DYNAMIC SPEECH EMOTION RECOGNITION USING A CONDITIONAL NEURAL PROCESS

Luz Martinez-Lucas and Carlos Busso

Multimodal Signal Processing (MSP) Lab, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
luz.martinez-lucas@utdallas.edu, busso@utdallas.edu

ABSTRACT

The problem of predicting emotional attributes from speech has often focused on predicting a single value from a sentence or short speaking turn. These methods often ignore that natural emotions are both dynamic and dependent on context. To model the dynamic nature of emotions, we can treat the prediction of emotion from speech as a time-series problem. We refer to the problem of predicting these emotional traces as dynamic speech emotion recognition. Previous studies in this area have used models that treat all emotional traces as coming from the same underlying distribution. Since emotions are dependent on contextual information, these methods might obscure the context of an emotional interaction. This paper uses a neural process model with a segment-level *speech emotion recognition* (SER) model for this problem. This type of model leverages information from the time-series and predictions from the SER model to learn a prior that defines a distribution over emotional traces. Our proposed model performs 21% better than a *bidirectional long short-term memory* (BiLSTM) baseline when predicting emotional traces for valence.

Index Terms— Speech Emotion Recognition, Dynamic Speech Emotion Recognition, Time-Continuous Emotional Traces.

1. INTRODUCTION

Studies on *speech emotion recognition* (SER) have often focused on recognizing an emotional category or predicting emotional attributes of a sentence or short speaking turn [1–4]. Although these models can give important information to improve *human-computer interaction* (HCI), they do not necessarily model the natural externalization of emotions during a conversation. In day-to-day life, emotional events happen within context-filled situations [5]. Furthermore, natural and nuanced emotions are dynamic throughout time [6]. This view of emotions promotes the formulation of SER problems as time-series problems, as opposed to treating each prediction or classification as an independent event. We coin the term *dynamic speech emotion recognition* (DSER) to refer to this SER setting.

Modeling emotional predictions from speech as a time-series problem has been a growing strategy [7, 8]. Much of the work on DSER has focused on single-distribution approaches, which treat each emotional time-series as coming from the same distribution as all other emotional traces. However, emotions occur in situations with various different contexts that can affect their trajectories. This study investigates a novel method that treats each emotional trace as coming from a distribution that is task-specific (e.g., an interview versus an argument). A step towards this strategy is using *neural process* (NP) models. In time-series problems, these models use observations from the series to characterize the desired function (i.e., a few values from the target emotional trace to guide the prediction models). These models use the observations to predict the entire emotional time-series, helping predict the remaining time-steps. The key step of the model is the prediction of these observations.

This paper proposes the use of a *conditional neural process* (CNP) model to the problem of DSER. We predict emotional traces of arousal (calm to active), valence (negative to positive), and dominance (weak to strong) from the conversational speech in the MSP-Conversation corpus [9]. We first explore different parameters of the CNP model in the ideal case of using ground-truth labels for our observations. We observe *concordance correlation coefficient* (CCC) values above 0.69 for all the emotional attributes. These high prediction scores are obtained with only five observations. Then, we implement a segment-level SER model to predict pseudo-labels for our observations. Our final proposed model has a 21% performance increase from a *bidirectional long short-term memory* (BiLSTM) baseline for valence predictions. This result is significant since detecting valence from speech is a difficult task [10, 11]. While the performances for arousal and dominance are similar to the baseline model, the high performance achieved when using better observations for these attributes (e.g., ground truth scores) indicates the importance of the SER predictions. This result suggests that improving the SER predictions for the key observations will lead to a significant increase in the performance of our proposed model.

