applied sciences

MDPI

Article

Make Patient Consultation Warmer: A Clinical Application for Speech Emotion Recognition

Huan-Chung Li [1], Telung Pan [2], Man-Hua Lee [1] and Hung-Wen Chiu [1,*]

[1] Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei 11031, Taiwan; wiselyli@iii.org.tw (H.-C.L.); m610107008@tmu.edu.tw (M.-H.L.)
[2] Bachelor Program in Interdisciplinary Studies, National Yunlin University of Science & Technology, Yunlin 64002, Taiwan; telung@yuntech.edu.tw
* Correspondence: hwchiu@tmu.edu.tw

Abstract: In recent years, many types of research have continued to improve the environment of human speech and emotion recognition. As facial emotion recognition has gradually matured through speech recognition, the result of this study provided more accurate recognition of complex human emotional performance, and speech emotion identification will be derived from human subjective interpretation into the use of computers to automatically interpret the speaker's emotional expression. Focused on use in medical care, which can be used to understand the current feelings of physicians and patients during a visit, and improve the medical treatment through the relationship between illness and interaction. By transforming the voice data into a single observation segment per second, the first to the thirteenth dimensions of the frequency cestrum coefficients are used as speech emotion recognition eigenvalue vectors. Vectors for the eigenvalue vectors are maximum, minimum, average, median, and standard deviation, and there are 65 eigenvalues in total for the construction of an artificial neural network. The sentiment recognition system developed by the hospital is used as a comparison between the sentiment recognition results of the artificial neural network classification, and then use the foregoing results for a comprehensive analysis to understand the interaction between the doctor and the patient. Using this experimental module, the emotion recognition rate is 93.34%, and the accuracy rate of facial emotion recognition results can be 86.3%.

Keywords: doctor-patient communication; speech emotion recognition; Mel Frequency Cepstrum Coefficients

check for updates

Citation: Li, H.-C.; Pan, T.; Lee, M.-H.; Chiu, H.-W. Make Patient Consultation Warmer: A Clinical Application for Speech Emotion Recognition. *Appl. Sci.* **2021**, *11*, 4782. https://doi.org/10.3390/app11114782

Academic Editor: José Machado

Received: 13 April 2021
Accepted: 19 May 2021
Published: 23 May 2021

1. Introduction

With technological advancement, artificial intelligence (AI) has been applied to process repetitive and complicated tasks and to predict future scenarios in a broad scope of industries. Currently, AI is employed in numerous medical areas, such as precision medicine, breast cancer imaging diagnosis, and medical care. In medical care, the behavior of medical personnel profoundly affects the emotional state of patients. AI can free up a physicians' cognitive and emotional space for their patients, and shift the focus away from transactional tasks to personalized care and medical service [1–3]. Because patients are vulnerable physiologically and psychologically, medical personnel should endeavor to facilitate positive feelings in patients to enable the smooth completion of disease diagnosis, treatment, and prevention processes, thereby improving the communication quality between the patient and the medical staff.

Because emotions enable people to better understand each other, efforts have been made to enable machines to understand human emotions. Smart mobile devices, which incorporate speech recognition, can receive and respond to voice commands with synthesized speech. Speech emotion recognition (SER) enables these devices to detect user emotions, and SER has been adopted for more than 20 years [4,5] in the fields of human-machine interactions [6], emotion integration in robots [7], computer games [8], and psychological assessment [9].

Although emotion recognition has been adopted in a broad way to daily living, emotions involve subjective perceptions making the identification of human emotions a challenge. The performance could not be improved beyond a certain limit as hand-crafted features are not very capable of capturing complex phenomena such as emotional state [10]. Over the past decade, deep learning techniques have been employed in several fields of research such as image classification, speaker recognition, speech and handwriting recognition, etc. because of their ability to capture high-level discriminative features which are otherwise not captured well in traditional handcrafted features. SER requires the use of a classifier, which usually employs a supervised learning algorithm in training a module to identify the emotion of a new voice signal. Supervised learning is crucial in emotion labeling. Voice signals must be preprocessed to highlight their features for extraction. These features consist of prosodies, frequency, and voice quality. In recent studies, AI in data analysis and classifiers integrated with deep learning algorithms have been prevalent [10–12].

