Hindawi
Mathematical Problems in Engineering
Volume 2023, Article ID 8837465, 16 pages
https://doi.org/10.1155/2023/8837465

Research Article

Feature Aggregation with Two-Layer Ensemble Framework for Multilingual Speech Emotion Recognition

Sangho Ough, Sejong Pyo, and Taeyong Kim

The Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University, Seoul 06974, Republic of Korea

Correspondence should be addressed to Taeyong Kim; kimty@cau.ac.kr

Received 8 March 2023; Revised 25 October 2023; Accepted 2 November 2023; Published 11 December 2023

Academic Editor: Jelena Nikolić

In this study, we present a framework for improving the accuracy of speech emotion recognition in a multilingual environment. In our prior experiments, where machine learning (ML) models were trained to predict emotions in Korean and then tested in English, as well as vice versa, we observed a dependency on language in emotion recognition, resulting in poor accuracy. We suspect that this may be related to the spectral differences in certain emotions between Korean and English and to the tendency for different formant values to have different acoustic frequencies. For this study, we investigated several different methods, including models with mixed databases, a single database, and bagging, boosting, and voting ML algorithms. Finally, we developed a framework consisting of two branches: one for the aggregation of high-dimensional features from multilingual data and one for a two-layered ensemble framework for emotion classification. In the ensemble framework for Korean and English (EF-KEN), features are extracted and ensemble models are trained, boosted, and evaluated by applying them to different spoken languages (English and Korean). The final experimental result demonstrates a meaningful improvement in an environment with two different languages.

1. Introduction

With the expansion of the global economy and the gradual end of the coronavirus disease 2019 (COVID-19) pandemic, worldwide mobility is once again on the rise. Fueled by the growth of Korean culture, tourism in South Korea continues to thrive, attracting increasing numbers of foreign visitors who are also showing a growing interest in the Korean language. Additionally, the number of South Korean travelers, immigrants, and international students heading toward English-speaking countries is steadily increasing.

In this global everyday-life context, the role of voice emotion recognition technology is becoming increasingly important. Being able to detect emotions while considering the speaker's culture and language can enhance mutual understanding and communication. This can help identify emotional cues that nonnative speakers might find challenging to express abroad, thus bridging the gap between language and culture. Moreover, detecting emotions in foreigners can enhance communication and human interaction in various fields, such as airport services, telephone guidance, and online education. Multilingual speech emotion recognition (SER) technology can overcome language barriers and facilitate effective communication in diverse societies.

Against this backdrop, this study aims to improve emotion recognition rates in English or Korean spoken by nonnative speakers. The expected outcome of this research is to aid nonnative speakers in their daily lives abroad, assisting them in communication and cultural adaptation. The goal is for SER technology to become the cornerstone for facilitating mutual understanding and communication among people from diverse linguistic and cultural backgrounds, thus serving as a universal service enhancer.

The contribution of this research is twofold: addressing crucial challenges in crosslingual emotion recognition and effectively countering emotion class imbalance:

(1) Improving emotion recognition in crosslingual environments: Korean and English exhibit distinct cultural and linguistic differences that result in varied ways of expressing emotions. Traditional voice emotion recognition models often fail to account for these disparities. To address this, an ensemble framework is proposed, which involves combining Korean

and English datasets to extract acoustic features and then developing models that consider the characteristics of each language. This framework aims to enhance emotion recognition performance in crosslingual environments.

(2) Resolving emotion class imbalance: The Interactive Emotional Motion Capture (IEMOCAP) dataset [1] suffers from an imbalance issue where certain emotion classes have fewer data instances compared to others. This imbalance can limit the performance of conventional voice emotion recognition models. The ensemble framework seeks to mitigate emotion class imbalance by using diverse models and data combinations. Thus, the proposed ensemble technique aims to balance the training of each class's data, thereby ameliorating the imbalance issue in the number of samples for each emotion class.

In summary, our research's notable contributions lie in its pioneering solution to enhance emotion recognition across different languages by harnessing an ensemble framework to tackle crosslingual disparities and its novel approach to alleviate emotion class imbalance concerns, ultimately advancing the field of crosslingual emotion recognition.

We focus on phonetic information to apply this model to multiple languages. We carry out experiments with the framework in a bilingual environment. The use of databases for both languages poses a formidable academic challenge because of the differences in the nature of phonograms in word orders. In English sentences, the subject is followed by the verb and then an object, whereas, in Korean, the subject is followed by an object and then the verb. We develop and experiment with a number of machine learning (ML) algorithms and ensemble approaches to see how different combinations of databases affect the model. In Section 2, related research on SER is reviewed. Section 3 describes the database that we used for both Korean and English, including preprocessing and the analysis of acoustic feature extraction. We introduce the framework for building high-dimensional features for multiple languages, namely high-dimensional feature mapping (HDFM). Section 4 describes a two-layer classification model for the HDFM called ensemble framework for Korean and English (EF-KEN). Section 5 provides the experimental results for comparisons of the types of databases and levels of the classification model.

2. Related Work on SER

Various techniques using emotion databases and artificial intelligence are employed to detect human emotions in speech.

(1) Acoustic analysis: Acoustic features, such as pitch, intensity, and duration, are analyzed to detect emotional cues in speech. For instance, high pitch, increased intensity, and prolonged duration can be associated with excitement or anger [2, 3].

(2) Language analysis: Words and phrases used in speech are analyzed to detect emotional content. Specific words and phrases, like "happy," "joyful," and "ecstatic," can indicate happiness [4, 5].

(3) Prosody analysis: Variations in pitch, intensity, and tempo are analyzed to detect emotional cues. For example, a rising pitch at the end of a sentence can imply a question or uncertainty [3, 6].

(4) Deep learning: Large datasets are analyzed and learned using artificial neural networks to identify speech patterns associated with specific emotions [7, 8].

Table 1: Classification of SER research: neural networks, feature representation, and multimodal.

Neural network	Features representation	Multimodal
CNN	Spectrogram	Speech + text
RNN	Numeric value	Speech + video
CNN + RNN	Spectrogram + numeric value	Speech + text + video

We classify research works on SER into three categories: neural network-based work, feature representation-based work, and multiple modality-based work. The categories are summarized in Table 1.

As the convolutional neural network (CNN) has contributed to research on image classification and regression, models have been used effectively to classify emotions by imaging voice signals through preprocessing [9, 10]. In order to learn voice emotion data using a CNN, it is necessary to image the characteristics of voice data [8, 11, 12]. One of the features of audio data is its spectral features. Learning emotions using the spectral features of voice has proved effective in previous studies [9, 13]. In this study, almost 200 high-dimensional acoustic features need to be converted to graphical images to be classified. However, this requires a significant amount of computing power for CNN to process such large amounts of data.

The long short-term memory (LSTM) is a recurrent neural network (RNN) [9, 14] learning model for solving the long-term dependency of RNN. LSTM can remember and connect information from the past to the present. Each unit has three gates: an input gate to learn what information is to be stored in memory, a forget gate to learn how long information is stored, and an output gate to learn when the stored information can be used [9]. The SER system receives the voice signal as input and preprocesses the data, and then the processed data enter the LSTM layer. It then connects all the nodes of the previous layer in the connected layer and outputs the resulting value through the softmax function [15–18].

The performance of voice-based emotion recognition is not satisfiable when an algorithm is implemented with one deep learning model. Therefore, in most cases, algorithms are created by connecting two or more deep learning models [19, 20]. The information gathered from speech and text has been developed into a methodology for a multimodal emotion recognition model with its speech features and text embeddings. Spectrograms generated from the voice signals are input to the CNN, which is integrated with an RNN

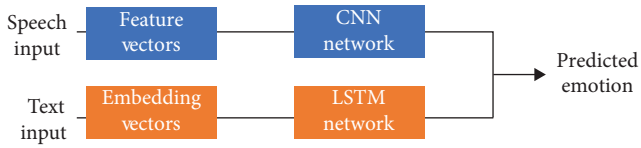

FIGURE 1: SER using feature vectors and embedding vectors with CNN and LSTM.

method for recognizing emotions using data extracted from content information in text format, as illustrated in Figure 1 [21]. In this study, CNN is not included because it requires excessive computing power to process large amounts of graphic data in a spectrogram. The RNN is also not included because it requires a text-processing model that could be applied universally across multiple languages. However, only a limited number of databases contain text information for LSTM multimodal systems. Hence, this study focuses solely on acoustic information. With this approach, we can easily expand our framework to other languages.

The main contributions of this study are (1) the introduction of a novel end-to-end multilingual framework for SER, (2) the creation of a methodology for extracting acoustic features from two different corpora and combining them to form a single training dataset, and (3) the development of a two-layered ensemble framework to improve the accuracy of emotion recognition in speech.

3. High-Dimensional Features for Multiple Languages

The research focuses on SER using ensemble techniques in both Korean and English environments. As shown in Figure 2, the proposed EF-KEN is structured with two main layers. The first layer, known as HDFM, involves the extraction and synthesis of high-dimensional Korean and English acoustic features. The second layer connects the preclassifiers of the ML algorithms and the ensemble voting (EV) metaclassifier [23–25]. When connecting HDFM and EF-KEN, the training is performed on a combined dataset containing both English and Korean data, whereas the testing is conducted separately for each language [26–28].

