applied sciences

MDPI

Article

A Study on a Speech Emotion Recognition System with Effective Acoustic Features Using Deep Learning Algorithms

Sung-Woo Byun [1] and Seok-Pil Lee [2,*]

[1] Department of Computer Science, Graduate School, SangMyung University, Seoul 03016, Korea; 123234566@naver.com

[2] Department of Electronic Engineering, SangMyung University, Seoul 03016, Korea

* Correspondence: esprit@smu.ac.kr

Abstract: The goal of the human interface is to recognize the user's emotional state precisely. In the speech emotion recognition study, the most important issue is the effective parallel use of the extraction of proper speech features and an appropriate classification engine. Well defined speech databases are also needed to accurately recognize and analyze emotions from speech signals. In this work, we constructed a Korean emotional speech database for speech emotion analysis and proposed a feature combination that can improve emotion recognition performance using a recurrent neural network model. To investigate the acoustic features, which can reflect distinct momentary changes in emotional expression, we extracted F0, Mel-frequency cepstrum coefficients, spectral features, harmonic features, and others. Statistical analysis was performed to select an optimal combination of acoustic features that affect the emotion from speech. We used a recurrent neural network model to classify emotions from speech. The results show the proposed system has more accurate performance than previous studies.

Keywords: speech emotion recognition; recurrent neural network; speech analysis

check for updates

Citation: Byun, S.-W.; Lee, S.-P. A Study on a Speech Emotion Recognition System with Effective Acoustic Features Using Deep Learning Algorithms. *Appl. Sci.* **2021**, *11*, 1890. https://doi.org/10.3390/app11041890

Academic Editors: Youngchol Choi and Yongseok Lee

Received: 1 January 2021
Accepted: 18 February 2021
Published: 21 February 2021

1. Introduction

Recently, with the technological development in the information society, high performance personal computers are becoming rapidly popularized. Consequently, the interactions between computers and humans have been actively altering into a bidirectional interface. Therefore, there is a need to better understand human emotions. It could improve the interaction systems between humans and machines [1,2]. In signal processing, for these reasons above, emotion recognition has become an attractive research topic [3]. The goal of the human interface is to recognize the user's emotional state precisely and to give personalized media according to user's emotions.

Emotion refers to a conscious mental reaction that one experiences subjectively, which, in other words, is strong feeling generally accompanied by physiological and behavioral changes in the body [4]. To recognize a user's emotional state, many studies have applied diverse forms of input, such as facial expression, speech, text, video, and others [5–11]. Among emotion recognition studies, a speech signal is one of the most natural ways of human communication. It contains linguistic content and implicit paralinguistic information, including the speaker's emotions. Several studies have reported that acoustic features, speech-quality features, and prosodic features imply abundant emotional significance [12]. In the speech emotion recognition study, the most important issue is the effective parallel use of the extraction of proper speech features and an appropriate classification engine. These features include formant, energy and pitch features [13–15]. Moreover, many studies for speech-emotion recognition utilize the Mel-frequency cepstrum coefficients (MFCC) feature representatively [16,17]. However, because there doesn't exist an overt and deterministic mapping between features and emotional state [18], speech emotion recognition still has a lower recognition rate than other emotion-recognition methods, such as facial emotion recognition (FER). Therefore, it is critical to combine appropriate audio features in speech-emotion recognition.

To accurately recognize and analyze emotions from speech, it is also important to construct a superior speech database. The key to these studies is to guarantee verified, reliable expressions of emotion. Therefore, statistical evaluation of whether speech data involves emotions is needed. To meet these needs, many studies have tried to construct speech databases, and a growing number of speech emotion databases have therefore become available [19–21]. However, there are no public Korean speech emotion databases. Since there are differences in the expression of emotions in speech which depend on the culture and language of the country, a Korean-based emotional speech database is essential to the analysis of emotions based on Korean speech. In this work, we constructed a Korean emotional speech database (K-EmoDB) for speech emotion analysis and proposed a feature combination that can improve emotion recognition performance, using a Recurrent Neural Network (RNN) model [22]. To do this, the database was recorded using two different methods. The emotion categories involved were ("Neutrality", "Happiness", "Anger", "Sadness", "Excitement", and "Fear"). All data were recorded in a professional studio, to maximize the sound quality by eliminating any background noise. To ensure the emotional quality and naturalness of the data, an evaluation test was carried out with seven subjects. The subjects listened to each sample and had to decide on an emotion score from 5 to 1. Based on the evaluation results, 150 emotion data points in each category were chosen to construct the final Korean emotional speech database. To investigate the acoustic features, which can reflect distinct momentary changes in emotional expression, we extracted fundamental frequency (F0), MFCC, chroma, spectral features, harmonic features, and others. Statistical analysis was performed to select an optimal combination of acoustic features that affect the emotion from speech. We used a long short-term memory model (LSTM) based model to classify emotions from speech. To compare the proposed method performance, two experiments using K-EmoDB and various international databases were conducted. The results demonstrated that the proposed method shows better performance than previous studies.

This paper is organized as follows. Section 2 introduces research into existing speech analysis technology. Section 3 introduces the database used in this research. Section 4 explains the proposed emotion recognition method. Section 5 presents the experiment description and results, and then concludes with Section 6.

2. Related Works

2.1. Emotional Speech Database

The Berlin Emotional Speech database (EmoDB) is a database of emotional speech recorded with German utterances [19]. The database was constructed using everyday sentences, so that priority was given to the naturalness of speech. Ten actors, five female and five male, recorded speech data according to specified emotions, producing 10 different sentences for seven kinds of emotion: anger, boredom, disgust, fear, happiness, sadness, and neutral. The special feature of EmoDB is that all the utterances in it are composed with emotionless words, which means that EmoDB is helpful for finding out the proper features of emotion due to its exclusion of the influence of the emotional information of words.

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database consists of about 12 h of recorded data [21]. Speech, facial, and landmark data were captured during each recording session. Each session was a sequence of conversations involving men and women. A total of 10 actors split into five pairs took part in the recording. All data were recorded in a professional cinema studio, to maximize the sound quality. Actors were seated across from each other at a "social" distance of three meters. This setup enables realistic communication. An evaluation test was carried out with seven subjects, to label each utterance based on both audio and video streams. The data was categorized using 10 labels: neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excited, or other. Speech data not recognized by the subjects as exhibiting the appropriate emotion were not included in the database. Data were only included in the database when there were more than half consistent answers. This process provided evidence for the credibility of human recognition.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a video database of emotional speech and songs in North American English, classified into eight emotions, including neutral, calm, happy, sad, angry, fearful, disgusted, and surprised [20]. The database comprises information from 24 professional actors, and each actor has 60 audio-visual (AV) items and 44 song items, for a total of 104 data points. Each recorded production of an actor was available in three modality formats: AV, video only, and audio only.

