IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 September 2022, accepted 8 November 2022, date of publication 10 November 2022,
date of current version 29 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3221453

RESEARCH ARTICLE

Chung-Ang Auditory Database of Korean Emotional Speech: A Validated Set of Vocal Expressions With Different Intensities

YOUNGJA NAM[1] AND CHANKYU LEE[1,2]
[1]Humanities Research Institute, Chung-Ang University, Seoul 06974, Republic of Korea
[2]Department of Korean Language and Literature, Chung-Ang University, Seoul 06974, Republic of Korea

Corresponding author: Chankyu Lee (leeck@cau.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean Government under Grant 2017S1A6A3A01078538.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Chung-Ang University, Republic of Korea.

ABSTRACT A growing body of evidence indicates that intensity plays a role in emotion perception. However, only a few databases have been explicitly designed to provide emotional stimuli that are expressed at varying intensities. We developed and validated a Korean audio-only database of emotional expressions. Eighteen actors were recorded using twenty-five sentences with strong and moderate intensities for "neutral," "happiness," "sadness," "anger," "fear," and "boredom" emotions. Twenty-five native Korean-speaking adults completed the emotion identification and naturalness rating tasks. All listeners were presented with the full set of 5400 recordings in a six-alternative forced-choice paradigm, yielding 135000 judgements for identification and naturalness, respectively. Raw and unbiased hit rates were calculated, with identification responses significantly above chance level for every emotion at both intensities. The overall raw hit rates reached 87% and 78% for the strong and moderate stimuli, respectively, indicating that strong emotional expressions were more accurately identified than their moderate counterparts. Similarly, a recognition advantage for strong intensity over moderate intensity was observed for each emotion at both intensities. High inter- and intra-rater reliabilities were found in listeners' identifying emotion categories and assigning naturalness ratings, respectively. Further, there was a strong association between identification accuracy and the degree of naturalness; more natural variants of an emotion were more accurately identified than its less natural counterparts. These results confirm that the proposed database will serve as a valuable source for emotion research. This database is available for research purposes upon request from the corresponding author.

INDEX TERMS Database, Korean, intensity, emotion identification, naturalness rating, raw hit rates, unbiased hit rates, inter-rater reliability, intra-rater correlation, identification accuracy and naturalness relationship.

I. INTRODUCTION

Emotional expression perception plays a fundamental role in facilitating interactions and communication. Accurately recognizing the emotional state of an interaction partner allows individuals to navigate the complex social world they encounter more adaptively [1], [2]. Emotional information can be conveyed verbally or via non-verbal cues, such as voice, facial expressions, gestures, and eye gaze. Voice is a particularly powerful tool for emotional communication because it contains rich information about a speaker's emotional state. Research has shown that distinct emotional states

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

in vocal expressions are characterized by specific acoustic profiles [3], [4]. For example, emotions such as anger and happiness with high excitation display a higher fundamental frequency, while sadness (with little excitation) shows a relatively lower fundamental frequency [5], [6], [7]. Relatedly, Kraus [8] reported that emotions in voice-only expressions are more accurately identified than those in facial expressions or in combined facial-vocal expressions.

Machine recognition of human emotions from voice has become increasingly prominent in human–computer interactions over the past two decades. Speech emotion recognition using speech signals as input has been an active area of research that has a wide spectrum of applications. For example, in the healthcare field, speech emotion recognition systems, which are less biased than humans, can be utilized as efficient and non-invasive screening tools for diagnosing depression and have the potential to provide early diagnosis and intervention [9], [10]. In intelligent transportation systems, accurate speech emotion recognition can help enhance the safety of autonomous vehicles and provide personalized in-vehicle experiences [11], [12].

In the study of speech emotion recognition, validated emotional speech databases constitute a crucial building block for developing and evaluating speech emotion recognizers [13]. To date, numerous emotional speech databases have been created in many languages, including Arabic [14], [15], Bangla [16], Mandarin Chinese [17], [18], Danish [19], English [20], [21], [22], German [23], [24], Italian [25], and Persian [26]. However, the majority of the databases come from high-resource languages such as English, Mandarin Chinese, and German [27]. Indeed, research on speech emotion recognition has mainly focused on these languages, whereas low-resource languages, including Korean, have received relatively less attention [28]. For Korean, there exists only one validated database [29]. It should be noted that the performance of emotion recognition algorithms generally varies according to language [30]. The research community requires more varied resources to enrich our understanding of both language-specific and language-universal aspects of emotion, which, in turn, will pave the way for robust emotion recognition models. Recently, there has been a growing need to address the role of emotional intensity in constructing emotional speech databases, given that emotional expressions with higher intensity tend to be more accurately recognized than their counterparts with lower intensity [31], [32], [33], [34]. However, only a handful of databases have been explicitly designed to manipulate the levels of intensity of vocal emotional expressions [17], [18]. Specifically, there exists no such resource for the Korean language. Therefore, the present study aimed to introduce Chung-Ang Auditory Database of Korean Emotional Speech (CADKES), a Korean emotional speech database that covers a range of vocal emotions collected at two levels of emotional intensity. For validation of emotion stimuli, listeners completed two tasks: emotion identification and naturalness rating. In particular, the naturalness rating task was conducted to explore whether utterances of an emotion category, which achieve higher naturalness ratings, are more accurately identified than their counterparts, which achieve lower naturalness ratings.

The remainder of this paper is organized as follows. Section II presents a brief overview of emotion theories, emotion elicitation approaches, and existing databases. Section III describes the design and creation process of CADKES. Section IV describes emotion identification and naturalness rating tasks for database validation. Section V describes the validation results that are obtained through a set of statistical analyses. Section VI is allocated for discussion, along with the limitations and directions for future research. Section VII offers concluding remarks.

II. BACKGROUND AND RELATED WORKS

A. EMOTION MODELS

With respect to speech emotion recognition and affective science, there are two influential conceptual perspectives about emotion representation: discrete emotion models and continuous dimensional emotion models. Within the discrete emotion models, emotions are perceived categorically and thus emotion category "X" is distinguished from emotion category "Y." These discrete perspectives posit the notion of a finite set of basic or fundamentally distinct emotion classes that can be recognized culture-universally rather than being culture-specific. Ekman, a prominent proponent of basic emotions, proposed the Big Six model [35] wherein anger, disgust, fear, sadness, surprise, and happiness constitute the six basic emotions. In [36], Ekman further proposed a set of basic emotion criteria, such as distinctive universal signals, presence in other primates, distinctive physiology, rapid onsets, and automatic appraisal. Within dimensional emotion models, emotions are identified by mapping them onto a two- or three-dimensional space. One of the most widely studied dimensional models is Russell's circumplex model of affect [37], wherein emotional experiences are described along the dimensions of arousal (which captures the amount of physiological activation ranging from calm or passive to excited or active) and valence (which captures the degree of pleasantness ranging from unpleasant or negative to pleasant or positive). More recently, a hybrid approach, which combines the discrete and dimensional emotion models, has gained increasing attention as an endeavor to explore the nature of emotion [18], [38], [39].

B. EMOTION ELICITATION APPROACHES

Speech emotion databases are generally classified into natural, induced, and acted emotional databases [3]. Natural speech emotions are typically collected from call-center conversations [40], human–robot interactions [23], podcasts [41], and TV talk shows [42], [43]. Thus, they have a high ecological validity. However, in such cases, speech intelligibility may be compromised due to background environmental noise. Moreover, in real-life conversations, emotions are often overlappingly expressed between

interlocutors. They are often not clearly articulated, which poses a substantial challenge in identifying and annotating the emotion expressed. In addition, it is difficult to collect natural datasets because it often involves copyright and privacy issues. Induced or elicited speech emotions [20], [44] are collected by placing speakers into certain emotion-eliciting situations. However, it is unclear whether such artificial situations elicit similar emotional states across speakers [3]. Acted or simulated emotional databases [19], [20], [22], [25] guarantee controlled recordings by asking professional actors or non-trained speakers to portray different target emotions using the same text. This allows for a comparative analysis of the acoustic properties of the recordings according to emotion type, phoneme type, or speaker. Recordings of acted databases generally take place in a laboratory environment that produces low noise and allows high-quality recordings suitable for automatic acoustic feature extraction. Acted databases are often preferably employed over natural or elicited databases for research on speech emotion recognition (for a review of database types used in speech emotion recognition research, the reader is referred to an excellent review provided by [27]).

C. EXISTING DATABASES

The Berlin Emotional Speech Database (EMO-DB) [24] is one of the most popularly used speech databases for speech emotion recognition. The EMO-DB initially consisted of approximately 800 utterances from 10 native German-speaking actors in 7 emotional states: anger, boredom, disgust, fear, joy, sadness, and neutrality. The final EMO-DB included 535 utterances screened based on 20 listeners' performance in emotion identification and naturalness judgment tasks.

The Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [20] is another prominent speech database. The IEMOCAP database is an audio-visual database that was recorded using 10 actors in English dyadic conversational sessions under scripted and improvisation-based hypothetical scenarios targeting 5 emotions (anger, frustration, happiness, sadness, and neutrality). Markers were attached to the actors' face, head, and hands during the recording to better understand the relationship between different communication modalities. The emotional content underwent emotion identification and dimensional ratings. Busso et al. [21] introduced another English audio-visual database, MSP-IMPROV, with a particular focus on controlling lexical content and securing naturalness of the recorded emotions. This database recorded 12 actors in 4 emotions under four types of dyadic conversational scenarios. The MSP-PODCAST database [41] consists of 197 speakers' spontaneous utterances, primarily in anger, contempt, disgust, fear, happiness, surprise, and neutral emotions, collected from 403 English podcasts. In [21] and [41], emotional content was evaluated using crowdsourced listeners.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [22] contains lexically matched emotional speech and songs recorded from 24 actors in North American English. The spoken utterances were produced in six emotions along with two baseline emotions. Each utterance has two versions of intensity (strong vs. normal), except for neutral expressions. The RAVDESS obtained its validity by performing perceptual tasks in which 247 listeners evaluated a subset of utterances, combined with a series of comprehensive statistical analyses. The King Saud University Emotions (KSUEmotions) database [14] consists of 23 lay Arabic speakers vocalizing anger, happiness, sadness, surprise, and neutral expressions. The database was validated by performing an emotion identification task and a variety of statistical analyses. Most recently, another database, whose verification is based on perception tests and statistical analyses, came from the Bangla language. The SUST Bangla Emotional Speech Corpus (SUBESCO) [16] involves 20 actors portraying 7 emotions.

