applied sciences

MDPI

Article

Multi-Modal Emotion Recognition Using Speech Features and Text-Embedding

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

[1] Department of Computer Science, Graduate School, SangMyung University, Seoul 03016, Korea; 123234566@naver.com (S.-W.B.); hui2666@naver.com (J.-H.K.)
[2] Department of Electronic Engineering, SangMyung University, Seoul 03016, Korea
* Correspondence: esprit@smu.ac.kr

Abstract: Recently, intelligent personal assistants, chat-bots and AI speakers are being utilized more broadly as communication interfaces and the demands for more natural interaction measures have increased as well. Humans can express emotions in various ways, such as using voice tones or facial expressions; therefore, multimodal approaches to recognize human emotions have been studied. In this paper, we propose an emotion recognition method to deliver more accuracy by using speech and text data. The strengths of the data are also utilized in this method. We conducted 43 feature vectors such as spectral features, harmonic features and MFCC from speech datasets. In addition, 256 embedding vectors from transcripts using pre-trained Tacotron encoder were extracted. The acoustic feature vectors and embedding vectors were fed into each deep learning model which produced a probability for the predicted output classes. The results show that the proposed model exhibited more accurate performance than in previous research.

Keywords: speech emotion recognition; emotion recognition; multi-modal emotion recognition

check for updates

Citation: Byun, S.-W.; Kim, J.-H.; Lee, S.-P. Multi-Modal Emotion Recognition Using Speech Features and Text-Embedding. *Appl. Sci.* **2021**, *11*, 7967. https://doi.org/10.3390/app11177967

Academic Editor: Arcangelo Castiglione

Received: 27 May 2021
Accepted: 26 August 2021
Published: 28 August 2021

1. Introduction

Recently, intelligent personal assistants, chat-bots, and AI speakers have been widely utilized as communication interfaces, leading to increased demand for more natural means of interaction. It is necessary to understand human emotions in order to improve human–machine interface (HMI) systems. Humans can express emotions in various ways, including by using voice tones or facial expressions; therefore, multimodal approaches for the recognition of human emotions have been used [1–6]. Emotion refers to a conscious mental reaction subjectively experienced as a strong feeling, typically accompanied by physiological and behavioral changes in the body [7]. To recognize a user's emotional state, there are many ways for machines to recognize emotions. They can use speech [8,9], facial expressions [10], or body signals such as ECG [11]. With the recent development of deep-learning algorithms, emotion recognition technologies based on DNN or CNN using high-level features from raw audio signals are currently under investigation [12,13]. However, methods using single signals for recognizing emotions have relatively low accuracy. Studies of multimodal technologies using more than two signal inputs are actively under way [14,15].

In this paper, we propose an emotion recognition method using speech and text data. The strengths of each type of data are utilized in this method. We used 43 feature vectors, including spectral features, harmonic features, and the Mel-frequency cepstrum coefficients (MFCC) from speech datasets. A total of 256 embedding vectors from transcripts obtained using the pre-trained Tacotron encoder [16] were also extracted. First, we extracted 43 feature vectors from the speech datasets and detected the probability of each emotion using these vectors with a softmax function, which consisted of a long short-term memory (LSTM) and fully connected layers. Then, after tokenizing the text data into consonants and vowels, the text-embedding vectors of 256 dimensions were extracted by embedding each token using a pre-trained Tacotron encoder. The extracted embedding vectors were input into the LSTM and fully connected layers, and the probability of each emotion was deduced by the final softmax layer in the same way as performed using the speech datasets. Finally, the average values for each emotion were calculated using the results of each softmax function from the speech and text data. Then, the appropriate emotion was selected, based on the emotion with the highest average value. To assess the performance of the emotion recognition model, Korean language emotion speech datasets were established. Transcripts of scenes from media such as TV dramas or movies were categorized according to different emotions. The transcripts were recorded by two professional actors and two professional actresses. The emotions used in the experiments fell into four categories: anger, happiness, sadness, and neutral, as used in many emotion studies. The proposed model produced more accurate performance than in previous research.

This paper is organized as follows. Section 2 introduces the database used in this research. Section 3 explains the proposed emotion recognition method. Section 4 presents the experiment results, which are then discussed and concluded with Section 5.

