Affective Priming in Emotional Annotations and its Effect on Speech Emotion Recognition

Luz Martinez-Lucas, *Student Member, IEEE,* Ali Salman, *Student Member, IEEE,* Seong-Gyun Leem, *Student Member, IEEE,* Woan-Shiuan Chien, *Student Member, IEEE,* Shreya G. Upadhyay, *Student Member, IEEE,* Chi-Chun Lee, *Senior Member, IEEE,* Carlos Busso, *Fellow, IEEE,*

***Abstract*—Emotional annotation of data is important in affective computing for the analysis, recognition, and synthesis of emotions. As raters perceive emotion, they make relative comparisons with what they previously experienced, creating "anchors" that influence the annotations. This unconscious influence of the emotional content of previous stimuli in the perception of emotions is referred to as the *affective priming effect*. This phenomenon is also expected in annotations conducted with out-of-order segments, a common approach for annotating emotional databases. Can the affective priming effect introduce bias in the labels? If yes, how does this bias affect emotion recognition systems trained with these labels? This study presents a detailed analysis of the affective priming effect and its influence on *speech emotion recognition* (SER). The analysis shows that the affective priming effect affects emotional attributes and categorical emotion annotations. We observe that if annotators assign an extreme score to previous sentences for an emotional attribute (valence, arousal, or dominance), they will tend to annotate the next sentence closer to that extreme. We conduct SER experiments using the most biased sentences. We observe that models trained on the biased sentences perform the best and have the lowest prediction uncertainty.**

***Index Terms*—Affective Computing, Emotional Annotations, Affective Priming, Emotional Attributes, Speech Emotion Recognition**

I. Introduction

The field of affective computing relies on emotional labels. These labels are often obtained from perceptual evaluations, where people give their emotional perceptions after listening to an audio or watching a video. Emotional perception is subjective and hard to describe. Therefore, assessing the reliability of emotional labels is important. In affective computing, reliability is measured by calculating the inter-evaluator agreement between all annotators that evaluated the stimuli [1], [2]. The higher the agreement, the higher the reliability. However, this method does not account for any potential issues with the annotation method itself. A common annotation method involves a rater sequentially evaluating many samples in a session, providing absolute categorical or attribute scores [3]–[6]. A less common method is annotating the samples in an ordinal manner, where a rater is asked to rank samples on an emotional scale or compare preferences between samples with respect to an emotional dimension (e.g., which sample is *happier*?). Yannakakis *et al.* [7], [8] observed that it is easier for people to record their emotional perception in that way, suggesting that those ordinal labels can better represent the person's perception. The hypothesis is that humans use an "anchor" when evaluating any emotional content, so when an explicit "anchor" is provided, a rater can provide more reliable labels. This hypothesis leads to some important questions about the sequential method of emotional annotations: do these annotators "anchor" to the samples previously annotated? Can we quantify this "anchoring" effect? Does this "anchoring" effect affect emotional tasks?

The effect of preceding stimuli on the emotional assessment of a current sample is known as affective priming [9]. In general, the priming effect refers to the observation that a *prime* (i.e., a preceding stimulus) can unconsciously influence a person's response to a *target*/subsequent stimulus, when there is some form of relationship between the two stimuli. Priming has been observed in the study of memory, where it is often demonstrated that target words are recalled faster when preceded by perceptually or conceptually related primes [10]. This phenomenon has also been observed in the human perception of speech. Bosker [11] showed that a target ambiguous speech segment could be perceived as one of two words depending on what audio the person was primed with. Affective priming refers to this priming phenomenon when the prime and target are related in the emotional space [9]. We focus on answering the following questions:

1) Does the affective priming effect occur in emotional attribute labeling of speech?
2) Is the affective priming effect observed in categorical labels?
3) Does the affective priming effect impact *speech emotion recognition* (SER) performance?

We analyze the affective priming effect on ratings of emotional attributes and emotional categories. For emotional attributes, we consider arousal (calm versus active), valence (negative versus positive), and dominance (weak versus strong)

L. Martinez-Lucas is with the Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas, Richardson, TX 75080 and a visitor at the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 E-mail: luz.martinez-lucas@utdallas.edu

A. Salman and S.-G. Leem are with the Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas, Richardson, TX 75080 E-mail: ali.salman@utdallas.edu and seong.leem@utdallas.edu

W.-S. Chien, S. G. Upadhyay, and C.-C. Lee are with the Department of Electrical Engineering, National Tsing Hua University, Taiwan E-mail: wschien@gapp.nthu.edu.tw, shreya@gapp.nthu.edu.tw, and cclee@ee.nthu.edu.tw

C. Busso is with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 E-mail: busso@cmu.edu

This study was funded by the National Science Foundation (Grant CNS-2016719), and the National Science and Technology Council, Taiwan (Grant: 110-2221-E-007-067-MY3, 111-2634-F-002-023)

[12]–[14]. For emotional categories, we consider the emotional classes of happiness, anger, sadness, and neutral state. Our study utilizes two publicly available databases to explore this effect: the MSP-Podcast corpus [5] and the BIIC-Podcast corpus [15]. The MSP-Podcast corpus contains 890,851 annotations of English sentences collected over 71,696 sessions (release 1.11). The large size of this corpus provides a multitude of examples to investigate the affective priming effect. The BIIC-Podcast corpus contains 147,949 annotations of sentences in Mandarin collected over 898 sessions (release 1.00). Although this database is smaller than the MSP-Podcast corpus, it allows us to validate the observed effects of affective priming in a different setting. For both databases, our approach involves comparing the label provided by an evaluator for a sentence to the average label of the other annotators who rated that sentence. We denote this average as the *expected label*. We then take the difference between the evaluator label and the expected label and condition it on the emotional scores of the previous sentences rated by the current evaluator. In our preliminary work [16], we explored how affective priming affects emotional annotations of speech during out-of-context sequential evaluations (i.e., annotators listen to unrelated emotional speech before rating new speech). This paper corroborates and extends our study on the affective priming effect by considering multiple databases, adding emotional categories in the analysis, and quantifying uncertainty to better understand the results.

Our analyses show that the affective priming effect is indeed observed in our databases' attribute annotations (Q1). When an annotator rates a sentence in one extreme of an emotional attribute, they will rate the next sentence closer to that extreme when compared to the expected label. Similarly, when an annotator rates a sentence with an emotional class, they are more likely to rate the next sentence with the same emotional class than the expected class obtained from other annotators (Q2). We observe that when annotators are primed with neutral sentences, i.e., they rated preceding sentences with neutral values for the emotional attributes, their attribute ratings are closer to the expected label (Q1). We see that affective priming results in a bias in the labels that is higher when the primes are more extreme, e.g., more extreme attribute scores or a consistent emotional class (Q1 and Q2). Therefore, we estimate the affective priming bias expected for each sentence in our dataset by looking at the emotional priming the raters for each sentence experienced during the annotation process. We then separate the sentences into groups with different levels of affective priming bias and perform SER modeling experiments on the groups (Q3). We see that models trained on labels affected by the affective priming effect perform best. These models also have lower variance in their performance metrics, suggesting a lower uncertainty at testing time. To further analyze model uncertainty, we use Monte Carlo Dropout to test the models. These uncertainty analyses support the idea that our performance results are due to the low uncertainty of the model when predicting labels affected by affective priming. We further analyze the performance of our models for different label values, showing that the higher performance of models trained with labels affected by affective priming is due to performance differences for extreme emotional labels.

II. Related Work

A. Contextual Influence In The Perception of Emotions

The idea that surrounding speech affects the perception or annotation of emotional speech has been explored in previous work. Cauldwell [17] observed that in some instances, a sentence was rated as angry in isolation but not angry when the sentence was heard within the full conversation. Jaiswal *et al.* [18] explored the differences between emotional annotations of sentences conducted when the sentences were in random order (out-of-context) and in the order they were spoken (in-context). They observed that in-context annotations were more similar to the self-reported emotions of the speakers. However, the out-of-context annotations were more easily predicted by a SER model. Overall, these studies demonstrate that including contextual information during the annotation process changes the emotional ratings enough to have an effect on SER results. However, our analysis in this paper does not focus on context. The databases we analyze give random "context" to the annotators, meaning when they rate a sentence, the next and previous sentences are probably unrelated to the current one. Can the presence and the emotional content of out-of-context sentences also affect annotations?

In the field of preference learning, the comparison of consecutive samples is often used to create ordinal labels [8]. These ordinal labels contain information about how the emotional value of one sentence compares to another. Therefore, they can be obtained when sentences are directly compared, creating explicit "anchors." In the field, there is an expectation that when annotators rate a sentence after another sentence, they will use the previous sentence to anchor their emotional rating [8]. Anchoring is a phenomenon where people are likely to give relative rather than absolute value judgments in uncertain situations [19], where these relative judgments are anchored to some stimuli with a given value. Considering the subjective and uncertain process of rating emotion, annotators are likely to anchor their value judgments to some previous stimuli to which they have already given a value. When we rely on common annotations collected in sequence during a session, the "anchors" are not explicit, leading to important questions on the validity of the labels.

Affective priming is a phenomenon where the emotional perception of a sample (audio, video, word) is affected by the emotional content of a previous stimulus [9], [20]. Studies on this phenomenon have found that priming with a clear emotional stimulus will push a person's valence perception of an ambiguous sample towards the valence value of the stimulus [9], [21], [22]. This effect has been studied mostly in emotional ratings of words [20] and images [21], [23]. Studies have mostly focused on the emotional attribute of valence. If the emotional priming effect occurs in the emotional annotations of speech, then the "anchors" that annotators use to rate sentences are also priming stimuli. Therefore, we expect that these anchors are not only used as a starting point for the emotional rating of the target sentences, but also alter the actual emotional perception of those sentences. In this paper,

we focus on exploring if the affective priming effect exists in the speech domain for the emotional attributes of arousal, valence, and dominance and emotional categories. We further evaluate how the resulting bias affects SER models by conducting various modeling experiments using annotations with large biases towards an emotional extreme or an emotional category.

B. Relation to Previous Work

This study is an extension of our preliminary work analyzing the affective priming effect [16]. Our previous paper showed that there was an affective priming effect on the emotional attribute annotations, conducting the evaluation on the MSP-Podcast corpus. This paper explores the same effect on a newer version of the corpus as well as on a different corpus, the BIIC-Podcast corpus. We show that the previously observed effect is consistently seen in both corpora. Furthermore, this paper also explores the affective priming effect on categorical labels in both corpora. In our previous work, we also conducted attribute-based SER modeling experiments on subsets of the MSP-Podcast corpus affected by affective priming. We conduct the experiments again using a larger amount of data to validate our previous findings. We also extend our modeling experiments to the prediction of the emotional category of each sentence. Finally, we aim to further explain our previous and current SER modeling results by studying the uncertainty of our SER models.

III. Resources

A. The MSP-Podcast Corpus, Version 1.11

The MSP-Podcast database [5] is a collection of emotionally diverse audio files sourced from audio and video sharing websites. This dataset is part of an ongoing project. In this study, we utilize version 1.11. The dataset contains 151,654 short audio files, each lasting between 3 to 11 seconds, and totals almost 238 hours. Every audio file in the dataset has been annotated by at least five annotators for both categorical emotions (e.g., anger, sadness, happiness, surprise, fear, disgust, contempt, and neutral state) and emotional attributes (e.g., valence, arousal, and dominance), using a 7-point Likert scale. Arousal ranges from calm to active, valence from negative to positive, and dominance from weak to strong. The MSP-Podcast corpus contains annotations partially completed by Amazon Mechanical Turk workers and UTD students. For both cases, we used an annotation website developed by our laboratory. During the annotation process, annotators work through a sequence of audio files during a session. An important aspect of this process is that the samples in the sequence and the order of the sequence each annotator rates are randomly chosen from a pool of audio files. Therefore, it is highly unlikely that the annotators were exposed to the same preceding samples when annotating a specific sample. For our experiments, we relied on the predefined splits in the dataset, which include training and development sets, as well as three distinct test sets. This study only utilizes Test 1.

