Differential Impacts of Monologue and Conversation on Speech Emotion Recognition

Woan-Shiuan Chien, *Student Member, IEEE*, Shreya G. Upadhyay, *Student Member, IEEE*, Wei-Cheng Lin, *Member, IEEE*, Carlos Busso, *Fellow, IEEE*, and Chi-Chun Lee, *Senior Member, IEEE*

***Abstract*—The advancement of *Speech Emotion Recognition* (SER) is significantly dependent on the quality of emotional speech corpora used for model training. Researchers in the field of SER have developed various corpora by adjusting design parameters to enhance the reliability of the training source. For this study, we focus on exploring communication modes of collection, specifically analyzing spontaneous emotional speech patterns gathered during conversation or monologue. While conversations are acknowledged as effective for eliciting authentic emotional expressions, systematic analyses are necessary to confirm their reliability as a better source of emotional speech data. We investigate this research question from perceptual differences and acoustic variability present in both emotional speeches. Our analyses on multi-lingual corpora show that, first, raters exhibit higher consistency for conversation recordings when evaluating categorical emotions, and second, perceptions and acoustic patterns observed in conversational samples align more closely with expected trends discussed in relevant emotion literature. We further examine the impact of these differences on SER modeling, which shows that we can train a more robust and stable SER model by using conversation data. This work provides comprehensive evidence suggesting that conversation may offer a better source compared to monologue for developing an SER model.**

***Index Terms*—Monologue, conversation, speech emotion recognition, emotion perception, acoustic variability.**

I. Introduction

THERE has been a notable surge in the development of speech emotion recognition (SER) systems intended for real-world applications. Given that most SER systems are reliant on data, the quality of the emotional speech corpora is critical in the pursuit of creating more effective systems. When assembling a database for SER, the selection of key design parameters can significantly impact the quality of the resulting corpus. These factors include the type of settings employed (e.g., monologue, dyad, and small group), the elicitation methods used (e.g., reading, improvisation, and scripted responses), and emotional descriptors used for annotating the corpus (e.g., categorical emotions or emotional attributes). As an example, the widely-known IEMOCAP database [1] is collected in dyadic conversations featuring both scripted and spontaneous interactions, and labeled with emotion categories as well as three major emotional attributes including valence, arousal, and dominance. The RECOLA [2] database relies on dyadic interactions while solving a collaborative task over video conferencing, and annotated with arousal, valence and different types of laughers. These design parameters have a direct impact on the speech data collected creating variations in perception and variability of acoustic features. For instance, F0-contour and voice quality would exhibit dissimilarities in recordings acquired from either prompted or unprompted settings [3], and acoustic intensity of speech is higher in scripted responses compared to spontaneous speech in dyadic interactions [4]. These differences may naturally affect the emotional perception of raters. In particular, F0 and pitch are closely associated with the perception of arousal and valence in emotional speech [5], and lower intensity is linked to calmer and more negative emotions [6]. A clear understanding of the optimal settings for emotional database collection is crucial for improving the quality and applicability of the resulting corpora.

Researchers have conducted extensive investigations on the impact of various design parameters on the quality of emotional speech corpora. Then, they strategically permuted these design parameters to create a comprehensive and representative corpus, taking into consideration goals that better resemble emotional expressions observed during daily interactions and ultimately affect downstream speech technology. Through this process, researchers have successfully produced diverse and distinctive corpora that reflect the complexity and variability of human emotional expression, where the acquisition of datasets has transitioned from laboratory environments (e.g., IEMOCAP [1], MSP-IMPROV [7], EMODB [8]) to more realistic settings (e.g., VAM [9], MSP-Podcast [10], MSP-Conversation [11]). Although there is a theoretical understanding of how design parameters may influence the quality of emotional speech corpora used to train SER systems, only a few studies have explored the particular impact of data collection designs on SER models. For instance, it is well-known that simulated datasets may cause overfitting in SER models, leading to poor

Received 11 May 2023; revised 23 November 2024; accepted 26 November 2024. Date of publication 28 November 2024; date of current version 27 May 2025. This work was supported in part by the NSTC under Grant 111-2634-F-002-023 and Grant 110-2221-E-007-067-MY3, and in part by the NSF under Grant CNS-2016719. Recommended for acceptance by Special Issue On ACII 2022. *(Corresponding author: Woan-Shiuan Chien.)*

Woan-Shiuan Chien, Shreya G. Upadhyay, and Chi-Chun Lee are with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu City 30013, Taiwan (e-mail: wschien@gapp.nthu.edu.tw; shreya@gapp.nthu.edu.tw; cclee@ee.nthu.edu.tw).

Wei-Cheng Lin is with the Erik Jonsson School of Engineering & Commputer Science, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: wei-cheng.lin@utdallas.edu).

Carlos Busso is with the Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: busso@cmu.edu).

Digital Object Identifier 10.1109/TAFFC.2024.3509138

performance in real-world scenarios [12], whereas datasets collected from natural settings tend to have higher generalization capabilities. Likewise, corpora mainly featuring acted samples may benefit from employing a specific normalization approach, as these samples often exhibit heightened expressiveness in their emotional content [13], even if such expressions are less accurate in reflecting genuine emotional states than spontaneous speech [14]. Gustafson-Capková [15] indicated that emotional sentences are more readily identifiable in acted databases compared to spontaneous ones, as a result of the exaggerated expressions present in the former. Collectively, these design factors shape the robustness and consistency of SER systems. We analyze the *communication modes* in emotional databases, specifically monologue versus conversation speech, as explored in our preliminary study [16]. By systematically comparing their perceptual differences and acoustic variability, we investigate the impact of these differences on SER model construction.

Monologue and conversation are two distinct modes of communication that may exhibit significant variations in emotional expression and perception. A growing body of psychological research indicates that the emergence and understanding of emotional reactions heavily rely on the interpersonal dynamics and social implications of communication interactions [17], [18], [19], [20]. The emotions expressed by one person often elicit responses from her/his interlocutor, creating a strong interplay between the emotions of those involved in the conversation [21]. Burgoon et al. [22] showed that people were more likely to disclose their emotions and express empathy to a person in a conversation. Furthermore, speaking face-to-face offers an optimal environment for individuals to express and perceive emotions more naturally, as evidenced in the context of marital relationships [23], teacher-student interaction [24], inter-organization interaction [25] and business negotiation [26]. Additionally, from a speech production viewpoint, the acoustic properties of a speaker differ when they are engaged in a conversation compared to when they are merely speaking to themselves [27], [28], [29]. Speakers in conversation tend to adjust their speech rate and pause more frequently to allow for turn-taking, resulting in a more variable speech rhythm [30]. Conversations can foster richer emotional interactions, with evidence suggesting that positive emotions emerge more frequently during conversations [31] due to eliciting similar responses in interlocutors [32]. These findings imply that emotional data from conversations might be more authentic and representative than monologues, potentially providing "better-quality" sources for training SER systems. In this context, "better-quality" specifically refers to the enhanced performance robustness across varied training conditions and stability during the training process. In our study, these criteria are critical for evaluating the effectiveness of data sources for SER systems.

Although a significant body of evidence indicates that conversations are superior to monologues for eliciting authentic emotions, our study is unique in investigating this research question from the viewpoint of a database for SER. In this work, we build upon our preliminary work [16] and extend the work to further examine our running hypothesis that conversation represents a "better-quality" mode for SER from two theoretical perspectives: perceptual differences and acoustic variability, which in turn contribute to the differential impact on SER model learning. Specifically, we only use a small size dataset in our previous work to examine our hypothesis. However, in this work, we consider multi-lingual SER databases, performing these analyses on both the MSP-Podcast [10] and BIIC-Podcast [33] corpora, which are large-scale naturalistic emotional databases in English and Tawanese-Manderin, respectively. For previous work, we only consider the primary emotion categories, we examine both the emotion categories and emotion attributes in this work. Furthermore, there are three analyses that are also important yet we did not consider them in our previous work. First, we compare the range of the dispersed samples in terms of their corresponding emotional attributes. Second, we analyze the segments that express clear emotional behaviors based on perceptions. Third, we also examine the acoustic features on emotion attributes and whether they align with the empirical expectations. Finally, we extend the SER modeling experiments to validate the robustness and stability of conversation and monologues as sources when building an SER system. Our findings in this study are consistent with Chien et al. [16] and provide more comprehensive evidence to indicate that: 1) categorical emotions in conversations are more likely to reach consensus and rated higher rater consistency, with anger and sadness occupying a more confined area in the Valence-Arousal (V-A) perceptual space, 2) the distribution patterns in the V-A perceptual space and acoustic features in conversation align more closely with expected trends documented in emotion research, and 3) extracted recordings from conversations offer a more robust and stable source of information for training SER models.

The rest of the paper is organized as follows. Section II lists related work discussing the data collection settings and the expectation of emotion and acoustic suggested by psychological literature in SER systems. Section III introduces the multilingual corpus and the split of conversation and monologue samples used in this study. Section IV presents our empirical experiments and analyses in terms of emotion perception (Section IV-A) and acoustic variability (Section IV-B). Section V demonstrates the impact of these differences on the learning of SER models. Section VI summarizes the insights obtained from this paper.

II. Background and Related Work

This paper focuses on the impact of conversation and monologue when training SER. We aim at analyzing the differences from two perspectives: emotion perception and acoustic variability. This section describes the data collection settings used to build the database and the expectation of emotion perception and acoustic variability suggested by psychological literature.

A. Collection Settings for Existing SER Databases

The quality and reliability of speech emotion databases rely on various factors associated with the collection settings utilized. The spectrum of settings for eliciting emotional speech ranges from controlled to naturalistic, leading to diverse types and intensities of expressed emotions. By tailoring collection settings

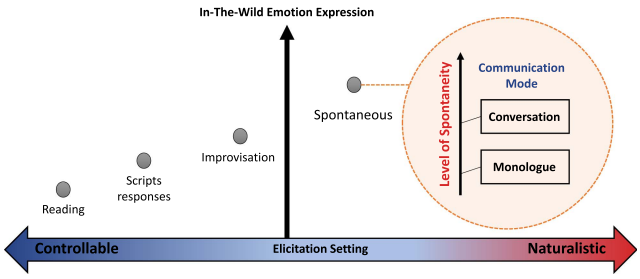


Fig. 1. A conceptual spectrum of widely used data collection settings for existing SER databases.

to align with research goals and objectives, researchers can effectively gather high-quality speech emotion data that accurately and reliably capture emotional expression in various settings. Fig. 1 illustrates the spectrum of collection settings for existing SER databases, spanning from controllable to naturalistic. This spectrum is designed to show how different data collection methods affect the spontaneity and authenticity of emotional expressions captured. On the left side, methods such as reading from scripts represent the most controlled environments, whereas improvisation and in-the-wild recordings, indicated towards the right, reflect more naturalistic settings. The placement of "Conversation" above "Monologue" in Fig. 1 highlights the typically higher spontaneity associated with interactive dialogue as compared to solo speech, which is more constrained by the lack of dynamic interaction.

Controllable parameters involve reading methods, allowing for a high degree of control over speech content and delivery, and may be suitable for eliciting specific emotions or speech features [37], [38]. Although reading is appropriate for eliciting discrete or categorical emotions, such as happiness or anger, and may provide a consistent and standardized approach to emotion elicitation [39], it may lack the naturalistic variability and prosody of speech and may not adequately represent the complexity and nuance of emotional expression in-the-wild [40]. Along with higher naturalistic settings, scripted responses and improvisation methods may produce highly natural and more diverse emotional behaviors [6], [41]. Acted speech is suitable for eliciting specific emotions and may provide a more controlled and consistent approach to emotion elicitation [41], yet it may be less authentic and may merely conform to the type of emotion [42].