2. RELATED WORK

Recurrent neural network (RNN) based models, such as *long short-term memory* (LSTM) [12], are a common method of predicting time-series signals [13]. These methods have been widely used in dynamic emotion recognition [7, 8] because emotions often depend on contextual information. However, the results of these models have been limited since model development for this problem is difficult. RNN based models are sequential in nature, and cannot use information from a full time-series to predict each time-step. Transformers are another sequence-to-sequence model that can be used to predict time-series, using positional encodings [14]. Transformers are able to use the full time-series when predicting a single time-step. However, they require many parameters to achieve this goal, so the contextual window is often limited to nearby frames. Transformers have been successfully used for segment-level SER [2, 15], but DSER uses a much longer stream of data than segment-level SER. Training a transformer with long time-series can be computationally expensive or intractable, depending on the resources. The *neural process* (NP) family of models has been recently applied to time-series problems [16]. Like transformers, *conditional neural process* (CNP) models can use information from the entire time-series to predict each time-step. Unlike transformers, a CNP model only needs a subset of the time-series to learn the rest of the time-steps.

NP models have been used in image generation [16], image classification [17, 18], and dynamic expression recognition [19]. As in

This work was supported by NSF under Grant CNS-2016719.

our work, Tellamekala *et al.* [20] apply a NP model to the task of DSER. In their work, they predict the pseudo-labels of the observations using an RNN model trained on the same data used for the NP model. This method gives the NP model a different setting during training and testing since the pseudo-labels are expected to be more accurate for the training samples than for the testing samples. In this paper, we apply a CNP model to the task of DSER and add positional encodings to introduce sequential information to the model. Furthermore, we use predicted pseudo-labels to condition the CNP model. These predictions come from a segment-level SER model trained on a separate dataset, ensuring similar settings during the training and testing of our DSER model.

3. METHODOLOGY

We propose a DSER model that takes feature vectors of dimension d_f at each time-step of a conversation, $\{x_i\}_{i=1}^{n}$ where $x_i \in \mathbb{R}^{d_f}$, and outputs a predicted emotional trace, $\{\hat{y}_i = f(x_i)\}_{i=1}^{n}$ where $\hat{y}_i \in \mathbb{R}$. In this section, we describe the two models we use to predict emotional traces from conversations. The first model (Sec. 3.1) is a CNP that predicts an emotional trace by conditioning on a sample of observed points on the trace. The second model (Sec. 3.2) is a segment-level SER model that predicts the pseudo-labels of the observed points for the first model. Figure 1 shows the final proposed DSER model that combines the segment-level and CNP models (Sec. 3.3).

3.1. Conditional Neural Process

The *Conditional neural process* (CNP) was introduced by Garnelo *et al.* [16]. It is a conditional stochastic process that conditions predictions on a set of observations or points of a function, $O = \{(x_i, \tilde{y}_i)\}_{i=1}^{n_o}$, where n_o is the number of observations. The observations are used to predict the distribution over the possible functions, $Q_\theta(f(x_i) \mid O, x_i)$. The CNP model has three steps: the first step embeds each observation using a neural network $h_\theta : \mathbb{R}^{d_f+1} \rightarrow \mathbb{R}^{d_h}$ (Eq. 1). The second step aggregates the embeddings into a single embedding using a commutative operation $\oplus$ (Eq. 2). In our paper, the commutative operation is the mean. In the third step, the model predicts two values for each point in the target set, $T = \{x_i\}_{i=1}^{n_o+n_t}$, using a neural network $g_\theta : \mathbb{R}^{d_f+d_h} \rightarrow \mathbb{R}^2$ (Eq. 3). The two values represent the mean and variance of a Gaussian distribution over the attribute value of the target point.