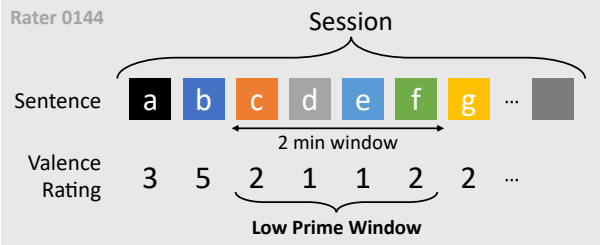


Fig. 1. Selection of prime window for the valence annotation of sentence "g" by rater 0144. The prime window of an annotation contains the annotations completed in the 2 minutes preceding the current annotation. In this example, the prime window is considered low since it only contains valence annotations with values of 1 and 2.

B. The BIIC-Podcast Corpus, Version 1.00

The BIIC-Podcast database [15] is a speech corpus for SER research in Taiwanese Mandarin, which mirrors the data collection techniques employed by the MSP-Podcast corpus. It is compiled from a variety of audio-sharing platforms, encompassing 157 hours of podcast speech samples. This corpus leverages the capabilities of *Label Studio* [24] to collect emotional annotations. During the annotation phase, annotators assess the emotional content of speech utterances by assigning attributes related to arousal, valence, and dominance. These attributes are evaluated using a 7-point Likert scale. Within this collection, *Label Studio* also logs the time spent on each annotation task, which aids our research objectives. The number of emotional annotations ranges from 3-7 per sample. In this study, we rely on release 1.00 of the BIIC-Podcast corpus, which contains 46,204 sentences (96 hours and 58 mins).

C. Data Preparation

The datasets explained above are designed for SER tasks [25]–[28]. Therefore, the data is structured so that the information about a particular sentence is easy to access. For this paper, we need information about each target annotation, including samples that were previously annotated. Therefore, the analysis in this paper considers the order in which each rater annotated the samples. First, we define the concept of an annotation *session*, which corresponds to all the annotations by one evaluator during an uninterrupted period of time. We order the annotations by the date and time they were conducted. Next, we define the *session* by looking at the time difference between annotations. We consider that a session ended if the given rater did not complete any new annotation in the next 15 minutes. The next annotation by this rater will mark the beginning of the next *session*.

Next, we define a *prime window* for each target rating in a session. We first assume that annotations closer in time are more likely to result in a priming effect. In the MSP-Podcast dataset, annotators in general take at most 1 minute to rate one sentence. Therefore, we only consider ratings done in the previous two minutes of a target sentence as possible primes. We refer to this two-minute segment as the *prime window*. Fig. 1 shows the process of selecting the prime window. The figure shows the prime window for the target valence rating

of sentence "g" by rater with ID 0144. In this case, the prime window for sentence "g" is $\{2, 1, 1, 2\}$. The prime window can contain any number of annotations, including zero. For the first two analyses in Section IV, we only consider target annotations with at least three ratings in their prime window.

IV. Affective Priming Effect Analyses

In this section, we conduct analyses to explore if the affective priming effect occurs in sentence-level emotional annotations. We compare ratings of arousal, valence, and dominance and categorical emotional ratings under different priming conditions. We explore whether the affective priming effect occurs in the emotional labeling of speech and strategies to quantify its impact on emotional attributes and categorical labels.

A. Low, Neutral, and High Attribute Priming

We want to explore if an affective priming effect exists in our annotations. We first define three types of prime windows for each emotional attribute: low, neutral, and high. A low prime window only contains attribute ratings of 1 and 2, i.e., calm for arousal, negative for valence, and weak for dominance. A neutral prime window only contains ratings of 3, 4, and 5. A high prime window only contains ratings of 6 and 7, i.e., active for arousal, positive for valence, and strong for dominance. In this section, we look at all target annotations with three or more ratings in their prime window, which is around 33.8% of the annotations in the MSP-Podcast corpus. Notice that we use a large corpora with several annotation sessions (we consider 71,696 sessions for the MSP-Podcast corpus and 898 sessions for the BIIC-Podcast corpus). While most cases do not qualify for low, neutral, and high prime windows, we still obtain enough examples for each condition. Out of the target annotations in the MSP-Podcast with three or more ratings in their prime window, 22.4% of arousal annotations, 26.1% of valence annotations, and 25.3% of dominance annotations demonstrate one form of prime window (low, neutral, or high). Table I shows the number of samples for each of these prime windows.

Next, we define the *attribute difference*. We want to compare annotations under a priming effect with annotations under no effect, i.e., expected annotations. We aim to measure the difference to quantify the affective priming effect. We assume that when all ratings of a sentence are averaged, any priming effect is mitigated. Therefore, we define the expected annotation of a target annotation as the average of all the ratings for the target sentence, excluding our target rating. Then, the attribute difference is the difference between the target rating and the expected annotation attribute value. Fig. 2 shows this process for the target valence rating of sentence "g" by rater 0144. In the example, the average score provided by the other four evaluators is 5.0. The score provided by rater 0144 is 2.0. Therefore, the difference is -3.0.

We calculated the attribute difference for each annotation in each prime window group. Then, we averaged the attribute differences in each priming group (i.e., low, neutral, and high). Fig. 3 reports the resulting average attribute differences for

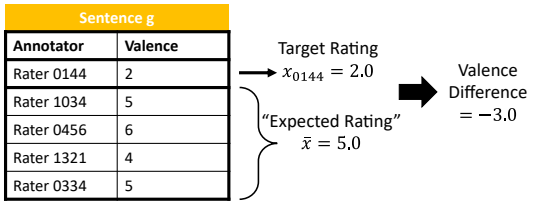


Fig. 2. Example of the calculation of the attribute difference of an annotation. The valence difference of the annotation of sentence "g" by rater 0144 is the difference between the valence target rating by rater 0144 and the valence expected rating, which is the average of the other valence annotations of sentence "g."

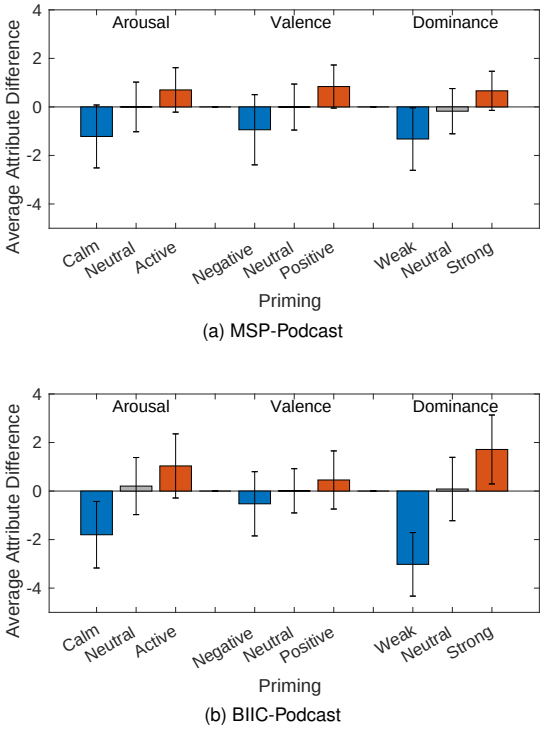


Fig. 3. Average attribute differences between target and expected annotations. The error bars show the standard deviation of the attribute differences for each type of priming.

the MSP-Podcast and BIIC-Podcast datasets. When annotators rate previous sentences with high attribute values, they will likely rate the current sentence with a higher score than the expected ranking. Similarly, when annotators rate previous sentences with low attribute values, they will likely rate the current sentence with a lower score than the expected ranking. Interestingly, when the previous sentences are rated with neutral scores, there is a minimal difference between the current rating and the expected ranking. To evaluate if the differences between the results of the priming groups are significant, we conducted t-tests (p-value < 0.05) between the attribute differences of each priming group within each attribute (all pairwise comparisons). These tests showed statistically significant differences between all groups tested. These results clearly show that we do have a priming effect in the attribute annotations. The results for the two datasets are similar, showing that the affective priming effect occurs in different environments and languages. However, the effects on

arousal and dominance are more extreme in the BIIC-Podcast corpus, while the effects on valence are more extreme in the MSP-Podcast corpus. The affective priming effect is similar between attributes, which coincides with previous affective priming results for the attribute of valence [9]. In general, the affective priming effect affects dominance the most and valence the least. We also conducted these experiments while varying the priming window length to 1, 1.5, 2, 2.5, and 3 minutes. We observed that all priming window lengths we tried showed a similar affective priming effect (Section I of supplemental materials).

In previous affective priming work, the target content also had an influence on the priming effect [9]. To explore how the content of our target sentences affects priming, we further group our target annotations into content bins according to their expected annotations. For example, the target annotation in Fig. 2 has an expected valence rating of 5.0, so it is placed in the "[5,6)" valence label bin. We conduct the same attribute difference experiments on these smaller groups, reporting the results on low, neutral, and high prime windows. Table I shows the number of examples considered in each case. Figs. 4 and 5 show the results for the MSP-Podcast corpus and BIIC-Podcast corpus, respectively. The trends are similar between the datasets, with the differences coming from the language or dataset size differences. We generally see larger attribute differences when the priming and content types are further apart. These results imply that when the prime window values are similar to the emotional content of the sentence (e.g., the prime window has positive ratings, and the sentence is also positive), there is a smaller priming effect. For the extreme bins, the affective priming effect is bounded by the range of the annotations (1 to 7). The priming effect pushes annotations to extremes, but if the average annotation is at a limit, the annotator cannot give a more extreme value, even if they perceive a more extreme emotion than the expected ranking.

We also examine the affective priming effect on individual annotators. We conduct similar experiments to those above to determine whether the affective priming effect varies depending on the reliability of the raters. We only conduct this analysis on the annotations of the MSP-Podcast corpus since we have 14,363 annotators in release 1.11. We focus on low and high prime windows, considering annotators with at least five annotations included in the specific prime window. We consider 177 raters with this criterion. We estimate their reliability by considering all of their annotations in the MSP-Podcast corpus. We use the Krippendorff's Alpha coefficient [1] as the agreement metric. Annotators with average agreements below 0.2 are clustered in the poor reliability group (orange bars in Fig. 6). Otherwise, they are clustered in the higher reliability group (blue bars in Fig. 6).

