applied sciences

MDPI

Article

Speech Emotion Recognition Based on Two-Stream Deep Learning Model Using Korean Audio Information

A-Hyeon Jo [1] and Keun-Chang Kwak [2,]*

[1] Electronic Engineering IT-Bio Convergence System Major, Chosun University, Gwangju 61452, Republic of Korea
[2] Electronic Engineering, Chosun University, Gwangju 61452, Republic of Korea
* Correspondence: kwak@chosun.ac.kr; Tel.: +82-062-230-6086

Abstract: Identifying a person's emotions is an important element in communication. In particular, voice is a means of communication for easily and naturally expressing emotions. Speech emotion recognition technology is a crucial component of human–computer interaction (HCI), in which accurately identifying emotions is key. Therefore, this study presents a two-stream-based emotion recognition model based on bidirectional long short-term memory (Bi-LSTM) and convolutional neural networks (CNNs) using a Korean speech emotion database, and the performance is comparatively analyzed. The data used in the experiment were obtained from the Korean speech emotion recognition database built by Chosun University. Two deep learning models, Bi-LSTM and YAMNet, which is a CNN-based transfer learning model, were connected in a two-stream architecture to design an emotion recognition model. Various speech feature extraction methods and deep learning models were compared in terms of performance. Consequently, the speech emotion recognition performance of Bi-LSTM and YAMNet was 90.38% and 94.91%, respectively. However, the performance of the two-stream model was 96%, which was a minimum of 1.09% and up to 5.62% improved compared with a single model.

Keywords: speech emotion recognition; human–computer interaction; two-stream; bidirectional long-short term memory; convolutional neural network

check for updates

Citation: Jo, A.-H.; Kwak, K.-C. Speech Emotion Recognition Based on Two-Stream Deep Learning Model Using Korean Audio Information. *Appl. Sci.* **2023**, *13*, 2167. https://doi.org/10.3390/app13042167

Academic Editors: Jan Egger and Chun-Xia Zhang

Received: 2 January 2023
Revised: 1 February 2023
Accepted: 7 February 2023
Published: 8 February 2023

1. Introduction

Different emotions, such as happiness, anger, sadness, and surprise, play an important role in communication among people, and speech (voice) is a communication means that easily conveys emotions in the most natural manner. Precisely recognizing human emotions using speech data has become an essential research area in human–computer interaction (HCI), particularly as artificial intelligence (AI) technology is progressing rapidly [1]. Speech emotion recognition (SER) technology can enhance the quality of HCI by accurately classifying human emotions and enabling machines to adequately grasp user intention [2]. This technology is receiving increasing attention from researchers because of its applicability in diverse fields, including medical and customer service robots. Deep learning technology has been successfully applied and advanced in the fields of image recognition and speech recognition; thus, studies have begun employing it in SER, and research is actively being conducted on deep learning-based SER algorithms [3–11].

Through experimentation, Kipyatkova [3] deduced that a long short-term memory (LSTM) network is effective in large-scale acoustic and speech sequence modeling for each layer of the network for the long-term dependency characteristics of speech sequences.

Basu [4] used features based on the mel frequency cepstral coefficients (MFCCs) as input and implemented an SER algorithm using a convolutional neural network (CNN), which is a deep learning model, and an LSTM network.

Peng [5] proposed an SER system in which a back-end deep learning model combining 3D convolution and attention-based sliding recurrent neural networks (ASRNN) was used with an auditory perception-based front-end using auditory signal processing and a temporal attention mechanism. In this system, the front-end is used to generate time-modulated signals, and the attention-based back end is used to identify emotional states in speech.

Bhosale [6] proposed an end-to-end model utilizing a multihead self-attention mechanism and convolutional layer that uses both encoded language and audio spectrogram as input.

Liu [7] proposed a deep learning-based feature fusion method for heightening the performance of an SER system. The extraction of hyperprosodic features (EHPF) involves extracting hyperparameter statistical features from prosodic features. Two-dimensional spectrogram features are extracted from the raw voice signal and then used as input for training the CNN network. Further, the spectra-based feature vectors are extracted using the CNN network and EHPF feature vectors are fused for emotion recognition via a deep neural network (DNN). The experiment confirmed that the proposed model effectively improved the SER accuracy.

Zayene [8] proposed a 3D convolutional recurrent global neural network (CRGNN) that uses log-mels (static, deltas, and delta-deltas for the log mel-spectrogram) as input for SER. The model consists of a CNN to extract log-mels and local invariant features, followed by a recurrent neural network (RNN) that learns temporal dependency between different time-step local invariant features. Finally, the most active feature is selected using the global max pooling mechanism. To evaluate model performance, experiments were conducted on four datasets, and the effectiveness of the model was demonstrated by significantly improving the SER accuracy, compared with that of other approaches.

Zhang [9] proposed a two-stream emotion-embedded autoencoder, which is a new type of autoencoder architecture, for extracting the features of deeper emotions. The first stream of the autoencoder is a basic autoencoder used to learn the best representation of speech. The second stream learns the information of strong emotions in labels through an emotion-embedding path. Finally, the autoencoder and emotion embedding are combined, and batch normalization (BN) is then converted to instance normalization (IN). In the emotion classification process, deep emotion features output by the two-stream autoencoder are fused with the IS10 feature set obtained from the openSMILE toolkit, and emotion classification is then performed by adopting a fully connected network based on the connected feature vectors. The proposed model was evaluated on the IEMOCAP database, resulting in a 71.56% classification accuracy.

Han [10] proposed a new parallel network for SER by connecting a ResNet-CNN-Transformer encoder. ResNet uses the Gaussian error linear unit (GELU) as an activation function to solve the gradient vanishing problem caused by the deepening of the network, while the CNN is used to compute fewer parameters to then improve the fitting expression capability of the network. Furthermore, a transformer encoder is used to predict the frequency distribution of various emotions using a multihead self-attention layer considering the continuity of speech over time. For the fusion of these three models, the outputs of the ResNet and CNN are planarized to a one-dimensional vector and connected with the output of the transformer encoder; subsequently, the fully connected layer and softmax layer are sequentially connected to classify eight types of emotions. This model was assessed on the RAVDESS dataset and achieved an 80.89% classification accuracy, which is higher than that observed in previous studies.

Kakuba [11] proposed a model that uses a hybrid of self and multihead attention mechanisms along with dilated convolutions and an LSTM to achieve a comparable performance. Such a method computes global context dependencies between features in parallel using multihead attention. Furthermore, the global context and long-term dependency are computed using the self-attention mechanism in the Bi-LSTM layer stack. Using a dilated convolution layer improves the receptive field because the increase in the number of parameters is low, compared with the number of layers. As models that use raw signals tend to confuse happiness with anger or neutral with sadness, they typically employ spectrum and audio quality features extracted from raw audio signals as input. In terms of the performance evaluation of the proposed model, 96.36% accuracy was achieved on the EMODB dataset and 88.96% accuracy was achieved on the RAVDESS dataset.

