Analyzing Continuous-Time and Sentence-Level Annotations for Speech Emotion Recognition

Luz Martinez-Lucas, *Student Member, IEEE,* Wei-Cheng Lin, *Student Member, IEEE,* and Carlos Busso, *Fellow, IEEE*

Abstract—The emotional content of several databases are annotated with *continuous-time* (CT) annotations, providing traces with frame-by-frame scores describing the instantaneous value of an emotional attribute. However, having a single score describing the global emotion of a short segment is more convenient for several emotion recognition formulations. A common approach is to derive *sentence-level* (SL) labels from CT annotations by aggregating the values of the emotional traces across time and annotators. How similar are these aggregated SL labels from labels originally collected at the sentence level? The release of the MSP-Podcast (SL annotations) and MSP-Conversation (CT annotations) corpora provides the resources to explore the validity of aggregating SL labels from CT annotations. There are 2,884 speech segments that belong to both corpora. Using this set, this study (1) compares both types of annotations using statistical metrics, (2) evaluates their inter-evaluator agreements, and (3) explores the effect of these SL labels on *speech emotion recognition* (SER) tasks. The analysis reveals benefits of using SL labels derived from CT annotations in the estimation of valence. This analysis also provides insights on how the two types of labels differ and how that could affect a model.

Index Terms—Emotional annotations, continuous time annotations, sentence level annotations, resources for emotion recognition, speech emotion recognition.

1 Introduction

Advances in *human-computer interaction* (HCI) have become increasingly effective, improving our daily interactions with technology. However, an open problem is the inclusion of systems that are aware of emotions to enable more effective communication between computers and humans, mirroring the sophisticated way we interact. Speech is an appealing modality for HCI, considering the increased development and use of speech-based interfaces. Hence, many studies have focused on creating *speech emotion recognition* (SER) systems [1], which require a vast amount of speech data accurately annotated with effective emotional labels.

The labeling of emotional speech is often done in one of two main ways: (1) *sentence-level* (SL) annotations, and (2) *continuous-time* (CT) annotations. SL annotations assign a single label to a speech segment, reflecting the global emotion in that segment. Speech is split into short segments (phrases/sentences), which are presented to evaluators who determine the emotion that they perceive for that segment. The duration of these segments is typically between three to fifteen seconds [2], [3], [4], [5]. CT annotations assign an emotional trace capturing the instantaneous values as the evaluators judge longer speech segments (conversations/monologues). This approach captures the instantaneous emotional perception of the evaluator at each frame, creating continuous traces for the emotional attribute [6], [7], [8], [9], [10], [11], [12]. SL annotations are easier and quicker to create than CT annotations. However, emotions and their causes are not always static in a sentence or limited to the moment they occur, since they are affected by what is taking place around the speaker. CT annotations are better at capturing contextual information since annotators have already listened to the previous parts of the conversation. Therefore, CT annotations create labels that are based on the current speech and the previous speech produced by speakers in the conversation. Furthermore, CT annotations can be subdivided and aggregated at different temporal resolutions as needed (phone, syllable, word, phrase, sentence, etc.), offering a flexible resource for analysis. A common approach used in previous studies is to create SL annotations

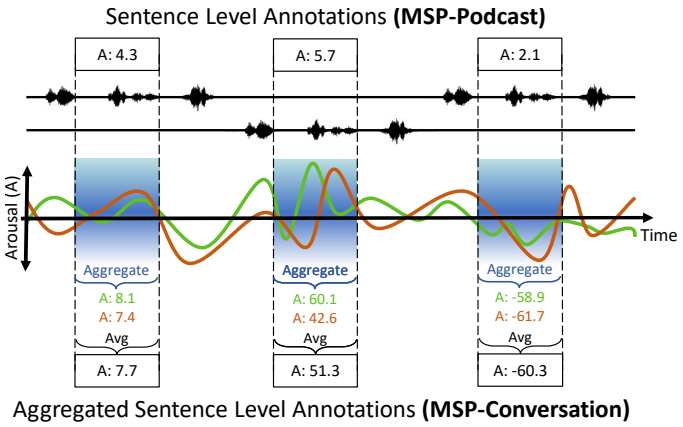


Fig. 1. Process of deriving SL labels from the CT annotations in the MSP-Conversation corpus. It shows arousal traces as an example (*A*: arousal). The green and orange traces represent two CT annotations from two individual evaluators. The SL annotations from the MSP-Podcast corpus provides the timing information for the segments used to split the traces. *Aggregate* refers to the different functions used to combine the CT labels over the segments times to obtain SL labels.

- *L. Martinez-Lucas, W.C. Lin and C. Busso are with the Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas, TX USA,*
E-mail: luz.martinez-lucas@utdallas.edu, wei-cheng.lin@utdallas.edu, busso@utdallas.edu

from CT annotations by segmenting the conversations into sentences and aggregating the CT values across annotators and across time [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24]. Figure 1 illustrates the process of creating SL labels from CT annotations. However, previous work [25] has established that aggregated SL labels and global annotations assigned to entire conversations are not interchangeable, which suggests that this issue could also exist for local SL labels. In contrast to global labels provided to an entire conversation, SL labels describe shorter speech segments of a few seconds capturing a measure of local emotion rather than a measure of a conversation's mood. Our study aims to analyze if aggregating CT labels, in the way it has been done in previous studies, is a valid way of approximating SL labels. Furthermore, we aim to provide insights into how the two types of annotations differ and how using one or the other could affect a model or task. The analysis in this study has implications for different affective computing tasks, including analysis, synthesis and recognition of emotions. To make our study more focused, we are particularly interested on the impact of these labels on emotion recognition tasks. We consider the following important questions about the relationship between SL and CT labels, and its impact on SER tasks:

- Question 1: What statistical measure, when used to aggregate CT emotional annotations into SL annotations, approximates actual SL labels the best?
- Question 2: How do the inter-evaluator agreements compare for SL annotations derived from CT annotations and actual SL annotations?
- Question 3: How well do SER models perform when they are trained with SL annotations derived from CT annotations compared with models trained with actual SL annotations?
- Question 4: How is each emotional attribute affected when SL annotations derived from CT annotations replace actual SL labels?

These questions are important for understanding the validity of aggregating CT traces into SL annotations, which is a common practice in emotion recognition studies [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], as well as understanding how using such annotations could affect models. By measuring how well aggregated SL annotations approximate SL annotations, we get a measure of similarity for different aggregation methods. We can obtain a measure of validity in aggregating CT annotations by combining those results with findings on inter-evaluator agreements and their impact on SER results. If we have similar inter-evaluator agreements and SER results between the two types of labels, then we can conclude that aggregated SL labels are as valid for our tasks as SL labels. If we have high similarity between the two types of labels, then we can also conclude that it is valid to use the aggregated SL labels as SL labels. Furthermore, even if they cannot be interchangeably used for each other, the analyses on aggregated SL labels versus SL labels can tell us how these labels compare, giving us insights on when to use the different types of labels to improve results for each emotional attribute. However, these questions are not easy to answer, since almost all the emotional databases are labeled either with SL or CT annotations, but not both. We have recently made two related corpora available to the public, offering the perfect resource to address these questions. The first database is the MSP-Podcast corpus [4], which contains SL annotations of both emotional attributes and emotional categories. These annotations are conducted with no-context on speech segments of podcasts obtained from audio-sharing websites. The second is the MSP-Conversation corpus [26], which contains CT labels of emotional attributes for conversations from the same podcasts. There are 2,884 speech segments that overlap with the recordings from the MSP-Conversation corpus. Therefore, the MSP-Conversation corpus is a perfect complement to the MSP-Podcast corpus. This study uses the overlapping segments to explore the relation between SL and CT annotations of the same speech. We transform CT annotations into SL annotations using various statistical methods and compare these new labels (MSP-Conversation) with the original labels from the SL annotations (MSP-Podcast). The annotation of the MSP-Conversation corpus contains the emotional attributes for arousal (calm to active), valence (negative to positive), and dominance (weak to strong). Therefore, the study focuses on these emotional dimensions, which are the most important attributes in the core emotion theory [27], [28], [29].

To address Questions 1 and 4, we use five statistical measures (mean, median, maximum, minimum, first quartile, and third quartile) to aggregate the CT annotations from the MSP-Conversation corpus over the speech segments that are found in the MSP-Podcast corpus. The aggregated SL labels are compared with the original SL labels from the MSP-Pocast corpus using the *mean squared error* (MSE) and the Pearson correlation coefficient (ρ). The results show that both SL labels are indeed correlated, with different relationships depending on the aggregation methods, but there are still important differences between them. For example, aggregated SL labels derived from CT annotations do not reach extreme values for the emotional attributes and are more skewed than the original SL annotations. These similarities and differences are reflected on the MSE and ρ results reported in this study. To address Questions 2 and 4, we also study the inter-evaluator agreement within each of these SL labels. For arousal and dominance, the inter-evaluator agreement is better for the original SL annotations. For valence, the inter-evaluator agreement is higher for the aggregated SL labels, which suggests that evaluators gain valuable information to assess valence from CT annotations. Finally, to address Questions 3 and 4, the analysis considers SER performances achieved when training a *deep neural network* (DNN) with the aggregated SL and original SL labels. We train SER models with the same architecture, but with different SL labels. Similar to the inter-evaluator agreement results, the models trained on arousal and dominance achieve better performance when trained with the original SL labels from the MSP-Podcast corpus, and the models for valence have better performance when trained with the aggregated SL labels from the MSP-Conversation corpus. In addition, we conduct cross-corpus SER evaluations with mismatched label conditions, where the SER models are trained with labels from one corpus, but tested with the labels from the other corpus. The mismatched label condition leads to negligible performance differences for the arousal

models, and a slight decrease in performance for the dominance models. The valence models have the best results when tested with the aggregated SL labels, regardless of the labels used to train the models. Collectively, these results suggest that aggregating SL labels from CT annotations is a valid approach, which can even help in the prediction of valence. However, the naive aggregation done to create the SL labels still results in important differences between them and the SL labels originally annotated after listening to only the speech segment.