With respect to Korean, [29] is the only validated Korean database. The database contains 5400 sentences with 10 non-professional actors uttering 45 sentences three times for four emotions. Recently, two databases were published by the same laboratory. One database [45] comprises 4 actors uttering 4 emotions, while the other database [46] includes 20 actors uttering 6 emotions. However, neither provides validation results. Arguably, examination of emotional quality based on perceptual tests plays a critical role in establishing quality databases. Thus, the scarcity of validated Korean databases motivated us to create a validated Korean speech database. The CADKES was evaluated particularly in light of the potential roles of emotional intensity, stimulus naturalness, and listener- and speaker-gender differences in emotion perception, along with inter- and intra-rater reliabilities.

III. PREPARATION OF CADKES STIMULI

A. SPEAKERS

Eighteen actors participated in stimulus recording (mean age, 26.7 years; range, 21–33 years; SD = 2.8; 10 males and 8 females). The inclusion criteria were as follows: (1) being a native speaker of Korean who was born and raised in Korea, and (2) speaker of the standard Seoul variety of Korean. Of the 18 actors, 11 were from the School of Performing Arts and Media at Chung-Ang University, and 7 were professional actors. All actors had 3–11 years of acting experience, with a mean length of 7.3 years (SD = 2.41). As reviewed in subsections II-B and II-C, experienced or professional actors are preferred over lay speakers in many popular databases [20], [21], [22], [24] because they produce more accurate and intelligible emotional productions while simultaneously fulfilling the naturalistic portrayal of emotions. Thus, CADKES relied on experienced actors to ensure that they are able to portray emotional expressions with different degrees of intensity more consistently across target emotions.

B. SELECTION OF EMOTIONS

The following six emotions were targeted: neutral, happiness, sadness, anger, fear, and boredom. Neutral was

TABLE 1. Target Sentences For CADKES.

	Sentence (English translation)
1	오늘 점심은 칼국수야. (Today's lunch is Kalguksu.)
2	이사님은 대구로 출장 가셨어.(The director went on a business trip to Daegu.)
3	꽃밭에 강아지가 있어. (There is a puppy in the flower garden.)
4	벚꽃이 바람에 휘날려 (Cherry blossoms are blowing in the wind.)
5	중대 정문에서 택시를 탔어.(I took a taxi from the main gate of Chung-Ang University.)
6	라디오를 듣고 있어. (I am listening to the radio.)
7	쌍문역에서 친구를 만났어.(I met my friend at the Ssangmun station.)
8	강변에서 산책해. (I am talking a walk near the riverside.)
9	방탄소년단과 약속이 있어.(I have an appointment with members of BTS Bangtan Boys.)
10	엄마가 파도를 타요. (My mom is riding the waves.)
11	미용실에서 머리를 감아.(I am washing hair at the hairdresser shop.)
12	노래방에서 노래를 불러.(I am singing a song at a karaoke.)
13	새빨간 딸기를 먹었어. (I ate a bright red strawberry.)
14	마루에 궤짝이 있어. (There is a crate on the floor.)
15	윤중로에 눈이 내려요. (It is snowing on Yunjung street.)
16	아기 사슴이 뛰어다녀. (A baby deer is running.)
17	원숭이가 낮잠을 자요. (The monkey is taking a nap.)
18	옆집에 의사가 살아. (A doctor lives next door.)
19	외가는 서울이야. (My mother's side of the family lives in Seoul.)
20	괭이로 땅을 파고 있어. (I am digging the ground with a hoe.)
21	편의점에서 우산을 샀어.(I bought an umbrella at the convenience store.)
22	창밖에 노랑나비가 있어.(There is a yellow butterfly outside the window.)
23	아이가 연을 날리고 있어.(The child is flying a kite.)
24	310관 2층에서 공부해. (I am studying on the second floor of Building 310.)
25	앞마당에 코스모스가 피었어.(Cosmos is blooming the front yard.)

considered the baseline emotional state. These emotions are found in many existing speech emotion databases [28], [50] and have been frequently explored in speech emotion recognition [27]. The first author designed 25 semantically neutral sentences for the CADKES source script. The script contains all Korean 19 initials and 10 monophthongs. This aspect is conducive to exploring emotional speech acoustics, for example, by measuring vowel acoustics as a function of emotional style. Table 1 provides the target Korean sentences and their respective English translations for CADKES. Each sentence consists of 7 to 12 syllables. Within CADKES, each actor produced 25 sentences for each emotion at 2 intensity levels, resulting in 5400 sentences (18 actors $\times$ 6 emotions $\times$ 2 intensities $\times$ 25 sentences). Each emotion at each intensity consisted of 450 sentences (18 actors $\times$ 25 sentences). Table 2 presents a brief summary of CADKES.

C. RECORDING PROCEDURE

The development of CADKES began with actor auditioning. Actors were recruited through personal contact with senior students at the School of Performing Arts and Media, Chung-Ang University, and a posting on the university website. The participants were asked to send audio files containing their emotional utterances via email. It should be noted that the actors recorded their own emotional expressions at home using smartphones because the recording took place during the acceleration phase of the COVID-19 pandemic in Seoul, and a professional recording studio at Chung-Ang University could not be booked. For the audition, 2 of the 25 target sentences were randomly selected by the first author. The actors were asked to produce 2 sentences in the 6 target emotions for each intensity (6 emotions $\times$ 2 intensities $\times$ 2 sentences $=$ 24 sentences). The actors were asked to portray the intended

TABLE 2. Summary of the cadkes stimuli.

Year of collection	2021
Language	Korean
Emotions	neutral, happiness, sadness, anger, fear, boredom
Speakers	18 actors (10 males)
Sentences	25
Modality	audio
Intensity	2 (strong, moderate)
Size	5400 sentences
Emotion size per intensity	450 sentences
Sampling rate	44.1 kHz
Software for editing	Praat
File duration	1 hour 58 min 3 sec (for strong-intensity expressions) 1 hour 43 min 42 sec (for moderate-intensity expressions)

emotions in a natural, genuine style in which they felt or experienced them. All utterances (30 actors × 24 sentences = 720 sentences) were subjected to a perceptual evaluation in which 10 listeners (5 males and 5 females) identified emotion categories and emotional intensity. Listeners performed the identification task on their home computers. They were given six answer options (''neutral,'' ''happiness,'' ''anger,'' ''boredom,'' ''fear,'' and ''sadness'') for emotion categories and two answer options (''strong'' and ''moderate'') for emotional intensity. The stimuli were presented in a random order using the Paradigm program [51]. All the actors and listeners who participated in this audition phase were paid for their time.

For the second recording, 18 actors who achieved the highest aggregate accuracy scores on performance evaluation were selected; for each actor, all utterances were identified as intended emotion and intended intensity for more than 80% of the instances. The actors were asked to record 25 sentences in 6 emotional styles with strong and moderate intensities at 2 different time points (4 to 8 days apart depending on their schedules) to ensure that any familiarity with one type of intensity would not affect their production at the other intensity level, as well as to compensate for fatigue. As mentioned above, the actors recorded themselves at home using smartphones (they all used Samsung Galaxy series smartphones). Emphasis was placed on the genuineness and naturalness of emotional expressions, and the actors were told that they start recording when they achieve the target emotions by naturally feeling or experiencing what they are expected to express. The actors were instructed to carry out their recordings in a quiet room. They were encouraged to say the name of the emotion type at the beginning of the recording of each emotion set and save the emotion set as a separate file. The actors were allowed to choose the order of producing the desired emotions to ensure that they produced emotional utterances when they felt comfortable with the target emotion. They were also asked to reproduce any mispronounced sentences and any emotion class set when they felt that they had not portrayed the intended emotion in a genuine and natural style. The actors were asked to take a self-paced break between emotion classes to ensure that they disengage from the recorded emotion class before proceeding to the next emotion class. All actors were compensated for their participation. The recording was carried out in accordance with relevant guidelines and regulations, and the recording methods of the database were approved by Chung-Ang University Institutional Review Board, Republic of Korea. Informed written consent was obtained prior to recording from all participants.

D. POST-PROCESSING

For post-processing, the M4A files created on smartphones were converted into WAV files to be loaded into Praat speech processing software [52] (44.1 kHz, 16 bit). Three paid expert listeners reviewed the sentence files. Actors were required to record the sentence(s) again in the following instances: a sentence was missing, was mispronounced, was corrupted by any unwanted audible noise, was distorted with peak-clipping, was unnatural by any unusually long pause between words, was unintelligible, or was not portraying the intended emotion in a natural style. One actor was asked to produce one emotion set at a moderate intensity due to one mispronunciation error and an unnaturally long pause between words.

IV. VALIDATION

A. LISTENERS

The listeners included 25 native Korean-speaking adults (mean age, 24.4 years; range, 20–35 years; SD = 4.0; 13 males and 12 females). All listeners were undergraduate and graduate students from Chung-Ang University. An additional three listeners (one male and two females) were excluded from the data analysis, in one case due to

withdrawal before completing the perceptual validation task and in two cases due to missing data due to technical issues. None of the listeners participated in the perceptual evaluation of the stimuli for the audition or the stimuli recording. No listener had a history of speech or hearing problems. All listeners were compensated for their time.

B. PROCEDURE

The validation experiment was implemented using the Paradigm program [51] run on a Windows computer. As mentioned earlier, due to the COVID-19 pandemic, our listeners performed the validation task at home on their own computers. They were asked to download the Paradigm Player program, a free experiment presentation desktop app, and were provided the experiment files via email. Once listeners selected the specific experiment files in their ''Paradigm Experiments'' folder, the experiments were ready to be implemented. Listeners were presented with emotional stimuli at a self-adjusted, comfortable listening level. They were instructed to listen to the stimuli over headphones or earphones in a quiet room.

The validation experiment involved emotion category identification and naturalness rating tasks. Listeners were instructed to first identify the emotion category of the sentence they heard using a six-alternative forced-choice paradigm, in which the six target emotions were available for response options. The emotion category choices were displayed horizontally in clickable rectangular boxes. Once the emotion identification decision was made, listeners had to judge the naturalness of their selected emotion category using a 5-point scale (1 = completely unnatural, 5 = very natural). In the experimental session, right after the listener made a response, the Paradigm program presented the next trial.

The identification and rating tasks consisted of 10 sessions. Half of the sessions consisted of strong-intensity trials, and the other half consisted of moderate-intensity trials; thus, listeners were presented with stimuli of the same intensity in one experimental session. Each session contained 540 test trials, in which all actors produced 6 emotions using 5 sentences at 1 type of intensity (18 actors × 6 emotions × 5 sentences = 540 trials). Each session was divided into six blocks. Each block contained 90 trials, in which 3 actors produced 6 emotions using 5 sentences (3 actors × 6 emotions × 5 sentences = 90 trials). Each trial consisted of identification and rating judgments. Notably, each of the 25 listeners identified and rated the full set of 5400 sentences. Thus, all sentences were identified and rated 25 times, resulting in 135000 responses (25 listeners × 5400 trials = 135000 responses). The series of strong-intensity sessions took place three to four days apart from that of moderate-intensity sessions. For each intensity type, five experimental sessions were conducted over two to three days. For each session, all blocks and stimulus presentations within each block were randomized for each listener. No feedback was provided during the experimental sessions. In each session, listeners were allowed to take a self-paced break (a maximum of 5 min each) after every 2 blocks of 180 trials. Each experimental session lasted approximately 50 min (on average). Informed consent was obtained from all participants involved in the validation experiment.

