IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 July 2023, accepted 18 July 2023, date of publication 21 July 2023, date of current version 28 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3297715

RESEARCH ARTICLE

Emotion-Aware Speaker Identification With Transfer Learning

KYOUNGJU NOH AND HYUNTAE JEONG
Electronics and Telecommunications Research Institute, Yuseong-gu, Daejeon 34129, Republic of Korea

Corresponding author: Kyoungju Noh (kjnoh@etri.re.kr)

This work was supported in part by the Electronics and Telecommunications Research Institute (ETRI) funded by the Korean Government, and in part by the Core Technology Research for Self-Improving Integrated Artificial Intelligence System under Grant 23ZS1100.

ABSTRACT Speech is a natural communication method used by humans. Speaker identification (SI) technology based on human speech has been used as an entry point for many human–computer-interaction applications. The performance of SI models can degrade when dealing with expressive speech uttered in emotional situations because emotion databases do not have sufficient data on expressive speech to train SI models for various emotional states. Generally, SI models are trained using relatively more samples of ''neutral'' speech than samples of other emotion classes. In this study, we propose an emotion-aware SI (em-SI) method that uses an emotion-embedding vector generated from a pre-trained speech emotion recognition (SER) model along with the acoustic features of speech data. We assess the performance of this method using individual English and Korean corpora and confirm that the proposed method provides an improved performance on multilingual corpora. The evaluation results show that the SI accuracy of em-SI on the Korean Emotion Multimodal Database (KEMDy19) improved by 3.2%, and the average speaker verification (SV) performance in terms of the equal error rate (EER) was improved by 1.3% compared to that of the baseline SI model. The visualization of the embedding vector of em-SI shows that em-SI maps speech data to an embedding space where both SI and emotional information are simultaneously represented. Through the experiments conducted in this study, we confirmed that the em-SI model, which learns by integrating emotion and speaker embedding information, improved the performance of SI for expressive speech.

INDEX TERMS Affective computing, emotion-aware speaker identification, Korean emotion database, multitask learning, speech emotion recognition, speech processing.

I. INTRODUCTION

Speaker recognition (SR), which distinguishes individuals based on speech, is a fundamental and performance-sensitive topic in natural human–computer interactions. Each human has a voiceprint, which is an acoustic characteristic of their unique voice-producing organs and speaking patterns. Regarding SR technology, it can be classified as either speaker identification (SI), which recognizes a speaker within a specific set of speakers, or speaker verification (SV), which determines whether a speaker is a specific person [1], [2].

The role of SI in emotional talking environments has garnered significant interest in human–computer interaction and affective computing research. Integrating emotion recognition and SI for expressive speech can improve the quality of computer responses, in terms of adaptability and reactivity to users in human–robot interfaces and intelligent call centers [1], [2], [3].

When a speaker utters expressive speech in an emotional situation, there are variations in the waveform, prosody, spectral characteristics, accent, speech rate, and syllable rate of the speech [4], [5], [6]. These variations in acoustic characteristics reduce the performance of SI during expressive speech and pose challenges when compared to neutral state utterances in terms of the emotional expression, speech rate, and loudness of vocalizations [1], [2], [4], [7], [8]. The performance degradation of SI during expressive speech is due to the difference between the emotion classes for

The associate editor coordinating the review of this manuscript and approving it for publication was Lin Wang.

enrollment data, which represent the training data, and the test data for each speaker in the SI model [9], [10]. Using speech data belonging to the same emotional state for both training and testing an SI model can improve the model's performance; however, collecting balanced speech data corresponding to the various emotions of each speaker is a challenging task owing to the associated high costs and time consumption.

Previous studies have explored the mutual dependencies between emotion and speaker recognition in the context of expressive speech [9], [10], [11]. Motivated by these dependencies, we investigate a multilabel learning structure that can simultaneously perform emotion recognition and SI based on an expressive speech dataset.

We hypothesize that, if an SI model could learn the emotional information and voiceprint of a speaker's utterance, it would help prevent the degradation of the SI performance for expressive utterances. Accordingly, this study proposes an emotion-aware SI (em-SI) method that incorporates an emotion-embedding vector into the acoustic features of speech data without neutralizing the emotional information expressed in a speech segment [9].

We evaluated the em-SI model, which is based on a deep-learning structure, for various expression utterances in three separate emotion databases in English and Korean. The enhanced performance of the em-SI model on all three databases was confirmed through single-corpus and multilingual corpora experiments. There was a 3.2% improvement in SI accuracy in the evaluation of 40 speakers from a Korean emotion multimodal database, namely the Korean Emotion Multimodal Database in 2019 (KEMDy19). Additionally, the average equal error rate (EER) of SV was reduced by 1.3% compared to that of the baseline SI model that did not use an emotion embedding vector. The contributions of this study are summarized as follows.

A. EMBEDDING SPACE FOR EMOTION AND SI

We evaluated SI performance based on the dependency between the emotions in training and test data for an SI model. We then proposed the em-SI method to improve the SI model for expressive utterances using an unbalanced emotion database. The proposed em-SI model uses the emotion embedding vector transferred from the pre-trained SER model. The deep-learning-based em-SI model learns the emotion- and speaker-embedding spaces of expressive speech without intentionally blurring the emotional information expressed in each utterance [9]. In the deep-learning structure of the em-SI model with the transferred SER model, the final embedding vector reflects the emotional context and voiceprint characteristics included in the utterance.

B. BIDIRECTIONAL LONG SHORT-TERM MEMORY (Bi-LSTM)-BASED PRE-TRAINED SER AND em-SI NETWORKS FOR MULTILABEL RECOGNITION

The Bi-LSTM-based pre-trained SER and em-SI networks operate on the same sequence of 56-dimensional (D) input features, which typically include the Mel-frequency cepstral coefficient (MFCC), Mel-spectrogram, zero-crossing rate, spectral roll-off, and spectral centroid. The Bi-LSTM networks of em-SI enable the simultaneous recognition of emotion labels and the SI of speech data.

C. TRANSFERRED SER MODEL OF A MULTITASK LEARNING STRUCTURE

We implemented the pre-trained SER model based on a multitask learning (MTL) structure using a weighted loss function to prevent overfitting to a specific emotion label. The MTL SER model could simultaneously learn categorical emotion labels and arousal and valence levels from speech data.

D. OPEN KOREAN EMOTION MULTIMODAL DATABASES

The proposed em-SI method was evaluated based on a single corpus and multilingual corpora with the English-speaking Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [12] and two Korean emotion multimodal databases, namely KEMDy19 [13] and Korean Emotion Multimodal Database in 2020 (KEMDy20) [14]. We published the Korean-based databases to be freely available on a data-sharing website [13], [14].

The remainder of the study is organized as follows. Section II describes related studies performed on expressive speech for SR and SER. Section III presents the operational flow of the proposed em-SI model based on Bi-LSTM. Section IV describes the data preparation and experimental setup procedures, and Section V thoroughly discusses the experimental results. Finally, Section VI concludes this study and suggests potential directions for future work.

II. RELATED STUDIES

A. SR ON EXPRESSIVE SPEECH

Many studies have reported that SR performance deteriorates in expressive speech uttered in emotional situations, particularly those involving extreme arousal and valence levels [1], [2], [4], [7], [8], [9], [10].

An x-vector is a representative method of fixed-length speaker embeddings [15] that is generated through a pooling layer that extracts statistical information about frame-level feature vectors extracted from a deep neural network (DNN). Many studies [15], [16], [17] have reported that x-vectors can outperform traditional generative models such as Gaussian mixture models (GMMs) [18] and i-vectors [19] in terms of SI performance.

Previous studies have analyzed the dependencies of emotion and speaker recognition on expressive speech using x-vectors or i-vectors, with the aim of improving the performance of SER or SR models specifically for emotional speech [9], [10]. Sarma et al. [9] proposed an approach for generating an emotion-invariant speaker embedding method for SI in emotional speech. This method transformed i-vectors extracted from different emotions into the i-vector space of neutral emotions through an encoder and decoder network structure. Using the generated emotion-invariant speaker embedding with an input of 39-D (13-base+13-Δ+13-ΔΔ) MFCC features of augmented speech data, an SI improvement of 2.6% on the IEMOCAP was observed.

Pappagari et al. [10] described the results of SER performance improvements by fine-tuning the pre-trained ResNet-based x-vector model using the 23-D MFCC of augmented speech data as an input vector. They determined that SV performance was sensitive to changes in emotion via experiments involving the EERs of SV. When test and enroll utterances of differing emotion classes were used, the ERR was worse than that when utterances of the same emotion class were used. This study deduced that the performance of SV on the Most Significant Point (MSP)-Podcast [20] corpus, which comprises longer utterances from many speakers, was better than the performance on IEMOCAP, which includes many short utterances within 4 s. Although CREMA-D [21] consisted of data shorter than those in IEMOCAP, the EER of SV for CREMA-D was better than that for IEMOCAP. Pappagari et al. attributed this observation to the phonetic content variability of CREMA-D being limited to only 12 sentences.

Recently, studies have been conducted based on end-to-end deep learning models that generate fixed-size segment embeddings from speech data by combining convolutional neural networks (CNNs) or recurrent neural networks (RNNs), often replacing the traditional DNN structures used in x-vectors. The results of these studies represent the latest advancements in SI performance [8], [10], [22], [23], [24], [25], [26]. Meftah et al. [22] evaluated SI performance through English and Arabic corpora based on the combined structure (CRNN) of CNN and LSTM models that used spectrogram input features. They demonstrated an improvement in the SI performance on the expressive speech of the proposed CRNN model in a single-corpus-based evaluation of English and Arabic. They attributed the degradation of the CRNN-SI performance on the multilingual corpora to the difference in corpus size between the Arabic and English-based databases.

