Sequential Modeling by Leveraging Non-Uniform Distribution of Speech Emotion

Wei-Cheng Lin, *Student Member, IEEE,* Carlos Busso, *Fellow, IEEE,*

***Abstract*—The expression and perception of human emotions are not uniformly distributed over time. Therefore, tracking local changes of emotion within a segment can lead to better models for *speech emotion recognition* (SER), even when the task is to provide a sentence-level prediction of the emotional content. A challenge to exploring local emotional changes within a sentence is that most existing emotional corpora only provide sentence-level annotations (i.e., one label per sentence). This labeling approach is not appropriate for leveraging the dynamic emotional trends within a sentence. We propose a framework that splits a sentence into a fixed number of chunks, generating chunk-level emotional patterns. The approach relies on emotion rankers to unveil the emotional pattern within a sentence, creating continuous emotional curves. Our approach trains the sentence-level SER model with a *sequence-to-sequence* formulation by leveraging the retrieved emotional curves. The proposed method achieves the best *concordance correlation coefficient* (CCC) prediction performance for arousal (0.7120), valence (0.3125), and dominance (0.6324) on the MSP-Podcast corpus. In addition, we validate the approach with experiments on the IEMOCAP and MSP-IMPROV databases. We further compare the retrieved curves with time-continuous emotional traces. The evaluation demonstrates that these retrieved chunk-label curves can effectively capture emotional trends within a sentence, displaying a time-consistency property that is similar to time-continuous traces annotated by human listeners. The proposed SER model learns meaningful, complementary, local information that contributes to the improvement of sentence-level predictions of emotional attributes.**

***Index Terms*—Emotion rankers, speech emotion recognition, chunk-level segmentation, sequence-to-sequence modeling.**

I. Introduction

ADVANCES in *human-computer interaction* (HCI) have improved the way we communicate with automatic systems, making the interaction more natural. In addition to recognizing the explicit information provided by users, these systems have attempted to infer the underlying emotional content, aiming to better understand the message beyond the verbal content. A critical remaining challenge is to build a reliable *speech emotion recognition* (SER) system [1]. A key barrier to building such a system is access to databases with informative labels. The most common approach for labeling emotional databases is to assign a single label per sentence, regardless of its duration [2]–[4]. Therefore, conventional SER formulations treat this problem as a *sequence-to-one* learning task, where features extracted from a sequence of frames are used to infer a single label for that sentence. Most models aim to extract good feature representations, given the sentence-level emotion label. However, it is difficult to leverage or uncover local emotional changes within a sentence with a single sentence-level label. Emotion is not uniformly distributed across the entire sentence. Various studies have shown that emotions are neither perceived nor expressed uniformly across a sentence [5]–[8]. For example, a sentence may be labeled with a given emotion, even though the emotional content in most frames is neutral. These studies suggest that we are missing important information by training an SER system with a sequence-to-one formulation that assumes that the global label is the same for all the frames in the sentence.

Recent studies have dealt with emotional variations within a sentence by formulating an SER task as a *sequence-to-sequence* problem. The most straightforward approach is to rely on time-continuous annotations of emotional attributes. These emotional traces are collected by asking annotators to continuously judge the emotional content by moving a cursor in a *graphical user interface* (GUI) while they listen or watch a stimulus. However, there are few emotional databases that have been annotated with time-continuous traces [9]–[11], so identifying approaches for leveraging local emotional variations with recordings annotated with sentence-level labels is still a challenge. The key assumption in these approaches is that the emotional content is not the same as the sentence-level label for all the segments within a sentence, exploring local emotional variations in the formulation. Some of these approaches have used models such as Markov chain [12] or *connectionist temporal classification* (CTC) [13]–[15] to construct an emotional sequence that contains different hidden states reflecting the local emotional variations. These hidden states are dynamically transited based on the input acoustic features of a sentence. The key idea is to learn a latent random variable that describes the emotion within a sentence. For example, this latent variable can have two discrete states to indicate that a given segment within a sentence is either emotional (*Emo*) or neutral (*Null*) [13], [14]. These approaches can effectively get rid of non-emotional frames and further improve SER accuracy. Nevertheless, the assumption of binary emotion distribution in the CTC framework is not able to reflect the dynamic variations of emotional intensity over time. Moreover, this formulation is not easy to implement for attribute-based emotional regression tasks such as arousal and valence [16], since discrete hidden states are not suitable for continuous emotional scores.