2.2. Speech Emotion Recognition

Speech signals are some of the most natural media of human communication, and they have the merit of real-time simple measurement. They contain implicit paralinguistic information and linguistic content, including the speaker's emotions. In the designing of a speech emotion recognition system, it is a challenging task to identify and extract different emotion related speech features. Since proper selection of the features affects the classification performance, it is critical to combine appropriate audio features in speech emotion recognition. There were many approaches to recognizing emotion from speech, and each study used different speech features. Linear Prediction Cepstrum Coefficients, MFCCs, and F0 have been widely used for the recognition of speech emotion. However, the question of whether these features are effective for classifying emotions is still under discussion in many studies [23]. A few studies extracted features from speech and recognized emotions utilizing Gaussian Mixture Models (GMMs) [24] and Hidden Markov Models (HMMs) [25]. Recently, with the rapid development of deep-learning algorithms, RNN have been applied to various fields of speech analysis [26–28]. The main focus of the approach [26] is to classify the emotions in an utterance, rather than classifying the emotions using the frames of the utterance. To extract global features, the authors fed 32-dimensional frame features directly into an RNN. Then, the global features were fed into an extreme learning machine to classify the emotions. This approach recognizes emotions from an utterance, rather than using the frames of the utterance, and therefore requires a lot of computational power to train the network. In another approach, Wieman et al. [29] found that binary decision trees can determine the features most relevant to emotions. However, their experiment was performed using a small dataset. A study [30] assumed that the characteristics of speech vary in each person, and that emotions are affected by the age, gender, and acoustic features of the speaker. The authors focused on speech emotion recognition by grouping speech data by age and gender. They proposed a hierarchical gender- and age-based model and utilized different feature vectors of OpenSmile [31] and eGeMAPS [32]. The results indicated that building a separate classifier for each gender and age group produces better performance than having one model for all genders and ages. Chernykh and Prikhodko extracted acoustic features by classifying the features into three categories [33]:

- 3 Time domain: zero crossing rate, energy, and entropy of energy
- 5 Spectral domain: spectral centroid, spectral spread, spectral entropy, spectral flux, and spectral roll-off
- 13 MFCCs
- 13 Chroma: 12-dimensional chroma vector, standard deviation of chroma vector

The total set of acoustic features is a sequence of 34-dimensional vectors for each utterance.

An RNN with a Connectionist Temporal Classification approach [34] was used to classify emotions from speech. Zhao et al. proposed two CNN LSTM networks created by stacking local feature learning blocks and other layers, to extract emotional features [9]. The 1D CNN LSTM network was intended to recognize speech emotion and extract deep features from a raw signal. The 2D CNN LSTM network focused mainly on learning global contextual information from a log-mel spectrogram. The CNN LSTM networks could recognize speech emotion effectively, without using hand-crafted features.

Many studies have demonstrated the correlation between emotional voices and acoustic features [35–40]. However, because there does not exist overt and deterministic mapping between features and emotional state, speech emotion recognition still has lower recognition rate than other emotion-recognition methods, such as FER. For this reason, finding the appropriate feature combination is a critical task in speech-based emotion recognition.

3. Korean Emotional Speech Database

The study of emotion recognition has rapidly developed over the last decade with broad interest from researchers in neuroscience, computer science, psychiatry, psychology, audiology, and computer science. The key of these studies is to secure the availability of validated and reliable expressions of emotion. Most emotion data sets include either facial expressions or speech recordings. Among data sets, few contain audio-visual recordings of speakers in Korean. This study constructs a K-EmoDB and reports validity and reliability of the data based on ratings from participants. The database was recorded with Korean utterances from professional actors. All recorded data was recorded in a professional studio, considering the sound quality of the data by eliminating any background noise. The database was recorded using two different methods: (1) A method of recording the emotions using 120 emotion-inducing scenarios (20 of each emotion) according to each emotion. (2) A method of recording 20 emotionless sentences that did not affect the emotions in the 6-emotion version.

The database was recorded with Korean utterances from 20 professional actors (M = 28.1, SD = 3.41, age range = 23–35, 11 males and 9 females). To be eligible, actors needed to have Korean as their first language, to speak with a neutral Seoul accent, and to not possess any distinctive features. Participants were also required to identify text presented at a distance of 1.5 m. Figure 1a shows an actual actor during recording.

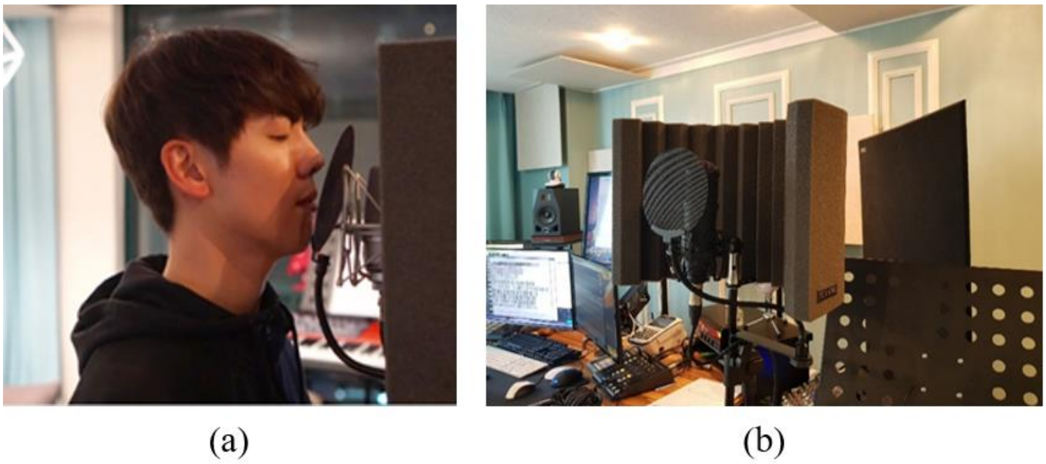

Figure 1. (**a**) One of the actors during recording (**b**) Recording equipment.