$$r_i = h_\theta(x_i, \tilde{y}_i) \quad \forall (x_i, y_i) \in O \tag{1}$$

$$r = r_0 \oplus r_1 \oplus \cdots \oplus r_{n_o-2} \oplus r_{n_o-1} \tag{2}$$

$$(\mu_i, \sigma_i^2) = g_\theta(x_i, r) \quad \forall x_i \in T \tag{3}$$

In this paper, the neural networks h_θ and g_θ are *multilayer perceptrons* (MLPs). We use the predicted mean, μ_i, as our prediction for the emotional attribute value at time-step i, $\hat{y}_i$. During training, we randomly select $n_t = 50$ time-steps of the conversation to use with the observations as the target set. While evaluating the model on the development and test sets, we use all the time-steps in the conversation as targets. A key step for the CNP model is the selection of the n_o observations. We choose the observations by splitting the conversation into n_o sections and randomly choosing a time-step in the first section. The rest of the observations are chosen by selecting the time-step $\frac{n}{n_o}$ time-steps away from the previous until we have n_o observations. We evaluate the approach under the ideal scenario where the observation labels are the attribute values from the ground-truth trace associated with each observation time-step (i.e., $\tilde{y}_i = y_i$; used in Secs. 5.1 and 5.2). Then, we evaluate the approach where the observation labels are the pseudo-labels predicted by the SER model ($\tilde{y}_i = \check{y}_i$; used in Sec. 5.3).

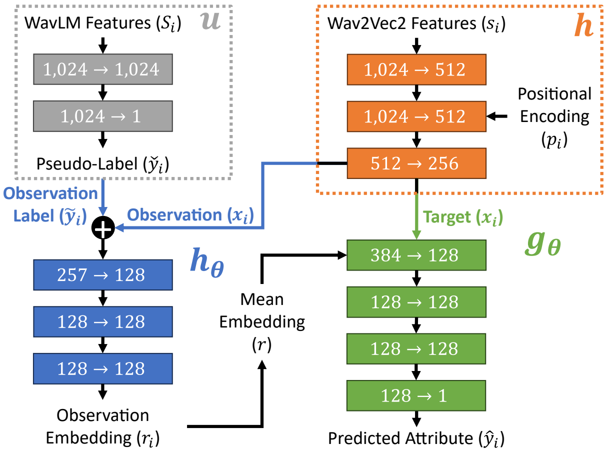


Fig. 1. Proposed DSER model. The segment-level SER model is shown in the top-left corner in grey. The CNP model feature encoder is shown in the top-right in orange. The CNP model is shown at the bottom in blue and green.

The CNP model does not encode any location information in its structure. To add this information, we use an MLP $h : \mathbb{R}^{d_s+d_p} \rightarrow \mathbb{R}^{d_f}$ to encode speech features $s_i \in \mathbb{R}^{d_s}$ and positional encodings $p_i \in \mathbb{R}^{d_p}$ into our time-step features x_i. We use the sine and cosine functions presented in [14] to calculate fixed positional encodings for each time-step. Figure 1 shows this model, highlighted in orange.

3.2. Segment-Level SER

The observations used in the CNP model initially include the ground-truth labels of time-steps ($\tilde{y}_i = y_i$). However, we expect this model to predict emotional traces without using any ground-truth labels. Therefore, we need a model that predicts the observation labels, creating pseudo-labels $\check{y}_i$. We achieve this goal with SER models that predict emotional attributes, which are used to create these pseudo-labels. We use speech features of a speech segment centered on the time-step (S_i) to predict the attribute value of the time-step. In this paper, we train the segment-level SER model first. Then, we predict the observation attribute values. Finally, the predictions are used as the observation labels to train the CNP model. The segment-level SER model we use is an MLP $u : \mathbb{R}^{d_s} \rightarrow \mathbb{R}$.

3.3. Proposed Model for Dynamic Speech Emotion Recognition

The CNP model consists of three networks shown in Figure 1. The first network is h, highlighted in orange in Figure 1. This function takes the feature vectors (we use a Wav2Vec2.0 feature extractor; explained in Sec. 4.2) and the positional encodings [14] at each time-step of a conversation part and outputs the encoded features at each time-step. The second network is h_θ, highlighted in blue in Figure 1. This function takes the encoded features belonging to the observation time-steps and concatenates them with their corresponding ground-truth labels (used in Secs. 5.1 and 5.2) or pseudo-labels from the SER model (used in Sec. 5.3). Then, the network passes the concatenated observations through three linear layers to get the embedding for each observation. The final network is g_θ (Eq. 3), highlighted in green in Figure 1. This function takes the encoded features from the first network belonging to the target time-steps and concatenates each with the mean of the observation embeddings. Then, each concatenated set is passed through four linear layers to predict the attribute value for the time-step (arousal, valence, or dominance).