The paper is structured as follows: Section 2 discusses related studies on SL and CT annotations and their relationships. Section 3 introduces the databases used in this study, and explains how we aggregate CT labels to create SL labels. Section 4 presents the results of the different statistical analyses done to compare the SL and CT annotations. Section 5 discusses the SER models and their performance when trained and tested with various SL labels in matched and unmatched conditions. Section 6 analyzes the role of context while analyzing CT labels in our results. Section 7 discusses the implications of this study, summarizing the main observations in this paper. Finally, Section 8 concludes the paper, suggesting future research directions.

2 Related Work

Although there are many datasets that contain either SL or CT annotations, few contain both types of annotations. Therefore, it is difficult to explore effective comparisons between them. However, previous studies have presented comparisons between emotional labels collected with context and without context. Cauldwell [30] compared SL annotations conducted in sequential order (with-context) and in random order (without-context). The study showed a definite difference in the judgements, where sentences labeled as *anger* in the without-context setting were perceived as *neutral* in the with-context setting, even when the same evaluators conducted the evaluations. This study corroborates the idea that context changes the emotional perception of speech. Jaiswal *et al.* [31] showed similar results, where SL annotations were also collected with-context and without-context. The two methods generated significantly different annotations. The with-context labels had higher agreement with the self-reported emotional labels of the speakers. However, the agreement between evaluators was higher for the without-context annotations. Jaiswal *et al.* [31] also used the two types of labels to train two static classifiers, and found that for arousal, the classifiers performed similarly, but for valence, the classifier trained using with-context labels performed worse. Moreover, when relevant context was added to the classifier by incrementally adding features from the previous speech segments to the input, the performance of all models increased, especially for the models trained with the labels annotated with-context. The results of these studies suggest that the presence of context when creating annotations will directly impact the labels, affecting the effectiveness and robustness of SER models. Our study differs from the previously mentioned studies since they only focus on label differences when the surrounding speech is considered by the annotators. Our study not only focuses on that type of added context, but also on the effect of annotations at different temporal resolutions. While annotators must provide instantaneous ratings during CT annotations, they have more time during SL annotations to think and summarize their annotations.

Metallinou and Narayanan [25] presented the most similar study to our work. CT annotations of the emotional attributes of arousal, valence, and dominance were conducted on the multimodal database CreativeIT [8], which contains theatrical performances ranging from 2 to 10 minutes in length. Evaluators were also asked to give a global annotation of the full performance immediately after finishing the CT annotation. The CT annotations were aggregated over time using the mean, median, maximum, minimum, and first and third quartiles. They estimated the MSE between the aggregated SL annotations and the global annotations. The lowest MSE for all attributes was when either the first or third quartiles were chosen, which implies that extreme values in longer recordings affect global perception the most. Notice that changes in emotions within the sessions are not represented in the global emotional score, so the temporal resolution provided in this study is not ideal to compare aggregated SL annotations and SL annotations. Our study differs in that the annotations in the MSP-Podcast and MSP-Conversation corpora are conducted on only speech data by different annotators, and the SL annotations are done on much shorter segments than the global annotations of the CreativeIT database. Global annotations encompass much longer sequences of speech (2-10 mins), which means that such values give a measure of mood instead of local emotion associated with a speaking turn. The SL annotations in our study focus on shorter speech segments, and, therefore, do measure the local emotion in these segments. We consider speech segments between three and eleven seconds which are representative of the common sentence duration in most speech emotional corpora [2], [3], [4], [5]. These resources allow us to focus on the validity of using SL labels aggregated from CT annotations as SL annotations that are meant to capture local emotion. Furthermore, these resources also give us insights into how CT annotations and SL annotations differ and how they could affect SER models.

3 Resources

3.1 The MSP-Podcast Corpus

The MSP-Podcast corpus [4], which is used as the source of the original SL labels, is a publicly available corpus of natural emotional speech sourced from podcasts on audio-sharing websites. The selected podcasts are chosen so that the corpus contains conversations and monologues of varying topics, a diverse group of speakers, and diverse emotional content. The speech is natural and spontaneous, as it was not recorded to specifically elicit emotional speech. For ease of sharing, the podcasts are all available under the less restrictive Creative Commons licenses (CC-BY and CC-0). The data protocol includes several steps. First, the podcasts are segmented into speaking turns using the Microsoft Azure Video Indexer. The next step is to identify segments without speaker overlap, music, or noise. All these steps are automatically implemented as part of the protocol. The selected segments have durations between 2.7s and 11s and

are annotated in a random order. This range of the duration of the speaking turns is selected to have speaking turns that are long enough to covey emotional cues that can be reliably annotated by human raters, and short enough to mitigate long sentences where the emotional content changes within the speaking turn. This process creates a speech repository from where we annotated the data.

For most of the corpus, the selection of sentences to be annotated follows the retrieval-based approach presented by Mariooryad *et al.* [32], where machine learning models are used to identify speech segments with emotional content. As explained in Section 3.2, this study uses the sentences that overlap with the recordings of the MSP-Conversation corpus. One of our goals is to maximize the number of speaking turns in the MSP-Podcast corpus that overlap with the recordings on the MSP-Conversation corpus. Therefore, we annotate all the segments contained in the conversations regardless of the emotional content of the speaking turns, as long as there is no overlapped speech, music or noisy recordings. The corpus contains SL annotations of the speech segments, collected with a crowdsourcing protocol inspired by the work of Burmania *et al.* [33]. Evaluators are asked to label the primary and secondary emotions of each speech segment with categorical emotional labels (e.g., happiness, sadness, anger). The annotations also include the emotional attributes valence (negative versus positive), arousal (calm versus active), and dominance (weak versus strong). For emotional attributes, we used a seven-point Likert scale. The data collection is an ongoing effort, where this study uses the release 1.9 of the corpus, which includes 86,389 speaking turns (137 hours, 11 minutes). For this study, we only consider the segments that overlap with the recordings of the MSP-Conversation corpus.

3.2 The MSP-Conversation Corpus

The MSP-Conversation corpus [26] was created as a complement to the MSP-Podcast corpus (Sec. 3.1) to study contextual information in SER tasks. The corpus is an ongoing effort that contains conversations sourced from the same podcasts identified in the MSP-Podcast corpus. These conversations are 10 to 20 minutes long and are chosen for their emotional and speaker content, focusing on a broad and balanced range of emotional content and gender diversity. The conversations are segmented into three to seven minute speech segments that are labeled with CT annotations for the three emotional attributes (arousal, valence, and dominance). Evaluators use the CARMA software [34] and a joystick to record their instantaneous emotional perception while listening to the conversation. The interface uses a scale of -100 to 100 for each emotional attribute. The first version of the MSP-Conversation corpus contains 74 conversations (15hrs, 9mins), and contains speech from 197 different speakers (87 female, 110 male). The corpus contains at least six annotations from different evaluators per conversation.

The MSP-Podcast corpus contains speech segments that overlap with the recordings in the MSP-Conversation corpus. To our knowledge, they are the only datasets that, combined, have both CT and SL labels of the same speech audio. When we consider version 1.9 of the MSP-Podcast corpus, and the first version of the MSP-Conversation corpus, there are 2,884 MSP-Podcast speech segments found in the MSP-Conversation speech data, which is more than 4 hours of data. We focus our analysis on these segments. The amount of speech data available is comparable to many early corpora that contain SL annotations [3], [35], [36], and this data will keep growing as both the MSP-Podcast and MSP-Conversation corpora are ongoing efforts. Furthermore, the corpora contain natural speech data collected from spontaneous conversations, and not recorded in a laboratory environment or purposely acted [2], [8], [9], [10], [37].

3.3 Creation of SL Annotations from CT Annotations

This study aims to compare aggregated SL labels, obtained from CT annotations of the MSP-Conversation corpus, with the original SL annotations of the MSP-Podcast corpus. This section explains the process of converting the emotional traces into SL labels for the speech segments overlapping with the MSP-Podcast corpus. First, we segment the CT labels from each evaluator using the timings of the 2,884 speech segments from the MSP-Podcast corpus. Then, those traces are aggregated over time for each evaluator. We use the aggregation methods used in Metallinou and Narayanan [25] by taking the following statistics on the traces for each emotional attribute during the target segments: *minimum* (m), *first quartile* (Q1), *median* (Q2), *mean* (μ), *third quartile* (Q3), and *maximum* (M). Then, the aggregated values of each annotator are averaged together to get the consensus labels for the MSP-Conversation corpus, which we refer to as *aggregated SL annotations*. Figure 1 demonstrates the full process, where the traces are the CT annotations that are aggregated into SL annotations to be compared with the labels in the MSP-Podcast corpus. We also account for the reaction lag of the evaluators while annotating CT labels (i.e., time an evaluator takes listening to and judging the emotional content before moving the joystick to record her/his perception [38], [39]). We use a constant time shift (T) to account for this reaction lag. For each annotator trace, we shift the annotation back by T seconds, effectively using the labels that were recorded T seconds after the annotator heard the utterance. Then, the rest of the aggregation process is performed. Figure 2 shows this process. We use time shifts of 2.8, 3.0, 3.6, 4.08, 5.44, and 5.6 seconds, as they have been found to be ideal delays by previous studies [39], [40]. In this study we focus on the simple task of constant time shifts for all annotators, which is sufficient as a first task. As found by Mariooryad and Busso [39], a constant delay results in a similar performance than using annotator and task specific delays.

4 Statistical Analysis

This section analyzes the MSP-Conversation SL labels for each time shift, aggregation method, and emotional attribute, comparing the results with the SL annotations from the MSP-Podcast corpus. Section 4.1 compares the SL labels from both corpora. Section 4.2 evaluates the inter-evaluator agreement within each corpus.

