Analyzing the Effect of Affective Priming on Emotional Annotations

Luz Martinez-Lucas[1], Ali Salman[1], Seong-Gyun Leem[1], Shreya G. Upadhyay[2], Chi-Chun Lee[2], Carlos Busso[1]
[1]*Department of Electrical and Computer Engineering, University of Texas at Dallas, USA*
[2]*Department of Electrical Engineering, National Tsing Hua University, Taiwan*

***Abstract*—In the field of affective computing, emotional annotations are highly important for both the recognition and synthesis of human emotions. Researchers must ensure that these emotional labels are adequate for modeling general human perception. An unavoidable part of obtaining such labels is that human annotators are exposed to known and unknown stimuli before and during the annotation process that can affect their perception. Emotional stimuli cause an affective priming effect, which is a pre-conscious phenomenon in which previous emotional stimuli affect the emotional perception of a current target stimulus. In this paper, we use sequences of emotional annotations during a perceptual evaluation to study the effect of affective priming on emotional ratings of speech. We observe that previous emotional sentences with extreme emotional content push annotations of current samples to the same extreme. We create a sentence-level bias metric to study the effect of affective priming on *speech emotion recognition* (SER) modeling. The metric is used to identify subsets in the database with more affective priming bias intentionally creating biased datasets. We train and test SER models using the full and biased datasets. Our results show that although the biased datasets have low inter-evaluator agreements, SER models for arousal and dominance trained with those datasets perform the best. For valence, the models trained with the less-biased datasets perform the best.**

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

I. INTRODUCTION

It is important that labels for emotion recognition problems are reliable. For most corpora, emotional labels are derived from perceptual evaluations, where annotators listen or watch a stimulus and report their perceived emotions using the provided descriptors (e.g., emotional attributes or categories). These perceptual evaluations often require a rater to sequentially annotate several samples in a session [1]–[4]. Previous studies on emotional labels for affective computing have observed that humans have an easier time recording their emotional perceptions during relative comparisons (e.g., which of the two samples is happier?). Yannakakis *et al.* [5] argued that annotations conducted in an ordinal manner, where annotators are asked to rank samples on a scale, are better at representing the underlying human perception. The hypothesis is that we consciously or unconsciously use "anchors" to assess the emotional content of a sample. Therefore, when the comparison is explicit (e.g., comparing two samples), we can provide more reliable labels. Since evaluators often rate multiple samples in a row, it is expected that samples that are previously annotated in a session play the implicit role of "anchoring" the emotional content for the next samples to be annotated. Can the emotion of previous samples affect the emotional judgment of the current sample? Can this "anchoring" effect be quantified? Is this "anchoring" effect important for emotion recognition tasks?

The influence of the emotional content of previous stimuli in the assessment of the emotional content of a sample is known as affective priming. Priming, in general, is a phenomenon in which information or an event affects how a person reacts to subsequent related information or events. Priming has been heavily observed in the study of memory, usually in studies showing how words are easier or faster to remember when they are preceded by perceptually or conceptually related words [6]. Priming has also been observed in studies focused on human perception [7]. Affective priming refers to the same phenomenon when the prime and target samples are specifically related within the emotional space [8]. This paper focuses on studying how affective priming (i.e., having annotators listen to emotional speech before rating new speech) affects the emotional annotations of speech segments. We also study how affective priming affects *speech emotion recognition* (SER) tasks.

We focus on analyzing affective priming in the rating of speech using the emotional attributes of arousal (calm versus active), valence (negative versus positive), and dominance (weak versus strong). These attributes have become popular in the field of SER since they allow for more nuanced emotional ratings than using emotional classes [9]–[11]. Our study uses the perceptual evaluation sessions of the publicly available MSP-Podcast corpus [3], consisting of over 850,000 annotations collected over more than 68,000 sessions (release 1.10). The large size of this perceptual evaluation effort provides the perfect platform to investigate affective priming. Our approach compares the label provided by a target evaluator with the average score provided by the other annotators for the sample. The key in our approach is to condition this difference with the emotional scores provided to the previous samples in the session by the target evaluator.

Our analyses show that affective priming has a clear effect on emotional annotations. When annotators rate a sentence in one extreme of the emotional attribute, they will tend to rate the next sentence with values that are closer to this

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)

979-8-3503-2743-4/23/$31.00 ©2023 IEEE

extreme value, deviating from the average score provided by other evaluators (e.g., expected rating). Since affective priming creates biases in emotional annotations, our next goal is to quantify this effect. We assign an expected affective priming bias to each annotation of an emotional sample based on the scores provided to previous samples. Then, we estimate a sentence-level affective priming bias by averaging the metric across the annotations. With this approach, we estimate the affective priming bias for all the sentences in the corpus. Then, we divide the dataset into groups with negative, neutral, and positive biases to explore how affective priming in emotional annotations affects SER modeling. We train and test an SER model on the groups. We see that for arousal and dominance, the biased groups give the best performances overall. However, the SER modeling experiments on valence show that the less-biased groups perform the best.