Frequency features are frequently examined as eigenvalues in voice recognition and have also been jointly investigated with prosodies in experiments. Human voices are filtered by sound channels when they are generated and, after filtering, the resulting voices are defined by their shapes. Any voice exhibits notable frequency features [13]. Frequency features are acquired by converting time-domain signals into frequency-domain signals through Fourier transform. Specifically, a speech is divided into blocks, and each block is converted to frequency-domain signals through discrete short-time Fourier transform. A mel-scale filter bank is applied to calculate the energy of sub-bands. The logarithms of the sub-bands are then calculated. Finally, inverse Fourier transform is conducted to acquire mel-frequency cepstrum coefficients (MFCCs), which are the most widely used frequency features [14]. Numerous studies on MFCCs have involved comparing prosodic features with spectral features and have shown that spectral features exhibit remarkable results [15]. Therefore, this study employed MFCCs as eigenvalues for voice signals. Voice quality comprises the physical properties of sound, such as jitter, shimmer, and harmonics to noise ratios.

MFCCs, which were proposed by Davis and Mermelstein in the 1980s [16], have been universally applied in automatic speech and speaker recognition. Before MFCCs were introduced, linear predictive coefficients and linear prediction cepstral coefficients were the primary eigenvalues employed in automatic speech recognition.

Artificial neural networks (ANNs) have been frequently used to solve classification problems by the structure consists of an input layer, an output layer, and multiple hidden layers. These layers are composed of nodes. The number of nodes in the input and output layers depend on the types of data representation marks, and the number of nodes in the hidden layers varies according to user needs. Each layer is connected to the next according to an initial, randomly selected weight. After a sample is selected from the training data set, the sample's value is loaded into the input layer and forwarded to the next layer. In the output layer, the weight is updated through backpropagation. Once the training is complete, new data are classified according to their weights.

Emotional voice signals can be identified from a set of retrieved data and employed MFCCs to identify emotions from the voice data [17]. The emotions were categorized as happiness, anger, or sadness. Many algorithms have been designed for feature extraction and speech signal testing. Likitha et al. [17], developed function-based MFCCs for the extract of features from voice samples, and the standard deviation of each emotion category was defined according to the means of MFCCs. In their research, SER can attain 80% accuracy based on 60 experiments. Lalitha, Geyasruti, Narayanan, and Shravani (2015), combined mel-frequency cepstrum eigenvectors with an ANN classifier to achieve an 85.7% classification accuracy [18]. However, voice sample training and test rates should be further improved to overcome erroneous classification.

The goal of this study was to develop an SER instrument using ANN and speech processing technology and to integrate other instruments to identify medical physicians' and patients' emotions in clinics.

2. Materials and Methods

2.1. Experimental Database

Emotions are subjective in concepts and have to be digitalized based on text descriptions for the future processing of machine learning. Studies have endeavored to explore the types of emotions in human beings and to analyze whether dimensional or label classification is more effective in debriefing emotional experience and classifying emotions into rough, noncontinuous emotional experiences [19]. Due to the special multilingual environment in Taiwan, a Taiwanese language emotional speech database was established [20]. The voices of patients were also included in the voice samples for each doctor, an emotion training model was constructed by the sample database. Four types of emotions were labeled and described as neutral, angry, happy, or sad in the 610 emotional voice samples. Then, 315, 100, 90, and 105 records were identified as neutral, angry, happy, and sad, respectively, in the measurement unit of seconds.

Before the experiment, the patient's consent was obtained, and the experiment analysis did not reveal the names and personal information of the participants (physicians and patients) to ensure the privacy and personal information of the participants.

Audio data were divided into blocks of seconds starting from the beginning of each sample as an observation unit during the identification process. An end block of less than 1 s in length was abandoned.

2.2. Audio Identification and Emotion Labeling

The recorded audio files collected from clinics were stored in iMovie format, audio files are in WAVE format (Waveform Audio File Format), 16 KHz, 16-bit, mono. Label files are ASCII text files, store with phoneme level and syllable levels data. The original audio files were not recorded separately, audio interference unrelated to the doctors and patients was present. The identities of the speakers had to be preliminarily identified and labeled manually with each second as a unit for matching the files with facial emotion recognition results. In this study, only the voice samples of doctors and patients were acquired, blocks without voice signals were eliminated, and emotional labels were manually added.

2.3. Feature Extraction Using MFCC (Mel Frequency Cepstral Coefficient)

The applicability of MFCCs in SER has been verified [21]. Therefore, MFCCs were employed in this study as eigenvalues, and the observation unit was defined as 1 s. In voice processing, mel-frequency cepstrum is the linear transformation of nonlinear mel-scale logarithmic energy spectra based on sound frequency. MFCCs are coefficients that constitute mel-frequency cepstra [17,18,22] and are derived from the cepstra of voice signal blocks. Mel-frequency cepstra, whose frequency bands are divided equidistantly on the mel scale, emulate the human hearing system more closely than do normal logarithmic cepstra, which feature linearly spaced frequency bands. The nonlinear frequency band division improves the expression of sound signals in multiple fields, such as audio compression.