4. EXPERIMENTAL SETTINGS

4.1. Emotional Databases

This study relies on two emotional datasets to train and evaluate the models. We use the MSP-Conversation corpus to train the DSER model to predict emotional traces. We use the MSP-Podcast corpus to train a segment-level SER model to predict pseudo-labels for the input observations for the DSER model.

The MSP-Conversation corpus [9] is a dataset of audio-only conversations between multiple speakers split into 3-7 minute parts. The conversations are sourced from podcasts used for the MSP-Podcast corpus [21]. Each conversation part is annotated by at least six annotators, creating emotional traces for the attributes of arousal (calm to active), valence (negative to positive), and dominance (weak to strong) [22, 23]. The annotation process involves raters using a joystick to record their instantaneous emotional perception as they listen to a conversation part. We use the proposed partitions from version 1.1 of the corpus, which contains 695 conversation parts (59 hrs 33 min), where 429 conversation parts are in the train set, 104 in the development set, and 162 in the test set.

The MSP-Podcast corpus [21] is a dataset of speech sentences obtained from publicly available audio sources. The database relies on the retrieval-based approach proposed by Mariooryad *et al.* [24] to identify emotional data to be annotated. Each speaking turn is between 3 and 11 seconds long and is annotated at the sentence level with values for arousal, valence, and dominance. Annotators rate each sentence using a 7-point Likert scale for each attribute. They also provided primary and secondary emotional categories. In this paper, we only use attribute-based annotations. We use version 1.11 of the corpus, which contains 151,654 sentences (237 hrs 56 min).

4.2. Feature Extraction

We use two types of speech features for our model. The speech features used to train the CNP model are extracted using the "wav2vec2-large-robust" architecture [25]. We use the pre-trained "wav2vec2-large-robust" model from the HuggingFace library [26] fine-tuned with Version 1.10 of the MSP-Podcast corpus (104,267 sentences) [21] as done in Martinez-Lucas *et al.* [27]. We obtain the outputs of the Wav2Vec2.0 model for each conversation part in the MSP-Conversation corpus. The wav2vec2 vectors are then split into 1-second chunks using a stride of 800 milliseconds and are averaged for each chunk. The emotional traces from the MSP-Conversation corpus are similarly split into 1-second chunks. We use 1-second chunks as time-steps for these emotional traces since they a short enough to capture the dynamic nature of the traces while minimizing noise from the annotations [28]. We obtain a wav2vec2 feature vector of dimension 1,024 for each 1-second time-step.

The speech features used to train the segment-level SER model are extracted using the "WavLM Large" architecture [29]. We use the pre-trained "wavlm-large" model from the HuggingFace library [26] and obtain the outputs for each MSP-Podcast sentence (Version 1.11). The WavLM vectors in the sentences from the MSP-Podcast corpus are averaged for each sentence. We also obtain outputs from the WavLM model for each conversation in the MSP-Conversation corpus. The WavLM vectors are split into 5-second segments centered around each 1-second time-step and averaged over the segments. We obtain a WavLM feature vector of dimension 1,024 for each MSP-Podcast sentence and MSP-Conversation segment.

Table 1. Test results of the CNP model using ground-truth labels for five observations.