II. Related Work

Affective priming is a phenomenon where the processing of emotional information is easier when preceded by stimuli that are similar in emotional content [8], [12]. Previous work on affective priming has focused on words [12], pictures [13], and faces [14]. It often includes just the priming effect of valence. Previous studies have shown that priming with an emotional stimulus will push people's perception of an ambiguous object towards the valence of the stimulus [8], [13], [15]. This result is observed with either positive or negative valence priming. Many aspects of the prime stimuli, such as how they are presented and the downstream task, can affect the resulting priming effect. More emotionally extreme stimuli seem to elicit clearer affective priming effects, especially during tasks that ask an annotator to evaluate the emotionality of the target stimulus, as opposed to evaluating an unrelated aspect such as its color [8]. The reported level of awareness from annotators also affects the results of affective priming. Lohse and Overgaard [13] described how the effect of affective priming in the emotional perception of ambiguous images increases when the reported level of awareness of the annotators is higher. However, Murphy and Zajonc [15] showed the opposite effect, where lower awareness leads to more pronounced priming effects in the evaluation of faces. This finding was explained by suggesting that giving too much time to the prime faces causes annotators to see other aspects of the prime images, which makes the prime emotionality more ambiguous and, therefore, has less effect on the rating of the target stimulus.

Although there have not been many works dealing with affective priming in the speech domain, some studies have looked at other causes of bias in emotional annotations. Nussbaum *et al.* [16] evaluated the perception of emotion by human annotators by changing vocal cues in the speech signals and observing how the perception changed. However, this study focused on emotional adaptation, which is a different effect. Adaptation is a phenomenon where a person's perception of a familiar object is *adapted* by showing slightly different versions of the object. Adaptation usually leads to a perceptual bias opposite to the adaptor, while priming leads to a bias towards the prime [14]. Chien and Lee [17] addressed the bias introduced by the gender of the annotator in the emotional perception of speech. They focused on how bias introduced by the annotator's gender causes issues in SER. While this type of bias is different from the bias introduced by the priming effect, the study presents a unique way of dealing with these biases in SER modeling. Instead of just minimizing the biases in a model, the strategy also introduces a "switch" that allows the model to become intentionally biased toward one gender's perception.

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

III. Resources

A. The MSP-Podcast Corpus

Given our research questions, we require a large database annotated by multiple raters. The perceptual evaluation in the corpus should have several sessions, including multiple sentences in each session. We also need that the presentation of the sentences is out of temporal order in a dialog to reduce the effect of context. The publicly available MSP-Podcast corpus [3] satisfies all these requirements, so it is ideal for this study. The corpus consists of sentences sourced from publicly available online audio sources. The labels are annotated with perceptual evaluations using a modified version of the crowdsourcing protocol introduced by Burmania *et al.* [18]. The labels include ratings for the emotional attributes of arousal (calm to active), valence (negative to positive), and dominance (weak to strong). These emotional attributes are annotated with a 7-point Likert scale (i.e., values between 1 and 7). The corpus has annotations for primary and secondary emotional categories, although we do not use them in this study. At least five annotators annotate each sentence. We rely on release 1.10 of the MSP-Podcast corpus, which contains 104,267 sentences (166 hours and 9 mins). The sentences are split into partitions. We have 63,076 sentences in the train set, 10,999 in the development set, and 16,903 in the test set. The corpus has a second test set that we do not use.

B. Ordered Perceptual Evaluation Sessions

We define a *session* as the set of all evaluations that are sequentially conducted by a given rater within a period of time. If the time between annotations was longer than 15 minutes, we split the evaluation into different sessions. Importantly for this study, we keep the ordering of the sentences as seen by the annotator for each of the sessions. Figure 1 illustrates this process, where a rater is annotating the valence score for sentence "g" after providing a score equal to either "1" or "2" to the previous four sentences (i.e., sentences "c" to "f" are perceived very negative). Then, we split these sessions into two-minute blocks called annotation *sequences*,

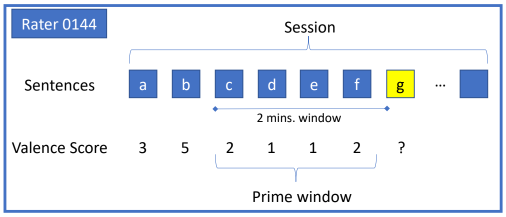
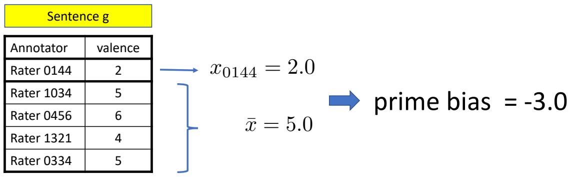


Fig. 1. Process of selecting the prime window for an annotation. In the figure, we see an annotation session done by Rater 0144. We select the rater's valence annotation for sentence "g" (in yellow). We find the prime window by looking at the previous valence annotations that happened within two minutes of the annotation of sentence "g".