We chose six emotions for the experiment: neutral, anger, happiness, sadness, fear, and excitement. Here, "Excitement" refers to a more delighted with excitement emotion than "happiness". Neutrality was selected as the baseline emotion, and the remaining states consisted of the set of five basic or fundamental emotions that are thought to be culturally universal. The concept of primary emotions has a long history in science and philosophy. While the discrete model of emotion has been criticized [41–43], it is a practical choice for the creation and labelling of emotion sets [21]. Therefore, the six emotions can be found in most existing sets [44–54].

K-EmoDB was created using the following procedure. Actors were recruited through online casting, and 20 actors were selected by audition. All data were recorded in a professional studio, in order to optimize the sound quality of the data by eliminating any background noise. The Neumann TLM 103, Neumann U87 Ai, Oktava MK-319 microphone, and Universal Audio LA-610 MK1, WARM AUDIO WA76, MPAC-01 mixers were used for recording, as shown in Figure 1b. Each person recorded 120 sentences. Each speech data point is approximately three to five seconds long. All speech emotion data were recorded at 48 kHz, and downsampling was performed with a 16 kHz downsampling rate in a PCM signed 16-bit format.

To ensure the emotional quality and naturalness of the data, an evaluation test was carried out with seven subjects. The subjects listened to each sample and had to decide on an emotion score from 5 to 1. If subjects felt that the emotion was clearly contained in the speech data, they gave it a score of 5. However, if subjects felt that the speech data did not include the emotion, they give it a score of 1. Using to the evaluation results, 150 emotion data points in each category were chosen to construct the final K-Emo DB database. The final data consists of male and female speech files in a ratio of 50 to 50.

4. Speech Emotion Recognition

Speech is not only the most natural and universal communication but also contains paralinguistic information such as emotions and tone. Several previous studies have used prosodic and acoustic speech features to recognize emotion, and the current study investigated acoustic features employed in previous voice recognition studies using the emotion model shown in Figure 2.

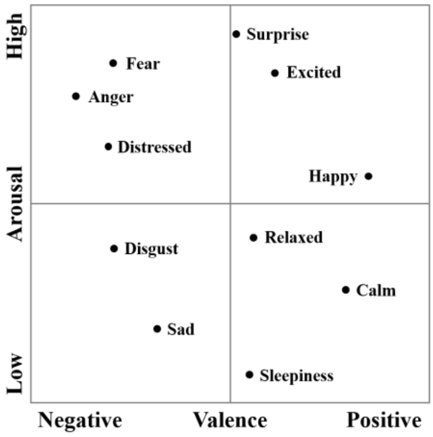

Figure 2. Two-dimensional emotion model [55].

Figure 2 shows the two-dimensional emotion model based on discrete emotion theory developed in previous studies. Horizontal and vertical axes refer to valence and arousal, respectively, where arousal represents excitement level (high or low), and valence represents the positive and negative emotions. For example, anger includes high excitement and negative emotion, and hence has high arousal and negative valence.

Most previous speech emotion studies mainly employed anger, disgust, fear, happiness, sadness, and surprise; whereas the present study added excitement to the basic emotions: anger, happiness, neutrality, sadness, and fear, providing six emotions describing speech purpose, reasonably evenly distributed over the arousal and valence scales. Generally, arousal is relatively easy to be distinguished whereas valence can be somewhat difficult. For example, voice pitch differs between happiness and sadness, making them relatively easy to distinguish; whereas voice pitch is high for both happiness and anger, making them more difficult to distinguish. Many studies have considered methods to difference valence from speech emotion recognition.

4.1. Feature Selection

Since speech acoustic features are important for emotion recognition, it is essential to select and analyze appropriate features. Lindstrom (2010) found that harmonic structures expressed positive and peaceful emotions, such as happy, cheerful, comfortable, and elegant, whereas dissonant intervals were closely related to negative and sad emotions, such as agitated, tense, angry, etc. Harmonic and dissonant speech intervals are closely related to the speech spectrum harmonic structure [56]. Table 1 shows harmonic features used to quantitatively analyze the harmonic structure. We extract the harmonic features using the features extraction tool Essentia which is open-source library and tools for audio and music analysis [57].

Table 1. Examples of sentences recorded for each emotion.

No	Features
1	Harmonic energy
2	Noise energy
3	Noiseness
4	F0
5	Inharmonicity
6	Tristimulus
7	Harmonic spectral deviation
8	Odd to even harmonic ratio

The harmonic energy feature presents the total sum of frequency peaks in a harmonic structure, when frequencies and magnitudes of harmonic peaks were given. Noise refers to the part besides harmonic structure in a spectrum, and the noise energy presents the subtraction of harmonic energy from the entire spectrum energy. The feature of Noiseness is that it represents the ratio of noise in a spectrum and it refers to the value of noise energy divided by the entire sum of spectrum. F0 feature means fundamental frequency of speech. Inharmonicity feature shows the degree of how much a spectrum does not form harmonic structure, and it is calculated as weighted sum of spectrum values which are the closest to multiples of fundamental frequency. Tristimulus feature is divided into 3 aspects; the first is to calculate relative weighted value, the second aspect is to calculate second, third and fourth harmonic's relative weighted value, and the third and the last is to calculate all the rest harmonics' relative weighted value. The detailed description of the harmonic features can be found in [57].

Table 2 shows the test result that distinguishes between Negative Valence and Positive Valence in the same Arousal.

Table 2. Examples of sentences recorded for each emotion.

	Features		Harmonic Feature Set	Accuracy (%)
	Positive	Negative		
High Arousal	Happiness	Anger	Noise energy, Noiseness, Inharmonicity, Tristimulus	65.5%
Low Arousal	Neutrality	Sadness	Noiseness, F0, Inharmonicity, Tristimulus	76%

We verified distinction in terms of accuracy as to the combination of features by utilizing Medium Gaussian support vector machines. Negative and positive emotions with high arousal were anger and happiness, respectively, and it yielded 65.5% in accuracy when we classified the two emotions by combining features of Noise energy, Noiseness, Inharmonicity, Tristimulus among harmonic features. Negative and positive emotions with low arousal were sadness and neutrality, respectively, and it yielded 76% in accuracy when we classified the two emotions by utilizing features of Noiseness, F0, Inharmonicity, Tristimulus among harmonic features. Through this result, it is clarified that Negative Valence and Positive Valence can be distinguished in the same Arousal when characteristics of Inharmonicity, Tristimuls, Harmonic energy, Noise energy, Noiseness are utilized in combination.