Attribute	CCC ↑	ρ ↑	MSE ↓
Arousal	0.741	0.749	78.4
Valence	0.787	0.787	152
Dominance	0.693	0.695	73.8

4.3. Training Details

For the conditional neural process, we use a *rectified linear unit* (ReLU) activation and dropout with a rate of 0.5 between each linear layer. We use the ADAM optimizer with a learning rate of 0.0001, and the loss function $\mathcal{L}(x, \hat{y}) = 1 - CCC(x, \hat{y})$, to maximize the *concordance correlation coefficient* (CCC) [30]. We train the model for 25 epochs with a batch size of 33 conversation parts on the MSP-Conversation training set. We select the model with the best CCC on the development set and report the CCC, *Pearson correlation coefficient* (ρ), and *mean squared error* (MSE) on the test set.

For the segment-level SER model, we consider two fully connected layers shown in Figure 1, highlighted in gray. It takes the sentence or segment WavLM features and outputs the prediction for a given emotional attribute (arousal, valence, dominance), which is used as the pseudo-label. We use dropout with a rate of 0.5 before both layers, and ReLU activation and layer normalization before the second layer. We use the same training optimizer and loss function as the CNP model. Since the MSP-Podcast corpus has speakers and sentences that overlap with the MSP-Conversation corpus, we do not use any sentences with speakers that overlap with the MSP-Conversation corpus. After removing those sentences, we obtain 58,965 sentences in the train set and 12,600 in the development set. We train the model for 15 epochs with a batch size of 32 sentences on the training set and choose the model with the best development CCC. The MSP-Podcast labels are scaled from 1 to 7 to -100 to 100 to match the range of the MSP-Conversation labels. As we have shown in our previous work [31], the labels of the MSP-Podcast and MSP-Conversation corpora are not interchangeable. Therefore, we also fine-tune the segment-level SER model with a subset of the MSP-Conversation corpus. This SER subset has 38 conversation parts in the train set and 11 conversation parts in the development set. We then train the CNP model with the CNP subset with 391 conversation parts in the train set and 93 conversation parts in the development set. The SER and CNP subsets of the MSP-Conversation are speaker-independent.

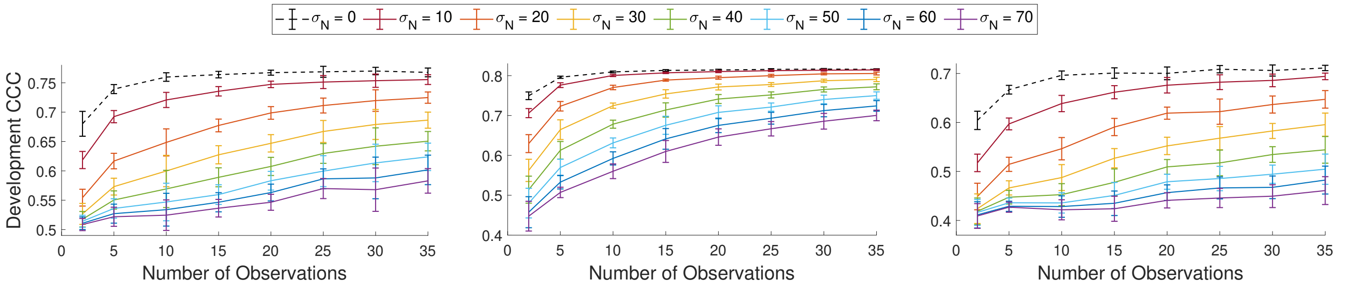
5. RESULTS

5.1. CNP Models with Accurate Observations

First, we are interested in evaluating the performance of the system in the case of accurate observations. Therefore, we obtain the observation labels from the ground truth traces. Table 1 shows the test results for the CNP model using five observations. We also show an example of a predicted trace in Figure 3. With accurate observations, the model is able to predict the emotional traces with high accuracy, reaching CCC values above 0.69 for all the emotional attributes. These results are achieved with only 5 observations.