Annotator	valence
Rater 0144	2
Rater 1034	5
Rater 0456	6
Rater 1321	4
Rater 0334	5

Fig. 2. Process of calculating the prime bias for a single annotation. In the figure, we see all the valence annotations available for sentence "g". The annotation we focus on is the one done by Rater 0144, x_{0144}. Therefore, we take the average, $\bar{x}$, of the annotations done by the other raters and subtract it from x_{0144}. This difference is our prime bias.

each sequence consists of a *current* rating and all previous ratings in the session started within two minutes of the start of the current rating. We call these previous ratings the *prime window*. After processing, we end up with 68,385 sessions. We obtain 853,097 sequences from these sessions, with an average of 3.15 sentences in each sequence.

IV. Quantifying Affective Priming

Our first question in this study is whether the emotionality of previously annotated emotional audio affects emotional ratings of the next stimuli (emotional priming). To answer this question, we first identify annotation sequences with low, neutral, and high priming for the emotional attribute ratings. For arousal, low and high priming represent calm and active priming, for valence negative and positive priming, and for dominance weak and strong priming. For this section, we only consider sequences with at least five annotations. We define sequences with low priming for an attribute as sequences in which all the scores in a prime window are either "1" or "2" for that attribute. Sequences with neutral priming have scores in the prime window between "3" and "5". Sequences with high priming have all the sentences in the prime window with scores of either "6" or "7". This requirement is imposed on all the sentences included in the prime window, whether it is 4, 5, or more sentences annotated within the last two minutes. The highlighted sequence in Figure 1 qualifies as a case of *low priming* since the annotations in the prime window are either "1" or "2".

A key question is how to quantify the affective priming. We consider all the annotations provided to a given sentence. We take the current rating of each sequence and subtract from it the average rating assigned to the sentence after excluding the annotation of the target rater. This estimation assumes that averaging the scores from other evaluators compensates for the effects of their own affective priming, providing a valid reference. Figure 2 illustrates this process, connecting it with the example in Figure 1. In this illustration, other raters assigned an average valence score of 5.0 for sentence "g". If the target rater provides a valence score of "2" for that sentence, the affective prime bias will be -3.0. Since the difference is negative, the current rating is lower than the expected current label for the sentence, which we attribute to the effect of anchoring her/his emotional perception on the negative emotions observed by the rater in the previous sentences.

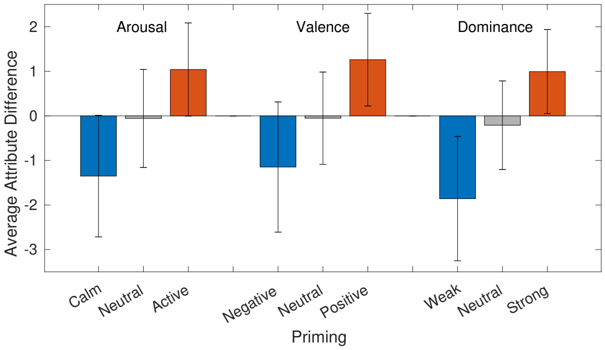
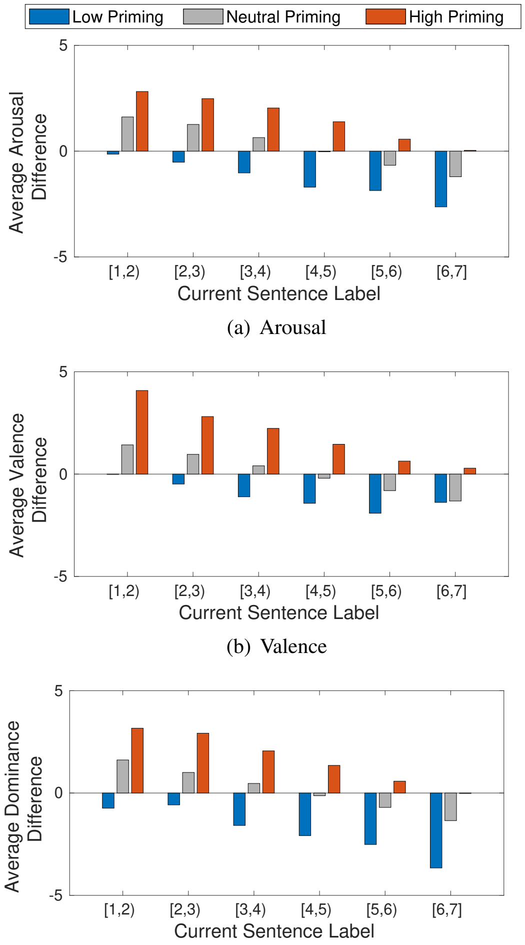


Fig. 3. Average difference between current ratings and the expected label for each sentence. The expected label is the average of all the ratings for the sentence excluding the current rating. The annotations are separated into priming groups depending on their prime window.