Garain et al. [24] introduced a golden-ratio-aided neural network (GRaNN) with an MTL structure in which emotion recognition, gender, and SI were processed simultaneously. They used a wrapper-filter-based feature-selection technique to select the input features for the three tasks with minimum redundancy and maximum relevance. They adopted the golden ratio [27] to determine the number of units in each layer and demonstrated the performance improvement of the GRaNN system based on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) corpus [28]. Similar to the MTL structure utilized in a previous study, each task model shared a common network and training dataset, making it difficult to expand or change the structure or training dataset for a single emotion or SI task [29].

Transfer-learning-based structures utilize features generated from a pre-trained model that has already been trained in a specific source domain for the target domain [10], [30], [31]. In the transfer learning model, the features generated from the pre-trained model can be used to improve performance in learning the target domain with limited data by reflecting the learning features of the source domain data. Zheng et al. [30] presented a pre-trained indeterminate speaker representation model (PRISM) consisting of a time-delayed neural network and convolutional transformer encoder layers. They suggested that the PRISM represented a speaker utterance as an indeterminate "floating" vector trained using frame-contrastive loss and that it could be transferred and used for various downtasks. The PRISM outperformed the fixed x-vector in the speaker diarization downtask.

In this study, we propose an em-SI structure that supports improvements in SI performance based on expressive speech. To the best of our knowledge, few studies have utilized pre-trained SER models for SI. The LSTM-based speaker embedding of em-SI learns the emotion embedding transferred from the pre-trained SER model along with the acoustic features of speech. The deep speaker embedding of em-SI transforms expressive speech into the speaker representation space that simultaneously reflects emotion and SI without neutralizing the emotional information included in a speech segment. The transfer learning structure of the em-SI method facilitates an independent optimization process such as learning about separate source domains other than the target domain or changing the network structure for each task.

B. SPEECH EMOTION RECOGNITION

The em-SI model uses the transferred emotion-embedding vector from the SER model along with the acoustic features of an utterance. Because an SER model accurately learns emotion representation using expressive speech data, the em-SI system can more effectively learn the speech-based emotional expression characteristics of each speaker.

Generally, SER aims to predict emotion labels defined in discrete [32] and dimensional emotion spaces [33]. In the discrete emotion space, emotion categories, namely "sad," "happy," "fear," "anger," "disgust," and "surprise," are defined. In the dimensional emotion space, valence and arousal are tagged as numerical values that present time-varying emotional states in a continuous emotion space [34]. Valence refers to the level of positive or negative emotional states, and arousal refers to the degree of emotional activation [35].

Recent state-of-the-art SER models are based on deep learning models, such as DNN [36], [37], RNN [38], [39], and CNN [40], as well as a combination of one or more DNN, CNN, and RNN systems [41], [42].

Considering the ambiguity of emotion labels and data imbalance of each emotion class in emotion databases is challenging when constructing a deep learning-based

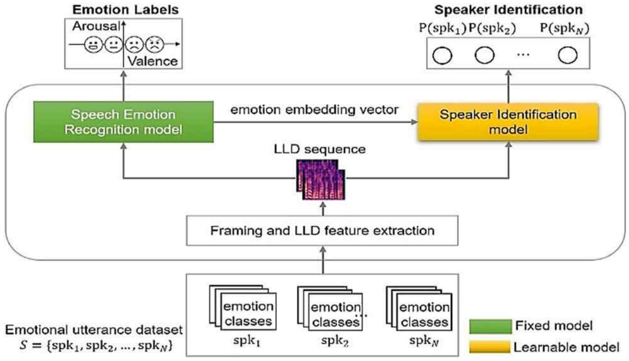

Emotion Labels
Arousal
Valence
Speaker Identification
P(spk$_1$) P(spk$_2$) P(spk$_N$)
...
Speech Emotion Recognition model
emotion embedding vector
Speaker Identification model
LLD sequence
Framing and LLD feature extraction
Emotional utterance dataset
$S = \{spk_1, spk_2, \ldots, spk_N\}$
emotion classes
spk$_1$
emotion classes
spk$_2$
emotion classes
spk$_N$
🟩 Fixed model
🟨 Learnable model

FIGURE 1. Workflow of the em-SI model, which uses an emotion embedding vector transferred from the pre-trained SER model.

SER model. This ambiguity of the emotion labels arises from the uncertainty of true labels. Emotion labels tagged by external observers may not be the same as those determined by self-reports of the same speech data [43]. Furthermore, it is expensive and time-consuming to collect large-scale balanced emotional speech data for the various emotional situations of a speaker [44], [45]; therefore, most emotion databases have data imbalance problems for emotion labels.

Mallol et al. [46] proposed a semi-supervised learning model that was trained to reduce the bias of annotators by manually combining and automatically tagged labels using a label classifier. This model, which used the emotion labels predicted by the label classifier model for emotion learning, responded to the uncertainty problem of emotion labels in a manner that did not entirely depend on the tagged emotion labels.

An MTL-based SER model trained to simultaneously predict multiple emotion labels can prevent the model from overfitting to a certain type of label or enhance the SER performance. Parthasarathy et al. [29], [47] presented an MTL-based model that could learn the arousal, valence, and dominance labels of a dimensional emotion space by placing weights on the corresponding emotion labels. Chen et al. [48] proposed an MTL-based SER model using bottleneck vectors from dimensional label learning networks to predict discrete labels.

The SER model used in this study was implemented based on an MTL structure to respond to the ambiguity of emotion labels and improve the generalization of emotion representation [29]. The MTL SER model can simultaneously learn emotion labels in discrete and dimensional emotion spaces. Moreover, we adopted the weighted cross-entropy (CE) [49] loss with weights for each emotion category in the discrete space by considering the imbalance problem in the training dataset for the MTL SER model. The applied weighted CE was intended to mitigate overfitting to emotion categories with relatively large amounts of data when the SER model learns emotion representations.

III. EMOTION-AWARE SI

A. SI WITH TRANSFERRED SER

In this study, we have proposed an em-SI model to improve SI for expressive speech uttered in various emotional situations. The operational flow in Fig. 1 reveals that the em-SI model uses the transferred emotion-embedding vector generated from a pre-trained SER model. The learnable em-SI model was trained using the transferred emotion-embedding vector from the fixed SER model and acoustic features for the corresponding speech data. The pre-trained SER and em-SI models used the same 56-D acoustic low-level descriptor (LLD) sequence of speech data as the input.

The speech sample and label spaces are denoted as X and Y, respectively; the emotional speech database is denoted as $D = \{D_1, D_2, \ldots, D_k\}$, where k is the name of the speech database. This study assumes a supervised learning environment wherein each speech sample is labeled with SI information as well as common emotion labels that include both the emotion category and the arousal and valence levels. Each emotion database consists of pairs, denoted by $D_k = \left\{\left(X_u^n, (y_{si,u}^n, (y_{c,u}^n, y_{a,u}^n, y_{v,u}^n))\right)\right\}_{n=1}^{N_k}$, where N_k is the number of speech samples in the k emotional database. The n-th speech input belonging to the speaker u, X_u^n, has multiple training labels of the SI, $y_{si,u}^n$; the emotion category is $y_{c,u}^n$ (e.g., "happy" and "sad"); and the arousal and valence levels are denoted by $y_{a,u}^n$ and $y_{v,u}^n$, respectively.

B. Bi-LSTM-BASED NETWORKS

In reference to our previous study [38] on SER, which resulted in an improved SER performance compared to the results of other SER models based on IEMOCAP, we utilized an LSTM-based network structure and 56-D input features in this study. We adopted a Bi-LSTM-based network comprising 128 cells in each direction for the SI and SER models, as illustrated in Fig. 2. We implemented the same Bi-LSTM-based network for both the SER and SI models, aiming to focus on and learn from the temporal features of the speech data using a frame-level RNN. The same network structure was used for both models to prevent the models from obtaining additional information extracted from network differences that could interfere with the results of the em-SI experiments.

The transferred SER and em-SI models, denoted as T_{SER} and E_{SI}, respectively, used the same 56-D LLD feature sequence frame-by-frame as the input for speech data. The 56-D LLD per frame of each speech comprised the 13-D MFCC and the 40-D Mel-spectrogram, along with 3-D time- and frequency-domain LLDs, such as the zero-crossing rate, spectral roll-off, and spectral centroid. The 56-D LLD per frame was extracted by applying sliding windows of 200 ms with a 50% shift for obtaining frequency decomposition results in the speech data. An LSTM structure that could learn inter-frame changes of utterances was applied, and delta features were not used as additional inputs to represent the inter-frame changes, as in the study by Pappagari et al. [10].

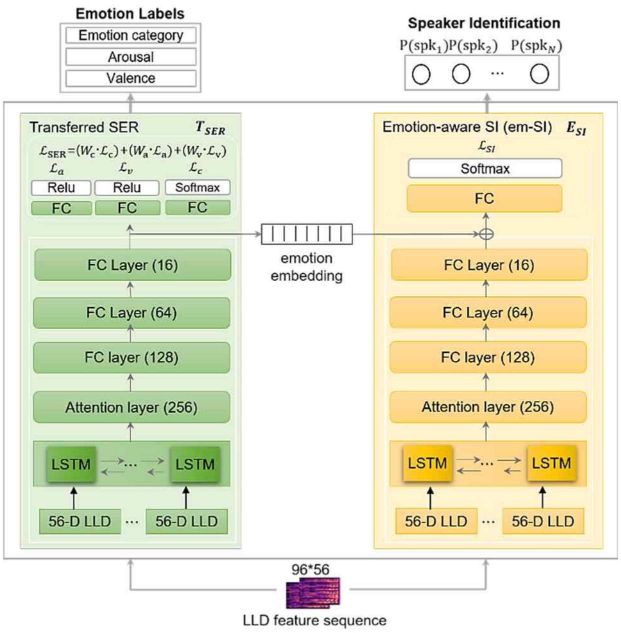


FIGURE 2. Architecture of the bidirectional long short-term memory (Bi-LSTM)-based network for the pre-trained SER and em-SI models.

We padded with zero values to obtain a fixed number of 96 frames and an input sequence of 96×56 per speech sample. The padded sequence was input into the T_{SER} and learnable em-SI model E_{SI}.