The voice samples were pre-emphasized before feature extraction to highlight the high-frequency portions. A high pass filter was employed to highlight the high-frequency formants in the samples. The samples were then blocked as several observation units called "frames" (0.03 × sampling rate per frame). To prevent excessive differences between two neighboring frames, an overlapping zone was set between the frames (0.02 × sampling rate), thereby enhancing the continuity between the two frames. Each frame was multiplied by a Hamming window to further enhance the continuity between its neighboring frames. The signals multiplied by the Hamming window exhibited notable formants in fast Fourier transform.

If the framed signals are S(n), n = 0, . . . , N − 1, then the framed signals multiplied by the Hamming window are denoted as S′(n) = S(n) × W(n); particularly, W(n) is expressed as W(n, a) = (1 − a) − a cos(2pn/(N − 1)), 0 ≤ n ≤ N − 1. Each a value generates a different Hamming window. Generally, a is defined as 0.46. Because the features of audio signals are difficult to detect in the time domain, they are usually converted to frequency-domain energy distributions for observation. Different energy distributions represent different audio features.

Each frame multiplied by a Hamming window must have its spectral energy distribution generated through fast Fourier transform. The modular square of the spectrum of the audio signal is then calculated to identify its power spectrum.

The energy spectrum was processed using a mel-scale triangular bandpass filter bank, which was defined as a bank with M filters, the number of filters must be close to the number of critical bands. This study employed 40 filters, constituting a filtering range of 133–6854 Hz. The interval between each pair of f(m) decreased as m decreased and increased as m increased. The purpose of a triangular bandpass filter is to smooth a frequency spectrum and eliminate the effect of harmonics, thereby highlighting the formants of the original audio sample. Therefore, the tones and pitches of a speech sample are not displayed in MFCC arguments; in other words, an MFCC-based speech recognition system is not affected by the tonal differences in input speeches. This also lowers the number of computation procedures required, and MFCCs can be obtained through discrete cosine transform. See Figure 1 for the MFCC calculation procedure. In this study, 13-dimensional MFCCs were acquired, and five eigenvectors (viz., maximum, minimum, median, mean, and standard deviation) were extracted from each MFCC, for a total of 65.

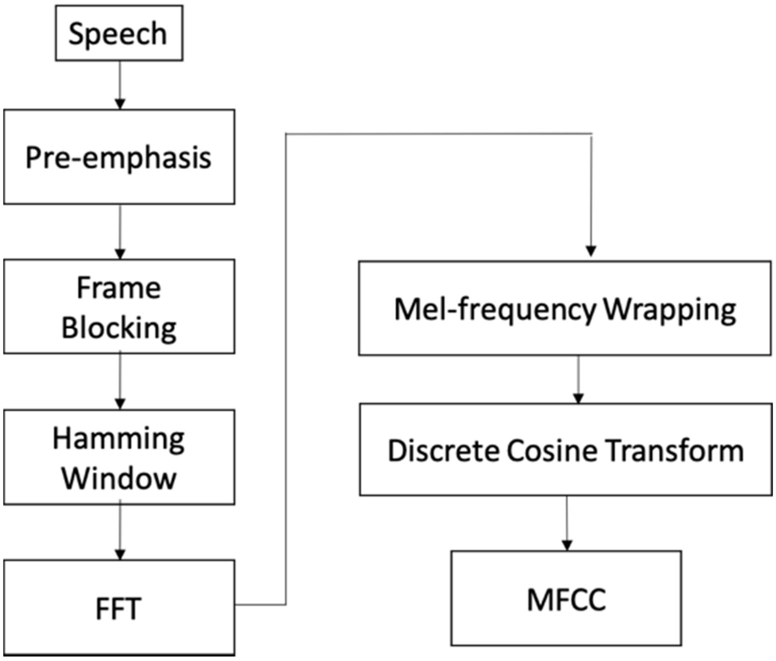

Figure 1. The MFCC calculation procedure.

2.4. ANN

This study employed the ANN MATLAB Neural Pattern Recognition classifier [13], which features a two-layer feedforward network with sigmoid hidden and SoftMax output neurons, enabling it to arbitrarily classify vectors with sufficient neurons provided in the hidden layers [23]. A Scaled conjugate gradient backpropagation training network was employed (Figure 2). A set of learning samples was employed for data training, and a classifier was established through argument matching. A set of validation data was applied to adjust the arguments in the classifier in the trained model; for example, the number of neurons in the hidden layers was determined in the neural network. The validation set was applied to confirm the network structure or control the arguments of model complexity. In addition, a set of testing data was employed to verify the recognition capacity of the trained model. In this study, 70%, 15%, and 15% of the samples were assigned to the training, validation, and testing sets, respectively. The number of neurons in the hidden layers was set as 10.

