Monologue versus Conversation: Differences in Emotion Perception and Acoustic Expressivity

Woan-Shiuan Chien[1], Shreya G. Upadhyay[1], Wei-Cheng Lin[2], Ya-Tse Wu[1], Bo-Hao Su[1], Carlos Busso[2], Chi-Chun Lee[1]

[1]*Department of Electrical Engineering, National Tsing Hua University, Taiwan*

[2]*Department of Electrical and Computer Engineering, University of Texas at Dallas, USA*

[1]{wschien, shreya, crowpeter, borrissu}@gapp.nthu.edu.tw,

[2]{wei-cheng.lin, busso}@utdallas.edu, [1]cclee@ee.nthu.edu.tw

2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII) | 978-1-6654-5908-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ACII55700.2022.9953814

***Abstract*—Advancing *speech emotion recognition* (SER) depends highly on the source used to train the model, i.e., the emotional speech corpora. By permuting different design parameters, researchers have released versions of corpora that attempt to provide a better-quality source for training SER. In this work, we focus on studying *communication modes* of collection. In particular, we analyze the patterns of emotional speech collected during interpersonal conversations or monologues. While it is well known that conversation provides a better protocol for eliciting authentic emotion expressions, there is a lack of systematic analyses to determine whether conversational speech provide a "better-quality" source. Specifically, we examine this research question from three perspectives: perceptual differences, acoustic variability and SER model learning. Our analyses on the MSP-Podcast corpus show that: 1) rater's consistency for conversation recordings is higher when evaluating categorical emotions, 2) the perceptions and acoustic patterns observed on conversations have properties that are better aligned with expected trends discussed in emotion literature, and 3) a more robust SER model can be trained from conversational data. This work brings initial evidences stating that samples of conversations may provide a better-quality source than samples from monologues for building a SER model.**

***Index Terms*—speech emotion recognition, emotion perception, acoustic expression, conversation, monologue**

I. Introduction

There is an increased interest in developing *speech emotion recognition* (SER) systems for everyday life applications. Most SER systems are data-driven, so the quality of the speech emotional corpora is crucial to build better systems. When collecting a database for SER, key design parameters have a profound impact on its quality, including the settings (monologue, dyad, and small group), the elicitation methods (read, improvisation and script), and the emotional descriptors used to annotated the corpus (categorical emotion, emotional attributes). As an example, the well-known IEMOCAP database [1] is setup in dyadic conversation with scripted and spontaneous interactions, and annotated with emotion categories and with three major emotional attributes (valence, arousal and dominance). These design parameters directly affect the actual speech collected in the data. For instance, the voice quality and F0-contour would differ in recordings collected in either prompted or unprompted settings [2], and this difference would naturally affect the emotional perception of raters. Understanding the best setting for collecting emotional databases is important to increase the quality and usability of the corpora.

Researchers have permuted several of these key design parameters to create diverse and distinct speech emotional corpora. Over the past years, the collections has moved from in-lab settings (e.g., IEMOCAP [1], MSP-IMPROV [3], EMODB [4]) to in-the-wild scenarios (e.g., VAM [5], MSP-Podcast [6], MSP-Conversation [7]), where the goal is to have databases that better resemble emotional expressions observed during daily interactions. While there is a general hypothesis on how these design parameters would affect the intended quality of these databases when being used to train a SER system, there is only a handful of studies exploring the impact of data collection designs on SER models. For example, we know that simulated datasets tend to result in an overfitted model [8]. Likewise, corpora containing mainly acted samples would favor using a certain normalization scheme due to a higher expressivity in the emotional content [9]. Gustafson-Capková [10] demonstrated that sad sentences are easier to recognize in acted databases than in spontaneous databases. This work aims to analyze the *modes of communication* used to record emotional databases. In particular, we focus on monologue versus conversation speech, systematically comparing their perceptual differences, acoustic variability, and influence on constructing an SER model.

Monologue and conversation have important differences. Psychology studies indicate that the emergence of most emotional reactions depends on the types of interpersonal interactions or social consequences [11]–[14]. The emotions of one person usually trigger the reactions from her/his interlocutor, forming a tight connection between the emotions of the people in the conversation [15]. Furthermore, speaking face-to-face provides a better setting for humans to naturally convey and perceive emotion, as demonstrated in marital relationship [16], classmates [17], parent-child interaction [18] and inter-organization interaction [19]. Moreover, from a speech production viewpoint, acoustic properties of a speaker are known to be different when she/he is participating in a conversation than when she/he is simply speaking to her/his own [20]–[22]. Conversations can stimulate more expressively-rich interactions. There are even evidences that a positive emotion emerges more frequently during conversations [23], where

978-1-6654-5908-2/22/$31.00 ©2022 IEEE

TABLE I: Statistics of monologue and conversation datasets.