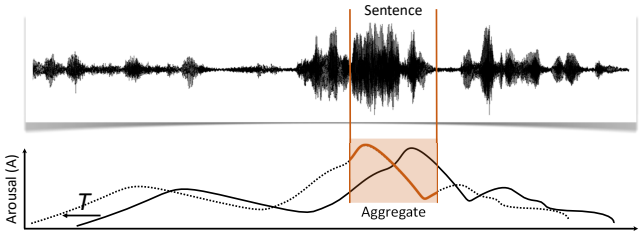


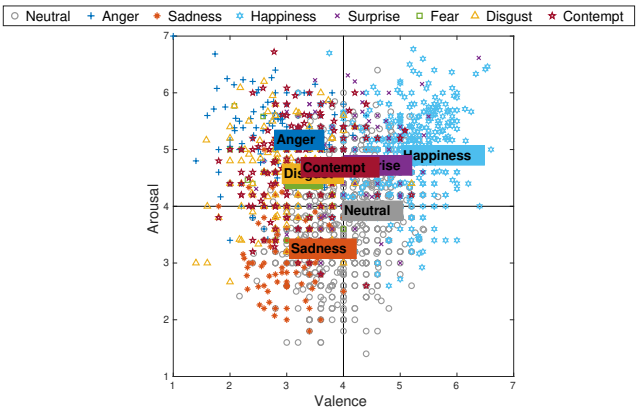
Fig. 2. Process of adding a time shift T to the aggregated SL labels. The trace of one annotator (curve in black) is shifted back by T seconds to get a shifted trace (dashed curve). The sentence timing is then used to aggregate the shifted trace over the appropriate time.

4.1 Aggregated SL Annotations versus SL Annotations

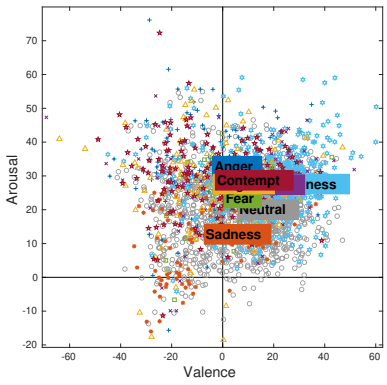
Our first analysis is to visualize the relation between the aggregated SL annotations from the MSP-Conversation corpus and the SL annotations from the MSP-Podcast corpus. First, we analyze the relationship between categorical and attribute-based labels assigned by the SL and CT labels. We plot the sentences on the arousal-valence space, grouping the sentences according to the categorical emotion labels provided by the MSP-Podcast corpus. We consider two conditions by using the emotional attributes either from the MSP-Podcast corpus (Fig. 3(a)) or aggregated from the MSP-Conversation corpus (Fig. 3(b)) using no time shift and the mean aggregation rule. Figure 3 shows that the SL labels and the aggregated SL labels have similar distributions over their given emotional space since the categorical groups are in similar positions relative to each other. However, the aggregated SL labels seem to have fewer extreme values, with a shift toward the positive side of the arousal space. One explanation is that annotators can easily react to more active behaviors. For less active behaviors, some raters may return to the origin instead of moving the joystick toward lower values of arousal.

Figure 4 shows the results when the SL and aggregated SL labels are directly compared, where each point corresponds to a sentence. The plots for different time shifts and aggregation methods look similar, so Figure 4 only reports the case with no compensation for the reaction lag, and with the mean aggregation rule. While the plots reveal a level of correlation between both annotation methods, they show that the MSP-Podcast labels cover a larger area of their range than the aggregated MSP-Conversation labels. Furthermore, using the maximum or minimum for aggregation does not change the look of the plots, suggesting that speech segments labeled with extreme values during SL annotations are not being labeled with extreme values during CT annotations. This result may be influenced by the time annotators are given to assign a value. Humans are notoriously less accurate at giving absolute values [41], [42], which would be exacerbated when being asked to give an instantaneous rating as done with the CT annotations. Using the SL labels derived from CT annotations would lead to models that predict emotional states closer to neutral than using original SL annotations, and skewed towards positive values of arousal and dominance.

We also compare the aggregated SL annotations from the MSP-Conversation corpus with the original SL annotations from the MSP-Podcast corpus using the MSE, and the *Pearson correlation coefficient* (ρ). For this analysis, the labels from the MSP-Podcast corpus are linearly scaled to match the range of the MSP-Conversation labels mapping the values from a range of 1 to 7 into a range of -100 to 100. Furthermore, looking at Figures 3 and 4, we observe that the MSP-Conversation labels have clear shifts, especially for the arousal and dominance dimensions. Therefore, we conduct some experiments with normalized labels. We use a method inspired by the normalization strategy proposed in Busso *et al.* [43]. The approach takes the average of the aggregated SL and SL labels for each attribute over the sentences labeled with the primary class neutral in the

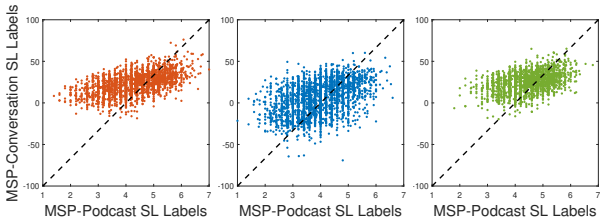


(a) Attributes from MSP-Podcast



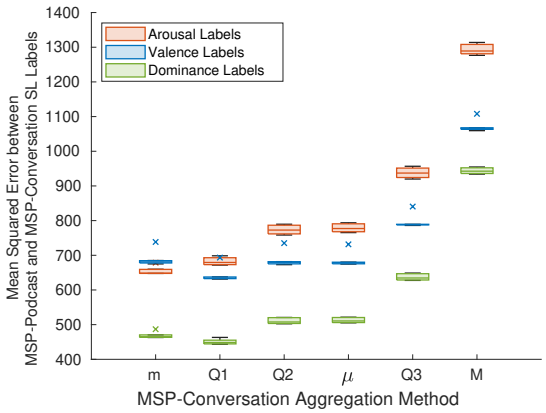
(b) Attributes from MSP-Conversation

Fig. 3. Valence-Arousal space with emotional attributes either from the (a) MSP-Podcast corpus or the (b) MSP-Conversation corpus. The categorical classes assigned to the MSP-Podcast corpus are added to the plots.

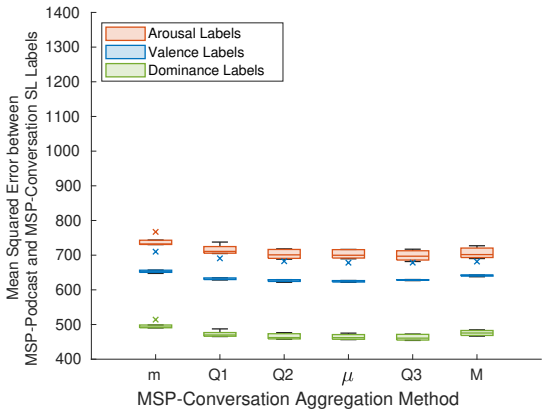


(a) Arousal (b) Valence (c) Dominance

Fig. 4. Plots of the MSP-Conversation SL labels (aggregated using the mean and no time shift) versus the MSP-Podcast SL labels. Each point corresponds to a speech segment.



(a) Original Labels



(b) Normalized Labels

Fig. 5. *Mean squared error* (MSE) between the MSP-Conversation SL labels and the MSP-Podcast SL labels. The values for each time shift are grouped to make the box plots. The traces are aggregated with the functions *minimum* (m), *first quartile* (Q1), *median* (Q2), *mean* (μ), *third quartile* (Q3), and *maximum* (M). The analysis is conducted with and without the label normalization described in Section 4.1

MSP-Podcast dataset. Then, those averages were subtracted from their corresponding labels, normalizing both the aggregated SL and SL labels to be centered around the neutral values of each attribute. This normalization mitigates the shift between the two annotation strategies. Our analysis considers the labels with and without this normalization. For each attribute and aggregation scheme, we estimate seven values corresponding to no time-shift, and the six delays mentioned in Section 3.3 for both the original and normalized labels. Figures 5 and 6 show these results.

Figure 5 shows the MSE between the MSP-Conversation SL labels and the MSP-Podcast SL labels. The outliers, marked with the 'x' symbol, correspond to time shifts of 0s. The overall MSE results for the dimensions follow the trends seen in Figure 4. Dominance has the most data points close to the black dashed line, which represents a MSE of 0, and arousal has the least. Therefore, dominance labels have the smallest MSE, while arousal labels have the highest MSE. These results are observed with and without label normalization. Without label normalization, the MSE is lower when the CT labels are aggregated over the segments with the functions minimum (m) and first quartile (Q1) (i.e., focus on lower extremes). However, the normalization *flattens* the results, showing that the shifts in the labels affect the MSE

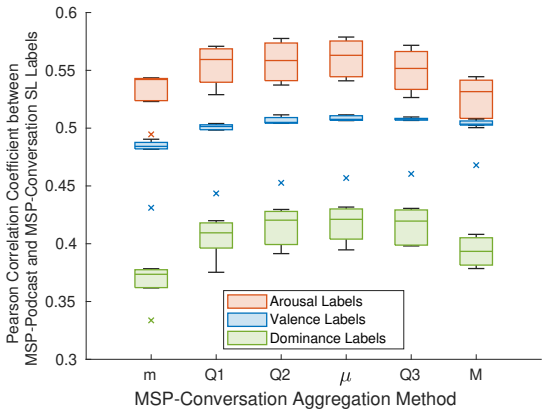


Fig. 6. Pearson correlation coefficient between the MSP-Conversation SL labels and the MSP-Podcast SL labels. We only present the results without the label normalization, since the normalization does not affect the correlation between the labels.

the most. This analysis also tells us that using the aggregated SL labels in models would lead to predictions that are different in absolute value when compared to predictions of models trained with original SL labels, even when they are predicting an emotional attribute with the same acoustic features.

