석 사 학 위 논 문
Master's Thesis

대화 과정에서 상대방 데이터를 활용한 멀티모달 감정 인식 개선

Improving Multi-modal Emotion Recognition with Counterpart Data in Dyadic Conversations

2022

김 용 신 (金 容 信 Kim, Yongshin)

한 국 과 학 기 술 원

Korea Advanced Institute of Science and Technology

석 사 학 위 논 문

대화 과정에서 상대방 데이터를 활용한 멀티모달 감정 인식 개선

2022

김 용 신

한 국 과 학 기 술 원

산업 및 시스템 공학과

대화 과정에서 상대방 데이터를 활용한 멀티모달 감정 인식 개선

김 용 신

위 논문은 한국과학기술원 석사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2022년 6월 13일

심사위원장 이 의 진 (인)

심 사 위 원 이 성 주 (인)

심 사 위 원 이 재 길 (인)

Improving Multi-modal Emotion Recognition with Counterpart Data in Dyadic Conversations

Yongshin Kim

Advisor: Uichin Lee

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Master of Science in Industrial and Systems Engineering

Daejeon, Korea
June 13, 2022

Approved by

Uichin Lee
Professor of School of Computing

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MKSE 김용신. 대화 과정에서 상대방 데이터를 활용한 멀티모달 감정 인식 개선. 산업 및 시스템 공학과 . 2022년. 27+iii 쪽. 지도교수: 이의진. (영문 논문)
Yongshin Kim. Improving Multi-modal Emotion Recognition with Counterpart Data in Dyadic Conversations. Department of Industrial and Systems Engineering . 2022. 27+iii pages. Advisor: Uichin Lee. (Text in English)

초 록

자연스러운 감정 인식 기술은 다양한 활용처가 있지만 많은 이전 연구들은 인간의 감정을 사람 간 상호작용 속에서 분석하지 않았다. 그 이유는, 대부분의 감정 관련 데이터셋은 대화 과정 속에서 수집되지 않았기 때문이다. 즉, 이전 연구는 인간의 감정을 예측하기 위해 상대방의 데이터가 아닌 화자의 데이터만 사용했다. 본 연구는 멀티모달 센서 데이터(음성 및 생체신호)를 활용하여 감정 상태를 자동 분류하는 CNN-LSTM 딥러닝 네트워크를 소개한다. 특히, 자연스러운 대화 상황을 가정한 환경에서 연속적인 감정을 수집한 K-EmoCon 데이터셋을 활용해 쌍방 대화에서 대화 파트너의 데이터를 활용하여 화자의 감정 분류의 정확도를 높일 수 있는 모델을 제시한다. 실험 결과 대화 상대방의 음성 및 생체신호가 발화자의 감정 예측 성능에 긍정적인 영향을 미친 것으로 나타났다. 이 논문을 통해 우리는 자연스러운 대화 과정에서 발화자 뿐만 아니라 상대방의 특성 또한 고려해야 된다는 것을 주장한다.

핵 심 낱 말 감정 인식, 감성 컴퓨팅, 일상대화, 딥러닝, 멀티모달

Abstract

Today, as we live in numerous interactions, many studies have tried to predict human emotions. Since our daily life consists of countless interactions, it is better to predict human emotions between interactions. However, most studies have focused only on the speaker's data, not the counterpart's data, to predict the speaker's emotions because datasets which labeled human emotions in the naturalistic conversation are rare. In this study, we propose a method for predicting the emotions of the speaker in the naturalistic conversation using a speaker encoder and counterpart encoder composed of CNN-LSTM deep learning networks. We used emotion-related data called K-EmoCon collected during the debate process to empirically evaluate our model. The results showed that the counterpart's speech and the physiological signals had a positive impact on predicting the speaker's emotions. Through this paper, we hope to be helpful in the study of predicting emotions in naturalistic conversation.

Keywords Emotion recognition, Affective computing, Naturalistic conversation, Interpersonal Features, Deep neural networks, Multimodal

Contents

List of Tables

List of Figures

Chapter 1. Introduction

Understanding emotions in our lives is very important in human-to-human communication. Emotion recognition in HCI allows user-centered systems to provide users with more natural and easier interactions. In recent years, methods of combining physiological signals, audio, video, hand gestures, and other forms such as body movements have contributed to progress in the field of emotional recognition. Automatic recognition of emotions provides a natural interface between humans and machines, allowing the system to understand, interpret, and respond accordingly.

Affective computing is the field of designing and developing systems and devices that can recognize, interpret, and process human emotions. Affective computing is important because emotions affect human physiological and psychological conditions, which are closely related to human mental health. In order to recognize emotions, we must first define emotions and quantify them. The basic definition of emotion was proposed by psychologists decades ago, and it's a way to divide emotions into discrete categories like joy and sadness. However, because of the fact that there are individual strengths in emotions, like a little joy, and a lot of joy, Russell [43] insisted that emotions can be classified into 2D spaces by arousal and valence. Arousal is related to whether a person is active or passive and valance is associated with whether a person is positive or negative. As shown in Figure.1.1, we used Russell's two-dimensional emotion model in this paper.

With significant advances in the field of machine learning, many studies have recently been conducted that allow us to automatically recognize human emotions [19, 24, 51]. However, all of these emotional recognition studies considered only individual speaker to predict that speaker's emotions. Although emotions have traditionally been regarded as a private and internally occurring phenomenon, recent studies on emotions suggest that emotions are inherently social. As Parkinson et al. [38] said, listening to someone reading in a happy, sad, or neutral voice induces similar feelings. Therefore, it is necessary that affective state estimation methods consider the social interaction scenario.

By trying to understand the counterpart's feelings during the conversation, we identify that person's attitude, feelings, and intentions, which leads to successful communication. However, the ability to recognize emotions varies from person to person, and there are cases where the counterpart's emotions are not recognized. These mistakes can lead to mutual misunderstanding, communication problems, and relationship deterioration [40].

While previous studies used human speech, and physiological signals for emotion recognition, most studies examined the emotions of isolated individuals without explicit interaction between subjects. The reason is that so far, there has been no dataset that has labeled human emotions in the natural conversation process of two or more people. In other words, most Speech Emotion Recognition (SER) systems had been developed using acted datasets [11]. This means that the dataset was not collected during a natural conversation, but was collected in a conversation that artificially created emotions. Therefore, they fail to detect emotions in natural utterances. The emotion distributions in the acted and natural speech do not match because an acted speech is recorded in a restricted environment, thereby lacking the variations in natural speech [11].

To address this gap, we present a model for emotion prediction of each utterance subject in the dialogue scenario of the two. In the conversation process, one's emotions may be influenced by counterpart's speech and physiological signals. In this paper, we went through a data preprocessing that considers

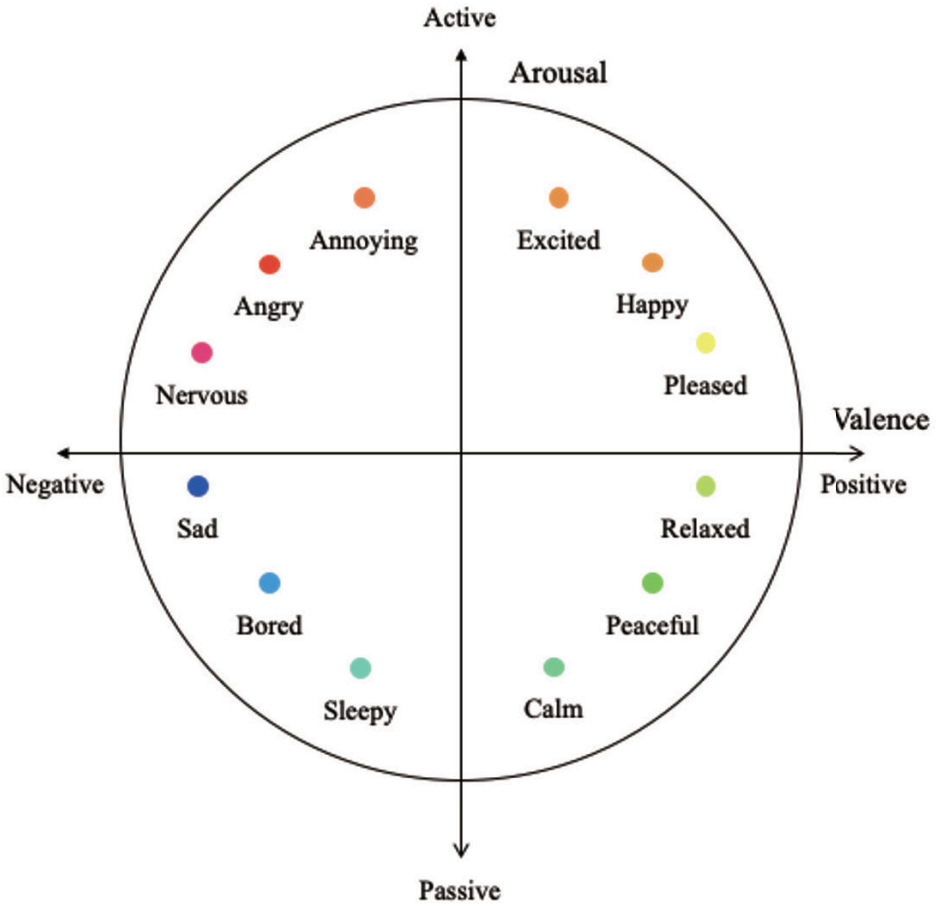

Figure 1.1: A diagram of an arousal-valence circumplex with approximate placements of emotional words on the space [36, 44]