We propose a framework that can retrieve chunk-level

W.C. Lin and C. Busso are with the Erik Jonsson School of Engineering & Computer Science, The University of Texas at Dallas, Richardson, TX 75080 USA (email: wei-cheng.lin@utdallas.edu, busso@utdallas.edu).

The code is available at GitHub: https://github.com/winston-lin-wei-cheng/Chunk-Level-Emotion-Retrieval

Manuscript received March XX, 2022; revised XXXX XX, XXXX.

emotional curves from sentence-level labels (i.e., arousal, dominance, and valence) using preference learning, creating an intermediate label representation. Our study is motivated by the expectation that leveraging local emotional variations can be useful for improving the performance of an SER system trained with sentence-level labels. By detecting emotional salient regions in a sentence [17], we can effectively build an SER system that is trained with this emotional-relevant data [18], [19]. Our proposed formulation starts by splitting a speaking turn into chunks. The chunks are created with the dynamic segmentation process presented by Lin and Busso [20], [21]. This process splits a sentence with an arbitrary duration into a fixed number of chunks with a fixed duration (e.g., one second) by dynamically changing the overlap between chunks. Then, we rely on a preference learning formulation to train rank-based classifiers (i.e., emotion rankers) with these data chunks using the methods presented by Parthasarathy and Busso [22]. This step aims to rank the emotional content conveyed within a sentence, imposing constraints so that the global emotion for the entire sentence matches the consensus sentence-level emotional label. We utilize the results of the trained rankers as an intermediate label representation to generate the continuous chunk-level emotional curves. We train our model with these chunk-level labels using a *sequence-to-sequence* SER formulation that directly considers the chunk-level local emotion sequence to recognize the sentence-level emotion. The chunk-level labels explicitly inform our model that not all the chunks within the sentence convey the same emotion, modeling the non-uniform externalization of emotions.

The main evaluation of the proposed approach is conducted on the MSP-Podcast corpus [4]. The results demonstrate that our approach significantly outperforms other baseline models, achieving the best *concordance correlation coefficient* (CCC) predictions for arousal (0.7120), dominance (0.6324), and valence (0.3125). The results are compared with the performance achieved by other existing *state-of-the-art* (SOTA) SER results. A key feature of the MSP-Podcast corpus is the complementary information provided by the MSP-Conversation corpus [10] for a subset of the recordings. The MSP-Conversation corpus relies on time-continuous annotations for a subset of the recordings included in the MSP-Podcast corpus. As a result, we have time-continuos evaluations (MSP-Conversation corpus) and sentence-level annotations (MSP-Podcast corpus) for some speaking turns. For these speaking turns, we can directly compare our chunk-level traces with actual emotional traces collected by annotators. We compute the correlation between the retrieved emotional curves and the time-continuous emotional traces (ground truth). We observe that the retrieved chunk-level curves are surprisingly correlated with the ground-truth emotional traces if the reaction lag in the annotations is considered [23], [24]. These results show that our model learns meaningful complementary information obtained from the retrieved chunk-level curves (i.e., non-uniform distribution of emotional content within a sentence), which leads to improvements in sentence-level recognition performance. Finally, we verify the benefits of the proposed approach with evaluations on two additional corpora: the IEMOCAP and MSP-IMPROV corpora. The main contributions of this study are:

- We propose a novel ranker-based framework to retrieve chunk-level continuous emotional curves solely relying on sentence-level labels. These retrieved curves have high similarity with human ground-truth annotations.
- We utilize the retrieved sequences to explicitly inform the model of local emotional variations, formulating the SER task as a sequence-to-sequence task. This approach achieves the best sentence-level recognition performance over all baselines for all emotion attributes.