We chose the emotion databases in English and Korean. Each database includes the characteristics and composition of the language dataset and the composition of the acoustic features for emotion recognition.

For the English data, we use voice-only waveform audion format (WAV) files from the IEMOCAP database developed by the University of Southern California. This database was designed for the collaborative analysis of speech and gestures [1]. It consists of 12 hr of audio and video data in English and consists of video, voice, text, and movement detection signals of the face, head, and hands. This includes a file recorded by 10 actors with a total of 10 emotions, such as happiness, anger, sadness, frustration, and neutral. There are five men and five female actors, and the database consists of data from five sessions recorded with one man and one woman. Regarding the Korean data, WAV files were collected by volunteers who naturally communicated with the internet application for a certain period of time using an emotional conversation application and were labeled with seven emotions (*happiness*, *anger*, *disgust*, *fear*, *sadness*, *surprise*, and *neutral*) by the Korea Electronics Technology Institute (KETI).

IEMOCAP is a well-established and widely used dataset for emotion recognition in English speech. It encompasses diverse emotional expressions and captures real-world scenarios, making it a reliable benchmark for English emotion recognition models. Likewise, the KETI dataset is a prominent resource for Korean SER, specifically tailored to capture the nuances of emotions expressed in the Korean language. Thus, we chose to use these language-specific datasets in our experimental design because they enabled us to capture the distinct cultural and linguistic characteristics that influence emotional expression in each language.

Furthermore, by using language-specific datasets, we ensure that our models are optimized to recognize emotions accurately within the linguistic and cultural contexts of each language. This approach enhances the generalization ability of our models when deployed in real-world scenarios where emotional expressions may differ considerably between languages. Leveraging language-specific datasets also enables us to tailor the model's architecture and hyperparameters according to the unique characteristics of each language, ultimately leading to improved performance.

3.1. Extraction of Acoustic Data. In our research, we aimed to extract as many acoustic features as possible from WAV files. We obtained 200 acoustic features and normalized them to values between 0 and 1. Some of the important features we extracted include the zero crossing rate (ZCR), the Mel frequency cepstral coefficient (MFCC), and chroma, which contain important frequency information [29, 30].

As humans can only perceive frequencies on a logarithmic scale, a Mel scale is used to represent perceptually relevant frequencies and amplitudes. A distance on the Mel scale represents the same perceptual distance. The frequency content of audio signals in speech and audio processing was obtained by converting the Mel scale value, m, into frequency, f, through Equations (1) and (2), where m is a dimensionless value corresponding to a linear frequency on the Mel scale:

$$m = 2,595 \log\left(1 + \frac{f}{500}\right), \tag{1}$$

$$f = 700\left(10^{m/2,595} - 1\right). \tag{2}$$

After the voice signal is converted to a Mel scale value, it is ready to acquire the MFCC by Fourier transform. In order to calculate the value of the MFCC, the human voice is divided into 25 ms frames, and Fourier transform is applied to each frame to extract the frequency information. The results of the Fourier conversion in each frame are called the Mel spectrum. The Mel spectrum is obtained by applying the Mel filter bank, which is sensitive to human speech recognition. The logarithmic Mel spectrum is called the log-Mel spectrum. The MFCC is obtained from the conversion of the frequency domain information into the time domain by

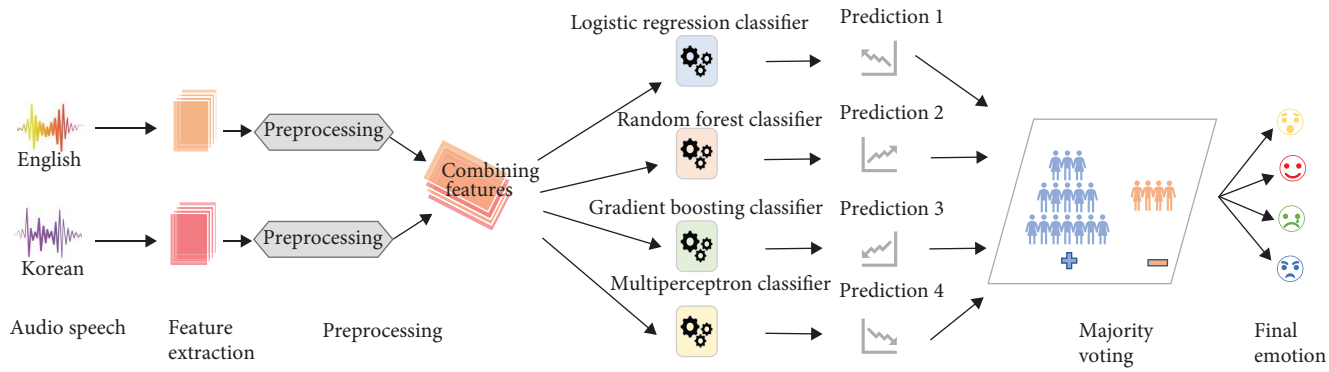

Figure 2: Graphical representation of proposed two-layered EF-KEN [22].

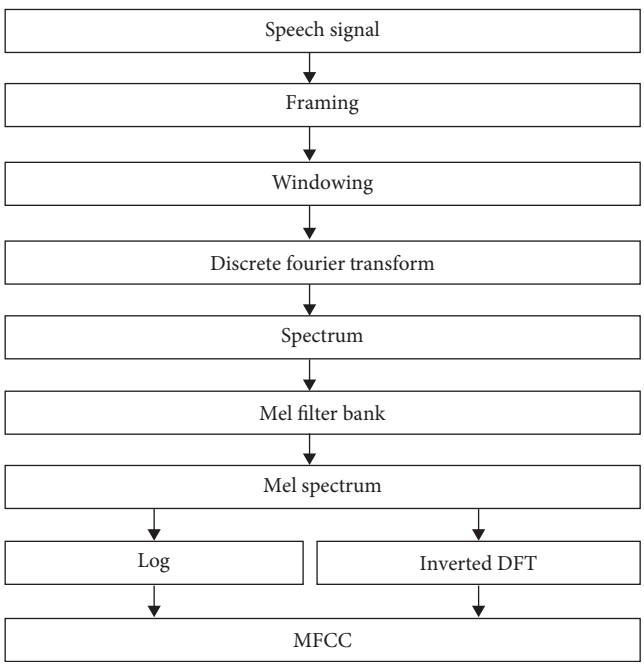

Figure 3: Process of extracting MFCC using Fourier transform.

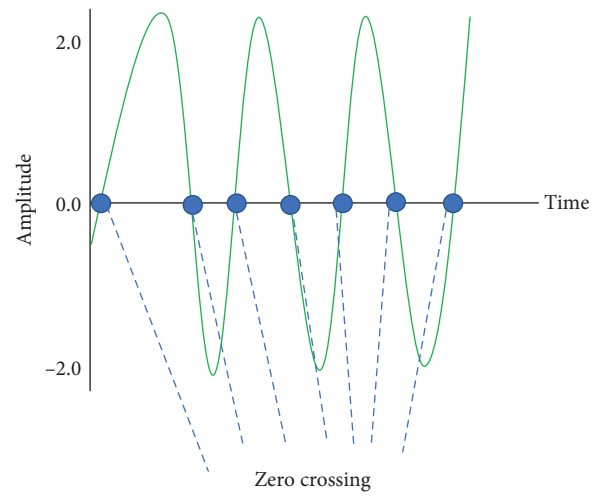

Figure 4: Method of calculating ZCR from a signal wave.

applying the inverse Fourier transform to the log-Mel spectrum. MFCC is also used as input to the Gaussian mixture model in the existing voice recognition system [30]. The mean of the MFCC features is calculated, and then short-time Fourier transform and Mel spectrogram features are obtained by setting the sampling rate for the audio files, and the number of MFCCs is set to 12 [27]. This entire process of extracting MFCCs is shown in Figure 3. In this study, the Librosa, Pandas, and NumPy libraries are used to perform feature extraction.

Other crucial features for emotion recognition are the chromagram and ZCR. The chroma features represent 12 pitch levels, including C, C#, D, D#, E, F, F#, G, G#, A, A#, and B. Chroma features are intended to represent the harmonic content of a short-lived sound window. Chroma features can show a high degree of robustness to changes in timbre. The number of chroma features is set to 12, the same as the pitch levels [31]. The ZCR shown in Figure 4 represents the number of times a voice signal from the human vocal tract crosses the horizontal axis [29].

In Equation (3), ZCR_t represents the ZCR at a specific time frame t. The variable t denotes the time frame or sample index for the calculation. K signifies the total number of samples or time frames and sets the upper limit for the summation. The summation, denoted by $\sum$, ranges from index k equal to $t{\cdot}K$ to $(t+1){\cdot}K-1$ representing the sum over a range of samples spanning t and the subsequent time frame $(t+1)$. The $\mathrm{sgn}(s(k))$ and $\mathrm{sgn}(s(k+1))$ are sign functions applied to the signal values at indices k and $k+1$, respectively. These functions return -1 for negative values, 0 for zero, and 1 for positive values. The $s(k)$ and $s(k+1)$ represent the values of the signal at the respective indices. The ZCR_t quantifies the frequency of zero crossings within the specified time frame, providing valuable information about the signal's waveform characteristics.

$$\mathrm{ZCR}_t = \frac{1}{2} \cdot \sum_{k=t\cdot K}^{(t+1)\cdot K-1} |\mathrm{sgn}(s(k)) - \mathrm{sgn}(s(k+1))|. \tag{3}$$

3.2. Preprocessing and Combining Feature Sets. These large numbers of features from IEMOCAP and KETI were preprocessed to equalize the number of emotions and the total number of each sample. After that, they were combined into one high-dimensional feature set called the HDFM. As shown in

Table 2: Comparison of original sample count and preprocessed sample count with extracted features.