not only individual signals but also the counterpart's signals for predicting speaker's emotions. The interpersonal features used in this study are defined as interaction activities (verbal or nonverbal) that occur consciously or unconsciously during communication which includes not only the speaker's signals but also the counterpart's signals. On the contrary, individual features are a characteristic of oneself that does not consider the counterpart in the conversation. Since emotion recognition in conversation is usually multi-party and there can be individual related features and interpersonal related features [14], we intended to investigate to what extent it can affect the emotion prediction model when only individual features are used and the counterpart's features are used together.

To the best of our knowledge, [40] is the only study that applied the influence of interpersonal features to emotion prediction. Quan et al. [40] analyzed how the counterpart's features affect the speaker's emotions during the conversation using visual and speech modalities. They predicted emotions every five seconds with the synchronization aspect between the time-lagged features of the counterpart and the features of the speaker using cosine similarity. They showed better emotional prediction performance when using synchronization for visual data, but for audio data, the degree of better is not as great as expected because when one person speaks, the other usually listens in silence [40]. In addition, finding similar signals between the speaker and the counterpart using cosine similarity does not necessarily provide evidence to say that the signal is a good description of the speaker's emotional state. For example, the heart rate when the counterpart shouts and the heart rate when we are surprised to hear it can be

completely different, but emotions can be transferred.

As a result, we analyzed the impact of interpersonal features using raw signals within 5-second segments without using the synchronization aspect. This is because, for audio, the synchronization effect is considered insignificant by [40], and physiological signals are signals that indicate personal conditions, not visual characteristics. Therefore, we set our research question as follows: Will the counterpart's raw speech and physiological signals improve the model performance in predicting the speaker's emotions?

In order to analyze the impact of the counterpart's signals on the speaker's emotions during the conversation process, we need a dyadic dataset in which the subjects speak spontaneous emotions during a naturalistic conversation. We used the K-EmoCon dataset to study continuous emotional states in the context of interactive communication. This dataset is one of the first datasets to include physiological signals and audiovisual records from both parties involved in social interaction [40]. These characteristics allow us to explore the possibility of improving emotional recognition accuracy by modeling emotional states using information from both sides of the communication. The dataset also includes self-reported and recognized emotion labels to observe the relationship between individual emotional states in communication.

To predict the speaker's emotions during the conversation process, we used speech and physiological signals. Speech, the most commonly used and quick communication in everyday life, convey useful information about a person's emotional state [25]. On the other hand, physiological signals that are rarely seen in our eyes also respond to emotional states. Using only speech data cannot guarantee the reliability of emotions because people can make enough facial expressions or speech while excluding real emotions. However, physiological signals which are from the autonomic nervous system (ANS) are involuntarily activated and cannot be easily controlled by people [47].

Therefore we propose an automatic multimodal emotion recognition model based on CNN-LSTM networks to predict arousal and valence using speech and physiological signals. Our paper contributes in that it examined how positively the raw data of counterpart's speech and physiological signals have an impact on model performance improvement in predicting the speaker's emotions during the naturalistic conversation. The rest of the paper is organized as follows: In Section 2, related recent studies are reviewed. Our proposed approach is described in Section 3. Experiment descriptions are in Section 4, and the results will be in Section 5. The discussion will be covered in Section 6 and lastly, Section 7 will talk about the conclusion.

Chapter 2. Related Works

Automatic emotion recognition has provided a natural interface between humans and machines, allowing systems to understand, interpret, and respond accordingly to emotions. This section reviews recent research in the field of emotion recognition. We introduce studies associated with speech and physiological signals since we present multimodal emotion recognition using these two modalities.

The most basic and easy approach to predicting emotions is to predict through the human voice. Many acoustic features have been studied to effectively perform emotion classification. Notable features include pitch-related features, energy-related features, Mel-Frequency Cepstral Coefficient (MFCC), and Linear Prediction Coefficient (LPC), etc [18]. Some studies have learned the distribution of these low-level features using generative models such as the Gaussian Mixed Model (GMM) and Hidden Markov Model (HMM), and then used Bayesian classifiers or maximal likelihood principles for emotion recognition [45, 28]. Another trend in speech emotion recognition is the application of statistical functions to these low-level acoustic features to calculate global statistical features for classification. Support Vector Machine is the most commonly used classifier for global functions [57, 34]. Some other classifiers such as decision tree [27] and K-Nearest Neighbor (KNN) [22] have also been used for speech emotion recognition. This approach requires a very high-level handcrafted function selected empirically. Meanwhile, deep learning is an emerging field in machine learning in recent years. A highly efficient characteristic of DNN is that it is possible to learn high-dimensional characteristics from raw data, and therefore, SER has also been studied using deep learning.

Assuncao et al. [2] used CNN with the various datasets to recognize human emotions. They also verified that emotion recognition improves when speakers' emotional prosody cues in speech are considered. Zhang et al. [55] used mel-spectrogram features to train Deep CNN (DCNN) model for predicting emotion. Zhao et al. [56] tried to use log-mel spectrogram from a human voice to predict emotions by using 1D and 2D CNN-LSTM networks. Mingke Xu et. al [52] used MFCC from the speech signal and utilized a multi-head self-attention model for speech emotion recognition. In addition to this, a comprehensive review of existing approaches to speech emotion recognition has already been included in the literature [10].

Research has been conducted to predict human emotions not only by speech data but also by using physiological signals. In this study, physiological signals refer to signals that occur internally such as electroencephalogram (EEG), temperature (T), electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR), respiration (RSP), etc. Signals from these autonomic nervous systems (ANS) are involuntarily activated, making it difficult for a person to deliberately control them. Because of these characteristics, emotional analysis using physiological signals is free from false or intentional emotional manipulation. Therefore many studies have predicted emotions using multiple physiological signals.

Like SER, we will also introduce emotion recognition with physiological signals using traditional machine learning and deep learning techniques. Das et al. [8] extracted Galvanic skin response and ECG features and predicted emotions using SVM, Naive Bayes, and KNN. They used physiological signals to predict three emotional states: joy, sadness, and neutrality and suggested which signals were efficient in predicting each emotion. Garcia et al. [13] used Gaussian process latent variable models (GP-LVM) which also contain an SVM to predict three levels of valence and arousal. Wang et al. [51] used KNN with four features extracted from electroencephalogram, electromyogram, skin conductivity, and respiration to

classify emotions.

In a case of predicting emotions by using deep learning, Martinez et al. [32] predicted four emotional states (relaxation, anxiety, excitation, and fun) by learning skin conduction and blood volume pulse signals through a CNN model [7]. Ranganathan et al. [41] introduced four Deep Belief Network (DBN) models for generating robust multimodal features for unsupervised emotion recognition by using speech, and facial expressions, body gestures, and physiological signals.

There are studies that predict emotions with only physiological signals, but few studies predict emotions using physiological signals and speech signals together as in our study. Bakhshi et al. [3] proposed an efficient architecture for recognizing emotion, utilizing the information acquired from raw speech signals and physiological signals. They predicted arousal and valence using the CNN-BiGRU model. They predicted emotions by including only speech data once, used only physiological signals once, and fused two at the end.

On the other hand, in addition to speech and physiological signals, gender information has been used to improve emotion prediction performance because when handling speech data, gender information can have a positive effect on model performance [4]. Zhang, Linjuan, et al. [54] checked that the performance of emotion recognition was improved by utilizing gender-related features. They showed that emotion prediction performance is improved by putting gender information in speech data. Similarly, Vogt and Andre [50] improve emotion recognition from speech signals by making use of automatic gender detection. Since our study also used speech data, we used the subject's gender data as well.