2. Dataset

2.1. Korean Emotional Speech Dataset

Well-defined speech databases are needed to accurately recognize and analyze emotions from data; therefore, many studies have tried to construct speech databases, and a growing number of speech emotion databases are now available. They include about 800 to 1000 data points and focus on the availability of validated and reliable expressions of emotion [17–19]. Additionally, among datasets, few contain audio–text recordings of speakers. This study considered collecting a large amount of emotional speech data rather than the emotional quality of the data. Therefore, unlike general speech databases, we did not conduct evaluation tests. This study constructed a Korean emotional speech database and reported validity and reliability of the data based on ratings from participants. The database was recorded with Korean utterances from professional actors. All speech data were recorded in a professional studio, considering the sound quality of the data by eliminating any background noise. The database was recorded using Korean scripts from dramas and movies, with four professional actors, two females and two males. The scripts were collected from actual scripts used in dramas and movies. The scripts consisted of 30 s long conversations between a male and a female per emotional scene. We defined four emotions: anger, happiness, neutrality, and sadness. The scripts consisted of 120 scenes per emotion. This database relied on professional actors and senior students from the Drama Department. Actors spoke Korean as their first language, spoke with a neutral Seoul accent, and did not possess any distinctive features. The actors recorded their voices from around 30 cm away from the microphone (Figure 1).

Figure 1. Two actors during recording, using a conversation between a male and a female.

After recording, the data were manually segmented by dialog turn (speaker turn) from continuous segments in which the actors were actively speaking. The scripts were segmented into sentences in advance, and used as references with which to split the data. The final audio recordings were divided into approximately 3–10 s long files. Each file was around two to three hours' worth per actor, producing a total of around 4000–5000 files. The total size of the database was 18,324 files, as shown in Table 1. All speech emotion data were recorded at 48 kHz and downsampled with a 24 kHz sampling rate in PCM signed 16-bit format.

Table 1. The amount of data per emotion category.

Emotion	Amount of Data
Anger	4299
Happiness	3675
Sadness	3897
Neutral	6453
Total	18,324

2.2. Pre-Processing

The need for determining whether a given speech signal should be classified as a voiced speech section or a silent 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. The signal energy value of the speech signal segment was larger than that of the non-speech signal segment; therefore, an absolute integral value (IAV) reflecting the energy value was used. The IAV value was computed by Equation (1) [20]:

$$\overline{X} = \sum_{i=1}^{N} |X(i\Delta t)| \quad (1)$$

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

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 a 10% difference between these two values. An example of determining the threshold is shown in Figure 2.

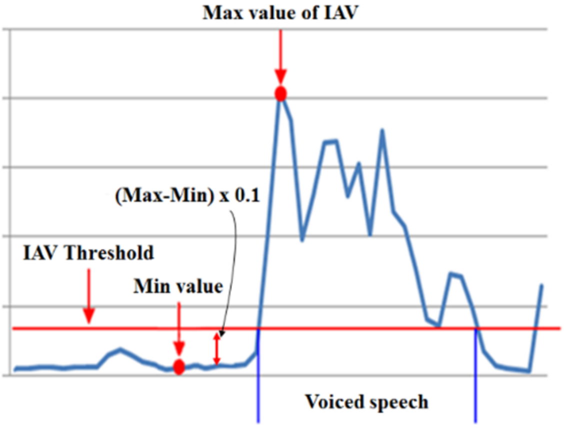

Figure 2. An example of determining threshold.

The signal threshold was computed to find the starting point of the signal in the window as well as the IAV threshold. The threshold of the signal was divided by the frame size at the IAV threshold value. The IAV value is the absolute integral value of all the signal values in the window; therefore, 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.

3. Proposed Method

3.1. Speech-Based Model

Verbal components convey one-third of human communication; therefore, using speech signals and text data simultaneously can be helpful in accurate and natural recognition [21,22]. Thus, we constructed a practical feature set to improve emotion recognition accuracy, complementing a text-based model. We surveyed acoustic features used for many speech-emotion recognition studies, and composed an optimal feature set by analyzing and combining the correlation of each feature. According to a previous study [23], harmonic structures expressed peaceful and positive emotions, such as happy, comfortable, and other, whereas dissonant intervals were closely related to sad and negative emotions. To quantify the harmonic structures, we extracted the harmonic features using the feature extraction tool Essentia, which is an open-source library tool for audio and music analysis; a detailed description of the harmonic features can be found in [24]. We used harmonic features reflecting the harmony of speech, which have not been extensively used in previous studies. We selected specialized features for emotion recognition through individual analysis, and then found the optimal feature set by re-combining features. The analysis showed that the harmonic feature set influenced the distinguishing of emotions, increasing the accuracy compared to that when not using the harmonic features.