		Mono	Conv
Segments	Overall	2415	2797
	Neutral	60.2%	51.3%
	Happiness	23.1%	28.4%
	Anger	7.0%	13.5%
	Sadness	9.6%	6.8%
Duration	Total	3.15 hr	4.09 hr
	Average	5.29 sec	5.26 sec
Average of (segments / Podcast)		50.31	58.27
Average of (speaker nums / Podcast)		1	3.17

"Laughters" have longer interval duration [20]. These studies suggest that collecting emotional data in conversations may evoke more authentic and representative emotional behaviors than collecting data with monologues, resulting in "better-quality" source for training SER systems.

While there exists substantial evidences showing that conversation provides a better protocol for eliciting authentic emotion versus monologue, none of these studies investigate this research question from the viewpoint of a database for SER. In this work, we examine our running hypothesis that conversation is a "better-quality" mode for SER from three different perspectives: perceptual differences, acoustic variability, and SER model learning. We conduct this analysis on the MSP-Podcast corpus [6], which is a large scaled naturalistic emotional database. Our findings suggest that: 1) categorical emotions are rated with higher rater consistency in conversations, where anger and sadness occupy a narrower region in the Valence-Arousal (V-A) perceptual plane, 2) the patterns for the spread on the V-A perceptual plane and acoustic characteristics in conversation are more aligned with expected trends reported in the emotion literature, and 3) recordings from conversations provide a more robust source of information for training a SER model.

II. Database

A. MSP-Podcast database

In this work, we use the release 1.8 of the MSP-Podcast corpus [6], which has a total of 113 hours of emotional speech. The recordings are obtained from podcasts available on audio-sharing websites. This database is becoming popular for SER related research [24]–[26] due to its scale and availability of emotionally balanced dialogues from many speakers. The database is naturalistic with diverse content including discussions about politics, movie review, science, technology, and economics. The collection of this corpus builds on the retrieval-based approach described in Mariooryad et al. [27]. The duration of each segment is between 2.75s and 11s. The emotional annotations are obtained using a modified version of the crowd-sourced protocol proposed by Burmania et al. [28]. Each segment is annotated by at least five workers with the primary emotions (e.g., the most dominant emotion perceived in the audio), secondary emotions (all emotional classes perceived in the recordings), and emotional attributes. This study uses primary emotions (anger, sadness, happiness, surprise, fear, disgust, contempt, neutral), and emotional attributes (arousal, valence and dominance). The attributes are annotated using a 7-Likert scale. The consensus labels are obtained with the plurality rule for primary emotions, and the averaged dimensional ratings for emotional attributes.

B. Monologue and Conversation Splits in the Corpus

Our goal is to understand the differences between emotional speech samples collected using two *modes of communication*: monologue and conversation. There is a total of 48 podcasts where all the speaking turns within the podcast are assigned to a single speaker. We consider all the speaking turns from these podcasts as representative examples of monologues, which we refer to as *Mono* in the rest of the paper. To obtain representative samples for conversations, we randomly select an equivalent number of podcasts having speaking turns assigned to two or more speakers. We refer to this set as *Conv* in our analysis. Table I shows key statistics about the speech samples belonging to the *Mono* and *Conv* sets considered in this study. The table shows that there are 2,415 speaking turns for *Mono*, and 2,797 speaking turns for *Conv*. Similar to most conventional SER model learning studies, we first focus our study on samples with emotional labels belonging to the four categorical emotion labels: *Neutral*, *Happiness*, *Anger* and *Sadness*.