Corpus	Language	Number of original samples	Number of preprocessed samples	Number of features
IEMOCAP	English	10,039	8,000	200
KETI	Korean	19,374	8,000	200

Table 2, there were 10,039 WAV files in English and 19,374 WAV files in Korean. Both databases were reduced to 8,000 random samples, and 200 voice features were extracted from each sample. We preprocessed the dataset in order to adjust the number of emotion classes.

In order to provide the same experimental environment, we equalize the number of samples from each corpus as well as the number of emotions. After the min–max scaling of values of features, two feature sets are combined into a training feature set called HDFM.

4. Ensemble Classification for Korean and English

The ensemble classification for Korean and English is composed of two layers: one for the preclassification of the speech emotion and one for the metaclassification of the earlier classifications. The preclassification consists of four classifiers: logistic regression (LR), random forest (RF), gradient boosting (GB), and multilayer perceptron (MLP). The metaclassifier is called the EV. Figure 5 shows the components of ensemble classification.

4.1. Preclassifiers. We introduce a novel approach to discover the optimal hyperparameters for each model and subsequently use them for predicting optimal values through EV, using grid search (GS) as the initial step within the ensemble framework, encompassing models such as RF, LR, MLP, and GB. Our focus here is on the advantages of GS in managing model complexity and addressing parameter uncertainty. By extensively testing various hyperparameter combinations, GS effectively manages model complexity, thus mitigating overfitting and enhancing generalization performance. Additionally, GS minimizes parameter uncertainty by considering all possible hyperparameter combinations, providing an opportunity to maximize model performance through the identification of optimal hyperparameter values.

LR is a supervised learning algorithm that uses regression to classify data into categories that are more likely to fall into a category. Most LR applications are used for binary classification. When there are three or more classes to be distinguished, LR analysis is an effective approach for multiple classification. The softmax function, represented by Equation (4), replaces the role of the sigmoid function in converting the z-values to probabilities in binary classification. It compresses the output values of multiple linear equations between 0 and 1 and adds the probability of all classes:

$$\sigma(\vec{z})_{\mathrm{i}} = \frac{e^{z_{\mathrm{i}}}}{\sum_{j=1}^{K} e^{z_{\mathrm{i}}}} \quad . \tag{4}$$

To compute the probability of the z-value, the softmax function applies the standard exponential function (e) to each element, $\sigma(\vec{z})_{\mathrm{i}}$ of the input vector and normalizes these values by dividing by the sum of all these exponentials. In detail, the softmax function, denoted $\sigma(\vec{z})_{\mathrm{i}}$, takes an input vector, $\vec{z}$, and computes a probability distribution over K classes (where K is the number of classes). Each element, $\sigma(\vec{z})_{\mathrm{i}}$, of the output vector represents the probability that the input belongs to the ith class. In the context of the softmax equation, the variable j serves as a summation index, ranging from 1 to K. It is used to represent the individual elements or components of the input vector, $\vec{z}$, which allows for the calculation of probabilities for each class.

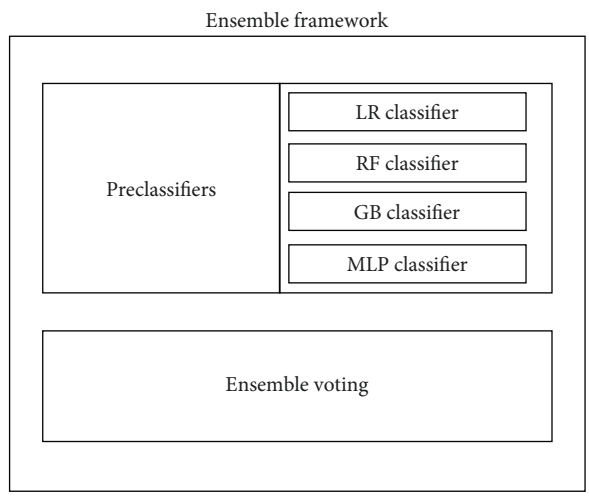

Figure 5: Components of a two-layered ensemble framework.

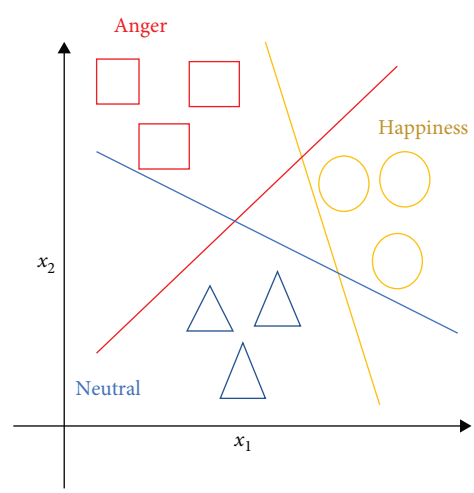

Figure 6: Multinomial emotion classification using linear regression algorithm.

Figure 6 shows the multinomial classification with linear regression in this study. The scikit-learn library is applied to

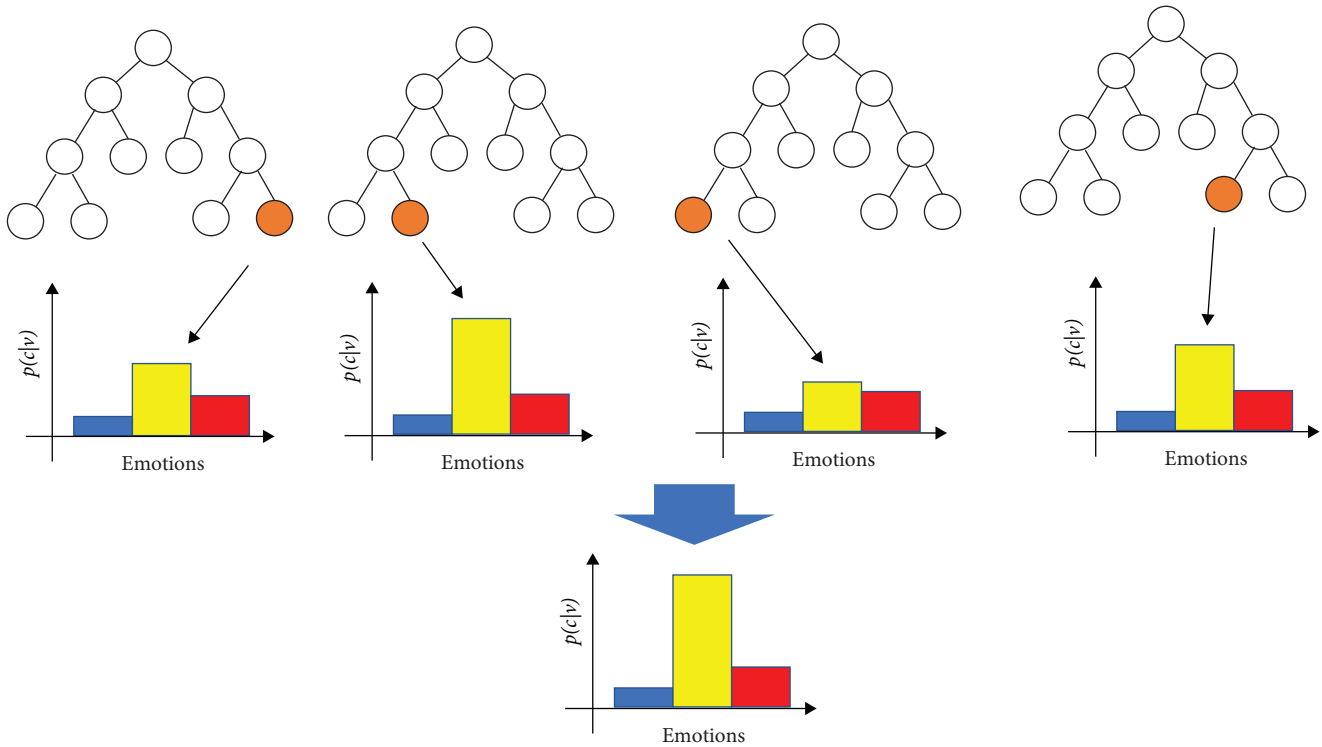

Figure 7: Graphical representation of RF algorithm in SER.

seven different kinds of multinomial emotions. The hyperparameters for LR include "penalty" (regularization term), C (inverse of regularization strength), and "solver" (optimization algorithm). These settings would impact the trade-off between model complexity and overfitting.