5.2. Analyzing Number and Precision of Observation Labels

Ultimately, we want to use predicted pseudo-labels for the observations instead of ground-truth labels. We expect these pseudo-labels to be noisy but still be somewhat close to the ground-truth labels. To explore how robust the CNP model is to noise in the observations, we add Gaussian noise to the ground-truth observation labels. We use



(a) Arousal

(b) Valence

(c) Dominance

Fig. 2. Average development CCC for a CNP model over 10 trials. We use a varying number of observations and a varying amount of label noise for the ground-truth observation labels. The error bars represent the STD over the 10 trails.

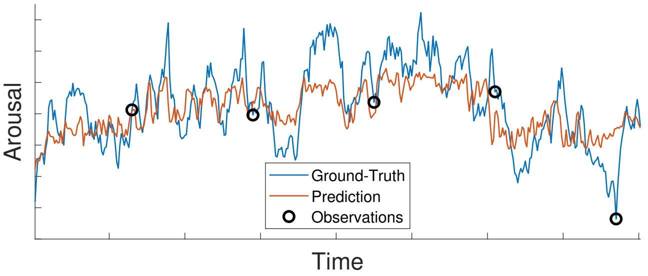


Fig. 3. Example of a prediction made by the CNP model. The blue line shows the arousal trace for one conversation part. The orange line shows the CNP predictions using accurate observations. The observations are highlighted in circles.

zero mean Gaussian noise with varying *standard deviation* (STD) (σ_N). Figure 2 shows the average development CCC over ten trials for CNP models trained with a varying number of observations (n_o) with labels contaminated with Gaussian noise. Figure 2 shows that more observations lead to better performance. However, for arousal and dominance, the trend breaks at a certain value for n_o when the level of noise in the labels is high. Overall, noisy observation labels affect arousal and dominance in similar ways. As the labels for the observations become noisy, the performance of the system is reduced. For valence, the noisy labels also reduce performance. However, unlike arousal and dominance, raising the number of observations mitigates the effect of noisy labels for the observations up to some extent. The CNP model is able to better compensate for the noisy labels when predicting valence with more observations.

5.3. CNP Models with Predicted Labels for the Observations

In this section, we look at the performance of the CNP model when using pseudo-labels for the observations. We train the CNP model with the attribute predictions from the fine-tuned segment-level SER model as pseudo-labels (SER+CNP). As a comparison, we also show the results when we obtain the label of the observation from the actual labels, denoting this case as CNP. As a baseline, we train a 2-layer BiLSTM network using the same training parameters as the CNP model. We used z-normalization on the labels for the baseline during training, which led to a more competitive model. Table 2 shows the results. All the models are trained using the CNP subset of the MSP-Conversation and tested with the full test set. For the CNP and SER+CNP models, we choose the number of observations according to the best development CCC of the SER+CNP model.

These results show that predicting the observation labels instead of using their ground truth labels has an effect on the performance of the CNP model, which is expected given the results in Figure 2 which show lower performance as we increase the noise in the labels. However, the proposed model still achieves better performance for valence and dominance. Furthermore, unlike the results in Section 5.2, we see that a higher number of observations does not necessarily lead to better performance. For the full SER+CNP model, the use of more observations leads to a tradeoff between more information to summarize the emotional trace and more noisy information to propagate errors from the SER model.

Table 2. Test results of the baseline and proposed models trained on the CNP subset of the MSP-Conversation corpus. The labels of the observations for the SER+CNP model are predicted by the segment-level SER.

Attribute	Model	# Obs.	CCC ↑	ρ ↑	MSE ↓
Arousal	BiLSTM		0.594	0.602	112
	CNP	20	0.766	0.767	74.3
	SER+CNP	20	0.560	0.574	118
Valence	BiLSTM		0.390	0.397	498
	CNP	35	0.802	0.803	149
	SER+CNP	35	0.474	0.478	441
Dominance	BiLSTM		0.435	0.440	124
	CNP	30	0.801	0.802	150
	SER+CNP	30	0.445	0.455	113

6. CONCLUSIONS

In this paper, we proposed a DSER model based on a CNP method. This model is able to predict each time-step in an emotional trace using an embedding learned from a set of time-steps (i.e., a few observations of the emotional trace). We show results where the embedding is learned using either ground-truth labels or pseudo-labels automatically predicted by an SER model. The results of the models using ground truth labels for the observations are very high, suggesting an undeniable potential for this framework. While the predictions drop when using the predicted pseudo-levels for the observations, the approach is still able to achieve high performance compared to a competitive baseline, especially for valence.