At the other end of the spectrum lies naturalistic settings, in which scholars in the field of SER suggest that the collected speech is highly natural and emotionally rich. Examples of naturalistic collection protocols encompass unscripted conversations and interviews [38]. These methods typically involve minimal guidance or prompts, allowing for a broad range of emotional expressions and dynamics. In this study, we specifically focus on monologues and conversations, both of which are forms of spontaneous speech but differ in their level of spontaneity. Monologues are typically more consistent in one specific emotion [35]. Whereas when a second speaker joins the conversation, he/she can introduce new topics, alter the conversation's direction, or provide unexpected responses that the first speaker may react to. This turn-taking interaction fosters more natural and spontaneous emotional expressions. In contrast, conversations are more dynamic and interactive, with speakers responding to each other in real-time. Consequently, conversations tend to have a higher level of spontaneity compared to monologues [43]. Our preliminary work [16] introduced the differences between monologue and conversation concerning emotion perception and acoustic expressiveness. To the best of our knowledge, this is the first study to systematically elaborate on the specific differences between conversations and monologues, with the former being a "better-quality" source for constructing an SER model.

B. Distribution of Emotional Categories Within the Valence-Arousal Space

Emotions can be described in terms of emotional attributes. The two most common dimensions are arousal and valence. The space formed by these two attributes is very appealing to analyzing emotional differences. In this study, we use the Valence-Arousal (V-A) space to describe emotional differences between conversations and monologues. The V-A space is extensively utilized in affective computing and serves as a guide for annotating emotional categories in speech datasets. Empirical studies in psychological literature have studied the expectations of emotion perception for speech emotion recognition. For instance, Scherer et al. [44] conducted a study to examine the consensus among participants concerning the location of emotional categories in the V-A space. Participants were exposed to a series of emotional stimuli, including speech samples, and were asked to rate the stimuli based on valence and arousal. The results revealed a high degree of agreement among participants regarding the location of emotional categories in the V-A space, with categories such as happiness, sadness, anger, and fear situated in specific quadrants of the space [45]: happiness in quadrant I (i.e., high valence, high arousal), anger and fear in quadrant II (i.e., low valence, high arousal), and sadness in quadrant III (low valence, low arousal). Specifically, happiness is generally found in the high valence, high arousal quadrant, while sadness is situated in the low valence, low arousal quadrant. Other emotional categories, such as anger and fear, are positioned in the low valence, and high arousal quadrants. Moreover, there is evidence that the location of emotional categories in the V-A space can be influenced by various factors. For example, research has demonstrated that the type of speech (spontaneous or read) can impact the location of emotional categories in the V-A space. Spontaneous speech is characterized by greater variability in pitch, loudness, and timing, which can affect the perception of emotional states [35].

C. Acoustic Patterns Related to Basic Emotions

Acoustic cues are essentially produced by speakers to express different emotions, and the corresponding features are crucially used for speech emotion recognition. Several empirical research on the vocal communication of emotion has explored the relationship between different acoustic features and emotional states, including dimensional emotions and discrete emotions.

TABLE I
EMPIRICAL ACOUSTIC PATTERNS OF THE BASIC EMOTIONS IN TERMS OF ACOUSTIC CUES, INCLUDING EMOTION CATEGORIES AND EMOTION ATTRIBUTES [34], [35], [36]

	Emotion Category			Emotion Attribute	
	Happiness	Anger	Sadness	Arousal	Valence
F0 mean	↗	↗	↘	+++	- -
Intensity	↗	↗	↘	+++	- - -
High-frequency energy	↗	↗	↘	+++	- - -
Speech rate	↗	↗	↘	- - -	++

For explanation of + and -, ++/- -: $p < 0.01$, +++/- - -: $p < 0.001$.

Scherer [34] introduced the expected emotion-specific modulations concerning acoustic patterns for *Happiness*, *Anger* and *Sadness* and the expected acoustic correlates of arousal and valence. Table I summarizes the empirical trends for popular acoustic patterns of basic emotions. For instance, one of the most studied features is the fundamental frequency (F0), which is closely related to the perception of pitch. The F0 is known to increase when expressing emotions such as anger and happiness/joy [46]. In contrast, sadness is associated with a decrease in F0 [47], [48]. Positive valence was associated with low F0, and the arousal was positively correlated with F0 [49]. Intensity is another frequent acoustic feature that has been studied in the context of speech emotion recognition. Emotions such as anger and excitement are often associated with increased intensity, while sadness and fear tend to exhibit lower intensity levels [47], [50]. Positive valence is generally associated with higher intensity, while the lower intensity of speech is commonly linked to negative valence [49], [50]). Short-term speech energy has been found to be positively related to the arousal level of emotions [51]. In addition, the shape of the vocal tract is modified by emotional states, and features such as speech rate vary depending on the specific emotion being expressed. For instance, angry males tend to speak more slowly compared to angry females [39]. The empirical research of acoustic expressivity in SER has provided valuable insights into the relationship between acoustic features and emotion. These insights can inform the development of more reliable SER systems.

III. RESROUCES

A. MSP-Podcast Database

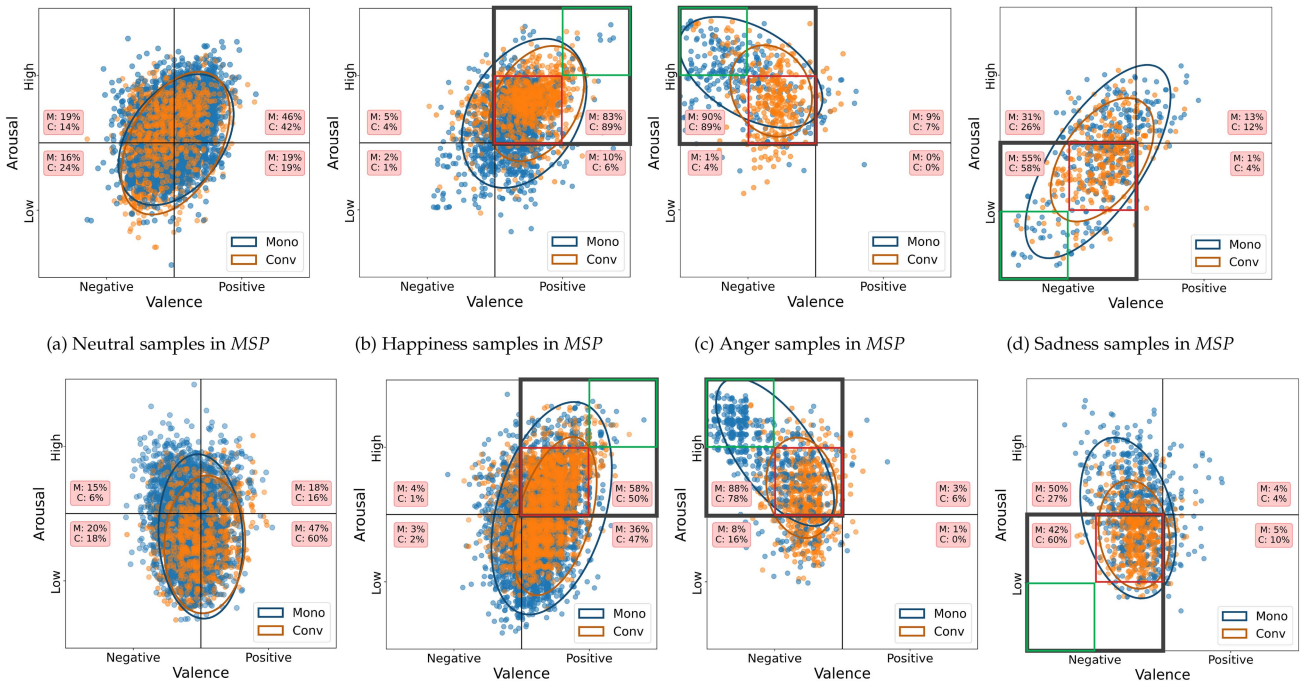
The MSP-Podcast is a collection of diverse and spontaneous speech samples with a range of emotional content from podcasts, which have been segmented into speaking turns to create a comprehensive repository of speech. This study employs version 1.10 of the corpus, containing 104,267 annotated utterances, making it an increasingly favored choice for SER-related research [52], [53], [54]. Its appeal stems from its large-scale, emotionally balanced dialogues from a variety of speakers, and diverse content spanning politics, movie reviews, science, technology, and economics. The corpus is assembled using a retrieval-based methodology as proposed by Mariooryad et al. [55] to retrieve target segments for emotional label annotation. Each segment has a duration ranging between 2.75 and 11 seconds. Emotional annotations are acquired through an adapted version of the crowd-sourced method presented by Burmania et al. [56]. A minimum of five workers annotate each segment, identifying primary emotions, secondary emotions and emotional attributes. A 7-point Likert scale is employed for annotating the attributes. The consensus labels are determined using the plurality rule for primary emotions and averaged dimensional ratings for emotional attributes.

B. BIIC-Podcast Database

We use a new SER database corresponding to the BIIC-Podcast corpus [33], which has a total of 145 hours of emotional speech. The collection of this corpus is an ongoing effort. The speech samples come from *Taiwanese Mandarin* podcasts, collected with a similar protocol as the one used for the MSP-Podcast corpus. The database was gathered from various audio-sharing platforms, which contain a wide range of diverse and naturalistic content, and carefully selected topics (sports, lifestyle, business, music, and more). The recordings include monologues and conversations, drama, interviews, casual conversations, etc. The duration of each segment is between 5 s and 16 s. The number of emotional annotations ranges from 3–7 per sample. Consistent with the majority of prior work, these segments have been annotated using eight primary emotion categories and three emotional attributes. The annotation protocol and the approach to obtaining consensus are the same as the one used for the MSP-Podcast corpus.

C. Monologue and Conversation Splits

We aim to comprehend the differential impacts on the SER between emotional speech samples obtained through two distinct *communication modes*: monologue and conversation. We first randomly selected 400 podcasts from the MSP-Podcast corpus and 500 podcasts from BIIC-Podcast corpus. Following the definition in our previous work [16], which includes 200 podcasts from the MSP-Podcast corpus and 250 podcasts from the BIIC-Podcast corpus, where these speaking turns are all given by one single speaker. We regard all speaking turns within these podcasts as representative instances of monologues, which we refer to as *Mono* in our analysis. Representative samples for conversations are obtained from the podcasts having speaking turns with two or more speakers, which we refer to this set as *Conv* in the rest of the paper. We manually checked that our selection of podcasts were genuine conversations and monologues to ensure the correctness of the experimental results and findings. Table II shows essential distributions regarding the speech samples from the *Mono* and *Conv* sets examined in this study. There are 10,004 utterances for *Mono* and 9,666 utterances for *Conv* in the MSP-Podcast corpus, and 10,869 utterances for *Mono*, and 12,966 utterances for *Conv* in the BIIC-Podcast corpus. Following our previous work [16], our study centers on samples with primary emotional labels that correspond to the four categorical emotional labels: *Neutral*, *Happiness*, *Anger* and *Sadness*. We additionally focus on the emotional attributes: *Arousal* and *Valence*. To ensure a fair comparison between these two communication modes without being influenced by data size, we further randomly select equal distributions of *Mono*



(a) Neutral samples in *MSP* (b) Happiness samples in *MSP* (c) Anger samples in *MSP* (d) Sadness samples in *MSP*

(e) Neutral for *BIIC* (f) Happiness samples in *BIIC* (g) Anger samples in *BIIC* (h) Sadness samples in *BIIC*

Fig. 2. Scatter plot depicting categorical samples on the valence-arousal (V-A) space for both *Mono* and *Conv* samples originating from *MSP* (2a-2d) and *BIIC* (2e-2h) sets. Each quadrant displays the occupancy rate for *Mono* (*M*) and *Conv* (*C*).