A. Annotation-Based Affective Priming Bias

We average the differences of each sequence over each priming group for each attribute. Figure 3 shows these average differences as well as the standard deviation of the differences. As expected, neutral priming resulted in current ratings that are very similar to the current labels, which also validates that using the average of the non-current ratings for a sentence is sufficient for approximating ratings resulting from neutral priming. The results for the low and high priming groups show that a specific extreme of emotional priming pushes the current ratings toward that extreme. Annotators who listen to consecutive sentences that are all perceived in one extreme of the emotional attribute will rate the following sentences closer to that extreme. If a corpus is heavily biased towards an extreme in the emotional space, we expect that the scores will be further pushed towards that extreme, worsening the bias. This is an important observation that few researchers have addressed in affective computing.

Previous studies have shown that the effect of priming also depends on the content of the current event [8], [16]. Our next question is whether the impact of emotional priming varies across sentences labeled with different emotional scores. For example, are sentences with more extreme emotions more susceptible to change in the presence of affective priming? To answer this research question, we conduct the same experiments as in Figure 3, but we bin the sequences according to the average score assigned to a sentence after excluding the



(a) Arousal

(b) Valence

(c) Dominance

Fig. 4. Average difference between current ratings and the expected label for each sentence. The annotations are separated into priming groups depending on their prime window, as well as separated into bins using the expected label for the sentence.

annotation of the target rater. Figure 4 shows the results for arousal, valence, and dominance. In general, sentences with the opposite emotional content as the priming are affected the most. As a result, the emotional perception of sentences with attribute scores in the extremes is more susceptible to being influenced by affective priming. Consistently, we observe that the low and high primings push the scores to their respective extremes.

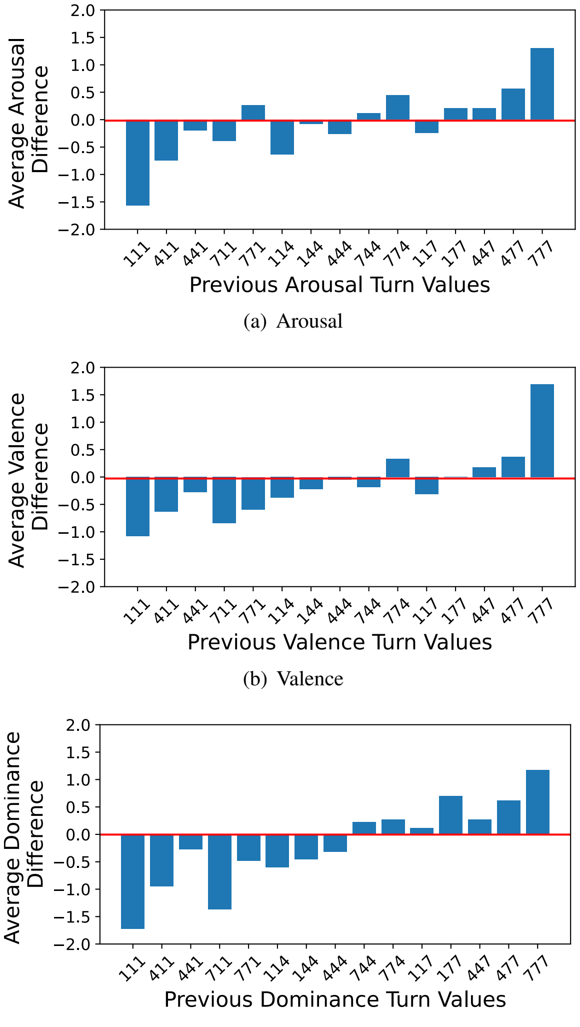
B. Sentence-Based Affective Priming Bias

We see from our previous experiments that the emotionality of sentences in the prime window has a clear effect on the rating of the current sentence in a sequence. This effect leads to a *bias* in the rating of a current sentence towards the emotion included in the prime window. This section aims to aggregate the effect of affective priming at the sentence level. Our strategy estimates the expected affective priming bias given the annotations of the previous sentences that were annotated. Our first step is choosing how many previous sentences we will use to quantify the affective priming bias. Lohse and Overgaard [13] showed that affective priming has an increased effect when there is longer exposure to the priming stimulus, but Murphy and Zajonc [15] showed the opposite result when using emotional faces as primes. Therefore, this study analyzes the effect of different emotional values and the number of sentences used to condition the emotion of the target sentence.

Consider the sequence in Figure 1. The last three previous scores before annotating sentence “g” are “2” (sentence “f”), “1” (sentence “e”) and “1” (sentence “d”). The bias for this sentence in Figure 2 is -3.0. Given the size of the corpus and the number of annotations, it is expected that the *priming patterns* 112 will appear multiple times in other sequences. Therefore, we can estimate the expected affective priming bias for this particular pattern by estimating its average across all the relevant sequences. We use this approach for all possible combinations, creating a table with the expected affective priming bias. At the beginning of a session, the sentences will not have three previous sentences previously annotated. Therefore, we also estimate the affective priming bias for patterns with two previous sentences and one previous sentence. A problem arises for the first sentence in a session. With the absence of a previous sentence, we set the affective priming bias to zero. This choice is justified by the results in Figure 5, which will be discussed next.