Figure 6 shows the results for the Pearson Correlation Coefficient (ρ). We only present the results without the label normalization, since the correlation does not change with this normalization. The outliers also correspond to time shifts of 0s. Overall, the aggregation methods focused on middle values show better correlation. Overall, arousal has the best ρ values and dominance has the worst ρ values. Focusing on aggregation methods that consider middle values may capture the relative trends between the annotations better (i.e., higher correlation). For arousal and valence, the aggregation method using the mean (μ) had the best ρ values. For dominance, using the third quartile (Q3) gave the best values. The best ρ values for each attribute were: ρ=0.555 for arousal (mean aggregation function and a time shift of 2.8s), ρ=0.498 for valence (mean aggregation function and a time shift of 3.6s), and ρ=0.421 for dominance (third quartile aggregation function and a time shift of 2.8s). Those ρ values show that the SL labels derived from CT annotations are definitely correlated with the original SL annotations, even when their absolute values are shifted. Models trained with either labels could predict emotional states that are ordered similarly, meaning both models could agree on which speech segment is more active or more positive than another.

Figures 4, 5 and 6 show that the SL annotations from the MSP-Podcast corpus and the aggregated SL labels from the MSP-Conversation corpus are correlated but have substantially different absolute values. We see from Figure 4 that the range of values for labels derived from CT annotations is narrower, without reaching the extremes. For dominance and arousal, the figure also shows that most of the annotations from the MSP-Conversation corpus are over 0, so the distributions are skewed to more active and stronger emotions. The annotations for the MSP-Podcast corpus are more balanced. Therefore, annotations that follow the same trends, but are not necessarily close in value can explain the

disagreement among the ρ and MSE results. As mentioned previously, these results show that using the aggregated SL labels to train models could lead to different absolute predictions when compared to models trained with original SL annotations, but both models could still agree on the emotional ordering of sentences. Another interesting result from Figures 5 and 6 is the lower spread of the boxes for valence. This result suggests that changes on the time shift to compensate for the reaction lag are less important when deriving aggregated SL annotations from the valence emotional traces. Arousal and dominance are more sensitive to the time shift.

The results from the normalization are encouraging for the typical methods of aggregating CT labels, since the results show that using the mean or median are the best at approximating SL labels when compared to the other simple aggregation methods, as shown by the correlation results (Fig. 6). Furthermore, although normalizing does make the MSE results better, they are not too different for the mean and median. Therefore, it is not necessary to conduct the normalization when categorical labels are not available (the normalization relies on selecting neutral samples). These results show that while normalization is useful to better understand the results obtained with different aggregation methods, normalization is not necessary when aggregating time-continuous traces. The statistical experiments point to the mean and median, with a time shift, being sufficient for CT labels approximating SL labels. When comparing the results between time shifts, we observe that consistently a time shift of 2.8s is best for approximating SL labels from CT labels, but the results from time shifts do not diverge too far from each other, supporting the results in Mariooryad and Busso [39]. The exception is when we do not use a time shift, which is consistently worse than any time shift. We recommend that CT labels continue to be used by taking into account the reaction time of the annotators by including at least a constant time shift.

TABLE 1
Krippendorff's Alpha coefficient to measure inter-evaluator for (1) the aggregated SL labels (MSP-Conversation), and (2) the original SL labels (MSP-Podcast). The traces are aggregated with the functions *minimum* (m), *first quartile* (Q1), *median* (Q2), *mean* (μ), *third quartile* (Q3), and *maximum* (M).

MSP-Podcast Labels		
Arousal=0.308	Valence=0.353	Dominance=0.228

MSP-Conversation Sentence Level Labels							
	Time	**Aggregation Method**					
	Shift	μ	**Q2**	**M**	**m**	**Q1**	**Q3**
Arousal	*0.00*	0.101	0.098	0.095	**0.129**	0.097	0.100
	2.80	0.100	0.095	0.095	0.121	0.100	0.099
	3.00	0.100	0.096	0.095	0.121	0.101	0.100
	3.60	0.101	0.096	0.097	0.121	0.102	0.101
	4.08	0.102	0.098	0.098	0.124	0.103	0.101
	5.44	0.103	0.100	0.100	0.125	0.102	0.103
	5.60	0.103	0.101	0.100	0.124	0.102	0.104
	Time	**Aggregation Method**					
	Shift	μ	**Q2**	**M**	**m**	**Q1**	**Q3**
Valence	*0.00*	0.396	0.405	0.381	0.393	0.407	0.402
	2.80	0.397	0.406	0.381	0.392	0.409	0.404
	3.00	0.397	0.406	0.381	0.392	**0.410**	0.404
	3.60	0.398	0.408	0.384	0.392	0.409	0.405
	4.08	0.398	0.408	0.384	0.394	0.410	0.406
	5.44	0.398	0.409	0.382	0.392	0.410	0.405
	5.60	0.398	0.409	0.382	0.392	0.409	0.405
	Time	**Aggregation Method**					
	Shift	μ	**Q2**	**M**	**m**	**Q1**	**Q3**
Dominance	*0.00*	0.070	0.067	0.079	0.075	0.071	0.079
	2.80	0.069	0.068	0.081	0.069	0.072	0.080
	3.00	0.069	0.066	0.083	0.070	0.071	0.079
	3.60	0.070	0.070	0.083	0.069	0.071	0.081
	4.08	0.070	0.070	0.083	0.068	0.071	0.081
	5.44	0.071	0.074	**0.086**	0.070	0.070	0.080
	5.60	0.071	0.073	0.085	0.070	0.069	0.080

4.2 Inter-Evaluator Agreement

This section compares the inter-evaluator agreements obtained after aggregating the CT annotations into SL labels. The agreements are compared with the inter-evaluator agreements from the SL annotations on the MSP-Podcast corpus for the 2,884 target sentences. Notice that this analysis compares the agreement across annotators within each of the labeling methods. We consider the reaction lags and aggregation methods presented in Section 3.3. Since the emotional attributes are rating values, we estimate the inter-evaluator agreement using the *Krippendorff's alpha coefficient* (K alpha) [44]. The normalization done in Sec. 4.1 does not affect the inter-evaluator agreement scores since it shifts all annotator labels equally. Therefore, we only report the agreements on the labels without normalization.

Table 1 shows the results. For arousal and dominance, the MSP-Podcast labels had the best inter-evaluator agreement ($\alpha_{arousal}$=0.308 and $\alpha_{dominance}$=0.228). These agreements are better than the best performance achieved with the MSP-Conversation labels ($\alpha_{arousal}$=0.129 and $\alpha_{dominance}$=0.086). For valence, in contrast, the MSP-Conversation SL labels had the best agreement ($\alpha_{valence}$=0.410), with the best agreement coming from the labels aggregated using the first quartile and a time shift of 3s. Using the median aggregation method is also very competitive for valence. Between the attributes, the labels of valence have the highest agreements for all labels and dominance have the lowest agreements. Valence often has the highest inter-evaluator agreement in perceptual evaluations among emotional attributes [2], [4], [9]. The definition of dominance is often more difficult to grasp than the definitions for arousal and valence, which may lead to lower inter-evaluator agreement. Evaluators seem to have more distinct arousal and dominance annotations when they are continuously annotating, but the large drop in agreement might also be partially due to the resolution given to the evaluators to annotate their emotional judgments. The annotators for the MSP-Podcast corpus only have seven choices for each attribute, which will force evaluators with slightly different perceptions to choose the same rating. On the other hand, the MSP-Conversation evaluators are given choices of effectively any number between -100 to 100, which would make it almost impossible for two evaluators to have the same annotation. Furthermore, the aggregation methods assume that all annotators have the same reaction lag, so the annotators might agree more if their actual reaction lags were used. All of those reasons make it surprising that

the valence agreement increases from the MSP-Podcast to the MSP-Conversation SL labels. This result suggests that the process for CT annotations gives evaluators important information for rating valence, which leads to more accurate annotations. This could be due to the added contextual information, since its importance for valence has been previously observed [45] (Sec. 6).

Cowie *et al.* [6] discussed the problem of identifying "appropriate tasks" for emotional annotation. If the task is appropriate then the annotations derived from it are viable. The viability of SL annotations has been validated by their successful usage, but the viability of aggregated SL labels is still in question. The authors mentioned that inter-rater agreement has been used for checking viability. However, they argue that this is only a good criterion if the annotators are a somewhat homogenous group annotating unambiguous samples. Although the annotators for the MSP-Conversation are relatively homogenous, as they were chosen from the same university with similar cultural backgrounds, the sentences being analyzed are not unambiguous. The agreement results in this section suggest that the arousal and valence aggregated SL labels are viable since the agreements are close to or higher than the agreements for the original SL labels. The dominance labels do not show a high enough agreement for viability, but since the sentences could be ambiguous for dominance, we cannot conclude that the aggregated SL dominance labels are not viable. Using the labels in simple SER tasks gives us a clearer idea of the viability of the aggregated SL labels for our specific tasks (Sec. 5).

5 Evaluation with SER Models

We evaluate the impact of aggregated SL labels derived from CT annotations on SER tasks, comparing the results with the ones obtained with the original SL annotations. We compare not only the effectiveness of the models in matched label conditions (Sec. 5.1), but also its generalization to mismatched label conditions by training with labels from one corpus and evaluating the results with labels from the other corpus (Sec. 5.2). The normalization done in Sec. 4.1 does not affect the model since any simple bias taken care of by the normalization is already accounted for in the bias terms included in our SER model. Therefore, we only report the SER results on the labels without normalization. Since our focus is on emotion recognition tasks, we define effectiveness of the labels as the ability of SER models to predict the labels assigned by the evaluators, using either SL or CT-based annotations. From a machine learning perspective, an SER model with good performance is able to learn patterns from the acoustic features to predict the labels. The consistency in the labels is expected to have a marked effect on the performance, considering that noisy, inconsistent SL labels in a classification task typically impair the learning capability of the models [37].