The attention mechanism [50] implemented in the Bi-LSTM network focuses on the more discriminative parts of the Bi-LSTM output sequence before the activation of emotion recognition or speaker detection. This attention layer focuses on the relevant parts of the Bi-LSTM output sequence by assigning weight scores and generating high-level contextual vectors.

The generated context vector was transmitted into three fully connected (FC) layers with hidden node sizes of 128, 64, and 16. The last FC layer was followed by an activation function for T_{SER} and E_{SI}. The 16-D embedding vectors in the last FC of T_{SER} and E_{SI} were $\mathbb{R}_{SER}^{16}$ and $\mathbb{R}_{SI}^{16}$, respectively.

The SER model, T_{SER}, was trained to predict the arousal and valence levels and the emotion categories of the corresponding speech data by inputting the 56-D LLD vector. The T_{SER} produced the 16-D emotion embedding vector on the embedding space denoted as $T_{SER}:X\rightarrow\mathbb{R}_{SER}^{16}$. The em-SI model E_{SI} performed SI based on the 32-D combined vector with the SI embedding vector $\mathbb{R}_{SI}^{16}$ and emotion-embedding vector $\mathbb{R}_{SER}^{16}$ for the n-th speech data. Regarding SI using the em-SI system based on the combined feature space, $E_{SI}:\mathbb{R}_{SER}^{16}\oplus\mathbb{R}_{SI}^{16}\rightarrow \mathrm{y}_{si,u}^{n}$.

This study assumes an SI model that uses only the SI embedding vector of $\mathbb{R}_{SI}^{16}$ without the emotion embedding as the baseline model, $\mathrm{Baseline}_{SI}:\mathbb{R}_{SI}^{16}\rightarrow \mathrm{y}_{si,u}^{n}$, for the experiments evaluating the em-SI performance.

C. MULTITASK LEARNING SER

We implemented an MTL SER model to prevent the overfitting of specific emotion labels by considering the uncertainty of emotion labels. The SER model was trained to predict the emotion category and arousal and valence levels for speech data using the shared and task-dependent layers. The MTL SER was trained based on the total loss $\mathcal{L}_{SER}$, which is the sum of the losses of each task multiplied by the weight of the task loss. $\mathcal{L}_c$, $\mathcal{L}_a$, and $\mathcal{L}_v$ are the losses of the recognition tasks for the emotion category, arousal level, and valence level, respectively, and W_c, W_a, and W_v are the weights for these losses, respectively.

In this study, the mean square error (MSE) loss for the arousal and valence recognition task losses were $\mathcal{L}_a$ and L_v,respectively, and W_c, W_a, andW_v were set as 0.5, 0.3, and 0.3, respectively, for $\mathcal{L}_{SER}$. The default values for these weights were arbitrarily determined to balance the categorical emotion loss with the dimensional loss for the arousal and valence levels, and these weights were not optimized for the employed dataset.

$$\mathcal{L}_{SER} = (W_c \cdot \mathcal{L}_c) + (W_a \cdot \mathcal{L}_a) + (W_v \cdot \mathcal{L}_v) \tag{1}$$

We implemented the transferred SER using the weighted CE for the emotion recognition loss $\mathcal{L}_c$. It compensated for model overfitting to the data in the "neutral" class by lowering the weight on the CE loss for this class. The weighted CE $\mathcal{L}_c$ is the sum of the loss CE (c) multiplied by the weights γ (c) of each emotion category c.

In our experiment, to compensate for model overfitting when the "neutral" class accounted for the majority of the training data, the weight value $\gamma\left(\text{“neutral''}\right)$ for the "neutral" class was set to 0.5, and the weights for the other three emotion classes ("angry," "sad," and "happy") were set to 1.0 for $\mathcal{L}_c$.

$$\mathcal{L}_c = \sum\nolimits_{c=1}^{C} CE(c) \cdot \gamma(c) \tag{2}$$

IV. EXPERIMENTAL SETUP

A. EMOTION DATABASES

We experimentally verified the performance of the proposed em-SI method on three separate emotion databases, namely IEMOCAP [12], which is a widely used English-based benchmark database for emotional speech research, and the Korean emotion multimodal databases, KEMDy19 and KEMDy20. We collected KEMDy19 and KEMDy20 using procedures approved by the Institutional Review Board of the Korea National Institute for Bioethics Policy (KoNIBP) in 2019 and 2020, respectively.

Table 1 lists the language, number of speakers, utterance type, data modality, distribution of speech data, and distribution of the arousal and valence levels for the four emotion categories in the three databases. We employed the entire voiced and unvoiced parts of the speech data [51] of the IEMOCAP and KEMD databases as inputs to the pre-trained SER and em-SI models.

TABLE 1. Emotion multimodal databases.

Property		IEMOCAP	KEMDy19[a]	KEMDy20[b]
Language		English	Korean	Korean
Speakers		10 (5 male, 5 female)	40 (20 male, 20 female)	80 (46 male, 34 female)
Utterance type		Acted (scripted/improvised)	Acted (scripted/improvised)	Free talking
Modality		Speech, transcription, motion-capture data	Speech, transcription, physiological data	Speech, transcription, physiological data
Speech data length	2–5 s	63%	47%	47%
	5–10 s	27%	33%	30%
	10–15 s	5%	10%	11%
	15–30 s	5%	10%	12%
	Total	100%	100%	100%
Emotion class	Neutral	35.3%	37.2%	86.5%
	Angry	24.4%	21.1%	1.1%
	Sad	24.4%	8.7%	0.9%
	Happy	13.1%	14.6%	9.5%
	Surprise	2.0%	9.9%	1.1%
	Fear	0.8%	4.3%	0.3%
	Disgust	0.1%	4.2%	0.5%
Valence (mean ± std.)	Angry	1.89 ± 0.53	1.78 ± 0.38	2.12 ± 0.31
	Happy	3.94 ± 0.45	4.33 ± 0.37	3.83 ± 0.33
	Neutral	2.95 ± 0.49	2.94 ± 0.60	2.99 ± 0.44
	Sad	2.24 ± 0.56	1.89 ± 0.53	2.44 ± 0.43
Arousal (mean ± std.)	Angry	3.69 ± 0.66	3.81 ± 0.59	3.73 ± 0.34
	Happy	3.16 ± 0.6	3.90 ± 0.53	3.25 ± 0.3
	Neutral	2.79 ± 0.52	2.99 ± 0.33	3.25 ± 0.36
	Sad	2.61 ± 0.61	2.63 ± 0.65	2.97 ± 0.44

[a] KoNIBP approval number for KEMDy19: P01-201907-22-010
[b] KoNIBP approval number for KEMDy20: P01-202009-12-001

The IEMOCAP database had a higher distribution than the KEMDy19 and KEMDy20 databases for short speech data less than 5 s. Compared with the data in KEMDy20, the multimodal data collected in IEMOCAP and KEMDy19 during the induced emotional vocal performance of an actor were more uniformly distributed for each emotion class. The data in KEMDy20 included the free talking of adults not engaged in acting; here, the speech data for the "neutral" emotion class accounted for 86.5% of the total speech data.

The IEMOCAP database was organized into five sessions, and the multimodal audio, visual, and textual data were collected during dyadic interactions involving ten voice actors. A pair of actors performed dialogue interactions based on scripted scenarios and improvised emotionally in multiple situational plays in the IEMOCAP sessions. The six human annotators evaluated the categorical emotion labels and the labels for the arousal and valence levels on a five-point Likert-like scale [34] for the speech data.

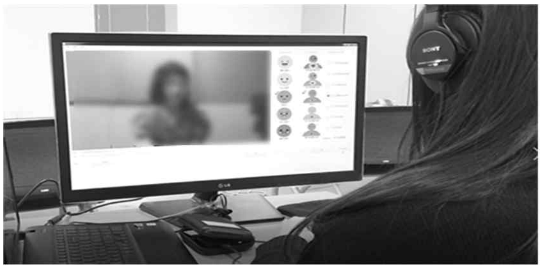
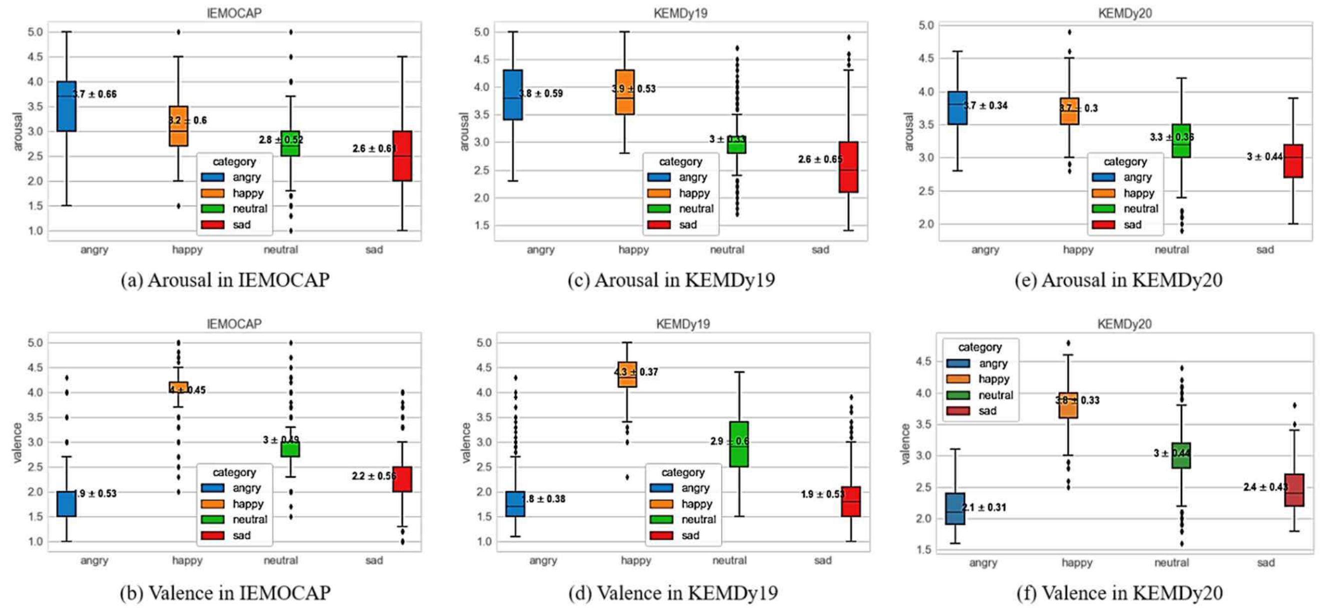


FIGURE 3. Emotion labels for the speech data are tagged while a recorded video of a speaker is being watched using the KEMD annotation application.