Features used for speech emotion recognition were also analyzed, including MFCC, LPC, Zero-Crossing Rate, Signal Energy, FFT Spectrum (magnitude), Mel Spectrum, Cepstral features, Pitch, Harmonic to Noise Ratio, Spectral features, Chroma and others, which have been widely used in previous studies. Through each feature's individual classification, we carried out an experiment to investigate features specialized in emotion recognition. In individual features, MFCC, in general, showed the highest accuracy (approximately 60%). We performed the analysis again, and redesigned feature sets by combining each feature according to its accuracy, starting with the highest, and after comparing the result with individual features, we could confirm that combined feature sets yielded improved result by 10% in average. Feature vector's combination, so called "basic feature combination" which shows the highest accuracy, is as shown below.

- 11 Spectral-domain: spectral centroid, spectral bandwidth, 7 spectral contrast, spectral flatness, and spectral roll-off
- 13 MFCCs
- 12 Chroma: 12-dimensional Chroma vector

Thus, seven harmonic feature vectors (inharmonicity, tristimulus, harmonic energy, noise energy, and noisiness) were added to the 36 individual feature vectors, making 43 feature vectors in total.

4.2. Pre-Processing

4.2.1. Speech Segment Extraction

The need for determining whether a given speech signal should be classified as a speech section or a silence section arises in many speech analysis systems. When non-speech sections are included in the learning or testing process, they can provide unnecessary information and become an obstacle. Since the signal energy value of the speech signal segment is larger than that of the non-speech signal segment, an absolute integral value (IAV) reflecting the energy value is used. The IAV value is computed by Equation (1)

$$\mathrm{IAV} = \sum\nolimits_{i=1}^{N} |X(i)|\,, \tag{1}$$

where X is the recorded signal, N is the number of samples, and i is the sample index.

The process of extracting speech segments is shown in Figure 3.

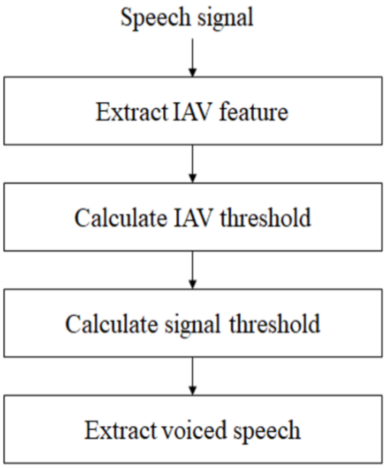

Figure 3. Flowchart of preprocessing.

The process of selecting the IAV threshold value is as follows. First, it is necessary to extract the IAV feature vector from the interval of the signal. Then, it is imperative to calculate the maximum value and the minimum value and determine the threshold value by 10% of the difference between these two values. An example of determining the threshold is shown in Figure 4.

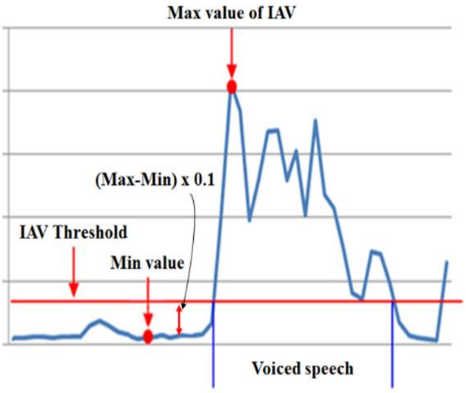

Figure 4. An example of determining threshold.

The signal threshold is computed to find the starting point of the signal in the window as well as the IAV threshold. The threshold of the signal is divided by the frame size at the IAV threshold value. As the IAV value is the absolute integral value of all the signal values in the window, the average signal value of the critical section can be obtained by dividing by the window size.

The process of extracting a speech interval includes a point at which the window is larger than the IAV value, and it determines a point at which the window is larger than the signal threshold value as a starting point. If an extracted IAV value is smaller than the IAV threshold, the end point is determined.

4.2.2. Feature Scaling

If the range of each attribute value of learning data differs greatly, the learning is not working efficiently. For example, if the range of the property A is 1 to 1000 and the range of the property B is 1 to 10, and if the value of A is larger, A is reflected as if it has a significant impact on the neural network, and B acts like it does not affect the network, relatively. Thus, transforming each property value into the same range is necessary before learning, and this process is referred as “feature scaling.” Among the various methods of feature scaling, a min–max scaling method that sets the range between 0 and 1 based on the maximum and minimum values within the range of each feature is most generally used. However, simply normalizing the range between 0 and 1 is not suitable for the feature used in this study, because the difference of each value will decrease excessively if the range changes into 0 to 1. In this study, we normalize the features using a standard-score method, which considers the range as well as the variation of the values. The formula for this scaling method is as follows (2)

$$x' = \frac{x - \overline{x}}{\sigma}, \tag{2}$$

where x' is the normalized feature vector, x is the input. $\overline{x}$ is the average of the input vector, and σ is the standard deviation of the input.

4.3. Emotion Recognition Model

This study used LSTM model to recognize emotional state from speech. LSTM is a modified RNN structure proposed by Hochreiter et al. [22], where hidden layer nodes are replaced by blocks. The LSTM block comprises memory cells with cyclic structure; and input, forget, and output gates. They differ from conventional RNN models in that LSTM blocks control flow by sending or not sending information appropriately using their gates depending on the specific circumstances. LSTM calculates the *i*-th value for each node at time *t* as

$$c_i(t) = tanh(r_i^c(t)), \qquad r_i^c(t) = \sum_j U_{ij}^c x_j(t) + \sum_j W_{ij}^c h_j(t-1) + b_i^c, \tag{3}$$

where $x_j(t)$ is the j-th input variable, $c_i(t)$ is the *i*-th input state, calculated by applying weights U, W, and b to hidden layer $h(t-1)$ at input times t and $t-1$. Equation (3) is applied to input, forget, and output gates g, f, and q, respectively, and the weighted sum of the constant term is converted into the range of 0 to 1 using a non-linear function. Gate output = 1 means that all information is maintained, whereas gate output = 0 means that all information is deleted.