Unfortunately, previous studies have limitations in that multidimensional feature information cannot be utilized as SER is conducted using only one of the one-dimensional spectral features or two-dimensional spectrogram features. Therefore, this study designed a two-stream-based SER model involving a Bi-LSTM network and CNN-based transfer learning model using multidimensional features (1D and 2D) of the Korean speech emotion database containing eight emotions, and then comparatively analyzed the emotion recognition model's performance according to diverse feature extraction methods. Since it uses multidimensional features, it is possible to utilize multidimensional feature information of audio, which has the advantage of being able to classify a person's emotional state more objectively and accurately. The remainder of this paper is organized as follows. Section 2.1 discusses the theoretical concept of the one-dimensional spectrum feature exaction method and two-dimensional spectrogram feature extraction method of speech data. Section 2.2 introduces the deep learning models used in this study, and Section 2.3 details the architecture of the proposed two-stream-based SER model employing Bi-LSTM and CNN. Section 3 overviews the Korean speech database and then discusses the comparative analysis of the results and SER performances. Section 4 discusses the findings of this study. Finally, Section 5 explains the outcomes of this study and proposes direction for future research.

2. Materials and Methods

2.1. Feature Extraction Methods of Speech Data

2.1.1. One-Dimensional Spectrum Feature Extraction

Extracting useful features is the most critical factor in implementing a model that uses speech data. In this study, the one-dimensional spectrum features of speech data were extracted using MFCCs and gammatone cepstral coefficients (GTCCs).

The MFCC is one of the most frequently used feature extraction methods in speech recognition. It extracts feature coefficients from the audio based on a mel filter bank reflecting the characteristics of the human auditory frequency range [12]. This is derived from the discrete cosine transform of the log power spectrum and is used to represent the short-term power spectrum of a sound. The purpose of extracting MFCC features from the speech is to determine the most compressed and beneficial set of features for improving efficiency [13]. In this study, the speech data were segmented into a frame unit, and these frames were overlapped by 25 ms and applied with a hamming window of 70 ms to extract the features of the MFCCs.

Unlike MFCC, GTCC is calculated by applying an equivalent rectangular bandwidth (ERB)-based gamma-tone filter bank rather than a mel filter bank. The ERB is a frequency scale used in psychoacoustics to measure the width of the auditory filter at each point. The gamma-tone filter bank is defined as an impulse response in the time domain, which provides an approximation to the bandwidth of the auditory filter of humans [14]. Similar to MFCCs, the speech data are segmented into a frame unit, and these frames are overlapped by 25 ms and applied with a hamming window of 70 ms to compute the GTCC values.

2.1.2. Two-Dimensional Spectrogram Feature Extraction

To train the 2D CNN model with speech data of the 1D time domain, we must convert the speech data into 2D spectrogram images by converting them from the time domain into the frequency domain. In this study, we extracted Bark spectrogram, ERB spectrogram, and log-mel spectrogram features, which are 2D images, based on the time–frequency conversion, and used them as inputs to the CNN transfer learning model.

The Bark spectrogram is based on the Bark frequency scale, which distinguishes sound characteristics. The Bark scale was first proposed by Barkhause as a subjective measurement of loudness [15]. The Bark spectrogram can be obtained by designing an auditory filter bank using the Bark frequency scale and then applying it to the spectrogram through a short-time Fourier transform. Figure 1 visualizes the Bark spectrogram for eight emotions in which the features are demonstrated.

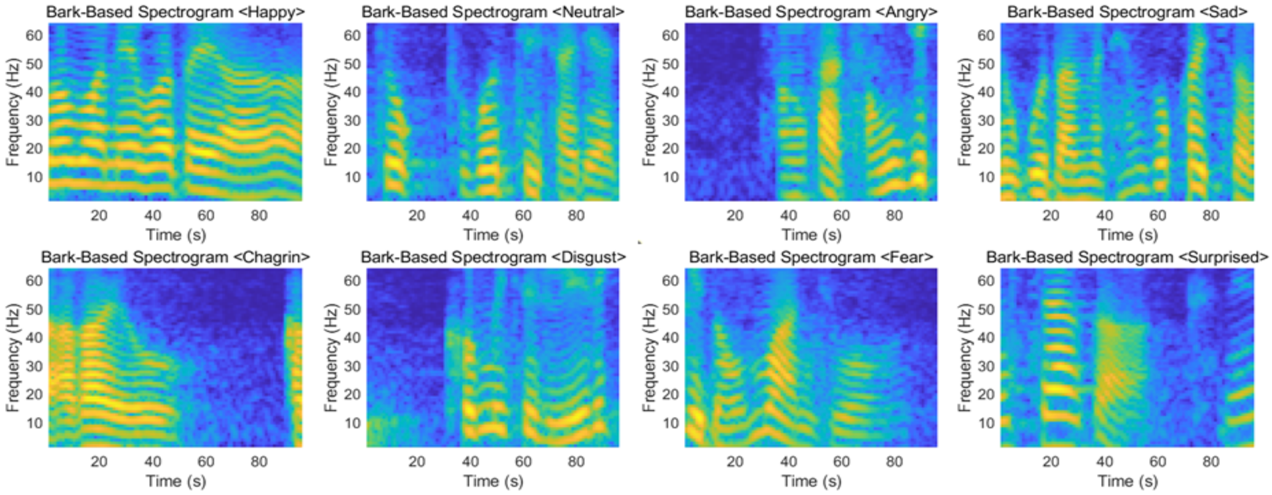


Figure 1. Visualization of Bark spectrograms for eight emotions.

The ERB spectrogram is based on the ERB scale, which is used in psychoacoustics. The ERB scale provides an approximation for the filter bandwidth in human hearing [16], and if an ERB auditory filter is designed based on it and is applied to the STFT spectrogram, an ERB spectrogram image can be obtained. Figure 2 presents a visualization of ERB spectrograms for eight emotions and shows the characteristics of the ERB spectrograms.

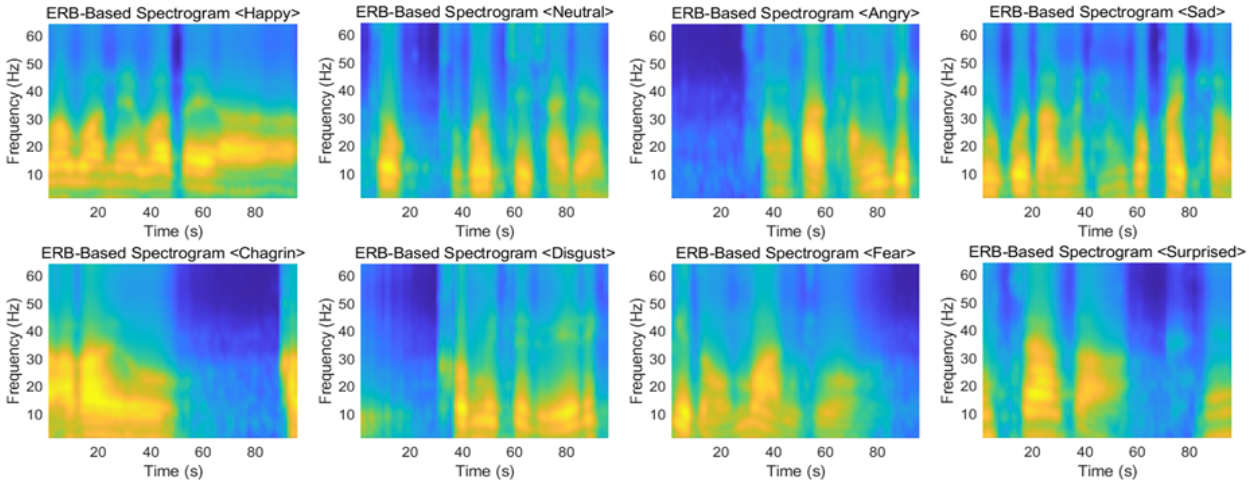


Figure 2. Visualization of ERB spectrograms for eight emotions.