V. RESULTS

The data were subjected to a set of analyses to assess identification accuracy, inter- and intra-rater reliabilities, the relationship between identification accuracy and naturalness degree, and the possible effects of intensity, listener gender, speaker gender, and emotion categories on emotion perception. All statistical analyses were performed using IBM SPSS (version 26), except for one-tailed one sample t-tests, which were performed using R statistical software (version 4.1.2).

A. IDENTIFICATION ACCURACY

The identification accuracy was measured in terms of raw and unbiased hit rates. Table 3 provides the identification performance across emotions and for each emotion at each intensity, along with the identification performance across intensities. The performance numbers were converted to raw hit rates (percentage of correct responses) and unbiased hit rates (H_u) [53] for each listener with respect to each emotion. Unbiased hit rates take into account false alarm biases, in which a perceiver tends to choose a particular emotion to identify a given stimulus in cases of doubt. Unbiased hit rates generally yield smaller values than the corresponding raw hit rates. The unbiased hit rates, which are defined in Eq (1), vary between 0 and 1, with higher values being more accurate. For example, for a target emotion ''anger,'' A indicates the number of correct responses identified as ''anger,'' B_1 indicates the number of instances in which *anger* was identified as *boredom*, S_1 indicates the number of instances in which *anger* was identified as *sadness*, B_2 indicates the number of instances in which *boredom* was identified as *anger*, and S_2 indicates the number of instances in which *sadness* was identified as *anger*.

$$UHR = \frac{A}{A + B_1 + S_1} \times \frac{A}{A + B_2 + S_2} \tag{1}$$

As shown in Table 3, the overall raw hit rates across emotions were 87% and 78% for the strong and moderate stimuli, respectively, while the overall unbiased hit rates across emotions for the stimuli were 76% and 64%, respectively. Raw hit rates ranged from 81% (fear) to 93% (neutral) for strong expressions, and from 69% (sadness) to 90% (neutral) for moderate expressions. Unbiased hit rates varied from 67% (boredom) to 84% (anger) for strong expressions, and from 56% (neutral and sadness) to 73% (happiness) for moderate expressions.

One sample t-tests were further performed on raw and unbiased hit rates to examine whether each emotion category was selected above a chance response level of 1/6 (since listeners were presented with six answer choices). For raw hit rates, one sample t-tests

TABLE 3. Raw hit rates (RHR) and unbiased hit rates (UHR) for identification accuracy measures.

	RHR % (SD %)		UHR % (SD %)		RHR % (SD %)	UHR % (SD %)
	Strong	Moderate	Strong	Moderate	Across intensities	Across intensities
Overall	87 (9)	78 (13)	76 (8)	64 (12)	82 (7)	69 (10)
Neutral	93 (6)	90 (10)	72 (12)	56 (13)	91 (8)	63 (13)
Happiness	83 (11)	79 (14)	79 (11)	73 (13)	81 (12)	76 (11)
Sadness	87 (6)	69 (12)	74 (8)	56 (13)	78 (8)	65 (10)
Anger	90 (5)	76 (11)	84 (6)	70 (10)	83 (8)	77 (7)
Fear	81 (9)	78 (11)	77 (11)	63 (16)	80 (10)	65 (11)
Boredom	86 (10)	77 (14)	67 (12)	64 (11)	81 (12)	70 (13)

TABLE 4. Confusion matrices for each intensity. The rows and columns represent intended and perceived emotions, respectively. Totals do not add up to 100% due to rounding.

	Actor intended emotion (strong)					
	Neutral	Happiness	Sadness	Anger	Fear	Boredom
Neutral	93	10	2	4	5	8
Happiness	1	83	1	2	1	0
Sadness	1	1	87	0	9	4
Anger	1	2	0	90	2	2
Fear	0	3	9	2	81	0
Boredom	3	1	1	2	1	86
	Actor intended emotion (moderate)					
	Neutral	Happiness	Sadness	Anger	Fear	Boredom
Neutral	90	18	11	13	9	15
Happiness	2	79	1	2	1	0
Sadness	2	1	69	1	9	5
Anger	1	1	1	76	1	2
Fear	0	1	14	2	78	1
Boredom	5	1	5	6	2	77

showed that all emotion categories were perceived as intended above chance at both intensities (strong, one-tailed $t(24)$s > 29, ps $< .001$, and moderate, one-tailed $t(24)$s > 21, ps $< .001$). For unbiased hit rates, one sample t-tests also indicated that all emotions were identified as intended above chance at both intensities (strong, one-tailed $t(24)$s > 21, ps $< .001$, and moderate, one-tailed $t(24)$s > 14, ps $< .001$).

Table 4 presents a brief overview of the confusion matrices according to intensity. Within strong intensity, misidentification was the highest for the happiness-neutral pair, in which happiness was misidentified as neutral 10% of the time. Within moderate intensity, the highest confusion was between the happiness-neutral pair, where happiness was perceived as neutral 18% of the time. This was followed by the boredom-neutral pair, where boredom was perceived as neutral 15% of the time, and the sadness-fear pair, where sadness was perceived as fear 14% of the time. The misidentification rate was above 10% for the sadness-neutral and anger-neutral pairs.

B. INTER-RATER RELIABILITY FOR EMOTION IDENTIFICATION

Inter-rater reliability for emotion category identification was assessed using Fleiss' Kappa [54]. Fleiss' Kappa is a chance-adjusted index of inter-rater agreement when there are more than two raters for classifying items or assigning categorical ratings to items. This statistical measure calculates the

TABLE 5. Overall and emotion-specific inter-rater reliability.

	Kappa (all ps $< .001$)		
	Strong	Moderate	Across intensities
Overall	0.76	0.64	0.70
Neutral	0.70	0.54	0.61
Happiness	0.79	0.75	0.77
Sadness	0.74	0.59	0.67
Anger	0.85	0.76	0.81
Fear	0.71	0.68	0.69
Boredom	0.76	0.61	0.68

extent to which the observed proportion of agreement among raters exceeds what would be expected if all raters made their ratings randomly. More specifically, the Kappa statistic measures the degree to which m raters concur in their respective identifications of n items into k categories. Recall that in this study, 25 listeners were presented with a full set of 5400 stimuli in the emotion identification task. The possible values of Kappa range from -1.0 (no agreement) to 1.0 (perfect agreement). Following the Landis and Koch guidelines [55], the magnitude of agreement for Kappa values was interpreted as follows: $\leq$0.0, poor agreement; 0.01–0.20; slight agreement, 0.21–0.40; fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.0, almost perfect agreement.

As shown in Table 5, all Kappa values were statistically significant ($p < .001$). This suggests that the agreement between the raters did not occur completely by chance. The overall mean Kappa values were 0.76 and 0.64, for strong and moderate intensity, respectively, suggesting that there was overall substantial agreement among listeners. Kappa values were also computed for each emotion category. For strong intensity, anger achieved almost perfect inter-rater agreement (0.85), and the remaining emotions showed substantial inter-rater agreement (0.70–0.79). For moderate intensity, boredom, fear, happiness, and anger revealed substantial inter-rater agreement (0.61–0.76), while neutral (0.54) and sadness (0.59) obtained moderate inter-rater agreement.

C. INTRA-CLASS CORRELATION COEFFICIENT FOR NATURALNESS RATINGS

The intraclass correlation coefficient (ICC) is a reliability index that reflects both the degree of correlation and agreement between two or more quantitative measurements [56], [57]. ICC values range from 0 to 1, with values below 0.40 indicating poor agreement, values between 0.40 and

TABLE 6. ICC for naturalness ratings using single- and average-rating, absolute-agreement, 2-way random effects models.

	ICC test		95% CI	*F*-test with True Value 0
Strong	Single (2, 1)	ICC = 0.14	CI (0.12, 0.15)	$F(2699, 64776) = 5.75, p < .001$
	Average (2, k)	ICC = 0.80	CI (0.78, 0.82)	$F(2699, 64776) = 5.38, p < .001$
Moderate	Single (2, 1)	ICC = 0.12	CI (0.10, 0,13)	$F(2699, 64776) = 5.75, p < .001$
	Average (2, k)	ICC = 0.77	CI (0.74, 0.79)	$F(2699, 64776) = 5.38, p < .001$

0.59 indicating fair reliability, values between 0.60 and 0.74 indicating good reliability, and values between 0.75 and 1.00 indicating excellent reliability [58]. The ICC values were separately calculated to assess the rater agreement for naturalness ratings according to emotional intensity. Table 6 presents the ICC values for assigning naturalness ratings to the vocal expressions for each intensity. ICC values (2, k) were computed for the average score of the k raters based on an absolute-agreement, two-way random effects model with 25 raters across 2700 trials, for strong and moderate intensity, respectively. Results showed excellent agreement in measuring the degree of naturalness of both strong- and moderate-intensity trials (0.80 and 0.77, respectively). Furthermore, ICCs were also computed using a single-rater, absolute-agreement, two-way random effects model. ICC (2, 1) results showed poor agreement across strong- and moderate-intensity trials (0.14 and 0.12, respectively).

D. IDENTIFICATION ACCURACY AND NATURALNESS RELATIONSHIP

The relationship between emotion identification accuracy and naturalness degree was evaluated using Pearson's correlation analysis: $r < 0.3$ = moderate, $0.3 \leq r \leq 0.5$ = moderate, and $r > 0.5$ = strong [59]. Table 7 shows the Pearson's correlation matrices. Strong positive correlations were observed for strong- and moderate-intensity expressions ($r = 0.74$ and $r = 0.62$, respectively). Within the strong-intensity stimuli, Pearson's r ranged from 0.58 (strong) to 0.99 (nearly perfect). More precisely, the happiness, sadness, anger, and boredom stimuli revealed nearly perfect positive correlations ($r = 0.96$–0.99, $ps \leq .005$), while the fear stimuli showed a very strong positive correlation ($r = 0.86$, $p = .03$). Neutral stimuli also yielded a strong positive correlation ($r = 0.58$); however, the relationship was not significant ($p = .15$). Within the moderate-intensity stimuli, the happiness, sadness, fear, and boredom stimuli showed nearly perfect positive correlations ($r = 0.94$–0.995, $ps \leq .01$), and the anger stimuli showed a very strong positive correlation ($r = 0.88$, $p = .025$). The neutral stimuli yielded a strong positive correlation ($r = 0.57$). However, the relationship was not significant ($p = .16$). The identification accuracy-naturalness relationship is captured in Fig. 1 by plotting the number of correctly identified emotion stimuli and their associated naturalness ratings.