We first group annotations by annotator. Then, we group annotations by prime window type. Then, we take the attribute difference from the expected scores for each sentence in each group, estimating the average of those differences. Fig. 6 shows the histogram of these average attribute differences for each annotator and priming type. The figure shows that most annotators have similar average attribute differences when primed similarly. These results show that for high attribute

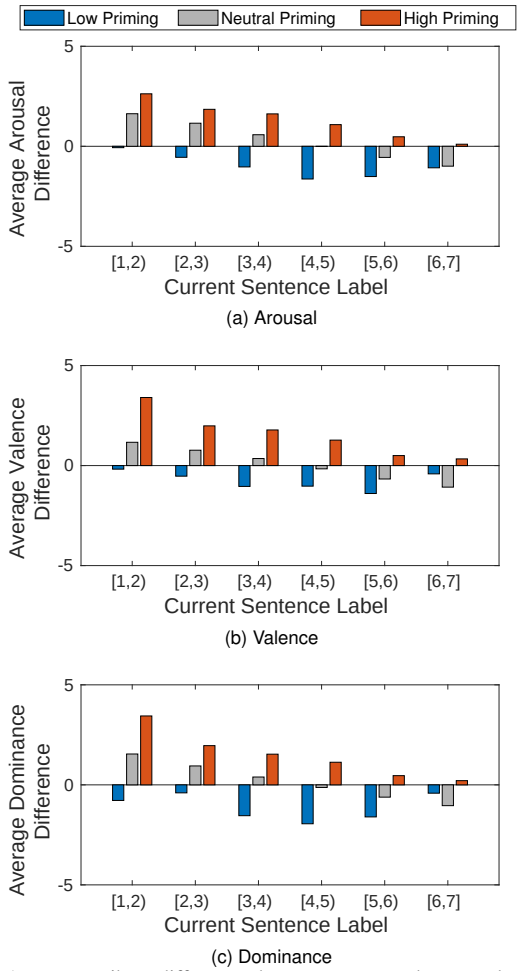


Fig. 4. Average attribute differences between target and expected annotations of the MSP-Podcast corpus. Before averaging, the attribute differences are binned according to their expected annotations.

priming, annotators in the poor reliability group have similar priming effects as the annotators in the higher reliability group. For the low attribute priming, the annotators in the poor reliability group have values further from zero, which suggests that they are more affected by the affective priming. However, there is a low amount of data available for the low priming cases, especially for dominance, which makes it hard to draw concrete conclusions from the graphs. We also conducted these experiments using two additional agreement thresholds, 0.4 and 0.6, where we observe similar trends (Section II in supplemental materials).

B. Emotional Category Priming

In the previous section, we showed that the affective priming effect occurs in our attribute annotations, but, does this effect occur in emotional category annotations? Since these annotations consist of emotional categories instead of a continuous value, as with the emotional attributes, defining priming types is different. In this section, we group annotations by the singular category in their prime window. For example, if a target label only has the label "Happy" in their prime window, the sample is placed in the happy priming group. If the

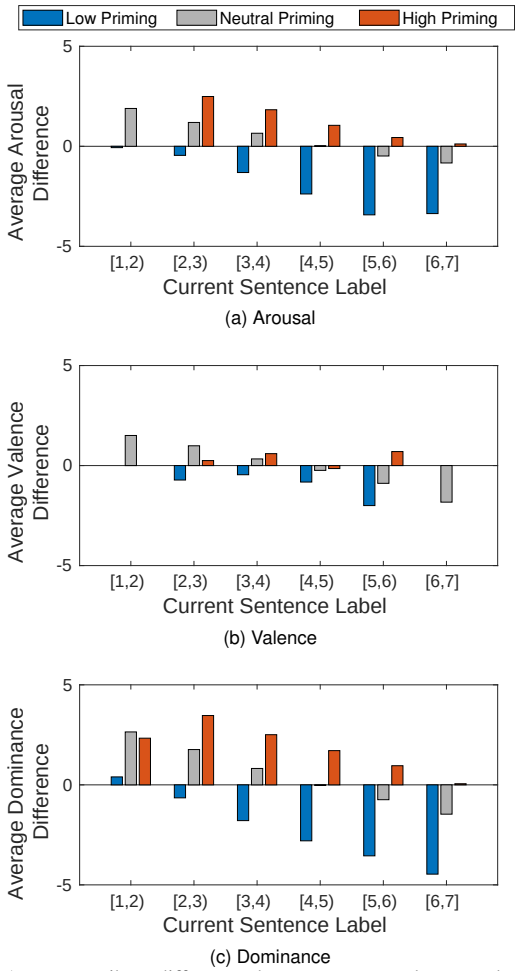


Fig. 5. Average attribute differences between target and expected annotations of the BIIC-Podcast corpus. Before averaging, the attribute differences are binned according to their expected annotations.

prime window contains less than three annotations or a mix of categories, the target annotation is not considered in this analysis.

Since we cannot easily measure a meaningful difference between categories, we compare the frequency or coverage of an emotional category in the labels. We first plot the coverage of each emotional category in our target labels for each priming type, shown in blue in Fig. 8. For example, 80% of the target labels are "Happy" in the MSP-Podcast happy priming group, while less than 10% are "Angry." Next, we plot the coverage in our expected labels. The expected labels are the emotional categories for the sentence, excluding the target label. Fig. 7 shows an example for sentence "g." This sentence has the following emotional category annotations: {Surprise, Happy, Happy, Neutral, Happy} and the target rater 0144 labeled it as "Surprise." The expected label for the target annotation will include the classes {Happy, Happy, Neutral, Happy}. We repeat this process for all the sentences primed by the target emotional class. Then, we estimate the histogram considering all the expected labels provided to the sentences (see Fig. 7). The coverage of the expected labels is shown in orange in Fig. 8. The results are similar between the two

TABLE I
NUMBER OF TARGET ANNOTATIONS IN EACH AFFECTIVE PRIMING GROUP AND EACH SENTENCE ATTRIBUTE LABEL BIN.

MSP-Podcast Corpus								
Affective Priming		*Expected Annotation Bins*						
		[1,2)	*[2,3)*	*[3,4)*	*[4,5)*	*[5,6)*	*[6,7]*	*Total*
Aro	Low	13	153	323	323	161	33	1,006
	Neutral	167	2,920	10,378	21,258	14,297	1,626	50,646
	High	7	96	472	1,224	3,340	465	5,604
Val	Low	13	132	214	164	79	14	616
	Neutral	559	6,447	18,505	25,327	11,004	898	62,740
	High	9	71	293	913	2,056	150	3,492
Dom	Low	29	73	61	102	35	4	304
	Neutral	53	1,388	9,955	29,336	17,337	1,227	59,296
	High	3	26	256	1,208	3,553	229	5,275
BIIC-Podcast Corpus								
Affective Priming		*Expected Annotation Bins*						
		[1,2)	*[2,3)*	*[3,4)*	*[4,5)*	*[5,6)*	*[6,7]*	*Total*
Aro	Low	5	241	354	407	231	24	1,262
	Neutral	72	1,408	4,360	5,267	2,448	312	13,867
	High	0	20	50	85	102	36	293
Val	Low	0	6	51	25	1	0	83
	Neutral	61	1,614	11,228	9,346	1,508	31	23,788
	High	0	1	13	14	10	0	38
Dom	Low	26	143	353	633	419	41	1,615
	Neutral	56	627	2,942	5,473	3,972	580	13,650
	High	4	126	615	1,297	1,047	180	3,269

datasets. Overall, if an annotation is primed with an emotional category, it is more likely to be labeled as that category. These results clearly show that the affective priming effect also occurs in emotional category annotations.

C. Priming Patterns

In the previous two sections, we explored the affective priming effect in general priming groups (e.g., high versus low for emotional attributes; happy, sad, angry, and neutral for categorical classes). However, there are many annotations in our datasets that do not have such a clear priming. An evaluator might see a very sad sentence, then a very excited one, then the target sentence, with no clear type of priming. In this section, we define a priming pattern for each annotation in our dataset.

The priming pattern of a target annotation is defined as the ratings or categorical labels of the most recent three or less annotations in its prime window. We choose the closest three or less annotations to ensure we keep as much priming information while still having enough target annotations in each priming pattern to analyze them. If we were to increase the maximum number of annotations in each priming pattern, we would have many patterns with almost no target ratings. This section reports the results on the MSP-Podcast corpus. The results for the BIIC-Podcast are described in the supplemental material. For example, the pattern "6 7 5" for a given emotional attribute indicates that the last sentence to be annotated before the target sentence was labeled with a score of "5". The other two previous sentences were labeled with "7" and "6", respectively. Similarly, the pattern "N N H" for an emotional category annotation indicates that the last sentence to be annotated before the target sentence was labeled as *Happy*, and the other two previous sentences as *Neutral*.

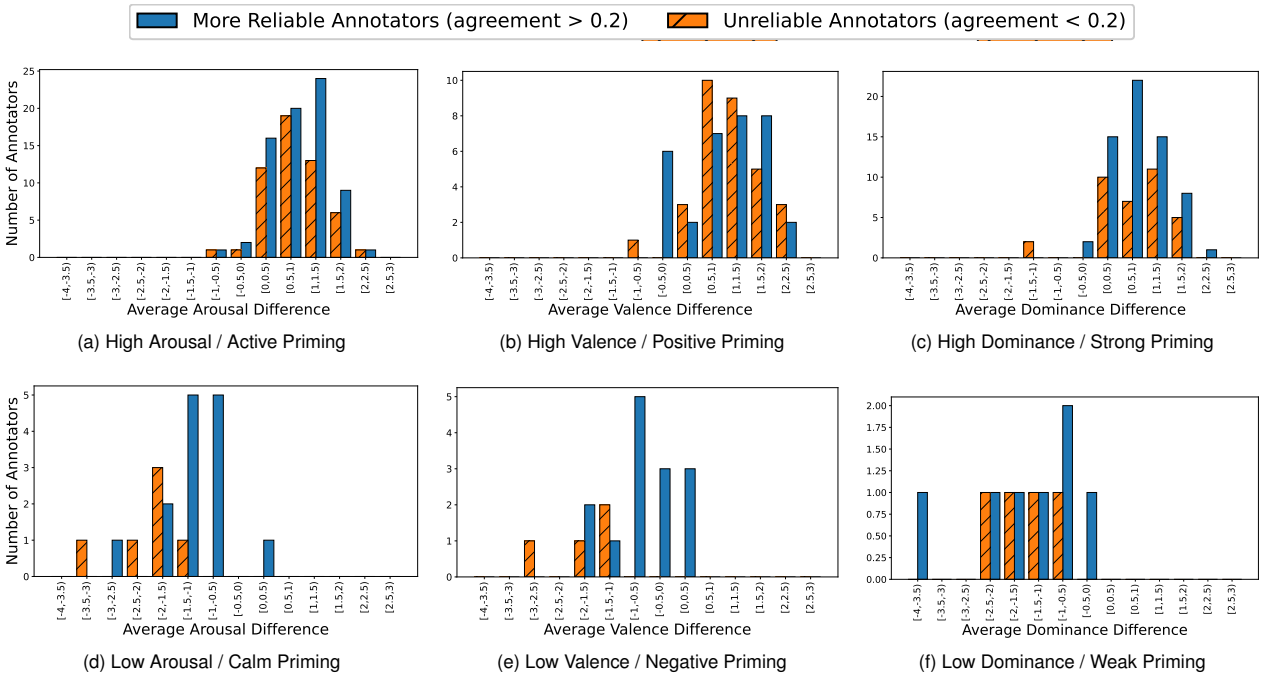


Fig. 6. Histograms of the average attribute differences between target and expected annotations of the MSP-Podcast corpus done by individual annotators. Annotators with low inter-evaluator agreements are highlighted in orange and with higher inter-evaluator agreements are highlighted in blue.

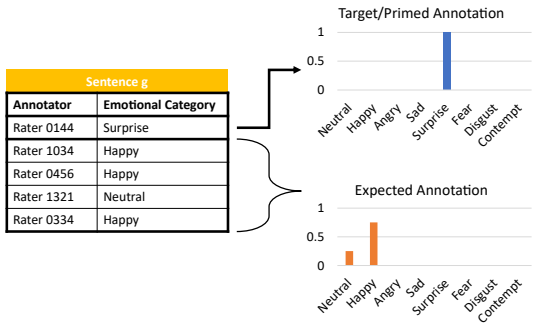


Fig. 7. Example of the process to obtain the vectors for the target/primed annotations and expected annotations using emotional categories. The categorical label of rater 0144 for sentence "g" is $Surprise$. Therefore, the one-hot vector for the primed annotation data has only a "1" for surprise and "0" for the other emotions. The vector for the expected annotations includes three selections for $Happy$ (0.75) and one for $Neutral$ (0.25).