The log-mel spectrogram is based on the mel scale, which was created to be most similar to the human auditory structure. The auditory filter bank was created based on the mel scale and applied to the STFT spectrogram to obtain the mel spectrogram, and log conversion is then applied to obtain the log-mel spectrogram. Figure 3 visualizes the log-mel spectrogram for eight emotions in which the features are shown.

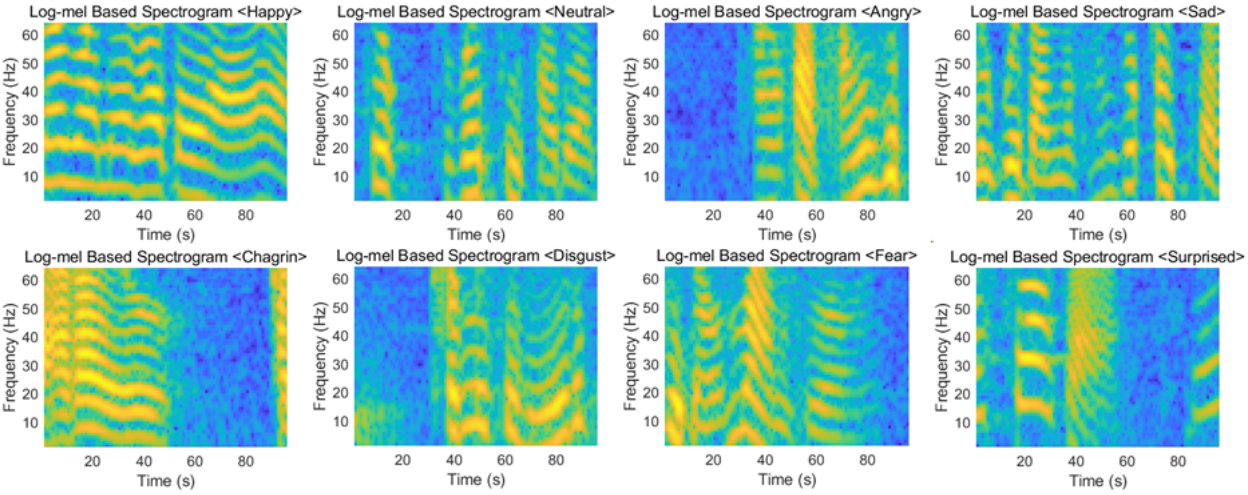


Figure 3. Visualization of log-mel spectrograms for eight emotions.

2.2. Deep Learning Models

2.2.1. One-Dimensional Audio Signal-Based LSTM and Bi-LSTM Models

LSTM is a type of RNN that iterates and maintains the information obtained from a previous step. It is a deep learning model explicitly designed to solve the long-term dependency problem of an RNN and can perform learning that requires a long dependency time. LSTM was first introduced by Hochreiter and Schmidhuber [17] and has been widely used across various fields owing to its advancements and popularity achieved.

All RNNs have a chain-like form that repeats neural network modules, and the repeated modules consist of one neural network layer. As shown in Figure 4, LSTM has a chain-like form; however, each repeated module has four layers exchanging information with each other in a specific manner. The core idea was designed based on a cell state. The cell state functions as a conveyor belt, and thus operates the entire chain continuously by applying a small linear interaction. It ensures that information flows without being altered. Furthermore, LSTM can add or remove certain elements to or from the cell state, which is controlled by a "gate" structure. The gate is an additional measure for delivering information and consists of a sigmoid layer and pointwise operation.

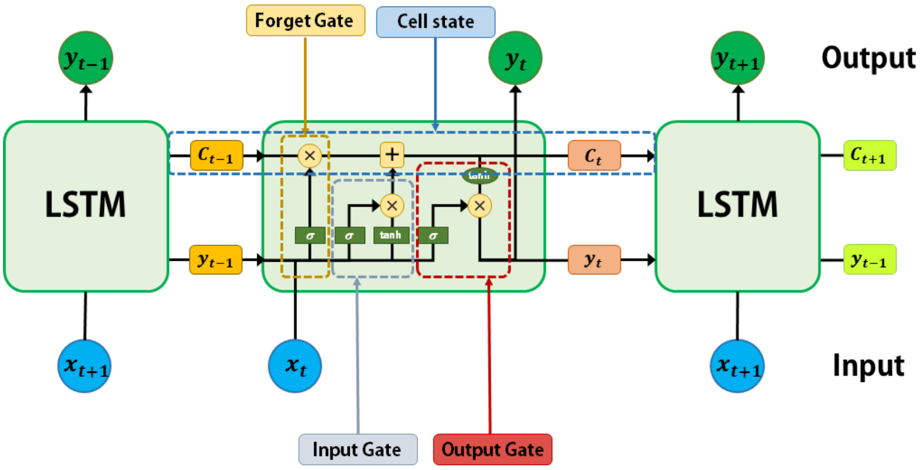


Figure 4. LSTM architecture.

Unlike the original LSTM, Bi-LSTM receives input from both directions and can utilize the information of both directions. LSTM can only perform forward learning in a sequential manner, whereas Bi-LSTM can learn forwards and backwards.

In Figure 5, the architecture of the Bi-LSTM indicates that the forward-direction LSTM in pink blocks is in parallel with the backward-direction LSTM in purple blocks. Bi-LSTM adds one more LSTM layer that reverses the direction of information flow. That is, the input sequence flows in the backwards direction in the additional LSTM layer. Then, the outputs of two LSTM layers are combined through various methods, including mean, sum, multiplication, and concatenation. Bi-LSTM is generally used when the task order must be designated. This type of network can be used for text classification, speech recognition, and prediction models.

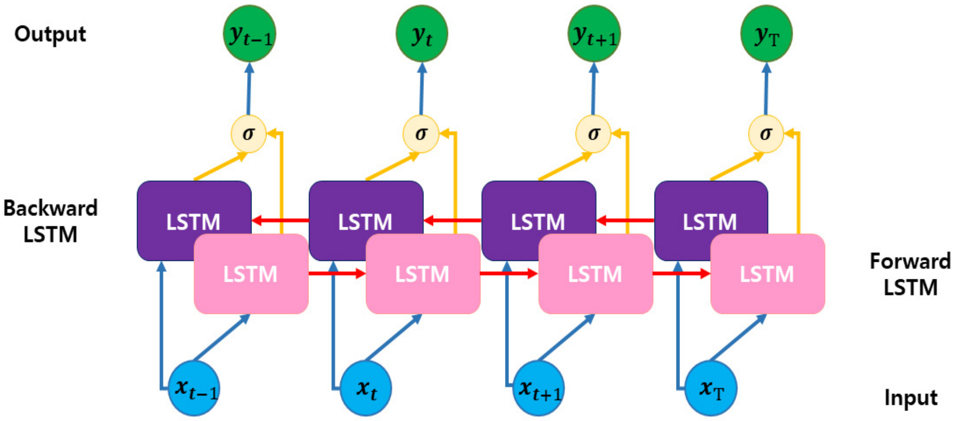


Figure 5. Bi-LSTM architecture.