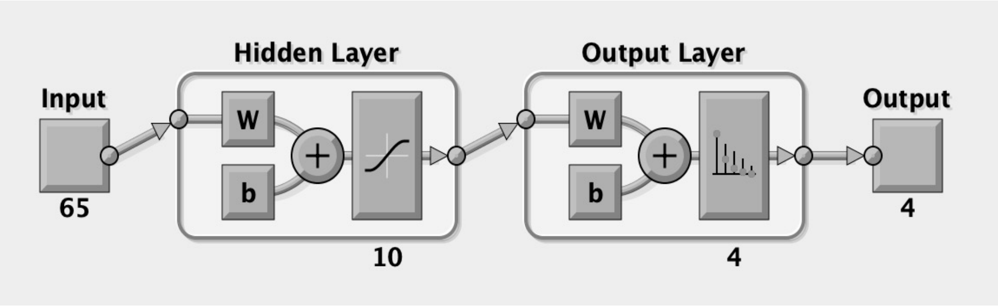

Figure 2. ANN model of this study.

2.5. Experimental Design and Procedures

A Chinese language SER system was constructed by using collected speech recordings to train the SER model and testing the model using clinical speech data.

A speech database referred to captured emotional speech files that could be assessed and that exhibited notable emotional expressions. The files were processed to extract the eigenvalues for training the Mandarin language SER model. Subsequently, the actual clinical speech was used in the model for block processing, identity recognition, emotion labeling, and MFCC acquisition. The SER results were then compared with the manual labeling results (Figure 3).

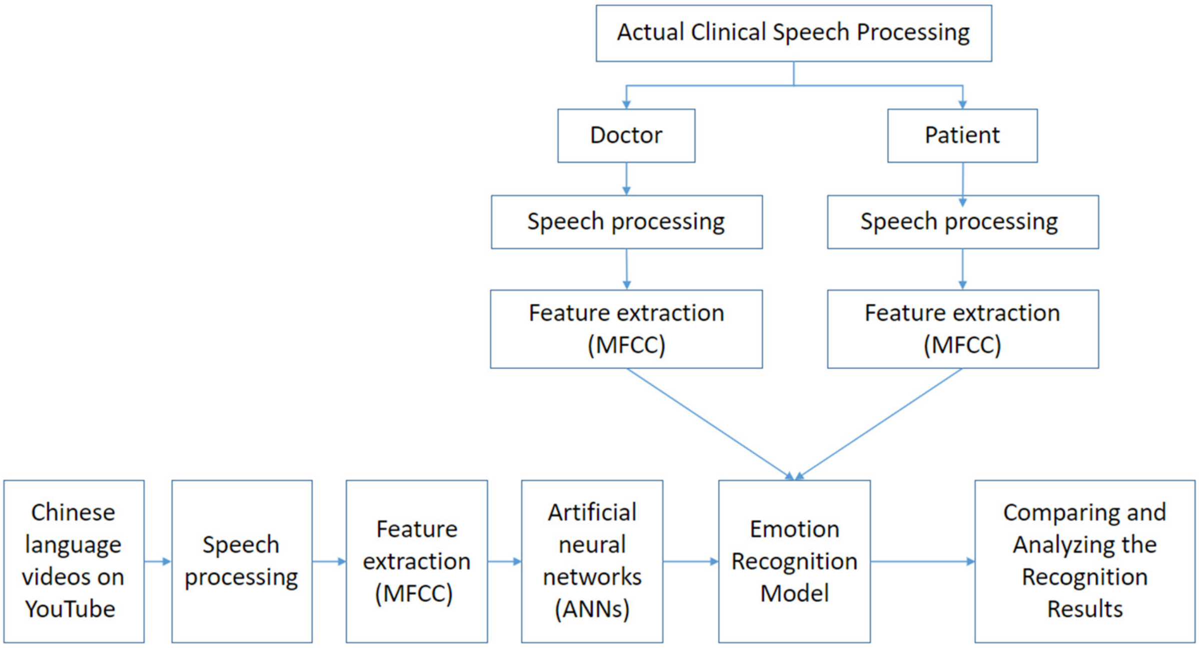
Actual Clinical Speech Processing
Doctor
Patient
Speech processing
Speech processing
Feature extraction (MFCC)
Feature extraction (MFCC)
Chinese language videos on YouTube
Speech processing
Feature extraction (MFCC)
Artificial neural networks (ANNs)
Emotion Recognition Model
Comparing and Analyzing the Recognition Results

Figure 3. Research framework.