In total, 43 features were extracted and used in this study:

- 13 MFCCs;
- 11 spectral-domain: spectral centroid, spectral bandwidth, seven spectral contrasts, spectral flatness, and spectral roll-off;
- 12 chroma: 12 dimensional chroma vectors;
- 7 harmonic features: inharmonicity, three tristimulus, harmonic energy, noise energy, and noisiness.

After windowing the speech signals, the signals were converted to acoustic features, and the features were input into the LSTM layers. The output of the LSTM layer was connected with the fully connected layers, and the last layer inferred the probability of each emotion through the softmax function. The whole speech-based model is illustrated in Figure 3. Weight-decay and dropout methods were used for regularization.

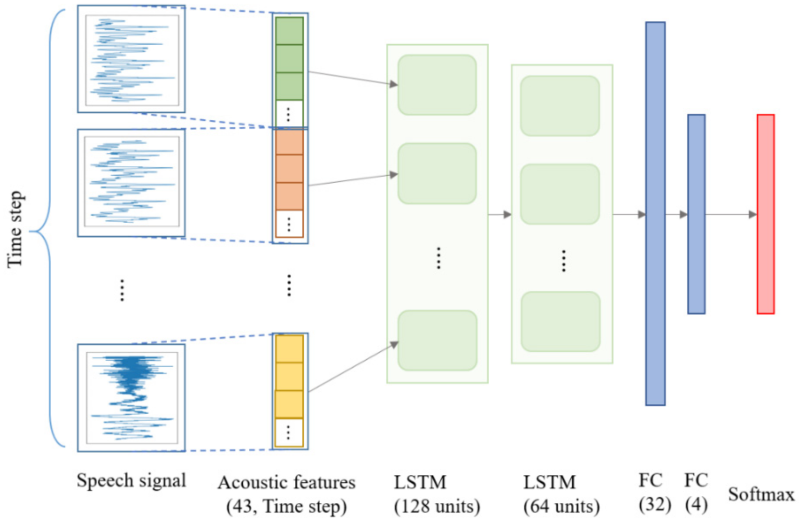

Figure 3. Structure of the model with acoustic features from speech data.

3.2. Text-Based Model

A text model was developed in order to process the text information of speech. We were using Korean language data; therefore, the model performed text analysis recognizing the characteristics of the Korean language. In English, when words are collected, the number of words is saturated at a specific level, based on the spaces between words. Therefore, English language can be tokenized using each word as a unit. However, because Korean is an agglutinating language, tokenized by word, the number of tokens increases indefinitely, and it is difficult to analyze the language in syllable-sized units. In this study, texts were therefore restructured into a sequence of onset, nucleus, and coda. For example, a word, "음성 [umsəŋ]" will be deconstructed into the consonants and vowels, " ㅇ ㅡ ㅁ ㅅ ㅓ ㅇ," and each separated unit will be embedded in a certain vector. The consonants of the onset and coda have different meanings; therefore, they were matched with different vectors. The language was categorized into 80 tokens, and each token was entered into a pre-trained Tacotron's encoder in the form of character embedding. Tacotron [16] is a Text-To-Speech (TTS) synthesis model developed by Google, described in Figure 4.

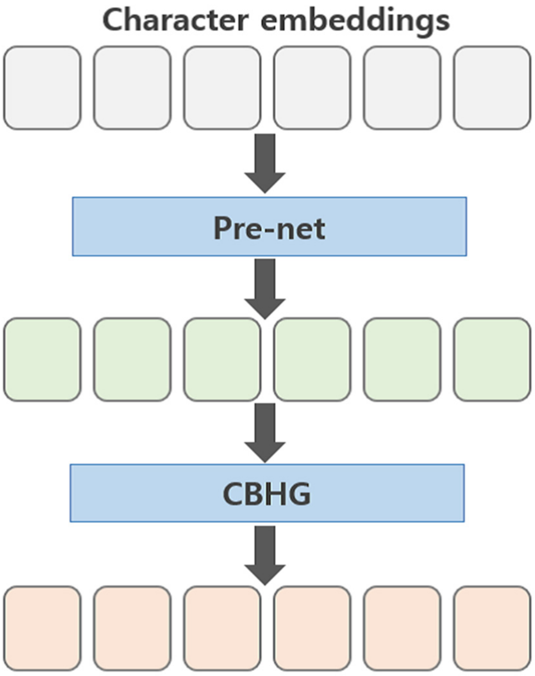

Figure 4. The architecture of Tacotron encoder [16].