As such, there are previous studies that use speech and physiological signals and gender information to predict emotions, but these studies have not considered the interaction between the two. The reason is that datasets which label human emotions in the natural conversation process are rare. Table 2.1 summarizes the human emotion-related dataset used in emotion research until recently. Looking at the table, all datasets except K-EmoCon are artificially inducing emotions. For example, it is to guide the subjects to get specific stimuli in the conversation for "delighted" feelings and conduct the experiment. In this case, it cannot be said that a person's emotions were predicted during a truly natural conversation.

However, when we think about emotion prediction in our real lives, we have to consider interaction in naturalistic conversation. To the best of our knowledge, Quan et al. [40] is the only study that applied the influence of interpersonal features to emotion prediction. This paper analyzed how the counterpart's characteristics affect the speaker's emotions during the conversation using visual and speech features. In addition, the author utilized the K-EmoCon dataset, as in our study, because the K-EmoCon is the only dyadic dataset in which the subjects show spontaneous emotions during naturalistic conversations [40]. One difference from us is that it analyzed the synchronization features between the time-lagged features of the counterpart and the features of the speaker using cosine similarity. Unlike this, we used the counterpart's raw features at that point without time-lagged to predict the speaker's emotions because of the following three reasons. First, Quan et al. [40] mentioned that while using synchronization for visual data showed better emotional prediction performance, for audio data, the better degree was not as large as expected because when one person spoke, the other usually heard in silence. Second, finding similar signals between the speaker and the counterpart using cosine similarity does not necessarily provide evidence to say that the signal is a good description of the speaker's emotional state. For example, the heart rate when the counterpart shouts and the heart rate when we are surprised to hear it can be completely different, but emotions can be transferred. Third, a segment has a length of 5 seconds, and we thought that five seconds was a time when the counterpart could sufficiently influence the speaker's emotional state. As a result, we analyze the effect of interpersonal features using raw signals within 5 seconds

without using the synchronization aspect. However, in a future study, we may also use synchronization to predict emotions and analyze whether there is room for improvement in model performance.

Table 2.1: Existing multimodal emotion recognition datasets. For annotation types, S = *self annotations*, P = *partner annotations*, and E = *external observer annotations*.

Name (year)	**Size**	**Modalities**	**Spon. vs. Posed**	**Natural vs. Induced**	**Annotation method**	**Context**
IEMOCAP (2008) [5]	10	Videos, face motion capture, gesture, speech (audio & transcribed)	Both	Induced	S, E	Dyadic
SEMAINE (2011) [33]	150	Videos, FAUs, speech (audio & transcribed)	Spon.	Induced	E	Dyadic
MAHNOB-HCI (2011) [48]	27	Videos (face and body), eye gaze, audio, biosignals (EEG, GSR, ECG, respiration, skin temp.)	Spon.	Induced	S	Individual
DEAP (2012) [23]	32	Face videos, biosignals (EEG, GSR, BVP, respiration, skin temp., EMG & EOG)	Spon.	Induced	S	Individual
DECAF (2015) [1]	30	NIR face videos, biosignals (MEG, hEOG, ECG, tEMG)	Spon.	Induced	S	Individual
ASCERTAIN (2016) [49]	58	Facial motion units (EMO), biosignals (ECG, GSR, EEG)	Spon.	Induced	S	Individual
DREAMER (2017) [21]	23	Biosignals (EEG, ECG)	Spon.	Induced	S	Individual
AMIGOS (2018) [7]	40	Vidoes (face & body), biosignals (EEG, ECG, GSR)	Spon.	Induced	S, E	Individual, Group
CASE (2019) [46]	30	Biosignals (ECG, respiration, BVP, GSR, skin temp., EMG)	Spon.	Induced	S	Individual
CLAS (2020) [31]	64	Biosignals (ECG, PPG, EDA), accelerometer	Spon.	Induced	Predefined†	Individual
RECOLA (2013) [42]	46	Audio	Spon.	Induced	S	Individual, group
K-EmoCon (2020)	*32*	*Videos (face, gesture), speech audio, accelerometer, biosignals (EEG, ECG, BVP, EDA, skin temp.)*	*Spon.*	*Natural*	*S, P, E*	*Dyadic*

† Predefined emotion categories of stimuli and success rates of participants in a set of purposefully selected cognitive tasks were used as ground-truth labels.

Chapter 3. Methodology

3.1 Dataset

This section describes the datasets and preprocessing we used. We also present a CNN-LSTM network containing a speaker encoder and counterpart encoder for interpersonal analysis in conversation.

We used the K-EmoCon dataset to study continuous emotional states in the context of interactive communication. This dataset is one of the first datasets to include physiological signals and audiovisual records from both parties involved in social interaction [40]. These characteristics allow us to explore the possibility of improving emotional recognition accuracy by modeling emotional states using information from both sides of communication.

Table 3.1: Steps for a data collection session, each session lasted approximately two hours.

Step	**Allocated time**	**Description**
Read and sign consent forms	10 min	Experimenters provided consent forms to participants, and two written consents each for participation and the collection of privacy-sensitive data were obtained.
Choose sides and the order	5 min	Participants were assigned to either argue in favor of or against accepting refugees and decided on the first speaker.
Prepare debate	15 min	Participants were provided with supplementary materials to prepare their arguments.
Equip sensors	10 min	Experimenters explained wearable devices to participants and assisted them in wearing devices.
Measure baseline	2 min	A baseline corresponding to a neutral state was measured for each participant.
Overview debate	5 min	The moderator explained the debate rules and notified participants that they are allowed to intervene.
Debate	10 min	Participants could speak for two consecutive minutes during their turns and they were notified twice at 30 and 60 seconds before the end of the debate.
Annotate emotions	60 min	Participants annotated emotions at intervals of every 5 seconds, watching footage of themselves and their partners.

The K-EmoCon dataset is a multimodal dataset containing various types of emotion annotations taken in a continuous manner [37]. It contains emotional annotations of three perspectives: self, argumentative partner, and external observer that differentiate the previous dataset. This dataset includes multimodal measurements of audiovisual images and physiological signals from 16 discussion sessions of approximately 10 minutes on social topics. While watching the discussion video, the commentators recorded emotional expressions every 5 seconds in terms of arousal and valence emotions. K-EmoCon is the first freely accessible emotional dataset to host more emotional evidence during social experiences [17]. Participants selected for the experiment were people between the ages of 19 and 36. The dataset aims to provide a counterpart participant's perspective on emotions and additional aspects of external reviewers to improve

Table 3.2: Collected emotion annotations categories of K-EmoCon dataset.

Emotion annotation categories	**Description**	**Measurement scale or method**
Arousal / Valence	Two affective dimensions from Russell's circumplex model of affect [43]	1: very low - 2: low - 3: neutral - 4: high - 5: very high
Cheerful / Happy / Angry / Nervous / Sad	Emotion states describing a subjective stress state [39]	1: very low - 2: low - 3: high - 4: very high
Boredom / Confusion / Delight / Engaged concentration / Frustration / Surprise / None	Commonly used Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) educationally relevant affective categories [35]	Choose one
Confrustion / Contempt / Dejection / Disgust / Eureka / Pride / Sorrow / None	Less commonly used BROMP educationally relevant affective categories [35]	Choose one

emotion classification.

Participants were asked to sit in a well-lit room and wore polar H7 heart rate sensors to detect ECG biological signals and E4 wristbands to indirectly measure triaxial acceleration, and heart rate via the PPG signals, galvanic skin response, and temperature. Table 3.1 shows the overall procedure for data collection. A total of 172.92 minutes of data were obtained. The five external evaluators selected to evaluate the participants' emotions were three men and two women, aged 22-27. The obtained data were synchronized over time and preprocessed. There are four different types of emotion annotations (target values) on the K-Emocon dataset: Self-Annotations, Partner Annotations, Self-Partner Mean Annotations, and Aggregate External Annotations.

Self annotations are labeled by the speaker about his feelings, and partner annotations are labeled by the counterpart about the speaker's feelings. At this time, the counterpart has the context information of the speaker because he/she is the person who conducted the discussion together and therefore labeled the emotion based on it. Finally, external annotations are a case of labeling the speaker's emotions from the perspective of a third party who did not participate in the discussion without any information about the context. In this study, emotion prediction was conducted using self-annotations, partner annotations, and self-partner mean annotations. For each emotion annotation, the emotion annotations categories shown in Table 3.2 were evaluated. In this study, emotions were predicted using only arousal and valence.