2.2.2. Two-Dimensional-Based CNN-Based Transfer Learning Model

In this study, two-dimensional spectrogram images were trained using VGGish and YAMNet among CNN-based models pretrained through audio data. VGGish is a CNN-based neural network proposed by Hershey [18] designed to classify audio classes by training the deep learning neural networks using audio data contained in large-scale video databases. The deep learning network is trained using audio content consisting of more than 2 million YouTube videos, including 527 audio classes. Adult voices, baby babbling, and animal sounds are included in the 527 audio classes [19]. This model is constructed based on the VGG network, one of the CNN-based pretrained models frequently used in the field of computer vision. VGGish consists of four convolution blocks and receives a spectrogram comprising an audio clip with a size of 96 × 64 × 1 as input. Each block includes a 2D-based convolution layer that performs the feature extractor role, ReLU activation function, and max pooling layer, which reduces the image dimensions while maintaining the image features. Two fully connected layers, embedding layers, and regression output layers, that function as a classifier, are included behind the four convolution blocks.

The YAMNet model is an audio detection model trained using an audio set with 521 classes, such as laughter, dog barking, or sirens, contained in more than 2 million video databases. The YAMNet model was proposed by Ellis and Chowdhry and is a computationally efficient model for the problem of classifying audio events. The VGGish model has a complicated computation because it involves over 72 million parameters. Conversely, the YAMNet model uses only 4.7 million parameters, thereby being computationally more efficient than the VGGish model. A lightweight model was designed using a convolution kernel capable of separating by depth to be used in the computer vision field; this model was designed based on the proposed MobileNet architecture in the study of Andrew [20]. The YAMNet model consists of 14 convolution layer blocks, and in this study, the pretrained model was modified by reassigning the number of classes by adding a fully connected layer after the last convolution layer and then replacing the classification layer.

2.3. Proposed Deep Learning Two-Stream-Based Emotion Recognition Model

Figure 6 illustrates the proposed architecture of the Bi-LSTM and CNN two-stream-based emotion recognition model using the Korean speech data. As shown in the figure, the softmax value is obtained by extracting 1D spectrum features using MFCCs and GTCCs and then training based on the Bi-LSTM model. In addition, two-dimensional spectrogram images such as Bark spectrogram, ERB spectrogram, and log-mel spectrogram are obtained on the basis of time–frequency conversion and then trained using CNN-based pretrained learning models, such as VGGish and YAMNet, to obtain the softmax value. The softmax probability values output from the two models are either added or multiplied using the late score fusion method to obtain the final classification value, and eight emotions are then classified based on the two-stream model.

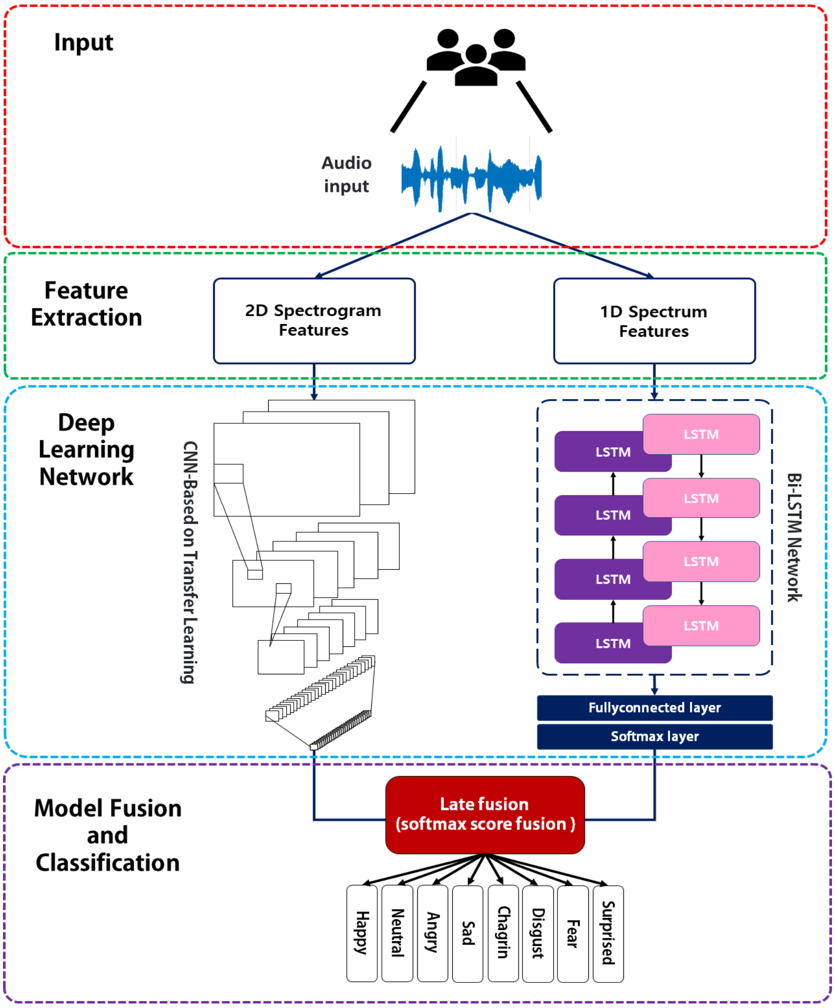


Figure 6. Structure of the Bi-LSTM and CNN two-stream-based SER model.

The Bi-LSTM network used for 1D-spectrum learning used four Bi-LSTM layers, and each layer had 100, 100, 50, and 30 hidden layers. In addition, the dropout layers (with probabilities of 0.5, 0.5, 0.3, and 0.3) were added in between to prevent overfitting during training. In the CNN-based transfer learning model, the number of classes was changed to eight in the fully connected layer for classifying eight emotions, and the existing classification layer was changed to a new classification layer.

3. Results

3.1. Korean Speech Emotion Database

The Korean speech emotion database built by Chosun University was used in this study. This database was built from 200 subjects and includes recorded speech files of eight emotions: happy, neutral, angry, sad, chagrin, disgust, fear, and surprised. The participants had a Sony ECM-CS3 stereo pin microphone attached to their collar and were recording in an environment that was maintained as quiet as possible without noise, as voice data are sensitive to noise. The data were collected by instructing the participants to act out short scripts appropriate for each type of emotion. The audio files were recorded at 48,000 Hz in a .wav file format. This database contains 10 files for each emotion type, which is 80 files per participant; therefore, 2000 files were collected for each emotion type, which is a total of 16,000 audio data files. Figure 7 visualizes the recorded speech file for the eight emotions of the 17th participant.