2.5.1. Emotion Recognition Model

To establish the Mandarin language SER training model, speech samples from Chinese language videos on YouTube were employed as training data. The emotional speech samples required for the study were divided into 1-s blocks and independently labeled with emotions. Subsequently, the maximum values, minimum values, medians, means, and standard deviations from the first to 13th dimensions of MFCCs were employed as 65 eigenvalues for model establishment. The MATLAB pattern recognition classifier was then used to set 70%, 15%, and 15% of the acquired samples as training, validation, and testing data, respectively, and the number of neurons in the hidden layers of ANN was set as 10. Thus, the Chinese language SER model was established (Figure 4).

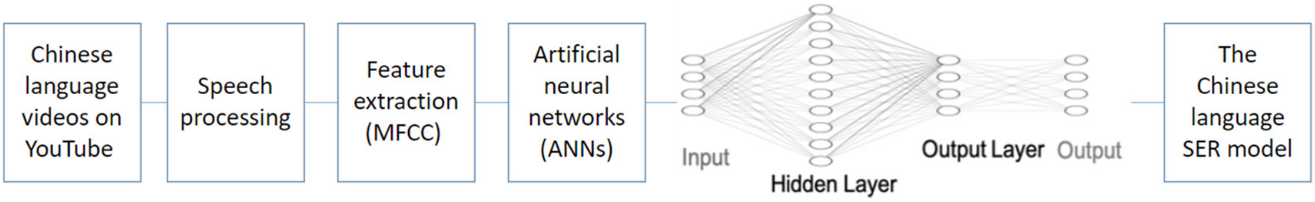
Chinese language videos on YouTube
Speech processing
Feature extraction (MFCC)
Artificial neural networks (ANNs)
Input
Hidden Layer
Output Layer
Output
The Chinese language SER model

Figure 4. SER model structure.

2.5.2. Actual Clinical Speech Processing

Because the clinical speech files were simultaneously recorded, to eliminate the presence of noise and identities not required for the study (e.g., other medical personnel and patient family members), the identities of the doctors and patients had to be recognized in speech processing with per second as an observation unit. The speech files were processed in the same manner as the SER model; that is, the files underwent speech processing, emotion labeling, and feature extraction. The processed speech files were then verified using the trained SER model, and the results were compared and analyzed.

2.5.3. Comparing and Analyzing the Recognition Results

Previous studies have employed emotional expression by professional actors or self-recorded and labeled emotional speech files. In this study, emotion recognition was conducted with a speech from actual clinical practice, and the participants were not professional actors; therefore, the results were consistent with actual interpersonal communications. The recognition results of the clinical speech files using the trained SER model were compared with the manually labeled emotional data to calculate the SER accuracy.

3. Results

To verify the eigenvalues acquired in previous studies as applicable for SER, the Berlin Database of Emotional Speech, the most frequently applied German SER database, was employed before the experiment. The other conditions of the German SER experiment, namely emotion categories, speech block processing, feature extraction, and model settings, were identical to those in the present Chinese language SER experiment. In the SER results, 84.7% of the samples were testing data, and the overall data accuracy was 92.4%; the area under the receiver operating characteristic (AUROC) was 0.90. As indicated in Figure 5, blue represents the receiver operating characteristic (ROC) curve of neutral emotion, green represents that of anger, dark yellow represents that of happiness, and purple that of sadness. If $0.5 < \text{AUROC} < 1$, then the ROC curve exhibits more satisfactory results than random guessing, verifying the classifier as effective in SER. According to the results of the experiment, the eigenvalues acquired in this study are applicable in SER research.

3.1. SER Model

The emotion recognition accuracy of the established SER model on the testing data was 91.3% (Table 1). Figure 6 illustrates the ROC curves in the model; in the training results, $0.5 < \text{AUROC} < 1$, indicating that the model is effective in SER. The trained model was then used to verify all the data sets, with an accuracy of 90.2% (Table 2). Figure 7 shows the ROC curves for all the data sets.