III. Empirical Analyses and Results

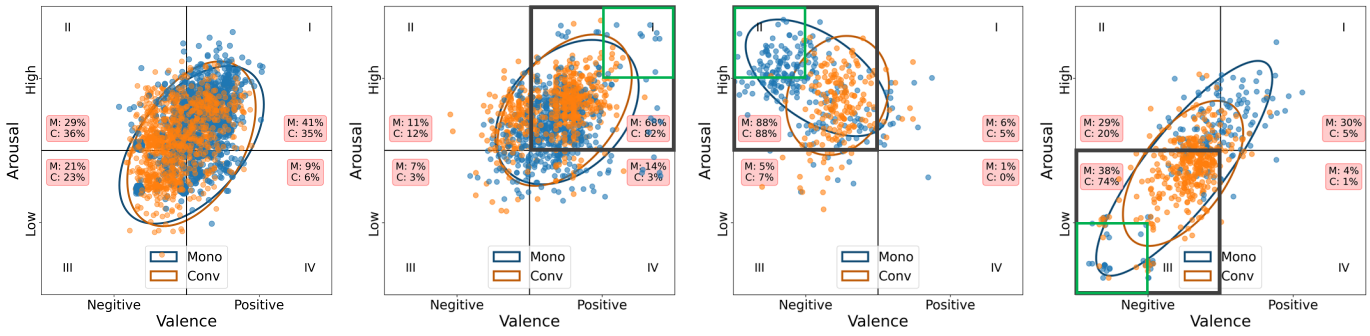
To evaluate the differences between speech recordings collected during conversation and monologues, we conduct analyses focusing on three different perspectives: perceptual differences (Sec. III-A), acoustic variability (Sec. III-B) and SER model learning (Sec. III-C).

A. Perceptual Differences

We investigate the emotional perception of categorical emotions in the valence-arousal (V-A) perceptual plane for sentences in the *Mono* and *Conv* sets. We also analyze the inter-annotator agreement for these sets.

1) Categorical Labels in the Valence-Arousal Plane: We first scatter the samples for *Neutral*, *Happiness*, *Anger* and *Sadness* on the V-A plane to visualize the perceptual differences between sentences in the *Mono* and *Conv* sets. Figure 1 shows four separate plots, one for each emotion. The blue-colored dots and the orange-colored dots are samples from the *Mono* and *Conv* sets, respectively. The ellipsoid regions are drawn such that they cover 80% of the samples. Psychology literature indicates that emotional categories are expected to be located in specific quadrants of the V-A perceptual plane [29]: *Happiness* in quadrant I, *Anger* in quadrant II and *Sadness* in quadrant III. Based on Figure 1, we ask the following questions by comparing samples in the *Mono* and *Conversation* sets for each emotion category:

1) Is there a difference between sets in terms of the location of the centroids of the ellipses? While the distributions for *Happiness* and *Neutral* are similar, we visually observe a noticeable difference in the distribution between *Mono* and



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

Fig. 1: Scatter plot for the categorical samples from *Mono* and *Conv* in valence-arousal (V-A) plane; each panel corresponds to a different categorical emotion and each quadrant also shows the occupancy rate for *Mono* (*M*) and *Conv* (*C*).

TABLE II: Statistical differences between *Mono* and *Conv*. "Mean ± SD" shows the statistics (mean and standard deviation) over the corresponding emotion-specific set. "p-value" shows the statistic differences between *Mono* and *Conv* by using two-tailed Mann-Whitney U rank test.

		Valence		Arousal	
		Mean ± SD	p-value	Mean ± SD	p-value
Neutral	*Mono*	4.12 ± 0.68	0.5566	4.13 ± 0.75	0.2361
	Conv	4.11 ± 0.61		4.15 ± 0.79	
Happiness	*Mono*	4.85 ± 0.68	0.9152	4.88 ± 0.69	0.0194
	Conv	4.80 ± 0.64		4.92 ± 0.72	
Anger	*Mono*	2.61 ± 0.90	**0.0007**	5.66 ± 0.74	**0.0001**
	Conv	3.08 ± 0.65		5.26 ± 0.76	
Sadness	*Mono*	3.18 ± 1.11	**0.0057**	3.59 ± 1.37	**0.0007**
	Conv	3.15 ± 0.76		3.48 ± 0.92	

Conv for *Anger* (Fig. 1c) and *Sadness* (Fig. 1d). To understand this difference, we first conduct a statistical testing using a two-tailed Mann-Whitney U rank test between samples in the *Mono* and *Conv* sets. We separately evaluate the scores for valence and arousal. Table II summarizes the results, which show that *Mono* and *Conv* have significant differences in the mean levels for arousal and valence (p-value ≤ 0.01 for both *Sadness* and *Anger*).