Figure 5 shows the results for 15 of the priming patterns. These patterns were chosen because they clearly show the effect of different exposures to affective priming. The figure shows that the last rating in the prime window plays a large role in the resulting bias (i.e., the most recent sentence annotated by the rater). However, the bias becomes more pronounced when the last rating is preceded by the same ratings. The plots show that the affective priming bias is higher in the “111” and “777” patterns, and it gets lower as we have less clear exposure (from “x11” to “xx1” and from “x77” to “xx7”). We also look at the resulting bias when a sentence has no prime window (e.g., at the beginning of each session, or after taking a break longer than two minutes). This result is shown in the red horizontal line in Figure 5. The red line shows a very small negative bias in all three attributes. Given that the red lines are almost zero, the choice of setting the expected affective priming bias to zero for sentences without previous annotated sentences is justified.

We estimate the priming bias for the sentence with the tabulated affective priming bias for each pattern. Figure 6 shows the process for sentence “g”, following the illustrations presented in Figures 1 and 2. The first rater (0144) has a priming pattern of “112” resulting in an affective priming bias equal to -0.749. The second rater (1093) started the session with the sentence “g” so there is no bias. The last step is to average the affective priming bias across the annotations, resulting in a sentence-level bias score. This process is independently calculated for arousal, valence, and dominance, resulting in three affective priming bias metrics per sentence. Figure 7 shows the distribution of the sentence biases for the MSP-Podcast 1.10 corpus. Most of the sentences in the dataset have a low bias. For example, the set of sentences

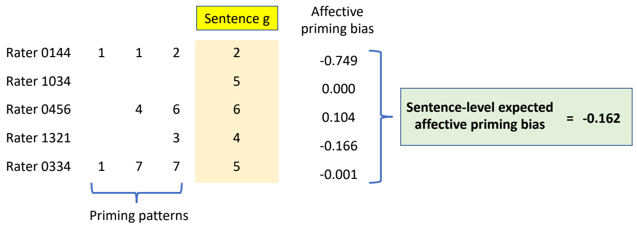


(a) Arousal

(b) Valence

(c) Dominance

Fig. 5. Average difference between current ratings and the expected label for each sentence. The annotations are separated into bins using the last three annotations in their prime window.



				Sentence g	Affective priming bias
Rater 0144	1	1	2	2	-0.749
Rater 1034				5	0.000
Rater 0456		4	6	6	0.104
Rater 1321			3	4	-0.166
Rater 0334	1	7	7	5	-0.001

Fig. 6. Figure showing how the bias measure is calculated for a sentence. It shows all the valence annotations for sentence “g”. Each annotation has a priming pattern, and each pattern has a corresponding affective priming bias. We assign the bias of the pattern to the annotation. Then, we average the biases for all the annotations, which becomes the bias sentence measure.

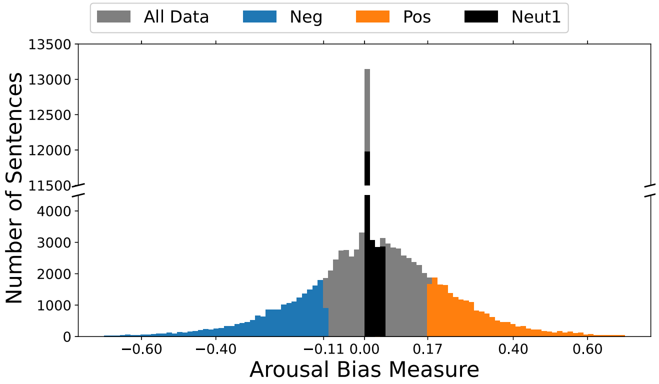


Fig. 7. Distribution of sentence-level expected affective priming biases for arousal in version 1.10 of the MSP-Podcast corpus. The bias distributions for valence and dominance are similar to the bias distribution for arousal.

TABLE I
NUMBER OF SENTENCES IN EACH MSP-PODCAST CORPUS SUBSET.

Corpus Subset	Subset Partition	Emotional Attribute		
		Arousal	*Valence*	*Dominance*
Neg	Train	13,207	12,807	12,936
	Dev	1,395	1,949	1,442
	Test	2,990	3,808	3,224
Pos	Train	12,586	12,823	12,759
	Dev	1,821	1,481	1,790
	Test	3,568	2,837	3,340
Neut1	Train	12,606	12,041	12,009
	Dev	3,900	3,648	3,639
	Test	3,004	2,814	2,855
Neut2	Train	11,079	12,543	11,587
	Dev	3,540	3,696	3,564
	Test	2,519	2,983	2,722

with absolute bias below 0.25 includes 81.7% of the corpus for arousal, 87.4% of the corpus for valence, and 86.1% of the corpus for dominance.