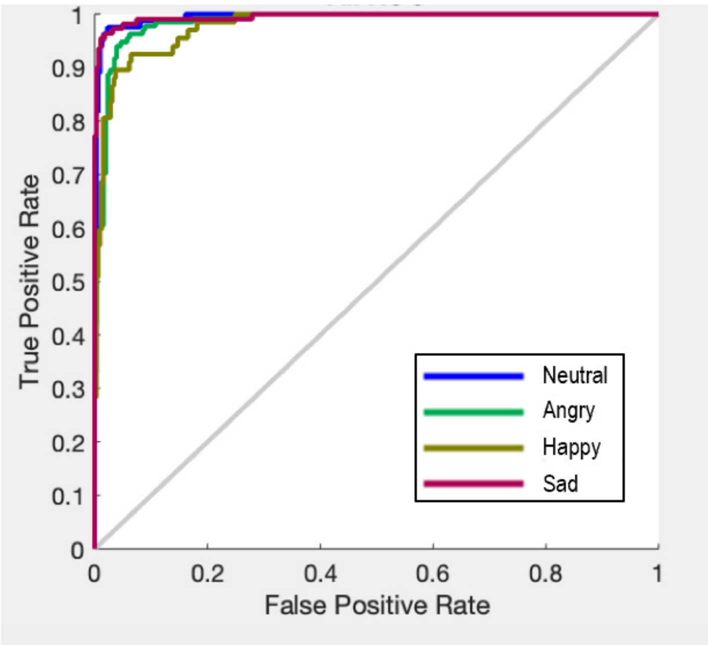

Figure 5. ROC test using the Berlin Database of Emotional Speech. Blue: Neutral. Green: Angry. Dark Yellow: Happy. Purple: Sad.

Table 1. SER model emotion categorization results on the testing data set (91.3% accuracy).

	Neutral	**Angry**	**Happy**	**Sad**	**Total**
Neutral	45	0	1	1	
Angry	0	15	2	0	
Happy	2	2	6	0	
Sad	0	0	0	18	
Accuracy	95.7%	88.2%	66.7%	94.7%	91.3%

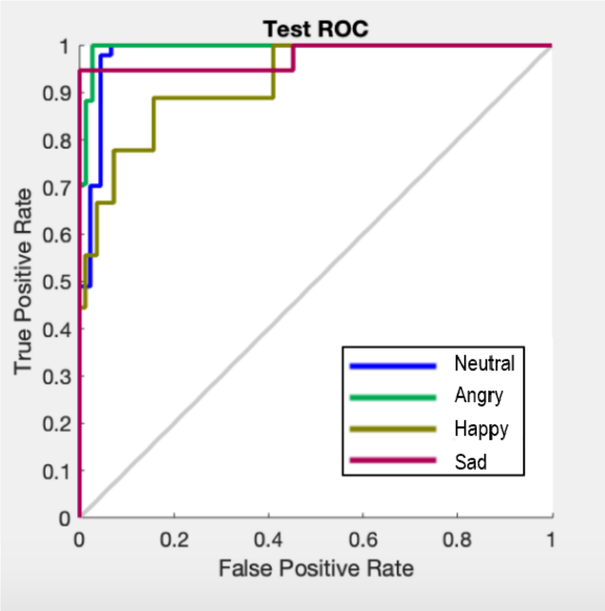

Figure 6. ROCs of the testing set. Blue: Neutral. Green: Angry. Dark Yellow: Happy. Purple: Sad.

Table 2. SER model emotion categorization results on all the data sets (90.2% accuracy).

	Neutral	**Angry**	**Happy**	**Sad**	**Total**
Neutral	299	0	22	7	
Angry	0	91	5	0	
Happy	16	9	63	1	
Sad	0	0	0	97	
Accuracy	94.9%	91.0%	70.0%	92.4%	90.2%

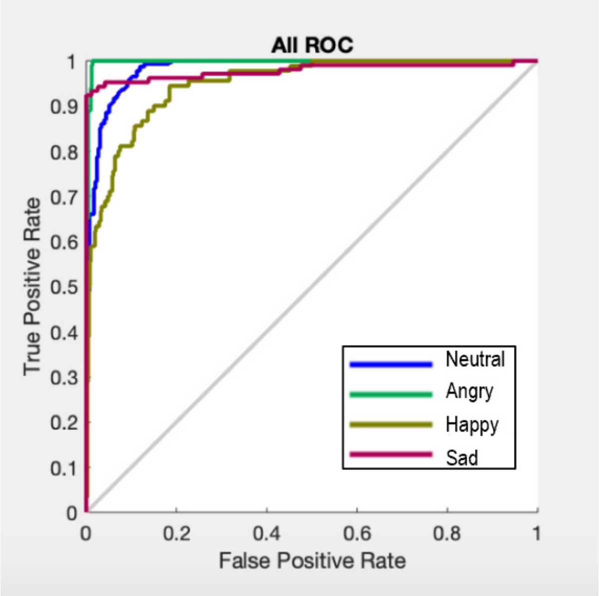

Figure 7. ROCs of all the data sets. Blue: Neutral. Green: Angry. Dark Yellow: Happy. Purple: Sad.