Cell state $s_i(t)$ at t is calculated from $c_i(t)$ and gates g and f,

$$\mathrm{s}_i(t) = f_i(t)s_i(t-1) + q_i(t)c_i(t), \tag{4}$$

where the *i*-th hidden node is

$$h_i(t) = q_i \tan\mathrm{h}(s_i(t)), \tag{5}$$

Thus, the first input value cannot influence the next value when gates *f* and *g* = 1 and 0, respectively, hence the cell state never changes, and short-term memory can be removed; whereas long-term memory can be maintained in the opposite case.

Data from previous time(s) influence current data for most time-series models. Therefore, RNN models predict status for current time *t* using data from previous times. However, current status can often be predicted more accurately using future data. For example, prediction accuracy may improve for Korean language objects if the verb, which appears after the object, is utilized rather than the subject. Thus, inferring in forward as well as backward directions may produce more meaningful results.

Equation (4) shows that conventional LSTM uses data for past time *t*-1 time to predict the data for current time *t*. Backward connection does not exist so the current time cannot be predicted using future time data. Therefore, bi-directional LTSM models are commonly employed to resolve this drawback, as shown in Figure 5.

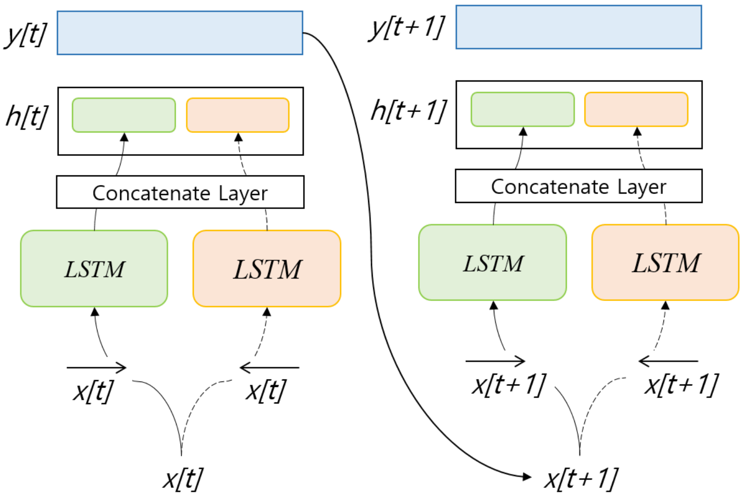

Figure 5. Example of Bi-directional LSTM network nodes.

This study used the feature vectors discussed in Section 4.1 and a bi-directional LSTM to recognize emotions from speech. Figure 6 shows the bi-directional LTSM model employed.

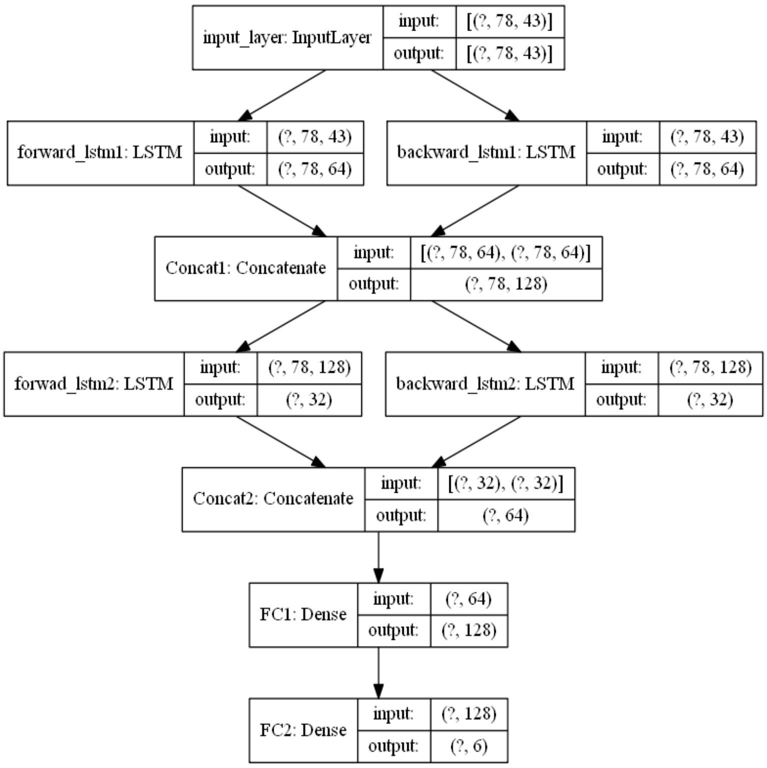

Figure 6. Proposed emotion recognition network structure.

Generally, emotion can be recognized when we hear the entire spoken sentence or phrase. However, it is very difficult to recognize emotion when we hear the speech into short segments (e.g., one second). Thus, emotion recognition accuracy generally improves for longer rather than shorter intervals. However, LSTM has a problem that a gradient is not delivered as the layer is deeper (meaning that speech is recognized at a longer interval). This study set LSTM interval as short as possible, and then calculated emotion probability over each section to decide the final emotion label, hence avoiding the gradient loss problem while still using longer speech intervals to recognize emotion.

Input speech data was divided into speech sections through the pre-processing procedure that searches the speech section as mentioned in Section 4.2.1. Combined feature vectors, discussed in Section 4.1, were extracted from the divided speech sections in chronological order, and overlapped as much as 0.3 for approximately 15 ms time units. Extracted feature vectors were re-configured into 78 vectors, emotion for each part was calculated approximately every second chronologically, and each frame was classified using the trained model. Using a voting mechanism on the classified frames, the final emotion label of the speech signal is detected.

5. Experiments and Results

Two experiments were conducted to compare the proposed method performance. The first experiment used the K-EmoDB database developed in Section 3. We chose six emotions for the experiment: neutral, anger, happiness, sadness, fear, and excitement. Randomly selected records (630 of 900, i.e., 70%) were used as training data, and the remaining 170 records as test data. We used 5-fold cross-validation, i.e., repeated the sample selection five times. The second experiment compared the proposed method with previous emotion recognition studies using various international databases. We used the various databases employed in the previous studies and compared accuracy directly and indirectly.