The KEMDy19 database is a Korean multimodal database created using a collection procedure similar to that of IEMOCAP. The KEMDy19 database consists of 20 sessions, each containing ten emotional situational plays lasting between 4 to 10 min. The six plays were based on scenarios written to elicit specific emotions, whereas the remaining four acting plays were improvised emotional situation plays. Twenty male and twenty female voice actors performed situational acting in Korean based on scripted and improvised situations. We collected speech data and physiological signals, such as the electrocardiogram (ECG) signals from the Refit patch U9 [52] and electrodermal activity (EDA) from the Empatica E4 [53] wristband device worn by the voice actor. The situations played by the voice actors were recorded as videos for emotion-label tagging. Ten external human annotators performed emotion label tagging for the data in KEMDy19 while watching the recorded video using a tagging application, as shown in Fig. 3. For each piece of speech data, the tagger assigned one of seven categorical emotion labels ("angry," "happy," "neutral," "sad," "surprised," "fear," and "disgust") and arousal and valence category labels based on the 5-point scale.

The KEMDy20 database is a Korean multimodal database comprising data from 80 adults who were not trained actors collected over 40 sessions. In each session, two participants watched a video on a specific topic for approximately 5 min and shared a conversation regarding the video topic, and then, they had a free conversation with their counterparts for approximately 5 min. Each pair of participants in a session repeated the process of watching and talking about six videos. The speech data and biosignals of the photoplethysmogram (PPG) generated from Empatica E4 were collected from all the participants during their free conversations. The emotion label of KEMDy20 was assigned to the emotion category and arousal and valence levels by ten external evaluators based on the video that had been recorded in the same manner as the videos in the KEMDy19 database.

Fig. 4 shows the mean and standard variation of the arousal and valence levels on the five-point scale for the four emotion categories "happy," "angry," "neutral," and "sad" of IEMOCAP, KEMDy19, and KEMDy20 (Table 1).



(a) Arousal in IEMOCAP

(c) Arousal in KEMDy19

(e) Arousal in KEMDy20

(b) Valence in IEMOCAP

(d) Valence in KEMDy19

(f) Valence in KEMDy20

FIGURE 4. Mean and standard variations of the arousal and valence levels on the five-point Likert-like scale for the four emotion classes of the IEMOCAP, KEMDy19, and KEMDy20 databases.

We could assert that the emotion categories had regular relationships with arousal and valence in all three databases. The arousal was highest in the emotion class "anger" and decreased in the order of "happy," "neutral," and "sad." Similarly, the valence was the highest in the emotion class "happy" and decreased in the order of "neutral," "sad," and "angry."

TABLE 2. Number of speech samples used in the experiments.

Database	Total (test)				
	Angry	*Happy*	*Neutral*	*Sad*	*Total*
IEMOCAP	1046 (212)	564 (118)	1516 (307)	1051 (214)	4177 (851)
KEMDy19	1621 (343)	1121 (243)	2853 (583)	669 (151)	6264 (1320)
KEMDy20	136 (48)	1162 (264)	7082 (1448)	115 (45)	8495 (1805)

TABLE 3. Evaluation procedure on a single database or multilingual corpora.

```
Input: a single database or multilingual corpora
Parameter:
    num_iter: Number of iterations (single database or multilingual corpora)
    num_speaker: Number of speakers of a database
Result: SER/SI performance metrics
for iter←1 to num_iter do:
| for s←1 to num_speaker do:
| | for c←1 to 4 emotion categories do:
| | | add 20% randomly selected data from speaker's emotion category
|         c to test dataset;
| | | add the remaining unselected 80% data from speaker's emotion
|         category c to training dataset;
| | end
| end
| perform SER/SI (baseline SI and em-SI) evaluation for 5 iterations;
end
calculate the average performance of SER/SI;
```

B. PREPARATION AND EVALUATION PROCEDURE

We used the speech data with a length of 2 to 30 s corresponding to the four emotion categories (i.e., "happy," "angry," "neutral," and "sad"). We assigned a unique SI to each speaker included in the three databases. To train the SER model, we categorized the emotions based on majority voting by the external taggers in all three databases. The average values of arousal and valence given by the external taggers based on the five-point scale were used as the training labels for arousal and valence.

Following the distribution of emotion categories of each database presented in Table 1, the data of KEMDy20 are severely imbalanced with most of the data residing in the "neutral" class. We downsampled the speech data corresponding to 70% of the "neutral" class of KEMDy20 for this experiment.

We did not adopt any data augmentation method for the speech data from any of the databases for the experiments focusing on the effectiveness of em-SI using an emotion-embedding vector in environments with data imbalances.

Table 2 lists the number of test and total speech samples used in the four emotion categories of IEMOCAP, KEMDy19, and KEMDy20. We applied z-normalization [54] to the speech data using the mean and standard deviation values of the databases to reduce the fluctuations of the speaker and speech signals. We implemented Bi-LSTM-

TABLE 4. Evaluation of the pre-trained SER model of the MTL structure.

Database	Loss (CE)	WA (%)	UA (%)	F1 score	CCC	
					Arousal	Valence
IEMOCAP	*w*-	66.3	62.5	0.621	0.614	0.448
	nw-	66.4	61.8	0.622	0.609	0.444
KEMDy19	*w*-	63.6	60.7	0.609	0.711	0.582
	nw-	64.2	59.8	0.608	0.718	0.574
KEMDy20	*w*-	84.6	42.9	0.450	0.600	0.382
	nw-	84.8	41.9	0.446	0.602	0.381

based SER and SI models using the TensorFlow toolkit [55] and trained the models using the Adam optimizer for 30 epochs.

Table 3 lists the evaluation procedure for SER and SI on each emotion database. The speech samples of the database were split into training and test datasets using stratified partitioning methods [56] by speaker-dependent emotion classes. We randomly selected 20% of the speech data corresponding to the emotion class of the speaker for the test data; the remaining 80% of the speech data were used for training. We configured a new training and test dataset for target single corpora or multilingual corpus combination until the counting reached the iteration number of the database, *num_iter*. Then, the evaluation of SER or SI was repeated five times with the same configured training and test dataset for each iteration to determine the average performance. The performance metric of the SER or SI for an emotion database was calculated using an average of 25 iterative tests through the evaluation procedure.

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. PRE-TRAINED SER

The pre-trained SER model of the MTL structure simultaneously predicted the emotion category and arousal and valence levels for speech data. We presented four SER performance metrics: the weighted accuracy (WA), unweighted accuracy (UA), macro F1 score [57], and concordance correlation coefficient (CCC) [58]. The metric WA is the overall accuracy, which is the ratio of correctly predicted samples to the total number of samples; UA is the recall average, which is an important performance indicator in evaluations based on imbalanced databases; and the F1 score is the harmonic mean of precision and recall. In this study, we used macro F1, which is the average of the F1 score for each label. We evaluated the CCC, which is a measure representing the degree of concordance between the predicted value and training label of the test dataset, for the SER performance metrics of arousal and valence.

Table 4 summarizes the average SER performance results of the transferred MTL-based SER and T_{SER} according to the weighted CE (*w*-) and non-weighted CE (*nw*-) for the categorical emotion classification $\mathcal{L}_c$. The results achieved by applying the weighted CE, *w*-, showed improvements in the UA of each class and decreases in the WA value compared to that achieved when applying the unweighted CE, *nw*-, in all three emotion databases. We adopted a weighted CE to train the transferred SER model in the SI experiments.

The transferred MTL-based SER model proposed in this study achieved an accuracy of 66.3% on IEMOCAP, which is comparable to the accuracy of 65.95% achieved by the fine-tuned ResNet-based SER model developed in a previous study [10].

The SER performances of the four emotion category classifications on the IEMOCAP and KEMDy19 databases compared to that on KEMDy20 were lower and higher in terms of WA and UA, respectively. It was inferred that IEMOCAP and KEMDy19 included relatively balanced speech data for each emotion class and well-expressed acoustic features in the data compared with the data of KEMDy20. The CCC performance for arousal and valence was also higher in IEMOCAP and KEMDy19 than that in KEMDy20.

Fig. 5 presents the confusion matrices of the results (values are rounded up to two decimals) for the four emotion category classifications and the line plots for the arousal and valence-level label recognition of the transferred MTL-based SER on IEMOCAP, KEMDy19, and KEMDy20.

The confusion matrices in Figs. 5(a), 5(b), and 5(c) clearly reveal that the "neutral" class, which has the most training data, shows the highest recognition accuracy. The "happy," "angry," and "sad" classes with relatively few training speech samples showed a high probability of being incorrectly predicted as the "neutral" class. Thus, it was inferred that the SER model had been potentially biased and trained in the prediction of a "neutral" class with a large amount of training data.

Figs. 5(c), 5(d), and 5(e) show the line plots for the arousal and valence levels, which are the training labels and predicted values of the pre-trained SER for 200 consecutive test speech samples. This indicated that the recognition performance of the valence level was lower than that of the arousal level in all three databases.

B. DEPENDENCY BETWEEN EMOTION AND SI

We performed dependency experiments between emotion and SI on expressive speech by distinguishing the emotion classes for the training data, which were used for enrollment, and the test data on the baseline SI model. Ablation experiments were conducted based on the baseline SI model, which did not use emotion embedding in the three emotion databases of IEMOCAP, KEMDy19, and KEMDy20.