3.2. SER Result Comparison

Table 3 provides the comparison between manual emotion labeling and SER on the emotions of doctors and patients. Most of the samples were identified as neutral emotions. According to Table 2, the manual emotion labeling results differed slightly from the SER results. For example, in the fourth second, the patient's emotion was sad according to the manual label but was identified as neutral in SER. A comparison between the manual labeling and SER results for 30 clinical speech files revealed a classification accuracy of 88.6%, indicating that the SER model developed in this study is effective in emotion recognition.

Table 3. Comparison between manual emotion labeling and SER. Overlapping speech of multiple individuals in clinics is labeled "x"; "0" indicates that the speech does not belong to particular speaker identity.

	Manual Emotion Labeling		**SER**	
	Patient	Doctor	Patient	Doctor
1	neutral	0	neutral	0
2	neutral	0	neutral	0
3	neutral	0	neutral	0
4	sad	0	neutral	0
5	happy	0	neutral	0

Table 3. *Cont.*

	Manual Emotion Labeling		**SER**	
	Patient	Doctor	Patient	Doctor
6	neutral	0	neutral	0
7	neutral	0	neutral	0
8	x	0	neutral	0
9	0	neutral	0	neutral
10	0	happy	0	happy
11	0	happy	0	neutral
12	neutral	0	neutral	0
13	neutral	0	neutral	0
14	0	happy	0	neutral
15	0	neutral	0	neutral
16	0	happy	0	neutral
17	neutral	neutral	neutral	0
18	neutral	neutral	neutral	0
19	0	neutral	0	neutral
20	neutral	0	neutral	0
21	neutral	0	neutral	0
22	neutral	0	neutral	0
23	neutral	0	neutral	0
24	neutral	0	neutral	0
25	neutral	0	neutral	0

4. Discussion and Conclusions

The detection rate for this study proves that emotion recognition has reached maturity in facial recognition and can be a practical approach for the situation when doctors wear masks or are outside the range of a camera, detection, and recognition of changes in facial emotions become impossible. In this study, an SER model was developed to address this shortcoming. The SER model can be applied to complement a facial emotion recognition system for a comprehensive emotion recognition system.

This study adopts SER to solve accents and pronunciation in a language that may lower recognition accuracy. A comprehensive mandarin language SER model was built for a multi-lingual environment. The eigenvalues from the first to the thirteenth dimensions of MFCCs were employed in conjunction with AI to enable a machine to learn the characteristics of each emotion and automatically identify it, eliminating the need for manual emotion labeling. The sentiment recognition system developed by the hospital is used as a comparison answer to compare the sentiment recognition results of the artificial neural network classification, and then use the foregoing results for a comprehensive analysis to understand the interaction between the doctor and the patient. Using this experimental module, the emotion recognition rate is 93.34%, and the accuracy rate of facial emotion recognition results was 86.3%.

With the evolution of medical AI, studies have been conducted in developing various symptom assessment systems, and natural language processing has been performed in medical history research. However, to improve patient–doctor relationships, technological application in medical AI empathy is expected. This study aimed to identify speaker emotions through SER. In the future, by combining SER technology with facial emotion recognition, emotion recognition accuracy can be improved, thereby achieving AI empathy.

Doctors should provide a comfortable and compassionate medical environment when treating patients and apply AI that is sensitive to interpersonal interactions. Patient confidence and persistence play a critical role in successful disease prevention, diagnosis, treatment, and management, and medical personnel behavior critically influences patient feelings. AI empathy should be implemented to assist medical personnel to correctly identify patient emotions and deliver appropriate emotional responses, thereby attaining personalized processing of patient-doctor relationships.

Author Contributions: Conceptualization, H.-C.L.; methodology, H.-C.L., T.P. and M.-H.L.; software, M.-H.L.; validation, T.P.; formal analysis, H.-C.L., T.P. and M.-H.L.; investigation, T.P. and M.-H.L.; resources, H.-C.L. and H.-W.C.; data curation, M.-H.L. and H.-W.C.; writing—original draft preparation, H.-C.L. and M.-H.L.; writing—review and editing, H.-C.L., T.P. and H.-W.C.; project administration, H.-C.L. and T.P.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: TMU-N201810028.

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: We want to thank that the project of Ministry of Science and Technology, Taiwan (MOST 108-2634-F-038-002) provided part of data for this research.