We start the analysis with emotional attributes. We conduct experiments similar to the ones reported in Section IV-A. Figs. 9a (arousal), 9b (valence), and 9c (dominance) show the average attribute differences for target annotations for some of the priming patterns. We observe similar results to our previous analyses, where "stronger" priming, i.e., priming with more extreme and consistent values, results in a more pronounced affective priming effect. For example, in Fig. 9a, the absolute arousal differences are higher for the priming patterns of "7 7 7" and "1 1 1" than for the priming patterns of "7 1 1," "1 7 7," and "4 4 4." In Fig. 9a, 9b, and 9c, the plotted red line shows the average attribute difference when there were no annotations in the priming window, e.g., at the beginning of a session, which is close to 0. We conduct similar experiments on the BIIC-Podcast corpus (Section III of supplemental materials), which show similar trends on the average attribute difference results.

We also analyze the patterns for categorical classes. The key requirement for this analysis is to define a strategy to quantify the affective priming effect. Emotional attributes have continuous values, so we can estimate the difference between the target score and the expected label. This is not straightforward for categorical classes. Instead, we transform the target and expected categorical labels into vectors. Then, we estimate the cosine distance between them. The label vectors for each target and expected categorical label are created by only considering the following categories: "Neutral" (N), "Happiness" (H), "Anger" (A), and "Sadness" (S). The rest of the emotional categories are added to the class "Other" (Oth).

The five-dimensional vector for the target category label is a one-hot vector with a single "1" in the location representing the emotional label [N,H,A,S,Oth]. The five-dimensional vector for the expected category label is created by estimating the proportion of labels assigned by other raters to the target sample. Fig. 10 shows this process for sentence "g" by rater 0144. The cosine distance function d is defined as:

$$d(\mathbf{x_r}, \bar{\mathbf{x}}) = 1 - \frac{\mathbf{x_r} \cdot \bar{\mathbf{x}}}{\|\mathbf{x_r}\|_2 \, \|\bar{\mathbf{x}}\|_2},$$

where $\mathbf{x_r}$ is the target label vector of rater r and $\bar{\mathbf{x}}$ is its expected label vector. We utilize the cosine distance as it is a standard method of quantifying the difference between vectors. The cosine distance has a minimum value of 0 when the vectors are pointed in the same direction and a maximum value of 2 when the vectors are pointed in opposite directions. In our case, the sum of the vector components are all 1. Therefore, the two vectors are the same if the distance is 0. All components

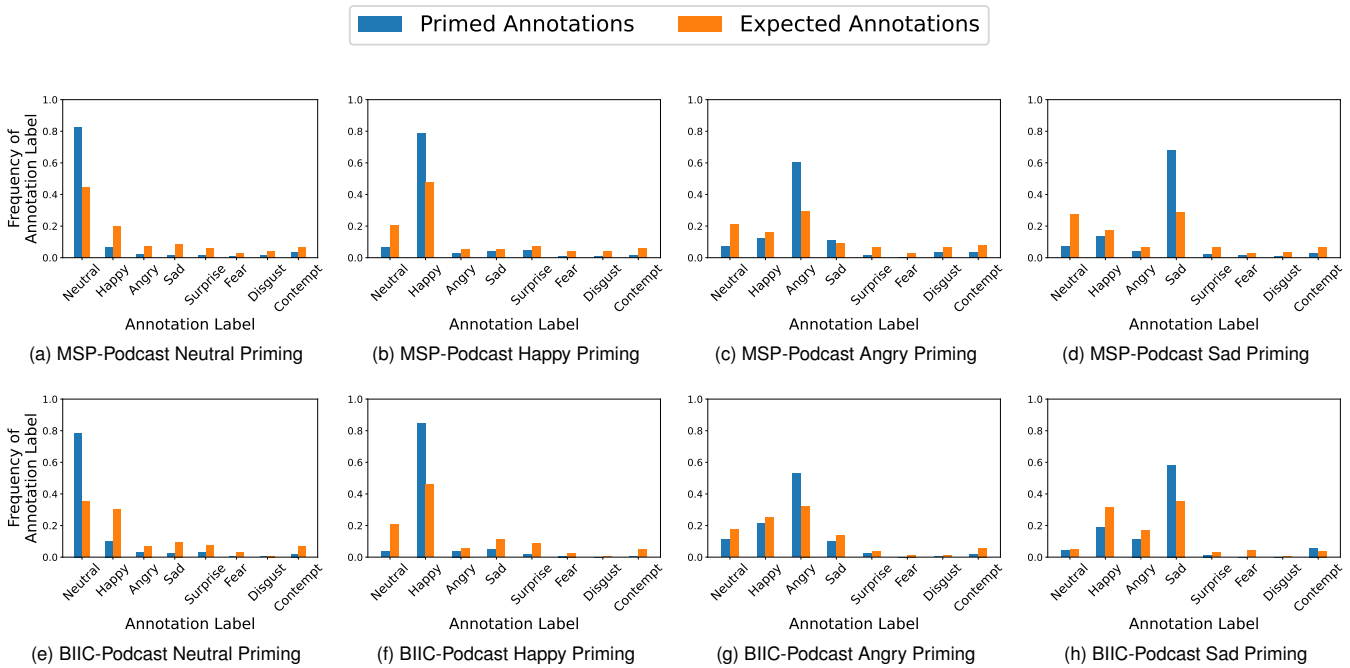


Fig. 8. Frequency of categories in the primed (target) annotations and the expected annotations of the emotional category of the sentences in the MSP-Podcast and BIIC-Podcast corpora.

also have values of 0 or more, which means our distance metric has a maximum value of 1.

Fig. 9d shows the average categorical differences for target annotations with a selection of priming patterns. Similarly to the attribute results, patterns with a clear emotion differ more than patterns with just neutral labels. However, unlike the attribute results, mixed emotions in priming do not lead to a lower difference. If the emotions are mixed with neutral, such as the priming patterns of "N N H" and "N N S," the difference does decrease. However, if the emotions are mixed with other emotions, such as "S S H" and "H S S," the difference can be higher. For emotional category priming, priming with any mixture of emotions leads to a more pronounced affective priming effect than neutral priming. Fig. 9d also shows the average categorical difference when there were no annotations in the priming window using a solid red horizontal line. Its value is around 0.3. This is the minimum categorical difference of all the priming patterns. Therefore, the average cosine distance for priming should be compared to this minimum instead of zero.

D. Creating an Expected Bias Measure

In the previous section, we observed that we could obtain measurements on how much affective priming affects each type of priming in the corpora. In this section, we use the average attribute and categorical differences for each priming pattern as the *expected bias* for each pattern. This expected bias can then be assigned to each annotation in the corpus with the same priming pattern seen during the annotation process (e.g., look-at-table approach). Then, we can average the expected biases of the annotations for every sentence in the corpus, which we call the *expected affective priming bias* of each sentence. Fig. 11 shows the process of calculating the expected affective priming bias of the valence label of sentence "g." For example, the pattern "1 1 2" has an expected bias of -0.687 for the attribute of valence. The valence annotation of sentence "g" by rater 0144 is preceded by the same pattern of valence annotations. Therefore, it has the same expected bias. We follow the same process for the rest of the annotations of sentence "g." Then, we average those expected biases to get the overall expected affective priming bias of sentence "g." The actual annotations of the sentence do not play a role in the expected bias calculation. This expected bias gives us an idea about how much the affective priming effect has affected the labels in a corpus.

After calculating the expected affective priming bias for each sentence in the corpus, we plot histograms of the biases of the attribute and emotional categories for the full MSP-Podcast corpus. Fig. 12 shows all the histograms of the MSP-Podcast corpus. For the attribute labels, Figs. 12a- 12c show that the affective priming biases are concentrated in the center around a bias of 0, where we see large peaks. These peaks contain sentences with either annotations with no prime window (e.g., all annotations done at the beginning of sessions) or annotations whose prime windows gave opposite effects (e.g., a sentence where half of its annotations were primed with "7 7 7" and the other half with "1 1 1"). We also calculate and plot the affective priming biases of the BIIC-Podcast corpus (Section IV of supplemental materials). The BIIC-Podcast histograms show that the attribute priming biases are also concentrated near zero. For the category labels, Fig. 12d shows that the affective priming biases have two peaks. The first peak is at the average categorical difference of the annotations with no annotations in their priming window (e.g., all annotations were done at the beginning of sessions). The second peak is around a bias of 0.425. This peak has

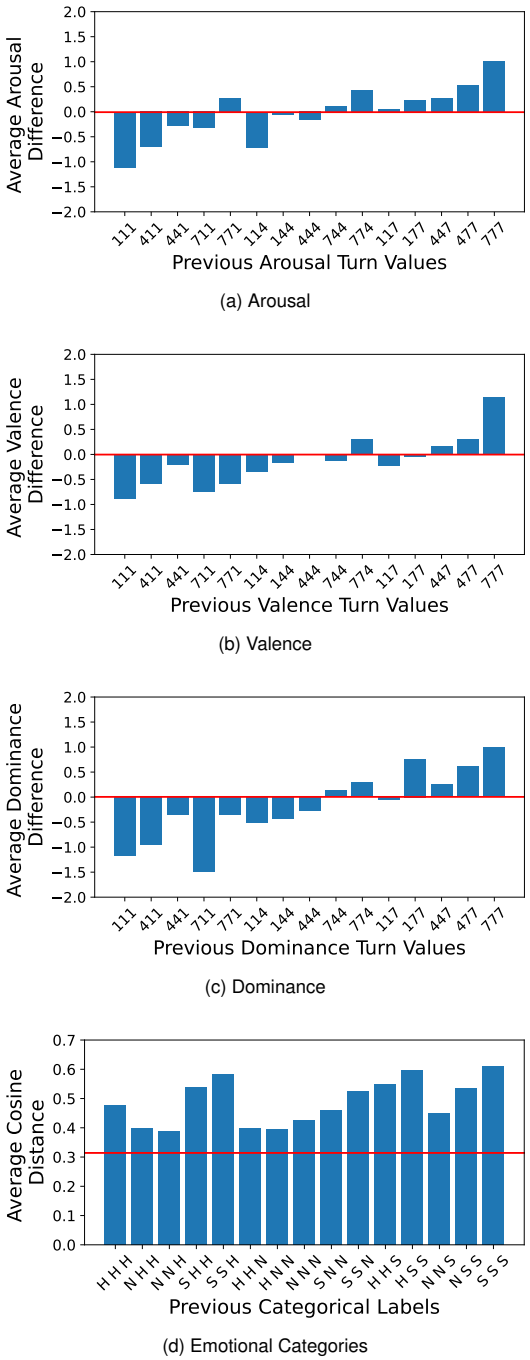


Fig. 9. Average attribute differences between target and expected annotations of the MSP-Podcast corpus. Before averaging, the attribute differences are grouped according to the priming patterns of their target annotations.

sentence annotations with a mixture of emotional and neutral annotations in their prime window. In the next section, we utilize the histograms to split the corpus into subsets that are expected to be more or less affected by the affective priming.

V. Modeling Experiments

This section explores the affective priming effect on SER modeling. We aim to assess whether the affective priming effect impacts SER performance. The experiments in this section are conducted on the MSP-Podcast corpus, since it is the larger corpus.

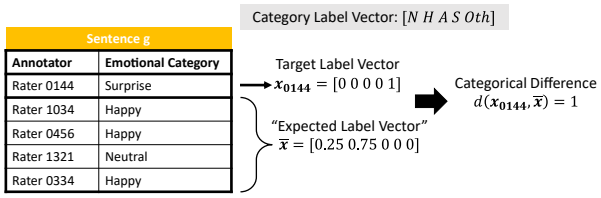


Fig. 10. Example of the calculation of the categorical difference of an annotation. The categorical difference of the annotation of sentence "g" by rater 0144 is the distance between the target label vector, created by the labeled category from annotator 0144, and the expected label vector, created by the other annotations of sentence "g."