The random forest classifier (RFC) is based on a model comprising several decision trees. It randomly draws data to create several small trees and combines them [7]. If a single decision tree predicts the y-value using all features as variables, an overfitting problem arises. Thus, the RFC is applied to alleviate overfitting concerns. For example, when 30 input variables exist, one input variable, A, is the most important for prediction, and the rest play a minor role. In this case, for the majority of the bagged trees, the input variable, A, is used for the top branch. Eventually, even though several trees can be used to improve performance, most trees have a similar form. Due to the characteristics of bagging, which takes the average of several trees, if the results of each prediction model are similar, the results are similar, even if the average is taken. Therefore, the RFC randomly selects five of them to make a tree, then randomly selects five features to make a second tree, and it makes several trees in this way. All the trees in the forest are independently trained, and in the test phase, the data point, $\boldsymbol{v}$, is simultaneously entered into all the trees to reach the end node. The number of predicted values is the same as the number of trees, and the result is selected through voting [32]:

$$p\left(c|\boldsymbol{v}| = \frac{1}{T}\sum_{t=1}^{T} P_t(c)\boldsymbol{v}\right). \tag{5}$$

In this study, the RFC algorithm as a preclassifier is used to avoid dependency on features that play a critical role. After 10 random features were selected out of 200 features per sample, and the leaf depth was set to 30, the final emotion was classified by a majority of emotions with 700 trees per sample [7]. Figure 7 represents the procedures of the RFC, from creating trees to voting classification from trees. The RFC ("rf_best") is tuned using GS with hyperparameters, such as "n_estimators" (number of trees in the forest), "criterion" (splitting criterion), "max_depth" (maximum depth of trees), and "max_features" (maximum number of features considered for splitting). The "class_weight" is set to "balanced" to handle class imbalance.

Boosting uses the results of a particular model as the input of the next model and calculates the results by giving weight between models. It is also called a sequential ensemble. In the first data, it can be seen in the order of 1-2-3 classifiers. According to the first model results, in the case of data with a large error, the weight is given to the next classifier. The data with a large error, named the weak learner, grant high weights, and the well-predicted data give low weights and are passed on to the next model. GB is a method of gradating the inputs handed over to the next

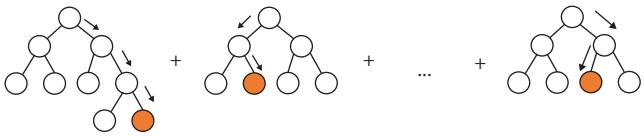

FIGURE 8: Application and equations of the GB algorithm in SER.

model and weighting them. GB uses the derivative value for the loss function (equal to the negative slope for the residual) to find the direction in which the loss function value decreases. By passing this result to the input of the new model, the new model is updated in the direction of reducing this value. That is, the models continue to learn in the direction of reducing the residual (difference between the actual value and the predicted value).

When y is the true value and $f(x)$ is the prediction of y, then:

$$\text{Loss function}: (y, f(x)) = \frac{1}{2}(y - f(x))^2 , \tag{6}$$

$$\begin{aligned}\text{Negative gradient}: \frac{\partial(y, f(x))}{\partial f(x)} &= \frac{\partial\left[\frac{1}{2}(y - f(x))^2\right]}{\partial f(x)} \\ &= -(f(x) - y) = y - f(x) .\end{aligned} \tag{7}$$

Therefore, the model performs residual fitting with a square error loss function.

In this study, we used the gradient boosting classifier (GBC) as one of the preclassifiers. With GBC, the number of weak learners is limited to 500 : 20 out of the 200 features are selected to configure the weak learner, and the deviance function is applied with a learning rate of 0.1 to increase the weight of the error value to reduce the value of the error when the prediction is made. By reducing the error, 500 trees are sequentially connected. In Equations(8)–(10), $A(x)$ is the first weak learner tree and E is the error in the corresponding model, that is, the residual, where E (residual) is again fitted with a weak learner named $B(x)$. Figure 8 shows the summing of the residual to the next tree to reduce the error rate:

$$\mathrm{F}(x) = A(x) + E , \tag{8}$$

$$E = B(x) + E' , \tag{9}$$

$$\mathrm{F}(x) = A(x) + B(x) + C(x) + \ldots + E'' . \tag{10}$$

The GBC ("gbm_best") is also optimized using GS. Parameters like "n_estimators" (number of boosting stages), "learning_rate" (step size for updates), "loss" (loss function), and "max_features" (maximum number of features considered) are explored. The "class_weight" is "balanced" for addressing the class imbalance.

The perceptron consists of an input layer and an output layer. At this point, the output layer is one node. The input layer is a $d+1$ node, where d is the dimension of the feature vector. The perceptron multiplies the input node by its weight and passes it to the output node. A bias node (denoted by node 0) is almost always included in the input layer to account for a constant offset in the data and has a constant value of 1, as depicted in Figure 9 [31, 33].

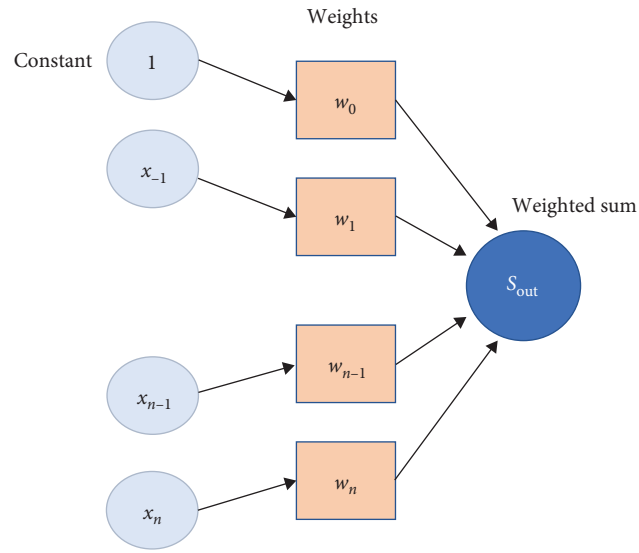

FIGURE 9: Application of MLP algorithm in SER.

As shown in Figure 9, we multiply all x_i and w_i and add them to a function called the activation function:

$$\mathrm{o} = \tau(\mathrm{s}) = \tau\left(\sum_{i=1}^{d} w_i x_i + w_0\right) . \tag{11}$$

In this study, the MLP classifier (MLP-C) uses rectified linear unit (ReLU) as an activation function and adaptive moment estimation (Adam) as the gradient-based solver for weight optimization. The entire input layers are set to 200, and the number of nodes in the hidden layer is limited to 500. The learning rate is set to an adaptive value, which is set to an average value if the training loss is not reduced. Adam optimization adopts gradient descent with momentum and root mean square propagation (RMSProp). We use ReLU instead of the sigmoid to activate the hidden layer. This function returns 0 if the value is less than 0 and the actual value if it is greater than 0 [33]. The MLP-C ("nnet_best") involves a wide range of hyperparameters set in the "params" dictionary. These include "activation" (activation function), "hidden_layer_sizes" (number of neurons in hidden layers), "alpha" (L2 regularization term), "solver" (optimization algorithm), "learning_rate" (learning rate schedule), "warm_start" (reuse the solution of the previous call), and "momentum" (momentum for gradient descent).

The hyperparameter settings of the preclassifier of the proposed model are thoroughly discussed in this section. Starting with LR, key hyperparameters, such as "penalty," C, and "solver," are meticulously selected to manage the trade-off between model complexity and overfitting. The RFC ("rf_best") is optimized using GS with parameters like "n_estimators," "criterion," "max_depth," and "max_features." The "class_weight" is set to "balanced" to tackle class

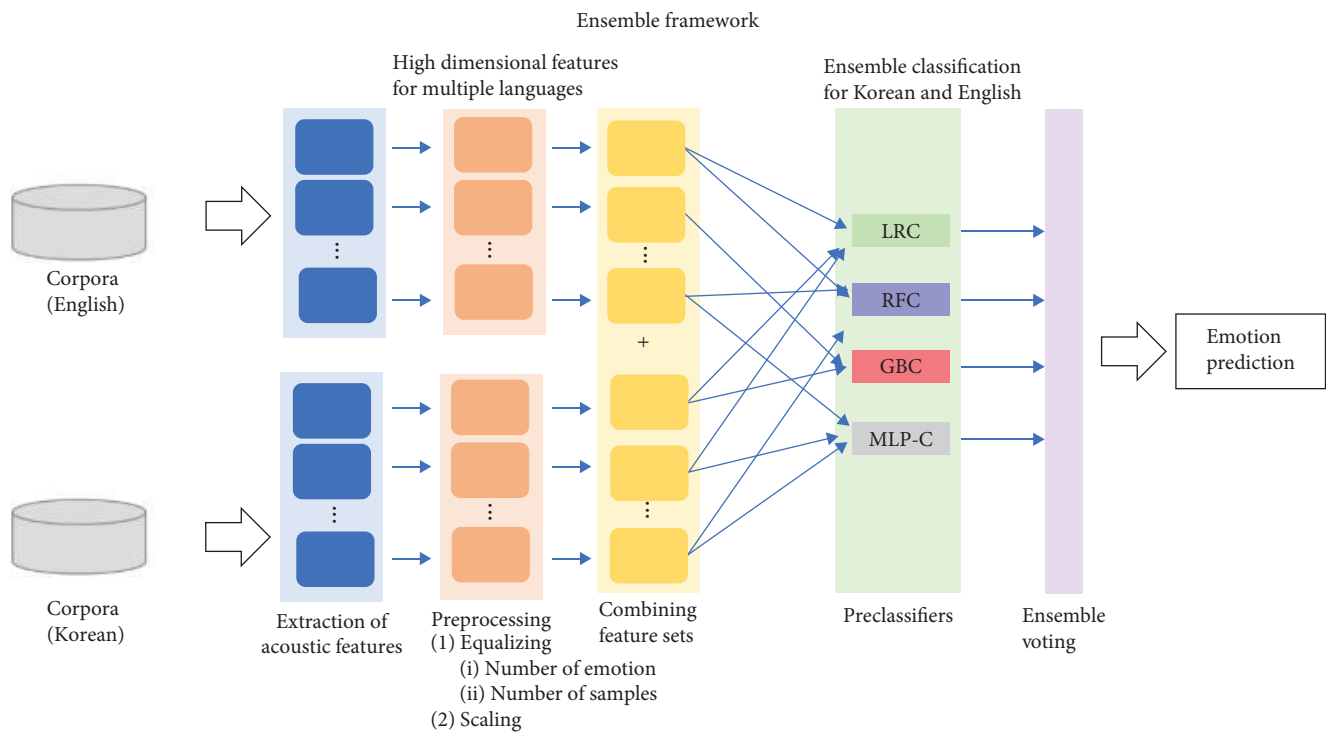

Figure 10: Procedures leading to EV in the ensemble framework.