The encoder creates a 256-dimensional embedding vector through the Pre-net and CBHG layers by taking the data of each consonant and vowel which is character-embedded after it has tokenized the text into syllables. The pre-net consists of a fully connected layer and a dropout layer, and the Convolution Bank Highway GRU (CBHG), consisting of a 1D convolution bank, a highway network, and a bidirectional GRU. The CBHG layer combines a CNN, which can extract abstracted features, and an LSTM, which is suitable for understanding the characteristics of time-series data. The CBHG layer also uses a Highway network in order to effectively express character units by extracting high-level features. The Highway network is a Residual network that utilizes a Gating structure. A model, through Gating, automatically makes decisions about what rate of Residual it should set. By converting or passing input signals, a model can be optimized, although the network becomes deeper. At this point, an encoder, pre-trained with the KSS dataset [25], a high-capacity corpus, is used. The generated embedding vector passed through the LSTM layer, which computes the probability values for each emotion through the softmax function in fully connected layer. To recognize emotions using speech and text data simultaneously, the model was improved in capacity to be a multi-modal emotion recognition model by combining a speech signal-based model and a text-based model. The structure of this model is shown in Figure 5.

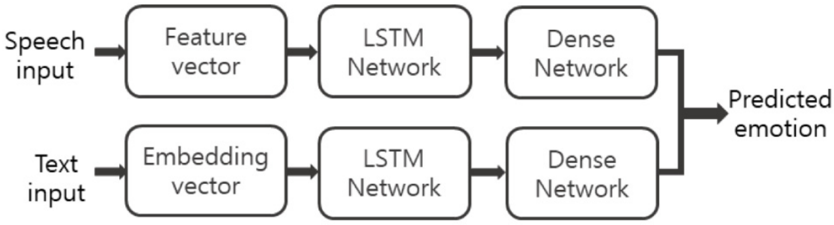

Figure 5. The architecture of the proposed model.

Speech and text were each processed through the speech signal-based model and the text-based model to create a 43-dimensional feature vector and a 256-dimensional text-embedding vector. Each generated vector passed through the LSTM layer of each model and computed probability values for each emotion using the softmax function in the fully connected layer. Then, by calculating the average value from the probability values of the speech and text for each emotion, the input was categorized into the emotion with the highest value.

4. Experiments

In order to evaluate capabilities of the system proposed in this paper, capability comparison proceeded between the system and other emotion recognition models from previous studies, using audio and text data.

Yoon [26] proposed a deep dual recurrent encoder model to proceed speech emotion recognition that utilized text and audio at the same time. After audio and text information was encoded using Audio Recurrent Encoder (ARE) and Text Recurrent Encoder (TRE), the emotion was predicted by combining encoded data in a fully connected layer. To extract speech signal's features, a 39-dimensional MFCC feature set and 35-dimensional prosodic feature set were extracted using the OpenSMILE toolkit [27]. The contents of each feature's sets are shown as follows:

- MFCC features: 12 MFCCs, log-energy parameter, 13 delta, 13 acceleration coefficient;
- Prosodic features: F0 frequency, voicing probability, loudness contours.

To analyze text, each sentence was tokenized into words and indexed to form sequences using the Natural Language Toolkit [28]. Each token was used to create a 300-dimensional embedding vector utilizing Glove [29], a pre-trained word-embedding vector. Lastly, by connecting ARE and TRE to fully connected layers, emotions were categorized. Atmaja [30] proposed a method to categorize emotions which uses speech feature extraction and word-embedding. In order to proceed with emotion recognition using two datasets simultaneously, a speech model that created feature vector by extracting speech feature and a text model that executed word-embedding by tokenizing text were designed, and they were set to perform their roles against two data points independently. The types of extracted features are as below:

- 13 MFCCs;
- 3 time domain features: zero crossing rate, energy, entropy of energy;
- 13 Chroma;
- 5 spectral domain features: spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off.

With the extracted feature vectors, emotions were categorized through fully connected layers. Each sentence was tokenized into words to convert text to embedding vectors. Each token was converted to a 300-dimensional embedding vector with the use of Glove. The generated embedding vector proceeded with dot-product calculation with one-hot encoding vector, thus preventing the model from over-fitting a certain word in a sentence.