5.1. K-EmoDB

Table 3 shows emotion recognition results for K-EmoDB. Chernykh [33] achieved approximately 67.9% accuracy using 34 acoustic features and LSTM. Shaqr [30] employed eGeMAPS features and a multi-layer perceptron model to achieve 67.14% accuracy. The study by Zamil [58] employed only MFCC as the acoustic feature and achieved 61.32% using the logistic model tree. The studies by George [59] and Jianfeng [9] proposed an emotion recognition method by end-to-end learning using a convolution layer. With the recognition using 1D convolution layer and LSTM with the input of time-series signals of speech data, their accuracy was 65.89%, and with the recognition using the LSTM layer by combining 1D convolution layer and 2D convolution layer with the input of Mel-spectrogram, their accuracy was 62.63%. When the basic feature combination proposed in this study was used, the accuracy was 70.51%.

Table 3. Experimental results from K-EmoDB.

Paper	Features	Algorithm	Accurary
Chernykh (2017)	• 3 Time-domain: zero crossing rate, energy, entropy of energy • 5 Spectral-domain: spectral centroid, spectral spread, spectral entropy, spectral flux, spectral rolloff • 13 MFCCs • 13 Chroma: 12-dimensional chroma vector, standard deviation of chroma vector	LSTM with CTC loss	67.90%
Shaqr (2019)	• eGeMAPS feature set	Multi-layer perceptron	67.14%
Zamil (2019)	• MFCC	Logistic Model Tree	61.32%
George (2016)	-	1D Conv layer + LSTM	65.89%
Jianfeng (2019)	Log mel-spectrogram	1D + 2D Conv layer + LSTM	62.63%
Proposed	• Spectral: Centroid, Bandwidth, Contrast, Flatness, Rolloff • MFCC • 12 Chroma • Harmonic: Inharmonicity, Tristimuls, Harmonic energy, Noise energy, Noiseness	LSTM + Voting mechanism	83.81%
Proposed	• Spectral: Centroid, Bandwidth, Contrast, Flatness, Rolloff • MFCC • 12 Chroma • Harmonic: Inharmonicity, Tristimuls, Harmonic energy, Noise energy, Noiseness	LSTM	75.46%
Proposed	• MFCC • Spectral: Centroid, Bandwidth, Contrast, Flatness, Rolloff • 12 Chroma	LSTM	70.51%

When emotion recognition was performed by combining the basic feature combination and harmonic features, the accuracy was 75.46%, which improved around 5% compared to that not using the harmonic features. Moreover, when the final emotion of the speech data was produced by determining the label of each section at a shorter interval, high accuracy of 83.81% was obtained.

5.2. International DB

We also compared emotion recognition using emotion speech databases widely used in previous studies. Table 4 compares prediction accuracy using the databases from the respective papers. Although databases and emotions differed between the various studies, accuracies achieved approximately 60–70%. However, the proposed method achieved 85.79% accuracy for EmoDB, and 87.11% for the RAVDESS database.

Table 4. Experimental results from International DB.

Paper	Database	Emotion	Features	Algorithm	Accuracy
Chernykh (2017)	(a) EmoDB (b) RAVDESS	(a) anger, excitement, neutral, sad (b) Anger, Neutral, Happiness, Sadness, Fear	• 3 Time-domain • 5 Spectral-domain • 13 MFCCs • 13 Chroma:	LSTM with CTC loss	(a) 54% (b) 71.08%
Shaqr (2019)	RAVDESS	neutral, calm, happy, sad, angry, fearful, disgust, surprised	• eGeMAPS	Multi-layer perceptron	74%
Zamil (2019)	(a) EmoDB (b) RAVDESS	(a) Anger, Boredom, Disgust, Anxiety, Joy, Neutral, Sad (b) Disgust, Fear, Happy, Neutral, Sad, Anger, Calm	• MFCC	Logistic Model Tree	(a) 64.51% (b) 70%
George (2016)	RECOLA	Arousal, Valence	-	1D Conv layer + LSTM	65.89%
Jianfeng (2019)	(a) EmoDB (b) IEMOCAP	(a) Anger, Boredom, Disgust, Anxiety, Joy, Neutral, Sad (b) Angry, Excited, Frustrated, Happy, Neutral, Sad	Log mel-spectrogram	1D + 2D Conv layer + LSTM	(a) 76.64% (b) 62.07%
Proposed	(a) EmoDB (b) RAVDESS	(a) Anger, Neutral, Happiness, Sadness, Fear (b) neutral, calm, happy, sad, angry, fearful, disgust, surprised	• MFCC • Spectral • 12 Chroma • Harmonic	(a) LSTM + Voting mechanism (b) 2D CNN(Facial) + Proposed(Speech)	(a) 85.79% (b) 87.11%

In the case of the result using the RAVDESS database, we used two kinds of deep neural networks for classifying emotions from speech and images, respectively. The networks reflected temporal representations from sequential data of images and speech. We fine-tuned the softmax functions of the pre-trained networks, considering the characteristic of each input, to maximize the ability of the networks. Consequently, the proposed method showed more accurate results than other models. Thus, higher accuracy can be acquired by combining the proposed acoustic features with a FER model.

6. Conclusions

In this paper, we constructed a Korean emotion speech database for speech emotion analysis and proposed a feature combination that could improve the emotion recognition performance using an RNN model. We carried out speech emotion recognition using the emotional speech database. To investigate whether harmonic and dissonant speech intervals affect positive and negative emotions, respectively, we extracted and analyzed harmonic features closely related to harmonic and dissonant speech. The experiments showed that harmonic features have an effect on distinguishing valence, which increased around 5% compared to that not using the harmonic features. This study set the LSTM interval to be as short as possible, and then calculated emotion probability over each section to decide the final emotion label, hence avoiding the gradient loss problem while still using longer speech intervals to recognize emotion. It has been shown experimentally that recognizing emotions after dividing speech into short intervals produces better recognition accuracy (83.81%) than recognizing speech emotion using one long LSTM model (75.51%).

In future works, in order to accurately learn deep running-based models, a considerable amount of data is required. Studies into speech emotion recognition use far less data than studies based on videos or texts. It is necessary to build high-quality databases in order to improve performance and generalization. In order to recognize emotions from speech, it is necessary to find an accurate mapping of acoustic features and the intensity of emotions, using large databases. Although studies have used a variety of acoustic features, they still have the problem of not knowing what emotional intensity is manifested when a certain feature is found at a certain level. It is necessary to solve these problems in order to recognize emotions and their intensity more accurately.