TABLE II
OVERVIEW OF THE NUMBER OF UTTERANCES FOR EACH PRIMARY EMOTION CATEGORY IN THE *MONO* AND *CONV* DATASETS

Datasets		Overall	Emotion Category			
			Neu	Hap	Ang	Sad
MSP-Podcast	*Mono*	10004	3977	1405	689	369
	Conv	9666	3360	1601	321	549
BIIC-Podcast	*Mono*	10869	4461	3970	1485	953
	Conv	12966	6062	4852	624	693
MSP	*Mono*	4600	3000	1000	300	300
	Conv	4600	3000	1000	300	300
BIIC	*Mono*	7200	3000	3000	600	600
	Conv	7200	3000	3000	600	600

and *Conv* samples for each primary emotion. We collectively refer to these subsets for each copus as *MSP* and *BIIC*.

IV. EMPIRICAL ANALYSES AND RESULTS

We conduct comprehensive analyses of the *MSP* and *BIIC* datasets to investigate the differences between speech recordings collected during conversations and monologues. We focus on evaluating the differences in emotional perception (Section IV-A) and acoustic variability (Section IV-B).

A. Emotional Perception

We conducted two analyses to evaluating emotional perception in the *Mono* and *Conv* sets on both *MSP* and *BIIC* datasets: 1) the categorical emotions in the valence-arousal (V-A) perceptual space and 2) measurements of the inter-annotator agreement.

1) Categorical Labels in the Valence-Arousal Space: To visualize the emotional perceptual differences between the *Mono* and *Conv* sets for each emotional category, we scatter the samples from the *MSP* and *BIIC* sets for *Neutral*, *Happiness*, *Anger* and *Sadness* in accordance with valence and arousal scores. Fig. 2 shows four by two separate plots for each emotion in the *MSP* and *BIIC* datasets. The blue points and the orange points represent samples from the *Mono* and *Conv* sets, respectively. Ellipsoidal regions are illustrated to encompass 80% of the samples. Based on Fig. 2, we enumerate the following observations by comparing the differences between samples in the *Mono* and *Conv* sets for each emotional category on both datasets.

1) Differences between sets in terms of the location of the centroids of the ellipses: While the distributions for *Neutral* appear to be similar, there is a visible difference in the distribution of *Happiness*, *Anger*, and *Sadness* emotions between *Mono* and *Conv* in both datasets. To understand and analyze these apparent differences, we first perform statistical testing using a two-tailed Mann-Whitney U rank test to compare the samples in the *Mono* and *Conv* sets. We separately evaluate the scores for valence and arousal separately. Table III presents these findings, revealing significant differences (p-value ≤ 0.01) in the mean arousal and valence levels between *Mono* and *Conv*. From the table, we can first observe that the standard deviation (SD) calculated for all emotions in the *Conv* are smaller than those in *Mono*. Second, for *Anger* and *Sadness*, the results indicate significant differences in both valence and arousal for *MSP* and *BIIC* sets. In *BIIC*, a significant difference in arousal is also present for the *Happiness*, while no significant differences are observed for *Netural*. This analysis demonstrates that clear differences exist in the statistical

TABLE III
STATISTICAL DIFFERENCES IN VALENCE AND AROUSAL BETWEEN *MONO* AND *CONV*

		Valence		Arousal	
MSP		Mean ± SD	p-value	Mean ± SD	p-value
Neutral	*Mono*	4.05 ± 0.64	0.3981	4.08 ± 0.76	0.3510
	Conv	4.03 ± 0.60		3.99 ± 0.75	
Happiness	*Mono*	4.71 ± 0.61	0.6733	4.72 ± 0.73	0.0167
	Conv	5.01 ± 0.58		4.96 ± 0.71	
Anger	*Mono*	2.42 ± 0.97	**0.0006**	5.73 ± 0.75	**0.0001**
	Conv	2.91 ± 0.71		5.27 ± 0.79	
Sadness	*Mono*	3.08 ± 0.83	**0.0044**	3.55 ± 1.11	**0.0011**
	Conv	3.17 ± 0.76		3.59 ± 0.91	
BIIC		Mean ± SD	p-value	Mean ± SD	p-value
Neutral	*Mono*	3.99 ± 0.47	0.2106	3.44 ± 0.92	0.3151
	Conv	4.09 ± 0.45		3.24 ± 0.77	
Happiness	*Mono*	4.73 ± 0.58	0.8125	4.15 ± 1.04	**0.0089**
	Conv	4.81 ± 0.51		3.96 ± 0.97	
Anger	*Mono*	2.36 ± 0.85	**0.0003**	5.53 ± 1.06	**0.0002**
	Conv	3.00 ± 0.57		4.66 ± 0.82	
Sadness	*Mono*	3.16 ± 0.58	**0.0022**	3.92 ± 0.97	**0.0010**
	Conv	3.30 ± 0.52		3.52 ± 0.79	

The bold in "p-value" indicates significant differences between *Mono* and *Conv* utilizing a two-tailed Mann-Whitney U rank test.

data for *Happiness*, *Anger*, and *Sadness* in their placement in the V-A space.

2) Comparison of the quadrant-specific occupancy rate between the two groups, which refers to the proportion of samples from an emotion category positioned in the expected V-A quadrant: As shown in Fig. 2 with a pink box, we compute the quadrant-specific occupancy rate. For *Happiness*, the occupancy rate in quadrant I is 89% (*MSP*) and 50% (*BIIC*) for samples in the *Conv* set, and 83% (*MSP*) and 58% (*BIIC*) for samples in the *Mono* set. We observe similar trends for *Anger*, where the occupancy rate in quadrant II of samples in the *Conv* set is 89% (*MSP*) and 78% (*BIIC*), but 90% (*MSP*) and 88% (*BIIC*) for samples in the *Mono* set. Likewise, for *Sadness*, the occupancy rate in quadrant III of samples in the *Conv* set is 58% (*MSP*) and 60% (*BIIC*), but only 55% (*MSP*) and 42% (*BIIC*) for samples in the *Mono* set. Although our previous work [16] has indicated that the majority of samples in the *Conv* set are located in the expected quadrant corresponding to their specific class, the difference is less clear possibly due to the increase in data size. Therefore, we investigate the range of the ellipses. We notice that the range of the ellipse for *Happiness*, *Anger*, and *Sadness* appears to be broader and more dispersed in *Mono*, whereas in *Conv*, it is significantly narrower and more concentrated; *Neutral* shows no such difference. We first calculate the area of each ellipse according to Fig. 2, which is summarized in Table IV. We observe that the area values of *Mono* are significantly larger than those of the *Conv*. Despite the different inspection angles, this insight still aligns with the results we previously obtained, indicating that the *Mono* set contains samples that further deviate from a specific quadrant. The emotional content of conversations is perceptually closer to the expected patterns.

3) Differences in the spread of the scattered samples on the V-A space: As illustrated in Fig. 2, it is evident that there are considerably more dispersed samples in *Mono*. Moreover, upon

TABLE IV
AREA FOR EACH ELLIPSE ACCORDING TO FIG. 2

		Neutral	Happiness	Anger	Sadness
MSP	*Mono*	7.06	8.75	7.47	12.98
	Conv	7.28	5.17	3.51	5.89
	Diff	-0.22	3.58	3.96	7.09
BIIC	*Mono*	6.52	12.02	8.19	7.11
	Conv	5.25	6.20	3.34	3.58
	Diff	1.27	5.82	4.85	3.53

"Diff" values obtained by subtracting *Mono* from *Conv*.

cross-referencing with Table IV, we find that the area values of *Mono* are notably larger than those of the *Conv*. Interestingly, we visually observe a noticeable similar trend for *Anger* in *Mono* on both datasets, which have larger values on the major axis of the ellipse. Namely, *Mono* have more extreme emotional values in their respective quadrant. Therefore, we further examine the scales of the samples with extreme emotional values (the green box) or neutral emotional values (the red box) in their respective quadrant. For *Anger*, this area of extreme values is defined as the occupancy rate in the area with valence ≤ 2.5 and arousal ≥ 5.5, and this area of neutral values is defined as the occupancy rate in the area with valence values from 2.5 to 4 and arousal values from 4 to 5.5. For *Sadness*, this area of extreme values is defined as the occupancy rate in the area with valence ≤ 2.5 and arousal ≤ 2.5, and this area of neutral values is defined as the occupancy rate in the area with valence and arousal both ranging from 2.5 to 4. A significant number of samples from the *Mono* set are in these extreme areas with more extreme values in the V-A space, i.e., in *MSP*, 58.7% for *Anger* and 23.6% for *Sadness*, and 23.6% for *Anger* in *BIIC*. In contrast, the proportion of samples in the extreme area for samples of the *Conv* set in *MSP* are only 3.8% for *Anger* and 1.8% for *Sadness*, and 8.2% for *Anger* in *BIIC*. Another point to consider is there are a significant proportion of samples from the *Conv* located in the inner regions, exhibiting more neutral values in the V-A space. i.e., in *MSP*, 76.3% for *Anger* and 40.6% for *Sadness*, and in *BIIC*, 60.8% for *Anger*. Whereas only 23.3% of *Anger* samples and 25.4% of *Sadness* samples from the *Mono* set in *MSP* and 30.9% of *Anger* samples in *BIIC* are located in the neutral area. It is interesting that the phenomenon of "scales" occurs identically across multi-lingual datasets. This phenomenon of "less-extreme" is supported by previous studies that individuals experience less sadness or anger when they have someone to talk to or participate in social interactions with [57]. This is because people tend to reduce or suppress negative emotions when engaging in social behaviors and interacting with someone [58].

2) Inter-Annotator Agreement: Inter-annotator agreement (IAA) is an important indicator to assess the perceptual differences between *Mono* and *Conv*, showing how easy the raters to clearly delineate the emotion category and how trustworthy the annotation is. It provides further evidence to evaluate the perceptual consistency with raters while perceiving from these recordings. We calculate the agreement values using raw data on the entire *Mono* and *Conv* sets, including emotions that were not included in the analysis (i.e., surprise, fear, disgust,

TABLE V
RESULTS OF INTER-ANNOTATOR AGREEMENT (κ) FOR CATEGORICAL EMOTIONS

		Overall	Consensus	Non-Consensus
MSP-Podcast [10]		0.229	0.265	-0.008
MSP	Mono	0.421	0.464	-0.022
	Conv	**0.456**	0.470	0.012
		Overall	Consensus	Non-Consensus
BIIC-Podcast		0.337	0.418	-0.014
BIIC	Mono	0.334	0.433	-0.079
	Conv	**0.375**	0.443	0.009

and contempt). *Fleiss' Kappa* (κ) [59] statistics is adopted as the evaluation metric since these are categorical emotions and it can have any number of annotators, where every segment is not necessarily annotated by each annotator. The κ result can be interpreted as follows [60]: values ≤ 0 as indicating no agreement and [0.01-0.20] as none to slight, [0.21-0.40] as fair, [0.41-0.60] as moderate, [0.61-0.80] as substantial, and [0.81-1.00] as almost perfect agreement. Based on Table V, we summarize two observations by investigating the similarities and differences between κ values in the *Mono* and *Conv* sets on both datasets.