imbalance effectively. Similarly, the GBC ("gbm_best") undergoes GS optimization involving "n_estimators," "learning_rate," "loss," and "max_features." The "class_weight" is also "balanced" to address the class imbalance. The MLP-C ("nnet_best") employs a diverse range of hyperparameters such as "activation," "hidden_layer_sizes," "alpha," "solver," "learning_rate," "warm_start," and "momentum."

These hyperparameters are meticulously tuned to optimize each classifier's performance while accounting for diverse model complexities, regularization effects, and data attributes. The application of GS ensures a systematic exploration of the hyperparameter space to identify the most favorable configuration. This experimentation is essential to strike a balance between model intricacy and generalization. However, it is noteworthy that GS bears the risk of overfitting due to evaluation across all hyperparameter combinations, potentially leading to overfitting on specific validation data and consequently compromising overall generalization performance. Furthermore, the inherent limitation of GS lies in its independent exploration of individual parameters without considering their interaction. This constraint may impede the optimization of model performance during hyperparameter search. To mitigate these limitations, we propose future investigations into more flexible exploration strategies. For instance, considering methods, such as randomized search (RandomizedSearchCV) or Bayesian optimization, can account for parameter interactions, enabling effective hyperparameter search. Through such endeavors, the complexities of model intricacies and parameter uncertainties would be navigated more adeptly.

The ensemble approach is further exemplified through the voting classifier (VC), which combines predictions from base classifiers ("log_best," "rf_best," "gbm_best," and "nnet_best") using a specified voting mechanism, particularly "soft" voting. While the VC itself has fewer hyperparameters to fine-tune, its efficacy relies heavily on the performance of its underlying base classifiers. These base classifiers were optimized using GS as well, each with distinct hyperparameter settings.

By amalgamating predictions from multiple classifiers, the VC aims to counteract individual classifier weaknesses while capitalizing on their strengths, leading to improved predictive accuracy. This ensemble technique, through a strategic selection of hyperparameters, serves to enhance classification outcomes, address class imbalance, and uncover intricate data patterns. The overall result is a more robust and potent classification framework capable of delivering enhanced results across diverse scenarios.

4.2. Ensemble Voting. In the framework, the EF-KEN is represented in Figure 10. The EV classifier makes the final decision about the predicted emotion among the recognized emotions from the preclassifiers with high-dimensional features. In the framework, five modules are executed serially to deliver the most likely prediction of emotions: preprocessing, extracted feature sets, normalization of emotion classes, combining feature sets, preclassifier layers, and EV, as shown in Figure 10.

EV is the metaclassifier that determines the final prediction result through voting [34]. It is the last layer before the preclassification layer, which is composed of the LRC, RFC, GBC, and multilayer perceptron classifier (MLP-C). EV collects the best parameters and predictions from the preclassifications

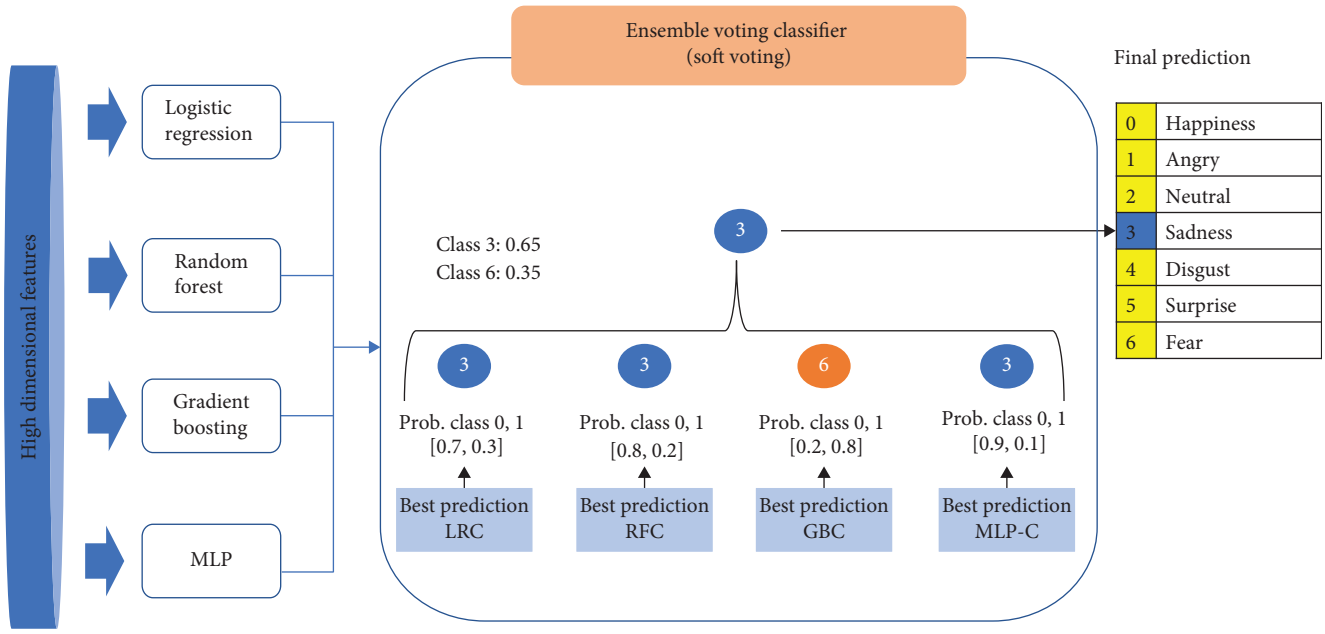

Figure 11: Soft voting process of emotion detection in EV.

layer and uses them as inputs for the VCs. This results in combining classifiers with a relatively higher probability of prediction. Voting is classified into two types: hard voting and soft voting. Hard voting follows most of the results of each classifier. That is, it follows the principle of the majority rule. Soft voting adds the probabilities of the classifier and averages each to select the result with the highest probability. In this study, we use soft VCs, as defined in Equation (12). Essentially, we combine predictions from different models (j) by multiplying them with their respective weights (w_j) and their corresponding scores or probabilities (p_{ij}). The summation over j aggregates these weighted predictions for each class, i, and the "argmax" operator selects the class (i) with the highest aggregated score as the final prediction, $\hat{y}$. The process of the final prediction is illustrated in Figure 11:

$$\widehat{y} = \arg \underbrace{\max}_{i} \sum_{j=1}^{m} w_j p_{ij}. \tag{12}$$

In a multilingual environment, we overcome the shortcomings of ML algorithms by combining various classifiers to learn various situations. As alluded to above, the advantage of EF-KEN is that it can maximize the effects of a biased trade-off while complementing the shortcomings. In 2021, Zehra et al. [28] conducted similar approaches for multilingual SER. They used the corpora of English, German, Italian, and Urdu with ensemble classifiers. The classifiers used for Zehra et al.'s [28] study were support vector machine based on sequential minimal optimization (SMO), RF, decision tree, and majority voting. Although the datasets are different from our study, we also used majority voting as a final classifier. Our approach proves that the concept of sequential layers of classifiers could have a significant impact on predicting emotions. In 2020, Heracleous et al. [35] combined audio features for emotion detection with three datasets of European languages (English, Italian, and Spanish) for detecting emotion.

Recall, precision, recognition accuracy (RA), and F-score metrics are used to measure the performance of the EF-KEN as evaluation metrics. We experiment with EF-KEN under balanced data conditions with an equal number of samples in both languages. RA is a commonly used metric in SER research and is chosen to evaluate the performance of the framework. The final RA was averaged across the RA results for each emotion. We also use recall, precision, and F1 scores for estimating each emotion.

5. Experimental Results

5.1. Preliminary Experiments. The classifiers designed for preclassification of the framework were evaluated in preliminary experiments. The English dataset was first tested using LRC, RFC, GBC, and MLP-C trained on English data, and the RFC and MLP-C achieved an RA of 36%. Similarly, the Korean dataset was tested using the same classifiers, and the RFC, GBC, and MLP-C achieved an RA of 41%. The RA values obtained in these experiments are presented in Table 3 as the baseline for comparison with the experimental results.