As per the evaluation procedure described in Table 3, 80% and 20% of the speech data from the emotion category of each speaker were used for enrolling and testing the baseline SI model, respectively. The pre-trained SER model generated emotion embeddings for the configured test dataset at each iteration of the emotion database. The baseline SI model was trained using speech data of emotion class

TABLE 5. Accuracy (%) of multitarget si according to emotion classes used for enrollment and testing for the baseline SI model.

Database	Enrollment	Testing							
		Neutral(N)		Angry(A)		Happy(H)		Sad(S)	
		acc. (%)	Δ	acc. (%)	Δ	acc. (%)	Δ	acc. (%)	Δ
IEMOCAP	N	81.1 ± 1.6		63.4 ± 2.5		69.9 ± 6.6		70.5 ± 2.9	
	N+A	86.0 ± 0.4	4.9	84.2 ± 2.0	**20.8**	81.3 ± 1.6	11.4	77.1 ± 2.1	6.6
	N+A+H	86.9 ± 1.1	0.9	83.5 ± 0.9	-(0.7)	84.1 ± 1.4	2.8	81.4 ± 0.7	**4.3**
	N+A+H+S	87.6 ± 1.1	0.7	86.1 ± 0.8	2.6	87.3 ± 0.9	3.2	86.6 ± 1.0	**5.2**
KEMDy19	N	71.2 ± 0.8		57.6 ± 2.0		39.3 ± 3.2		59.4 ± 2.1	
	N+A	76.3 ± 1.4	5.1	74.9 ± 1.1	**17.3**	52.5 ± 1.6	13.2	67.9 ± 1.6	8.5
	N+A+H	77.0 ± 1.4	0.7	76.4 ± 1.9	1.5	61.6 ± 0.8	**9.1**	71.1 ± 0.6	3.2
	N+A+H+S	78.9 ± 0.3	1.9	77.3 ± 1.3	0.9	61.6 ± 2.0	0.0	76.8 ± 1.8	**5.7**
KEMDy20	N	86.0 ± 0.4		90.5 ± 2.2		78.7 ± 1.3		85.5 ± 1.6	
	N+A	84.4 ± 1.1	-(1.6)	92.8 ± 1.6	**2.3**	76.3 ± 2.1	-(2.4)	86.3 ± 3.1	0.8
	N+A+H	85.7 ± 0.6	1.3	92.6 ± 0.7	-(0.2)	80.7 ± 0.7	**4.4**	86.8 ± 1.4	0.5
	N+A+H+S	86.6 ± 2.7	0.9	94.3 ± 1.6	1.7	87.0 ± 3.1	**6.3**	88.8 ± 1.5	2.0

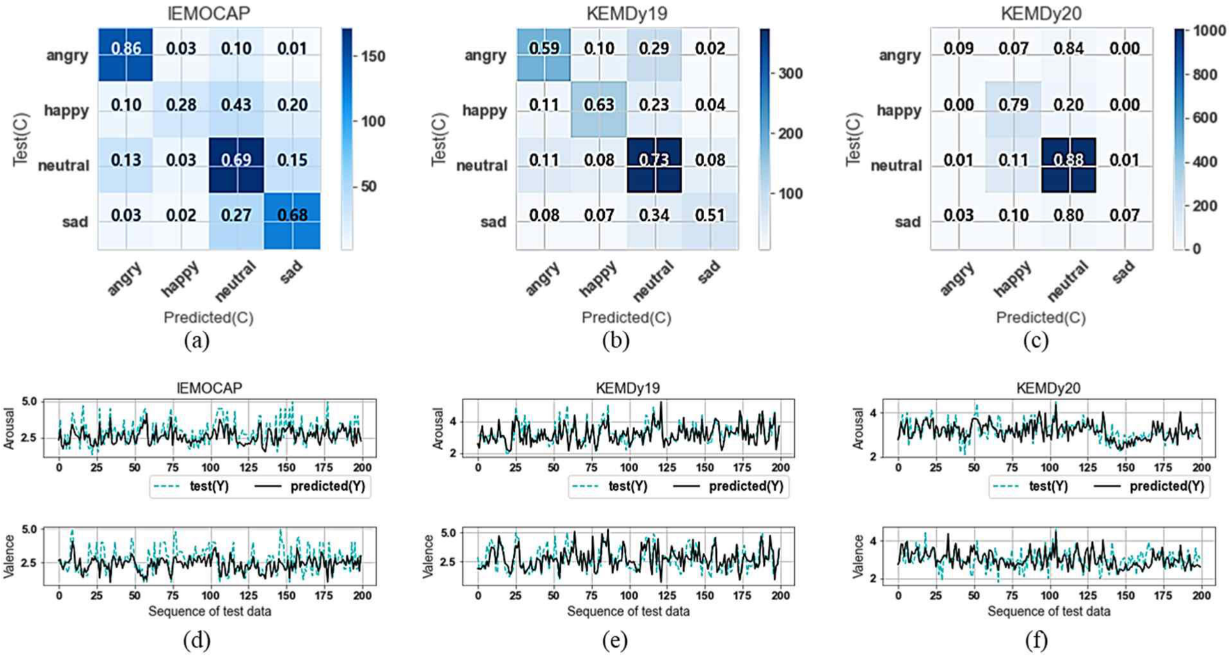


FIGURE 5. Visualization of the pre-trained SER results for categorical emotion classification and arousal and valence level recognitions on the IEMOCAP, KEMDy19, and KEMDy20 databases. Confusion matrices for the recognition of four emotion classes on the (a) IEMOCAP, (b) KEMDy19, and (c) KEMDy20 databases. Line plots for the training label and recognition values of arousal and valence levels on the (d) IEMOCAP, (e) KEMDy19, and (f) KEMDy20 databases.

combinations, where the emotion class is added in the order of "neutral" (N), "angry" (A), "happy" (H), and "sad" (S). The SI performance was evaluated with 20% of the test data belonging to each of the four emotion classes based on the trained baseline SI model by each combination of emotion classes.

Table 5 lists the experimental SI accuracy (in units of %) based on the emotion categories used for enrollment and testing. When a specific emotion class is added, the change in SI performance compared to that for the previous emotion class combination without using that emotion class is indicated in a separate column (Δ). As summarized in the evaluation procedure in Table 3, for each emotion class combination used for enrollment, we randomly selected the training and test datasets for the four emotion classes data per speaker.

In the table, for cases where speech data from each emotion class are additionally used for enrollment, the SI

TABLE 6. Evaluation results of baseline SI and EM-SI.

Database	Speakers	SI model	Accuracy (%)	EER (%)
IEMOCAP	10	baseline SI	86.9 ± 0.7	3.7 ± 0.2
		em-SI	87.5 ± 0.9	3.3 ± 0.4
KEMDy19	40	baseline SI	74.5 ± 0.5	6.8 ± 0.3
		em-SI	77.7 ± 0.7	5.5 ± 0.3
KEMDy20	80	baseline SI	85.1 ± 1.3	2.6 ± 0.3
		em-SI	87.5 ± 0.9	1.8 ± 0.2
IEMOCAP +KEMDy19	50	baseline SI	79.6 ± 0.5	6.5 ± 0.3
		em-SI	82.0 ± 0.3	5.2 ± 0.2
IEMOCAP +KEMDy20	90	baseline SI	85.2 ± 0.2	3.2 ± 0.1
		em-SI	86.9 ± 0.5	2.6 ± 0.2

improvement with the highest enhancement in performance is highlighted in bold. These results confirmed that, when the data of the emotion class to which the test data belonged were used for enrollment, there was a rapid improvement in the SI performance of the corresponding emotion class in most test cases. Our evaluation showed the highest SI accuracy when speech data corresponding to all four emotion categories were used for enrollment, which was common to all the emotion categories in the three databases.

The SI performance of each emotion class in this study was higher according to the order "neutral," "happy," "sad," and "angry" when using 80% training data of the "neutral" class in IEMOCAP. This was the same result as that achieved in previous studies [9], [10].

Unexpected experimental results were also obtained, such as in the case of IEMOCAP when the data of "N+A+H" were trained; the SI performance of "angry" was slightly lower than that when the enrollment data of "N+A" were trained. The results presented in [9] indicated that the SI performance for the "sad" class of IEMOCAP was the maximum when the training emotion was neutral. Regarding the experimental results obtained in previous research [9] and this study, they can be attributed to the training and test datasets relying on a combination of the variability of the sampled data, including the number of samples of each class, data length, and acoustic features of the expressive speech data that constitute the database.

The IEMOCAP database, which comprised data from 10 speakers, had a relatively high proportion of short utterances within 5 s, while KEMDy20, which comprised longer speech data uttered from 80 speakers, had a high proportion of the "neutral" class and low proportions of the "angry" and "sad" classes.

The performance in terms of SI accuracy of the "neutral" test data of IEMOCAP was similar to that of the "neutral" class data of KEMDy20. KEMDy19 comprised speech data with higher deviations in the arousal and valence levels across various emotion classes, as shown in Table 1 and Fig. 4. Although KEMDy19 showed the lowest SI performance among the three databases, when each emotion class was used for enrollment, it displayed the most noticeable SI improvement for the test data of the corresponding class.

The SI experimental results in Table 5 reveal the improvement in the SI performance for expressive data, which have a high acoustic variation from "neutral" class data, when utterances expressing similar emotional levels are used for enrollment on the SI model.

Fig. 6 shows a bar graph of the SI accuracy according to the emotion class of speech data used for enrollment and testing on the IEMOCAP, KEMDy19, and KEMDy20 databases, as described in Table 5. The results shown in Table 5 and Fig. 6 confirm that the performance of the deep-learning-based SI model is greatly affected by the distribution and scale of speech data in the different emotion categories for each speaker in the training dataset.

C. EMOTION-AWARE SI

To evaluate the improvements in SI performance achieved by em-SI, we constructed training and test datasets using the speaker-dependent stratified partitioning methods for all four emotion classes based on three single emotional corpora and multilingual corpora according to the procedure listed in Table 3. The SI performances of the LSTM-based baseline SI method without the emotion embedding vectors and the em-SI method that used the emotion embedding vectors transferred from the pre-trained SER were evaluated. The training and test datasets that were randomly selected in each repetition for the target corpus were commonly used for the SI evaluation of the baseline SI and em-SI models and pre-training of SER.