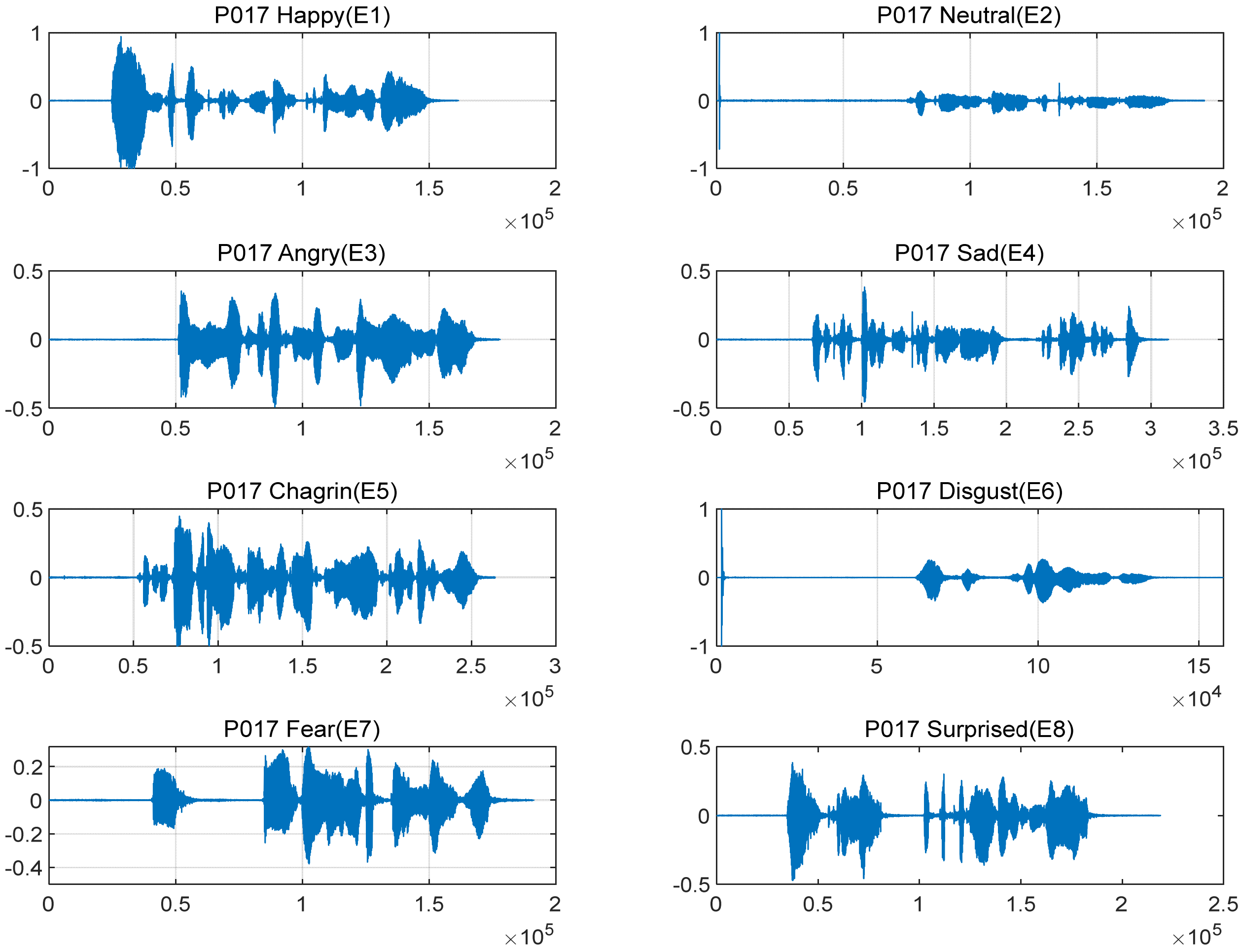


Figure 7. Visualization of speech data for eight emotions.

3.2. Results and Performance Analysis

This study comparatively analyzed the performance of a two-stream-based SER model using the Bi-LSTM and CNN two-stream-based transfer learning model, as well as various other deep learning models, by extracting multidimensional features of speech data. Table 1 overviews the experimental environment. The first experiment was conducted for emotion recognition based on the LSTM and Bi-LSTM using the spectrum features of the Korean speech data. Three different feature extraction methods were used for the LSTM and Bi-LSTM models to compare the emotion recognition performance. Table 2 lists the learning parameters used for the models.

Table 1. Experimental environment.

Division		Use
Hardware	CPU	Intel Core i9 10900K @ 3.70 GHz
	GPU	NVIDIA GeForce RTX 2080 SUPER
	RAM	128 GB
Software	OS	Windows 10
	Programming Language	Matlab 2022b

Table 2. Learning parameters of LSTM and Bi-LSTM.

Training Options Parameter	Optimization Function	Gradient Threshold	Mini-Batch Size	Epoch
Parameter values	Adam	2	300	50

Table 3 compares the performances of the LSTM and Bi-LSTM according to the spectrum feature extraction method. Three cases of feature extraction were separately conducted: MFCC, GTCC, and both MFCC and GTCC. When LSTM and Bi-LSTM are compared, the overall performance was higher for the Bi-LSTM. As shown in Table 3, the best performance of 90.38% resulted when learning was performed by the Bi-LSTM model and the features were extracted from both the MFCC and GTCC spectrum. Figure 8 shows the relevant confusion matrix, in which emotions are fairly accurately classified at an overall high probability; however, fear has a relatively lower classification accuracy compared with the other emotion types.

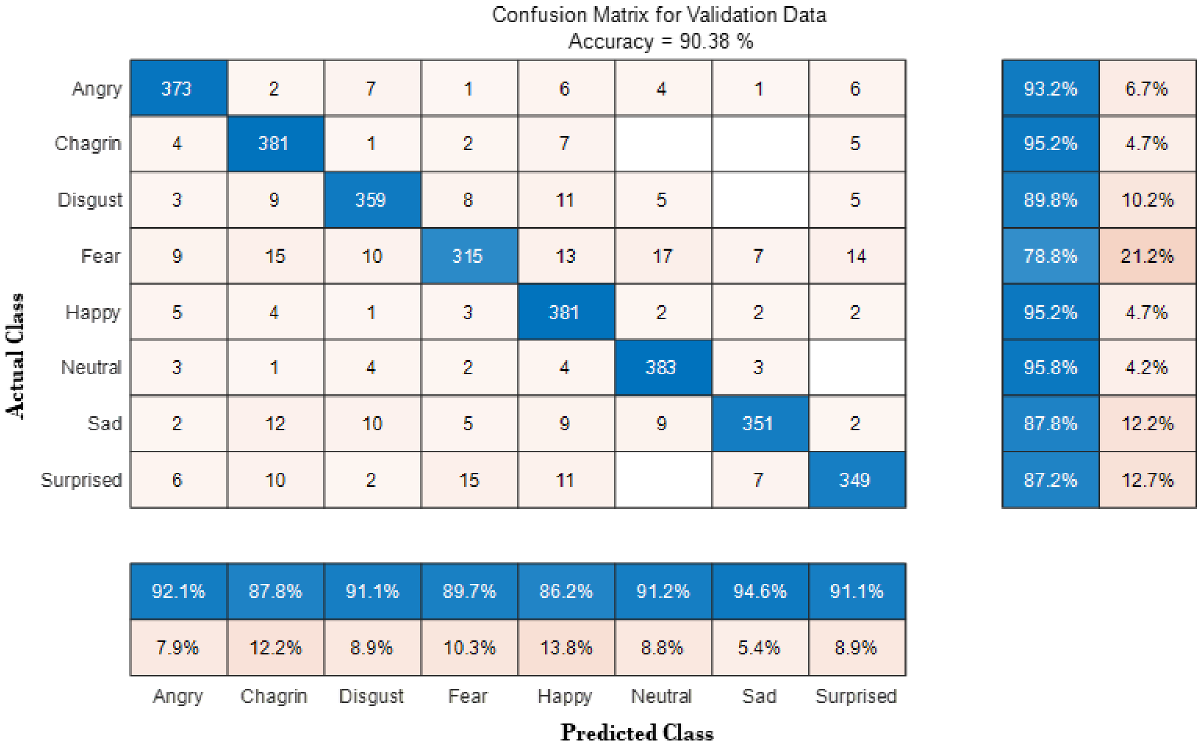
Table 3. Accuracy of 1D spectrum feature-based LSTM and Bi-LSTM models.