1) Differences in the raters' consistency: Table V presents the original inter-annotator agreement for the MSP-Podcast and BIIC-Podcast datasets before selecting the subsets and the results of the inter-annotator agreement for both the *Mono* and *Conv* sets. A comparison reveals higher agreement for categorical emotions in the *Conv* set than the *Mono* set. This is evident for both *MSP* and *BIIC*, with absolute κ value increases of 0.035 and 0.041, respectively. It indicates that the emotions expressed in sentences from the *Conv* set are more consistently perceived by annotators. This increased consistency value suggests that these sentences convey less ambiguous emotions that align more closely with the annotators' expected emotion impressions. Notably, the inter-annotator agreement findings correspond with the outcomes discussed in Section IV-A1 concerning the distribution of dispersed samples in the V-A space analysis. In the *Conv* set, sentences representing *Happiness*, *Anger*, and *Sadness* show a concentrated area within the expected quadrant on the V-A space and higher overall inter-annotator agreement. These emotional sentences demonstrate better consistency in categorical labels and V-A space positioning. These findings suggest that they are more perceptually coherent than those in the *Mono* set and also align with [16].

2) Differences and similarities in the proportion of consensus: The MSP-Podcast and BIIC-Podcast databases differ in their mechanism for achieving consensus among evaluators. In the BIIC-Podcast database, the minimum number of raters required for providing ratings is set to three. If a consensus label is not reached under the majority rule, the number of raters is increased to achieve a consensus label. We define the samples with a number of evaluators greater than three in BIIC-Podcast as non-consensus samples in this analysis. On the other hand, the MSP-Podcast database requires at least 5 raters to provide ratings, irrespective of whether a consensus label is achieved under the majority rule. It would not further process the sample without consensus.

The purpose of this analysis is to determine whether there exist inherent similarities or differences in conveying or being perceived emotions between conversation and monologue samples. Hence, we examine the proportion of consensus achieved. For this analysis, we use all segments originating from their respective podcasts, which are assigned to a single speaker (monologue) and more than two speakers (conversation). The results show that in the BIIC-Podcast, 84% of conversation samples achieve consensus with the minimum number of raters, while in the MSP-Podcast, 78% of conversation samples reach consensus. In the BIIC-Podcast, 44% of monologue samples achieve consensus with the minimum number of raters, while in the MSP-Podcast, 55% of monologue samples reach consensus. The higher proportions suggest that conversation sets convey emotions more clearly, with less ambiguity, and are more consistent with annotators' emotional impressions.

Furthermore, based on the kappa values in Table V for datasets split by the presence or absence of consensus among raters, there is virtually no difference in inter-annotator agreement for consensus segments between conversation and monologue sets, with the maximum difference being 0.01. It has been observed that samples without consensus (i.e., those difficult to evaluate) exhibit poorer agreement between conversation and monologue sets, with a level of less-than-chance agreement. This analysis indicates that *Conv* and *Mono* sets all contain samples that are difficult to evaluate, but those are less in *Conv* set. These findings are consistent with the previously discussed spread of scattered samples (Sec), i.g. the distribution of *Mono* is scattered with more extreme values, and the distribution of *Conv* is more central.

B. Acoustic Variability

This section uses two datasets to investigate potential acoustic feature differences between *Mono* and *Conv* set samples. We focus on five key acoustic features: fundamental frequency (F0), intensity, high-frequency energy (HFEnergy), word-level speech rate (WordSR), and phone-level speech rate (PhoneSR). We use Praat tool [61] to extract LogF0, intensity, and high-frequency energy (cut-off 500 Hz). The normalization for LogF0 is performed per speaker, while the normalization for intensity and high-frequency energy is conducted per podcast to reduce channel effects. Speech rate features are calculated by averaging the number of spoken words and phoneme classes per second using forced alignment results. More specifically, we employed the *Montreal Forced Aligner* [62] to extract phones and words for *MSP* and trained a *Taiwanese Mandarin* forced aligner [63] using the *Formosa* database to obtain phones and words for *BIIC*. The process for *BIIC* was conducted in accordance with the methodology outlined in [64]. These alignments are based on human transcriptions.

1) Sample Distribution of Acoustic Features: Fig. 3 presents an analysis of acoustic features for emotional categories in the *Mono* and *Conv* sets. The mean value of frames per sentence is used as the sentence-level feature descriptor. In Fig. 3, error bars

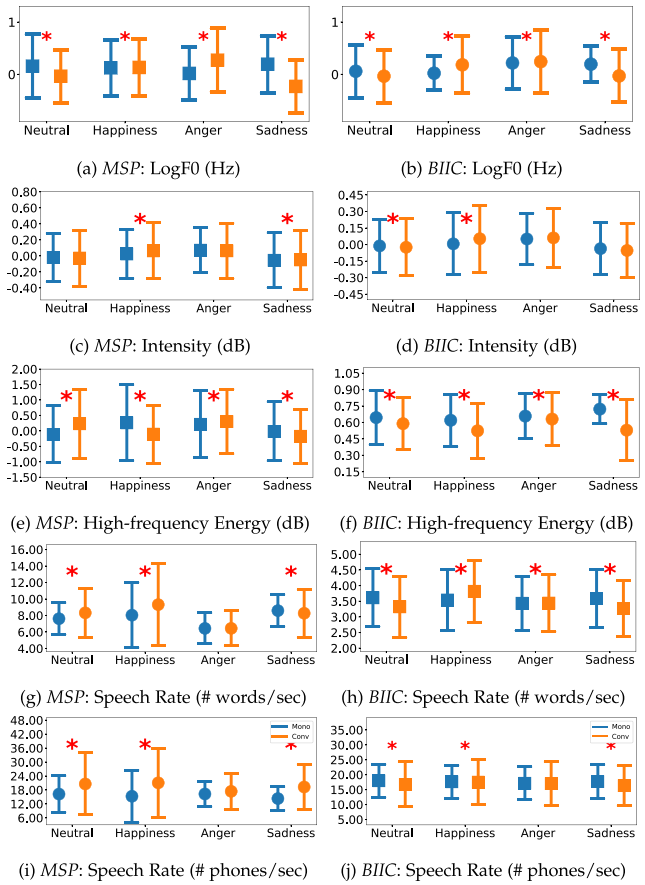


Fig. 3. Results of acoustic analyses on the *Mono* and *Conv* sets, plotted for each primary emotion, with each panel corresponding to a different acoustic cue. Error bars show the standard deviation from the mean. $*$-tagged results denote the statistical significance (two-tailed T-test, p-value ≤ 0.01) between the *Mono* and *Conv* features.

illustrate the statistics (mean and standard deviation) for each emotion-specific set of sentences (e.g., the *Anger* set). There are two major points to be noticed: 1) significant differences in acoustic expressiveness between the *Conv* and the *Mono* set for each emotional category, and 2) the acoustic expression levels, as measured by acoustic features of *Conv* are generally higher and more variable than those of *Mono*. First, we perform a two-tailed T-test for statistical comparison between the *Mono* and *Conv* sets. We find significant differences with $p \leq 0.01$ in acoustic features for each emotional category. These differences in acoustic patterns align with the perceptual differences observed in Section IV-A. Second, and more importantly, we visually observe an apparent difference in the standard deviation values of the acoustic variability. Namely, the acoustic properties of speech can vary more in *Conv* than in *Mono*. It is noteworthy to align with most of the literature suggesting that interactive conversations can produce more diverse and complex acoustic expressions [65], [66]. In particular, this variability is more visible at speech rates (Fig. 3(i) and (j)) due to the presence of turn-taking dynamics and varying cognitive loads during interaction [67], [68].

2) Consistency in Relation to Empirical Expectations: In Section II-C, we examined the empirical patterns of acoustic

TABLE VI
CORRELATION BETWEEN DIFFERENT ACOUSTIC FEATURES AND EMOTION ATTRIBUTES

		MSP				BIIC			
		Arousal		Valence		Arousal		Valence	
		R	*p*	*R*	*p*	*R*	*p*	*R*	*p*
F0	***M***	0.324	++	-0.245	-	0.330	++	0.011	+
	C	0.268	++	-0.332	- - -	0.249	++	-0.117	-
Intensity	***M***	0.312	++	-0.310	- -	0.371	+++	-0.224	- -
	C	0.139	+	-0.289	- -	0.213	++	-0.354	- - -
HFEnergy	***M***	0.180		0.203	++	0.088	+	0.144	
	C	0.202	+	-0.212	- -	0.101	+	-0.122	-
WordSR	***M***	-0.208	-	-0.043		0.023		0.089	+
	C	-0.233	-	0.200		0.200	+	0.133	++
PhoneSR	***M***	-0.055	-	0.247	++	0.102	+	0.145	+
	C	-0.293	- -	0.308	++	-0.193	- -	0.287	++

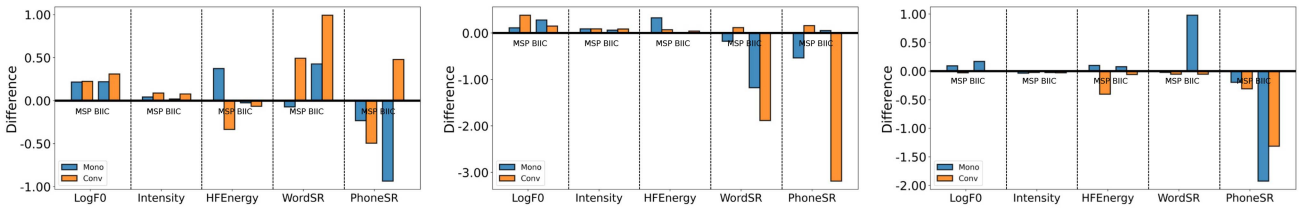
For explanation of + and -, ++/- -: $p < 0.01$, +++/- - -: $p < 0.001$.

TABLE VII
A SUMMARY TABLE OF RELATIVE DIFFERENCES OF DIFFERENT ACOUSTIC FEATURES OF *MONO* OR *CONV*, WHICH IS CONSISTENT WITH RESPECT TO THE EMPIRICAL EXPECTATIONS

Emotion	Happiness		Anger		Sadness		Arousal		Valence	
MSP	***M***	***C***	***M***	***C***	***M***	***C***	***M***	***C***	***M***	***C***
LogF0	✓	✓	✓	✓		✓	✓	✓	✓	✓
Intensity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
HFEnergy	✓		✓	✓		✓		✓		✓
WordSR		✓		✓	✓	✓	✓	✓		
PhoneSR				✓	✓	✓	✓	✓	✓	✓
BIIC	***M***	***C***	***M***	***C***	***M***	***C***	***M***	***C***	***M***	***C***
LogF0	✓	✓		✓	✓	✓	✓	✓	✓	✓
Intensity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
HFEnergy				✓	✓	✓	✓	✓		✓
WordSR	✓	✓				✓			✓	✓
PhoneSR		✓	✓		✓	✓		✓	✓	✓

features correlating to emotions, including categorical emotions and emotion attributes. From the previous section, it is evident from Fig. 3 that the emotion-specific modulation of acoustic patterns for sentences in the *Conv* set aligns more closely with empirical expectations than those in the *Mono* set. First, with respect to emotional categories, we measure the relative changes in acoustic features for *Happiness*, *Anger*, and *Sadness* sentences compared to those from *Neutral* sentences. In this analysis, reference values include all sentences labeled as *Neutral* in the MSP-Podcast and BIIC-Podcast corpora. We investigate if the relative acoustic differences between emotional and neutral sentences align with expected trends and present the results of relative feature values in Fig. 4.