2) Is there a difference in the spread of the scattered samples on the V-A plane? We observe that the spread of the ellipse for *Sadness* and *Anger* in *Mono* is wider and scattered compared to *Conv*, which is much narrower and orderly; *Neutral* and *Happiness* has no such difference. We further examine samples with extreme emotional values in their respective quadrant. For *Anger*, this area is defined as the occupancy rate in the area with valence ≤ 2.5 and arousal ≥ 2.5 (green box in Fig. 1c). For *Sadness*, this area is defined as the occupancy rate in the area with valence ≤ 2.5 and arousal ≤ 2.5 (green box in Fig. 1d). A significant number of samples from the *Mono* set are in these green boxes with more extreme values in the V-A plane, i.e., 41.6% for *Anger* and 22.4% for *Sadness*. In contrast, the proportion of samples in the green boxes for samples of the *Conv* set are only 2.1% for *Anger*, and 6.1% for *Sadness*. This "less-extreme" phenomenon is interesting and corroborates past studies stating that a person feels less sadness or anger when they have someone to talk to or interact with [30].

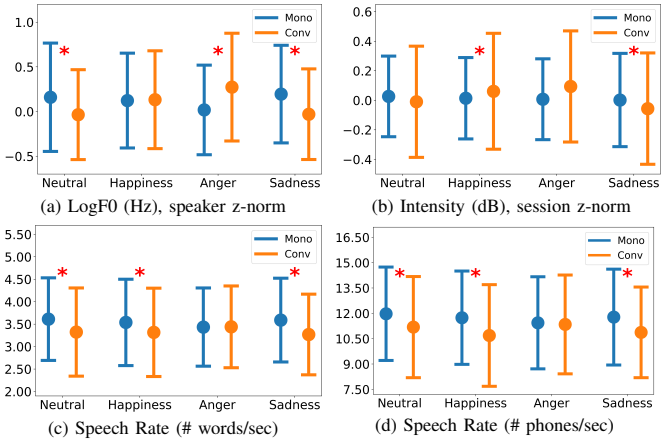
TABLE III: A summary of the agreement measurement of inter-annotator with categorical emotions.

Categorical Emotions (κ)	Overall	Neutral	Happiness	Anger	Sadness
MSP-Podcast [6]	0.229	-	-	-	-
Mono	0.448	0.218	0.108	0.184	0.113
Conv	0.468	0.285	0.199	0.197	0.221

3) Is there a difference in the quadrant-specific occupancy rate between both groups (i.e., the proportion of samples from an emotion category that is positioned in the expected V-A quadrant)? We compute the quadrant-specific occupancy rate. For *Happiness*, the occupancy rate in quadrant I is 82% for samples in the *Conv* set, but only 68% for samples in the *Mono* set. Similar trends are observed for *Sadness*, where the occupancy rate in quadrant III of samples in the *Conv* set is 74%, but only 38% for samples in the *Mono* set. This analysis indicates that the majority of samples in the *Conv* set are located on the expected quadrant for a given class, indicating that the emotional content of conversations is better perceptually aligned with the expected patterns.

2) Inter-Annotator Agreement: Inter-annotator agreement is an important indicator to assess the perceptual differences between *Mono* and *Conv*. It provides further evidences to assess rater's perceptual consistency while listening to these recordings. We compute the inter-annotator agreement on the entire *Mono* and *Conv* sets, including emotions that were not included in the analysis (i.e., surprise, fear, disgust, and contempt). We also separately estimate the agreement for *Happiness*, *Sadness*, *Anger* and *Neutral*. We adopt the *Fleiss' Kappa* (κ) [31] statistics since these are categorical emotions.

Table III shows the original inter-annotator agreement on MSP-Podcast dataset before selecting the subset and the results of the inter-annotator agreement for *Mono* and *Conv* sets. We observe that the inter-rater agreement of categorical emotions in the *Conv* set is higher than in the *Mono* set. This result is especially clear for *Happiness* and *Sadness*, which present absolute increases in their κ values of 0.091 and 0.108,



(a) LogF0 (Hz), speaker z-norm (b) Intensity (dB), session z-norm

(c) Speech Rate (# words/sec) (d) Speech Rate (# phones/sec)

Fig. 2: Results of acoustic analyses of the *Mono* and *Conv*, plotted for each of the primary emotions. Each panel corresponds to a different acoustic cue. Error bars show the standard deviation from the mean. Results that are tagged in $*$ indicates the statistical significance (two-tailed T-test, p-value ≤ 0.01) between the *Mono* and *Conv* features.