The evaluation uses the 2,884 speech segments mentioned in Section 3.2. We use the partition suggested for the MSP-Podcast corpus for training, development, and testing, which gives us 1,688 speech segments in the train set, 903 speech segments in test set, and 293 speech segments in the development set. The size of this set is significantly smaller than the full corpus, so the SER results reported here cannot be directly compared with other studies using the entire MSP-Podcast corpus [46], [47], [48], [49], [50], [51], [52]. The acoustic feature vector corresponds to the feature set proposed for the Interspeech 2013 *computational paralinguistics challenge* (ComParE) [53]. The feature set is extracted with the openSMILE toolkit [54]. The set computes statistical descriptions (e.g., mean) over frame-level features such as the fundamental frequency and energy. The set generates a 6,373 dimensional feature vector to represent the speech segment, regardless of the duration of the sentence. Given the reduced size of the training set, we build a simple *deep neural network* (DNN) for our evaluation. State-of-the-art SER models are large, need an even larger amount of data, and take time to fine-tune and train. Using a simple neural network allows us to quickly get and validate our results, reducing the chances of overfitting the network. The model consists of two fully connected hidden layers implemented with 128 nodes, and an output layer with a single node. We use $tanh$ as the activation function, and dropout for the second layer with a dropout rate of $p = 0.3$ for arousal and dominance and a dropout rate of $p = 0.5$ for valence. The higher dropout rate for valence follows the recommendations by Sridhar *et al.* [55]. The output layer has a linear activation with one node since this is a regression task. We build separate models for arousal, valence, and dominance. The models are trained to maximize the *concordance correlation coefficient* (CCC), which is also our evaluation metric. We use the Adam optimizer with a learning rate of 0.0001, and early-stopping criterion on the development set. The best model is evaluated on the test set. We train for 100 epochs with a batch size of 128.

5.1 SER Results with Matched Label Conditions

Table 2 shows the results for the matched label condition. The best models trained with the aggregated SL labels from the MSP-Conversation corpus for each attribute are shown in bold. The results show that for arousal and dominance, the models trained with the SL labels from the MSP-Podcast corpus have better CCC values than all the models trained with the aggregated SL labels from the MSP-Conversation corpus. The models have an easier time finding the relationship between the features and the labels when the labels come from SL annotations. For arousal and dominance, the decrease in performance could be explained by the lower agreement between the evaluators for the aggregated SL labels (Sec. 4.2). However, this performance drop is not as drastic as expected from the low inter-evaluator agreements for arousal and dominance for aggregated SL annotations. We conducted further SER model testing to explore the relationship between inter-evaluator agreement and our model's performance. For this analysis, we only use the labels from the MSP-Conversation corpus. We split the test set of the MSP-Conversation corpus into two equal sets by considering sentences with low and high agreement. We order the sentences according to the standard deviation of their attribute values assigned by the evaluators, splitting the sentences into low and high agreement sets. We do not retrain the SER models. Instead, we only test the models on these two sets. We consider the best SER models in Table 2, highlighted in bold. Figure 7 shows the results. We observe

TABLE 2
CCC values of SER models in matched label conditions. Aggregation functions: *minimum* (m), *first quartile* (Q1), *median* (Q2), *mean* (μ), *third quartile* (Q3), and *maximum* (M).

MSP-Podcast Labels						
Arousal=0.556		Valence=0.083			Dominance=0.462	
Models Trained with MSP-Conversation SL Labels						
	Time	**Aggregation Method**				
	Shift	μ	**Q2**	**M**	**m**	**Q1**	**Q3**
Arousal	*0.00*	0.494	0.498	0.463	0.417	0.461	0.448
	2.80	0.517	**0.522**	0.495	0.421	0.488	0.474
	3.00	0.510	0.515	0.495	0.421	0.482	0.469
	3.60	0.452	0.500	0.485	0.444	0.474	0.454
	4.08	0.447	0.494	0.475	0.432	0.462	0.448
	5.44	0.458	0.456	0.440	0.417	0.441	0.453
	5.60	0.453	0.452	0.437	0.413	0.434	0.448
	Time	**Aggregation Method**				
	Shift	μ	**Q2**	**M**	**m**	**Q1**	**Q3**
Valence	*0.00*	0.231	0.226	**0.280**	0.191	0.214	0.246
	2.80	0.228	0.226	0.261	0.191	0.221	0.233
	3.00	0.228	0.223	0.262	0.191	0.218	0.234
	3.60	0.225	0.220	0.257	0.190	0.217	0.236
	4.08	0.226	0.222	0.258	0.193	0.214	0.237
	5.44	0.232	0.229	0.259	0.197	0.224	0.236
	5.60	0.233	0.231	0.260	0.196	0.227	0.235
	Time	**Aggregation Method**				
	Shift	μ	**Q2**	**M**	**m**	**Q1**	**Q3**
Dominance	*0.00*	0.307	0.310	0.243	0.298	0.312	0.299
	2.80	0.361	**0.369**	0.278	0.353	0.359	0.336
	3.00	0.360	0.364	0.277	0.355	0.360	0.336
	3.60	0.356	0.358	0.277	0.359	0.368	0.329
	4.08	0.349	0.349	0.278	0.345	0.368	0.322
	5.44	0.337	0.338	0.263	0.359	0.360	0.302
	5.60	0.335	0.337	0.262	0.360	0.359	0.300

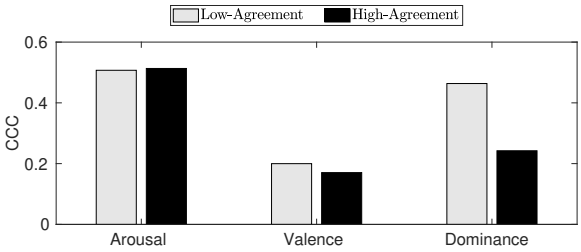


Fig. 7. CCC values of the best MSP-Conversation SER models (bolded in Table 2) for each attribute tested with the low and high agreement sentence sets.

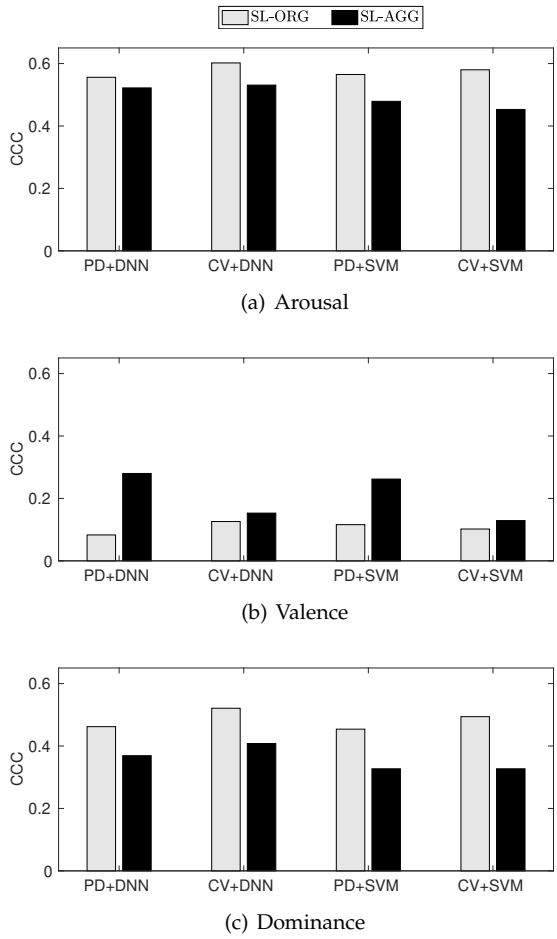
that higher agreements do not necessarily lead to labels that are easier to predict. Therefore, even though arousal and dominance do not show high agreement, their model results can still be good. In contrast, Table 2 shows opposing results for valence, where the aggregated SL labels result in better predictions than using the original SL labels. We hypothesize that these results can be explained by the extra context combined with the instantaneous reactions captured by CT labels. We extend the analysis on the role of context in Section 6. The CCC results for valence are in general lower than the other attributes. This result is expected since valence is usually the most difficult attribute to predict from acoustic features [56], [57], [58], [59]. The lower performance for valence is consistently reported across SER studies. We also observe that the CT annotations make it easier for the SER model to learn the connection between the valence labels and acoustic features.

We conduct two additional experiments to validate the robustness of SER results presented in this section. First, we use a simpler SER model to evaluate the results using *principal component analysis* (PCA), and *support vector machine* (SVM). PCA is set to reduce the feature dimension from 6,373D to 128D (same as the DNN model), and the SVM model (linear kernel, penalty=1) is trained on this 128D feature vector and their corresponding emotion labels. This PCA-SVM model is used here as a naïve model for SER. By comparing the PCA-SVM results with the DNN results, we both evaluate the generalizability of the conclusions drawn in this section as well as provide a baseline for the DNN results. Second, we design a speaker-independent five-fold cross-validation (CV) setting to evaluate the SER results. This approach aims to evaluate whether the results in Table 2 are due to specific speech files included in the predefined train, development, and test partitions. This CV process is implemented for the DNN and PCA-SVM models. Figure 8 shows the results for the two models with the *predefined partition* (PD) and with the CV setting. For the models trained with the MSP-Conversation corpus, we only report the best results achieved with different aggregation strategies and time shifts for each of the evaluation approaches. We consistently observe similar performance trends across different evaluation approaches. We observe better valence performance, but worse arousal and dominance performance when using the CT labels compared to the corresponding SL labels. These results validate the findings reported in this section.

5.2 SER Results with Mismatched Label Conditions

To explore the generalizability of the SER models, this section discusses the mismatched label condition where the models are trained with labels from one corpus and tested with labels from the other corpus. We test six of the models with labels from the corpus they were not trained on. The first three models are the ones trained on the MSP-Podcast annotations. The other three models are the best SER systems trained on the aggregated SL labels from the MSP-Conversation corpus for each attribute (conditions highlighted in bold in Table 2). We train all six models 10 times with different network initialization settings reporting the average results. We assess whether the difference between matched and mismatched conditions are statistically significant using the two-tailed t-tests, asserting significance with p-value< 0.05 (the degree of freedom is 9).