Table 4 shows the accuracy rate of the RFC in the second preliminary experiment, which aimed to calculate the phonetic correlation between the English and Korean databases by testing the Korean database with the model trained on the English database. The results indicate a low accuracy rate of only 13% for the RFC.

In order to investigate the cause of the low accuracy in predicting emotions across different languages, Praat software [36, 37], commonly used for phonetic research, was used to measure the formant frequency. The formant

Table 3: RA of different classifiers on the same training/testing dataset.

Database (training and testing)	LRC (%)	RFC (%)	GBC (%)	MLP-C (%)
English (IEMOCAP)	32	36	37	36
Korean (KETI)	27	41	41	41

Table 4: RA of the training and testing on cross-Korean and -English datasets.

Training database	Test database	LRC (%)	RFC (%)	GBC (%)	MLP-C (%)
Korean (KETI)	IEMOCAP (English)	20	12	12	14
English (IEMOCAP)	Korean (KETI)	14	13	16	16

Table 5: F1–F4 frequency differences in anger emotion between Korean and English.

Training database	Emotion	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)
IEMOCAP	*Anger*	640	1,768	2,783	3,808
KETI	*Anger*	370	1,864	2,832	4,311

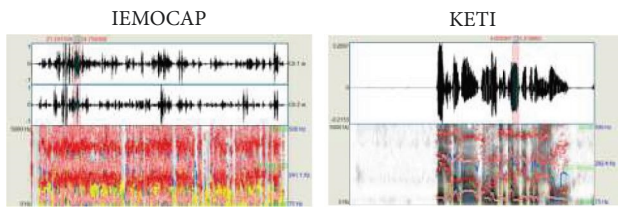

Figure 12: Formant frequency differences in anger emotion between Korean and English.

Table 6: RA results using HDFM as training dataset and Korean KETI as testing dataset.

Training database	Testing database	LRC (%)	RFC (%)	GBC (%)	MLP-C (%)
HDFM of IEMOCAP and KETI	KETI	26	23	20	25

frequency for each emotion was randomly measured to observe any linguistic or social differences between the two languages. The results in Table 5 and Figure 12 suggest that there are significant differences in the F1 frequency for the emotion of *anger*.

Table 5 and Figure 12 demonstrate that the F1 frequency for the emotion of anger was twice as high for the English speaker compared to the Korean speaker. This discrepancy could point to the differences in vowel frequencies between Korean and English for the same emotion as a possible explanation for the low RA obtained by classifiers in cross-training and testing with different corpora [38, 39].

5.2. Preclassifiers with HDFM. In the process of constructing the HDFM feature set, 80% of each of the 8,000 data samples in English and Korean were randomly selected and trained by the preclassifiers of the ML algorithms. To balance the number of training samples, the English testing dataset was tested with only 90% of the randomly selected 2,039 data, and the Korean testing dataset was also randomly tested with only 90% of the 2,039 data samples. The results of the experiment, as shown in Table 6, gave a better prediction rate than the result shown in Table 4 with the crossed design of different training and testing corpus. The prediction rate of the classifiers decreased in the order of LRC, RFC, GBC, and MLP-C, as shown in Table 6.

Table 7 shows the accuracy rates for each emotion using the four different classifiers: LRC, RFC, GBC, and MLP-C. The LRC has a high accuracy rate for *sadness* and *happiness*, whereas the RFC has a high accuracy rate for *happiness* and *anger*. The GBC has a high accuracy rate for *fear*, *happiness*, and *neutral*, whereas the MLP-C has a high accuracy rate for *happiness*, *anger*, and *fear*. Overall, the emotion recognition for *happiness* and *anger* has a high accuracy rate across all classifiers.

When tested on English datasets using the same training model, as shown in Table 8, the LRC and GBC performed better on the English datasets compared to the Korean tests, whereas the RFC and MLP-C showed better results on the Korean datasets.

Table 9 shows that the LRC has a high recall rate, above 60%, for the emotions of *sadness* and *anger*. The RFC has an

Table 7: Preclassifiers' RA results for seven emotions using HDFM as training dataset and Korean KETI dataset as testing dataset.

	Precision	Recall	F1-score
LRC			
Anger	0.14	0.08	0.11
Disgust	0.20	0.18	0.19
Fear	0.16	0.33	0.22
Happiness	0.38	0.50	0.43
Neutral	0.17	0.03	0.05
Sadness	0.27	0.38	0.32
Surprise	0.08	0.02	0.03
Accuracy			0.26
Macro Average	0.20	0.22	0.19
Weighted Average	0.23	0.26	0.23
RFC			
Anger	0.15	0.63	0.24
Disgust	0.00	0.00	0.00
Fear	0.44	0.07	0.13
Happiness	0.33	0.51	0.40
Neutral	0.18	0.04	0.06
Sadness	0.70	0.05	0.10
Surprise	0.33	0.01	0.01
Accuracy			0.23
Macro Average	0.30	0.19	0.13
Weighted Average	0.30	0.23	0.17
GBC			
Anger	0.12	0.11	0.12
Disgust	0.12	0.06	0.08
Fear	0.25	0.38	0.30
Happiness	0.33	0.33	0.33
Neutral	0.14	0.29	0.19
Sadness	0.23	0.15	0.18
Surprise	0.33	0.01	0.01
Accuracy			0.20
Macro Average	0.18	0.20	0.18
Weighted Average	0.20	0.20	0.19
MLP-C			
Anger	0.18	0.41	0.25
Disgust	0.23	0.15	0.18
Fear	0.21	0.34	0.26
Happiness	0.36	0.37	0.36
Neutral	0.06	0.00	0.01
Sadness	0.35	0.15	0.21
Surprise	0.16	0.23	0.19
Accuracy			0.25
Macro Average	0.22	0.24	0.21
Weighted Average	0.25	0.25	0.23

80% recall rate for the *angry* emotion. The GBC has a recall rate of 44% and 100% for the *disgust* and *fear* emotions, respectively. The MLP-C has a higher recall rate for the *anger* and *sadness* emotions compared to other emotions. In summary, the results suggest that the recognition of emotional *anger* is consistent across the classifiers.

With HDFM, the Korean testing dataset showed an improvement of at least 4% and up to 10% in the RA of preclassifiers compared to the results of the crossdataset shown in Table 4. Similarly, the English testing dataset showed a positive effect with an increase of 3%–7% compared to the results of the crossdataset.

5.3. Completion of EV Classifier with HDFM. The final step of the HDFM framework involves the EV classifier, which takes the output from the preclassifiers as input. Soft voting is used to determine the highest results of preclassification. The RA of the Korean testing dataset increased by 15% from the average RA of Korean emotion recognition in the crossdataset in Table 4. Table 10 shows the improved result with the EV. The final RA of the English testing dataset increased by 13% from the average RA of English emotion recognition in the crossdataset. However, there is an exceptional RA of 32% in the LRC of preclassification, as shown in Table 8. This LRC shows a high prediction rate, particularly for the emotions of *anger*, *disgust*, and *neutral* in Table 9. The precision data overall show a high prediction rate.

Table 11 shows that the recall rate for the emotions of *anger* and *happiness* are higher than other emotions, while the F1 scores of the emotions of *happiness* and *sadness* are higher than other emotions. The macroaverage of precision, recall, and F1-score is lower than the RA, while the weighted average of precision and recall is the same as the RA.

In Table 12, the weighted average of precision in English testing is higher than the overall RA, indicating that the EV shows better precision for some emotions in the English testing dataset. Specifically, the emotions of *disgust* and *sadness* have higher precision rates with EV than the other emotions. However, the emotions of *fear* and *surprise* have a 0% prediction rate, likely due to the smaller sample size compared to other emotions.

5.4. Comparison with Other Studies. The research on SER with the IEMOCAP dataset is ongoing. Table 13 represents the state-of-the-art benchmarks for SER using IEMOCAP, as curated by "paper with code." Many of these multimodal approaches achieve recognition rates in the early 80% range.

However, the experiment conducted by Liu et al. [44], shown in Table 14, reveals somewhat unexpected outcomes. Cross-experimenting between the same English emotion data from IEMOCAP and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) for four emotions results in recognition rates below 40% despite the data being in the same language. Table 15 compares our study with Liu et al.'s [44] experiment, highlighting the superiority of our approach targeting both Korean and English emotions over English-only emotion data.

Zehra et al.'s [28] model obtained the results by using up to seven emotions with positive and negative valence. In order to compare our approach with Zehra et al.'s [28] model, we also differentiated our model's emotions into positive and negative valence, as presented in Table 16.

Table 8: RA results using HDFM as training dataset and English IEMOCAP as testing dataset.

Training database	Testing database	LRC (%)	RFC (%)	GBC (%)	MLP-C (%)
HDFM of IEMOCAP and KETI	IEMOCAP	32	20	24	21

Table 9: Preclassifier RA results for seven emotions using HDFM as training dataset and English IEMOCAP dataset as testing dataset.