Fig. 11. Example of the calculation of the expected affective priming bias of a sentence. The expected bias of the valence label of sentence "g" is the average of the expected biases of each valence annotation of the sentence. The expected affective priming bias of each annotation is based on the priming pattern seen by the rater preceding the current annotation.

A. Creating Biased Data Subsets

In Section IV-D, we defined affective priming biases for the sentences in the MSP-Podcast corpus. Our goal in this section is to use those biases to create biased and unbiased subsets of the MSP-Podcast corpus to ultimately explore how the affective priming effect affects SER modeling. First, we sort all the sentences according to their biases. We have four independent affective priming bias lists for each sentence, one for each of the three attributes and one for the emotional categories.

The subsets for the emotional attributes are as follows. We create four subsets for each emotional attribute: the *Neg* and *Pos* subsets are the sets of sentences with the bottom 20% and top 20% of affective priming biases, respectively. We call these the *biased subsets*. The *Neut1* subset is the set of sentences whose bias is between the 40% and 60% quantiles, and the *Neut2* subset has sentences in which the bias is between -0.02 and 0.02. These are two different ways to define sentences with low affective priming biases, one by choosing the sentences with middle bias values (Neut1) and the other by choosing sentences with bias values around zero (Neut2). We call these sets the *unbiased subsets*. The subsets are visualized in Fig. 12a, 12b and 12c. The *Neg* subset is shown in blue, the *Pos* subset is shown in orange, and the *Neut1* subset is shown in black. We do not show the *Neut2* subset since it mostly overlaps with the *Neut1* subset. Table II shows the number of sentences in each partition of each attribute subset.

The subsets for the emotional categorical labels are as follows. Since the emotional category affective priming bias does not have negative values, we define different subsets: the *Neut1* and *High* subsets are the sets of sentences with the bottom 20% and top 20% of affective priming biases, respectively. The *Mid* subset is the set of sentences whose

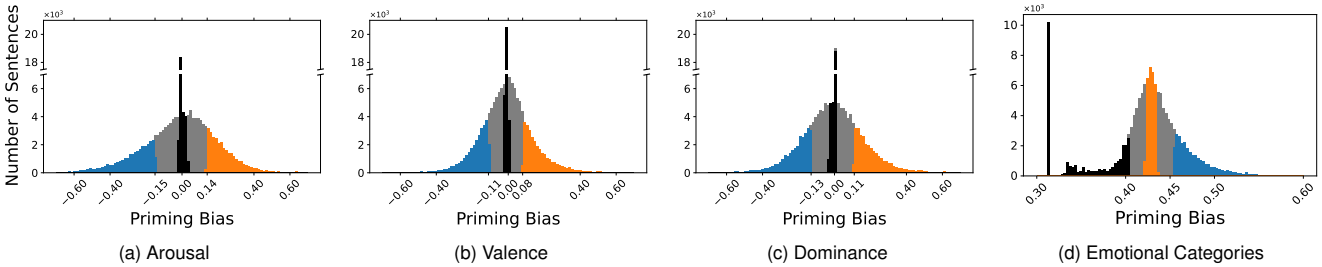


Fig. 12. Histograms of the expected affective priming biases (MSP-Podcast corpus). The sentences are split according to their bias values. In Fig. 12a, 12b, and 12c, the sentences with the bottom 20% of the bias values are shown in blue, the top 20% are shown in orange, and the 20% centered around the median are shown in black. In Fig. 12d, the sentences with the bottom 20% of bias values are also shown in blue, the top 20% are shown in black, and the 20% centered around the median are shown in orange. The rest of the sentences are shown in grey. Most sentences have low expected affective priming bias.

TABLE II
NUMBER OF SENTENCES IN EACH ATTRIBUTE SUBSET FOR THE MSP-PODCAST CORPUS.

Corpus Subset	Subset Partition	Emotional Attribute		
		Arousal	*Valence*	*Dominance*
Neg	Train	18,402	17,260	18,019
	Dev	3,438	4,346	3,443
	Test	5,106	6,216	5,577
Pos	Train	16,612	17,318	16,778
	Dev	3,593	2,921	3,555
	Test	6,331	5,335	6,123
Neut1	Train	15,139	15,232	15,331
	Dev	5,285	5,151	5,200
	Test	6,389	6,371	6,296
Neut2	Train	13,049	16,875	14,531
	Dev	4,786	5,465	5,047
	Test	5,619	6,958	6,053

TABLE III
NUMBER OF SENTENCES IN EACH EMOTIONAL CATEGORY SUBSET IN THE MSP-PODCAST CORPUS.

Corpus Subset	Subset Partition		
	Train	*Dev*	*Test*
High	16,898	3,312	5,247
Mid	15,891	2,898	5,987
Neut1	13,927	4,823	5,832
Neut2	14,051	4,846	5,893

bias is between the 40% and 60% quantiles, and the *Neut2* subset has sentences whose bias is between 0 and 0.405. Fig. 12d visualizes the subsets, where the *High* subset is shown in blue, the *Mid* subset is shown in orange, and the *Neut1* subset is shown in black. Table III shows the number of sentences in each partition of each emotional category subset.

1) Inter-Evaluator Agreements of Subsets: We estimate the inter-evaluator agreements of the different subsets. The inter-evaluator agreement of the labels gives us an idea of the reliability of the labels, especially for SER tasks. Noisy and inconsistent emotional labels often lower the performance of SER models [29]. We use the Krippendorff's Alpha coefficient [1] to evaluate the agreement of the attribute labels, and we use the Fleiss' Kappa [30] to evaluate the agreement of the emotional category labels. Table IV shows the calculated inter-evaluator agreements. The unbiased subsets have the highest agreements, followed by the full MSP-Podcast dataset (*Full*). For the attribute labels, the biased subsets have the lowest agreements. This result is expected since their range of biases

TABLE IV
INTER-EVALUATOR AGREEMENTS FOR THE DIFFERENT SUBSETS. WE USE THE KRIPPENDORFF'S ALPHA FOR EMOTIONAL ATTRIBUTES AND THE FLEISS' KAPPA FOR EMOTIONAL CATEGORIES.

Corpus Subset	Emotional Attribute			Corpus Subset	Emotional Categories
	Arousal	*Valence*	*Dominance*		
Full	0.413	0.372	0.358	*Full*	0.174
Neg	0.306	0.327	0.255	*High*	0.128
Pos	0.351	0.257	0.235	*Mid*	0.070
Neut1	0.622	0.601	0.652	*Neut1*	0.446
Neut2	0.653	0.582	0.665	*Neut2*	0.444

is much larger (Fig. 12). Therefore, the labels are *pushed* by different amounts during the annotation. For the emotional categories, we observe an opposite result. The *Mid* subset has the lowest agreement, possibly due to more confusion during more neutral priming. Annotators might agree more often if they are all strongly primed (e.g., with only *Happy* sentences).

2) Preparing Data for Experiments: The subsets shown in Tables II and III do not have the same emotional distributions. They also do not have the same distribution as the full MSP-Podcast dataset. When training and testing models, the emotional distribution differences could affect the comparisons between the models. For example, Sridhar and Busso [31] showed that predictions for arousal, dominance, and valence have varying uncertainty for different values of the emotional attributes. To isolate the effect of the affective priming bias, we normalize the emotional distributions of the subsets using an under-sampling strategy. We sample each subset according to a defined emotional distribution. Since all we require is the same or similar emotional distribution between each subset, we could choose an arbitrary emotional distribution. We decide to follow the emotional distribution of the full MSP-Podcast corpus. We fit a Gaussian distribution for each attribute: arousal $\mathcal{N}(4.37,0.96)$, valence $\mathcal{N}(3.99,0.90)$, and dominance $\mathcal{N}(4.50,0.59)$. Next, we bin the labels according to the following edges: $[1, 2, 2.5, 3, 3.25, 3.5, 3.75, 4, 4.25, 4.5, 4.75, 5, 5.5, 6, 7]$. We use these non-uniform bins since most of the data is concentrated around the neutral attribute values (3-5). Next, we calculate the area under the fitted Gaussian for each bin. We take the ratio of the area of each bin to the area under the Gaussian truncated at 1 and 7 (the label limits). We are left with the percentage of

sentences to be randomly selected in each attribute label bin. To define the desired emotion distribution for the emotional categories, we count the number of sentences with the same consensus labels and divide by the full number of sentences to get the rate for each category. We use the following categories: *[Anger, Sadness, Happiness, Surprise, Fear, Disgust, Contempt, Neutral, Other]* with "Other" containing any other label not listed, including "no agreement." We are left with a percentage of sentences for each emotional category.

To sample each subset, we first take all the sentences in each subset partition (train, development, and test) and bin their labels using the predefined edges or by category. Then, we sample sentences for each partition. For the attributes, we choose around 7,000 sentences for each train set, 700 for each development set, and 1,400 for each test set. For the emotional categories, we choose around 10,000 sentences for each train set, 1,000 for each development set, and 2,000 for each test set. We randomly sample (without replacement) sentences from each bin or category to match the desired percentage of each bin of their corresponding attribute or emotional category. For example, if the wanted percentage for valence bin 3 is 9%, we will randomly sample $7,000 * 0.09 = 630$ sentences from bin 3 of each valence train set. Most subset bins have more sentences than we need in the bin or category. However, some have fewer sentences than needed. In those cases, we choose all sentences available in that bin or category and then randomly sample (with replacement) the same sentences until we have the required number.

B. SER Model Details

We implement our approach with the "wav2vec2-large-robust" architecture [32], which showed the best recognition performance in the study of Wagner *et al.* [33] among the variants of the Wav2vec2.0 model [34]. We first import the pre-trained "wav2vec2-large-robust" model from the HuggingFace library [35]. Then, we fine-tune the transformer encoder with the downstream head using specific subsets of the corpus depending on the evaluation, as explained in Sections V-C, V-D, and V-E. For the attribute models, the downstream head consists of one fully connected layer and a linear output layer. The fully connected layer has 1,024 nodes, layer normalization, and the *rectified linear unit* (ReLU) as the activation function. The linear output layer has three nodes to predict the emotional attribute scores for arousal, dominance, and valence. For the emotional category models, we use the same downstream head used for emotional attributes. The only difference is the output layer, which has four nodes to predict category probabilities, followed by a softmax layer. We predict one of four categories, *[Anger, Sadness, Happiness, Neutral]*, in the emotional category experiments.

We freeze the convolutional feature encoder during fine-tuning as it performs better than updating all the parameters [36]. We aggregate the frame-level representations of the transformer encoder to the sentence-level representation by using average pooling per utterance. Then, we feed the representation to the downstream head. For the regularization, dropout is applied to all the hidden layers, with a rate set to $p = 0.5$. To fine-tune the pre-trained SER model, we use specific subsets of the MSP-Podcast corpus depending on the evaluation, as explained in Sections V-C, V-D, and V-E. We apply Z-normalization to the raw waveform by using the mean and standard deviation estimated over the training set, and min-max normalization to the emotional attribute labels, mapping them into the range of 0 to 1. We use the Adam optimizer [37] with a learning rate of 0.0001. We use 32 utterances per mini-batch and update the model for ten epochs. We employ early stopping using the development set to avoid overfitting.