3.2 Preprocessing

The K-EmoCon data consists of a total of 16 discussion data, with 32 subjects paired with two. The K-EmoCon dataset is available upon request on Zenodo (https://doi.org/10.5281/ zenodo.3931963). To analyze interpersonal features, we had to choose a case in which the audio and physiological signals of the two conversational subjects were fully accessible. Unfortunately, unlike perfectly accessible audio data, physiological signals have in many cases limited accessibility. As a result, we used a total of 12 subjects' data included in the six discussions. The user code included in the experiment is 9, 10, 13, 14, 15, 16, 25, 26, 27, 28, 31, and 32. Table 3.3 shows these subjects' personal information and the distribution of self-annotated emotion labels.

Emotional states were collected during the discussion period of 10 to 15 minutes at 5-second intervals. Therefore, we proceeded with the data preprocessing every 5 seconds. Table 3.4 shows the physiological

Table 3.3: Subjects's personal information.

ID	Counterpart ID	Gender	Age	Discussion Length	Arousal (mean/std)	Arousal (high)	Valence (mean/std)	Valence (high)
P9	P10	M	21	10:16	2.55 (0.66)	0.471	3.14 (0.45)	0.967
P10	P9	M	22	10:16	2.55 (0.53)	0.555	2.56 (0.55)	0.529
P13	P14	M	22	10:11	3.26 (0.97)	0.717	2.98 (0.55)	0.842
P14	P13	F	21	10:11	3.43 (1.06)	0.793	2.29 (1.09)	0.353
P15	P16	M	30	10:15	2.93 (1)	0.708	2.96 (1.01)	0.625
P16	P15	F	26	10:15	3.07 (0.75)	0.75	3.47 (0.5)	1
P25	P26	F	26	10:23	2.27 (0.75)	0.374	3.8 (0.46)	1
P26	P25	M	25	10:23	3.73 (1)	1	3.33 (1.07)	0.756
P27	P28	F	24	10:47	3.66 (0.7)	1	2.99 (0.61)	0.808
P28	P27	F	23	10:47	2.74 (0.77)	0.533	2.83 (0.73)	0.625
P31	P32	M	24	10:58	2.38 (0.77)	0.205	3.34 (0.81)	0.795
P32	P31	F	19	10:58	4.39 (0.78)	1	2.87 (0.57)	0.867

signals provided by the K-EmoCon dataset. Among these, we used physiological signals present in the K-Emocon dataset captured by Empatica E4 and Polar H7 heart rate sensors, namely blood volume pulse (BVP) (64 Hz), skin temperature (ST) (4 Hz), electrocardiograph (ECG) (1 Hz), and electrodermal activity (EDA) (4 Hz). For feature extraction, since the sampling rate of each biosignal is different, we used resampling techniques of four biosignals each, four resampling techniques per second. In this case, the BVP was reduced from 64 to 4, and the ECG was increased from 1 to 4. By doing so, the weight of each physiological signal could be equally learned in the model. In addition, we normalized the physiological signals for each user so that the gap in signals between users was eliminated.

Since each discussion conversation is in one audio file, the mixed voice of the two speakers can cause many problems in voice processing. Therefore, we manually separated the audio data of the two people to overcome as in Quan et al [40]. As shown in Figure 3.1, the conversation data of the two people is divided into audio data of each person.

When one person speaks during the conversation, the other person is mainly listening in silence. Therefore, when preprocessing audio data, we inevitably face a silent section without any voice. Since emotions exist even in the silent section, how to deal with this part can be an important criterion. This will be covered in more detail in the discussion section. In this study, the silence section was left untouched. As a result, of the two people's audio data, only the speaker's audio data was used for individual audio features analysis, and both speaker and counterpart's data were used for interpersonal features analysis.

We also applied pre-trained models for audio data since from previous studies, it can be seen that using pre-trained models can lead to high accuracy in deep learning models even on small datasets [9, 20]. The dataset used by K-EmoCon is a dataset of 12 people, which is relatively less than the number to learn deep learning. Therefore, we used a widely used transfer learning model called VGGish [20] which is popular pretrained convolutional audio architecture to create audio feature embeddings. The pre-trained VGGish model transforms speech recordings into a mel-spectrogram processed by a multilayer convolutional network, extracting an embedding vector of size 128 every second to form a 2D array for use in the classification layer [20]. Then this 2D array is fine-tuned for the emotion prediction model.

Table 3.4: Data collected with each wearable device, with respective sampling rates and signal ranges.

Devices	**Collected data**	**Sampling rate**	**Signal range [min, max]**
Empatica E4 Wristband	3-axis acceleration	32Hz	[-2g, 2g]
	BVP (PPG)	64Hz	n/a
	EDA	4Hz	[0.01μS, 100μS]
	Heart rate (from BVP)	1Hz	n/a
	IBI (from BVP)	n/a	n/a
	Body temperature	4Hz	[$-40\,^{\circ}$C, $115\,^{\circ}$C]
NeuroSky MindWave Headset	Brainwave (fp1 channel EEG)	125Hz	n/a
	Attention & Meditation	1Hz	[0, 100]
Polar H7 Heart Rate Sensor	HR (ECG)	2Hz	n/a

3.3 Proposed Architecture

To analyze the difference between interpersonal features and independent features, we present a deep learning model consisting of CNN-LSTM networks as shown in Figure 3.2. The basic configuration of the model contains a speaker encoder and a counterpart encoder. For the speaker encoder, speech data, physiological signal data, and gender data of the speaker are extracted from the dataset. First, to explain the speech data, as described above, the speech data is converted into a mel-spectrogram through a pretrained model called VGGish. This is soon extracted via Convolutional Neural Networks(ConvNet). Convolutional neural networks process incoming data in the form of multiple arrays, such as images composed of an audio spectrogram or a video. The architecture of the general ConvNet consists of the following. The first few steps consist of a convolution layer and a pooling layer. A unit of a convolutional layer consists of a feature map that connects to a local patch from a feature map of the previous layer through a weight set called a filter bank [26]. The results of this local weighted sum are transferred via a nonlinear function. Different feature maps of layers use different filter banks, as groups of local values in array data generally form unique local patterns that are highly correlated and easily detectable, and local statistics of images and other signals are invariant with the ocation. After passing the ConvNet model, the audio data passes through the LSTM deep learning model. LSTM is designed to avoid long-term dependency problems. Existing RNNs had the disadvantage of showing effect only on relatively short sequences. As the sequence of input data increases, the RNN loses the amount of information of the initial input value. To overcome these disadvantages, the LSTM uses input gates, forget gates, and output gates in the memory cells of the hidden layer to keep information stored for a long time. In this paper, we used Bidirectional LSTM, which is advanced from the original LSTM. In a bidirectional LSTM, each training sequence is presented forward and backward to separate recurrent nets. Both sequences are connected to the same output layer. Bidirectional LSTMs show higher performance than traditional LSTMs because they have complete information about all points in a given sequence and everything before and after it [15].

In the case of physiological signals, unlike speech data, the LSTM model is directly passed without going through ConvNet. The reason is that we conducted the experiment with data from 12 people as explained earlier. In this case, the data of 12 people are shown as small data to train a deep learning

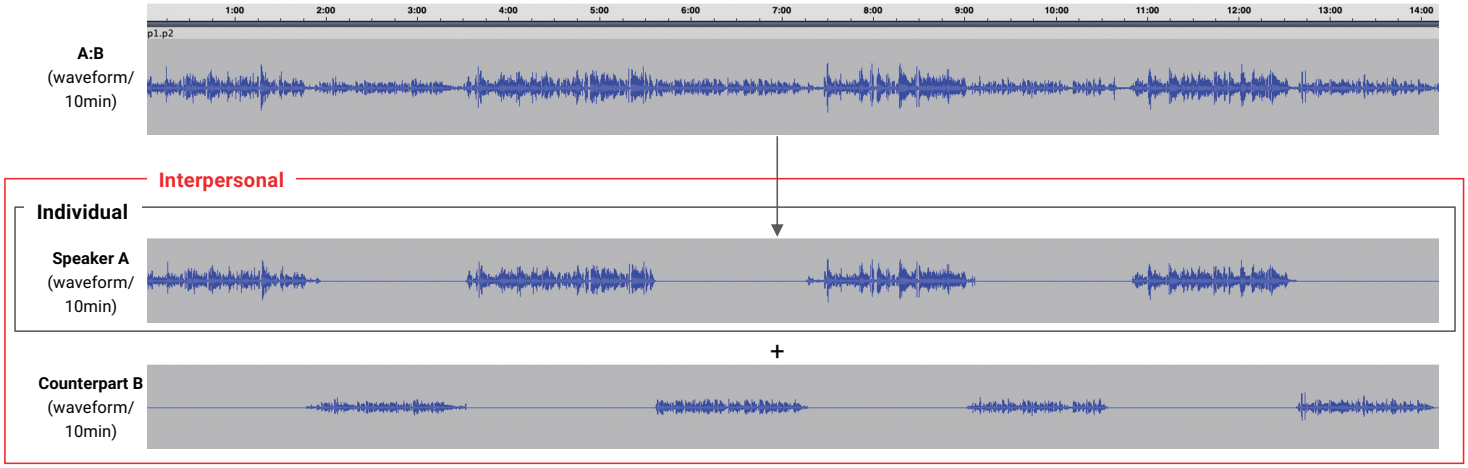

Figure 3.1: The mixed voice data of A and B was preprocessed by dividing it into voice data of A and B. Individual feature uses A's own audio data to predict A's emotions, whereas interpersonal feature uses both A and B's audio data to predict A's emotions.