After giving affective priming bias measures to each sentence in the MSP-Podcast corpus, we sorted the sentences and created four subsets of the corpus for each emotional attribute: *Neg*, *Pos*, *Neut1* and *Neut2*. The *Neg* subset contains the sentences with the bottom 20% of the affective priming bias. The *Pos* subset contains the top 20% of the affective priming bias. The *Neut1* subset has sentences included between the 40% quantile and 60% quantile. The *Neut2* subset has sentences with affective priming bias measures in the range -0.02 to 0.02. Table I shows the number of sentences included in each partition for each subset. Figure 7 shows the subsets in the distributions (*Neg* subset in blue, the *Pos* subset in orange, and the *Neut1* subset in black). We will use these subsets in the rest of the evaluation.

C. Inter-Evaluator Agreement

We measure the inter-evaluator agreements for the different subsets made from the bias calculations. The agreement between annotators plays an important role in both validating annotations and how well SER models perform. We use Krippendorff’s Alpha coefficient to measure the agreements. Table II shows that the *Neut1* and *Neut2* subsets have the highest agreements, while the *Neg* and *Pos* subsets have the lowest agreements. These results are expected since the subsets with near zero bias contain annotations that are more similar to the average of other annotators than the more biased subsets. The biased subsets have low agreements since the effects of priming are not uniform for all annotators. Even if many of the annotations were pushed toward the same extreme, the degree of the affective priming bias experienced by them may not be

TABLE II
KRIPPENDORFF'S ALPHA COEFFICIENT TO MEASURE INTER-EVALUATOR AGREEMENT FOR THE DIFFERENT SUBSETS.

Corpus Subset	**Emotional Attribute**		
	Arousal	***Valence***	***Dominance***
Full	0.376	0.335	0.327
Neg	0.279	0.287	0.193
Pos	0.268	0.186	0.179
Neut1	0.559	0.575	0.596
Neut2	0.595	0.559	0.612

TABLE III
AVERAGE TEST CCC RESULTS OVER 10 TESTING TRIALS USING THE MODEL TRAINED WITH THE FULL MSP-PODCAST TRAIN SET.

Sampled Test Set	**Emotional Attribute**		
	Arousal	***Valence***	***Dominance***
Full	0.650	**0.543**	0.531
Neg	**0.684**	0.531	**0.609**
Pos	0.650	0.523	0.546
Neut1	0.576	0.514	0.436
Neut2	0.555	0.530	0.436

consistent. Also, since both the bias measures and ratings are averaged, one or two very different annotations could push both the bias and agreement measures to an extreme.

V. SER MODELING EXPERIMENTS

The objective of this section is to analyze the role of the affective priming bias in SER experiments. The evaluation consists of training and testing the models with the subsets defined in Sections IV-B. We use the "wav2vec2-large-robust" architecture [19] as the core architecture of our SER model. This model showed the best recognition performance in the study of Wagner *et al.* [20] among the variants of the Wav2vec2.0 model [21]. 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, valence, and dominance. First, we import the pre-trained "wav2vec2-large-robust" model from the HuggingFace library [22]. Then, we fine-tune the transformer encoder with the downstream head using a specific subset of the corpus. We aggregate the outputs of the Wav2vec2.0 model 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 $p = 0.5$. We use the train set of the MSP-Podcast corpus to fine-tune the pre-trained SER model. 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 labels, mapping them into the range of 0 to 1. We use the Adam optimizer [23] with a learning rate of 0.0001. We use 32 utterances per mini-batch and update the model for 10 epochs.

The subsets for *Neg*, *Pos*, *Neut1* and *Neut2* shown in Table I have different emotional distributions. Those differences could affect the model results, resulting in prediction differences that are not dependent on the effect of affective priming but on the underlying emotional distributions of the subsets. To isolate the effect of the biases, we sample the different subsets to have the same emotional distributions across subsets. First, we fit a normal distribution to the full MSP-Podcast set for each emotional attribute. Then, we bin the data using the following edges $[1, 2, 2.5, 3, 3.25, 3.5, 3.75, 4, 4.25, 4.5, 4.75, 5, 5.5, 6, 7]$. This strategy creates non-uniform bins for sampling the sets, increasing the resolution for data around the central values of the attributes, where more sentences are located. Next, we calculate the area under the fitted Gaussian for each bin and the area of the Gaussian truncated at 1 and 7. We get a ratio for each bin by dividing the bin area and the area of the truncated Gaussian. Then, we bin the sentences of each subset partition (train, development, and test) according to the defined edges. We calculate the number of sentences needed for each bin by multiplying the bin ratio with the desired number of sentences in each partition. Then, we sample each bin without replacement. In few cases, we do not have enough sentences in each bin. For these cases, we include all the sentences in the bin and then sample them with replacement to fill the desired number of samples per bin. We choose to have 5,007 sentences in each train partition, 507 sentences in each development partition, and 1,006 sentences in each test partition. These numbers were chosen to have enough sentences for fine-tuning the models while also minimizing the number of bins with repeated sentences. When we compare the subsets with the full corpus in the experiments, we also sample the full corpus partitions to have the same number of sentences and the same distribution as the subset partitions.