E. ANALYSIS OF VARIANCE (ANOVA) ANALYSIS ON IDENTIFICATION ACCURACY

Normality of data was screened using the Shapiro–Wilk test prior to ANOVA. Some of the data were non-normally distributed. The raw hit percentages were transformed into arcsine values to satisfy the ANOVA assumption of normal distribution [60]. ANOVA was performed on the arcsine-transformed data. For readability, the untransformed raw hit rates are reported in Figs. 2–5. Two-way repeated measures ANOVAs and two-way mixed ANOVAs were conducted to assess the possible effects of emotional intensity and gender differences on emotion identification performance. Partial eta squared (η_p^2) is reported as a measure of effect size, indicating the extent to which an independent variable affected the dependent variable. The level of significance was set at 0.05. Significant interaction effects were decomposed by performing simple effects analyses.

1) EMOTION IDENTIFICATION ACCORDING TO INTENSITY

Identification percentages were submitted to a 2 (Intensity: strong vs. moderate) × 6 (Emotion: neutral, happiness, sadness, anger, fear, boredom) repeated measures ANOVA (two within-subjects factors) to assess how intensity affected listeners' emotion identification performance. As illustrated in Fig. 2, there was a main effect of Intensity ($F(1, 120) = 132.59$, $p < .001$, $\eta_p^2 = 0.85$), with listeners performing more accurately overall on strong expressions (87%) than on moderate expressions (78%) (see Table 3). The main effect of Emotion was significant ($F(5, 120) = 12.12$, $p < .0001$, $\eta_p^2 = 0.34$). Bonferroni-adjusted pairwise comparisons of accuracies revealed that neutral (93%) was more accurately identified than all other emotions (81%–90%): neutral > happiness, sadness, anger, fear, and boredom ($ps \leq .029$). No such differences were observed between any other emotion pairs ($ps \geq .10$). There was also a significant interaction between Intensity and Emotion ($F(5, 120) = 25.93$, $p < .001$, $\eta_p^2 = 0.52$). Table 8 summarizes the ANOVA results.

The Intensity × Emotion interaction was further probed using the follow-up tests. Simple effects analyses using two-tailed paired t-tests revealed a significant effect of Intensity, showing that strong expressions were more accurately identified than moderate expressions for all emotions ($ps \leq .015$). Table 9 provides a summary of the t-test results.

A one-way repeated measures ANOVA with Bonferroni post hoc comparisons was separately conducted to explore the differences in identification accuracy between emotions. There was a simple effect of Emotion ($F(5, 120) = 11.48$, $p < .001$, $\eta_p^2 = .32$). For strong intensity, pairwise comparisons showed that neutral > happiness, sadness, and fear ($ps \leq .012$), and anger > happiness and fear ($ps \leq .023$). No such differences were found for any of the other between-emotions comparisons ($ps \geq .058$). For moderate intensity, pairwise comparisons revealed that

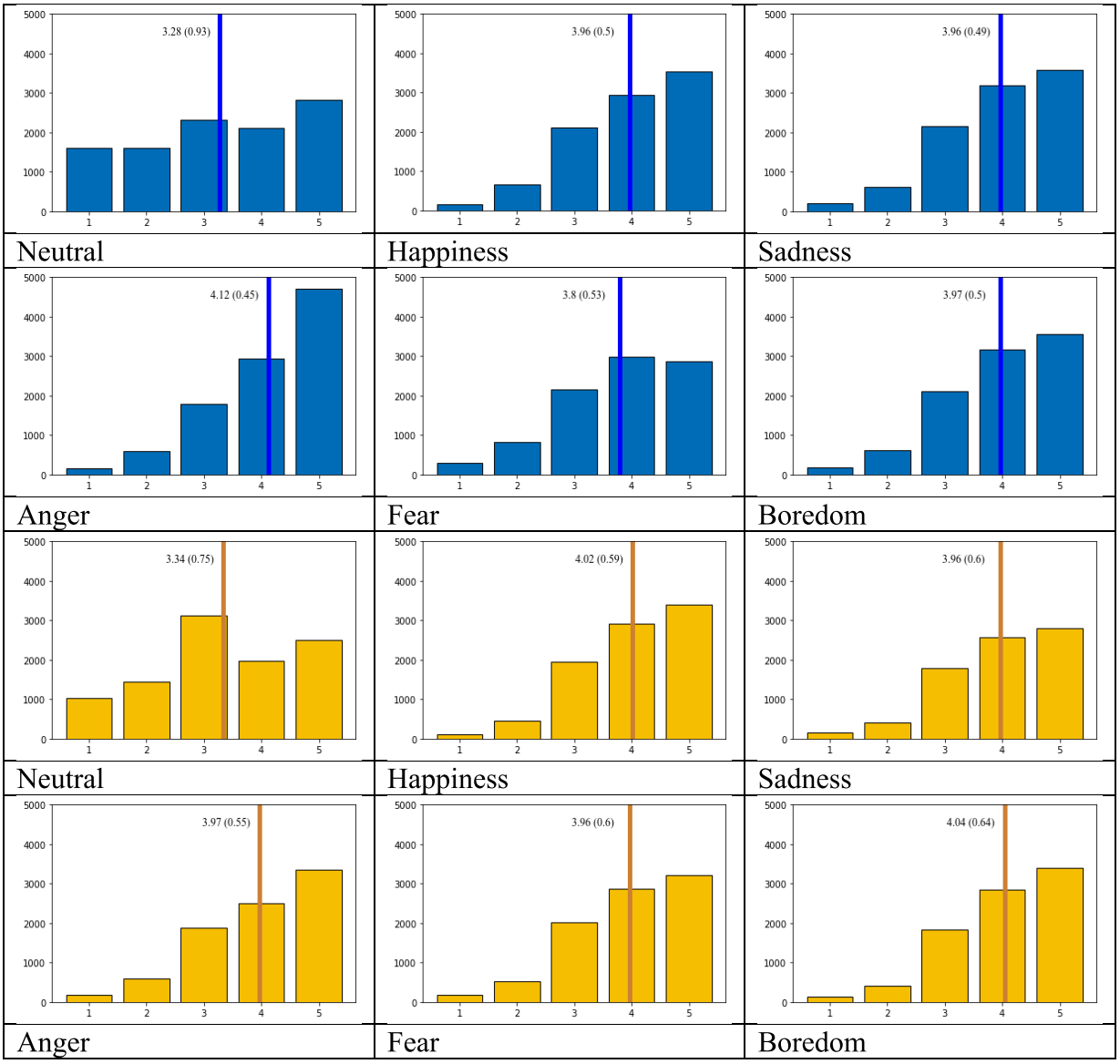

FIGURE 1. Naturalness rating distribution, along with mean ratings (indicated by vertical lines) and standard deviations (in parentheses), for each emotion within each intensity. The upper and lower two planes plot ratings for strong and moderate intensities, respectively. The vertical and horizontal axes represent the number of correctly identified emotion stimuli and their associated naturalness ratings, respectively. Naturalness ratings range from 1 (completely unnatural) to 5 (very natural).

TABLE 7. Pearson's correlations between identification accuracy and naturalness ratings.

	Strong		Moderate	
	Naturalness rating (SD)	r	Naturalness rating (SD)	r
Overall	3.85 (0.64)	0.74 ($p < .001$)	3.88 (0.66)	0.62 ($p < .001$)
Neutral	3.28 (0.93)	0.58 ($p = .15$)	3.34 (0.75)	0.57 ($p = .16$)
Happiness	3.96 (0.50)	0.98 ($p = .002$)	4.02 (0.59)	0.94 ($p = .01$)
Sadness	3.96 (0.49)	0.96 ($p = .005$)	3.96 (0.60)	0.96 ($p = .005$)
Anger	4.12 (0.45)	0.99 ($p = .001$)	3.97 (0.55)	0.88 ($p = .025$)
Fear	3.80 (0.53)	0.86 ($p = .03$)	3.96 (0.60)	0.96 ($p = .005$)
Boredom	3.97 (0.50)	0.99 ($p = .001$)	4.04 (0.64)	0.995 ($p < .001$)

neutral > sadness, anger, fear, and boredom ($ps \leq .022$), and happiness, fear, boredom > sadness ($ps \leq .039$). None of the other emotions differed from each other ($ps \geq .06$). The pairwise comparisons are summarized in Tables 10 and 11.

2) INTERPLAY OF LISTENER AND SPEAKER GENDER IN EMOTION IDENTIFICATION

Fig. 3 illustrates the raw hit rates according to listener and speaker gender for each intensity. A two-way mixed

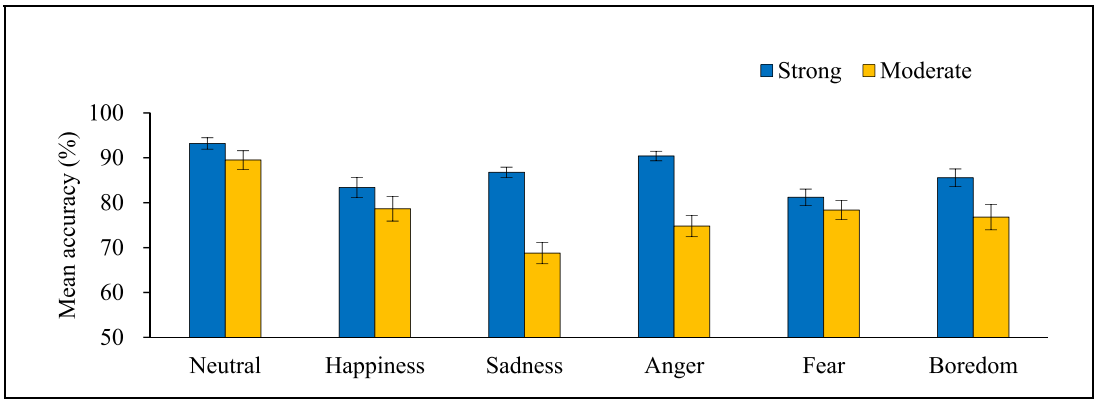

FIGURE 2. Raw hit rates for each emotion according to intensity. Error bars indicate standard errors of the means.