model. As a result, the ConvNet model was not included in the physiological signals data as part of an effort to make the model as simple as possible. The reason why ConvNet is put only in audio data despite the small amount of data is that ConvNet is a deep learning model specialized in photographic and audio data [16].

We designed to allow the inclusion of personal gender information directly into the final classification layer of the framework. It has been observed that gender is an important confounding factor in detecting depression from voice [29, 30, 53]. This personal information is part of the dataset. K-EmoCon data provide age information as well as a person's gender. However, since the ages of all 12 subjects we will use are different, we found that the model performance was rather low in the process of embedding the age variable. That is why we removed the age information. In terms of gender, the subjects we will use are made up of six males and six females as shown in Table 3.3.

Meanwhile, to check the interpersonal features' impact on the speaker's emotion recognition, we created a counterpart encoder. The audio data is the same as the speaker encoder until it passes through the ConvNet. The gender data is also the same. However, in the case of Counterpart Encoder, audio and physiological signals are concatenated to enter a single BiLSTM network. The reason is to differentiate itself from the speaker encoder. If the speaker encoder and counterpart encoder have the same structure, the deep learning model will have the same sequence value in the same structure, but only the labels will be different when the discussion data of the two subjects are put in each. Thus, we have differentiated the counterpart encoder structure, leading the model to analyze interpersonal features.

Note that in the case of interpersonal features, the data of the speaker and counterpart are used, so the dimension of features is doubled compared to individual features. Therefore, in order to match the number of feature dimensions, individual features were included in both the speaker encoder and counterpart encoder when analyzing individual features.

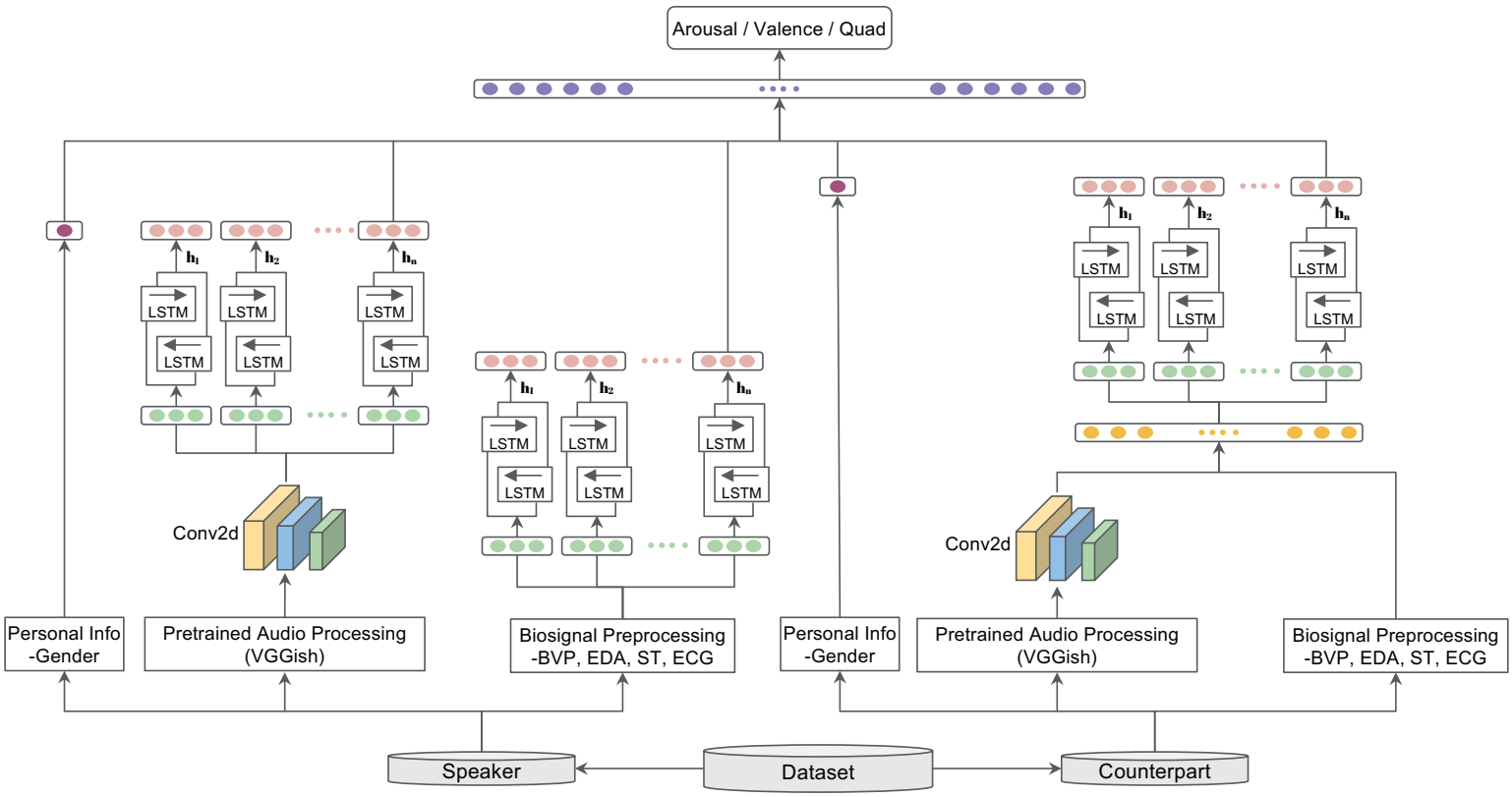

Figure 3.2: Speaker encoder and counterpart encoder for individual and interpersonal analysis.

Chapter 4. Experiment

4.1 Experiment Setup

We were able to obtain a total of 1256 samples by sampling each of the 12 subjects at 5-second intervals. We used Adam optimizer [24] for deep learning training and 0.0001 for the learning rate. We used a batch size of 128. The total epoch was performed 500 times, and to prevent overfitting, the dropout rate was set to 0.4 for both audio and physiological signals.

In the case of ConvNet, features were extracted using the number of filter channels of 1, 32, 64, 128, and 256. Within each ConvNet block, padding and stride were used with kernel sizes of (3, 3). In LSTM, the hidden size was set to 32, and the number of layers was set to 4. All experiments were implemented in the Pytorch framework of Python programming. The models were trained using NVIDIA TITAN RTX GPU.

In this study, emotion prediction was defined as a binary classification that classifies arousal and valence into low and high classes [57]. Since both arousal and valence were collected using a 5-point Likert scale, we defined values corresponding to 1 and 2 as low and values corresponding to 3, 4, and 5 as high. For combined annotations, the self and partner ratings are accumulated and adjusted from 2 to 10, and then converted to 2, 3, 4, 5 as low and 6, 7, 8, 9, and 10 as high. Furthermore, emotions are classified into quad classes with the combination of high arousal high valence (HAHV), high arousal low valence (HALV), low arousal high valence (LAHV), and low arousal low valence (LALV) [57].

For the model evaluation, we used hold-out cross validation for model training and validation. In this case, 80 percent of the data was extracted from each user for training. Therefore, out of 1256 samples, 1000 data were used as training data, and 256 data were used as test data.

To evaluate the performance of classification models, we use accuracy as the underlying metric and use two additional metrics: weighted F1 score and weighted AUROC, the area under the receiver operating characteristic (ROC) curve [12]. F1 Score is a metric combining precision and recall into one with the harmonic mean.