	Precision	Recall	F1-score
LRC			
Anger	0.34	0.62	0.44
Disgust	0.39	0.41	0.40
Fear	0.00	0.00	0.00
Happiness	0.22	0.06	0.10
Neutral	0.44	0.02	0.04
Sadness	0.35	0.61	0.45
Surprise	0.04	0.14	0.06
Accuracy			0.32
Macro Average	0.18	0.19	0.13
Weighted Average	0.28	0.20	0.18
RFC			
Anger	0.19	0.80	0.30
Disgust	0.33	0.12	0.17
Fear	0.00	0.00	0.00
Happiness	0.08	0.19	0.11
Neutral	0.27	0.10	0.15
Sadness	0.41	0.14	0.20
Surprise	0.00	0.00	0.00
Accuracy			0.20
Macro Average	0.18	0.19	0.13
Weighted Average	0.28	0.20	0.18
GBC			
Anger	0.00	0.00	0.00
Disgust	0.36	0.44	0.39
Fear	0.25	1.00	0.40
Happiness	0.14	0.21	0.17
Neutral	0.25	0.16	0.19
Sadness	0.43	0.14	0.21
Surprise	0.00	0.00	0.00
Accuracy			0.24
Macro Average	0.20	0.28	0.20
Weighted Average	0.27	0.24	0.24
MLP-C			
Anger	0.23	0.34	0.28
Disgust	0.31	0.21	0.25
Fear	0.00	0.00	0.00
Happiness	0.09	0.20	0.12
Neutral	0.39	0.05	0.08
Sadness	0.30	0.45	0.36
Surprise	0.00	0.00	0.00
Accuracy			0.21
Macro Average	0.19	0.18	0.16
Weighted Average	0.29	0.21	0.21

As presented in Table 17, the EF-KEN model outperformed Zehra et al.'s [28] model in recognizing emotions in English. In addition, when comparing the emotion recognition rates for Urdu and Korean separately, the EF-KEN model also showed better performance.

Zehra et al.'s [28] study used EV with RF, decision tree (J48), and SMO as preclassifiers, differing from our approach. The dataset in Zehra et al.'s [28] study comprises Urdu and English languages, focusing on positive and negative emotions. The results indicate a 43% recognition rate for English and a 45% recognition rate for Urdu. Under the same conditions, our model exhibited improved results of 60% for English and 57% for Korean emotions.

While the languages used and the databases involved differ, our ensemble framework, including HDFM and preclassifiers, can be considered superior to previous work in terms of performance. The primary distinction in our research model lies in the incorporation of diverse preclassifiers and the mitigation of emotion class imbalance within the IEMOCAP dataset. These represent the most significant differentiating factors that set our model apart.

5.5. Further Studies. Studies have indicated that even within the same language, variations in training and testing datasets can impact the accuracy of SER. This phenomenon was evident in Liu et al.'s [44] study, where different datasets for training and testing in English led to decreased emotion prediction accuracy. In our research, we confronted similar issues. Despite training on combined datasets of both Korean and English, we found that the RA fell short when evaluating each language individually.

The dataset's diversity and balance are critical factors in accurate emotion recognition. Without encompassing a range of contexts and environments in the training data, models might struggle to generalize effectively. Moreover, an imbalanced distribution of emotion categories within the dataset can result in reduced accuracy for specific emotions. To address this challenge, we integrated Korean and English data within the HDFM process to mitigate imbalanced issues in each emotion class. Attentive dataset collection and preprocessing are essential to tackle these challenges successfully.

As a result, our study implemented a reduction in the number of emotion categories to enhance the dataset's diversity and balance. This modification yielded improved recognition performance. This methodology can be applied to different languages and environments, thereby serving as a valuable approach to further the field of SER [48, 49].

Table 10: Comparison of RA: crosslingual dataset without HDFM, preclassifier with HDFM, and EV with HDFM.

Test language	Average of crosslanguage (%)	Average of preclassifier (%)	EV classifier (%)
Korean	15	24	30
English	15	24	27

Table 11: EV RA results for seven emotions using HDFM as training dataset and Korean KETI dataset as testing dataset.

	Precision	Recall	F1-score
Anger	0.19	0.34	0.24
Disgust	0.29	0.22	0.25
Fear	0.27	0.32	0.29
Happiness	0.37	0.49	0.42
Neutral	0.21	0.05	0.07
Sadness	0.37	0.30	0.33
Surprise	0.26	0.16	0.20
Accuracy			0.30
Macro average	0.28	0.27	0.26
Weighted average	0.30	0.30	0.28

Table 12: RA of English emotion for EV in the HDFM.

	Precision	Recall	F1-score
Anger	0.25	0.38	0.30
Disgust	0.36	0.36	0.36
Fear	0.00	0.00	0.00
Happiness	0.13	0.22	0.16
Neutral	0.27	0.01	0.02
Sadness	0.35	0.53	0.42
Surprise	0.00	0.00	0.00
Accuracy			0.27
Macro average	0.19	0.21	0.18
Weighted average	0.29	0.27	0.24

Table 13: Papers with high rankings in state-of-the-art SER using IEMOCAP [40].

Authors	WA (%)	UA (%)	Paper
Lian [41]	82.7	–	Context-dependent domain adversarial neural network for multimodal emotion recognition
Gat [42]	81.0	–	Speaker normalization for self-supervised SER
Jalal [43]	80.5	74.0	Empirical interpretation of speech emotion perception with attention-based model for SER

Table 14: Crossdataset evaluation results for English language.

Training dataset	Testing dataset		
	IEMOCAP (%)	RAVDESS [45] (%)	MSP-IMPROV [46] (%)
IEMOCAP	77.8	40.8	36.5

Table 15: Comparison of crossdataset evaluation results and EF-KEN performance for English and Korean.

Training dataset	Testing dataset (four emotions)			
	IEMOCAP (%)	RAVDESS [45] (%)	MSP-IMPROV [46] (%)	KETI (%)
IEMOCAP	77.8	40.8	36.5	–
EF-KEN	39	–	–	43

Table 16: Emotion classification between Zehra et al.'s [28] ensemble model and EF-KEN's multilingual model.

Model	Corpus	Language	Positive valence	Negative valence
EF-KEN	IEMOCAP	English	*Happiness*, *excitement*, *surprise*, *neutral*	*Anger*, *sadness*, *fear*, *disgust*, *others*
	KETI	Korean	*Happiness*, *surprise*, *neutral*	*Anger*, *sadness*, *fear*, *disgust*
Zehra et al. [28]	SAVEE [47]	English	*Happiness*, *surprise*, *neutral*	*Anger*, *sadness*, *fear*, *disgust*
	Urdu [28]	Urdu	*Happiness*, *neutral*	*Anger*, *sadness*

Table 17: Performance comparison between Zehra et al.'s [28] ensemble model and EF-KEN's multilingual model.

Testing languages	Models	
	EF-KEN (%)	Zehra et al. [28] (%)
English	61	43
Korean	57	–
Urdu	–	45

6. Conclusions

In this study, evidence was presented of distinctive formant differences for specific emotions between English and Korean, hypothesizing that these differences posed additional challenges in emotion prediction for both languages. To address this, an ensemble framework was developed for English and Korean emotion recognition, using high-dimensional feature integration and two layers of ensemble classifiers. This framework comprised the HDFM and EV connected through normalized feature sets from Korean and English speech databases. The HDFM feature set was constructed for training and evaluation on a mixed emotion database from Korean and English databases, significantly alleviating the inherent problem of emotion class imbalance observed in the IEMOCAP emotion data. Moreover, the advantages of the ensemble framework included intuitive design, low computational demands, and improved prediction speed when training in one language and testing in another.

The framework's preliminary classifiers enhanced the RA by approximately 9% for Korean and 10% for English across seven emotions. The overall framework improved the final prediction accuracy by about 15% for Korean and 13% for English. The results demonstrated that the EV provided superior predictive performance compared to ML algorithms alone. The EF-KEN model was compared to emotion data research in English, confirming that its feature set construction and model design contributed to enhanced predictive performance. Particularly, diverse configurations of the preliminary classifier yielded improved results compared to other studies.

The proposed approach's strengths lie in its ability to be easily deployed in lightweight, stand-alone, or minimally resource-intensive environments using ML algorithms. However, one limitation is that the approach relies solely on acoustic features and does not include aspects like context modeling or the flow of context. As a result, it may have limitations in practical applications where the context of emotions, such as in psychological counseling, holds significant importance.

In closing, this study has laid the foundation for cross-lingual emotion recognition with promising results. Future research will focus on enhancing the robustness, multimodality, and real-world applicability of these systems, fostering a deeper understanding of emotions across languages and cultures [50, 51]. Additionally, the use of more advanced deep learning architectures, including transformers and attention mechanisms, could be investigated to capture complex dependencies and temporal relationships in speech data. These models have shown promise in various natural language processing tasks and may enhance the performance of emotion recognition systems.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the Chung-Ang University research grant in 2023.

References

[1] C. Busso, M. Bulut, C.-C. Lee et al., "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[2] S. G. Koolagudi and K. Sreenivasa Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, pp. 99–117, 2012.

[3] B. T. Atmaja, A. Sasou, and M. Akagi, "Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion," *Speech Communication*, vol. 140, pp. 11–28, 2022.

[4] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," in *IEEE Access*, vol. 9, pp. 47795–47814, 2021.