Upon cross-referencing Fig. 4 with Table I, it becomes evident that the expected trends in Table I consistently align with the modulation patterns exhibited by the *Conv* sentences in Fig. 4. For instance, we observe a higher LogF0, intensity and high-frequency energy for *Anger* sentences and lower intensity, high-frequency energy and speech rates for *Sadness* sentences. These observations are consistent across both *MSP* and *BIIC*. Contrarily, sentences in the *Mono* set display several conflicting trends. For example, the *Sadness* sentences have much higher speech rates than *Anger* sentences in *MSP* and much higher high-frequency energy than *Anger* sentences in *BIIC*. By and large, sentences in the *Mono* set are more inconstant as compared to sentences in *Conv* set.



(a) Happiness (b) Anger (c) Sadness

Fig. 4. Relative differences toward the *Neutral* emotion of different acoustic features for *Happiness*, *Anger* and *Sadness*.

Additionally, with respect to emotional attributes, we evaluate the Pearson correlation between acoustic features and the value of emotion attributes. As an example, in *MSP*, we compute the Pearson correlation between the intensity features and the values of *Arousal* in the *Mono* set, obtaining they have a significantly positive correlation of $R = 0.312$ and $p = 0.002$. Table VI presents the results of correlation values (R) and p-value (p) for *Conv* and *Mono* in both datasets. By cross comparing Tables I and VI, we can observe that the modulation patterns of the *Conv* sentences in Table VI are in a consistent manner with the expected correlations in Table I. For instance, higher LogF0 values are associated with higher *Arousal*, leading to positively correlated outcomes. Likewise, lower high-frequency energy values lead to higher *Valence* values, resulting in a negative correlation. Conversely, sentences in the *Mono* set exhibit an inverse correlation between speech rates and *Arousal* values, or even lack significant correlation.

These analyses show that the acoustic patterns in the *Conv* set are better consistent to the empirical emotion-specific modulation. This consistency is not always observed in the *Mono* set, which exhibits non-intuitive and irregular patterns. Table VII summarizes two parts of our acoustic analyses for *Mono* and *Conv*: the relative differences of emotion categories and the correlation of emotion attributes, which matches the empirical expectations. According to our results, out of the 25 evaluation metrics (i.e., five acoustic features corresponding to five emotions), 88% of the Conv set in the *MSP* conform to empirical expectations, and 84% meet these expectations in *BIIC*. However, in the *Mono* set, only 64% of the metrics align with empirical expectations for both *MSP* and *BIIC* sets. The acoustic patterns and evaluator-perceived results (Section IV-A) indicate that the samples in the *Conv* set contain emotional information that is better aligned with findings in emotion literature. Collectively, these deep findings once again provide evidence that emotional information conveyed in the *Conv* set is of "better-quality" (with regard to perceptual assessment and acoustic expression) than that in the *Mono* set. In addition, we can corroborate the same findings across the two different language datasets. In the following section, we will conduct a more extensive comparison between *Conv* and *Mono* for training SER systems.

V. Impacts on SER Model Learning

In Section IV, we carry on comprehensive empirical analyses of emotion perception (Section IV-A) and acoustic variability (Section IV-B), which demonstrate that samples in the *Mono* and the *Conv* differ in their emotional perceiving to the raters and emotional conveying from the speakers. We carry out SER experiments to further investigate the impact of utilizing sentences from either the *Mono* or *Conv* sets when training a SER system.

A. Experimental Setup

In this experiment, the podcasts are randomly divided into training (85%), validation (5%), and testing (10%) subsets to ensure the podcast-independent scheme is applied. We use vq-wav2vec representation [69] as input for training our models. Categorical emotion recognition is performed as binary classification, targeting *Neutral*, *Happiness*, *Anger*, and *Sadness* emotions. Emotion attribute recognition experiments are conducted as regression tasks.

1) Baseline Experiments: Four networks are selected for all the basic experiments: DNN, Convolutional Neural Network (CNN) [70], Gated Recurrent Unit (GRU) [71], and Transformer (Trans.) [72]. For both CNN and GRU, we employ a two-layer model with 256 hidden nodes. The Transformer utilizes a two-layer, two-head self-attention mechanism. The learning rate and decay factor are 0.001, and we use the Adam optimizer. Binary classification tasks use binary cross-entropy loss, while regression SER tasks employ mean squared error (MSE) loss. The networks are trained for up to 50 epochs, with a batch size of 64 and early stopping.

2) Fine-Tuned Pre-Trained Models Experiments: Speech Self-Supervised Learning (SSLM) has emerged as a particularly promising direction in the development of speech models, with recent studies demonstrating its SOTA performance in SER tasks [73], [74]. In light of these advancements, our study leverages the pre-trained vq-wav2vec model [69] for fine-tuning. To validate our research hypothesis, we conduct fine-tuning experiments using two separate instances of the pre-trained vq-wav2vec model for investigating the differential impacts of *Conv* and *Mono* data on SER performance. The first model ($\mathrm{FT}_{\mathbf{Conv}}$) is fine-tuned exclusively with *Conv* samples, while the second model ($\mathrm{FT}_{\mathbf{Mono}}$) is fine-tuned with *Mono* samples. This approach allows us to understand and compare the impact differences in how each *communication mode* contributes to the effectiveness of SER models.

3) Evaluations: To investigate the potential effects of different *communication modes* as sources for training SER models. To validate our hypothesis, We follow the two same evaluation scenarios from [16]: the "Matched" and "Mismatched"

TABLE VIII
SER PERFORMANCES IN UAR(%) WITH SD FOR EACH EMOTIONAL CATEGORY WITH DIFFERENT SCENARIOS

Scenario		*MSP*				*BIIC*			
		Matched		Mismatched		Matched		Mismatched	
Task	Model	$M \rightarrow M$	$C \rightarrow C$	$M \rightarrow C$	$C \rightarrow M$	$M \rightarrow M$	$C \rightarrow C$	$M \rightarrow C$	$C \rightarrow M$
		UAR ± SD	UAR ± SD	UAR ± SD	UAR ± SD	UAR ± SD	UAR ± SD	UAR ± SD	UAR ± SD
Neutral	DNN	53.07 ± 1.91	63.92 ± 1.21	50.79 ± 3.29	61.02 ± 1.23	**64.17** ± 2.71	63.20 ± 1.26	**68.51** ± 0.24	56.30 ± 0.12
	CNN	**60.64** ± 4.65	**67.07** ± 3.97	**62.19** ± 3.47	63.39 ± 1.61	61.81 ± 1.19	**64.05** ± 1.35	67.34 ± 0.13	**57.86** ± 0.30
	GRU	58.30 ± 1.73	58.81 ± 1.86	52.38 ± 3.24	61.75 ± 2.20	54.90 ± 1.84	59.04 ± 2.20	55.94 ± 0.92	53.16 ± 0.57
	Trans.	59.18 ± 3.69	64.87 ± 1.33	54.30 ± 6.34	**64.24** ± 2.25	62.94 ± 1.46	63.12 ± 1.95	62.47 ± 1.39	55.85 ± 1.33
Happiness	DNN	60.43 ± 2.39	60.39 ± 1.95	55.31 ± 3.44	55.65 ± 2.02	**65.90** ± 2.37	66.11 ± 1.51	**58.56** ± 2.42	72.30 ± 1.47
	CNN	58.20 ± 2.98	60.20 ± 2.44	54.50 ± 4.88	59.59 ± 2.42	63.23 ± 0.42	**66.63** ± 1.36	58.30 ± 1.20	**72.50** ± 1.75
	GRU	60.33 ± 3.41	55.65 ± 2.59	55.30 ± 2.64	57.48 ± 1.83	53.86 ± 0.97	58.04 ± 1.89	49.60 ± 2.27	65.52 ± 2.52
	Trans.	**64.57** ± 3.37	**66.54** ± 2.50	**61.00** ± 3.64	**62.20** ± 1.43	63.94 ± 1.19	64.32 ± 2.77	58.14 ± 1.72	69.37 ± 1.73
Anger	DNN	59.85 ± 0.48	61.43 ± 1.41	55.00 ± 2.00	61.58 ± 3.33	59.00 ± 0.83	60.76 ± 1.27	**59.28** ± 1.93	**62.83** ± 1.35
	CNN	63.20 ± 4.01	63.51 ± 1.06	54.35 ± 3.30	58.30 ± 1.73	59.90 ± 3.23	**61.80** ± 1.88	58.14 ± 1.10	62.49 ± 3.33
	GRU	61.22 ± 1.21	59.12 ± 1.24	54.97 ± 2.04	53.37 ± 1.74	56.55 ± 1.34	60.29 ± 1.23	53.24 ± 0.89	56.98 ± 0.61
	Trans.	**63.13** ± 4.91	**65.74** ± 3.33	**57.56** ± 3.77	**61.64** ± 3.25	**61.17** ± 2.59	61.79 ± 1.52	56.78 ± 2.65	61.51 ± 1.51
Sadness	DNN	60.00 ± 2.77	55.00 ± 2.11	55.00 ± 2.00	57.99 ± 1.03	55.00 ± 1.32	58.00 ± 1.33	53.45 ± 2.96	56.75 ± 1.58
	CNN	**62.64** ± 3.02	54.94 ± 1.03	54.85 ± 2.43	57.93 ± 1.97	56.16 ± 1.56	56.20 ± 2.04	52.18 ± 2.71	**56.76** ± 1.66
	GRU	60.00 ± 3.42	**65.20** ± 2.26	55.00 ± 2.00	60.00 ± 1.55	54.93 ± 0.32	58.35 ± 2.11	49.95 ± 1.83	56.33 ± 1.47
	Trans.	59.87 ± 0.43	58.84 ± 1.23	**54.94** ± 2.34	**62.93** ± 2.09	**58.67** ± 2.61	**58.92** ± 2.20	**54.71** ± 2.37	59.02 ± 2.42