C. Testing SER Model with Biased Subsets

For the experiments in this section, we train the model with the full MSP-Podcast train set. Then, we test the models with the sampled test subsets. The subset *All* is sampled from the full MSP-Podcast test set in the same manner as the other sampled subsets (Section V-A2). The subset *Neg+Pos* is sampled from the combined *Neg* and *Pos* subsets. Table V shows the average attribute results of ten testing trials (sampling test sets ten times) with the *standard deviation* (STD) of the ten trials in brackets. For all three attributes, the model had the highest test CCC when testing on the *Neg* subset. The model can predict the labels of the biased subsets (i.e., *Neg*, *Pos*, and *Neg+Pos*) with higher performance than the labels of the unbiased subsets (i.e., *Neut1* and *Neut2*). Table VI shows the average emotional category results of ten testing trials. Like the attribute results, the model can predict the biased subset labels (e.g., *High* and *Mid*) better than the unbiased subset labels. The model had the highest test macro F1 score and *unweighted average recall* (UAR) when testing on the *Mid* subset (i.e., *Neut1* and *Neut2*). Our hypothesis for these results comes from Sridhar and Busso [31], where they showed that SER models are uncertain when predicting neutral emotional values and more certain when predicting extreme emotional values. The biased subsets have labels that are shifted to extremes compared to those without priming.

> We hypothesize that (a) our models follow the uncertainty pattern seen in Sridhar and Busso [31], making them more certain of the biased labels, as they are more extreme than the unbiased labels. Therefore, we hypothesize that (b) due to the extreme biased labels, the models are also more accurate when predicting and learning from biased labels as opposed to unbiased ones.

The STDs shown in Tables V and VI show that, in general, the model is more uncertain when predicting the unbiased labels, but this pattern is not as clear as the lower prediction results. We further analyze these results in Section V-G, where we explore the model's performance for specific attribute values to test our hypothesis on the accuracy of our model.

D. Training SER Model with Biased Subsets

Next we trained the model with all the different sampled subsets and tested with the full MSP-Podcast test set. Tables VII and VIII show the average test results for ten trials

TABLE V
AVERAGE TEST CCC RESULTS OVER 10 TESTING TRIALS USING THE MODEL TRAINED WITH THE FULL MSP-PODCAST TRAIN SET AND TESTED ON THE SAMPLED ATTRIBUTE PARTITIONS. THE STD OF THE 10 TRIALS ARE SHOWN IN BRACKETS.

Sampled Test Set	Emotional Attribute		
	Arousal	*Valence*	*Dominance*
All	0.694 [0.02]	0.586 [0.01]	0.641 [0.01]
Neg	**0.782** [0.01]	**0.638** [0.01]	**0.702** [0.01]
Pos	0.748 [0.01]	0.566 [0.02]	0.690 [0.01]
Neg+Pos	0.761 [0.01]	0.592 [0.02]	0.698 [0.01]
Neut1	0.550 [0.02]	0.505 [0.01]	0.508 [0.02]
Neut2	0.522 [0.02]	0.496 [0.01]	0.495 [0.01]

TABLE VI
AVERAGE TEST MACRO F1 AND UAR RESULTS OVER 10 TESTING TRIALS USING THE MODEL TRAINED WITH THE FULL MSP-PODCAST TRAIN SET AND TESTED ON THE SAMPLED EMOTIONAL CATEGORY PARTITIONS. THE STD OF THE 10 TRIALS ARE SHOWN IN BRACKETS.

Sampled Test Set	Metric	
	Macro F1	*UAR*
All	0.478 [0.006]	0.507 [0.006]
High	0.481 [0.007]	0.509 [0.008]
Mid	**0.504** [0.006]	**0.536** [0.007]
Neut1	0.408 [0.009]	0.430 [0.009]
Neut2	0.415 [0.008]	0.437 [0.008]

TABLE VII
AVERAGE TEST CCC RESULTS OVER 10 TRAINING TRIALS USING THE FULL MSP-PODCAST TEST SET AND TRAINED ON THE SAMPLED ATTRIBUTE PARTITIONS. THE STD OF THE 10 TRIALS ARE SHOWN IN BRACKETS.

Sampled Train Set	Emotional Attribute		
	Arousal	*Valence*	*Dominance*
All	0.639 [0.02]	0.337 [0.04]	0.541 [0.05]
Neg	0.640 [0.02]	**0.397** [0.03]	**0.549** [0.04]
Pos	0.642 [0.01]	0.339 [0.06]	0.538 [0.06]
Neg+Pos	**0.654** [0.01]	0.383 [0.05]	0.544 [0.02]
Neut1	0.615 [0.01]	0.322 [0.06]	0.530 [0.04]
Neut2	0.602 [0.03]	0.294 [0.08]	0.522 [0.05]

TABLE VIII
AVERAGE TEST MACRO F1 AND UAR RESULTS OVER 10 TRAINING TRIALS USING THE FULL MSP-PODCAST TEST SET AND TRAINED ON THE SAMPLED EMOTIONAL CATEGORY PARTITIONS. THE STD OF THE 10 TRIALS ARE SHOWN IN BRACKETS.

Sampled Train Set	Metric	
	Macro F1	*UAR*
All	0.429 [0.005]	0.465 [0.005]
High	**0.437** [0.004]	**0.474** [0.005]
Mid	0.431 [0.004]	0.468 [0.004]
Neut1	0.389 [0.020]	0.432 [0.012]
Neut2	0.403 [0.027]	0.446 [0.018]

(sampling test sets and initializing the network ten times) with the STD in brackets. These results are similar to our previous results in Table V, where the unbiased subsets show the worst results. The model has a harder time learning an effective connection between the speech and emotional labels when using unbiased labels, even when most test labels are unbiased (the full test set). As opposed to our previous results, the models trained on the *All* subset perform similarly to the models trained on the biased subsets. These results suggest that having at least some biased labels in the training set keeps the performance high, where more biased labels improve the model but have a minimal effect compared to not having any biased labels. We will analyze in more detail the uncertainty of our models for different prediction values in Section V-F to test our hypothesis on the uncertainty of our models.

E. Training and Testing SER Model with Biased Subsets

We also test and train all the models with the sampled subsets. The results of these experiments are shown in Tables IX and X. The best result for each sampled test set is shown in bold, and the best result for each sampled train set is shown in italics. In general, the attribute models perform best when predicting on the *Neg* test subset, for all attributes. For arousal, the models trained with the *Neg+Pos* subset perform the best for all test sets. For valence and dominance, the models trained with the *Neg* subset perform best. The arousal and dominance results show that model performance drops when training with the unbiased subsets. However, there are bigger performance drops when testing with the unbiased subsets. All the models can predict the biased labels much better than the unbiased labels. However, training with the unbiased labels still gives information on predicting the biased labels. The valence results show the opposite, where the performance drop is larger when training with unbiased subsets than when testing on them. For valence, models trained with unbiased labels cannot predict unbiased or biased labels, while those trained with biased labels can predict unbiased labels.

The emotional category models show similar trends, where testing with the *Mid* subset gives the best results and training with the *High* subset gives the best results. Like the arousal and dominance results, the performance drop when going from testing with biased to unbiased subsets is larger than the drop when going from training with the biased to unbiased subsets. All emotional category models have a hard time predicting the unbiased labels, while training with the unbiased labels still gives the model information about the biased labels. The attribute and emotional category results support *Hypothesis (b)*. We consistently observe that the models are more accurate when predicting and learning from biased labels. To support the rest of our hypothesis, we further explore our prediction and uncertainty results in the next two sections.

F. Uncertainty Analysis of Models

The results in the previous three sections show that even though the unbiased subsets of the MSP-Podcast corpus have a much higher inter-evaluator agreement, the models trained with the unbiased subsets perform the worst. Previously, we hypothesized that the results were due to the model becoming more uncertain when the train labels were less extreme. Previous work has shown that an SER model's uncertainty depends on the values being predicted [31]. We conduct a similar analysis here, where we take the attribute predictions of the full MSP-Podcast test set done by the models trained on the sampled attribute subsets and calculate the variance of each prediction. Our analysis relies on one training trial. To get the variance of the predictions, we test the model 100 times using Monte Carlo Dropout [38] and use the mean of

TABLE IX
AVERAGE TEST CCC RESULTS OVER 10 TRAINING TRIALS, TRAINED AND TESTED WITH SAMPLED MSP-PODCAST ATTRIBUTE PARTITIONS.

	Sampled Train Set	Sampled Test Set					
		All	*Neg*	*Pos*	*Neg+Pos*	*Neut1*	*Neut2*
Arousal	*All*	0.654	*0.740*	0.698	0.716	0.526	0.503
	Neg	0.652	*0.745*	0.705	0.719	0.524	0.501
	Pos	0.657	*0.746*	0.712	0.726	0.527	0.499
	Neg+Pos	**0.669**	***0.759***	**0.725**	**0.737**	**0.534**	**0.507**
	Neut1	0.630	*0.703*	0.670	0.682	0.500	0.479
	Neut2	0.616	*0.694*	0.650	0.666	0.500	0.477
Valence	**Sampled Train Set**	**Sampled Test Set**					
		All	*Neg*	*Pos*	*Neg+Pos*	*Neut1*	*Neut2*
	All	0.297	*0.311*	0.285	0.294	0.280	0.274
	Neg	**0.354**	***0.373***	**0.343**	**0.354**	**0.326**	**0.319**
	Pos	0.300	*0.310*	0.297	0.302	0.279	0.273
	Neg+Pos	0.342	*0.358*	0.334	0.346	0.318	0.312
	Neut1	0.276	*0.295*	0.271	0.281	0.265	0.264
	Neut2	0.256	*0.271*	0.249	0.254	0.243	0.240
Dominance	**Sampled Train Set**	**Sampled Test Set**					
		All	*Neg*	*Pos*	*Neg+Pos*	*Neut1*	*Neut2*
	All	0.564	*0.625*	0.617	0.620	0.466	0.454
	Neg	**0.576**	***0.641***	**0.636**	**0.634**	0.470	**0.458**
	Pos	0.567	0.616	*0.624*	0.615	0.462	0.448
	Neg+Pos	0.571	*0.631*	0.628	0.629	0.468	0.455
	Neut1	0.544	*0.599*	0.581	0.585	**0.476**	0.454
	Neut2	0.533	*0.581*	0.568	0.568	0.459	0.444

TABLE X
AVERAGE TEST UAR RESULTS OVER 10 TRAINING TRIALS, TRAINED AND TESTED WITH SAMPLED MSP-PODCAST EMOTIONAL CATEGORY PARTITIONS.

Sampled Train Set	Sampled Test Set				
	All	*High*	*Mid*	*Neut1*	*Neut2*
All	0.456	0.458	*0.484*	0.397	0.403
High	**0.465**	**0.469**	***0.490***	**0.405**	**0.413**
Mid	0.457	0.462	***0.487***	0.393	0.399
Neut1	0.429	0.436	*0.447*	0.380	0.386
Neut2	0.444	0.445	*0.457*	0.390	0.395

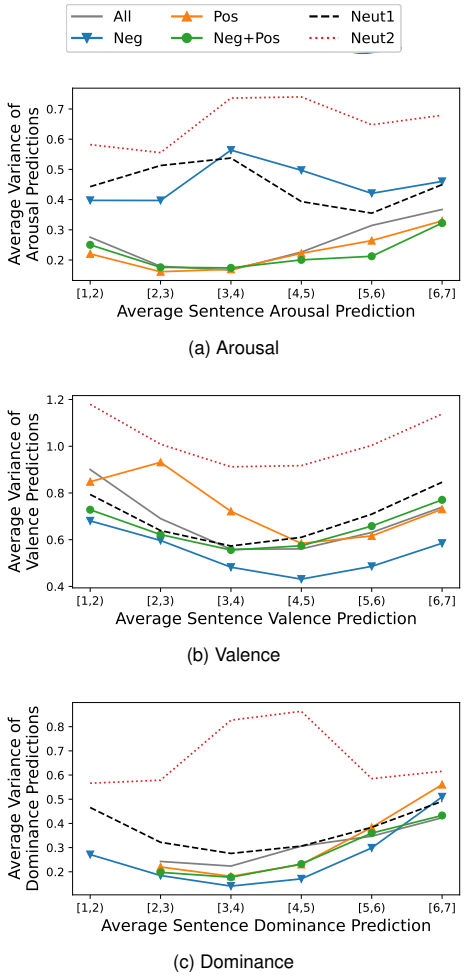


Fig. 13. Plots of the mean of the variance of the attribute predictions over 100 Monte Carlo dropout testing trials. Before taking the mean, the variances were binned according to their corresponding average sentence attribute predictions. The models were tested on the full MSP-Podcast test partition.

the 100 trials as the prediction and the variance as a measure of uncertainty.