4.2 Baseline models

As previously mentioned, we predicted emotion prediction by binary classification or quad classification, distinguishing emotion from low and high. However, since the K-EmoCon dataset is a relatively recently released dataset, there are not many cases of utilizing this dataset before. Therefore, we proceeded with classification using baseline models to find out if this dataset is a suitable dataset for classification tasks and to what extent our proposed model performs well. We used the most classic machine learning methods including support vector machine (SVC), logistic regression (LR), K-Nearest Neighbor (KNN), and eXtreme Gradient Boosting (XGBoost) [6].

The support vector machine identifies the hyperplane that best divides the two data classes. The kernel function can be used to improve the performance of the SVC by converting the data into another space that can be further divided. We used an SVC kernel with Radial Basis Function (RBF), and gamma equals to 0.5. The logistic regression is a linear model that captures the relationships between the features and classes. We used LR with a solver equal to “lbfgs” and max iteration equal to 8000. The

K-Nearest Neighbor identifies the nearest neighbor in the feature space and uses the class labels of these neighbors to determine the class of the new data instance. Lastly, XGBoost is an algorithm based on gradient boosting, as an ensemble technique, where many weak learners create one huge robust learner while learning additively. Each new learner is based on the residual of the previous learner until further improvement is not possible. For XGBoost, we used default parameters to train the model.

Chapter 5. Results

We performed model training and classification based on two levels of arousal and valence (high or low). Then, we extend it to the quad-class classification based on the four quadrants of the arousal-valence space. Table 5.3 shows the overall accuracy, f1-score, and AUROC results of the proposed model and baseline models using different annotations. Table 5.1 shows the result of comparing the emotion classification performance of the proposed model with the highest value among the baseline models. The parts marked in red show that the performance of the proposed model is superior to the baseline models. From the table, it can be seen that the proposed model using CNN-LSTM shows superior classification performance than baseline models in all respects, regardless of whether it has individual or interpersonal features.
We set the research question to see if the model performance improves when the counterpart's speech and physiological signals are used to predict the speaker's emotions during the naturalistic conversation. Table 5.2 shows the analysis results. In the table, the parts marked in red show the result of subtracting the model performance when using individual features and the model performance when using interpersonal features.

First of all, in the case of majority classification, the same performance is shown regardless of whether it is individual features or interpersonal features because it is classified with the highest frequency label. Therefore, the value of the majority classification appears as 0 in all matrices. For XGBoost, it can be seen that the accuracy and f1 score of the valence does not always conform to our hypothesis in partner annotations. However, the degree to which interpersonal features influence the model is negative is less than 0.01 in all cases. And if only this part is excluded, it can be confirmed that the influence of interpersonal features is positive in all cases when using XGBoost. In the case of SVM, it can be seen that there was no difference in model performance between individual features and interpersonal features in the quad classification. However, in Table 5.3, it can be seen that the SVM performed the classification task to perfectly match the majority classification. In other words, we can say that SVM did not learn properly when performing Quad classification.

In the case of Linear Regression, like XGBoost, the impact of interpersonal features on accuracy and some parts of the F1-score was found to be negative when classifying valence. However, except for this, it was found that in all cases, the influence of the interpersonal features had a positive impact on predicting the speaker's emotions. For KNN and the proposed model, the positive influence of interpersonal features could be found in all matrices.

Table 5.1: The result of comparing the emotion classification performance of the proposed model with the highest value among baseline models. The red-colored portions show that the performance of the proposed model is superior to that of the baseline models.

			Highest performance of any baseline model (A)			Proposed Model – A		
			Self	Partner	Combine	Self	Partner	Combine
Arousal	Individual	Accuracy	0.682	0.693	0.723	0.127	0.086	0.086
		f1 score	0.665	0.664	0.711	0.144	0.099	0.102
		AUC	0.715	0.696	0.770	0.103	0.008	0.003
	Interpersonal	Accuracy	0.785	0.779	0.786	0.041	0.077	0.094
		f1 score	0.776	0.766	0.777	0.050	0.085	0.104
		AUC	0.812	0.812	0.824	0.057	0.001	0.103
Valence	Individual	Accuracy	0.767	0.844	0.752	0.043	0.100	0.043
		f1 score	0.720	0.810	0.712	0.077	0.131	0.084
		AUC	0.672	0.722	0.702	0.147	0.096	0.023
	Interpersonal	Accuracy	0.769	0.836	0.759	0.063	0.131	0.051
		f1 score	0.729	0.805	0.730	0.108	0.162	0.070
		AUC	0.719	0.770	0.752	0.127	0.134	0.027
Quad	Individual	Accuracy	0.533	0.594	0.535	0.059	0.113	0.093
		f1 score	0.478	0.585	0.506	0.095	0.123	0.140
		AUC	0.675	0.687	0.711	0.064	0.012	0.022
	Interpersonal	Accuracy	0.627	0.664	0.596	0.041	0.106	0.064
		f1 score	0.578	0.649	0.572	0.048	0.132	0.109
		AUC	0.760	0.778	0.776	0.039	0.014	0.051

Table 5.2: The difference in emotion classification performance between individual and interpersonal features. The red-colored portions show that the counterpart's audio and physiological signals have a positive effect on the speaker's emotion classification.

		Arousal			Valence			Quad		
		Accuracy	f1 score	AUC	Accuracy	f1 score	AUC	Accuracy	f1 score	AUC
Majority	Self	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Partner	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Combine	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
XGBoost	Self	0.103	0.111	0.097	0.002	0.009	0.047	0.094	0.100	0.085
	Partner	0.086	0.098	0.116	-0.008	-0.005	0.048	0.076	0.093	0.100
	Combine	0.063	0.066	0.054	0.007	0.018	0.050	0.061	0.066	0.065
SVM	Self	0.096	0.102	0.121	-0.006	0.008	0.036	0.000	0.000	0.064
	Partner	0.113	0.114	0.145	-0.020	-0.008	0.040	0.000	0.000	0.089
	Combine	0.043	0.041	0.023	0.002	0.012	0.064	0.000	0.000	0.083
LR	Self	0.109	0.110	0.104	-0.009	0.009	0.072	0.046	0.058	0.105
	Partner	0.102	0.107	0.133	-0.015	-0.002	0.072	0.080	0.093	0.109
	Combine	0.035	0.035	0.029	-0.003	0.013	0.075	0.031	0.040	0.049
KNN	Self	0.094	0.091	0.119	0.066	0.059	0.045	0.121	0.112	0.103
	Partner	0.087	0.083	0.111	0.009	0.010	0.031	0.057	0.064	0.091
	Combine	0.032	0.030	0.040	0.030	0.027	0.035	0.079	0.076	0.073
Proposed model	Self	0.017	0.017	0.051	0.022	0.040	0.027	0.076	0.053	0.060
	Partner	0.077	0.088	0.109	0.023	0.026	0.086	0.063	0.073	0.093
	Combine	0.071	0.068	0.154	0.015	0.004	0.054	0.032	0.035	0.094

Table 5.3: Emotion classification Accuracy, F1-score, and AUROC results of proposed models and baseline models using different annotations.