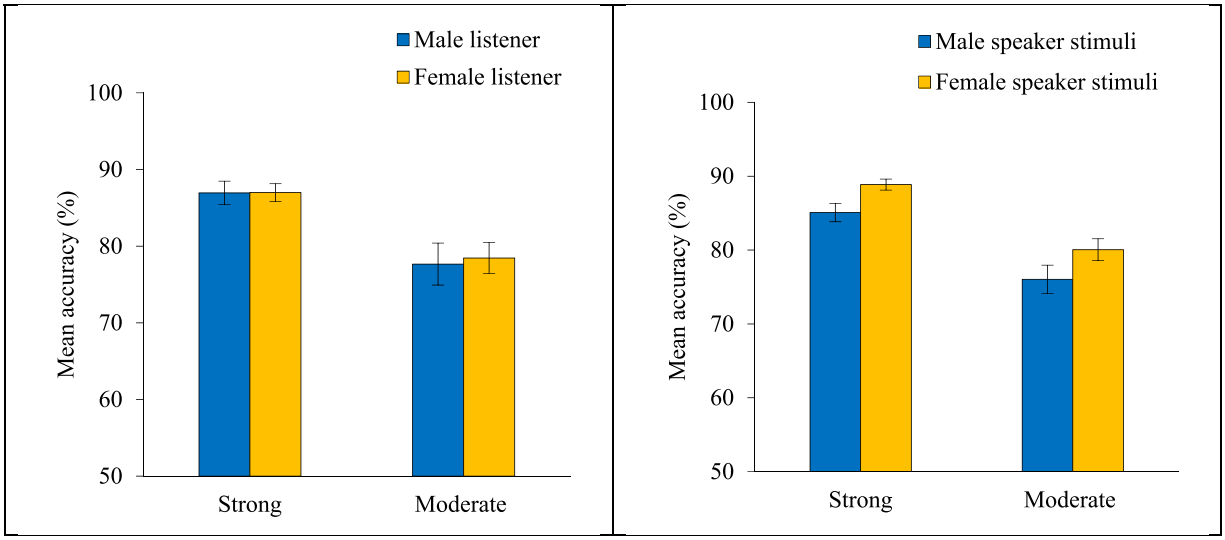

FIGURE 3. Raw hit rates for each intensity according to listener gender (left) and speaker gender (right). Error bars indicate standard errors of the means.

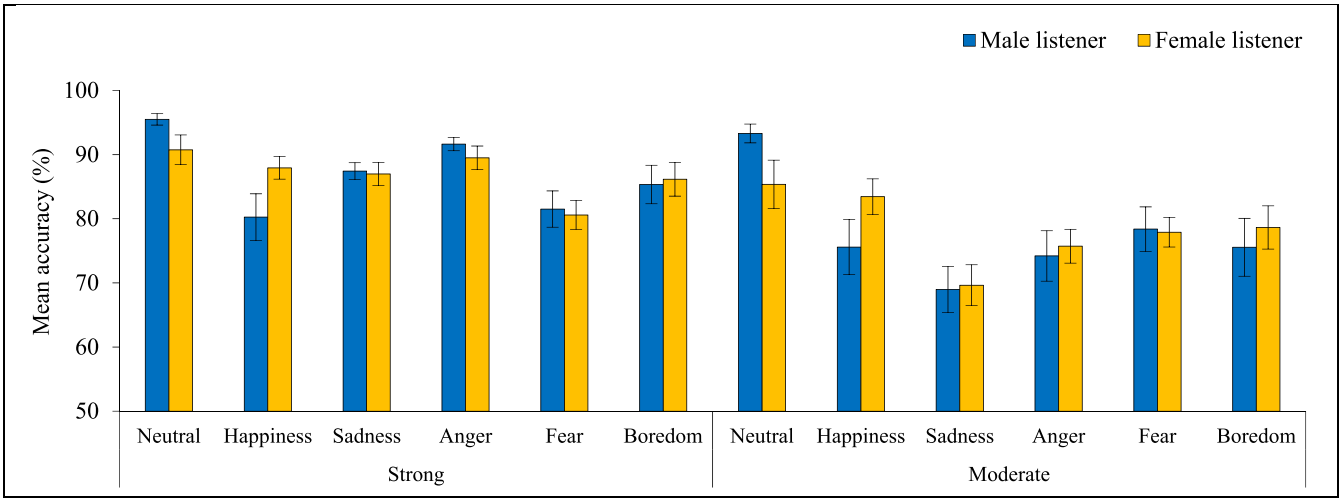

FIGURE 4. Raw hit rates for each emotion category according to listener gender at each intensity. Error bars indicate standard errors of the means.

ANOVA with Listener Gender (2: male vs. female) as a between-subjects factor and Speaker Gender (2: male vs. female) as a within-subjects factor was separately performed for each intensity to examine whether listener gender interacted with speaker gender. Table 12 summarizes the ANOVA results.

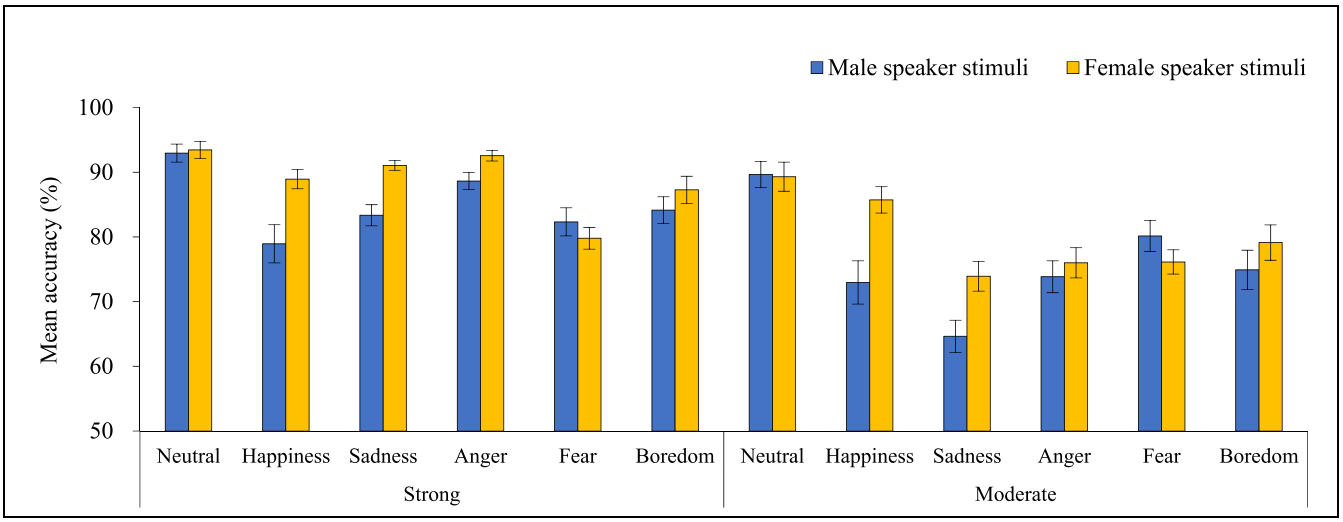

FIGURE 5. Raw hit rates for each emotion category according to speaker gender at each intensity. Error bars indicate standard errors of the means.

TABLE 8. 2 (Intensity) × 6 (Emotion) repeated measures ANOVA.

	Source	Sum of Squares	*df*	Mean Square	*F*	*p*	η_p^2
Within	Intensity	1.03	1	1.03	132.59	$p < .001$	0.85
	Emotion	1.38	2.7	0.51	12.12	$p < .001$	0.34
	Intensity × Emotion	0.42	5	0.08	25.93	$p < .001$	0.52
	Error	0.39	120	0.003			

TABLE 9. Paired t-tests for performance differences according to intensity (Strong–Moderate).

	Mean difference	*df*	*t*	*p*
Neutral	0.06	24	3.79	0.001
Happiness	0.06	24	2.63	0.015
Sadness	0.22	24	12.67	$p < .001$
Anger	0.21	24	10.17	$p < .001$
Fear	0.03	24	2.74	0.011
Boredom	0.11	24	7.33	$p < .001$

For strong intensity, the analysis yielded no main effect of Listener Gender ($F(1, 23) = 0.01$, $p = .92$). There was a main effect of Speaker Gender ($F(1, 23) = 34.37$, $p < .001$, $\eta_p^2 = 0.60$), showing that emotional expressions from female speakers (89%) were more accurately identified than those from male speakers (85%). The Listener Gender × Speaker Gender interaction was not significant ($F(1, 23) = 0.01$, $p = .92$).

For moderate intensity, the analysis also revealed only a significant Speaker Gender effect ($F(1, 23) = 44.95$, $p < .001$, $\eta_p^2 = 0.66$), showing that female speakers' expressions (80%) > male speakers' expressions (76%). There was no main effect of Listener Gender ($F(1, 23) = 0.002$, $p = .96$) or Listener Gender × Speaker Gender interaction ($F(1, 23) = 1.02$, $p = .32$).

3) EMOTION IDENTIFICATION ACCORDING TO LISTENER GENDER

Fig. 4 plots identification performance for each emotion category at each intensity based on the listener's gender. Data were submitted to a two-way mixed ANOVA with Listener Gender (2: male vs. female) as a between-subjects factor and Emotion (6: neutral, happiness, sadness, anger, fear, boredom) as a within-subjects factor to explore whether emotion identification differed according to listeners' gender. Table 13 provides a summary of the ANOVA results.

For strong intensity, the analysis yielded no main effect of Listener Gender ($F(1, 23) = 0.51$, $p = .48$) suggesting no differences between male and female listeners. The main effect of Emotion was significant ($F(5, 115) = 11.29$, $p < .001$, $\eta_p^2 = .33$), showing that some emotions were more accurately identified than others: neutral > happiness, sadness, and fear ($ps \leq .015$), and anger > happiness and fear ($ps \leq .026$). No such differences were observed between any other emotions ($ps \geq .062$). The Listener Gender × Emotion interaction was not significant ($F(5, 115) = 0.78$, $p = .57$).

For moderate intensity, the analysis revealed a significant Emotion effect ($F(5, 115) = 15.49$, $p < .001$, $\eta_p^2 = .40$), showing that performance varied according to emotion: neutral > sadness, anger, fear, and boredom ($ps \leq .023$), and happiness, fear, and boredom > sadness ($ps \leq .026$). No such differences were observed between any other emotions ($ps \geq .067$). Neither the Listener Gender effect ($F(1, 23) = 1.01$, $p = .33$) nor the Listener Gender × Emotion interaction ($F(5, 115) = 1.05$, $p = .39$) was significant.

4) EMOTION IDENTIFICATION ACCORDING TO SPEAKER GENDER

Fig. 5 illustrates the identification performance for each emotion category according to the speaker's gender at each intensity. Data were submitted to a 2-way repeated measures ANOVA with Speaker Gender (2: male vs. female) and

TABLE 10. Pairwise comparisons at strong intensity.