Deep Learning Model	Feature Extraction Method	Accuracy
LSTM	MFCC	87.22%
	GTCC	86.28%
	MFCC + GTCC	88.47%
Bi-LSTM	MFCC	89.59%
	GTCC	87.72%
	MFCC + GTCC	90.38%

The second experiment examined the emotion recognition performance of the CNN-based transfer learning model using the time–frequency conversion-based 2D-spectrogram features of the Korean speech data; Table 4 presents the learning parameters of this model.

Table 4. Learning parameters of CNN transfer learning model.

Training Options Parameter	Optimization Function	Gradient Threshold	Mini-Batch Size	Epoch
Parameter values	Adam	1	512	5



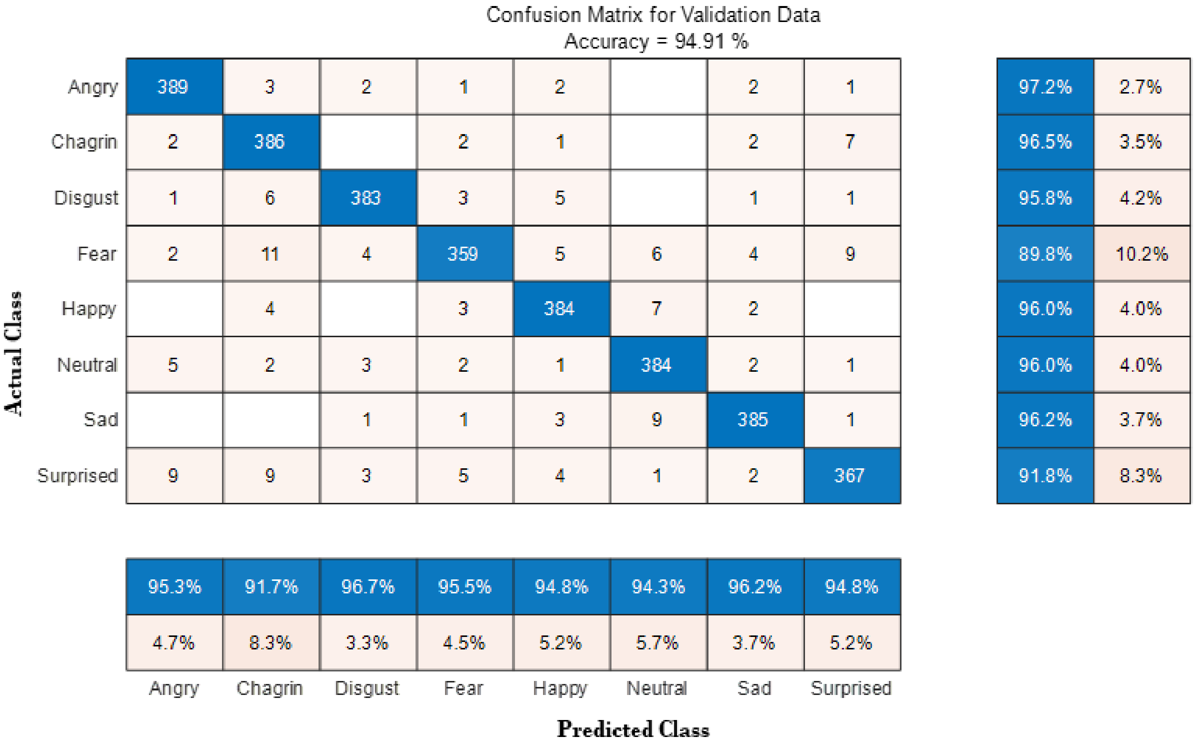
	Angry	Chagrin	Disgust	Fear	Happy	Neutral	Sad	Surprised		
Angry	373	2	7	1	6	4	1	6	93.2%	6.7%
Chagrin	4	381	1	2	7			5	95.2%	4.7%
Disgust	3	9	359	8	11	5		5	89.8%	10.2%
Fear	9	15	10	315	13	17	7	14	78.8%	21.2%
Happy	5	4	1	3	381	2	2	2	95.2%	4.7%
Neutral	3	1	4	2	4	383	3		95.8%	4.2%
Sad	2	12	10	5	9	9	351	2	87.8%	12.2%
Surprised	6	10	2	15	11		7	349	87.2%	12.7%
	92.1%	87.8%	91.1%	89.7%	86.2%	91.2%	94.6%	91.1%		
	7.9%	12.2%	8.9%	10.3%	13.8%	8.8%	5.4%	8.9%		

Figure 8. Confusion matrix of Bi-LSTM SER model.

Table 5 presents the performance of VGGish and YAMNet, among the CNN-based transfer learning models, according to the spectrogram feature extraction method. The features were extracted for the Bark, ERB, and log-mel spectrograms for the experiment. Here, the performance of the YAMNet model is overall more outstanding compared with the VGGish model. In this experiment, the best performance of 94.91% resulted when ERB spectrogram features were extracted and learning was performed based on the YAMNet model. Figure 9 shows the relevant confusion matrix. Similar to the CNN-based model, each type of emotion was classified with an overall high probability of 90% or higher; however, fear has a relatively lower classification accuracy than the other types of emotion.

Table 5. Accuracy of CNN transfer learning model based on 2D spectrogram features.

Deep Learning Model	Feature Extraction Method	Accuracy
VGGish	Bark Spectrogram	89.19%
	ERB Spectrogram	88.53%
	Log-mel Spectrogram	92.31%
YAMNet	Bark Spectrogram	93.47%
	ERB Spectrogram	94.91%
	Log-mel Spectrogram	93.66%



	Angry	Chagrin	Disgust	Fear	Happy	Neutral	Sad	Surprised		
Angry	389	3	2	1	2		2	1	97.2%	2.7%
Chagrin	2	386		2	1		2	7	96.5%	3.5%
Disgust	1	6	383	3	5		1	1	95.8%	4.2%
Fear	2	11	4	359	5	6	4	9	89.8%	10.2%
Happy		4		3	384	7	2		96.0%	4.0%
Neutral	5	2	3	2	1	384	2	1	96.0%	4.0%
Sad			1	1	3	9	385	1	96.2%	3.7%
Surprised	9	9	3	5	4	1	2	367	91.8%	8.3%
	95.3%	91.7%	96.7%	95.5%	94.8%	94.3%	96.2%	94.8%		
	4.7%	8.3%	3.3%	4.5%	5.2%	5.7%	3.7%	5.2%		

Figure 9. Confusion matrix of YAMNet SER model.