First, we evaluate the models trained with the entire training corpus using the sampled subsets from the test set. We repeat the sampling of the subsets to balance the emotional distribution of the attributes 10 times creating different trials. Table III reports the average *concordance correlation coefficient* (CCC) across the 10 trials. The biased subsets have the best prediction results for arousal and dominance. For valence, the prediction results are not very different between the subsets. Sridhar and Busso [24] showed that SER models are more uncertain when predicting attribute values closer to neutral, especially for arousal and dominance. Although we have the same labeled emotion distribution for all our subsets, the way the labels relate to the speech features is different. We can think of the emotion distribution of the *Neg* subset as being shifted to the left from what it would be without the affective priming, and to the right for the *Pos* subset. Therefore, when the model predicts extreme values with more certainty, it is more likely to be closer to the labeled values of the biased subsets than the less-biased subsets. This is not seen in valence, since SER models do not show the same extreme uncertainty differences between extreme and neutral predictions for valence [24].

Second, we conduct regression experiments by fine-tuning the SER models with the sampled subsets from the train partition. This evaluation aims to explore further insights into how affective priming in emotional labels affects SER models.

These models adapted with the biased and unbiased subsets are evaluated with the sampled test partitions. We conducted this experiment with all the subsets eight times. For each trial, we used a different seed to sample the subsets and initialize the network. Table IV shows the average test CCC results over the eight trials. For arousal and dominance, the biased and full subsets perform better than the neutral subsets for both testing and training. For arousal, fine-tuning with the *Neg* subset results in the best CCC results for each test partition (second row). Evaluating the models on the *Neg* set also leads to the best performance (second column). This result is consistent with the previous experiment showing that arousal and dominance are better predicted using the biased subsets. Since arousal and dominance SER models are more certain of extreme predictions [24], the models built with the biased subsets can more confidently predict extreme values. Therefore, even if the biased models are not good at predicting the exact labels from the less-biased subsets, they could still consistently predict the high and low values as high and low. In contrast, the models built with less-biased subsets could become less certain of their predictions. These less-biased models might get closer to the unbiased labels, but they could be less consistent at the extreme values compared to the biased models. For valence, the neutral subsets perform the best. The best performance is achieved by either fine-tuning the model with *Neut2* (fifth row) or testing the model with *Neut2* (fifth column). Valence does not show a large uncertainty difference between extreme and neutral predictions [24], so the less-biased subsets perform better. This result is expected since the less-biased subsets both represent the average human perception better and have better inter-evaluator agreements.

TABLE IV
AVERAGE TEST CCC RESULTS OVER 8 TRAINING TRIALS, TRAINED AND TESTED WITH SAMPLED MSP-PODCAST PARTITIONS.

	Sampled	**Sampled Test Set**				
	Train Set	*Full*	*Neg*	*Pos*	*Neut1*	*Neut2*
Arousal	*Full*	0.574	0.590	0.573	0.505	0.500
	Neg	0.586	0.610	0.598	0.516	0.511
	Pos	0.544	0.561	0.544	0.485	0.477
	Neut1	0.448	0.465	0.445	0.410	0.403
	Neut2	0.472	0.482	0.469	0.433	0.422
	Sampled	**Sampled Test Set**				
	Train Set	*Full*	*Neg*	*Pos*	*Neut1*	*Neut2*
Valence	*Full*	0.173	0.156	0.147	0.191	0.204
	Neg	0.196	0.163	0.151	0.209	0.222
	Pos	0.169	0.157	0.139	0.182	0.201
	Neut1	0.190	0.168	0.154	0.224	0.233
	Neut2	0.223	0.202	0.181	0.255	0.254
	Sampled	**Sampled Test Set**				
	Train Set	*Full*	*Neg*	*Pos*	*Neut1*	*Neut2*
Dominance	*Full*	0.406	0.451	0.385	0.388	0.379
	Neg	0.423	0.466	0.407	0.378	0.374
	Pos	0.386	0.425	0.376	0.348	0.348
	Neut1	0.279	0.297	0.252	0.311	0.297
	Neut2	0.277	0.291	0.257	0.305	0.291

VI. DISCUSSION AND CONCLUSIONS

In this study, we analyzed the effect of affective priming on emotional annotations of speech. We observed affective priming in the typical emotional speech annotation process of evaluating sentences in random sequences. The effect of the priming was consistent with previous studies on valence priming [8], [13], [15], indicating that previous samples annotated with extreme values of valence push the annotations toward that extreme. Interestingly, we also observe the same effect for arousal and dominance. We also observe that the resulting biases from affective priming lead to lower inter-evaluator agreements. Further studies on SER modeling showed that the resulting biases lead to higher SER performances for arousal and dominance and lower performances for valence.