Strong	Neutrality	Happiness	Sadness	Anger	Fear	Boredom
Neutral	1	0.166* [0.03, 0.31]	0.127* [0.03, 0.22]	0.068 [-0.03, 0.17]	0.202* [0.09, 0.31]	0.133 [0.00, 0.27]
Happiness		1	-0.0384 [-0.13, 0.05]	-0.097* [-0.19, -0.01]	0.036 [-0.04, 0.12]	-0.0328 [-0.13, 0.06]
Sadness			1	-0.059 [-0.13, 0.01]	0.074 [-0.01, 0.16]	0.006 [-0.10, 0.11]
Anger				1	0.133* [0.06, 0.20]	0.064 [-0.02, 0.15]
Fear					1	-0.069 [-0.16, 0.02]
Boredom						1

Asterisk denotes a significant difference ($p < .05$).

TABLE 11. Pairwise comparisons at moderate intensity.

Moderate	Neutral	Happiness	Sadness	Anger	Fear	Boredom
Neutral	1	0.165 [0.00, 0.33]	0.290* [0.16, 0.41]	0.219* [0.08, 0.35]	0.175* [0.06, 0.29]	0.186* [0.02, 0.36]
Happiness		1	0.124* [0.03, 0.22]	0.054 [-0.04, 0.15]	0.01 [-0.07, 0.09]	0.021 [-0.06, 0.10]
Sadness			1	-0.071 [-0.16, 0.02]	-0.114* [-0.20, -0.03]	-0.104* [-0.20, 0.00]
Anger				1	-0.0436 [-0.14, 0.05]	-0.033 [-0.13, 0.06]
Fear					1	0.011 [-0.07, 0.10]
Boredom						1

Asterisk denotes a significant difference ($p < .05$).

TABLE 12. 2 (Listener Gender (LG)) × 2 (Speaker Gender (SG)) mixed ANOVA.

Strong	Source	Sum of squares	*df*	Mean square	*F*	*p*	η_p^2
Within	SG	0.04	1	0.04	34.37	$p < .001$	0.60
	LG × SG	0.000	1	0.000	0.001	0.98	0.000
	Error	0.02	23	0.001			
Between	LG	0.0001	1	0.0001	0.010	0.92	0.000
	Error	0.22	23	0.010			
Moderate	**Source**	**Sum of squares**	***df***	**Mean square**	***F***	***p***	η_p^2
Within	SG	0.027	1	0.027	44.95	$p < .001$	0.66
	LG × SG	0.001	1	0.001	1.02	0.32	0.04
	Error	0.01	23	0.001			
Between	LG	0.0001	1	0.0001	0.002	0.96	0.000
	Error	0.5	23	0.02			

Emotion (6: neutral, happiness, sadness, anger, fear, and boredom) as within-subjects factors.

For strong intensity, all statistical tests reached significance: Speaker Gender ($F(1, 120) = 37.52$, $p < .001$, $\eta_p^2 = 0.61$), Emotion, ($F(5, 120) = 11.75$, $p < .001$, $\eta_p^2 = 0.33$), and Speaker Gender × Emotion ($F(5, 120) = 16.37$, $p < .001$, $\eta_p^2 = 0.41$). Table 14 summarizes the ANOVA results.

Follow-up paired t-tests revealed a simple effect of Speaker Gender, showing that female speaker stimuli > male speaker stimuli for happiness, sadness, and anger ($ps \leq .014$). No such differences were observed for neutral and fearful expressions ($ps \geq .055$). Table 15 summarizes the results of the t-tests. Simple effects of Emotion with one-way repeated measure ANOVAs were significant for male and female speaker stimuli ($F(5, 120) = 10.65$, $p < .001$, $\eta_p^2 = 0.31$, and

TABLE 13. 2 (Listener Gender (LG)) × 6 (Emotion) mixed ANOVA.

Strong	Source	Sum of squares	*df*	Mean square	*F*	*p*	η_p^2
Within	Emotion	0.04	5	0.04	34.37	$p < .001$	0.60
	Emotion × LG	0.000	5	0.000	0.001	0.98	0.000
	Error	1.31	115	0.01			
Between	LG	0.0001	1	0.0001	0.010	0.92	0.000
	Error	0.69	23	0.03			
Moderate	Source	Sum of squares	*df*	Mean square	*F*	*p*	η_p^2
Within	Emotion	0.04	5	0.04	34.37	$p < .001$	0.60
	Emotion × LG	0.000	5	0.000	0.001	0.98	0.000
	Error	1.68	115	0.01			
Between	LG	0.0001	1	0.0001	0.010	0.92	0.000
	Error	1.46	23	0.06			

TABLE 14. 2 (Speaker Gender (SG)) × 6 (Emotion) repeated measures ANOVA.

	Source	Sum of squares	*df*	Mean square	*F*	*p*	η_p^2
Strong	SG	0.23	1	0.23	37.52	$p < .001$	0.61
	Emotion	1.30	5	0.26	11.75	$p < .001$	0.33
	SG × Emotion	0.25	5	0.05	16.37	$p < .001$	0.41
	Error	0.37	120	0.003			
	Source	Sum of squares	*df*	Mean square	*F*	*p*	η_p^2
Moderate	SG	0.17	1	0.17	48.93	$p < .001$	0.67
	Emotion	2.25	5	0.45	15.21	$p < .001$	0.39
	SG × Emotion	0.39	5	0.08	39.32	$p < .001$	0.62
	Error	0.24	120	0.002			

TABLE 15. Paired t-tests for the performance differences according to speaker gender (Male–Female).

Strong	Mean difference	*df*	*t*	*p*	Moderate	Mean difference	*df*	*t*	*p*
Neutral	-0.01	24	-0.82	0.42	Neutral	0.01	24	0.33	0.74
Happiness	-0.13	24	-7.19	$p < .001$	Happiness	-0.16	24	-12.80	$p < .001$
Sadness	-0.11	24	-6.79	$p < .001$	Sadness	-0.10	24	-12.05	$p < .001$
Anger	-0.07	24	-6.15	$p < .001$	Anger	-0.03	24	-2.62	0.015
Fear	0.04	24	2.02	0.055	Fear	0.06	24	3.69	0.001
Boredom	-0.05	24	-2.66	0.014	Boredom	-0.05	24	-3.60	0.001

$F(5, 120) = 14.43$, $p < .001$, $\eta_p^2 = 0.38$, respectively). For male speaker stimuli, Bonferroni pairwise comparisons revealed that neutral > happiness, sadness, fear, and boredom ($ps \leq .04$), and anger > happiness and fear ($ps \leq .049$). No such differences were observed between any other emotions ($ps \geq .09$). For female speaker stimuli, Bonferroni pairwise comparisons showed that neutral, happiness, sadness, and anger > fear ($ps < .001$) and fear > boredom ($p = .005$). No such differences were observed between any other emotions ($ps \geq .17$).

For moderate intensity, all statistical tests reached significance: Speaker Gender ($F(1, 120) = 48.93$, $p < .001$, $\eta_p^2 = 0.67$), Emotion, ($F(5, 120) = 15.21$, $p < .001$, $\eta_p^2 = 0.39$), and Speaker Gender × Emotion, ($F(5, 120) = 39.32$, $p < .001$, $\eta_p^2 = 0.62$).

As shown in Table 15, follow-up paired *t*-tests revealed a simple effect of Speaker Gender, showing that female speaker stimuli > male speaker stimuli for happiness, sadness, anger, and boredom ($ps \leq .015$), but male speaker stimuli > female speaker stimuli for fear ($p = .001$). No such difference was observed for neutral ($p = .74$). Simple effects of Emotion with one-way repeated measure ANOVAs reached significance for male and female speaker stimuli ($F(5, 120) = 20.11$, $p < .001$, $\eta_p^2 = 0.46$ and $F(5, 120) = 13.21$, $p < .001$, $\eta_p^2 = 0.35$, respectively). For male speaker stimuli, Bonferroni pairwise comparisons revealed that neutral > happiness, sadness, anger, fear, and boredom; fear > happiness and anger; and fear and boredom > sadness ($ps \leq .038$). No such differences were observed between any other emotions ($ps \geq .08$). For female speaker stimuli, Bonferroni pairwise comparisons showed that neutral > sadness, anger, and fear ($ps \leq .003$) and sadness, anger, and fear > happiness ($ps \leq .001$). No such differences were observed between any other emotions ($ps \geq .11$).

VI. DISCUSSION

This study aimed to develop and validate CADKES, a Korean audio-only database of emotional expressions at different levels of intensity, through a set of statistical analyses. For validation, 25 listeners identified emotion categories and rated their degree of naturalness. The overall hit rate was 82% across intensities. This accuracy rate is comparable to that of

the RAVDESS multimodal database [22], which reported an overall hit rate of 62% across intensities in the audio-only speech set. This value is also comparable to the existing audio-only emotional speech databases that do not consider the intensity factor: 84.24% for the Arabic KSUEmotions database [14], 75.5% for the Bangla SUBESCO database [16], 67.3% for the Danish DES database [19], 85% for the German EMO-DB database [24], and 80% for the Italian EMOVO database [25]. The unbiased hit rates were also computed to account for response biases, which are commonly used to assess the validity of emotional speech databases [16], [17], [22], [61]. The CADKES obtained overall unbiased hit rates of 69% across intensities, which were comparable to the RAVDESS stimuli in the audio modality with overall unbiased hit rates of 46.5% across intensities. These results confirmed that CADKES was successfully validated with respect to the overall raw and unbiased hit rates across intensities.

A. PERFORMANCE ON EMOTION IDENTIFICATION ACCURACY

1) IDENTIFICATION OF EMOTION CATEGORIES

Examination of identification performance revealed that some emotion categories are easier to identify than others, which was typically observed in previous research on human emotion perception [16], [22], [49]. For the strong stimuli, the accuracy rates were 81.2%–93.2%. These rates are comparable to the MES-P [18] strong stimuli, in which accuracy rates ranged between 83.85% and 91.74%, and the RAVDESS audio-only strong stimuli [22], in which the accuracy rates were 44%–91%. The moderate stimuli achieved accuracy rates of 68.8%–89.5%, which is comparable to MES-P [18] moderate stimuli with accuracy rates ranging from 75.74% to 87.35%, and the RAVDESS moderate stimuli in the audio-only set with accuracy rates ranging from 29% to 79%. The accuracy rates of CADKES are also comparable to those of RAVDESS, which achieved accuracy rates of 72%–94% and 56%–89% for strong and moderate emotional expressions, respectively, in the audio-visual stimulus set in which listeners can benefit from the integration of auditory and visual information to decode emotion.