Finally, the experiment compared the performance of the two-stream ensemble-based SER model. Table 6 presents the performances of the two-stream based emotion recognition models measured by two different late score fusion methods according to the feature extraction method; Figure 10 visualizes the performance. The two-stream model's performance was compared by selecting the LSTM, Bi-LSTM, and YAMNet models that demonstrated the best performances in previous experiments. As shown in Table 6 and Figure 10, the proposed Bi-LSTM and CNN two-stream-based model outperformed the single model. In particular, the highest accuracy of 96.00% was achieved when emotion recognition was performed based on the two-stream model by extracting 1D features for both the 2D MFCC and GTCC features for the ERB spectrogram and then inputting them into the Bi-LSTM and YAMNet models, respectively, and multiplying the last softmax value.

Figure 11 shows the relevant confusion matrix. The classification accuracy for fear remains insufficient; however, all other types of emotion were classified with a fairly high probability overall. This result indicates that the classification performance improves when SER is performed based on the two streams through an ensemble of two models because the performance improved by at least 1.09% and up to 5.62% compared with single models in previous studies.

Table 6. Accuracy of two-stream-based SER models.

Deep Learning Model	Feature Extraction Method	Late Fusion	Accuracy
LSTM + YAMNet	1D: MFCC + GTCC 2D: ERB Spectrogram	sum	95.69%
		product	95.34%
Bi-LSTM + YAMNet	1D: MFCC + GTCC 2D: ERB Spectrogram	sum	95.94%
		product	96.00%

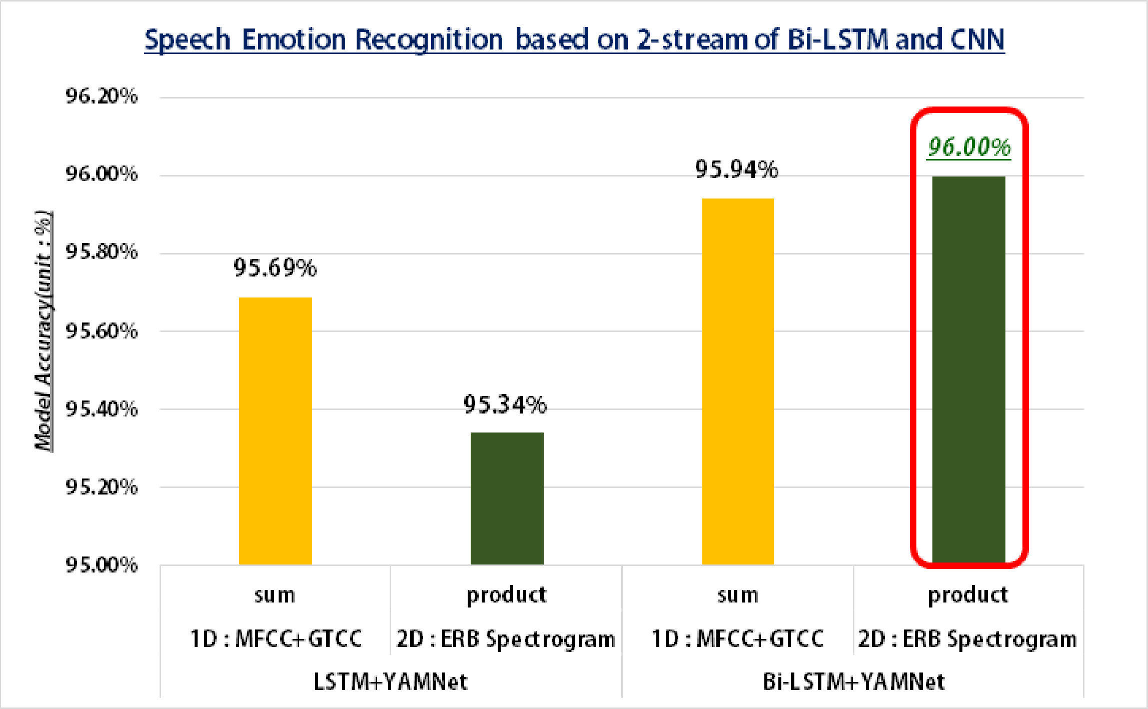
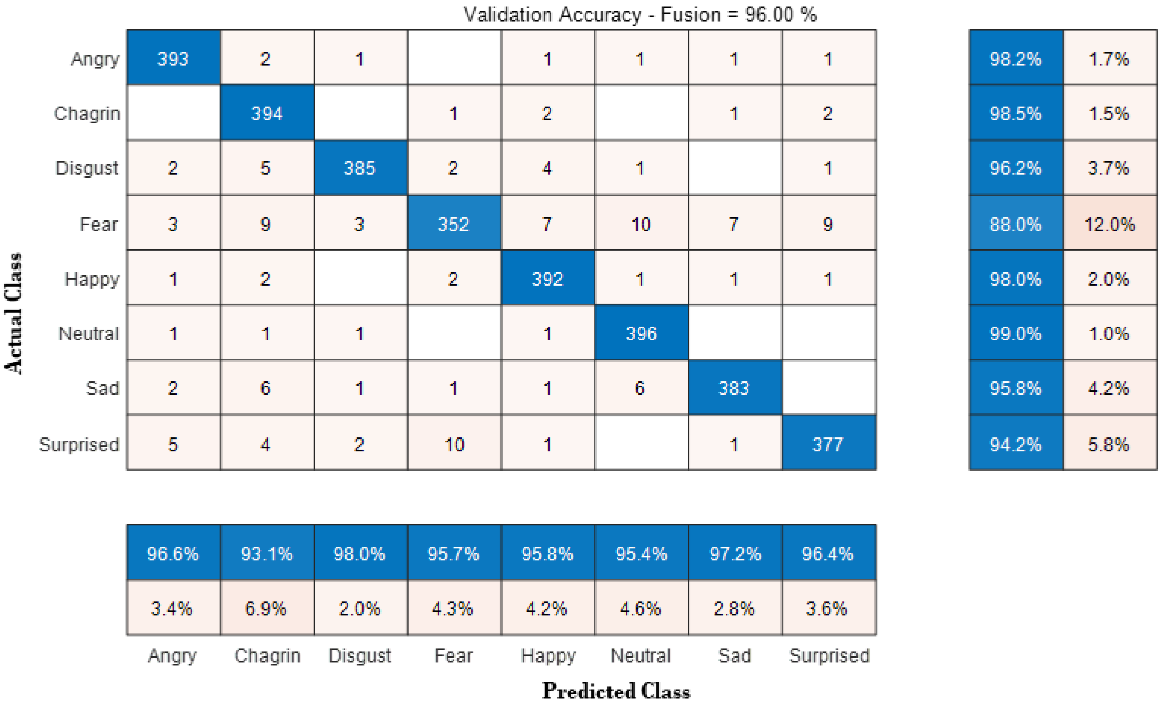


Figure 10. Accuracy of Bi-LSTM and CNN two-stream-based SER model.