In general, biases are considered undesirable aspects of data. A way to mitigate the bias from affective priming could be to have multiple annotations per sentence. Burmania and Busso [25] argued that consensus labels do not radically change after adding five annotations per stimulus. Our study suggests that reducing affective priming is another valid reason to annotate each sample by multiple annotators. The good news is that most of the sentences in the corpus have a low affective priming bias when the data is annotated by at least five workers, as in the MSP-Podcast corpus. This analysis also supports two practices for perceptual evaluations that are relevant to data collection: (1) presenting the data in out-of-temporal order, and (2) randomizing the order of the sentences presented to multiple raters. These two approaches can avoid having the same affective priming bias across annotators. As argued in the paper, corpora that are biased to some emotions may be more sensitive to affective priming biases (e.g., databases containing relationship problems between disruptive family members, or databases containing colloquial conversations between friends). The results of this study highlight the importance of having emotionally balanced databases.

In general, we do not recommend removing data that is biased by affective priming. The modeling results show that these biases are not detrimental to all SER models. Sometimes, the bias can even help the SER performance. Furthermore, these biases represent natural emotional perceptions, since affective priming affects people in their day-to-day lives. Completely removing the effects of affective priming from affective computing could be detrimental to the applicability of emotional models. The results also support using an ordinal formulation for SER studies [26]–[31], relying on relative labels where the "anchors" are explicit.

For our future work, we want to more deeply explore the cause of the SER model results. We hypothesize that the higher uncertainty of more neutral predictions of arousal and dominance results in the higher performance results of the biased subsets. One way to validate our hypothesis is to explore the uncertainty of the models trained on the differently-biased subsets. We also want to explore different options for dealing with the biases from affective priming. For example, we can use biases to create models that can adapt to different contexts, as presented in the study of Chien and Lee [17]. If an agent is able to detect the type of affective priming a subject has been exposed to, it can switch to a similarly primed model to engage the subject in a more effective manner.

Ethical Impact Statement

Bias in machine learning models has been an often talked about topic when it comes to ethical problems in the field. In this paper, we study the role of a type of bias on emotional speech annotations and SER models. Although we focus on biases arising from affective priming, the methods we use in our paper can be used to isolate other types of biases. Isolating biases, such as cultural or gender biases, could be useful in dismantling inequalities in models. However, the isolation of these biases can also be used to create more unequal models. Furthermore, our methods can also be used to detect certain biases in data. Such detection could be used to discover characteristics of people that have not been disclosed to the researchers. In our study, the identity of the annotators is not available so this risk is limited.

Our results offer preliminary research on building SER models that can adapt to a person's emotional characteristics. In our conclusion, we mention that we can use our findings to build systems that can change their behavior to more closely resemble a specific human subject's behavior. Models that adapt to people's emotional states can also be used to manipulate them. Having agents that can make people feel connected or understood can lead to strong attachments to such agents. Although not necessarily negative, such attachments can be exploited by the researchers in control of the agents.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] J.-S. Lee, J. Choi, J. 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, September 2014.

[7] H. Bosker, "Putting Laurel and Yanny in context," *The Journal of the Acoustical Society of America*, vol. 144, no. 6, pp. EL503–EL508, December 2018.

[8] K. Klauer and J. Musch, "Affective priming: Findings and theories," in *The psychology of evaluation: Affective processes in cognition and emotion*, J. Musch and K. Klauer, Eds. Mahwah, NJ, USA: Lawrence Erlbaum Associates Publishers, November 2002, pp. 9–50.

[9] 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.

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

[11] 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.

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

[13] M. Lohse and M. Overgaard, "Emotional priming depends on the degree of conscious experience," *Neuropsychologia*, vol. 128, pp. 96–102, May 2019.

[14] R. Mueller, S. Utz, C.-C. Carbon, and T. Strobach, "Face adaptation and face priming as tools for getting insights into the quality of face space," *Frontiers in Psycholog*, vol. 11, pp. 1–17, February 2020.

[15] S. Murphy and R. Zajonc, "Affect, cognition, and awareness: affective priming with optimal and suboptimal stimulus exposures," *Journal of Personality and Social Psychology*, vol. 64, no. 5, pp. 723–739, May 1993.

[16] C. Nussbaum, C. von Eiff, V. Skuk, and S. Schweinberger, "Vocal emotion adaptation aftereffects within and across speaker genders: Roles of timbre and fundamental frequency," *Cognition*, vol. 2019, p. 104967, February 2022.

[17] W.-S. Chien and C. C. Lee, "Achieving fair speech emotion recognition via perceptual fairness," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes island, Greece, June 2023, pp. 1–5.

[18] 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.

[19] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *ArXiv e-prints (arXiv:2104.01027)*, pp. 1–9, April 2021.

[20] 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. Early Access, 2023.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] A. Burmania and C. Busso, "A stepwise analysis of aggregated crowdsourced labels describing multimodal emotional behaviors," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 152–157.

[26] 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.

[27] S. Parthasarathy, R. Lotfian, and C. Busso, "Ranking emotional attributes with deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 4995–4999.

[28] H. Martinez, G. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 314–326, July-September 2014.

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

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

[31] R. Lotfian and C. Busso, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 490–494.