Table 6 presents the evaluation results of the baseline SI and em-SI models in terms of the SI accuracy (%) and EER (%) of SV for the multitarget speakers included in the three single corpora and multilingual corpora. The SI accuracy is the average speaker classification accuracy across all the speakers enrolled in the SI model. The presented EER was also calculated using the average EER of the test data belonging to all the speakers enrolled in the SI model. The baseline SI model in this study showed an average accuracy of 86.9% on IEMOCAP, surpassing the maximum average accuracy of 81.7% achieved by the baseline model using the i-vector demonstrated in a prior study by Sarma et al. [9]. The proposed em-SI model outperformed the baseline SI model on the three single corpora and multilingual corpora.

A previous study [22] attributed the decreased SI performance of the proposed deep learning-based SI model in the multilingual corpora experiment compared to that in the single corpus scenario to the difference in the size of the two corpora. The amount of data in KEMDy20 used for the evaluation in this study was more than twice that of IEMOCAP; however, most data in KEMDy20 were speech data belonging to the "neutral" class. In the SI evaluation of

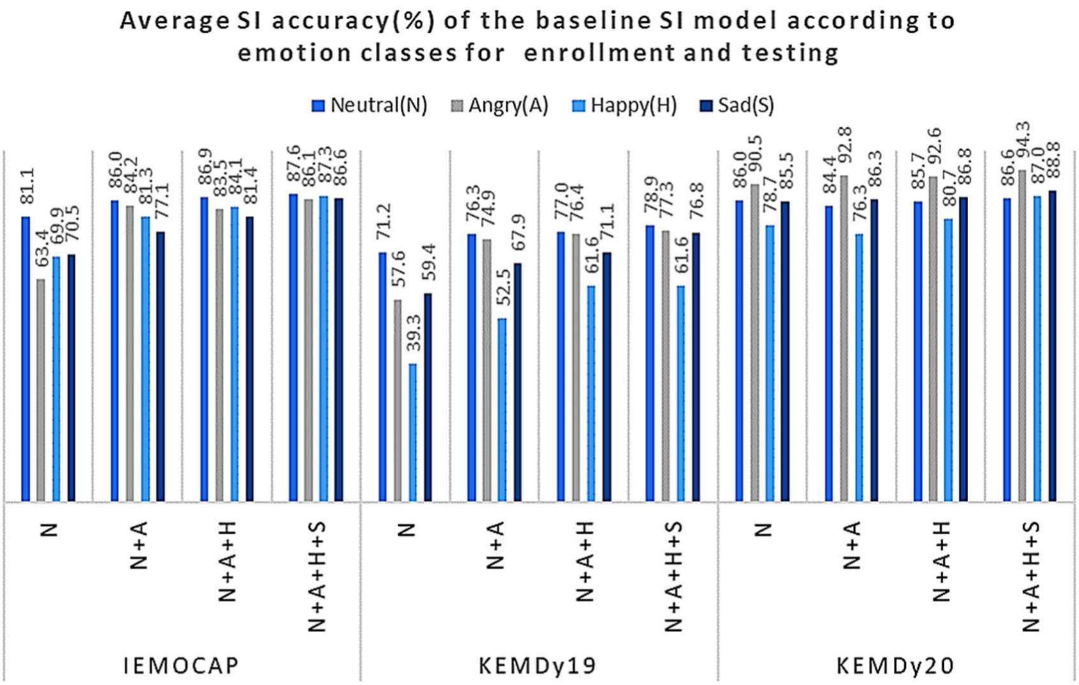


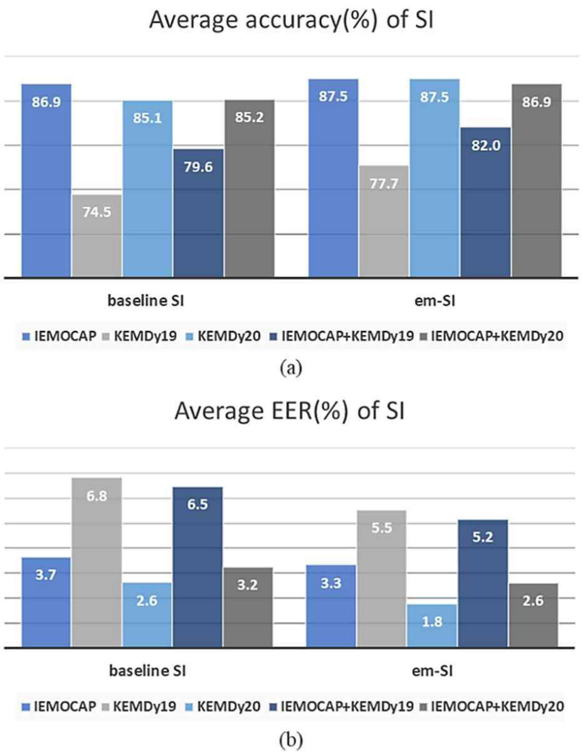
FIGURE 6. Average SI accuracy (%) based on the emotion classes used for enrollment and testing on the IEMOCAP, KEMDy19, and KEMDy20 databases by baseline SI.

the multilingual corpora using IEMOCAP and KEMDy20, em-SI showed a higher SI performance than that of the baseline SI. The SI performance in the cross corpus-based SI was slightly lower than that of the single corpus-based SI, as also shown by the results obtained in a previous study [22].

The average accuracies of the multitarget SI of the baseline SI and em-SI models for the ten speakers of the IEMOCAP were 86.9% and 87.5%, respectively. The performance of em-SI on KEMDy19 comprising 40 speakers was 3.2% higher in terms of accuracy, and the em-SI model achieved a performance improvement in terms of a decrease in the EER of approximately 1.3% compared to that of the baseline. In the SI evaluation of the data of the 80 speakers from KEMDy20, the average accuracy of em-SI was approximately 2.4% higher than that of the baseline model.

Fig. 7 shows the SI accuracy (%) and EER (%) of the baseline SI and em-SI models in the three single and multilingual corpora. In the experiments on English and Korean corpora, the proposed em-SI model showed a better performance than that of the baseline SI model, similar to that of a single corpus-based model. The em-SI model showed accuracy improvements of 2.4% and 1.7% in the multilingual corpora test for IEMOCAP and KEMDy19 and for IEMOCAP and KEMDy20, respectively, compared to the corresponding values of the baseline model.

Fig. 8 shows the average receiver operating characteristic (ROC) curves of the baseline SI and em-SI models for multitarget SI on IEMOCAP, KEMDy19, and KEMDy20. Each ROC curve was plotted for a test dataset for each speaker, and the average ROC curve and EER for all



(a)

(b)

FIGURE 7. SI performance of the baseline SI and em-SI models in terms of (a) SI accuracy (%) and (b) EER (%).

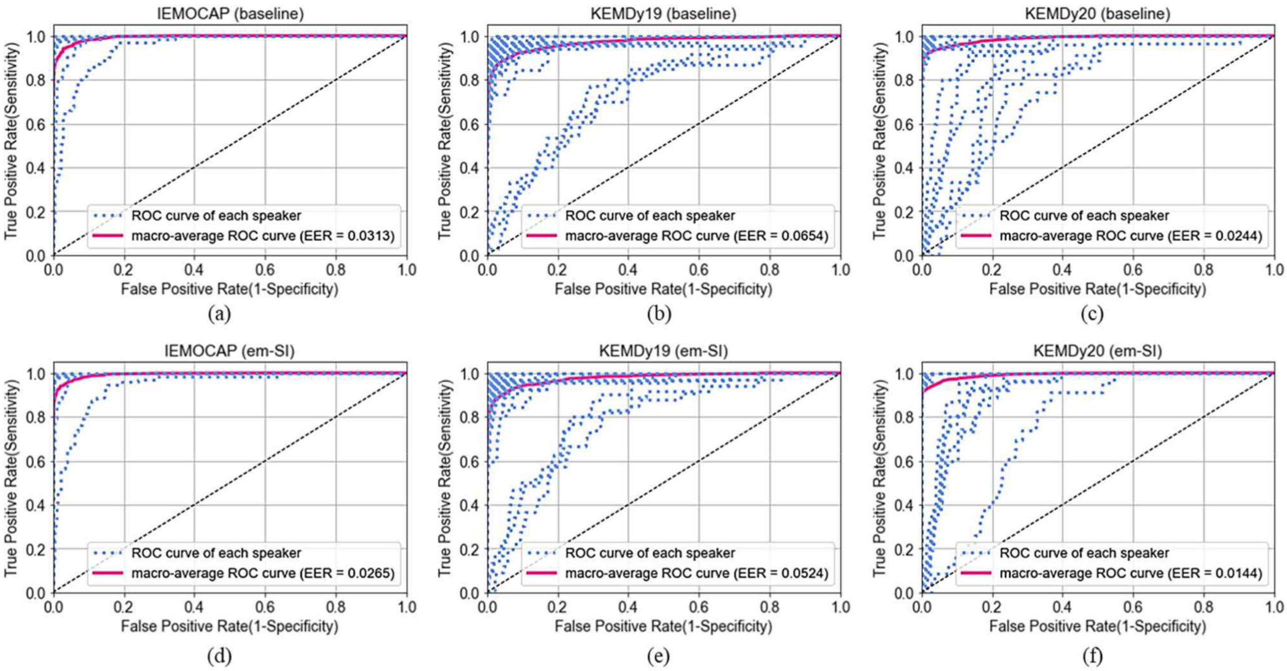


FIGURE 8. Average ROC curves of the baseline SI model for multitarget SI on the (a) IEMOCAP, (b) KEMDy19, and (c) KEMDy20 databases. Average ROC curves of em-SI on the (d) IEMOCAP, (e) KEMDy19, and (f) KEMDy20 databases.