The paper is organized as follows. Section II discusses related studies to our work. Section III presents our proposed framework. Section IV describes the experimental resources, acoustic features, baselines, and implementation settings. Section V presents our main experimental results including generalization test, SOTA comparisons, model computational analysis, and the evaluation of the retrieved emotion curves. Lastly, Section VI concludes the paper, discussing future research directions.

II. Related Work

A. Nonuniform Externalization of Emotion

Studies have observed the nonuniform expression and perception of emotions. Lee et al. [25] explored phoneme-level patterns in the expression of emotion. Their analysis demonstrated that vowels such as /aa/ have larger variability across emotions than vowels such as /iy/, which are more constrained by the articulatory movements (i.e., placement of the tongue in the palette). Similarly, Busso et al. [7] demonstrated that nasal phonemes have more restricted emotional modulation than other phonemes given the rigid configuration of the nasal cavity. The modulation of emotions is also dictated by the strong relationship between facial expressions and acoustic features [26]. Busso and Narayanan [6] found that the externalization of emotions in prosody and facial expression is emphasized when the vocal tract is physically constrained by speech articulation, showing a multimodal compensatory effect. These differences in the externalization of emotion also happen in suprasegmental acoustic features. Wang et al. [8] discovered a common pattern of raising the fundamental frequency at the end of a sentence for happy sentences. Cowie et al. [5] stated that emotions can change either gradually or sharply over time. They designed the FEELTRACE toolkit to continuously annotate the localized changes of emotions. This toolkit generates traces reflecting the instantaneous dynamic perception of an annotator while watching or listening to a stimulus. Collectively, these studies indicate that emotions are not uniformly externalized and that emotional modulation depends on the actual emotion and modality.

Based on these findings, studies have proposed different frameworks to leverage the nonuniform externalization of emotions to improve SER performance. Lee et al. [25] used phoneme-class dependent HMM classifiers trained with *Mel-frequency cepstral coefficients* (MFCCs) features. Their results showed that the phoneme-class dependent model obtained better discriminability compared to the generic utterance-level

TABLE I
SUMMARY OF THE MAIN DIFFERENCES BETWEEN OUR APPROACH AND OTHER EXISTING SER MODELING FRAMEWORKS DISCUSSED IN SECTION II-B. THE SYMBOLS ✓AND ✗INDICATE WHETHER A MODELING ASPECT IS APPLICABLE FOR A GIVEN STUDY.

SER Modeling Scheme	Sentence-Level Labels	Time-Continuous Labels	Considers Local Emotion Variants	Classification Task	Regression Task
Seq-to-One: [30]–[32]	✓	✗	✗	✓	✓
Seq-to-Seq: [33], [34]	✗	✓	✓	✓	✓
Seq-to-Seq (CTC): [14], [15]	✓	✗	✓	✓	✗
our approach	✓	✗	✓	✗	✓

HMM classifier. Lotfian and Busso [27] used synthetic speech as a reference for emotionally neutral speech to contrast localized emotional differences in the sentences. Another direction is to directly identify emotionally salient regions (i.e., hotspots) within a sentence [18]. Parthasarathy and Busso [17], [28] proposed a deep learning framework to recognize these emotional *hotspots* in speech. They first defined hotspot regions using the concept of *qualitative agreement* (QA) [29], searching for emotional trends. Then, an ensemble of *bidirectional long short-term memory* (BLSTM) regressors was trained and combined to detect the emotional hotspots. The high recognition performance (F1-score=60.9%) demonstrated a promising application to continuously track emotions during human interaction.