[5] R. Vempati and L. D. Sharma, "A systematic review on automated human emotion recognition using electroencephalogram signals and artificial intelligence," *Results in Engineering*, vol. 18, Article ID 101027, 2023.

[6] Y. B. Singh and S. Goel, "A systematic literature review of speech emotion recognition approaches," *Neurocomputing*, vol. 492, pp. 245–263, 2022.

[7] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human–robot interaction," *Information Sciences*, vol. 509, pp. 150–163, 2020.

[8] L. Guo, L. Wang, J. Dang, E. S. Chng, and S. Nakagawa, "Learning affective representations based on magnitude and dynamic relative phase information for speech emotion recognition," *Speech Communication*, vol. 136, pp. 118–127, 2022.

[9] A. H. Jo and K. C. Kwak, "A trend analysis on emotional recognition technology using speech signals," *Spring Conference on Korean Smart Media Journal*, 2020.

[10] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: a review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[11] Mustaqeem and S. Kwon, "1D-CNN: speech emotion recognition system using a stacked network with dilated CNN features," *Computers, Materials & Continua*, vol. 67, no. 3, pp. 4039–4059, 2021.

[12] M. Ishaq Mustaqeem and S. Kwon, "A CNN-assisted deep echo state network using multiple time-scale dynamic learning reservoirs for generating short-term solar energy forecasting," *Sustainable Energy Technologies and Assessments*, vol. 52, Article ID 102275, 2022.

[13] A. Aftab, A. Morsali, S. Ghaemmaghami, and B. Champagne, "LIGHT-SERNET: a lightweight fully convolutional neural network for speech emotion recognition," in *ICASSP. 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6912–6916, IEEE, Singapore, 2022.

[14] S. Chamishka, I. Madhavi, R. Nawaratne et al., "A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling," *Multimedia Tools and Applications*, vol. 81, no. 24, pp. 35173–35194, 2022.

[15] B. T. Atmaja, K. Shirai, and M. Akagi, "Speech emotion recognition using speech feature and word embedding," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 519–523, IEEE, Lanzhou, China, 2019.

[16] V. Chernykh and P. Prikhodko, "Emotion recognition from speech with recurrent neural networks," *ArXiv*, Article ID ArXiv.1701.08071, 2017.

[17] I. Aliyu and C. G. Lim, "EEG dimensional reduction with Stack Auto Encoder for emotional recognition using LSTM/RNN," *The Journal of the Korea Institute of Electronic Communication Sciences*, vol. 15, no. 4, pp. 717–724, 2020.

[18] Y. J. Ko and Y. J. Kim, "Performance improvement of speech emotion recognition model using generative adversarial networks," *The Journal of Korean Institute of Information Technology*, vol. 17, no. 11, pp. 77–85, 2019.

[19] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5089–5093, IEEE, Calgary, AB, Canada, 2018.

[20] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning," *Interspeech*, vol. 2021, pp. 4508–4512, 2021.

[21] J.-H. Kim and S.-P. Lee, "Multi-modal emotion recognition using speech features and text embedding," *The Transactions of the Korean Institute of Electrical Engineers*, vol. 70, no. 1, pp. 108–113, 2021.

[22] Mustaqeem and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, Article ID 183, 2020.

[23] R. Vempati and L. D. Sharma, "EEG rhythm based emotion recognition using multivariate decomposition and ensemble machine learning classifier," *Journal of Neuroscience Methods*, vol. 393, Article ID 109879, 2023.

[24] V. K. Velpula and L. D. Sharma, "Multi-stage glaucoma classification using pre-trained convolutional neural networks and voting-based classifier fusion," *Frontiers in Physiology*, vol. 14, 2023.

[25] M. J. Seo and M. H. Kim, "Ensemble method of emotion classifier for speech emotion recognition," *Journal of The Korea Society of Information Technology Policy & Management*, vol. 11, no. 2, pp. 1187–1193, 2019.

[26] V. Scotti, F. Galati, L. Sbattella, and R. Tedesco, "Combining deep and unsupervised features for multilingual speech emotion recognition," in *Pattern Recognition. ICPR International Workshops and Challenges*, A. D. Bimbo, R. Cucchiara, and S. Sclaroff, et al., Eds., vol. 12662 of *Lecture Notes in Computer Science*, pp. 114–128, Springer, Cham, 2021.

[27] S. Lalitha, D. Gupta, M. Zakariah, and Y. A. Alotaibi, "Investigation of multilingual and mixed-lingual emotion recognition using enhanced cues with data augmentation," *Applied Acoustics*, vol. 170, Article ID 107519, 2020.

[28] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," *Complex and Intelligent Systems*, vol. 7, pp. 1845–1854, 2021.

[29] T. K. Amol and R. M. R. Guddeti, "Multiclass SVM-based language-independent emotion recognition using selective speech features," in *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1069–1073, IEEE, Delhi, India, 2014.

[30] A. Bhavan, P. Chauhan, Hitkul, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Systems*, vol. 184, Article ID 104886, 2019.

[31] A. A. Alnuaim, M. Zakariah, P. K. Shukla et al., "Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier," *Journal of Healthcare Engineering*, vol. 2022, Article ID 6005446, 12 pages, 2022.

[32] S. Yan, L. Ye, S. Han, T. Han, Y. Li, and E. Alasaarela, "Speech interactive emotion recognition system based on random forest," in *International Wireless Communications and Mobile*

Computing (IWCMC), pp. 1458–1462, IEEE, Limassol, Cyprus, 2020.

[33] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, and G.-Z. Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271–280, 2018.

[34] M. M. V. Chalapathi, M. R. Kumar, N. Sharma, and S. Shitharth, "Ensemble learning by high-dimensional acoustic features for emotion recognition from speech audio signal," *Security and Communication Networks*, vol. 2022, Article ID 8777026, 10 pages, 2022.

[35] P. Heracleous, Y. Mohammad, and A. Yoneyama, "Integrating language and emotion features for multilingual speech emotion recognition," in *Human-Computer Interaction, Multimodal and Natural Interaction, HCII 2020*, M. Kurosu, Ed., vol. 12182 of *Lecture Notes in Computer Science*, pp. 187–196, Springer, Cham, 2020.

[36] T. W. Kim, S. J. Han, M. S. Kim, and J. H. Lee, "A design and implementation of speech recognition preprocessing system using formant frequency," in *Proceedings of the Korea Information Science Society Conference 1999*, pp. 198–200, 1999.

[37] S. P. Yi, "Study on pitch contour extracted from Korean emotional speech using momel," *Journal of Language Sciences*, vol. 25, no. 3, pp. 191–209, 2018.

[38] S. P. Yi, "An analysis of formants extracted from emotional speech and acoustical implications for the emotion recognition system and speech recognition system," *Phonetics and Speech Sciences*, vol. 3, no. 1, pp. 45–50, 2011.

[39] H. Chun, J. Chung, B. Kim, and Y. Lee, "An analysis on the pitch variation of the emotional speech," in *Proceedings of the Acoustical Society of Korea Conference*, pp. 93–96, 1999.

[40] https://paperswithcode.com/sota/speech-emotion-re cognition-on-iemocap.

[41] Z. Lian, J. Tao, B. Liu, J. Huang, Z. Yang, and R. Li, "Context-dependent domain adversarial neural network for multimodal emotion recognition," in *21st Annual Conference of the International Speech Communication Association, Virtual Event, Interspeech 2020*, H. Meng and T. F. Zheng, Eds., pp. 394–398, Shanghai, China, 25–29 October 2020.

[42] I. Gat, H. Aronowitz, W. Zhu, E. Morais, and R. Hoory, "Speaker normalization for self-supervised speech emotion recognition," *ICASSP. 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7342–7346, 2022.

[43] M. A. Jalal, R. Milner, and T. Hain, "Empirical interpretation of speech emotion perception with attention based model for speech emotion recognition," in *Interspeech 2020*, pp. 4113–4117, 2020.

[44] Z.-T. Liu, A. Rehman, M. Wu, W.-H. Cao, and M. Hao, "Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence," *Information Sciences*, vol. 563, pp. 309–325, 2021.

[45] S. R. Livingstone, F. A. Russo, and J. Najbauer, "The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, Article ID e0196391, 2018.

[46] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: an acted corpus of dyadic interactions to study emotion perception," in *IEEE Transactions on Affective Computing*, vol. 8, pp. 67–80, IEEE, 2017.

[47] P. Jackson and S. Haq, *Surrey Audio-Visual Expressed Emotion (savee) Database*, University of Surrey, Guilford, 2014.

[48] J. Li, N. Yan, and L. Wang, "Unsupervised cross-lingual speech emotion recognition using pseudo multilabel," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 366–373, IEEE, Cartagena, Colombia, 2021.

[49] Y. Xia and F. Xu, "Study on music emotion recognition based on the machine learning model clustering algorithm," *Mathematical Problems in Engineering*, vol. 2022, Article ID 9256586, 11 pages, 2022.

[50] L. D. Sharma and A. Bhattacharyya, "A computerized approach for automatic human emotion recognition using sliding mode singular spectrum analysis," *IEEE Sensors Journal*, vol. 21, no. 23, pp. 26931–26940, 2021.

[51] H. Aouani and Y. B. Ayed, "Speech emotion recognition with deep learning," *Procedia Computer Science*, vol. 176, pp. 251–260, 2020.