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

References

1. Lin, S.Y.; Mahoney, M.R.; Sinsky, C.A. Ten ways artificial intelligence will transform primary care. *J. Gen. Int. Med.* **2019**, *34*, 1626–1630. [CrossRef] [PubMed]
2. Pan, T. A Health Support Model for Suburban Hills Citizens. *Appl. Syst. Innov.* **2021**, *4*, 8. [CrossRef]
3. Pan, T.; Fang, K. Ontology-based formal concept differences analysis in radiology report impact by the adoption of pacs. In *International Conference on Formal Concept Analysis*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 267–283.
4. Kerkeni, L.; Serrestou, Y.; Mbarki, M.; Raoof, K.; Mahjoub, M.A.; Cleder, C. Automatic Speech Emotion Recognition Using Machine Learning. In *Social Media and Machine Learning*; IntechOpen: London, UK, 2019.
5. Schuller, B.W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* **2018**, *61*, 90–99. [CrossRef]
6. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [CrossRef]
7. Huahu, X.; Jue, G.; Jian, Y. Application of Speech Emotion Recognition in Intelligent Household Robot. In Proceedings of the 2010 International Conference on Artificial Intelligence and Computational Intelligence, Sanya, China, 23–24 October 2010; Volume 1, pp. 537–541.
8. Szwoch, M.; Szwoch, W. Emotion recognition for affect-aware video games. In *Image Processing & Communications Challenges 6*; Springer: Cham, Switzerland, 2015; pp. 227–236.
9. Low, L.S.A.; Maddage, N.C.; Lech, M.; Sheeber, L.B.; Allen, N.B. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Trans. Biomed. Eng.* **2010**, *58*, 574–586. [CrossRef] [PubMed]
10. Pandey, S.K.; Shekhawat, H.S.; Prasanna, S.R.M. Deep Learning Techniques for Speech Emotion Recognition: A Review. In Proceedings of the 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), Pardubice, Czech Republic, 16–18 April 2019; pp. 1–6.
11. Liu, Y.Y.; Pan, T.; Cheng, B.W. Volume of surgery and medical quality: A big data analysis of hip hemiarthroplasty. In Proceedings of the 2018 IEEE International Conference on Applied System Invention (ICASI), Chiba, Japan, 13–17 April 2018; pp. 943–945.
12. Pan, T.; Fang, K.; Tsai, Y. Discriminant based analysis of unplanned 14 days readmission patients of hospital. *World Rev. Sci. Technol. Sustain. Dev.* **2010**, *7*, 86–99.
13. Koolagudi, S.G.; Rao, K.S. Emotion recognition from speech: A review. *Int. J. Speech Technol.* **2012**, *15*, 99–117. [CrossRef]
14. Kuchibhotla, S.; Vankayalapati, H.D.; Vaddi, R.; Anne, K.R. A comparative analysis of classifiers in emotion recognition through acoustic features. *Int. J. Speech Technol.* **2014**, *17*, 401–408. [CrossRef]
15. Sato, N.; Obuchi, Y. Emotion recognition using mel-frequency cepstral coefficients. *Inf. Media Technol.* **2007**, *2*, 835–848. [CrossRef]
16. Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 357–366. [CrossRef]

17. Likitha, M.S.; Gupta, S.R.R.; Hasitha, K.; Raju, A.U. Speech Based Human Emotion Recognition Using MFCC. In Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 22–24 March 2017.
18. Lalitha, S.; Geyasruti, D.; Narayanan, R.; Shravani, M. Emotion detection using MFCC and cepstrum features. *Procedia Comput. Sci.* **2015**, *70*, 29–35. [CrossRef]
19. Costanzi, M.; Cianfanelli, B.; Saraulli, D.; Lasaponara, S.; Doricchi, F.; Cestari, V.; Rossi-Arnaud, C. The effect of emotional valence and arousal on visuospatial working memory: Incidental emotional learning and memory for object-location. *Front. Psychol.* **2019**, *10*, 2587. [CrossRef] [PubMed]
20. Chiou, B.C. Cross-Lingual Automatic Speech Emotion Recognition. Master's Thesis, National Sun Yat-sen University, Kaoshiung, Taiwan, 2014.
21. Mohino-Herranz, I.; Gil-Pita, R.; Alonso-Diaz, S.; Rosa-Zurera, M. MFCC based enlargement of the training set for emotion recognition in speech. *arXiv* **2014**, arXiv:1403.4777. [CrossRef]
22. Mel Frequency Cepstral Coefficient (MFCC) Tutorial. Available online: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/ (accessed on 22 May 2021).
23. Umamaheswari, J.; Akila, A. An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN. In Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 14–16 February 2019; pp. 177–183.