	Angry	Chagrin	Disgust	Fear	Happy	Neutral	Sad	Surprised		
Angry	393	2	1		1	1	1	1	98.2%	1.7%
Chagrin		394		1	2		1	2	98.5%	1.5%
Disgust	2	5	385	2	4	1		1	96.2%	3.7%
Fear	3	9	3	352	7	10	7	9	88.0%	12.0%
Happy	1	2		2	392	1	1	1	98.0%	2.0%
Neutral	1	1	1		1	396			99.0%	1.0%
Sad	2	6	1	1	1	6	383		95.8%	4.2%
Surprised	5	4	2	10	1		1	377	94.2%	5.8%
	96.6%	93.1%	98.0%	95.7%	95.8%	95.4%	97.2%	96.4%		
	3.4%	6.9%	2.0%	4.3%	4.2%	4.6%	2.8%	3.6%		

Figure 11. Confusion matrix of Bi-LSTM and CNN two-stream-based SER model.

4. Discussion

This study aimed to analyze the information of emotions by extracting multidimensional features from speech data and design a deep learning model capable of effectively identifying the emotional state of a person. In previous studies, speech emotion recognition was conducted using one of the 1D spectrum features and time-frequency-based 2D spectrogram features of audio data, but in this study, we confirmed that the proposed model is significant for speech emotion recognition by achieving 96% accuracy as two deep learning models are fused using 1D and 2D features. Consequently, the two-stream deep learning model used to classify eight types of emotion was proven to be more effective in identifying the emotional states of a person than a single model. However, fear had a relatively lower classification accuracy compared with the other seven types of emotion, which is suggested to be due to limitations in how to express emotions in detail because a person expresses emotions directly. Therefore, further research is required on methods that can more objectively and accurately identify emotional states from speech data.

5. Conclusions

This study designed a two-stream-based emotion recognition model based on a Bi-LSTM and CNN-based transfer learning model using the Korean speech emotion database, and the emotion recognition performance was compared. For the experiment, the Korean speech emotion database containing eight types of emotions—happy, neutral, angry, sad, chagrin, disgust, and fear—was built by 200 participants from Chosun University. Various experiments confirmed that the performance improves by at least 1.09% and up to 5.62% when emotion recognition is performed by the proposed Bi-LSTM and CNN-based two-stream ensemble transfer learning model compared with a single model. Accordingly, the proposed two-stream based ensemble model is more effective for emotion recognition using speech data. In the future, further research should be conducted on methods that can more objectively and accurately identifying emotional states from speech data, as well as on methods for recognizing emotions using multimodal data, including both speech and text data.

Author Contributions: Conceptualization, A.-H.J. and K.-C.K.; methodology, A.-H.J. and K.-C.K.; software, A.-H.J. and K.-C.K.; validation, A.-H.J. and K.-C.K.; formal analysis, A.-H.J. and K.-C.K.; investigation, A.-H.J.; resources, K.-C.K.; data curation, K.-C.K.; writing—original draft preparation, A.-H.J.; writing—review and editing, K.-C.K.; visualization, A.-H.J. and K.-C.K.; supervision, K.-C.K.; project administration, K.-C.K.; funding acquisition, K.-C.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Healthcare AI Convergence R&D Program through the National IT Industry Promotion Agency of Korea (NIPA) funded by the Ministry of Science and ICT (No. S0254-22-1001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Akçay, M.B.; Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **2020**, *116*, 56–76. [CrossRef]
2. An, X.D.; Ruan, Z. Speech Emotion Recognition algorithm based on deep learning algorithm fusion of temporal and spatial features. *J. Phys. Conf. Ser.* **2021**, *1861*, 1–6.
3. Kipyatkova, I. LSTM-Based Language Models for Very Large Vocabulary Continuous Russian Speech Recognition System. In Proceedings of the SPECOM 2019: Speech and Computer, Istanbul, Turkey, 20–25 August 2019; Volume 11658, pp. 219–226.
4. Basu, S.; Chakraborty, J.; Aftabuddin, M. Emotion recognition from speech using convolutional neural network with recurrent neural network architecture. In Proceedings of the 2017 2nd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 19–20 October 2017; pp. 333–336.
5. Peng, Z.; Li, X.; Zhu, Z.; Unoki, M.; Dang, J.; Akagi, M. Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends. *IEEE Access* **2020**, *8*, 16560–16572. [CrossRef]
6. Bhosale, S.; Chakraborty, R.; Kopparapu, S.K. Deep encoded linguistic and acoustic cues for attention based end to end speech emotion recognition. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7189–7193.

7. Liu, G.; He, W.; Jin, B. Feature Fusion of Speech Emotion Recognition based on Deep Learning. In Proceedings of the 2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC), Guiyang, China, 22–24 August 2018; pp. 193–197.
8. Zayene, B.; Jlassi, C.; Arous, N. 3D Convolutional Recurrent Global Neural Network for Speech Emotion Recognition. In Proceedings of the 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia, 2–5 September 2020; pp. 1–5.
9. Zhang, C.; Xue, L. Two-stream Emotion-embedded Autoencoder for Speech Emotion Recognition. In Proceedings of the 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, Canada, 21–24 April 2021; pp. 1–6.
10. Han, S.; Leng, F.; Jin, Z. Speech Emotion Recognition with a ResNet-CNN-Transformer Parallel Neural Network. In Proceedings of the 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), Xiamen, China, 28–30 July 2021; pp. 803–807.
11. Kakuba, S.; Han, D.S. Speech Emotion Recognition using Context-Aware Dilated Convolution Network. In Proceedings of the 2022 27th Asia Pacific Conference on Communications (APCC), Jeju Island, Republic of Korea, 19–21 October 2022; pp. 601–604.
12. Chu, S.; Narayanan, S.; Kuo, C.C.J. Environmental sound recognition with time-frequency audio features. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 1142–1158. [CrossRef]
13. Rawat, A.; Mishra, P.K. Emotion Recognition through Speech Using Neural Network. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **2015**, *5*, 422–428.
14. Liu, J.M.; You, M.; Li, G.Z.; Wang, Z.; Xu, X.; Qiu, Z.; Xie, W.; An, C.; Chen, S. Cough signal recognition with Gammatone Cepstral Coefficients. In Proceedings of the 2013 IEEE China Summit and International Conference on Signal and Information Processing, Beijing, China, 6–10 July 2013; pp. 160–164.
15. Zwicker, E. Subdivision of the audible frequency range into critical bands. *J. Acoust. Soc. Am.* **1961**, *33*, 248. [CrossRef]
16. Torben, P. Acoustic Communication. Hearing and Speech. Version 2.0. In *31230 Acoustic Communication*; Online Research Database in Technology: Lyngby, Denmark, 2005; pp. 1–94.
17. Hochreiter, S.; Schmidhuber, J. Long Short-term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
18. Hershey, S.; Chaudhuri, S.; Ellis, D.P.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN architectures for large-scale audio classification. In Proceedings of the 2017 IEEE International Conference on Acoustics Speech and Signal Processing, New Orleans, LA, USA, 5–9 March 2017; pp. 131–135.
19. Syed, Z.S.; Memon, S.A.; Memon, A.L. Deep acoustic embeddings for identifying parkinsonian speech. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 726–734. [CrossRef]
20. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.