To compare the model uncertainty of the differently biased models, we take the mean of the variance values of the sentences in the same attribute prediction bin for all models trained on the subsets. We plot the mean of the variances versus the attribute prediction bins in Fig. 13. For all attributes, the models trained on the *Neut2* subsets show the highest uncertainty for all prediction values. This coincides with the performance results, where the models trained with the *Neut2* subsets performed the worst. The models trained with the *Neut1* subsets also performed badly. However, their prediction uncertainties are not consistently higher than the better-performing models. The models trained on the *All* subsets perform much better than those trained with the unbiased subsets but still worse than the biased subsets. The prediction uncertainties of those models are not consistently above all the better performing models. Overall, the uncertainty of the model predictions somewhat coincides with the performance differences between models trained with biased and unbiased subsets. By comparing the uncertainty of specific prediction values, we observe that most models show lower uncertainty when predicting neutral values. These results contradict *Hypothesis (a)* since our models are generally more certain about neutral emotions as opposed to extreme emotions. A key observation is that the study of Sridhar and Busso [31], which motivated *Hypothesis (a)*, was conducted with hand-crafted features using fully connected layers. The analysis of this study uses Wav2vec2.0, which is a transformer-based *self-supervised learning* (SSL) model. The architectural differences can explain the disagreements in the relation between emotional attributes and uncertainty level observed in this study and Sridhar and Busso [31]. However, we note that our results do support *Hypothesis (b)*. They show lower prediction uncertainty when learning from more biased labels.

G. Prediction of Labels Analysis

The previous section showed that the uncertainty of our models coincides with the performance differences between the biased (*Neg* and *Pos*) and unbiased (*Neut1* and *Neut2*) labels. This section explores the model performance for different label values. In Section V-C, we hypothesized that the reason for the higher performance of models trained with biased labels is that the model can more easily learn from and predict more extreme labels (*Hypothesis (b)*). Since the biased labels

are pushed to more extreme values, they should be easier to predict. We take the model trained with the full MSP-Podcast train set and test it with each unsampled subset (shown in Table II) 10 times. We then split the test sets into label bins and randomly chose 50 sentences in each bin and took the CCC between the labels and the predictions. This procedure is done 10 times for all subsets and then the CCC values are averaged within each subset. We chose this procedure because the CCC is sensitive to the number of samples compared.

We plot the average CCC values versus the bin values in Fig. 14. In general, the model is more accurate for sentences with labels in the $[4, 6)$ range for arousal and dominance and $[2, 6)$ for valence as opposed to the more extreme labels close to 1 and 7. However, we gain further insight by noting that the performance differences for arousal and dominance between the biased and unbiased labels are due to the extreme emotion labels. When we compare the performance differences between biased and unbiased models, we observe larger differences in the extreme values of the attributes. The model has a higher accuracy at the extremes for the biased labels (*Neg* and *Pos*), which shows that the extreme emotions of the unbiased labels are more difficult to predict than the extreme emotion biased labels for arousal and dominance. The large gap supports the results in Table V. For valence, we do not see a subset that is easier to predict than the other subsets across most bins. The unbiased labels with negative values are easier to predict while the biased labels with positive values are easier to predict. Like in Section V-F, these results contradict *Hypothesis (a)*. The model can better predict neutral emotion labels as opposed to more extreme labels. However, our results partly support *Hypothesis (b)*. The model is more accurate when predicting biased labels as opposed to more unbiased ones. Moreover, this difference seems to be due to their respective extreme emotion labels. We can conclude that the models can better predict and learn from biased labels due to their more extreme labels. However, the reason for this cannot be explained by *Hypothesis (a)*. Further research into this problem is needed to fully explain these findings.

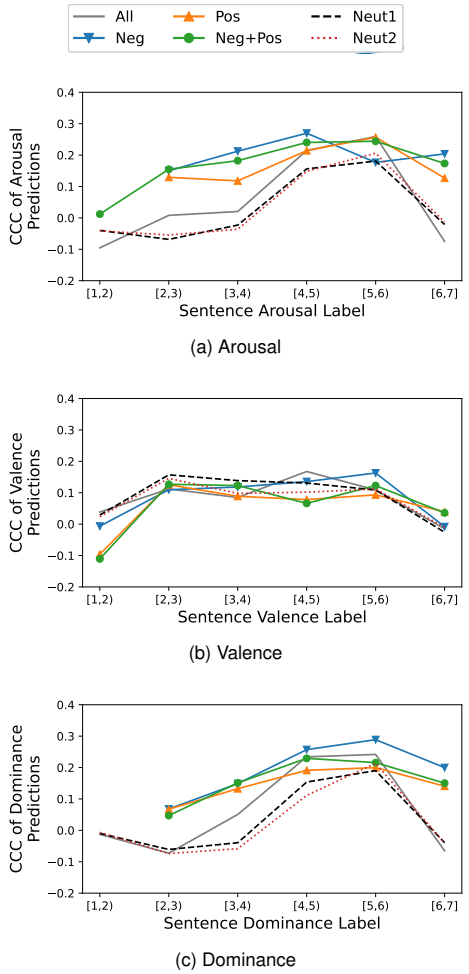


Fig. 14. Plots of the average test CCC results over 10 testing trials using the model trained with the full MSP-Podcast train set and tested on the unsampled subsets.

VI. Discussion

In Section I, we introduced the three questions we wanted to answer with this paper. The first and second questions are: *does the affective priming effect occur in emotional attribute labeling of speech?* and *is the affective priming effect observed in categorical labels?* We answered these questions in Section IV. Our experiments in Section IV-A explored if we could observe affective priming in two emotional datasets, the MSP-Podcast and BIIC-Podcasts corpora. We indeed observed that when evaluators annotated previous sentences with extreme values, they were more likely to annotate the next sentence towards that extreme, which is consistent with the affective priming effect. In Section IV-B, we explored if we could see a similar effect in the emotional category annotations of both corpora. We were also able to observe such an effect, where annotators were more likely to choose a category as the primary emotion if they had been primed with that category.

The third question is *Does the affective priming effect impact SER performance?* To answer this question, we first created an expected affective priming bias measure for the MSP-Podcast corpus (Sec. IV-C) to be able to define subsets of the corpus with clear affective priming effects (Sec. V-A). The biases for the attribute labels were concentrated around a 0 bias, showing that most of the attribute labels of the MSP-Podcast corpus are not greatly impacted by the affective priming effect. The SER evaluations were presented in Section V. We used a Wav2Vec 2.0-based SER model (Sec. V-B) to perform our experiments. In Section V-C, we trained the model with the full MSP-Podcast train partition and tested the model with the differently biased subsets. The results showed that the model can predict the biased labels with higher performance than the unbiased labels. In Section V-D, we trained the model with the differently biased subsets to explore how the affective priming effect affects SER modeling. In Section V-E, we also evaluate training and testing the SER models with biased and unbiased partitions.

Similar to the previous results, all the models, regardless of which subset was used to train them, performed better when testing on the biased subsets. The models had much worse results when testing on the unbiased subsets. All results clearly show that adding labels biased by affective priming leads to better SER results, even though the biased labels have much

worse inter-evaluator agreements (Sec. V-A1). We hypothesize that a reason for the SER modeling results is that the SER model becomes more confused by the unbiased labels. We explore this possibility in Section V-F, where we calculated the variance of the model predictions using Monte Carlo dropout to give us a measure of the uncertainty of the model when it was trained using differently biased labels. In general, the models trained with unbiased labels show higher uncertainty. However, the pattern is not as clear as the performance results. The uncertainty of the models does not completely explain the performance increase when going from a model trained with unbiased labels to more biased labels. We also explored the possibility that the model has a harder time predicting the unbiased labels because they are less extreme. We explore this in Section V-G, where we took the CCC between labels and predictions from the model trained with the full MSP-Podcast train set for different attribute values using bins. We observed that the model does seem to have a harder time predicting the unbiased extreme labels compared to the biased extreme labels, at least for arousal and dominance. However, the model can best predict the neutral labels without much difference between the biased and unbiased labels. More evaluations are needed to understand these interesting results fully.

The consistent and surprising performance results shown in Sections V-C, V-D and V-E could be explained by the tendency of people to anchor themselves when dealing with uncertain situations [19]. When evaluators are not given some form of prime or a similarly uncertain prime (e.g., no previous sentences or more nuanced previous sentences), they might become even more uncertain. They might anchor to a part of the sentence that is less related to the emotional content (e.g., volume, accent, gender, or audio quality). Other annotators could consistently anchor to a similar aspect of the sentence, explaining the higher inter-evaluator agreement, but each sentence could have different anchors. This could lead to having some emotional labels unbiased by the affective priming effect that the model connects to aspects of the audio unrelated to emotional content, leading to somewhat confident predictions that are not good emotional predictions. Furthermore, emotionally unrelated aspects of the speech audio could be easier for the model to extract, as the Wav2Vec 2.0 model is a general speech model. The model might focus on the "easier" label-feature pairs, especially when they are more abundant. Encouragingly, the models trained with a mixture of biased and unbiased labels perform close to the biased subsets, even when the ratio of biased to unbiased is low. The performance gap is closed when considering the larger amount of data available when combining all available labels.

Our experiments and analyses focused on the presence of the affective priming effect in emotional annotations, which were conducted by providing isolated sentences in random order. Our analysis on annotator agreement (Section V-A1; Fig. 6) showed that this effect is present in all annotators, regardless of their measured reliability. Affective priming is an unconscious phenomenon that is hard to "train away," and would be cost-prohibitive to do so. However, our analysis on the expected affective priming bias of the MSP-Podcast corpus (Fig. 12) shows that most sentences experience neutral priming since the ordering of sentences is inconsistent between annotators and most spoken natural sentences are neutral. Furthermore, our modeling results show that a mix of primed and unprimed sentences still have good modeling results. Therefore, increasing the number of annotations in an emotional corpus provides a good way of diminishing the affective priming effect on the whole corpus, as long as those annotations are subject to a variety of priming. The out-of-order and inconsistent annotation sequences used in the MSP-Podcast and BIIC-Podcast corpora ensure that annotators experience a variety of affective priming.

Another common annotation method is annotating consecutive sentences in their original context, i.e., annotating each sentence in a conversation in the order of the original conversation. This method does not eliminate affective priming but instead ensures that all annotators receive the same priming. In real-life conversations, affective priming still occurs. Therefore, the second annotation method is a valid annotation process that does not need to remove affective priming. However, the labels derived from these annotations are not necessarily consistent with the labels from out-of-context annotations [17], [18], [39]. Moreover, labels derived from in-context annotations often require a model that takes context into account; SER models that do not incorporate context into their predictions perform better for out-of-context labels [18]. Out-of-context labels are useful for applications where context is not available, and in those cases, inconsistent ordering during annotation is necessary to reduce the effect of affective priming.