The accuracy rates achieved in this study were also comparable to the ADFES-BIV [34] audio-visual dataset, in which low-, intermediate-, and high-intensity expressions achieved accuracy rates of 27%–90%, 37%–90%, and 41%–96%, respectively. In CADKES, the unbiased identification accuracy ranged from 69% to 84% for strong expressions and from 56% to 73% for moderate expressions. These values are comparable to RAVDESS [22] audio-only stimuli, which reported unbiased accuracy rates of 42%–64% across intensities and the ADFES-BIV [34] audio-visual dataset, which obtained 17%–63%, 23%–69%, and 30%–70% for low-, intermediate-, and high-intensity expressions, respectively. Furthermore, the identification responses were significantly above chance for every emotion category at both intensities. Overall, these results lend validity to our strong and moderate expressions.

2) RELIABILITY OF EMOTION CATEGORIES

Fleiss' Kappa statistic showed that overall agreement in identifying emotion categories was substantial at 0.76 and 0.64 for strong- and moderate-intensity expressions, respectively. All Kappa values for the emotion categories were within the substantial-to-almost-perfect-agreement range for strong expressions (0.70–0.85), and the Kappa values fell within the moderate-to-substantial-agreement range for moderate expressions (0.54–0.76). In RAVDESS [22], the auditory-only stimuli yielded overall Kappa values of 0.52 and 0.41 for strong intensity and moderate intensity, respectively. Their overall Kappa values for emotions varied between 0.53 and 0.67 across intensities; [22] did not provide emotion-specific Kappa values according to intensity. Notably, each stimulus was identified 10 times within RAVDESS. The SUBESCO auditory-only database [16], which was developed in two phases, reported overall Kappa values of 0.58 and 0.69 for Phase 1 and Phase 2, respectively. It should also be noted that, in SUBSECO, each stimulus was judged by just two raters. These results suggest that our listeners were substantially consistent in terms of emotion category identification, regardless of emotional intensity. Together, these Kappa values support the validity of our vocal emotional stimuli.

3) RELIABILITY OF NATURALNESS RATINGS

Rater agreement for ratings of perceived naturalness was also assessed using ICC. The single-measure ICC values were within the poor range for strong- and moderate-intensity expressions (0.14 and 0.12, respectively), whereas the average-measure ICC values were within the excellent range for strong- and moderate-intensity expressions (0.80 and 0.77, respectively). The ICC for a single rater generally yields smaller values than its corresponding ICC for the mean of multiple raters [57]. The observed rating patterns indicated that the ratings of perceived naturalness were more consistent within the same raters than between different raters, showing that individual raters were highly stable in assigning ratings across multiple instances, whereas ratings varied according to raters. The variability between raters suggests that each listener may have a different perspective of naturalness with respect to emotional categories. In RAVDESS [22], emotion stimuli were assessed in terms of genuineness ratings. Livingstone and Russo [22] reported that single-measure ratings of genuineness fell within the poor range of reliability (0.07) for speech (collapsed across intensities and modalities), and average-measure ratings of genuineness fell within the fair range of reliability (0.42). In this study, none of the emotion stimuli was repeatedly represented at each intensity. Nevertheless, our listeners displayed excellent consistency when examined through the multiple-rater ICC, which validates the CADKES stimuli.

B. IDENTIFICATION ACCURACY IS ASSOCIATED WITH NATURALNESS RATINGS

The effect of degree of naturalness on emotion category perception was assessed using Pearson's correlation. Results showed a strong positive relationship between identification accuracy and naturalness degree for both strong and moderate expressions; that is, the more accurately emotional expressions were categorized, the more naturally they were perceived (see Fig. 1). More precisely, nearly perfect positive correlations were observed for the happiness, sadness, anger, and boredom stimuli at strong intensity, and for the happiness, sadness, fear, and boredom stimuli at moderate intensity. These strong-to-nearly-perfect correlations suggest that the listeners performed systematically, and not arbitrarily, in judging the degree of naturalness of the emotion stimuli, providing support for the validity of CADKES. The observation that neutral emotion exhibited the lowest correlation coefficients at both strong and moderate intensities suggests that determining the degree of naturalness was more challenging for the neutral emotion than for the other emotion categories. To the best of our knowledge, no previous research has addressed the relationship between emotion identification accuracy and degree of naturalness to validate emotional speech databases. This is the first study to provide evidence that naturalness ratings affect identification accuracy rates, suggesting that future research should evaluate the degree of naturalness as a validation measure for emotional speech stimuli.

C. INTENSITY PLAYS A ROLE IN EMOTION PERCEPTION

The analysis of raw hit rates showed that strong expressions (87%) were significantly more accurately identified than their moderate counterparts (78%). This pattern aligns with the RAVDESS audio-only stimulus set [22], which reported raw hit rates of 67% and 58% at strong and moderate intensities, respectively. A similar pattern was observed in the Mandarin Chinese MES-P database [18], which achieved overall hit rates of 86.73% and 83.25% at intense and moderate intensities, respectively. CADKES obtained overall unbiased hit rates of 76% and 64% for strong and moderate expressions, respectively, which are comparable to the RAVDESS audio-only set [22], which achieved overall unbiased hit rates of 50% and 43% for strong and moderate expressions, respectively.

Furthermore, in CADKES, the differences in raw hit rates between strong and moderate expressions remained significant for each emotion category. The role of intensity in emotion perception was particularly evident for sadness and anger, showing 18% and 15% higher hit rates for the strong versions of sadness and anger, respectively, than their moderate versions. These results provide further evidence for the notion that intensity plays a role in emotional speech perception while validating distinctions between strong- and moderate-intensity levels in CADKES. It is also noteworthy that, for each emotion, the rating patterns were highly similar between strong and moderate expressions (see Table 7). In addition, all emotion stimuli were perceived as good-to-very-good variants of the specific emotion categories at both strong intensity (3.28–3.96) and moderate intensity (3.34–4.04). These patterns further validate the differential intensity levels produced in CADKES.

D. GENDER EFFECT IN EMOTION PERCEPTION

Results showed no interaction between listener gender and speaker gender, suggesting that male and female listeners performed similarly, regardless of speakers' gender. In addition, performance was not affected by the listener's gender at either intensity, with male and female listeners performing similarly, overall, at strong intensity (86.9% vs. 87%) and moderate intensity (77.7% vs. 78.5%). In contrast, speaker gender affected identification performance, showing that emotion stimuli were more accurately identified when female speakers produced the emotions (88.9%) than when male speakers produced the emotions (85.1%) at strong intensity. This pattern remained the same at moderate intensity (male speaker expressions, 76% vs. female speaker expressions, 80%). Previous research has produced mixed results regarding the gender effect in emotion identification. For example, the KSUEmotions database [14], which was constructed during two phases, reported that performance on male speaker expressions was better than on female speaker expressions in Phase 2, but no gender differences were found in Phase 1, whereas male listeners outperformed female listeners in both phases. Collignon et al. [48] found that female participants outperformed male participants in emotion identification for all auditory, visual, and audio-visual presentation modes, and performance was better for female speaker expressions than male speaker expressions.

E. LIMITATIONS AND FUTURE DIRECTIONS

This study has several limitations. First, listeners identified emotion categories in a six-alternative, forced-choice paradigm that does not allow for the possibility to answer "none." As can be seen in Table 3, neutral expressions achieved the highest raw hit rates at both strong and moderate intensities while showing large differences between the raw and unbiased hit rates at both intensities. These patterns are in line with those of the study of Sultana et al. [16], who employed a forced-choice format involving only the target emotions. In addition, all non-neutral emotions, with one exception, were most frequently misidentified as neutral at strong intensities, which became more evident at moderate intensities. These patterns suggest that listeners are biased toward neutral emotion when they are unsure of the target non-neutral emotions. This response bias was addressed by calculating the bias-corrected hit rates. However, comparing the results obtained in different paradigms may provide a meaningful avenue for future research to better understand emotion perception.

Another limitation is that strong neutral expressions were more accurately identified than their moderate counterparts,

although strong and moderate neutral expressions were considered similar in terms of naturalness (strong, 3.28 vs. moderate, 3.34). Thus, it remains unclear whether such an accuracy difference reflects variations in emotional intensity or whether this is because strong and moderate expressions were recorded at different time points. Further evidence is required to identify what drives this difference between the two intensities; thus, this result should be interpreted with caution. Meanwhile, this result raises an interesting question of whether neutral expressions may have different emotional intensities to a certain degree, given that emotions are perceived in a gradient manner rather than in an all-or-none manner, suggesting another important direction for future research.

Finally, it will be informative to assess valence and arousal ratings in addition to naturalness ratings to further explore how listeners process emotions with differential emotional intensities. It will be particularly informative to examine whether identification accuracy is associated with valence or arousal ratings and whether naturalness ratings are associated with valence or arousal ratings.

VII. CONCLUSION

This paper introduced CADKES, a Korean audio-only emotional speech database with strong and moderate intensities. CADKES is the only Korean database of emotional expressions produced at different levels of emotional intensity and the only Korean database that has been subjected to a set of validation measures. Validation measures yielded high levels of inter- and intra-rater reliability. In this study, every recording was identified and rated 25 times by asking all listeners to evaluate all stimuli, resulting in 135000 identification and 135000 naturalness rating responses. This is unlike the existing databases of comparable size, in which several groups of listeners rated a subset of stimuli. This further supports the validity and reliability of the CADKES database. Importantly, the present study adds to the literature pointing toward the role of intensity in emotion perception by observing that stronger emotional expressions are more easily identified than their more moderate counterparts. It also highlights the association between identification accuracy and naturalness degrees in emotion category distinctions by revealing that a more natural variant of an emotion is more accurately identified than its less natural variant. Overall, these validation results confirm that this dataset can serve as a valuable source for researchers in human emotion perception and machine-based emotion recognition. CADKES is available for research purposes upon request from the first author.