B. Sequence-to-One and Sequence-to-Sequence SER Tasks

This section focuses on the discussion of the main differences between our approach with other existing SER modeling frameworks. Table I summarizes the key distinctions between the formulations of previous studies and our approach

The most common formulation for SER is a *sequence-to-one* formulation, where a sequence of frames is mapped into a label representing the emotion for the entire sentence. A popular approach consists of extracting *low-level descriptors* (LLDs). LLDs are frame-level acoustic features such as the fundamental frequency, energy, and *Mel-frequency cepstral coefficients* (MFCCs). Then, a high dimensional sentence-level feature representation is created by extracting *high-level descriptors* (HLDs) or functionals (e.g., mean or variance of the fundamental frequency) [35]–[37]. A recent trend in SER studies is to directly extract feature representations from either raw features (i.e., raw waveform or spectrogram) or LLDs by using deep learning techniques. A key advantage of end-to-end training in deep learning models is the ability to jointly optimize the feature extraction and target discriminative task. This approach generally results in a powerful feature representation space and provides state-of-the-art performance. Various deep learning architectures have been applied for SER that have their own benefits: *convolutional neural networks* (CNNs) can capture local variations between acoustic patterns [31], and *recurrent neural network* (RNN), often implemented with *long short-term memory* (LSTM), are effective for dealing with temporal dynamic information [32]. Some hybrid architectures such as CNN-RNN can handle both spatial and temporal information in the features [30], [38]. However, these methods assume that the emotional content within a sentence can be represented with a single descriptor, which consistently reflects the emotion across the sentence.

An alternative formulation consists of modeling SER as a *sequence-to-sequence* task. Most of these studies use time-continuous annotations of emotional attributes [39]. Le et al. [33] proposed a multi-task BLSTM model to recognize time-continuous sequences, which were discretized with the k-mean algorithm at different resolutions. They incorporated an emotional *n-gram language model* for decoding the final predicted sequence. The sequential decoding procedure smooths the sequence, leading to further improvement in *concordance correlation coefficient* (CCC). Huang et al. [34] utilized an end-to-end *convolutional LSTM* (ConvLSTM) network to handle spatiotemporal information for continuous emotion recognition. The proposed ConvLSTM network outperformed a 3D convolutional model, showing its effectiveness in modeling temporal information. However, these studies require time-continuous emotional labels to train the model, which are only available in a few emotional databases. In contrast, our work does not rely on emotional traces. Instead, it predicts them with chunk-level continuous emotional curves derived from sentence-level labels.

Some studies have aimed to learn a hidden sequence of emotions that aligns with the input acoustic frames using only sentence-level labels. The CTC model [14], [15] is the most conventionally used method to achieve this goal. It requires the neural network architecture using recurrent units (e.g., LSTM) to produce sequential predictions for the desired categorical labels, which correspond to a sequence of probability distributions over time (i.e., softmax outputs). Then, the model optimizes the CTC loss [40] to choose the most probable label sequence over all possible paths based on the probability distributions (i.e., maximizing the posterior probability). These approaches assume that the CTC sequence consists of a series of transitional binary states between the sentence-level emotion state (*Emo*) and a non-emotional state (*Null*). This approach accounts for the fact that not every segment in an emotional sentence is emotional. The recognition performance of CTC models generally outperforms sequence-to-one modeling approaches [13], [14]. While this binary dichotomization and discrete state assumption might be suitable for categorical descriptors, these assumptions may not be the most appropriate formulation for modeling continuous emotion attributes such as valence, arousal, and dominance. Han et al. [41] proposed a framework built on a similar concept as our work. It leverages segment-level information to construct a sentence-level emotion recognizer. The approach first extracts the segment-level features by stacking neighbor frames together to train a DNN model. Then, they used the trained DNN model to compute the emotional state probability

distribution for every small segment in each sentence, forming a continuous segment-level probability curve for each emotion category. Finally, the sentence-level classification model was built with statistic descriptions of these probability curves. Unlike this method, our approach is formulated for regression tasks, exploiting rankers to directly retrieve continuous emotional curves as the training target sequences.