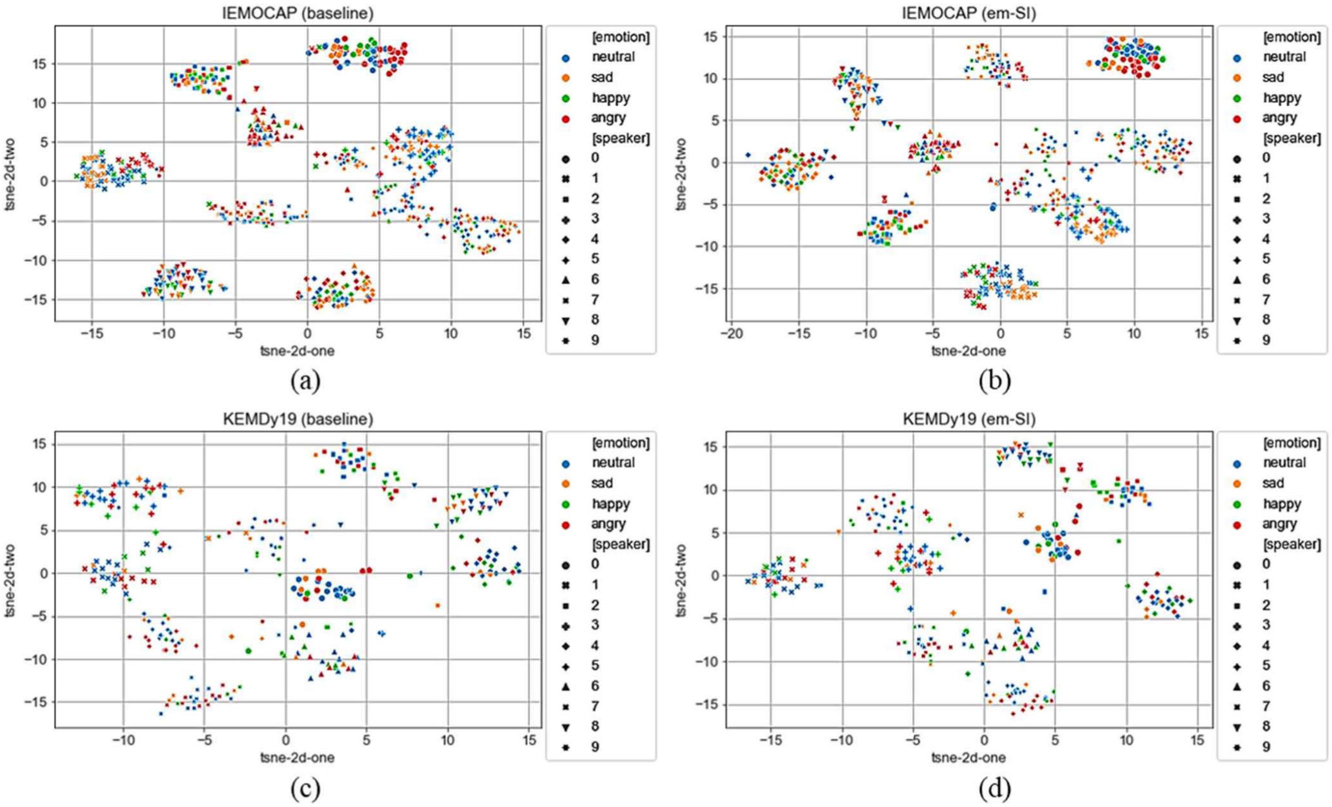


FIGURE 9. Embedding space of SI of the baseline SI and em-SI using the t-SNE: (a) and (b) SI embedding space of all ten speakers of IEMOCAP; (c) and (d) SI embedding space of ten speakers on KEMDy19.

speakers were calculated. The average EER for the em-SI model evaluation on all the three databases, as shown in Figs. 8(d), 8(e) and 8(f), was lower than that of the baseline EER, as shown in Figs. 8(a), 8(b), and 8(c). The

ROC curve for several speakers below the average ROC curve of the baseline SI model was improved, and the average EER of the em-SI models was lower than that of the baseline.

Fig. 9 shows the 2-D reduction of the SI embedding vectors of the test dataset for the baseline SI and em-SI models using t-distributed stochastic neighbor embedding (t-SNE). Figs. 9(a) and 9(b) show the SI embedding of the baseline SI and em-SI models for the test dataset for all ten speakers included in IEMOCAP, and Figs. 9(c) and 9(d) show the embeddings of the ten speakers in KEMDy19. The evaluation results for the baseline SI and em-SI models shown in Fig. 9 were obtained with the same training and test data in the corresponding database.

We ensured that the em-SI model learned the SI embedding space that included both the SI and emotion class information of the corresponding speech data, as shown through the visualization in Fig. 9.

In both Figs. 9(a) and 9(b), the cluster boundary for the SI embedding of speaker "9" is unclear. The SI embedding vectors of the same emotion category in the SI embedding of the em-SI model in Fig. 9(b) are closer to each other within the corresponding speaker cluster than they are in the baseline model shown in Fig. 9(a). The cluster of emotion classes in the SI embedding space of em-SI was also observed in the KEMDy19-based dataset shown in Fig. 9(d) and compared with that for the baseline model in Fig. 9(c).

The evaluation results on KEMDy19 presented in Table 5 indicate that em-SI with the transferred emotion embedding achieved an average accuracy improvement of 3.2% compared to that of the baseline SI model. The SI embedding space visualization of em-SI in Fig. 9(d) showed that the speaker cluster for the test dataset was evident, and that the test data of the same emotion category were located closer to each other in the speaker cluster than in the case of the baseline model in Fig. 9(c). Although the SI performance on KEMDy19 was the lowest among the three databases, the improvement in SI performance for this database through em-SI was the highest.

Fig. 9 confirmed that the em-SI model learned the embedding space that reflected the SI and emotional information of the speech data. The speaker embedding of em-SI, which simultaneously reflected the speaker and emotion embeddings in Figs. 9(b) and 9(d), showed a better SI performance for expressive speech than that of the baseline SI that did not use the emotion embeddings, as shown in Figs. 9(a) and 9(c).

VI. CONCLUSION

We presented an em-SI model that learns the speaker-embedding space and simultaneously embeds SI and emotional information from speech data. The experiments evaluating the em-SI system based on the multilingual emotion database IEMOCAP and the freely available Korean emotion multimodal databases KEMDy19 and KEMDy20 confirmed that the proposed em-SI method improves the SI performance in expressive speech.

To improve the SI performance of a deep-learning-based SI model for expressive speech uttered in various emotional situations, training data for various emotions for each speaker are required. However, such emotion databases incur a high cost for collecting sufficient data for training deep-learning-based SI models in individual emotional situations for individual speakers.

The proposed em-SI model could learn the emotion-embedding vector transferred from the pre-trained SER model along with the acoustic features. The separate SI and SER models of the em-SI model were independent for each task and easy to cross-reference. Each of the SER and SI models could be expanded using other speech databases and optimization methods; they could be changed to a different task-dependent network structure or combined with existing models.

We implemented the MTL SER model using a weighted loss to handle the problem of emotional data imbalance and label uncertainty for emotion transfer learning within the em-SI system. The em-SI model combined with the transferred SER model exhibited an improved SI performance for expressive speech in disproportionate emotion databases.

In this study, we implemented the SI and SER models in the em-SI system based on the same emotion database using the same Bi-LSTM-based network structure. This was performed to minimize the gain from the network and database of the transferred SER for SI operations. We also confirmed the effect of emotion-embedding learning on the SI model. In the future, we will attempt to evaluate various network structures of em-SI for enhancing SI in expressive speech using multiple-language databases.

REFERENCES

[1] A. B. Nassif, N. Alnazzawi, I. Shahin, S. A. Salloum, N. Hindawi, M. Lataifeh, and A. Elnagar, "A novel RBFNN-CNN model for speaker identification in stressful talking environments," *Appl. Sci.*, vol. 12, no. 10, p. 4841, May 2022, doi: 10.3390/app12104841.

[2] I. Shahin, "Speaker identification in emotional talking environments based on CSPHMM2s," *Eng. Appl. Artif. Intell.*, vol. 26, no. 7, pp. 1652–1659, Aug. 2013, doi: 10.1016/j.engappai.2013.03.013.

[3] S. Gong, Y. Dai, J. Ji, J. Wang, and H. Sun, "Emotion analysis of telephone complaints from customer based on affective computing," *Comput. Intell. Neurosci.*, vol. 2015, p. 15, Nov. 2015, doi: 10.1155/2015/506905.

[4] S. Parthasarathy, C. Zhang, J. H. L. Hansen, and C. Busso, "A study of speaker verification performance with expressive speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 5540–5544, doi: 10.1109/ICASSP.2017.7953216.

[5] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 582–596, May 2009, doi: 10.1109/TASL.2008.2009578.

[6] M. Abdelwahab and C. Busso, "Evaluation of syllable rate estimation in expressive speech and its contribution to emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, South Lake Tahoe, NV, USA, Dec. 2014, pp. 472–477, doi: 10.1109/SLT.2014.7078620.

[7] A. B. Nassif, I. Shahin, S. Hamsa, N. Nemmour, and K. Hirose, "CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions," *Appl. Soft Comput.*, vol. 103, May 2021, Art. no. 107141, doi: 10.1016/j.asoc.2021.107141.

[8] S. Hamsa, I. Shahin, Y. Iraqi, E. Damiani, and N. Werghi, "Speaker identification from emotional and noisy speech data using learned voice segregation and speech VGG," 2022, *arXiv:2210.12701*.

[9] B. D. Sarma and R. K. Das, "Emotion invariant speaker embeddings for speaker identification with emotional speech," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Auckland, New Zealand, 2020, pp. 610–615.

[10] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 7169–7173, doi: 10.1109/ICASSP40776.2020.9054317.

[11] W. Zheng, P. Yang, R. Lai, K. Zhu, T. Zhang, J. Zhang, and H. Fu, "Exploring multi-task learning based gender recognition and age estimation for class-imbalanced data," in *Proc. Interspeech*, Incheon, South Korea, 2022, pp. 1983–1987, doi: 10.21437/Interspeech.2022-682.