TABLE IX
SER PERFORMANCES IN AN AVERAGE OF CCC WITH SD FOR *AROUSAL* AND *VALENCE* WITH DIFFERENT SCENARIOS

Scenario		*MSP*				*BIIC*			
		Matched		Mismatched		Matched		Mismatched	
Task	Model	$M \rightarrow M$	$C \rightarrow C$	$M \rightarrow C$	$C \rightarrow M$	$M \rightarrow M$	$C \rightarrow C$	$M \rightarrow C$	$C \rightarrow M$
		CCC ± SD	CCC ± SD	CCC ± SD	CCC ± SD	CCC ± SD	CCC ± SD	CCC ± SD	CCC ± SD
Arousal	DNN	0.285 ± 0.17	0.343 ± 0.84	0.131 ± 1.09	**0.347** ± 0.70	0.439 ± 1.16	**0.613** ± 0.24	**0.449** ± 1.29	**0.545** ± 1.41
	CNN	0.274 ± 0.25	0.352 ± 0.72	**0.218** ± 1.97	0.335 ± 0.47	**0.464** ± 1.59	0.593 ± 0.23	0.414 ± 1.20	0.509 ± 1.44
	GRU	0.315 ± 2.02	0.162 ± 1.21	0.176 ± 3.24	0.250 ± 3.96	0.271 ± 1.60	0.369 ± 1.40	0.183 ± 1.55	0.297 ± 2.82
	Trans.	**0.319** ± 0.74	**0.383** ± 0.59	0.188 ± 1.08	0.338 ± 0.47	0.361 ± 0.95	0.510 ± 1.03	0.376 ± 1.13	0.456 ± 1.64
Valence	DNN	0.146 ± 0.76	0.343 ± 1.21	0.109 ± 2.02	0.134 ± 0.59	**0.232** ± 1.12	**0.345** ± 0.66	0.116 ± 1.30	0.323 ± 0.53
	CNN	0.097 ± 2.04	0.219 ± 0.50	0.128 ± 2.55	0.198 ± 1.37	0.199 ± 1.11	0.307 ± 0.50	0.170 ± 1.45	**0.411** ± 0.53
	GRU	0.195 ± 1.16	0.245 ± 3.13	0.193 ± 2.15	**0.281** ± 0.47	0.220 ± 1.68	0.281 ± 4.35	0.127 ± 2.24	0.241 ± 2.38
	Trans.	**0.282** ± 3.01	**0.365** ± 0.53	**0.245** ± 2.13	0.262 ± 0.85	0.152 ± 2.86	0.247 ± 1.52	**0.177** ± 1.98	0.228 ± 0.87

scenarios. In the "Matched" scenario, both the training and testing sets share the same *communication modes*, which are *Mono* ($\boldsymbol{M} \rightarrow \boldsymbol{M}$) and *Conv* ($\boldsymbol{C} \rightarrow \boldsymbol{C}$). Conversely, in the "Mismatched" scenario, the training and testing sets come from different *communication modes* ($\boldsymbol{M} \rightarrow \boldsymbol{C}, \boldsymbol{C} \rightarrow \boldsymbol{M}$). Each classifier is trained ten times with random initializations. To assess the SER performance of these models, the evaluation metrics used for the SER categorical binary classification and regression tasks are unweighted average recall (UAR) and concordance correlation coefficient (CCC) respectively. The standard deviation (SD) is employed to examine the consistency of performance across the ten models.

B. Experimental Results and Analyses

1) Experiment Performance Comparisons: Table VIII presents the emotion recognition performances for each emotional category and Table IX presents the emotion recognition performances for each emotional attribute. To highlight the comparison, Fig. 5 shows the differences in performance between the specified experiments. In this analysis, the objective is to evaluate both the accuracy and stability of using different *communication modes* as training sources for SER. First, we consider the best results per emotion from four models observed in Table VIII for the "Matched" scenario (i.e., compared between $\boldsymbol{M} \rightarrow \boldsymbol{M}$ and $\boldsymbol{C} \rightarrow \boldsymbol{C}$). We notice that *Conv* leads to higher UAR than *Mono* with performance gains of 6.43%, 1.96%, 2.54% and 2.56% for *Neutral*, *Happiness*, *Anger*, and *Sadness* in *MSP*, and performance gains of 0.73%, 0.63%, and 0.25% for *Happiness*, *Anger*, and *Sadness* in *BIIC*, respectively. These initial findings align with the results discussed in Section IV-A and IV-B, where sentences in the *Conv* have higher inter-annotator consistency and better alignment with V-A impressions and acoustic expressiveness suggested in emotion literature.

Next, we compare the mismatched condition results of using different communication modes as SER training sources: $\boldsymbol{M} \rightarrow \boldsymbol{M} \Rightarrow \boldsymbol{C} \rightarrow \boldsymbol{M}$ and $\boldsymbol{C} \rightarrow \boldsymbol{C} \Rightarrow \boldsymbol{M} \rightarrow \boldsymbol{C}$. First, we find that using the *Conv* as sources ($\boldsymbol{C} \rightarrow \boldsymbol{M}$) can produce more competitive UAR performance in comparison to the matched condition ($\boldsymbol{M} \rightarrow \boldsymbol{M}$), with only a slight decrease of 2.30% for *Happiness*, 1.60% for *Anger*, 0.001 for *Valence* in *MSP* and 6.60% for *Happiness* in *BIIC*. The rest even show a slight increase in performance. In contrast, the mismatched scenario where *Mono* is utilized as the source ($\boldsymbol{M} \rightarrow \boldsymbol{C}$) results in substantially larger performance gaps relative to the matched condition ($\boldsymbol{C} \rightarrow \boldsymbol{C}$). In this case, all UARs decrease significantly, as in *MSP*, declining by 5.54% for *Happiness*, 8.18% for *Anger*, 10.20% for *Sadness*, 0.165 for *Arousal* and 0.119 for *Valence*. Similar trend in *BIIC*, with

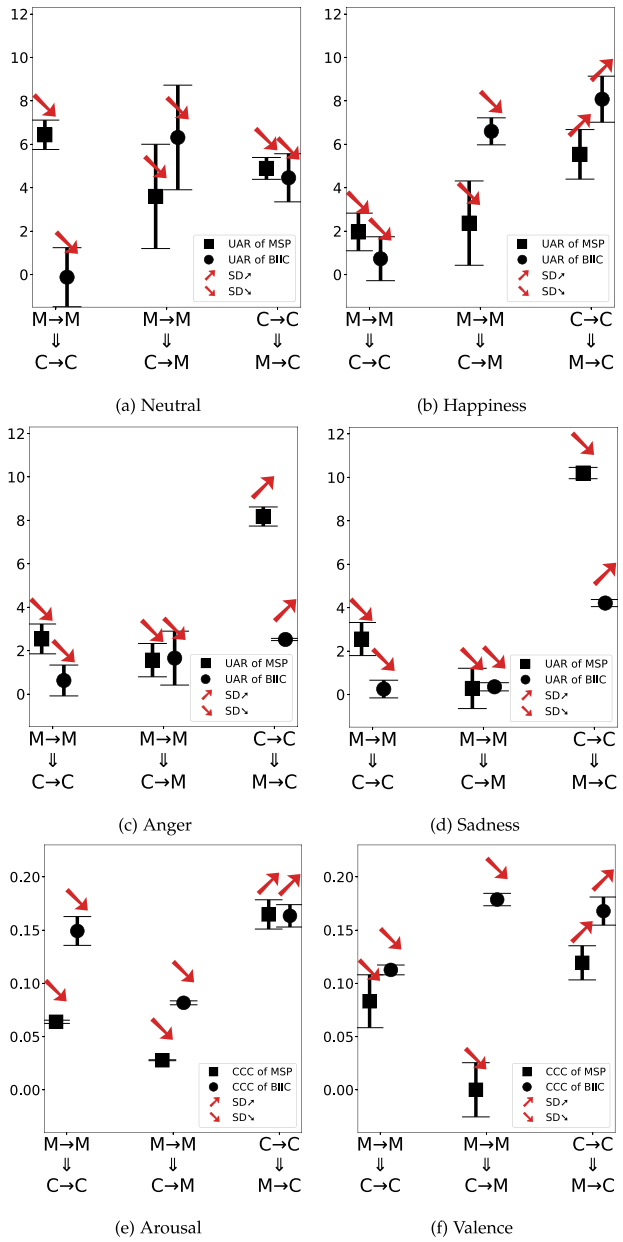


Fig. 5. Highlighting a comparison of different scenarios with performance in UAR(%) for each emotional category and performance in CCC for each emotion attributes. Red arrows indicate the trends of differences in SD under different scenarios.

a drop of 8.08% for *Happiness*, 2.52% for *Anger*, 4.21% for *Sadness*, 0.163 for *Arousal*, 0.168 for *Valence*. This outcome suggests that employing *Conv* as training data leads to a more robust SER capable of better handling mismatched conditions compared to using *Mono* as training data. Literature can further explain that the *Conv* sets exhibit greater inclusiveness, and as a result, using *Conv* as a source for training SER models can better fit to *Mono* samples [75].

Last but not least, we further evaluate the stability of the model. We highlight the trend using red arrows in Fig. 5, revealing that models trained using the *Conv* as the source exhibit less variability in performance compared to those trained with

TABLE X
SER PERFORMANCES IN AN AVERAGE OF UAR/CCC WITH SD FOR EMOTION CATEGORIES AND ATTRIBUTES WITH DIFFERENT SCENARIOS

		MSP		*BIIC*	
		Mono	***Conv***	***Mono***	***Conv***
Task	Fine-tune	UAR ± SD	UAR ± SD	UAR ± SD	UAR ± SD
Neutral	***Mono***	71.20 ± 0.17	68.65 ± 0.26	72.22 ± 0.22	78.38 ± 0.30
	Conv	73.90 ± 0.32	75.66 ± 0.20	71.24 ± 0.45	79.62 ± 0.16
Happiness	***Mono***	69.24 ± 0.29	68.05 ± 0.34	70.28 ± 0.19	68.24 ± 0.50
	Conv	67.26 ± 0.35	74.88 ± 0.30	70.25 ± 0.51	73.50 ± 0.37
Anger	***Mono***	70.16 ± 0.34	67.27 ± 0.34	68.54 ± 0.30	65.87 ± 0.60
	Conv	68.22 ± 0.27	67.26 ± 0.32	70.62 ± 0.46	66.20 ± 0.16
Sadness	***Mono***	68.69 ± 0.23	64.82 ± 0.34	64.62 ± 0.31	66.88 ± 0.40
	Conv	67.62 ± 0.15	72.10 ± 0.23	71.21 ± 0.21	73.82 ± 0.08
		CCC ± SD	CCC ± SD	CCC ± SD	CCC ± SD
Arousal	***Mono***	0.366 ± 0.33	0.346 ± 0.44	0.462 ± 0.35	0.378 ± 0.50
	Conv	0.374 ± 0.22	0.411 ± 0.30	0.446 ± 0.25	0.633 ± 0.20
Valence	***Mono***	0.285 ± 0.29	0.343 ± 0.44	0.131 ± 0.47	0.347 ± 0.37
	Conv	0.325 ± 0.32	0.383 ± 0.42	0.377 ± 0.40	0.436 ± 0.31

the *Mono*. Specifically in terms of numbers, employing *Conv* instead of *Mono* as the training data in the matched condition reduces the SD by 0.68% for *Neutral*, 0.87% for *Happiness*, 0.69% for *Anger*, 0.76% for *Sadness*, 1.35% for *Arousal* and 0.46% for *Valence* in *MSP*. Similar in *BIIC*, a decrease of SD by 1.36% for *Neutral*, 1.01% for *Happiness*, 0.71% for *Anger*, 0.41% for *Sadness*, 0.15% for *Arousal* and 2.48% for *Valence*. Moreover, we observe similar trends under mismatched conditions, where the standard deviations all decrease, depicting that the SER system has been trained in a more stable manner. Whereas using *Mono* as the training data under the mismatched condition tends to result in less stable and more fluctuated models. Generally, the analyses and SER results from this experiment suggest that utilizing sentences from the *Conv* set leads to SER models with enhanced robustness and consistency. The more consistent agreement among annotators' ratings, improved consistency in acoustic expressions, and potentially better-behaved emotion manifestations position *Conv* as a "better-quality" source for constructing SER databases. Most importantly, our findings exhibit consistent results across both datasets, demonstrating that the conversational communication pattern is a "better-quality" source regardless of the language involved.