			Majority			XGBoost			SVM			LR			KNN			Proposed model		
			Self	Partner	Combine	Self	Partner	Combine	Self	Partner	Combine	Self	Partner	Combine	Self	Partner	Combine	Self	Partner	Combine
Arousal	Individual	Accuracy	0.670	0.705	0.705	0.682	0.693	0.723	0.638	0.644	0.670	0.648	0.665	0.686	0.615	0.639	0.666	0.809	0.779	0.809
		f1 score	0.538	0.583	0.583	0.665	0.664	0.711	0.633	0.642	0.673	0.647	0.659	0.685	0.627	0.653	0.680	0.809	0.763	0.813
		AUROC	0.500	0.500	0.500	0.715	0.696	0.770	0.644	0.632	0.711	0.673	0.653	0.715	0.646	0.646	0.690	0.818	0.704	0.773
	Interpersonal	Accuracy	0.670	0.705	0.705	0.785	0.779	0.786	0.734	0.757	0.713	0.757	0.767	0.721	0.709	0.726	0.698	0.826	0.856	0.880
		f1 score	0.538	0.583	0.583	0.776	0.762	0.777	0.735	0.756	0.714	0.757	0.766	0.720	0.718	0.736	0.710	0.826	0.851	0.881
		AUROC	0.500	0.500	0.500	0.812	0.812	0.824	0.765	0.777	0.734	0.777	0.786	0.744	0.765	0.757	0.730	0.869	0.813	0.927
Valence	Individual	Accuracy	0.759	0.834	0.734	0.767	0.844	0.752	0.702	0.807	0.684	0.717	0.811	0.700	0.579	0.742	0.612	0.810	0.944	0.795
		f1 score	0.655	0.759	0.622	0.720	0.810	0.712	0.692	0.798	0.673	0.699	0.800	0.682	0.609	0.765	0.634	0.797	0.941	0.796
		AUROC	0.500	0.500	0.500	0.672	0.722	0.702	0.606	0.697	0.605	0.582	0.670	0.605	0.594	0.682	0.626	0.819	0.818	0.725
	Interpersonal	Accuracy	0.759	0.834	0.734	0.769	0.836	0.759	0.696	0.787	0.686	0.708	0.796	0.697	0.645	0.751	0.642	0.832	0.967	0.810
		f1 score	0.655	0.759	0.622	0.729	0.805	0.730	0.700	0.790	0.685	0.708	0.798	0.695	0.668	0.775	0.661	0.837	0.967	0.800
		AUROC	0.500	0.500	0.500	0.719	0.770	0.752	0.642	0.737	0.669	0.654	0.742	0.680	0.639	0.713	0.661	0.846	0.904	0.779
Quad	Individual	Accuracy	0.504	0.585	0.502	0.533	0.588	0.535	0.504	0.585	0.502	0.477	0.537	0.483	0.450	0.594	0.463	0.592	0.707	0.628
		f1 score	0.338	0.432	0.335	0.478	0.532	0.506	0.338	0.432	0.335	0.463	0.521	0.474	0.452	0.585	0.469	0.573	0.708	0.646
		AUROC	0.500	0.500	0.500	0.675	0.674	0.711	0.435	0.408	0.412	0.614	0.627	0.653	0.643	0.687	0.652	0.739	0.699	0.733
	Interpersonal	Accuracy	0.504	0.585	0.502	0.627	0.664	0.596	0.504	0.585	0.502	0.523	0.617	0.514	0.571	0.651	0.542	0.668	0.770	0.660
		f1 score	0.338	0.432	0.335	0.578	0.625	0.572	0.338	0.432	0.335	0.521	0.614	0.514	0.564	0.649	0.545	0.626	0.781	0.681
		AUROC	0.500	0.500	0.500	0.760	0.774	0.776	0.499	0.497	0.495	0.719	0.736	0.702	0.746	0.778	0.725	0.799	0.792	0.827

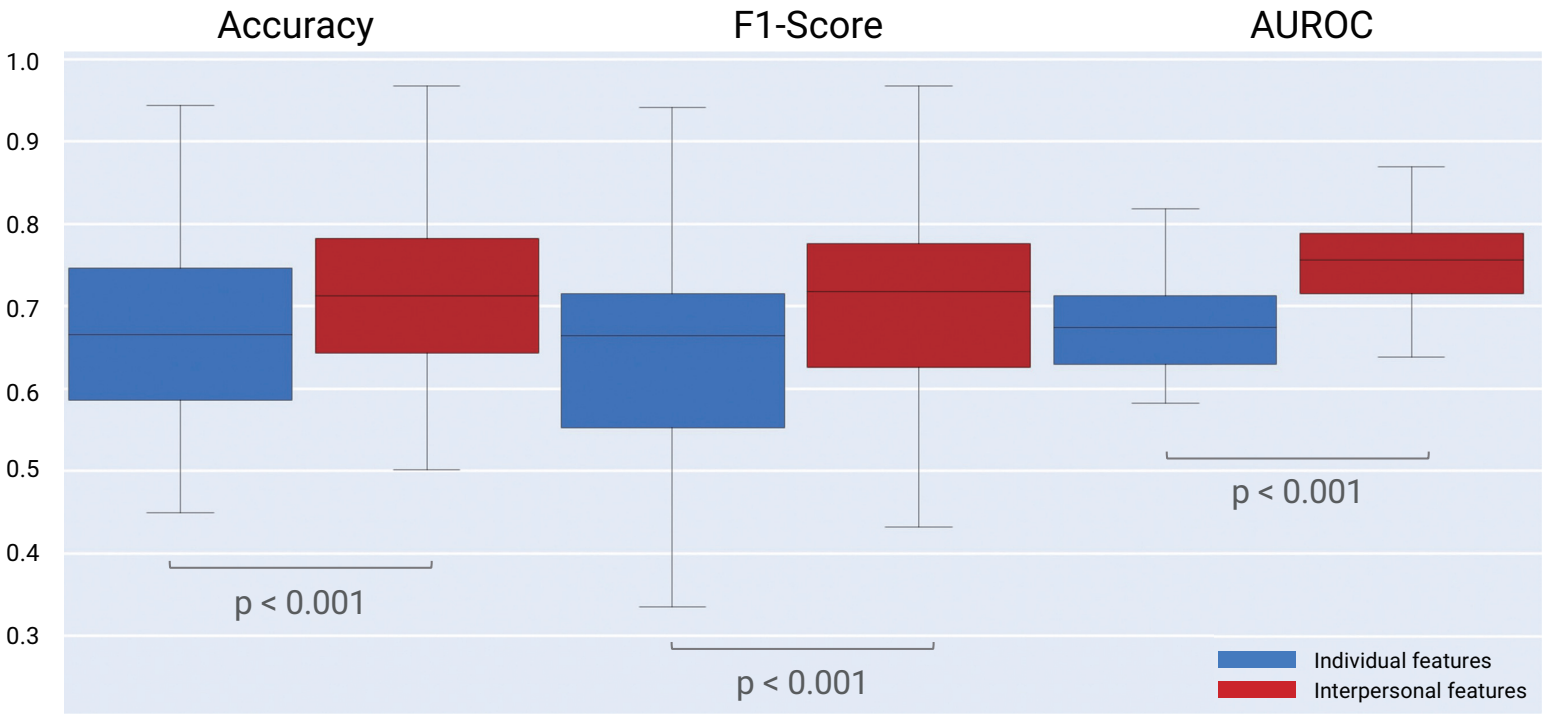

Figure 5.1: Analysis of whether the difference between Accuracy, F1-score, and AUC of using individual features and using interpersonal features is statistically significant.

We also conducted a two-tailed Wilcoxon signed-rank test which is a non-parametric alternative to paired t-test to statistically analyze the impact of interpersonal features dependency. In Figure 5.1, we classified all accuracy, f1-score, and AUROC values into the case of those obtained using individual features and those obtained using interpersonal features. We conducted the test using the SPSS statistical program. As a result, it was confirmed that the p-value was less than 0.001 in all three matrices. Therefore, we were able to statistically prove the positive impact of interpersonal features on the speaker's emotion prediction.

Chapter 6. Discussion

As described in the Methodology section, we conducted experiments using blood volume pulse (BVP), skin temperature (ST), electrocardiograph (ECG), and electrodermal activity (EDA) physiological signals. Including more diverse physiological signals is one of our future studies. In addition, research on which physiological signals have a more important influence on emotion recognition by alternately subtracting physiological signals is also our future work.

There are some points that we would like to discuss the K-EmoCon dataset. This dataset is rare data that labels human emotions every five seconds during a conversation. Quan et el. [40] also mentioned that the K-EmoCon dataset is the only dyadic dataset in which the subjects show spontaneous emotions during naturalistic conversations.

However, some ambiguity was found in the process of preprocessing the dataset. First, the dataset provides a baseline measurement for physiological signals data of each subject for generalization. The physiological signals of the subjects were measured for 1.5 to 2 minutes before the discussion began. However, it seemed that in the process of measuring the physiological sensor, the subjects made an error in turning on the sensor before attaching the sensor. The reason is that if we look at each subject's baseline measurement data, the first few seconds for each subject are always classified as outliers of the data. Since it does not specify how many seconds of error each subject has, it is difficult to use this data as it is. Because of this, the criteria for baseline measurement became ambiguous, so this study did not use this part. Instead of this, we normalized the physiological signals of each subject for generalization.

Second, as shown in Figure 3.4, K-EmoCon provides about nine kinds of physiological signal data. Unlike speech data, where all data are completely accessible without missing values, there are some missing values and errors for physiological signals. In order to train a deep learning model, the more data, the better, so it is necessary to consider which physiological signals to use by referring to the K-EmoCon description paper [37]. In the case of our study, not only the speaker but also the counterpart's physiological signals should be accessible. Therefore, we explored ways to make good use of the characteristics of the physiological signals while preserving the maximum number of data and consequently used a total of four signals(BVP, ST, ECG, EDA) which were fully accessible in 6 sessions out of the total 16 sessions. Since there are only about 1500 data for model training at most, we have simplified the model using only the CNN-LSTM networks even though we originally planned to build the model by applying state-of-the-art deep learning models such as attention.