REFERENCES

[1] M. N. Dalili, I. S. Penton-Voak, C. J. Harmer, and M. R. Munafò, "Meta-analysis of emotion recognition deficits in major depressive disorder," *Psychol. Med.*, vol. 45, no. 6, pp. 1135–1144, Apr. 2015.
[2] J. A. Hall, S. A. Andrzejewski, and J. E. Yopchick, "Psychosocial correlates of interpersonal sensitivity: A meta-analysis," *J. Nonverbal Behav.*, vol. 33, no. 3, pp. 149–180, Sep. 2009.
[3] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, nos. 1–2, pp. 227–256, 2003.
[4] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *J. Personality Social Psychol.*, vol. 70, no. 3, p. 614, 1996.
[5] M. Carl, M. Icht, and B. M. Ben-David, "A cross-linguistic validation of the test for rating emotions in speech: Acoustic analyses of emotional sentences in english, German, and Hebrew," *J. Speech, Lang., Hearing Res.*, vol. 65, no. 3, pp. 991–1000, Mar. 2022.
[6] P. Laukka, D. Neiberg, M. Forsell, I. Karlsson, and K. Elenius, "Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation," *Comput. Speech Lang.*, vol. 25, no. 1, pp. 84–104, Jan. 2011.
[7] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, Z. Deng, S. Lee, S. Narayanan, and C. Busso, "An acoustic study of emotions expressed in speech," in *Proc. Interspeech*, Oct. 2004, pp. 1–4.
[8] M. W. Kraus, "Voice-only communication enhances empathic accuracy," *Amer. Psycholog.*, vol. 72, no. 7, p. 644, Oct. 2017, doi: 10.1037/amp0000147.
[9] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 829–837, Jul. 2000.
[10] L. Hansen, Y. Zhang, D. Wolf, K. Sechidis, N. Ladegaard, and R. Fusaroli, "A generalizable speech emotion recognition model reveals depression and remission," *Acta Psychiatrica Scandinavica*, vol. 145, no. 2, pp. 186–199, Feb. 2022.
[11] C. M. Jones and I. M. Jonsson, "Performance analysis of acoustic emotion recognition for in-car conversational interfaces," in *Proc. Int. Conf. Universal Access Hum.-Comput. Interact.* Berlin, Germany: Springer, Jul. 2007, pp. 411–420.
[12] M. F. M. Idros, A. H. A. Razak, S. Al Junid, A. K. Halim, and N. Khairudin, "Capability of voice recognition system for automatic signal in autonomous vehicle (AV) application," in *Proc. IEEE 5th Int. Conf. Smart Instrum., Meas. Appl. (ICSIMA)*, Nov. 2018, p. 16, doi: 10.1109/ICSIMA.2018.8688755.
[13] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Commun.*, vol. 48, no. 9, pp. 1162–1181, 2006.
[14] A. H. Meftah, Y. A. Alotaibi, and S.-A. Selouani, "Evaluation of an Arabic speech corpus of emotions: A perceptual and statistical analysis," *IEEE Access*, vol. 6, pp. 72845–72861, 2018.
[15] L. Abdel-Hamid, "Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features," *Speech Commun.*, vol. 122, pp. 19–30, Sep. 2020.
[16] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "SUST Bangla emotional speech corpus (SUBESCO): An audio-only emotional speech corpus for Bangla," *PLoS ONE*, vol. 16, no. 4, Apr. 2021, Art. no. e0250173.
[17] B. Gong, N. Li, Q. Li, X. Yan, J. Chen, L. Li, X. Wu, and C. Wu, "The Mandarin Chinese auditory emotions stimulus database: A validated set of Chinese pseudo-sentences," *Behav. Res. Methods*, vol. 54, pp. 1–19, May 2022.
[18] Z. Xiao, Y. Chen, W. Dou, Z. Tao, and L. Chen, "MES-P: An emotional tonal speech dataset in Mandarin with distal and proximal labels," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 408–425, Jan. 2022.
[19] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a Danish emotional speech database," in *Proc. 5th Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Sep. 1997, pp. 1–4.
[20] 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, 2008.
[21] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan. 2017.
[22] 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.
[23] A. Batliner, C. Hacker, S. Steidl, E. N?th, S. D'Arcy, M. J. Russell, and M. Wong, "'You stupid tin box'-children interacting with the AIBO robot: A cross-linguistic emotional speech corpus," in *Proc. Lrec*, 2004, pp. 171–174.

[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, vol. 5, Sep. 2005, pp. 1517–1520.
[25] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO corpus: An Italian emotional speech database," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC), Eur. Lang. Resour. Assoc. (ELRA)*, 2014, pp. 3501–3504.
[26] N. Keshtiari, M. Kuhlmann, M. Eslami, and G. Klann-Delius, "Recognizing emotional speech in Persian: A validated database of Persian emotional speech (Persian ESD)," *Behav. Res. Methods*, vol. 47, no. 1, pp. 275–294, Mar. 2015.
[27] M. B. Mustafa, M. A. M. Yusoof, Z. M. Don, and M. Malekzadeh, "Speech emotion recognition research: An analysis of research focus," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 137–156, Mar. 2018.
[28] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: A review," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 93–120, 2018.
[29] B. S. Kang, "A text-independent emotion recognition algorithm using speech signal," M.S. thesis, Dept. Electr. Comput. Eng., Yonsei Univ., Seoul South Korea, 2001.
[30] Y. Nam and C. Lee, "Cascaded convolutional neural network architecture for speech emotion recognition in noisy conditions," *Sensors*, vol. 21, no. 13, p. 4399, Jun. 2021.
[31] H. Mukherjee, H. Salam, A. Othmani, and K. C. Santosh, "How intense are your words? Understanding emotion intensity from speech," in *Proc. IEEE 21st Int. Conf. Commun. Technol. (ICCT)*, Oct. 2021, pp. 1280–1286.
[32] N. Holz, P. Larrouy-Maestri, and D. Poeppel, "The paradoxical role of emotional intensity in the perception of vocal affect," *Sci. Rep.*, vol. 11, no. 1, pp. 1–10, Dec. 2021.
[33] P. N. Juslin and P. Laukka, "Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion," *Emotion*, vol. 1, pp. 381–412, Dec. 2001, doi: 10.1037/1528-3542.1.4.381.
[34] T. S. H. Wingenbach, C. Ashwin, and M. Brosnan, "Validation of the Amsterdam dynamic facial expression set—Bath intensity variations (ADFES-BIV): A set of videos expressing low, intermediate, and high intensity emotions," *PLoS ONE*, vol. 11, no. 1, Jan. 2016, Art. no. e0147112, doi: 10.1371/journal.pone.0147112.
[35] P. Ekman, "An argument for basic emotions," *Cognit. Emotion*, vol. 6, nos. 3–4, pp. 169–200, 1992.
[36] P. Ekman, "All emotions are basic," in *The Nature of Emotion: Fundamental Questions*, P. Ekman and R. J. Davidson, Eds. New York, NY, USA: Oxford Univ. Press, 2019, pp. 15–19.
[37] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, p. 1161, Dec. 1980.
[38] M. Mehu and K. R. Scherer, "Emotion categories and dimensions in the facial communication of affect: An integrated approach," *Emotion*, vol. 15, no. 6, p. 798, Dec. 2015.
[39] S. D. Morgan, "Categorical and dimensional ratings of emotional speech: Behavioral findings from the Morgan emotional speech set," *J. Speech, Lang., Hearing Res.*, vol. 62, no. 11, pp. 4015–4029, Nov. 2019.
[40] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.
[41] 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.
[42] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera Am Mittag German audio-visual emotional speech database," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jun. 2008, pp. 865–868, doi: 10.1109/ICME.2008.4607572.
[43] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross lingual speech emotion recognition: Urdu vs. Western languages," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, Dec. 2018, pp. 88–93, doi: 10.1109/FIT.2018.00023.
[44] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2007, p. 733, doi: 10.1109/ICASSP.2007.366340.
[45] S.-W. Byun, J.-H. Kim, and S.-P. Lee, "Multi-modal emotion recognition using speech features and text-embedding," *Appl. Sci.*, vol. 11, no. 17, p. 7967, Aug. 2021.
[46] S.-W. Byun and S.-P. Lee, "A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms," *Appl. Sci.*, vol. 11, no. 4, p. 1890, Feb. 2021.
[47] A. Lausen and A. Schacht, "Gender differences in the recognition of vocal emotions," *Frontiers Psychol.*, vol. 9, p. 882, Jun. 2018.
[48] O. Collignon, S. Girard, F. Gosselin, D. Saint-Amour, F. Lepore, and M. Lassonde, "Women process multisensory emotion expressions more efficiently than men," *Neuropsychologia*, vol. 48, no. 1, pp. 220–225, Jan. 2010.
[49] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the radboud faces database," *Cognit. Emotion*, vol. 24, no. 8, pp. 1377–1388, Jun. 2010.
[50] S. Zhang, R. Liu, X. Tao, and X. Zhao, "Deep cross-corpus speech emotion recognition: Recent advances and perspectives," *Frontiers Neurorobotics*, vol. 15, Nov. 2021, Art. no. 784514.
[51] C. Turner and T. James, *Paradigm, (Version 1.0.2.479) [Computer Software]*. Lawrence, Kansas: Perception Research Systems, 2010.
[52] P. Boersma and D. Weenink,. (2010). *Praat: Doing Phonetics by Computers. Version 5.1.44*. Accessed: Sep. 13, 2021. [Online]. Available: https://www.fon.hum.uva.nl/praat/
[53] H. L. Wagner, "On measuring performance in category judgment studies of nonverbal behavior," *J. Nonverbal Behav.*, vol. 17, no. 1, pp. 3–28, 1993, doi: 10.1007/bf00987006.
[54] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, p. 378, 1971.
[55] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
[56] R. Müller and P. Büttner, "A critical discussion of intraclass correlation coefficients," *Statist. Med.*, vol. 13, nos. 23–24, pp. 2465–2476, Dec. 1994.
[57] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *J. Chiropractic Med.*, vol. 15, no. 2, pp. 155–163, 2016.
[58] D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," *Psychol. Assessment*, vol. 6, no. 4, p. 284, Dec. 1994.
[59] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1988.
[60] B. Winer, *Statistical Principles in Experimental Design*, 2nd ed. New York, NY, USA: McGraw-Hill, 1971.

YOUNGJA NAM received the B.S. degree in English language and literature from Kyungpook National University, South Korea, the M.S. degree in speech-language pathology from Daegu University, South Korea, and the Ph.D. degree in speech perception/phonetics from McGill University, Montreal, Canada.

She is currently a Research Professor with the Humanities Research Institute, Chung-Ang University, Seoul, South Korea. Her research interests include speech emotion recognition, machine translation, infant and adult speech perception, bi- and multilingualism, and psychoacoustics.

CHANKYU LEE received the B.S., M.S., and Ph.D. degrees from Chung-Ang University, Seoul, South Korea. He is currently a Professor with the Department of Korean Language and Literature, Chung-Ang University, where he also works as the Director of the Humanities Research Institute (HRI), which is dedicated to fostering AI-related interdisciplinary research, including collaboration between computer science and social sciences and humanities. HRI research lines include, but are not limited to, deep learning, machine learning, natural language processing, big data analysis, computer vision, speech recognition and synthesis, emotion recognition from speech and visual information, AI ethics, and social sciences and humanities-based approaches to AI. His research interests include AI-based chatbots, natural language processing, and multimodal emotion recognition.