[12] 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," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Nov. 2008, doi: 10.1007/s10579-008-9076-6.

[13] K. J. Noh and H. Jeong. *KEMDy19*. Accessed: Jul. 24, 2023. [Online]. Available: https://nanum.etri.re.kr/share/kjnoh2/KEMDy19?lang=En_us

[14] K. J. Noh and H. Jeong. *KEMDy20*. Accessed: Jul. 24, 2023. [Online]. Available: https://nanum.etri.re.kr/share/kjnoh2/KEMDy20?lang=En_us

[15] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 5329–5333, doi: 10.1109/ICASSP.2018.8461375.

[16] K. A. Lee, Q. Wang, and T. Koshinaka, "Xi-vector embedding for speaker recognition," *IEEE Signal Process. Lett.*, vol. 28, pp. 1385–1389, 2021, doi: 10.1109/LSP.2021.3091932.

[17] F. Kelly, A. Alexander, O. Forth, and D. V. D. Vloed, "From I-vectors to X-vectors—A generational change in speaker recognition illustrated on the NFI-FRIDA database," in *Proc. 25th Int. Assoc. Forensic Phonetics Acoust. (IAFPA)*, 2019, pp. 1–28.

[18] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995, doi: 10.1109/89.365379.

[19] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011, doi: 10.1109/TASL.2010.2064307.

[20] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, Oct. 2019, doi: 10.1109/TAFFC.2017.2736999.

[21] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct. 2014, doi: 10.1109/TAFFC.2014.2336244.

[22] A. H. Meftah, H. Mathkour, S. Kerrache, and Y. A. Alotaibi, "Speaker identification in different emotional states in Arabic and English," *IEEE Access*, vol. 8, pp. 60070–60083, 2020, doi: 10.1109/ACCESS.2020.2983029.

[23] N. Simić, S. Suzić, T. Nosek, M. Vujović, Z. Perić, M. Savić, and V. Delić, "Speaker recognition using constrained convolutional neural networks in emotional speech," *Entropy*, vol. 24, no. 3, p. 414, Mar. 2022, doi: 10.3390/e24030414.

[24] A. Garain, B. Ray, F. Giampaolo, J. D. Velasquez, P. K. Singh, and R. Sarkar, "GRaNN: Feature selection with golden ratio-aided neural network for emotion, gender and speaker identification from voice signals," *Neural Comput. Appl.*, vol. 34, no. 17, pp. 14463–14486, Sep. 2022, doi: 10.1007/s00521-022-07261-x.

[25] H. Zeinali, L. Burget, and J. Cernocky, "Convolutional neural networks and x-vector embedding for DCASE2018 acoustic scene classification challenge," 2018, *arXiv:1810.04273*.

[26] Y. Tang, G. Ding, J. Huang, X. He, and B. Zhou, "Deep speaker embedding learning with multi-level pooling for text-independent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6116–6120, doi: 10.1109/ICASSP.2019.8682712.

[27] M. Avriel and D. J. Wilde, "Optimally proof for the symmetric Fibonacci search technique," *Fibonacci Quart. J.*, vol. 10, pp. 265–269, Jan. 1966.

[28] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391, doi: 10.1371/journal.pone.0196391.

[29] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*.

[30] S. Zheng, H. Suo, and Q. Chen, "PRISM: Pre-trained indeterminate speaker representation model for speaker diarization and speaker verification," in *Proc. Interspeech*, Sep. 2022, pp. 1431–1435, doi: 10.21437/Interspeech.2022-289.

[31] N. Vaessen and D. van Leeuwen, "Training speaker recognition systems with limited data," in *Proc. Interspeech*, Incheon, South Korea, 2022, pp. 4760–4764, doi: 10.21437/Interspeech.2022-135.

[32] P. Ekman, "Universal facial expressions in emotion," *Stud. Psychol.*, vol. 15, no. 2, pp. 140–147, 1973.

[33] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Pers.*, vol. 11, no. 3, pp. 273–294, 1977, doi: 10.1016/0092-6566(77)90037-X.

[34] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and M. Yang, "Deep learning for continuous multiple time series annotations," in *Proc. Audio/Visual Emotion Challenge Workshop*, Oct. 2018, pp. 91–98, doi: 10.1145/3266302.3266305.

[35] B. T. Atmaja, A. Sasou, and M. Akagi, "Speech emotion and naturalness recognitions with multitask and single-task learnings," *IEEE Access*, vol. 10, pp. 72381–72387, 2022, doi: 10.1109/ACCESS.2022.3189481.

[36] B. T. Atmaja and M. Akagi, "Evaluation of error- and correlation-based loss functions for multitask learning dimensional speech emotion recognition," *J. Phys., Conf. Ser.*, vol. 1896, no. 1, Apr. 2021, Art. no. 012004, doi: 10.1088/1742-6596/1896/1/012004.

[37] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," 2018, *arXiv:1801.06353*.

[38] K. J. Noh, C. Y. Jeong, J. Lim, S. Chung, G. Kim, J. M. Lim, and H. Jeong, "Multi-path and group-loss-based network for speech emotion recognition in multi-domain datasets," *Sensors*, vol. 21, no. 5, p. 1579, Feb. 2021, doi: 10.3390/s21051579.

[39] Y. Wang, J. Zhang, J. Ma, S. Wang, and J. Xiao, "Contextualized emotion recognition in conversation as sequence tagging," in *Proc. Special Interest Group Discourse Dialogue*, 2020, pp. 186–195.

[40] C.-C. Lu, J.-L. Li, and C.-C. Lee, "Learning an arousal-valence speech front-end network using media data in-the-wild for emotion recognition," in *Proc. Audio/Visual Emotion Challenge Workshop*, Oct. 2018, pp. 99–105, doi: 10.1145/3266302.3266306.

[41] S. Lee, D. K. Han, and H. Ko, "Fusion-ConvBERT: Parallel convolution and BERT fusion for speech emotion recognition," *Sensors*, vol. 20, no. 22, p. 6688, Nov. 2020, doi: 10.3390/s20226688.

[42] V. Dissanayake, H. Zhang, M. Billinghurst, and S. Nanayakkara, "Speech emotion recognition 'in the wild' using an autoencoder," in *Proc. Interspeech*, Oct. 2020, pp. 526–530, doi: 10.21437/Interspeech.2020-1356.

[43] T. S. Buda, M. Khwaja, and A. Matic, "Outliers in smartphone sensor data reveal outliers in daily happiness," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 1, pp. 1–19, Mar. 2021, doi: 10.1145/3448095.

[44] J. Bang, T. Hur, D. Kim, T. Huynh-The, J. Lee, Y. Han, O. Banos, J.-I. Kim, and S. Lee, "Adaptive data boosting technique for robust personalized speech emotion in emotionally-imbalanced small-sample environments," *Sensors*, vol. 18, no. 11, p. 3744, Nov. 2018, doi: 10.3390/s18113744.

[45] A. C. L. Ngo, R. C.-W. Phan, and J. See, "Spontaneous subtle expression recognition: Imbalanced databases and solutions," in *Proc. Asian Conf. Comput. Vis.*, Singapore, 2014, pp. 33–48, doi: 10.1007/978-3-319-16817-3_3.

[46] A. Mallol-Ragolta, N. Cummins, and B. W. Schuller, "An investigation of cross-cultural semi-supervised learning for continuous affect recognition," in *Proc. Interspeech*, Oct. 2020, pp. 511–515, doi: 10.21437/Interspeech.2020-2641.

[47] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proc. Interspeech*, Aug. 2017, pp. 1103–1107, doi: 10.21437/Interspeech.2017-1494.

[48] J.-M. Chen, P.-C. Chang, and K.-W. Liang, "Speech emotion recognition based on joint self-assessment manikins and emotion labels," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, San Diego, CA, USA, Dec. 2019, pp. 1–4, doi: 10.1109/ISM46123.2019.00073.

[49] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[50] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[51] S. Kumar, "Real-time implementation and performance evaluation of speech classifiers in speech analysis-synthesis," *ETRI J.*, vol. 43, no. 1, pp. 82–94, Feb. 2021, doi: 10.4218/etrij.2019-0364.

[52] Solmitech Corporation. *Refit Patch U7D*. Accessed: Jul. 24, 2023. [Online]. Available: https://en.solmitech.com/20/?idx=39

[53] Empatica Corporation. *E4*. Accessed: Jul. 24, 2023. [Online]. Available: https://www.empatica.com/en-int/research/e4/

[54] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, Jan. 2020, doi: 10.1016/j.specom.2019.12.001.

[55] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*.

[56] V. L. Parsons, "Stratified sampling," in *Wiley StatsRef: Statistics Reference Online*, 2017, pp. 1–11, doi: 10.1002/9781118445112.stat05999.pub2.

[57] Z. C. Lipton, C. Elkan, and B. Narayanaswamy, "Thresholding classifiers to maximize F1 score," 2014, *arXiv:1402.1892*.

[58] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 5, pp. 255–268, Mar. 1989.

KYOUNGJU NOH received the B.S. and M.S. degrees in computer science from Chonbuk National University, Jeonju, Republic of Korea. Since 2001, she has been a Principal Researcher with the Artificial Intelligence Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea. She has been involved in communications and personal-computing-based science-related projects for over a decade. She is currently developing affective computing technology for human understanding. Her research interests include human understanding, affective computing, artificial intelligence, and emotion recognition.

HYUNTAE JEONG received the B.S. and M.S. degrees in electronic engineering from Chungnam National University, Daejeon, Republic of Korea, in 1993 and 1995, respectively.

From 1995 to 2000, he was with the Samsung Heavy Industries Research and Development Center, Daejeon, developing control systems for ship experiments. Since 2001, he has been with the Electronics and Telecommunications Research Institute, Daejeon. He has been involved in the development of mobile and wearable computing technologies. His current research interests include cognitive computing, artificial intelligence, wearable computing, and HCI.