Third, in this study, the original audio data containing the voices of the two speakers were divided into the voices of each speaker. In this process, we muted the counterpart's utterance part to extract only the speaker's voice, that is, the utterance part of the counterpart no longer has any voice data at all. However, human emotions are recorded even at that moments. This can be an obstacle that hinders model training. When dealing with audio data, we should consider how to cope with this situation of silence before putting it into the model. Solutions can be trying to replace the nearest voice with silence or delete the segments altogether when it is silent. In this study, in the case of interpersonal features, physiological signals and audio signals were combined to minimize the absence of audio data to avoid such cases.

Emotion prediction in naturalistic conversations has many applications. For example, we can think of a loving relationship between a robot and a human. As seen in the movie Her, it is essential to accurately

grasp human emotions in order to elicit communication between robots and humans. Since love involves countless interactions in everyday conversations, using our model to recognize emotions will be of great help in performing successful interactions. Emotional prediction in the conversation will also play an important role in parents who educate children. It is very important for an educator to understand the emotions of the person he/she teaches. Young children, in particular, learn by listening passively, so the parents will be able to communicate more accurately if they use not only children's characteristics but also their voice and physiological signals to predict children's emotions.

Chapter 7. Conclusion

Inspired by the fact that humans recognize emotions through individual features and interpersonal features, this study explored whether interpersonal features are beneficial for emotion prediction. Specifically, we constructed an interpersonal model using speech and physiological signals. We then analyzed whether the characteristics of counterpart positively influence the speaker's emotion prediction by comparing the use of individual features alone with the use of interpersonal features with the K-EmoCon dataset. Our key experimental results show that the model performance when using interpersonal features is higher than that of using individual features. Through our work, we have shown that not only one's own data but also the data of another interacting person can play an important role in predicting a person's emotions. Since humans constantly communicate with others, we support integrating interpersonal features for automatic emotion recognition in natural communication settings. In addition, our study used only audio and physiological signals, but there will be several other things that can affect emotions. Visual factors such as human expressions or contextual factors such as weather and time can also affect emotions. Research that predicts human emotions using these various factors is also considered as our future study.

Bibliography

[1] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe. Decaf: Meg-based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing*, 6(3):209–222, 2015.

[2] G. Assunção, P. Menezes, and F. Perdigão. Speaker awareness for speech emotion recognition. *Int. J. Online Biomed. Eng.*, 16(4):15–22, 2020.

[3] A. Bakhshi and S. Chalup. Multimodal emotion recognition based on speech and physiological signals using deep neural networks. In *International Conference on Pattern Recognition*, pages 289–300. Springer, 2021.

[4] S. Bhukya. Effect of gender on improving speech recognition system. *International Journal of Computer Applications*, 179(14):22–30, 2018.

[5] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008.

[6] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[7] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 2018.

[8] P. Das, A. Khasnobish, and D. Tibarewala. Emotion recognition employing ecg and gsr signals as markers of ans. In *2016 Conference on Advances in Signal Processing (CASP)*, pages 37–42. IEEE, 2016.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587, 2011.

[11] M. Fahad, A. Deepak, G. Pradhan, J. Yadav, et al. Dnn-hmm-based speaker-adaptive emotion recognition using mfcc and epoch-based features. *Circuits, Systems, and Signal Processing*, 40(1):466–489, 2021.

[12] T. Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine learning*, 31(1):1–38, 2004.

[13] H. F. García, M. A. Álvarez, and Á. Á. Orozco. Gaussian process dynamical models for multimodal affect recognition. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 850–853. IEEE, 2016.

[14] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*, 2019.

[15] A. Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

[16] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

[17] P. Gupta, S. A. Balaji, S. Jain, and R. Yadav. Emotion recognition during social interactions using peripheral physiological signals. In *Computer Networks and Inventive Communication Technologies*, pages 99–112. Springer, 2022.

[18] K. Han, D. Yu, and I. Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech 2014*, 2014.

[19] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access, 2018.

[20] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[21] S. Katsigiannis and N. Ramzan. Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE journal of biomedical and health informatics*, 22(1):98–107, 2017.

[22] Y. Kim and E. M. Provost. Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3677–3681. IEEE, 2013.

[23] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[24] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel. Emotion recognition and its application in software engineering. In *2013 6th International Conference on Human System Interactions (HSI)*, pages 532–539. IEEE, 2013.

[25] S. G. Koolagudi and K. S. Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117, 2012.

[26] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[27] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9-10):1162–1171, 2011.

[28] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan. Emotion recognition based on phoneme classes. In *Eighth international conference on spoken language processing*, 2004.

[29] Z. Liu, D. Wang, L. Zhang, and B. Hu. A novel decision tree for depression recognition in speech. *arXiv preprint arXiv:2002.12759*, 2020.

[30] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586, 2010.

[31] V. Markova, T. Ganchev, and K. Kalinkov. Clas: A database for cognitive load, affect and stress recognition. In *2019 International Conference on Biomedical Innovations and Applications (BIA)*, pages 1–4. IEEE, 2019.

[32] H. P. Martinez, Y. Bengio, and G. N. Yannakakis. Learning deep physiological models of affect. *IEEE Computational intelligence magazine*, 8(2):20–33, 2013.

[33] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17, 2011.

[34] E. Mower, M. J. Matarić, and S. Narayanan. A framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1057–1070, 2010.

[35] J. Ocumpaugh. Baker rodrigo ocumpaugh monitoring protocol (bromp) 2.0 technical and training manual. *New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences*, 60, 2015.

[36] G. Paltoglou and M. Thelwall. Seeing stars of valence and arousal in blog posts. *IEEE Transactions on Affective Computing*, 4(1):116–123, 2012.

[37] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):1–16, 2020.

[38] B. Parkinson. Interpersonal emotion transfer: Contagion and social appraisal. *Social and Personality Psychology Compass*, 5(7):428–439, 2011.

[39] K. Plarre, A. Raij, S. M. Hossain, A. A. Ali, M. Nakajima, M. Al'Absi, E. Ertin, T. Kamarck, S. Kumar, M. Scott, et al. Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pages 97–108. IEEE, 2011.

[40] J. Quan, Y. Miyake, and T. Nozawa. Incorporating interpersonal synchronization features for automatic emotion recognition from visual and audio data during communication. *Sensors*, 21(16):5317, 2021.

[41] H. Ranganathan, S. Chakraborty, and S. Panchanathan. Multimodal emotion recognition using deep learning architectures. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[42] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.

[43] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[44] K. R. Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[45] B. Schuller, G. Rigoll, and M. Lang. Hidden markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 2, pages II–1. Ieee, 2003.

[46] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific data*, 6(1):1–13, 2019.

[47] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang. A review of emotion recognition using physiological signals. *Sensors*, 18(7):2074, 2018.

[48] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 3(1):42–55, 2011.

[49] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 9(2):147–160, 2016.

[50] T. Vogt and E. André. Improving automatic emotion recognition from speech via gender differentiaion. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 2006.

[51] Y. Wang and J. Mo. Emotion feature selection from physiological signals using tabu search. In *2013 25th Chinese Control and Decision Conference (CCDC)*, pages 3148–3150. IEEE, 2013.

[52] M. Xu, F. Zhang, and S. U. Khan. Improve accuracy of speech emotion recognition with attention head fusion. In *2020 10th annual computing and communication workshop and conference (CCWC)*, pages 1058–1064. IEEE, 2020.

[53] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli. Decision tree based depression classification from audio video and language information. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 89–96, 2016.

[54] L. Zhang, L. Wang, J. Dang, L. Guo, and Q. Yu. Gender-aware cnn-blstm for speech emotion recognition. In *International Conference on Artificial Neural Networks*, pages 782–790. Springer, 2018.

[55] S. Zhang, S. Zhang, T. Huang, and W. Gao. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, 20(6):1576–1590, 2017.

[56] J. Zhao, X. Mao, and L. Chen. Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical signal processing and control*, 47:312–323, 2019.

[57] M. S. Zitouni, C. Y. Park, U. Lee, L. Hadjileontiadis, and A. Khandoker. Arousal-valence classification from peripheral physiological signals using long short-term memory networks. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 686–689. IEEE, 2021.