2) Performance Evaluation of Fine-Tuned SSLMs: In addition to evaluating the base models, we employ advanced fine-tuning techniques on pre-trained models to rigorously test our research hypothesis. In this analysis, we aim to evaluate the accuracy and stability of using different *communication modes* as sources to fine-tune the pre-trained model for SER. Table X presents the emotion recognition performances of these fine-tuned models (fine-tuned with *Mono* or *Conv* samples), detailing the results across each emotional category and attribute.

At the outset, it is clear that while the SD does not exhibit the same level of variability as the base models shown in Section V-B1, the models fine-tuned with *Conv* data as the source (highlighted with a gray background) manifest significantly greater stability. Furthermore, upon examining the performance metrics, it is evident that the models fine-tuned with *Conv* data consistently outperform those fine-tuned with *Mono* data. The recognition performance for *Conv* as the source across both *MSP* and *BIIC* datasets demonstrates superiority over the *Mono*-sourced models. This set of experiment reinforces our research

TABLE XI
SER PERFORMANCES IN AN AVERAGE OF UAR/CCC WITH SD WITH DIFFERENT SCENARIOS

		MSP		*BIIC*	
		Mono	***Conv***	***Mono***	***Conv***
Task	Model	UAR ± SD	UAR ± SD	UAR ± SD	UAR ± SD
Neutral	DNN	62.24 ± 0.27	63.65 ± 0.40	71.28 ± 0.42	74.38 ± 0.19
	CNN	73.00 ± 0.27	71.35 ± 0.36	73.22 ± 0.50	77.30 ± 0.33
	GRU	68.22 ± 0.77	78.65 ± 0.56	74.22 ± 0.62	76.38 ± 0.40
	Trans.	75.20 ± 0.17	73.65 ± 0.66	74.64 ± 0.32	74.38 ± 0.50
Happiness	DNN	66.26 ± 0.45	68.65 ± 0.37	77.11 ± 0.39	73.38 ± 0.44
	CNN	64.20 ± 0.77	62.30 ± 0.64	67.82 ± 0.46	78.38 ± 0.50
	GRU	66.20 ± 0.34	68.60 ± 0.18	70.22 ± 0.42	76.28 ± 0.30
	Trans.	66.28 ± 0.17	65.37 ± 0.37	68.02 ± 0.32	68.62 ± 0.40
Anger	DNN	66.00 ± 0.36	67.31 ± 0.27	68.12 ± 0.19	68.38 ± 0.42
	CNN	66.44 ± 0.37	68.65 ± 0.36	70.44 ± 0.29	68.38 ± 0.40
	GRU	62.46 ± 0.37	67.65 ± 0.36	68.12 ± 0.52	67.48 ± 0.43
	Trans.	70.20 ± 0.27	73.65 ± 0.31	70.12 ± 0.62	69.06 ± 0.25
Sadness	DNN	67.10 ± 0.23	68.34 ± 0.46	73.12 ± 0.28	77.80 ± 0.36
	CNN	60.90 ± 0.15	67.67 ± 0.32	78.42 ± 0.52	71.28 ± 0.33
	GRU	62.80 ± 0.37	63.45 ± 0.27	73.52 ± 0.40	76.88 ± 0.22
	Trans.	68.20 ± 0.17	68.52 ± 0.26	67.37 ± 0.38	70.82 ± 0.21
		CCC ± SD	CCC ± SD	CCC ± SD	CCC ± SD
Arousal	DNN	0.306 ± 0.35	0.326 ± 0.54	0.360 ± 0.40	0.388 ± 0.48
	CNN	0.342 ± 0.33	0.346 ± 0.54	0.315 ± 0.75	0.338 ± 0.60
	GRU	0.266 ± 0.34	0.316 ± 0.24	0.402 ± 0.30	0.408 ± 0.37
	Trans.	0.406 ± 0.39	0.378 ± 0.34	0.431 ± 0.23	0.388 ± 0.33
Valence	DNN	0.266 ± 0.23	0.226 ± 0.54	0.196 ± 0.15	0.288 ± 0.37
	CNN	0.266 ± 0.33	0.246 ± 0.46	0.246 ± 0.52	0.321 ± 0.60
	GRU	0.288 ± 0.13	0.271 ± 0.16	0.380 ± 0.33	0.349 ± 0.30
	Trans.	0.290 ± 0.33	0.284 ± 0.71	0.222 ± 0.36	0.298 ± 0.40

We combine *Conv* and *Mono* as the training source.

hypothesis that conversation data serves as a better source for SER.

3) Results of Multi-Sources Training: In traditional SER database construction, it is a common practice to combine various data sources to train models, utilizing the collective diversity of the dataset to enhance recognition capabilities. To align our research with these standard practices and provide a comprehensive analysis, this section delves into the outcomes of training with multiple data sources. We train the base models with both *Conv* and *Mono* sets and the results are summarized in Table XI.

In these results, our focus is on assessing the robustness of models trained using multiple sources. When cross-compared to the results of prior experiments (Section V-B1), we observe that while the recognition performance is relatively higher, the SD when tested on *Conv* tends to be lower than when tested on *Mono*. This could suggest that the inclusion of *Mono* as a training source may have somewhat compromised the adaptability of models. From this analysis, we not only provide a comparative benchmark but also reinforce the notion that utilizing *Conv* as the training source for SER models leads to enhanced stability.

VI. DISCUSSION AND CONCLUSION

In this study, our objective was to systematically examine the influence of *communication modes*, specifically conversation and monologue speech, on the collection of emotional data for building SER models. On the basis of our previous work, we deeper extend these analyses to multi-lingual corpora, including the MSP-Podcast and BIIC-Podcast, to evaluate the hypothesis that the conversation mode offers a "better-quality" medium for creating SER databases compared to the monologue mode. We investigated this hypothesis by examining rater perceptions and acoustic expressiveness. Our findings reveal several interesting insights: 1) conversation samples exhibit a higher consensus rate among annotators, as well as a higher percentage of samples occupying the expected Valence-Arousal (V-A) space based on their categorical emotion. Conversation samples are more concentrated and display higher inter-annotator agreement compared to the more widely dispersed monologue samples in the V-A space; 2) we found that differences exist not only in rater perception but also in acoustic cues, with conversation samples demonstrating more consistent patterns that align with the emotion-specific modulation and emotion attributes of acoustic patterns documented in the existing literature. Our SER model learning derived from experiments conducted from multiple perspectives indicates that utilizing speech samples from conversations contributes to a more robust and less variable SER model, collectively serving to enhance our research hypothesis.

As a result, these analyses suggest that the conversation mode of communication might serve as a "better-quality" medium for creating SER databases, potentially leading to more stable and reliable SER models. The extensibility of this work is broad, with several crucial future directions to further enhance our understanding of the impact of monologues and conversations on SER systems. An intuitive one is to conduct a more fine-grained analysis of specific emotions and their manifestation in conversation and monologue speech to better understand the nuances of emotional expression in different communication modes. Integrating linguistic modalities into our examination of emotional expressiveness will be a vital next step in this research.

REFERENCES

[1] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[2] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.

[3] R. Jürgens, K. Hammerschmidt, and J. Fischer, "Authentic and play-acted vocal emotion expressions reveal acoustic differences," *Front. Psychol.*, vol. 2, 2011, Art. no. 180.

[4] C.-C. Lee, C. Busso, S. Lee, and S. S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 1983–1986.

[5] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digit. Signal Process.*, vol. 22, no. 6, pp. 1154–1160, 2012.

[6] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the geneva multimodal expression corpus for experimental research on emotion perception," *Emotion*, vol. 12, no. 5, 2012, Art. no. 1161.

[7] 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, First Quarter, 2017.

[8] F. Burkhardt et al., "A database of german emotional speech," in *Proc. Interspeech*, 2005, pp. 1517–1520.

[9] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *Proc. 2008 IEEE Int. Conf. Multimedia Expo*, 2008, pp. 865–868.

[10] 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, Fourth Quarter, 2019.

[11] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *Proc. Interspeech*, 2020, pp. 1823–1827.

[12] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, 2021, Art. no. 1249.

[13] R. Böck, O. Egorow, I. Siegert, and A. Wendemuth, "Comparative study on normalisation in emotion recognition from speech," in *Proc. Int. Conf. Intell. Hum. Comput. Interaction*, Springer, 2017, pp. 189–201.

[14] F. Chenchah and Z. Lachiri, "Speech emotion recognition in acted and spontaneous context," *Procedia Comput. Sci.*, vol. 39, pp. 139–145, 2014.

[15] S. Gustafson-Capková, "Emotions in speech: Tagset and acoustic correlates," *Speech Technol.*, pp. 1–13, 2001.

[16] W.-S. Chien et al., "Monologue versus conversation: Differences in emotion perception and acoustic expressivity," in *Proc. IEEE 10th Int. Conf. Affect. Comput. Intell. Interact.*, 2022, pp. 1–7.

[17] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behav. Brain Sci.*, vol. 27, no. 2, pp. 169–190, 2004.

[18] N. A. Roberts, J. L. Tsai, and J. A. Coan, "Emotion elicitation using dyadic interaction tasks," in *Handbook of Emotion Elicitation and Assessment*. London, U.K.: Oxford Univ. Press, 2007, pp. 106–123.

[19] D. Keltner and J. Haidt, "Social functions of emotions at four levels of analysis," *Cogn. Emotion*, vol. 13, no. 5, pp. 505–521, 1999.

[20] H. P. Branigan, C. M. Catchpole, and M. J. Pickering, "What makes dialogues easy to understand?," *Lang. Cogn. Processes*, vol. 26, no. 10, pp. 1667–1686, 2011.

[21] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 183–196, Second Quarter, 2013.

[22] J. K. Burgoon, D. B. Buller, L. Dillman, and J. B. Walther, "Interpersonal deception: Iv. effects of suspicion on perceived communication and nonverbal behavior dynamics," *Hum. Commun. Res.*, vol. 22, no. 2, pp. 163–196, 1995.

[23] R. W. Levenson and J. M. Gottman, "Marital interaction: Physiological linkage and affective exchange," *J. Pers. Social Psychol.*, vol. 45, no. 3, 1983, Art. no. 587.

[24] A. Alghamdi, A. C. Karpinski, A. Lepp, and J. Barkley, "Online and face-to-face classroom multitasking and academic performance: Moderated mediation with self-efficacy for self-regulated learning and gender," *Comput. Hum. Behav.*, vol. 102, pp. 214–222, 2020.

[25] C. Hardy, T. Lawrence, and N. Phillips, "Talking action: Conversations, narrative and action in interorganizational collaboration," *Discourse Org.*, vol. 65, 1998, Art. no. 83.

[26] R. J. Lewicki, B. Barry, and D. M. Saunders, *Negotiation*, 7th ed. New York, NY, USA: McGraw-Hill, 2015.

[27] J. Vettin and D. Todt, "Laughter in conversation: Features of occurrence and acoustic structure," *J. Nonverbal Behav.*, vol. 28, no. 2, pp. 93–115, 2004.

[28] X. Zhu and G. Penn, "Comparing the roles of textual, acoustic and spoken-language features on spontaneous-conversation summarization," in *Proc. Hum. Lang. Technol. Conf. NAACL*, 2006, pp. 197–200.

[29] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Commun.*, vol. 53, no. 1, pp. 36–50, 2011.

