiments were conducted using only the RAVDESS dataset, which consists of acted emotional speech in English. Therefore, further validation is required using spontaneous speech data, multilingual datasets, different accents, and noisy real-world environments. Future studies may improve the proposed approach by incorporating additional acoustic features, applying data augmentation techniques, and testing the model in real-time or edge-device environments.

Such extensions may contribute to the development of more reliable and practically applicable speech emotion recognition systems for healthcare, education, customer service, and human–computer interaction.

Acknowledgement
This article was derived from the master's thesis entitled "The Use of Deep Learning Approaches in Speech-Based Emotion Recognition", prepared by Neslihan GÜNDOĞAN under the supervision of Assist. Prof. Dr. Buket İŞLER.

Conflicts of interest
The author declares no conflicts of interest.

Ethics Statement:
This study was conducted using the open-access RAVDESS dataset. No human or animal subjects were involved, and only existing audio recordings were analyzed. Therefore, ethical approval was not required.

References

[1] Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, 47, 312-323.

[2] Makhmudov, F., Kutlimuratov, A., & Cho, Y. I. (2024). Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition. *Applied Sciences*, 14(23), 11342.

[3] Dinesh, P., Mishra, S. P., Warule, P., & Deb, S. (2024, December). Speech Emotion Recognition with DNN and Combination of CNN-LSTM. In TENCON 2024-2024 IEEE Region 10 Conference (TENCON) (pp. 1159-1162). IEEE.

[4] Yang, Z., Li, Z., Zhou, S., Zhang, L., & Serikawa, S. (2024). Speech emotion recognition based on multi-feature speed rate and LSTM. *Neurocomputing*, 601, 128177.

[5] Shetty, K. J., Shetty, S., & Shetty, M. (2024, April). Speech emotion recognition using lstm. In 2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) (pp. 1-6). IEEE.

[6] Şeker, A., Diri, B., Balık, H. H. (2017). Derin öğrenme yöntemleri

ve uygulamaları hakkında bir inceleme. *Gazi Mühendislik Bilimleri Dergisi*, *3*(3), 47–64.

[7] Ezz-Eldin, M., Khalaf, A. A. M., Hamed, H. F. A., & Hussein, A. I. (2021). Efficient feature-aware hybrid model of deep learning architectures for speech emotion recognition. *IEEE Access*, *9*, 19999–20011.

[8] Jakubec, M., Lieskovska, E., Jarina, R., Spisiak, M., & Kasak, P. (2024). Speech emotion recognition using transfer learning: Integration of advanced speaker embeddings and image recognition models. *Applied Sciences*, *14*(21), 9981.

[9] Rajapakshe, T., Rana, R., Khalifa, S., Sisman, B., Schuller, B. W., & Busso, C. (2024). emodarts: Joint optimisation of cnn & sequential neural network architectures for superior speech emotion recognition. *IEEE Access*.

[10] Ulgen, I. R., Du, Z., Busso, C., Sisman, B. (2024). Revealing emotional clusters in speaker embeddings: A contrastive learning strategy for speech emotion recognition. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12081–12085.

[11] Madan, P., Dixit, D., Shobana, J. (2024). Speech emotion recognition system using MLP Classifier Algorithm. *2024 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, 1–6.

[12] Şengül, F., Akkaya, S. (2024). A Modified MFCC-based deep Learning method for emotion classification from speech. *International Advanced Researches and Engineering Journal*, *8*(1), 33–42.

[13] Akinpelu, S., Viriri, S. (2024). Deep Learning Framework for Speech Emotion Classification: A Survey of the State-of-the-Art. *IEEE Access*.

[14] Chowdhury, J. H., Ramanna, S., & Kotecha, K. (2025). Speech emotion recognition with light weight deep neural ensemble model using hand crafted features. *Scientific Reports*, *15*(1), 11824.

[15] Donuk, K., Hanbay, D. (2022). Konuşma Duygu Tanıma için Akustik Özelliklere Dayalı LSTM Tabanlı Bir Yaklaşım. *Computer Science*, *7*(2), 54–67.

[16] Yetkin, A. K., Köse, H. (2023, July). Speech-Based Emotion Analysis Using Log-Mel Spectrograms and MFCC Features. In 2023 31st Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[17] Kılıç, F. R. (2023). Modlar Arası Transfer Öğrenimi İle Ses Sinyallerinden Duygu Tanıma.

[18] Mountzouris, K., Perikos, I., & Hatzilygeroudis, I. (2023). Speech emotion recognition using convolutional neural networks with attention mechanism. Electronics, 12(20), 4376.

[19] Middya, A. I., Nag, B., Roy, S. (2024). Effective MLP and CNN based ensemble learning for speech emotion recognition. *Multimedia Tools and Applications*, 83(36), 83963-83990.

[20] Kumar, C. A., Maharana, A. D., Krishnan, S. M., Hanuma, S. S. S., Lal, G. J., Ravi, V. (2022, December). Speech emotion recognition using cnn-lstm and vision transformer. In International Conference on Innovations in Bio-Inspired Computing and Applications (pp. 86-97). Cham: Springer Nature Switzerland.

[21] Zeng, Y., Mao, H., Peng, D. Yi, Z. Spectrogram based multi-task audio classification. Multimed. *Tools Appl*. 78, 3705–3722 (2019).

[22] Issa, D., Demirci, M. F. & Yazici, A. Speech emotion recognition with deep convolutional neural networks. Biomed. Signal Process. *Control* 59, 101894 (2020).

[23] Mustaqeem, M., Sajjad, M., & Kwon, S. (2020). Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. IEEE Access, 8, 79861–79875.

[24] Kim, S., Lee, S.-P. (2023). A BiLSTM–Transformer and 2D CNN Architecture for Emotion Recognition from Speech. *Electronics*, 12(19), 4034.

[25] Akinpelu, S., Viriri, S., Adegun, A. (2023). Lightweight deep learning framework for speech emotion recognition. *IEEE Access*, 11, 77086-77098.