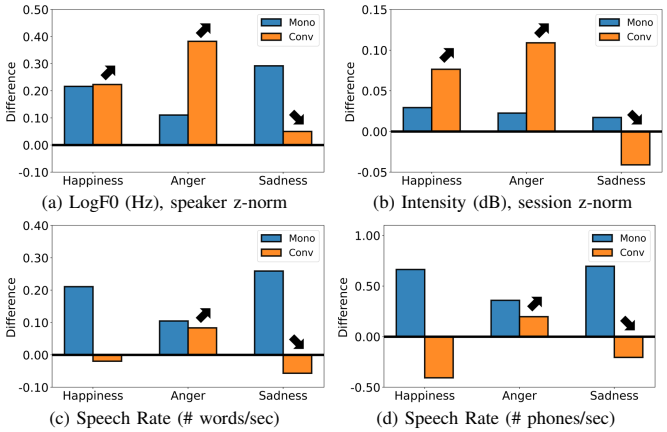
respectively. This result shows that emotions conveyed by sentences in the *Conv* set are more consistently perceived by the annotators (particularly evident for *Happiness* and *Sadness*). This higher consistency value provides another evidence that these sentences convey clearer (less ambiguous) emotions that better fit the expected emotion impressions for the annotators. It is interesting to see that the inter-annotator agreement results agree with the results reported in Sec. III-A1 for the quadrant-specific occupancy rate analysis. The sentences of *Happiness* and *Sadness* in the *Conv* set occupy with a higher rate the expected quadrant in the V-A plane. At the same time, they result in a higher inter-annotator agreement. The better consistency in the labels and the narrower and more accurate V-A plane occupancy signify that the emotional sentences in the *Conv* set are more perceptually consistent and, potentially, more authentic than the sentences in the *Mono* set.

B. Acoustic Variability

This section analyzes whether there is a difference in the acoustic features between samples in the *Mono* and *Conv* sets. We consider four acoustic features: fundamental frequency (F0), intensity, word-level and phone-level speech rate. We use the Praat [32] tool to extract the logarithm F0 (LogF0) and intensity. We perform the z-normalization on both features. For LogF0, the normalization is done per speaker, neutralizing the variations between speakers. For the intensity, the normalization is done per podcast, reducing channel effects from different settings across podcasts. For the speech rate features, we estimate the average number of spoken words and phoneme classes per second using the forced alignment results obtained with the *Montreal Forced Aligner* [33]. These alignments are based on human transcriptions. Figure 2 provides the acoustic feature analysis with respect to different emotional categories for the sentences in the *Mono* and *Conv* sets. The mean

TABLE IV: Empirical acoustic patterns of the four basic emotions [34].

	Happiness / Joy	Anger	Sadness
F0 mean	↗	↗	↘
Intensity	↗	↗	↘
Speech and articulation rate	↗	↗	↘



(a) LogF0 (Hz), speaker z-norm (b) Intensity (dB), session z-norm

(c) Speech Rate (# words/sec) (d) Speech Rate (# phones/sec)

Fig. 3: Relative differences toward the *Neutral* emotion of different acoustic features for *Happiness*, *Anger* and *Sadness*.

value of frames in a sentence represents the sentence-level feature description, where the error bars in Figure 2 shows the statistics (mean and standard deviation) over the corresponding set of emotion-specific sentences (e.g., *Happiness* set).

Figure 2 shows two major points to be noticed: 1) the difference in the sample distribution of acoustic features, and 2) the consistency with respect to the empirical expectations. First, we observe significant differences of acoustic expressiveness between the *Conv* and the *Mono* set for each emotion category. The differences in the acoustic patterns reflect the perceptual differences observed in Section III-A. Second, and more importantly, we further observe that the emotion-specific modulation on acoustic patterns for the *Conv* sentences are more consistent with the empirical expectations than for the *Mono* ones. Scherer [34] presented the expected emotion-specific modulations on acoustic patterns for *Happiness*, *Anger* and *Sadness*. Table IV summarizes the trends for F0, intensity and speech and articulation rate. For instance, we expect higher F0, intensity and speech rate for *Anger* sentences than for *Neutral* sentences. *Sadness* sentences are expected to have lower F0 value [35].

We estimate the relative changes of the acoustic features for *Happiness*, *Anger* and *Sadness* sentences with respect to features derived from *Neutral* sentences. For this part of the analysis, the reference values consider all the sentences in the MSP-Podcast corpus labeled with the label *Neutral*. We compare if the relative acoustic differences between emotional and neutral sentences fit the expected trends. Figure 3 illustrates the relative differences of the feature values. By cross referencing Figure 3 and Table IV, we can see that the expected trends in Table IV consistently match the modulation patterns of the

TABLE V: Speech emotion recognition performances in UAR(%) with SD for each emotion category with different scenarios.