[30] B. Ludusan, R. Mazuka, M. Bernard, A. Cristia, and E. Dupoux, "The role of prosody and speech register in word segmentation: A computational modelling perspective," in *Proc. 55th Annu. Meeting Assoc. for Comput. Linguistics*, 2017, pp. 178–183.

[31] B. Campos, D. Schoebi, G. C. Gonzaga, S. L. Gable, and D. Keltner, "Attuned to the positive? Awareness and responsiveness to others' positive emotion experience and display," *Motivation Emotion*, vol. 39, no. 5, pp. 780–794, 2015.

[32] N. E. Joby and H. Umemuro, "Effect of group identity on emotional contagion in dyadic human agent interaction," in *Proc. 10th Int. Conf. Hum.-Agent Interaction*, 2022, pp. 157–166.

[33] S. G. Upadhyay et al., "An intelligent infrastructure toward large scale naturalistic affective speech corpora collection," in *Proc. IEEE 11th Int. Conf. Affect. Comput. Intell. Interact.*, 2023, pp. 1–8.

[34] K. R. Scherer, T. Johnstone, and G. Klasmeyer, *Vocal Expression of Emotion*. London, U.K.: Oxford Univ. Press, 2003.

[35] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, no. 1–2, pp. 227–256, 2003.

[36] M. Schröder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. Gielen, "Acoustic correlates of emotion dimensions in view of speech synthesis," in *Proc. 7th Eur. Conf. Speech Commun. Technol.*, 2001, pp. 87–90.

[37] T. Bänziger and K. R. Scherer, "The role of intonation in emotional expressions," *Speech Commun.*, vol. 46, no. 3–4, pp. 252–267, 2005.

[38] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47 795–47 814, 2021.

[39] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Commun.*, vol. 48, no. 9, pp. 1162–1181, 2006.

[40] R. Nakatsu, J. Nicholson, and N. Tosa, "Emotion recognition and its application to computer agents with spontaneous interactive capabilities," in *Proc. 7th ACM Int. Conf. Multimedia*, 1999, pp. 343–351.

[41] R. Cowie et al., "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[42] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 186–202, 2015.

[43] K. Mangalam and T. Guha, "Learning spontaneity to improve emotion recognition in speech," 2017, *arXiv: 1712.04753*.

[44] K. R. Scherer, A. Schorr, and T. Johnstone, *Appraisal Processes in Emotion: Theory, Methods, Research*. London, U.K.: Oxford Univ. Press, 2001.

[45] J. A. Russell and G. Pratt, "A description of the affective quality attributed to environments," *J. Pers. Social Psychol.*, vol. 38, no. 2, 1980, Art. no. 311.

[46] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?," *Psychol. Bull.*, vol. 129, no. 5, 2003, Art. no. 770.

[47] K. R. Scherer and P. H. Tannenbaum, "Emotional experiences in everyday life: A survey approach," *Motivation Emotion*, vol. 10, pp. 295–314, 1986.

[48] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Commun.*, vol. 40, no. 1–2, pp. 5–32, 2003.

[49] P. Laukka, P. Juslin, and R. Bresin, "A dimensional approach to vocal expression of emotion," *Cogn. Emotion*, vol. 19, no. 5, pp. 633–653, 2005.

[50] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *J. Pers. Social Psychol.*, vol. 70, no. 3, 1996, Art. no. 614.

[51] K. R. Scherer, "Vocal affect expression: A review and a model for future research," *Psychol. Bull.*, vol. 99, no. 2, 1986, Art. no. 143.

[52] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2697–2709, 2020.

[53] A. Triantafyllopoulos and B. W. Schuller, "The role of task and acoustic similarity in audio transfer learning: Insights from the speech emotion recognition case," in *Proc. 2021 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7268–7272.

[54] R. Pappagari, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak, "Copypaste: An augmentation method for speech emotion recognition," in *Proc. 2021 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 6324–6328.

[55] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Proc. Interspeech*, Singapore, 2014, pp. 238–242.

[56] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 374–388, Fourth Quarter, 2016.

[57] S. A. Salskov, J. Backström, and K. Creutz, "From angry monologues to engaged dialogue? On self-reflexivity, critical discursive psychology and studying polarised conflict," in *The Far-Right Discourse of Multiculturalism in Intergroup Interactions*. Berlin, Germany: Springer, 2022, pp. 163–187.

[58] D. T. Nguyen and S. R. Fussell, "Effects of conversational involvement cues on understanding and emotions in instant messaging conversations," *J. Lang. Social Psychol.*, vol. 35, no. 1, pp. 28–55, 2016.

[59] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, 1971, Art. no. 378.

[60] J. R. Landis and G. G. Koch, "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers," *Biometrics*, vol. 33, pp. 363–374, 1977.

[61] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam, Netherlands, Tech. Rep. 132, 1996. [Online]. Available: http://www.praat.org

[62] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Proc. Interspeech*, 2017, pp. 498–502.

[63] Y.-F. Liao et al., "Formosa speech recognition challenge 2018: Data, plan and baselines," in *Proc. IEEE 11th Int. Symp. Chin. Spoken Lang. Process.*, 2018, pp. 270–274.

[64] S. G. Upadhyay et al., "Phonetic anchor-based transfer learning to facilitate unsupervised cross-lingual speech emotion recognition," in *Proc. 2023 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.
[65] K. Hammerschmidt and U. Jürgens, "Acoustical correlates of affective prosody," *J. Voice*, vol. 21, no. 5, pp. 531–540, 2007.
[66] A. Weisser, J. M. Buchholz, and G. Keidser, "Complex acoustic environments: Review, framework, and subjective model," *Trends Hear.*, vol. 23, 2019, Art. no. 2331216519881346.
[67] A. Weise, S. I. Levitan, J. Hirschberg, and R. Levitan, "Individual differences in acoustic-prosodic entrainment in spoken dialogue," *Speech Commun.*, vol. 115, pp. 78–87, 2019.
[68] I. Gessinger, "Phonetic accommodation of human interlocutors in the context of human-computer interaction," Ph.D. dissertation, 2022.
[69] A. Baevski, S. Schneider, and M. Auli, "VQ-WAV2Vec: Self-supervised learning of discrete speech representations," 2019, *arXiv: 1910.05453*.
[70] T. Liu, S. Fang, Y. Zhao, P. Wang, and J. Zhang, "Implementation of training convolutional neural networks," 2015, *arXiv:1506.01195*.
[71] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS 2014 Workshop Deep Learn.*, 2014.
[72] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[73] J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10745–10759, Sep. 2023.
[74] S.-W. Yang et al., "Superb: Speech processing universal performance benchmark," 2021, *arXiv:2105.01051*.
[75] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

Woan-Shiuan Chien (Student Member, IEEE) received the BS degree in electrical engineering from Chung Yuan Christian University, Taiwan in 2015 and the MS degree in electrical engineering from the National Chung Cheng University (CCU), Taiwan in 2016. She is currently working toward the PhD degree with the Electrical Engineering (EE) Department of National Tsing Hua University (NTHU), Taiwan. Her research interests are in human-centered behavioral signal processing and automatic speech emotion recognition. She was the recipient of the Outstanding Doctoral Students Program sponsored by the Taiwan Science and Technology Council (NSTC) (2022). She is a student member of the AAAC, ACM, ACLCLP, and IEEE Signal Processing Society (SPS).

Shreya G. Upadhyay (Student Member, IEEE) received the BE degree in computer engineering from Mumbai University, India in 2013, and the MTech degree in computer engineering from the K. J. Somaiya College of Engineering, India in 2018. She is currently working toward the PhD degree in electrical engineering with National Tsing Hua University (NTHU), Taiwan. Her research interests include behavioral speech signal processing, speech emotion recognition, automatic speech recognition, and acoustic sound event detection. She is a student member of International Speech Communication Association (ISCA), AAAC, European Association for Signal Processing (EURASIP), and the IEEE Signal Processing Society (SPS).

Wei-Cheng Lin (Member, IEEE) received the BS degree in communication engineering from the National Taiwan Ocean University (NTOU), Taiwan in 2014 and the MS degree in electrical engineering from the National Tsing Hua University (NTHU), Taiwan in 2016. She is currently working toward the PhD degree with the Electrical and Computer Engineering Department of The University of Texas at Dallas (UTD). His research interests are in human-centered behavioral signal processing (BSP), deep learning, and multimodal/speech signal processing. He is also a student member of the IEEE Signal Processing Society (SPS) and International Speech Communication Association (ISCA).

Carlos Busso (Fellow, IEEE) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is a professor at the Language Technologies Institute, Carnegie Mellon University, where he is also the director of the Multimodal Speech Processing (MSP) Laboratory. His research interest is in human-centered multimodal machine intelligence and application, focusing on the broad areas of speech processing, affective computing, and machine learning methods for multimodal processing. He has worked on speech emotion recognition, multimodal behavior modeling for socially interactive agents, in-vehicle active safety systems, and robust multimodal speech processing. He was selected by the School of Engineering of Chile as the best electrical engineer who graduated in 2003 from Chilean universities. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. His students received the third prize IEEE ITSS Best Dissertation Award (N. Li) in 2015, and the AAAC Student Dissertation Award (W.-C. Lin) in 2024. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He received the Best of IEEE Transactions on Affective Computing Paper Collection in 2021 (with R. Lotfian) and the Best Paper Award from IEEE Transactions on Affective Computing in 2022 (with Yannakakis and Cowie). In 2023, he received the Distinguished Alumni Award in the Mid-Career/Academia category by the Signal and Image Processing Institute (SIPI) at the University of Southern California. He received the 2023 ACM ICMI Community Service Award. He is a member of AAAC and a senior member of ACM. He is an ISCA fellow.

Chi-Chun Lee (Senior Member, IEEE) received the BS and PhD degrees in electrical engineering from the University of Southern California, USA in 2007 and 2012, respectively. He is an professor with the Department of Electrical Engineering of the National Tsing Hua University (NTHU), Taiwan. His research interests are in speech and language, affective computing, health analytics, and behavioral signal processing. He is an associate editor for *IEEE Transaction on Affective Computing* (2020-), *IEEE Transaction on Multimedia* (2019-2020), *Journal of Computer Speech and Language* (2021-), *APSIPA Transactions on Signal and Information Processing* and a TPC member for APSIPA IVM and MLDA committee. He serves as the general chair for ASRU 2023, an area chair for Interspeech 2016, 2018, 2019, senior program committee for ACII 2017, 2019, publicity chair for ACM ICMI 2018, late breaking result chair for ACM ICMI 2023, sponsorship and special session chair for ISCSLP 2018, 2020. He is the recipient of the Foundation of Outstanding Scholar's Young Innovator Award (2020), the CIEE Outstanding Young Electrical Engineer Award (2020), the IICM K. T. Li Young Researcher Award (2020), the NTHU Industry Collaboration Excellence Award (2021), and the MOST Futuretek Breakthrough Award (2018, 2019). He led a team to the 1st place in Emotion Challenge in Interspeech 2009, and with his students won the 1st place in Styrian Dialect and Baby Sound subchallenge in Interspeech 2019. He is a coauthor on the best paper award/finalist in Interspeech 2008, Interspeech 2010, IEEE EMBC 2018, Interspeech 2018, IEEE EMBC 2019, APSIPA ASC 2019, IEEE EMBC 2020, and the most cited paper published in 2013 in Journal of Speech Communication. He is also an ACM and ISCA member.