The embedding vector categorized emotions through the LSTM layer, and audio and text models were connected as fully connected layers.

Pepino [31] proposed different approaches for classifying emotions from speech using acoustic- and text-based features. They obtained contextualized word embeddings with BERT to represent the information contained in speech transcriptions and combined the audio and text modalities. To extract the features of speech signals, pitch, jitter, shimmer, log harmonic to noise ratio, loudness, and 13 MFCC features were extracted using OpenSMILE toolkit.

To combine audio and text information, the fusion model consisted of two parallel branches processing audio and text separately up to a layer where the information from the two branches was merged.

All models were evaluated using the Korean language speech database that was mentioned in Section 2. Therefore, as for the text model, a preparatory process relevant to Korean language proceeded. Unlike the English language, where the morphemes are separated by spacing among words, in the Korean language, it is impossible to analyze morphemes by a word as a unit because of the usage of particles. As a result, each morpheme should be categorized after separating particles from each word; therefore, we tokenized sentences into morphemes in the Korean language. For each token, a 100-dimensional embedding vector was processed using the Korean-based Glove model and BERT. Table 2 shows a recognition rate of audio, text, and multi-modal model of the model in evaluation.

Table 2. Accuracy comparison with other papers.

Model	**Amount of Data**		
	Speech	**Text**	**Multi-Modal**
Yoon [26]	89.13%	63.68%	91.35%
Atmaja [30]	92.67%	65.98%	93.34%
Pepino [31]	75.26%	68.11%	81.31%
Proposed	94.86%	68.42%	95.97%

In the audio model, the proposed model showed the highest recognition rate, with 94.86%. Additionally, in the text model, the proposed model shows the highest recognition rate, with 68.42%. Additionally, in the integrated emotion recognition model, it resulted in 95.97%, which confirms that the proposed model also shows the best capability among previous multi-modal emotion recognition models.

5. Conclusions

As intelligent personal assistants, chat-bots, and AI speakers are rapidly popularized, these approaches can be applied to human interfaces to precisely recognize the user's emotional state and to provide personalized media according to the user's emotions. They can also develop a better understanding of human emotions which could improve human–machine interaction systems. In terms of perceptual and cognitive sciences, using speech signals and text data simultaneously can be helpful in accurate and natural recognition. However, because the characteristics of the methods to recognize emotions from speech and text sequences are different, combining the two inputs is still a challenging issue in the area of emotion-recognition. Additionally, to accurately train deep learning models, considerable amounts of data are required; however, previous speech emotion recognition studies used far less data than other approaches. In this paper, we discussed an emotion recognition method which increased accuracy by using both speech and text data, utilizing the strength of each type of data. To do this, we extracted features from both speech and text data. We extracted 43 feature vectors from the speech datasets, and the probability of each emotion was deduced using these vectors through the softmax function, which consisted of a long short-term memory (LSTM) and fully connected layers. Then, after tokenizing the text data into consonants and vowels, text-embedding vectors of 256 dimensions were extracted by embedding each token using a pre-trained Tacotron encoder. The extracted embedding vectors entered the input of the LSTM and fully connected layers, and the probability of each emotion was deduced by the final softmax layer in the same way as the speech datasets. Lastly, the average values for each emotion were calculated with the results of each softmax function from the speech and text data. Then, the desired emotion was selected based on the emotion that had the highest average value. We also established a Korean language emotion speech dataset. Transcripts of scenes from media such as TV dramas or movies were categorized according to different emotions. The emotions used in the experiments fell into four categories: anger, happiness, sadness, and neutral, as used in many emotion studies. The proposed model had the highest accuracy, 95.97%, in recognizing emotions.

The number of studies using Korean language corpora and embedding is conspicuously smaller than those using English-based natural language processing. In the future, studies using Korean text processing and embedding can be expected to combine accurate emotion recognition technology and communication interface technology by utilizing both speech and text information.

Author Contributions: Conceptualization, S.-W.B.; methodology, J.-H.K. and S.-P.L.; investigation, J.-H.K.; 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 Sangmyung University.

Institutional Review Board Statement: Not applicable.

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

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported by the 2021 Research Grant from Sangmyung University.