Table 3 shows the results in both mismatched and matched label conditions for all six models (three emotional attributes $\times$ two types of labels). In the table, the values at the diagonals show the test results for the matched conditions for each attribute, and the out-of-diagonal values are the mismatched results. The standard deviation across the 10 runs are reported in brackets. We compare each matched test result with the result of the same model tested in the mismatched condition. For example, we compare the arousal model trained and tested with the original SL labels to the same model tested with the aggregated SL labels. Two matched test results are shown to be significantly better than their mismatched counterparts: the valence matched model



(a) Arousal

(b) Valence

(c) Dominance

Fig. 8. CCC values of SER models with different evaluation approaches to show the consistency of the results. We compare models trained with the MSP-Podcast original SL labels (SL-ORG) and the MSP-Conversation aggregated SL labels (SL-AGG) The models used for the MSP-Conversation evaluations are the best models for each attribute and model design. The evaluation approaches are composed of either a PD or CV partition design and either a DNN or SVM model setup.

trained with aggregated SL labels (CCC = 0.185) and the dominance matched model trained with original SL labels (CCC = 0.483). One mismatched test result is shown to be significantly better than its matched counterpart: the valence mismatched model trained with aggregated SL labels (CCC = 0.145). For arousal, the model trained with SL labels from the MSP-Podcast corpus increased performance when tested with the aggregated SL labels from the MSP-Conversation corpus. The model trained with the aggregated SL labels achieves lower performance when tested with the original SL labels. However, the performance differences between matched and unmatched conditions were not statistically significant. Therefore, we conclude that the original and aggregated SL arousal labels share relevant information for this SER task, which agrees with our statistical analyses in Figure 6 that show that arousal has the most correlated labels when considering both labels. For valence, however, the models tested with the aggregated MSP-Conversation SL labels had the best performance, with the best model being trained and tested with the aggregated SL labels. It appears that the SER model is learning some valuable and similar connections from training with both types of

TABLE 3
Mean and standard deviation (in brackets) of CCC values of SER models under matched and mismatched label conditions. We denote the MSP-Podcast SL labels as *SL-ORG* and the MSP-Conversation aggregated SL labels as *SL-AGG*. We indicate the aggregation method and time shift in brackets. Results tagged with $*$ indicate that the average CCC values across 10 trials for the mismatched condition are statistically significantly better than the results for the matched condition (two-tailed t-test, p-value < 0.05). Results tagged with † indicate the matched condition results in statistically significantly better performance than the mismatched condition.

		Model Train Set	
Model Test Set	**Arousal**	*SL-ORG*	*SL-AGG (T=2.8, Q2)*
	SL-ORG	0.531 (0.031)	0.509 (0.015)
	SL-AGG (T=2.8, Q2)	0.548 (0.025)	0.510 (0.013)
	Valence	*SL-ORG*	*SL-AGG (T=0, M)*
	SL-ORG	0.088 (0.018)	0.051 (0.014)
	SL-AGG (T=0, M)	$0.145^{\dagger}$(0.051)	0.185^{*}(0.074)
	Dominance	*SL-ORG*	*SL-AGG (T=2.8, Q2)*
	SL-ORG	0.483^{*}(0.017)	0.378 (0.033)
	SL-AGG (T=2.8, Q2)	0.215 (0.016)	0.351 (0.024)

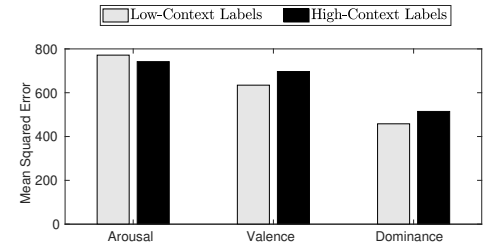
labels. However, those connections appear to hold better on the aggregated SL labels. For dominance, the models tested with the original MSP-Podcast SL labels have the best performance. The best model is when we train and test with the original SL labels. Only the result of the models trained with the original SL labels has a statistically significant difference compared to the mismatched condition. The models trained with the aggregated SL labels have very similar performance when tested with both types of labels, showing that the labels have some similarities. For dominance, the connections learned by the model hold best for the original SL labels, although the performance differences are not as large as for valance. Dominance values are the most similar in value between the datasets (Figure 5), so it is natural that the SER dominance results are good in mismatched conditions even with low correlations (Figure 6).

While the aggregated SL and original SL labels have clear differences, the overall results in this section indicate that they still share important information relevant to SER tasks, especially for arousal labels. The low standard deviation in the CCC values across the 10 trials (Table 3) shows that the results are consistent, validating the observations discussed in this section.

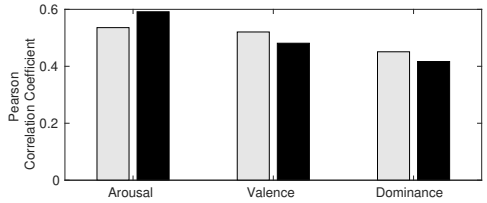
6 Context Analysis

Our experiments in Sections 4 and 5 show that there are differences between the aggregated SL labels and the SL labels, especially for the attribute of valence. These differences could be due to many factors in the distinct annotation processes, including the extra context that is available when annotating CT labels (i.e., SL labels in the MSP-Podcast corpus are annotated out-of-context). This section analyzes in more detail the role of context.

We compare sentences annotated at different context levels during the CT annotation process. We divide our sentences into two groups. The first group includes sentences that happen in the first 60 seconds of a conversation, which are referred to as *low-context* sentences. The second group includes sentences that occur after three minutes in a conversation, which are referred to as *high-context* sentences.



(a) Average Mean Squared Error



(b) Average Pearson Correlation Coefficient

Fig. 9. Statistical measures between the aggregated SL labels (MSP-Conversation) and original SL labels (MSP-Podcast) for low-context and high-context sentences. We only report the results with the mean aggregation method.

TABLE 4
CCC values of SER models under matched and mismatched label conditions trained and evaluated on low-context or high-context sentences. We denote the MSP-Podcast original SL labels as *SL-ORG* and the MSP-Conversation aggregated SL labels as *SL-AGG*.

Low-Context Sentences			
Model Test Set	Model Train Set		
	Arousal	*SL-ORG*	*SL-AGG (T=2.8, Q2)*
	SL-ORG	0.492	0.484
	SL-AGG (T=2.8, Q2)	0.569	0.470
	Valence	*SL-ORG*	*SL-AGG (T=0, M)*
	SL-ORG	0.008	0.048
	SL-AGG (T=0, M)	0.104	0.156
	Dominance	*SL-ORG*	*SL-AGG (T=2.8, Q2)*
	SL-ORG	0.557	0.494
	SL-AGG (T=2.8, Q2)	0.413	0.417
High-Context Sentences			
Model Test Set	Model Train Set		
	Arousal	*SL-ORG*	*SL-AGG (T=2.8, Q2)*
	SL-ORG	0.502 (0.044)	0.421 (0.043)
	SL-AGG (T=2.8, Q2)	0.469 (0.033)	0.443 (0.039)
	Valence	*SL-ORG*	*SL-AGG (T=0, M)*
	SL-ORG	0.045 (0.072)	0.022 (0.095)
	SL-AGG (T=0, M)	0.182 (0.108)	0.261 (0.092)
	Dominance	*SL-ORG*	*SL-AGG (T=2.8, Q2)*
	SL-ORG	0.369 (0.017)	0.307 (0.041)
	SL-AGG (T=2.8, Q2)	0.201 (0.051)	0.241 (0.077)

Comparing the labels of these two groups can give us insights into the role of context and the role of the continuous annotation process in our results. There are 514 low-context sentences, so to perform the analyses, we randomly select 514 high-context sentences to effectively compare the two groups.

6.1 Statistical Analysis

We perform similar experiments to Section 4.1 using low-context and high-context sentences on the labels without normalization. Figures 9(a) and 9(b) show the results of the statistical analyses using the mean for aggregation of the CT labels. The MSE and ρ values are averaged over all the time shifts considered in this study. There is a slightly bigger difference between the CT and SL labels for valence and dominance for the high-context sentences as opposed to the low-context sentences. This result is expected since the SL labels are conducted without context matching the low-context setting. For arousal, however, context decreases the differences between aggregated and original SL labels. The similarity between the CT and SL labels is higher for the high-context sentences than for the low-context sentences. The results in this section show slight differences in the patterns for aggregated and original SL labels in low-context and high-context sentences. However, the aggregated and original SL labels are still different in both conditions, so we conclude that the differences cannot be exclusively attributed to context.

6.2 Evaluation with SER Models

We replicate SER experiments from Section 5.2 on the low-context and high-context sentences. Given the reduced size of the corpus, the experiment results are based on the PCA-SVM method mentioned in Section 5.1. We assign sentences to the train, development, and test partitions based on the original partition of the MSP-Podcast corpus. To ensure that our results are not affected by the choice of high-context sentences, we repeat the random sampling process of the high-context set five times, reporting the average results. Since there are only 514 low-context sentences, we only have one repetition for the low-context set. The number of sentences is too low to perform multiple samples and train an SER model.

Table 4 shows our results. If context plays a role in the performance for matched and mismatched conditions, we would observe a smaller percentage difference in the low context setting compared to the high context setting in Table 4. For arousal, we observe that models trained on the low-context sentences show slight CCC improvements from matched to mismatched conditions (from 0.492 to 0.569 when trained with original SL labels; from 0.470 to 0.484 when trained with aggregated SL labels). The models trained on high-context sentences show the opposite trends, with slightly worse performance going from matched to mismatched conditions (from 0.502 to 0.469 when trained with original SL labels; from 0.443 to 0.421 when trained with aggregated SL labels). These results suggest that aggregated SL labels for arousal are better for using in conjunction with SL labels for SER tasks when they are conducted in a low-context setting. Context seems to make the CT labels diverge from SL labels for SER tasks, which is expected since SL labels are conducted out-of-context. However, the results in Tables 2 and 3 show that the two types of labels can be successfully used together in these tasks. Although context might change the labels slightly, these changes are not large enough to have a drastic effect on the results.