Author Contributions: conceptualization, S.-W.B.; methodology, S.-W.B. and S.-P.L.; investigation, S.-W.B.; writing—original draft preparation, S.-W.B.; writing—review and editing, S.-P.L.; project administration, S.-P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data sharing is not applicable to this article.

Acknowledgments: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2019R1F1A1050052).

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

References

1. El Ayadi, M.; Kamel, M.S.; Karray, F. Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. *Pattern Recognit.* **2011**, *44*, 572–587. [CrossRef]
2. Shin, B.; Lee, S. A Comparison of Effective Feature Vectors for Speech Emotion Recognition. *Trans. Korean Inst. Electr. Eng.* **2018**, *67*, 1364–1369.
3. Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech Emotion Recognition using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Trans. Multimed.* **2017**, *20*, 1576–1590. [CrossRef]
4. Domínguez-Jiménez, J.A.; Campo-Landines, K.C.; Martínez-Santos, J.; Delahoz, E.J.; Contreras-Ortiz, S. A Machine Learning Model for Emotion Recognition from Physiological Signals. *Biomed. Signal Process. Control.* **2020**, *55*, 101646. [CrossRef]
5. Liu, M.; Li, S.; Shan, S.; Wang, R.; Chen, X. Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis. In Proceedings of the Asian Conference on Computer Vision, Singapore, 1–5 November 2014; pp. 143–157.
6. Xiong, X.; De la Torre, F. Supervised Descent Method and its Applications to Face Alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 532–539.
7. Jia, X.; Li, W.; Wang, Y.; Hong, S.; Su, X. An Action Unit Co-Occurrence Constraint 3DCNN Based Action Unit Recognition Approach. *KSII Trans. Internet Inf. Syst.* **2020**, *14*, 924–942.
8. He, J.; Li, D.; Bo, S.; Yu, L. Facial Action Unit Detection with Multilayer Fused Multi-Task and Multi-Label Deep Learning Network. *KSII Trans. Internet Inf. Syst.* **2019**, *13*, 5546–5559.
9. Zhao, J.; Mao, X.; Chen, L. Speech Emotion Recognition using Deep 1D & 2D CNN LSTM Networks. *Biomed. Signal Process. Control.* **2019**, *47*, 312–323.
10. Swain, M.; Routray, A.; Kabisatpathy, P. Databases, Features and Classifiers for Speech Emotion Recognition: A Review. *Int. J. Speech Technol.* **2018**, *21*, 93–120. [CrossRef]
11. Gilbert, C.; Hutto, E. Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14), Ann Arbor, MI, USA, 1–4 June 2014; p. 82. Available online: http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf (accessed on 20 April 2016).
12. Ma, Y.; Hao, Y.; Chen, M.; Chen, J.; Lu, P.; Košir, A. Audio-Visual Emotion Fusion (AVEF): A Deep Efficient Weighted Approach. *Inf. Fusion* **2019**, *46*, 184–192. [CrossRef]
13. Scherer, K.R. Vocal Communication of Emotion: A Review of Research Paradigms. *Speech Commun.* **2003**, *40*, 227–256. [CrossRef]
14. Lee, C.; Lui, S.; So, C. Visualization of Time-Varying Joint Development of Pitch and Dynamics for Speech Emotion Recognition. *J. Acoust. Soc. Am.* **2014**, *135*, 2422. [CrossRef]
15. Wu, C.; Yeh, J.; Chuang, Z. Emotion perception and recognition from speech. In *Affective Information Processing*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 93–110.

16. Lotfian, R.; Busso, C. Curriculum Learning for Speech Emotion Recognition from Crowdsourced Labels. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 815–826. [CrossRef]
17. Song, P.; Zheng, W. Feature Selection Based Transfer Subspace Learning for Speech Emotion Recognition. *IEEE Trans. Affect. Comput.* **2018**, *11*, 373–382. [CrossRef]
18. Jing, S.; Mao, X.; Chen, L. Prominence features: Effective emotional features for speech emotion recognition. *Digit. Signal Process.* **2018**, *72*, 216–231. [CrossRef]
19. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A Database of German Emotional Speech. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisboa, Portugal, 4–8 September 2005.
20. Livingstone, S.R.; Russo, F.A. 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* **2018**, *13*, e0196391. [CrossRef] [PubMed]
21. Busso, C.; Bulut, M.; Lee, C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Lang. Resour. Eval.* **2008**, *42*, 335. [CrossRef]
22. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
23. Bou-Ghazale, S.E.; Hansen, J.H. A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress. *IEEE Trans. Speech Audio Process.* **2000**, *8*, 429–442. [CrossRef]
24. Lee, C.M.; Yildirim, S.; Bulut, M.; Kazemzadeh, A.; Busso, C.; Deng, Z.; Lee, S.; Narayanan, S. Emotion Recognition Based on Phoneme Classes. In Proceedings of the Eighth International Conference on Spoken Language Processing, Jeju Island, Korea, 4–8 October 2004.
25. Schuller, B.; Rigoll, G.; Lang, M. Hidden Markov Model-Based Speech Emotion Recognition. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, China, 6–10 April 2003.
26. Lee, J.; Tashev, I. High-Level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
27. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic Speech Emotion Recognition using Recurrent Neural Networks with Local Attention. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2227–2231.
28. Chen, R.; Zhou, Y.; Qian, Y. Emotion Recognition using Support Vector Machine and Deep Neural Network. In Proceedings of the National Conference on Man-Machine Speech Communication, Lianyungang, China, 11–13 October 2017; pp. 122–131.
29. Wieman, M.; Sun, A. Analyzing Vocal Patterns to Determine Emotion. Available online: http://www.datascienceassn.org/content/analyzing-vocal-patterns-determine-emotion (accessed on 3 September 2016).
30. Shaqra, F.A.; Duwairi, R.; Al-Ayyoub, M. Recognizing Emotion from Speech Based on Age and Gender using Hierarchical Models. *Procedia Comput. Sci.* **2019**, *151*, 37–44. [CrossRef]
31. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.
32. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* **2015**, *7*, 190–202. [CrossRef]
33. Chernykh, V.; Prikhodko, P. Emotion Recognition from Speech with Recurrent Neural Networks. *arXiv* **2017**, arXiv:1701.08071.
34. Graves, A.; Fernandez, S.; Gomez, F.J.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural nets. In Proceedings of the 23rd International Conference on Machine Learning, New York, NY, USA, 25–29 June 2006; pp. 369–376.
35. Iliou, T.; Anagnostopoulos, C. Statistical Evaluation of Speech Features for Emotion Recognition. In Proceedings of the 2009 Fourth International Conference on Digital Telecommunications, Colmar, France, 20–25 July 2009; pp. 121–126.
36. Kao, Y.; Lee, L. Feature Analysis for Emotion Recognition from Mandarin Speech Considering the Special Characteristics of Chinese Language. In Proceedings of the Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006.
37. Luengo, I.; Navas, E.; Hernáez, I.; Sánchez, J. Automatic Emotion Recognition using Prosodic Parameters. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisboa, Portugal, 4–8 September 2005.
38. Rao, K.S.; Koolagudi, S.G.; Vempada, R.R. Emotion Recognition from Speech using Global and Local Prosodic Features. *Int. J. Speech Technol.* **2013**, *16*, 143–160. [CrossRef]
39. Schuller, B.; Batliner, A.; Steidl, S.; Seppi, D. Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge. *Speech Commun.* **2011**, *53*, 1062–1087. [CrossRef]
40. Schuller, B.; Steidl, S.; Batliner, A.; Vinciarelli, A.; Scherer, K.; Ringeval, F.; Chetouani, M.; Weninger, F.; Eyben, F.; Marchi, E. INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In Proceedings of the 14th Annual Conference of the International Speech Communication Association, Lyon, France, 25–29 August 2013; Interspeech: Lyon, France, 2013.
41. Russell, J.A. Is there Universal Recognition of Emotion from Facial Expression? A Review of the Cross-Cultural Studies. *Psychol. Bull.* **1994**, *115*, 102. [CrossRef]
42. Ortony, A.; Collins, T. What's Basic about Basic Emotions? *Psychol. Rev.* **1990**, *97*, 315–331. [CrossRef] [PubMed]