Scenario		Matched		Mismatched	
Task	Model	$M \rightarrow M$	$C \rightarrow C$	$M \rightarrow C$	$C \rightarrow M$
		UAR ± SD	UAR ± SD	UAR ± SD	UAR ± SD
Neu.	CNN	**59.96** ± 3.80	56.21 ± 2.20	57.37 ± 5.49	**56.38** ± 1.67
	GRU	57.05 ± 3.08	**60.22** ± 2.41	**57.58** ± 2.49	55.61 ± 2.41
	Trans.	57.79 ± 3.07	58.18 ± 2.52	56.54 ± 2.68	55.27 ± 2.05
Hap.	CNN	64.66 ± 5.99	66.40 ± 3.77	**62.42** ± 4.28	**63.87** ± 3.18
	GRU	62.15 ± 3.98	66.35 ± 3.72	55.00 ± 2.01	59.59 ± 2.64
	Trans.	**64.93** ± 5.79	**66.60** ± 4.44	59.87 ± 3.53	56.62 ± 2.84
Ang.	CNN	**68.00** ± 4.67	**69.55** ± 3.28	66.32 ± 2.58	**67.38** ± 3.50
	GRU	55.60 ± 1.77	70.31 ± 3.62	**66.54** ± 6.10	67.02 ± 2.78
	Trans.	55.41 ± 2.72	63.63 ± 3.82	67.27 ± 6.17	63.97 ± 2.82
Sad.	CNN	**56.46** ± 3.42	57.60 ± 1.36	**53.39** ± 3.16	50.30 ± 3.01
	GRU	55.58 ± 5.49	**66.77** ± 1.30	50.61 ± 5.27	**54.76** ± 2.46
	Trans.	54.20 ± 3.40	60.05 ± 1.62	49.03 ± 2.95	53.76 ± 1.05

Conv sentences in Figure 3. For instance, we observe a higher LogF0 for *Anger* sentences and a lower intensity and speech rates for *Sadness* sentences. In contrast, sentences in the *Mono* set show some contradicting trends. For instance, the *Sadness* sentences have a much higher LogF0 and speech rates than for *Anger* sentences.

This analysis demonstrates that the acoustic patterns in sentences in the *Conv* set follow the expected emotion-specific modulation. This result is not always the case for sentences in the *Mono* set, which show non-intuitive and inconsistent patterns. The results of both the acoustic patterns and the rater perceptions (Sec. III-A) indicate that samples in the *Conv* set contain emotional information that is better aligned with findings from the emotion literature. Collectively, these findings may provide evidences that the sentences in the *Conv* set contains "better-quality" emotion information (in terms of acoustic expressiveness and perceptual judgements) when compared to sentences in the *Mono* set. We will further compare the *Conv* and *Mono* sets when training a SER system in the next section.

C. SER Model Learning

In Section III-A and III-B, our analyses indicate that samples in the *Mono* and the *Conv* differ in their rater perceptions and acoustic expressiveness. We conduct emotion recognition experiments to further investigate the effect of using either sentences in the *Mono* or *Conv* sets for training a SER system.

1) Experimental Setup: We randomly split the 48 podcasts into train (40), validation (3), and test (5) sets. The vq-wav2vec representation [36] is utilized as feature input to train our models. We carry out the emotion recognition experiments as a binary classification problem, i.e., *Neutral*, *Happiness*, *Anger* and *Sadness* detectors. We choose three different network architectures to perform binary emotion recognition: *convolutional neural network* (CNN) [37], *gated recurrent unit* (GRU) [38], and the Transformer (Trans.) [39]. For CNN and GRU, we use a model with two layers with 256 hidden nodes. For the Transformer, we use a two-layer, two-head self-attention. We consider the Adam optimizer with a learning rate with a decaying factor of 0.001. The loss function is implemented with the binary cross-entropy loss with a maximum of 50 epochs. The batch size is set to 16 with early stopping.

TABLE VI: Differences of performance in UAR(%) with SD for each emotion category with comparison of different scenarios.