For dominance, the low and high-context results support previous results that show lower CCC results for models trained and/or tested with the aggregated SL labels. We conducted two-tailed paired t-tests (p-value < 0.05) for each attribute to check if the differences between the low and high-context results are statistically significant. Only

the dominance results show a statistically significant difference between the two context settings, where the low-context setting has significantly better results. However, when comparing matched and mismatched results for each dominance model, we see that the differences between the two testing results are similar for both context settings. The percentage differences for the models trained using the original SL labels are 29.7% (0.557 versus 0.413 CCC values) and 16.9% (0.369 versus 0.201 CCC values) for the low and high-context settings, respectively. For the models trained using the aggregated SL labels, the percentage differences are 16.9% (0.494 versus 0.417 CCC values) and 24.1% (0.307 versus 0.241 CCC values), respectively. The addition of context to the aggregated SL labels does not seem to affect how much worse the models perform when using mismatched dominance labels. Ultimately, context seems to have a minimal effect on the differences between the aggregated SL and SL labels for dominance in SER tasks.

For valence, the low-context and high-context results show similar trends to previous results in Tables 2 and 3. Testing with the MSP-Conversation SL labels results in much better predictions. The high-context results show better CCC values than the low-context results as well. These experiments suggest that valence results could be improved by using in-context labels, as suggested by Lee *et al.* [45] which showed that the addition of contextual information to valence models increases accuracy significantly more than for arousal. However, the addition of context in the aggregated SL labels does not affect the performance differences between the models tested with matched and mismatched labels. Similar to dominance, the percentage differences between the matched and mismatched results for each model are similar between the two context settings. For the models trained with the original SL labels, the percentage differences are 171% (0.008 versus 0.104 CCC values) and 121% (0.045 versus 0.182 CCC values) for the low and high-context settings, respectively. For the models trained with the aggregated SL labels, the percentage differences are 106% (0.048 versus 0.156 CCC values) and 169% (0.022 versus 0.261 CCC values), respectively. The still large difference between the prediction performances of models tested with the original SL valence labels and the low-context aggregated SL valence labels implies that the differences between the types of labels are due to more than just context. These preliminary experiments point to the idea that CT annotations add information, through context and other aspects, that is lost when conducting out-of-context SL annotations for valence.

The addition or lack of context seems to affect the attribute of arousal the most, but for both valence and dominance, there seems to be an aspect innate to CT annotations independent of context that makes the aggregated SL labels and the original SL labels different for SER tasks. Cowie *et al.* [6] discuss the unique aspects of CT annotations, the main one being the instantaneous reactions that are captured and cannot be extracted from more thoughtful annotations. Furthermore, Cowie *et al.* [6] argued that emotional attributes that can be numerically captured can also capture more instantaneous reactions as opposed to categorical emotions. Putting emotions into language changes the emotional state of humans in a way that does not happen when just experiencing the emotions [60]. Although both types of annotations analyzed in our experiments use dimensional labels, the SL annotations from the MSP-Podcast corpus are captured using a questionnaire. The questionnaire not only gives an annotator more time for a thoughtful annotation, but also asks raters to label sentences with an emotional category (putting emotion into language). These differences do not seem to affect the attribute of arousal, but they do affect valence and dominance, affecting valence at a much higher rate for our SER task.

The instantaneous reactions captured by the CT annotations are beneficial for the prediction of valence in SER tasks as well as the context added. Thoughtful and out-of-context annotations seem to not only make raters diverge in their ratings of valence (Section 4.2), but also detangle the labels from the speech features in a way that makes them harder to predict with simple SER models (Section 5). Dominance SER models seem to have the opposite trend, where thoughtful annotations are better. Further research on the rating of emotion at the personal level is needed to explain these phenomena, but these results do indicate important considerations for the field. As mentioned previously, valence is historically harder to predict than the other attributes. The role of context in increasing valence predictability has been previously observed and used [45], but the role of instantaneous reactions has not. The combination of both aspects in modeling and data collection could significantly help valence SER predictions.

7 Discussion

The comparisons made between the CT and SL labels are informative, even when one corpus was collected in context and the other was collected out of context. From a practical viewpoint, out-of-context SL labels make up a significant part of the emotional annotations used in the speech emotion recognition field. Therefore, much of the infrastructure and models are based on such labels. Furthermore, many new datasets are including CT labels [8], [10], [11], [12], which are expensive and time-consuming to obtain. If the field can find a way to use the new CT labels in the same framework as the previously used SL labels, then the field can consolidate much of its resources and not start from scratch when it comes to using the CT labels. Our results show that both types of labels can be used together, which is an encouraging outcome, since that suggests that current resources and models used for SL labels can be modified for use with CT labels with less time and effort than if they were widely different. Furthermore, several studies blindly combine databases collected with different labels [61], [62], including databases collected with CT and SL annotations [63]. Our study is instrumental to determine the validity of this approach.

Our stated goal introduced in Section 1 included four questions we wanted to answer with our analysis. To answer Questions 1 and 4, we evaluated different aggregation functions used to aggregate the CT labels into SL labels. We compared the original SL (MSP-Podcast corpus) and aggregated SL (MSP-Conversation corpus) labels using the MSE and correlation metrics. We also compare the labels after normalizing both types of labels. The results showed that

the labels are correlated. When analyzing the aggregation functions, we observe that using the mean or median as aggregation functions results in aggregated SL labels that approximate the original SL labels the best. This suggests that annotators use average emotional values to make their SL judgements as opposed to the more extreme values.

The inter-evaluator agreements for the original SL annotations (MSP-Podcast corpus) and for the SL annotations derived from CT labels (MSP-Conversation corpus) were calculated to answer Questions 2 and 4. The results for the aggregated SL labels showed much lower agreement for arousal and dominance, as expected, given the larger number of choices for values in the CT annotations. Surprisingly, the aggregated SL labels for valence had the highest agreement, even above the original SL labels, showing how the two types of labels affect emotional attributes differently.

The analysis also explored the effect of using either the aggregated SL annotations or the original SL annotations in SER tasks to respond to Questions 3 and 4. We trained SER models, comparing their performances when trained and tested with both types of SL annotations. Like the inter-evaluator agreement results, the arousal and dominance models had lower performance when using the aggregated MSP-Conversation SL labels. The valence models had greater performance when using the aggregated MSP-Conversation SL labels for training. Collectively, the results of the SER models and the inter-evaluator agreement show that the emotional attribute of valence gains the most when SL labels are created from CT annotations. The results in mismatched conditions, when training with labels from one corpus and testing with labels from the other corpus, led to very similar results for the attribute of arousal, mixed results for dominance, and significantly different results for valence. The mismatched results highlight that the two types of labels have similarities in the information relevant to an SER model, so aggregating CT labels to make SL labels is valid for simple SER tasks.

We conducted further analyses to understand the role of context. In particular, we compared the MSP-Conversation SL labels and MSP-Podcast SL labels for sentences with low and high-context during the MSP-Conversation annotations. The results showed that although the context does play a role in the differences between the two types of labels, there are features unique to CT annotations that also affect the labels. The differences in our simple SER task are more drastic for valence and dominance labels, suggesting that context, instantaneous reactions, and thoughtful annotations affect these SER tasks the most.

Overall, the results from this study indicate that aggregated SL labels obtained from CT labels can be a valid approach, since they are correlated with the original SL labels, have comparable agreements, and can create SER models with similar performances. The statistical experiments point to the mean and median, with a time shift, being sufficient for CT labels approximating SL labels. However, when adding the agreement and modeling results, we see that there are clear aggregation methods that are better for each attribute. The results across time shifts are more consistent. We recommend that the field use multiple aggregation methods when evaluating models using aggregated SL labels, with a constant time shift also being used. This time shift can be three seconds, recommended in previous literature [39], or tuned to the specific dataset or annotator. The aggregation method can also be tuned to the specific task by using a small subset of the dataset. However, it is important to be aware that the two types of labels have differences, as they are not interchangeable in all cases using the aggregation methods used in this study. Comparing models using both types of annotations can be done, but the comparison is not as valid as comparing models using only SL labels since there are performance differences (using one type of label will not always lead to lower performance) depending on the attribute. These biases can be considered when comparing models. The results of the analysis could also be used to develop new aggregation methods that can better approximate original SL labels. We observed that the normalization strategy used in Section 4.1 was successful in removing the shift between the labels, indicating that more sophisticated normalization methods can be effective for this task. Moreover, our analysis shows that the attribute of valence, which is famously the lowest performing attribute in SER models [57], [59], gains the most from the CT labels. We observed that the reason for the higher performance is a mix between the added context and instantaneous reactions of CT annotations. Further research in leveraging these aspects of CT labels could be extremely helpful in closing the gap between valence and the other attributes.

8 Conclusions

This paper evaluated the approach of aggregating SL annotations from CT annotations, which is a common approach used in emotion recognition tasks. The analysis leveraged the annotations of two publicly available databases that provide the perfect resource to study the validity of this approach. We used the CT annotations of the MSP-Conversation corpus [26], and the SL annotations of the MSP-Podcast corpus [4], focusing on 2,884 speech segments that are present in both corpora. The CT labels of the MSP-Conversation corpus were aggregated over the target speech segments to create SL annotations, which were compared with the original SL annotations from the MSP-Podcast corpus. Comparing the SL labels from both corpora indicated that although they are clearly different in absolute values, they are definitely correlated and can be used together for various tasks. With further developments in aggregation approaches, they could be used interchangeably. We left as future work the analysis of when the labels are interchangeable or when using one type of label over the other is best. Our analysis supports the idea that for the emotional dimension of valence, using SL labels derived from CT annotations leads to better performance. However, this is not the case for arousal and dominance. Therefore, further work needs to be done to validate these observations.

There are several implications and recommendations that can be drawn from this study. We recommend that both types of annotations, CT and SL, continue to be collected to aid in the development and testing of models. We recommend using aggregated SL labels by tuning the aggregation over time and taking into account the reaction time of the annotators by including a constant time shift, although careful considerations of the differences noted

in this paper should be mentioned. Our analysis of the role of context in the differences between the two types of labels suggests that context and the differences between instantaneous and thoughtful annotations both play a role in our results. Considering how the two annotation styles differ could make the labels more valuable for specific tasks. For example, if the ultimate usage of an SER model is to engage a human in a game or a conversation, then labels that capture instantaneous reactions (CT annotations) could lead to better engagement [64]. On the other hand, if the application calls for more thoughtful considerations of human emotions to educate or calm a person, then SL annotations that require putting emotions into words could be more useful.