VII. Conclusion

In this paper, we showed that the affective priming effect affects emotional annotations of speech when conducted out-of-context. We showed this effect for both the MSP-Podcast corpus and the BIIC-Podcast corpus, demonstrating that the affective priming effect is general and we should expect it regardless of the language. Furthermore, we showed that affective priming creates a bias in the emotional labels of the MSP-Podcast corpus that affects SER models trained on the data. SER models trained on labels affected by priming have more certain predictions and better performance results. We partially explain these results by exploring the certainty of the predictions. Fortunately, the MSP-Podcast corpus and similar datasets contain a mixture of labels that are affected to different degrees by the affective priming effect. The SER modeling results in this paper suggest that SER models can successfully learn with such a mixture of labels.

Further exploration of the affective priming effect on emotional annotations is useful, as this paper only focuses on annotations of speech. Many studies have proposed multimodal strategies to recognize emotions [40]–[42]. It would be interesting to understand how the level of priming changes as new modalities are presented to the evaluators. Another interesting direction is to understand the effect of priming in databases annotated with time-continuous annotations [43]. Since the field of affective computing relies on effective emotional labels, we hope that our work in this paper encourages the field

to explore this effect and other subconscious phenomena that can affect the emotional labels that we rely on. Understanding these effects is important to bring new insights to design better emotional computational models.

REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] G. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 248–255.

[8] ——, "The ordinal nature of emotions: An emerging approach," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 16–35, January-March 2021.

[9] K. C. Klauer and J. Musch, "The psychology of evaluation: Affective processes in cognition and emotion," *Affective priming: Findings and theories*, pp. 7–49, 2003.

[10] J.-S. Lee, J. Choi, J. H. Yoo, M. Kim, S. Lee, J.-W. Kim, and B. Jeong, "The effect of word imagery on priming effect under a preconscious condition: an fmri study," *Human brain mapping*, vol. 35, no. 9, pp. 4795–4804, 2014.

[11] H. R. Bosker, "Putting laurel and yanny in context," *The Journal of the Acoustical Society of America*, vol. 144, no. 6, pp. EL503–EL508, 12 2018.

[12] 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.

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

[14] 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.

[15] S. G. Upadhyay, W.-S. Chien, B.-H. Su, L. Goncalves, Y.-T. Wu, A. N. Salman, C. Busso, and C.-C. Lee, "An intelligent infrastructure toward large scale naturalistic affective speech corpora collection," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.

[16] L. Martinez-Lucas, A. Salman, S.-G. Leem, S. Upadhyay, C.-C. Lee, and C. Busso, "Analyzing the effect of affective priming on emotional annotations," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*, Cambridge, MA, USA, September 2023.

[17] 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.

[18] 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.

[19] B. Seymour and S. M. McClure, "Anchors, scales and the relative coding of value in the brain," *Current Opinion in Neurobiology*, vol. 18, no. 2, pp. 173–178, 2008, cognitive neuroscience. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959438808000676

[20] R. H. Fazio, "On the automatic activation of associated evaluations: An overview," *Cognition and Emotion*, vol. 15, no. 2, pp. 115–141, 2001.

[21] M. Lohse and M. Overgaard, "Emotional priming depends on the degree of conscious experience," *Neuropsychologia*, vol. 128, pp. 96–102, 2019, neural Routes to Awareness in Vision, Emotion and Action: A tribute to Larry Weiskrantz. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0028393217304049

[22] S. T. Murphy and R. B. Zajonc, "Affect, cognition, and awareness: affective priming with optimal and suboptimal stimulus exposures." *Journal of personality and social psychology*, vol. 64, no. 5, p. 723, 1993.

[23] R. Mueller, S. Utz, and C.-C. Carbon, "Face adaptation and face priming as tools for getting insights into the quality of face space," *Frontiers in Psychology*, vol. 11, 02 2020.

[24] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov, "Label Studio: Data labeling software," 2020-2022, open source software available from https://github.com/heartexlabs/label-studio. [Online]. Available: https://github.com/heartexlabs/label-studio

[25] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 415–420.

[26] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Not all features are equal: Selection of robust features for speech emotion recognition in noisy environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 6447–6451.

[27] L. Goncalves, A. Salman, A. Reddy Naini, L. Moro-Velazquez, T. Thebaud, P. Garcia, N. Dehak, B. Sisman, and C. Busso, "Odyssey 2024 - speech emotion recognition challenge: Dataset, baseline framework, and results," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, Quebec, Canada, June 2024, pp. 247–254.

[28] H.-C. Chou, W.-C. Lin, C.-C. Lee, and C. Busso, "Exploiting annotators' typed description of emotion perception to maximize utilization of ratings for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 7717–7721.

[29] 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.

[30] J. Fleiss, *Statistical methods for rates and proportions*. New York, NY, USA: John Wiley & Sons, 1981.

[31] K. Sridhar and C. Busso, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *IEEE international conference on acoustics, speech and signal processing (ICASSP 2020)*, Barcelona, Spain, May 2020, pp. 8384–8388.

[32] W.-N. Hsu *et al.*, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *ArXiv e-prints (arXiv:2104.01027)*, pp. 1–9, April 2021.

[33] 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," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, September 2023.

[34] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS 2020)*, vol. 33, Virtual, December 2020, pp. 12 449–12 460.

[35] T. Wolf *et al.*, "HuggingFace's transformers: State-of-the-art natural language processing," *ArXiv e-prints (arXiv:1910.03771v5)*, pp. 1–8, October 2019.

[36] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned Wav2vec 2.0/HuBERT benchmark for speech emotion recognition, speaker verification and spoken language understanding," *ArXiv e-prints (arXiv:2111.02735)*, pp. 1–7, November 2021.

[37] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, San Diego, CA, USA, May 2015, pp. 1–13.

[38] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *International*

Conference on Machine Learning (ICML 2016), New York, NY, USA, June 2016, pp. 1050–1059.

[39] L. Martinez-Lucas, W.-C. Lin, and C. Busso, "Analyzing continuous-time and sentence-level annotations for speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 1754–1768, July-September 2024.

[40] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso, "Versatile audio-visual learning for emotion recognition," *IEEE Transactions on Affective Computing*, vol. 16, no. 1, pp. 306–318, January-March 2025.

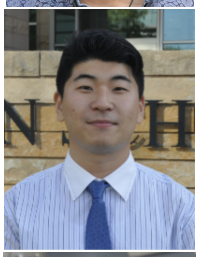
[41] L. Goncalves and C. Busso, "Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2156–2170, October-December 2022.

[42] ——, "AuxFormer: Robust approach to audiovisual emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 7357–7361.

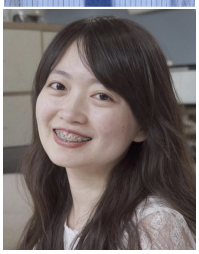
[43] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *Interspeech 2020*, Shanghai, China, October 2020, pp. 1823–1827.

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 is currently a visiting scholar in the Language Technologies Institute at Carnegie Mellon University (CMU). 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 AAAC.

Ali N. Salman Received his B.S. and M.S. Degrees in Computer Science at Indiana State University in 2015 and 2017, respectively. He is currently pursuing his Ph.D. degree in Electrical Engineering at the University of Texas at Dallas. His current research interests include affective computing, deep learning, and facial analysis.



Seong-Gyun Leem (S'21) received his B.S. and M.S. degree in Computer Science and Engineering at Korea University, Seoul, South Korea in 2018 and 2020, respectively. He is currently pursuing his Ph.D. degree in Electrical Engineering at the University of Texas at Dallas. His current research interests include speech emotion recognition, noisy speech processing, and machine learning.



Woan-Shiuan Chien (S'23) currently is a PhD student at the Electrical Engineering (EE) Department of National Tsing Hua University (NTHU), Taiwan. She received her B.S. degree in electrical engineering from Chung Yuan Christian University, Taiwan in 2015 and her M.S. degree in electrical engineering from the National Chung Cheng University (CCU), Taiwan in 2016. Her research interests are in human-centered behavioral signal processing and automatic speech emotion recognition. She was the recipient of the Outstanding Doctoral Students Program sponsored by the Taiwan Science and Technology Council (NSTC) (2022), and the travel grant sponsored by IEEE Signal Processing Society (2023) and the Association for Computational Linguistics and Chinese Language Processing (ACLCLP) (2023). She is a Student Member of the AAAC, ACM, ACLCLP, and IEEE Signal Processing Society (SPS).

Shreya G. Upadhyay (S'23) is currently pursuing a PhD degree in Electrical Engineering at National Tsing Hua University (NTHU), Taiwan. She obtained her BE degree in computer engineering from Mumbai University, India in 2013, and her MTech degree in computer engineering from K. J. Somaiya College of Engineering, India in 2018. Her research interests include behavioral speech signal processing, speech emotion recognition, automatic speech recognition, and acoustic sound event detection. She is a student member of International Speech Communication Association (ISCA), AAAC, European Association for Signal Processing (EURASIP), and the IEEE Signal Processing Society (SPS).

Chi-Chun Lee (M'13, S'20) is a Professor at the Department of Electrical Engineering of the National Tsing Hua University (NTHU), Taiwan. He received his B.S. and Ph.D. degree both in Electrical Engineering from the University of Southern California, USA in 2007 and 2012. His research interests are in speech and language, affective computing, health analytics, and behavioral signal processing. He is an associate editor for the IEEE Transaction on Affective Computing (2020-), the IEEE Transaction on Multimedia (2019-2020), the Journal of Computer Speech and Language (2021-), the APSIPA Transactions on Signal and Information Processing and a TPC member for APSIPA IVM and MLDA committee. He serves as the general chair for ASRU 2023, an area chair for Interspeech 2016, 2018, 2019, senior program committee for ACII 2017, 2019, publicity chair for ACM ICMI 2018, late breaking result chair for ACM ICMI 2023, sponsorship and special session chair for ISCSLP 2018, 2020. He is the recipient of the the NSTC Outstanding Research Award (2024), the Foundation of Outstanding Scholar's Young Innovator Award (2020), the CIEE Outstanding Young Electrical Engineer Award (2020), the IICM K. T. Li Young Researcher Award (2020), the NTHU Industry Collaboration Excellence Award (2021), and the MOST Futuretek Breakthrough Award (2018, 2019). He led a team to the 1st place in Emotion Challenge in Interspeech 2009, and with his students won the 1st place in Styrian Dialect and Baby Sound subchallenge in Interspeech 2019. He is a co-author on the best paper award/finalist in Interspeech 2008, Interspeech 2010, IEEE EMBC 2018, Interspeech 2018, IEEE EMBC 2019, APSIPA ASC 2019, IEEE EMBC 2020, and the most cited paper published in 2013 in Journal of Speech Communication. He is also an ACM and ISCA member.

Carlos Busso (S'02-M'09-SM'13-F'23) is a Professor at Language Technologies Institute, Carnegie Mellon University, where he is also the director of the *Multimodal Speech Processing* (MSP) Laboratory. He 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. His research interest is in human-centered multimodal machine intelligence and applications, focusing on the broad areas of speech processing, affective computing, multimodal behavior generative models, and foundational models for multimodal processing. He was selected by the School of Engineering of Chile as the best electrical engineer who graduated in 2003 from Chilean universities. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. His students received the third prize IEEE ITSS Best Dissertation Award (N. Li) in 2015, and the AAAC Student Dissertation Award (W.-C. Lin) in 2024. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He received the Best of IEEE Transactions on Affective Computing Paper Collection in 2021 (with R. Lotfian) and the Best Paper Award from IEEE Transactions on Affective Computing in 2022 (with Yannakakis and Cowie). In 2023, he received the Distinguished Alumni Award in the Mid-Career/Academia category by the *Signal and Image Processing Institute* (SIPI) at the University of Southern California. He received the 2023 ACM ICMI Community Service Award. He is currently a Senior Area Editor of IEEE/ACM Speech and Language Processing. He is a member of AAAC and a senior member of ACM. He is an IEEE Fellow and an ISCA Fellow.