Scenario	Matched		Mismatched			
Task	$M \rightarrow M \Rightarrow C \rightarrow C$		$M \rightarrow M \Rightarrow C \rightarrow M$		$C \rightarrow C \Rightarrow M \rightarrow C$	
	UAR	SD	UAR	SD	UAR	SD
Neu.	0.26 ↗	1.39 ↘	3.58	2.13 ↘	2.64	0.08 ↗
Hap.	1.67 ↗	1.35 ↘	1.06	2.61 ↘	4.18	0.16 ↘
Ang.	1.55 ↗	1.39 ↘	0.62	1.17 ↘	3.01	2.82 ↘
Sad.	10.31 ↗	2.12 ↘	1.70	0.96 ↘	13.38	1.86 ↗

We design two scenarios for the evaluation: the "Matched" and "Mismatched" scenarios. In the "Matched" scenario, the train and test sets have the same *communication modes* (*Mono* ($\boldsymbol{M \rightarrow M}$) and *Conv* ($\boldsymbol{C \rightarrow C}$)). In the "Mismatched" scenario, the train and test sets belong to different *communication modes* ($\boldsymbol{M \rightarrow C}$, $\boldsymbol{C \rightarrow M}$). We train each classifier ten times with random initialization. The *unweighted average recall* (UAR) is used as the metric to evaluate the SER performance of these models. We evaluate the variability in the performances across the ten models using the *standard deviation* (SD) to assess consistency.

2) Experimental Results and Analyses: Table V shows the emotion recognition performances for each emotional category. To highlights the trends, Table VI shows the differences in performance between the indicated experiments. In this analysis, our goal is to examine both the accuracy and the stability of using different *communication modes* for SER. First, we consider the best results per emotion observed in Table VI for the "Matched" scenario. We observe that *Conv* ($\boldsymbol{C \rightarrow C}$) leads to higher UAR than *Mono* ($\boldsymbol{M \rightarrow M}$) with performance gains of 0.26%, 1.67%, 1.55%, and 10.31% for *Neutral*, *Happiness*, *Anger*, and *Sadness*, respectively. These preliminary results are in agreement with the finding reported on Sections III-A and III-B, where sentences in the *Conv* have higher inter-annotator consistency and are more aligned with V-A impressions and acoustic expressiveness in emotion literature.

Furthermore, we observe that the mismatched condition ($\boldsymbol{C \rightarrow M}$) results in competitive UAR performances compared to the matched condition ($\boldsymbol{M \rightarrow M}$) with only a slight drop of 1.06% for *Happiness*, 0.62% for *Anger* and 1.70% for *Sadness*. In contrast, the mismatched case of using the *Mono* as source ($\boldsymbol{M \rightarrow C}$) leads to much larger recognition performance gaps compared to the matched condition ($\boldsymbol{C \rightarrow C}$), where the UAR drops 4.18% for *Happiness*, 3.01% for *Anger*, and 13.38% for *Sadness*. This result depicts that using *Conv* as the training data leads to a more robust SER that can better handle the mismatched conditions than using *Mono* as training data.

We further consider the stability of the model. Table VI shows that models trained with *Conv* as the source have less variability in the performances than models trained with *Mono*. Specifically, using *Conv* instead of *Mono* as the training data in the matched condition decreases the SD by 1.39% for

Neutral, 1.35% for *Happiness*, 1.39% for *Anger* and 2.12% for *Sadness*. A similar trend is also observed in the mismatched condition, where the SD decreases 2.13% for *Neutral*, 2.61% for *Happiness*, 1.17% for *Anger*, and 0.96% for *Sadness*. In this experiment, the overall analyses and SER results indicate that using sentences in the *Conv* set leads to SER models with higher robustness and consistency. The more consistent agreement between annotators' ratings, better consistency in the acoustic expressions, and potentially more well-behaved emotion manifestations have positioned *Conv* as a "better-quality" source for building SER databases.

IV. Discussion and Conclusion

In this work, our goal was to systematically investigate the effect of the *modes of communications* used in the collection of emotional data used to build SER models, focusing our analysis on conversation and monologue speech. Our running hypothesis was that collecting speech data in conversation *mode* helps elicit a "better-quality" data for SER than collecting data using monologue. Here, we investigated this running hypothesis under the perspectives of the rater's perceptions, acoustic expressiveness and SER model building. Our study reveals interesting insights: 1) we observed that the conversation samples occupy higher percentage of the expected V-A plane given their categorical emotion, and receive ratings with higher inter-annotator agreement as compared to monologue samples that are more widely scattered over two or more quadrants in the V-A plane; 2) we observed that the differences occur not only in the rater's perception, but also in the acoustic feature space, where the samples from conversation present more consistent patterns, matching the emotion-specific modulation of acoustic patterns reported in the literature; and 3) we observed that using speech samples drawn from conversation helps a SER model to be more robust and less variable. Collectively, these analyses point toward the fact that the communication mode of conversation may provide a "better-quality" medium for creating SER databases, potentially leading to a more robust and stable SER model. As the extensions of this work, an immediate one would be verifying our hypothesis on different corpora. Furthermore, it continues to be important to better understand the effect of these monologues and conversations on the performances of SER systems, and incorporating linguistic modalities into our analysis of emotional expressiveness will be our next step.