In our future work, we will include more speech data to compare SL and CT annotations. We only had 2,884 speech segments to use for this analysis, which is limited especially for performing automatic recognition analyses. The data collection of both corpora are ongoing efforts, so the expectation is that new releases of these corpora will increase the number of speech segments contained in both databases. This study only considers emotional attributes. Some emotional corpora are annotated with CT annotations for emotional categories (e.g., the SEMAINE database [9]). A similar analysis may be conducted for categorical CT labels. Furthermore, our analysis explores out-of-context SL annotations and typical CT annotations, which does not completely isolate the differences between the two types. Collecting additional in-context SL annotations and comparing all three types of annotations would allow us to explore which aspects of the annotation process affect the differences between them and give us more insight into how humans perceive emotion in speech.

ACKNOWLEDGMENTS

This study was funded by the National Science Foundation (NSF) under grants CNS-2016719 and CNS-1823166.

REFERENCES

[1] C. Busso, M. Bulut, and S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds. New York, NY, USA: Oxford University Press, November 2013, pp. 110–127.

[2] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[3] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *IEEE International Conference on Multimedia and Expo (ICME 2008)*, Hannover, Germany, June 2008, pp. 865–868.

[4] 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, October-December 2019.

[5] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, October-December 2014.

[6] R. Cowie, G. McKeown, and E. Douglas-Cowie, "Tracing emotion: An overview," *International Journal of Synthetic Emotions*, vol. 3, no. 1, pp. 1–17, January-June 2012.

[7] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Newcastle, Northern Ireland, UK: ISCA, September 2000, pp. 19–24.

[8] A. Metallinou, Z. Yang, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations," *Journal of Language Resources and Evaluation*, vol. 50, no. 3, pp. 497–521, September 2016.

[9] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, January-March 2012.

[10] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, Shanghai, China, April 2013, pp. 1–8.

[11] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, K. Star, E. Hajiyev, and M. Pantic, "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[12] L. Stappen, A. Baird, G. Rizos, P. Tzirakis, X. Du, F. Hafner, L. Schumann, A. Mallol-Ragolta, B. Schuller, I. Lefter, E. Cambria, and I. Kompatsiaris, "MuSe 2020 – the first international multimodal sentiment analysis in real-life media challenge and workshop," in *The Multimodal Sentiment in Real-life Media Challenge (MuSe 2020)*, Seattle, US, October 2020.

[13] J. Arias, C. Busso, and N. Yoma, "Energy and F0 contour modeling with functional data analysis for emotional speech detection," in *Interspeech 2013*, Lyon, France, August 2013, pp. 2871–2875.

[14] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011- the first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction (ACII 2011)*, ser. Lecture Notes in Computer Science, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Memphis, TN, USA: Springer Berlin / Heidelberg, October 2011, vol. 6975/2011, pp. 415–424.

[15] R. Lotfian and C. Busso, "Emotion recognition using synthetic speech as neutral reference," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015)*, Brisbane, Australia, April 2015, pp. 4759–4763.

[16] ——, "Practical considerations on the use of preference learning for ranking emotional speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5205–5209.

[17] M. Wöllmer, A. Metallinou, N. Katsamanis, B. Schuller, and S. Narayanan, "Analyzing the memory of BLSTM neural networks for enhanced emotion classification in dyadic spoken interactions," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012, pp. 4157–4160.

[18] S. Mariooryad and C. Busso, "The cost of dichotomizing continuous labels for binary classification problems: Deriving a Bayesian-optimal classifier," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 119–130, January-March 2017.

[19] S. Parthasarathy and C. Busso, "Predicting emotionally salient regions using qualitative agreement of deep neural network regressors," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 402–416, April-June 2021.

[20] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015)*, Brisbane, Australia, April 2015, pp. 5058–5062.

[21] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, Hong Kong, China, November 2019, pp. 154–164.

[22] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6818–6825.

[23] D. Sheng, D. Wang, Y. Shen, H. Zheng, and H. Liu, "Summarize before aggregate: A global-to-local heterogeneous graph inference network for conversational emotion recognition," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), December 2020, pp. 4153–4163.

[24] W. Li, W. Shao, S. Ji, and E. Cambria, "Bieru: Bidirectional emotional recurrent unit for conversational sentiment analysis," *Neurocomputing*, vol. 467, pp. 73–82, 2022.

[25] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, Shanghai, China, April 2013, pp. 1–8.

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

[27] J. Russell and L. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant," *Journal of personality and social psychology*, vol. 76, no. 5, pp. 805–819, May 1999.

[28] J. Russell, "Core affect and the psychological construction of emotion," *Psychological review*, vol. 110, no. 1, pp. 145–172, January 2003.

[29] J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, December 2007.

[30] R. Cauldwell, "Where did the anger go? the role of context in interpreting emotion in speech," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, September 2000, pp. 127–131.

[31] M. Jaiswal, Z. Aldeneh, C.-P. Bara, Y. Luo, M. Burzo, R. Mihalcea, and E. Mower Provost, "MuSE-ing on the impact of utterance ordering on crowdsourced emotion annotations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, Brighton, United Kingdom, May 2019, pp. 7415–7419.

[32] 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.

[33] 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.

[34] J. M. Girard, "CARMA: Software for continuous affect rating and media annotation," *Journal of Open Research Software*, vol. 2, no. 1, pp. 1–6, July 2014.

[35] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 1517–1520.

[36] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Sixth International Conference on Multimodal Interfaces ICMI 2004*. State College, PA: ACM Press, October 2004, pp. 205–211.

[37] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower 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, January-March 2017.

[38] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *Affective Computing and Intelligent Interaction (ACII 2013)*, Geneva, Switzerland, September 2013, pp. 85–90.

[39] ——, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, April-June 2015, special Issue Best of ACII.

[40] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *International Workshop on Audio/Visual Emotion Challenge*, Amsterdam, The Netherlands, October 2016, pp. 3–10.

[41] G. A. Miller, "The magical number seven, plus or minus two: some limits on our capacity for processing information," *Psychological Review*, vol. 63, no. 2, pp. 81–97, March 1956.

[42] G. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 16–35, January-March 2021.

[43] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative feature normalization scheme for automatic emotion detection from speech," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 386–397, October-December 2013.

[44] K. Krippendorff, "Bivariate agreement coefficients for reliability of data," *Sociological methodology*, vol. 2, pp. 139–150, 1970.

[45] C.-C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 1983–1986.

[46] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane, "Gender de-biasing in speech emotion recognition," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2823–2827.

[47] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 815–826, April 2019.

[48] 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, September 2020.

[49] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *ArXiv e-prints (arXiv:2203.07378)*, pp. 1–25, March 2022.

[50] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3698–3702.

[51] W.-C. Lin and C. Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," *IEEE Transactions on Affective Computing*, vol. Early Access, 2022.

[52] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*. Calgary, AB, Canada: IEEE, April 2018, pp. 5084–5088.

[53] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[54] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[55] K. Sridhar, S. Parthasarathy, and C. Busso, "Role of regularization in the prediction of valence from speech," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 941–945.

[56] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5200–5204.

[57] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 1179–1182.

[58] W.-C. Lin, K. Sridhar, and C. Busso, "DeepEmoCluster: A semi-supervised framework for latent cluster representation of speech emotions," in *IEEE international conference on acoustics, speech and signal processing (ICASSP 2021)*, Toronto, ON, Canada, June 2021, pp. 7263–7267.

[59] K. Sridhar and C. Busso, "Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 1959–1972, October-December 2022.

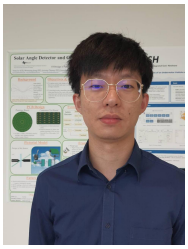
[60] J. Pennebaker and C. Chung, "Expressive writing: Connections to physical and mental health," *The Oxford Handbook of Health Psychology*, vol. H.S. Friedman, pp. 417–437, August 2011.

[61] B. Schuller, Z. Zhang, F. Weninger, and F. Burkhardt, "Synthesized speech for model training in cross-corpus recognition of human

emotion," *International Journal of Speech Technology*, vol. 15, no. 3, pp. 313–323, September 2012.
[62] H. Sagha, J. Deng, M. Gavryukova, J. Han, and B. Schuller, "Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5800–5804.
[63] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *IEEE international conference on acoustics, speech and signal processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 5688–5691.
[64] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, January 2001.

Carlos Busso (S'02-M'09-SM'13-F'23) 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 (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is an associate professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He was selected by the School of Engineering of Chile as the best electrical engineer graduated in 2003 across Chilean universities. At USC, he received a provost doctoral fellowship from 2003 to 2005 and a fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory [http://msp.utdallas.edu]. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. In 2015, his student received the third prize IEEE ITSS Best Dissertation Award (N. Li). 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 in 2022 (with Yannakakis and Cowie). He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interest is in human-centered multimodal machine intelligence and applications. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, nonverbal behaviors for conversational agents, in-vehicle active safety system, and machine learning methods for multimodal processing. His work has direct implication in many practical domains, including national security, health care, entertainment, transportation systems, and education. He was the general chair of ACII 2017 and ICMI 2021. He is a IEEE Fellow. He is a member of ISCA, AAAC, and a senior member of ACM.

Luz Martinez-Lucas (S'21) is a PhD Student in the Electrical and Computer Engineering Department at the University of Texas at Dallas (UTD). She did her Bachelor's in Electrical Engineering at UTD. Her research interests include affective computing, speech technology, and machine learning. She is a student member of IEEE and SIAM.



Wei-Cheng Lin (S'16) currently is a PhD student at Electrical and Computer Engineering Department of The University of Texas at Dallas (UTD). He received his B.S. degree in communication engineering from the National Taiwan Ocean University (NTOU), Taiwan in 2014 and M.S. degree in electrical engineering from the National Tsing Hua University (NTHU), Taiwan in 2016. His research interests is 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).