43. Barrett, L.F. Are Emotions Natural Kinds? *Perspect. Psychol. Sci.* **2006**, *1*, 28–58. [CrossRef] [PubMed]
44. Ekman, P. *Pictures of Facial Affect*; Consulting Psychologists Press: Palo Alto, CA, USA, 1976.
45. Lundqvist, D.; Flykt, A.; Ohman, A. *Karolinska Directed Emotional Faces*; Database of Standardized Facial Images; Psychology Section, Department of Clinical Neuroscience, Karolinska Hospital: Stockholm, Sweden, 1998; Volume S-171, p. 76.
46. Wang, L.; Markham, R. The Development of a Series of Photographs of Chinese Facial Expressions of Emotion. *J. Cross-Cult. Psychol.* **1999**, *30*, 397–410. [CrossRef]
47. Kanade, T.; Cohn, J.F.; Tian, Y. Comprehensive Database for Facial Expression Analysis. In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. no. PR00580), Grenoble, France, 28–30 March 2000; pp. 46–53.
48. Tottenham, N.; Tanaka, J.W.; Leon, A.C.; McCarry, T.; Nurse, M.; Hare, T.A.; Marcus, D.J.; Westerlund, A.; Casey, B.; Nelson, C. The NimStim Set of Facial Expressions: Judgments from Untrained Research Participants. *Psychiatry Res.* **2009**, *168*, 242–249. [CrossRef] [PubMed]
49. Tracy, J.L.; Robins, R.W.; Schriber, R.A. Development of a FACS-Verified Set of Basic and Self-Conscious Emotion Expressions. *Emotion* **2009**, *9*, 554. [CrossRef] [PubMed]
50. Langner, O.; Dotsch, R.; Bijlstra, G.; Wigboldus, D.H.; Hawk, S.T.; Van Knippenberg, A. Presentation and Validation of the Radboud Faces Database. *Cogn. Emot.* **2010**, *24*, 1377–1388. [CrossRef]
51. Simon, D.; Craig, K.D.; Gosselin, F.; Belin, P.; Rainville, P. Recognition and Discrimination of Prototypical Dynamic Expressions of Pain and Emotions. *PAIN* **2008**, *135*, 55–64. [CrossRef] [PubMed]
52. Castro, S.L.; Lima, C.F. Recognizing Emotions in Spoken Language: A Validated Set of Portuguese Sentences and Pseudosentences for Research on Emotional Prosody. *Behav. Res. Methods* **2010**, *42*, 74–81. [CrossRef] [PubMed]
53. Zhang, X.; Yin, L.; Cohn, J.F.; Canavan, S.; Reale, M.; Horowitz, A.; Liu, P.; Girard, J.M. Bp4d-Spontaneous: A High-Resolution Spontaneous 3d Dynamic Facial Expression Database. *Image Vis. Comput.* **2014**, *32*, 692–706. [CrossRef]
54. LoBue, V.; Thrasher, C. The Child Affective Facial Expression (CAFE) Set: Validity and Reliability from Untrained Adults. *Front. Psychol.* **2015**, *5*, 1532. [CrossRef] [PubMed]
55. Emotion Classification. Available online: https://en.wikipedia.org/wiki/Emotion_classification (accessed on 29 January 2021).
56. Gabrielsson, A.; Lindström, E. The Role of Structure in the Musical Expression of Emotions. In *Handbook of Music and Emotion: Theory, Research, Applications*; Series in affective science; Juslin, P.N., Sloboda, J.A., Eds.; Oxford University Press: Oxford, UK, 2010; pp. 367–400.
57. Bogdanov, D.; Wack, N.; Gómez Gutiérrez, E.; Gulati, S.; Boyer, H.; Mayor, O.; Roma Trepat, G.; Salamon, J.; Zapata González, J.R.; Serra, X. Essentia: An audio analysis library for music information retrieval. In Proceedings of the 14th International Society for Music Information Retrieval Conference, Curitiba, Brazil, 4–8 November 2013; pp. 493–498.
58. Zamil, A.A.A.; Hasan, S.; Baki, S.M.J.; Adam, J.M.; Zaman, I. Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames. In Proceedings of the 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), Dhaka, Bangladesh, 10–12 January 2019; pp. 281–285.
59. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu Features? End-to-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204.