Ethical Impact Statement

Our pilot study on the MSP-Podcast examines the hypothesis stating that the conversation *mode* provides a "better-quality" medium for creating SER databases than the monologue *mode*. We demonstrate this hypothesis by exploring three different perspectives. However, it is clear that the three chosen perspectives are not necessarily exhaustive. Simply stating that conversation *mode* provides a better-quality source from our initial results may be ethically fraught. This hypothesis should be further investigated from different angles to prevent blind acceptance in constructing a SER database and in leading to blind development of any deployable SER model. Additionally, podcast is actually a specific speaking *mode*, which may not comprehensively represent monologues and conversations. We hope that our findings and research methodology will have a positive impact on considering the usage of conversation as an elicitation technique for building a SER model.

Acknowledgements

This work was supported by the NSTC under Grants 110-2634-F-002-050 and 110-2221-E-007-067-MY3 and the NSF under Grant CNS-2016719.

References

[1] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[2] R. Jürgens, K. Hammerschmidt, and J. Fischer, "Authentic and play-acted vocal emotion expressions reveal acoustic differences," *Frontiers in psychology*, vol. 2, p. 180, 2011.

[3] 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 Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[4] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[5] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *2008 IEEE international conference on multimedia and expo*. IEEE, 2008, pp. 865–868.

[6] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

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

[8] 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, p. 1249, 2021.

[9] R. Böck, O. Egorow, I. Siegert, and A. Wendemuth, "Comparative study on normalisation in emotion recognition from speech," in *International Conference on Intelligent Human Computer Interaction*. Springer, 2017, pp. 189–201.

[10] S. Gustafson-Capková, "Emotions in speech: tagset and acoustic correlates," *Speech Technology, term paper*, pp. 1–13, 2001.

[11] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[12] N. A. Roberts, J. L. Tsai, and J. A. Coan, "Emotion elicitation using dyadic interaction tasks," *Handbook of emotion elicitation and assessment*, pp. 106–123, 2007.

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

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

[15] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 183–196, April-June 2013.

[16] R. W. Levenson and J. M. Gottman, "Marital interaction: physiological linkage and affective exchange." *Journal of personality and social psychology*, vol. 45, no. 3, p. 587, 1983.

[17] M. J. Mallen, S. X. Day, and M. A. Green, "Online versus face-to-face conversation: An examination of relational and discourse variables." *Psychotherapy: Theory, Research, Practice, Training*, vol. 40, no. 1-2, p. 155, 2003.

[18] K. H. Lagattuta and H. M. Wellman, "Differences in early parent-child conversations about negative versus positive emotions: implications for the development of psychological understanding." *Developmental Psychology*, vol. 38, no. 4, p. 564, 2002.

[19] C. Hardy, T. Lawrence, and N. Phillips, "Talking action: Conversations, narrative and action in interorganizational collaboration," *Discourse and organization*, vol. 65, p. 83, 1998.

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

[21] X. Zhu and G. Penn, "Comparing the roles of textual, acoustic and spoken-language features on spontaneous-conversation summarization," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 2006, pp. 197–200.

[22] 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 Communication*, vol. 53, no. 1, pp. 36–50, 2011.

[23] 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 and Emotion*, vol. 39, no. 5, pp. 780–794, 2015.

[24] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 2697–2709, 2020.

[25] 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 *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7268–7272.

[26] R. Pappagari, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak, "Copypaste: An augmentation method for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6324–6328.

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

[28] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[29] J. A. Russell and G. Pratt, "A description of the affective quality attributed to environments." *Journal of personality and social psychology*, vol. 38, no. 2, p. 311, 1980.

[30] 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*. Springer, 2022, pp. 163–187.

[31] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

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

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

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

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

[36] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[37] T. Liu, S. Fang, Y. Zhao, P. Wang, and J. Zhang, "Implementation of training convolutional neural networks," *arXiv preprint arXiv:1506.01195*, 2015.

[38] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.