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

References

1. Liu, M.; Li, S.; Shan, S.; Wang, R.; Chen, X. Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis. In *Asian Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 143–157.
2. 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.
3. 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.
4. 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. (TIIS)* **2019**, *13*, 5546–5559.
5. 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]
6. Hutto, C.; Gilbert, E. Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In Proceedings of the International AAAI Conference on Web and Social Media. Available online: https://ojs.aaai.org/index.php/ICWSM/article/view/14550 (accessed on 31 December 2018).
7. Byun, S.; Lee, S. Human Emotion Recognition Based on the Weighted Integration Method using Image Sequences and Acoustic Features. *Multimed. Tools Appl.* **2020**, 1–15. [CrossRef]
8. Jin, Q.; Li, C.; Chen, S.; Wu, H. Speech Emotion Recognition with Acoustic and Lexical Features. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 4749–4753.
9. Kumbhar, H.S.; Bhandari, S.U. Speech Emotion Recognition using MFCC Features and LSTM Network. In Proceedings of the 2019 5th International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 19–21 September 2019; pp. 1–3.
10. Jain, N.; Kumar, S.; Kumar, A.; Shamsolmoali, P.; Zareapoor, M. Hybrid Deep Neural Networks for Face Emotion Recognition. *Pattern Recog. Lett.* **2018**, *115*, 101–106. [CrossRef]
11. Shin, D.; Shin, D.; Shin, D. Development of Emotion Recognition Interface using Complex EEG/ECG Bio-Signal for Interactive Contents. *Multimed. Tools Appl.* **2017**, *76*, 11449–11470. [CrossRef]
12. Zhao, J.; Mao, X.; Chen, L. Speech Emotion Recognition using Deep 1D & 2D CNN LSTM Networks. *Biomed. Signal Process. Control.* **2019**, *47*, 312–323.

13. Han, K.; Yu, D.; Tashev, I. Speech Emotion Recognition using Deep Neural Network and Extreme Learning Machine. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singerpore, 14–18 September 2014.
14. Ko, K.; Sim, K. Development of Context Awareness and Service Reasoning Technique for Handicapped People. *J. Korean Inst. Intell. Syst.* **2009**, *19*, 34–39. [CrossRef]
15. Huang, Y.; Yang, J.; Liao, P.; Pan, J. Fusion of Facial Expressions and EEG for Multimodal Emotion Recognition. *Comput. Intell. Neurosci.* **2017**, *2017*, 2107451. [CrossRef] [PubMed]
16. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S. Tacotron: Towards End-to-End Speech Synthesis. *arXiv* **2017**, arXiv:1703.10135.
17. 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, Lisbon, Portugal, 4–8 September 2005.
18. 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]
19. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The enterface'05 audio-visual emotion database. In Proceedings of the 22nd International Conference on Data Engineering Workshops, Atlanta, GA, USA, 3–7 April 2006.
20. Byun, S.; Lee, S. Emotion Recognition using Tone and Tempo Based on Voice for IoT. *Trans. Korean Inst. Electr. Eng.* **2016**, *65*, 116–121. [CrossRef]
21. Mehrabian, A. Communication without words. *Psychol. Today* **1968**, *2*, 53–56.
22. Kaulard, K.; Cunningham, D.W.; Bülthoff, H.H.; Wallraven, C. The MPI facial expression database—A validated database of emotional and conversational facial expressions. *PLoS ONE* **2012**, *7*, e32321.
23. Byun, S.; Lee, S. A Study on a Speech Emotion Recognition System with Effective Acoustic Features Using Deep Learning Algorithms. *Appl. Sci.* **2021**, *11*, 1890. [CrossRef]
24. Essentia. Available online: https://essentia.upf.edu/index.html (accessed on 31 December 2018).
25. Park, K. KSS Dataset: Korean Single Speaker Speech Dataset. Available online: https://kaggle.com/bryanpark/korean-single-speaker-speech-dataset/ (accessed on 31 December 2018).
26. Yoon, S.; Byun, S.; Jung, K. Multimodal Speech Emotion Recognition using Audio and Text. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 112–118.
27. 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.
28. Loper, E.; Bird, S. Nltk: The Natural Language Toolkit. *arXiv* **2002**, arXiv:cs/0205028.
29. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
30. Atmaja, B.T.; Shirai, K.; Akagi, M. Speech Emotion Recognition using Speech Feature and Word Embedding. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 519–523.
31. Pepino, L.; Riera, P.; Ferrer, L.; Gravano, A. Fusion Approaches for Emotion Recognition from Speech Using Acoustic and Text-Based Features. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020.