We can view the CNP results using ground-truth labels as the limit in performance for the CNP model. Therefore, increasing the performance of the SER model used to predict the pseudo-labels is a good next step to improve the model. We can also intentionally choose observations whose pseudo-labels are predicted with higher confidence, expecting that they will be closer to the ground-truth labels. We can also focus on raising the performance limit set by the CNP. For example, this paper uses MLPs for all the CNP networks. Therefore, we expect better performance if we use more sophisticated models to build these functions by including attention mechanisms or RNN layers.

7. REFERENCES

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

[2] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B.W. 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. 10745–10759, September 2023.

[3] M. Abdelwahab and C. Busso, "Active learning for speech emotion recognition using deep neural network," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2019)*, Cambridge, UK, 2019, pp. 441–447.

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

[5] W.-C. Lin and C. Busso, "Sequential modeling by leveraging non-uniform distribution of speech emotion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1087–1099, February 2023.

[6] R. Cowie et al., "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, January 2001.

[7] F. Ringeval others, "AVEC 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *International on Audio/Visual Emotion Challenge and Workshop*, Nice, France, Oct. 2019, pp. 3–12.

[8] A. Mallol-Ragolta, N. Cummins, and B.W. Schuller, "An investigation of cross-cultural semi-supervised learning for continuous affect recognition," in *Interspeech 2020*, Shanghai, China, October 2020, pp. 511–515.

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

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

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

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[13] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Interspeech 2014*, Singapore, September 2014, pp. 338–342.

[14] A. Vaswani et al., "Attention is all you need," in *In Advances in Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, December 2017, pp. 5998–6008.

[15] X. Wang, M. Wang, W. Qi, W. Su, X. Wang, and H. Zhou, "A novel end-to-end speech emotion recognition network with stacked transformer layers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)*, Toronto, ON, Canada, June 2021, pp. 6289–6293.

[16] M. Garnelo et al., "Conditional neural processes," in *Proceedings of Machine Learning Research (PMLR 2018)*, J. Dy and A. Krause, Eds., vol. 80, pp. 1704–1713. PMLR, Stockholm, Sweden, July 2018.

[17] J. Wang, T. Lukasiewicz, D. Massiceti, X. Hu, V. Pavlovic, and A. Neophytou, "NP-match: When neural processes meet semi-supervised learning," in *Proceedings of Machine Learning Research (PMLR 2022)*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 97, pp. 22919–22934. PMLR, Baltimore, Maryland, USA, July 2022.

[18] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y.W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *Proceedings of Machine Learning Research (PMLR 2019)*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97, pp. 3744–3753. PMLR, Long Beach, CA, USA, June 2019.

[19] E. Sanchez, M. K. Tellamekala, M. Valstar, and G. Tzimiropoulos, "Affective processes: stochastic modelling of temporal context for emotion and facial expression recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, Nashville, TN, USA, June 2021, pp. 9070–9080.

[20] M.K. Tellamekala et al., "Stochastic process regression for cross-cultural speech emotion recognition," in *Interspeech 2021*, Brno, Czech Republic, Aug.-Sep. 2021, pp. 3390–3394.

[21] 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.

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

[23] J.A. Russell and L.F. 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.

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

[25] 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.

[26] 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.

[27] L. Martinez-Lucas, A. Salman, S.-G. Leem, S.G. 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.

[28] R. Cowie and G. McKeown, "Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme," September 2010, SEMAINE Report D6b.

[29] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, October 2022.

[30] L.I.K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, March 1989.

[31] 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. Under submission, 2024.