C. Preference Learning in SER tasks

Our formulation relies on preference learning, which exploits the ordinal nature of emotions [42]. Conventional SER systems predict absolute emotional labels, which are commonly collected with perceptual evaluations from multiple annotators. However, the differences in emotional perception across annotators [43] and the ambiguity in emotions [44] lead to poor agreement in the consensus labels, which affect SER systems. Studies have shown that more reliable systems can be obtained by learning trends in the labels [45]–[48]. Preference learning solutions have been presented for emotional categories [48], [49] (e.g., is one sentence happier than another?) and emotional attributes [22], [46], [50], [51] (e.g., is one sentence more active than another?). Since relative labels relying on trends can be more robust than absolute labels, rank-based SER systems are often more effective. We demonstrate in this paper that this formulation is ideal for retrieving relative emotional differences within a sentence. This study adopts the preference learning method proposed by Parthasarathy and Busso [22], which is based on the QA method [29], to train rankers to build our chunk-level continuous emotion sequences.

III. Proposed Methodology

The key contribution of this study is the ranker-based method for retrieving chunk-level continuous emotional curves. Our approach purely relies on the sentence-level annotations without requiring time-continuous annotations. The retrieved curves can effectively reflect local dynamic emotional changes in sentences, and, thus, improve recognition accuracy of a sentence-level SER model. The framework consists of four steps: (1) segmenting the data into chunks, (2) building the chunk-level emotional rankers, (3) retrieving the chunk-level continuous curves, and (4) constructing the sequence-to-sequence SER model with these curves. This section presents the detailed descriptions of each step.

A. Data Chunks Segmentation

We first apply the segmentation method proposed by Lin and Busso [21] to split the feature map into small data chunks for every sentence by formulating the sentence duration as a function of the data chunk segmentation process. The approach splits a sentence into a fixed number of chunks having the same length w_c, regardless of the duration of the sentence. The key to achieving this objective is the varied step size of the chunks Δc_i, which depends on the duration of the sentence i (i.e., T_i). The approach has two parameters. The first parameter is w_c, which is the desired duration of the chunks. The second parameter is T_{max}, which is the maximum duration of sentences in the corpus. We can obtain the minimum number of chunks, C, using Equation 1 with these two parameters.

$$C = \left\lceil \frac{T_{\max}}{w_c} \right\rceil \tag{1}$$

The variable Δc_i determines the overlap between chunks and it is estimated with Equation 2. The overlap between chunks is larger for shorter sentences. Equation 2 includes an optional scaling factor n ($1 \leq n$, $n \in \mathbb{N}$) to increase the number of chunks (i.e., nC). Equation 2 shows that as we increase n, Δc_i decreases, resulting in more overlap between chunks. This case is particularly useful for longer sentences. This scaling factor is relevant to our study, since adding more chunks will make our chunk-level curves denser.

$$\Delta c_i = \frac{T_i - w_c}{nC - 1} \tag{2}$$

B. Chunk-level Emotional Rankers

Our novel formulation to create chunk-based curves requires to identify relative differences in emotional content between chunks. We implement this approach with emotional rankers to establish preferences between chunks. Parthasarathy and Busso [22] proposed a method that defines ordinal labels based on sentence-level attribute-based annotations by leveraging the concept of *qualitative agreement* (QA) [29]. We follow this approach to build the chunk-level rankers for learning emotional trends. The key concept of this approach is to compare the trends across annotators, rather than estimating absolute scores. Figure 1 shows an example of how this approach defines the preference label. The example assumes that evaluators used a Likert scale to score each sentence. As an example, consider the emotional attribute valence with a scale ranging from 1 (very negative) to 7 (very positive). Consider that we have two sentences that we want to establish preference, which were not necessarily annotated by the same raters. In fact, the sentences can even be annotated by a different number of annotators. The approach creates a $N_1 \times N_2$ preference matrix, where N_1 and N_2 are the arbitrary number of raters for sentence one and sentence two. This matrix includes the comparison between all pairs of annotations and each entry can be a positive trend ($\uparrow$), negative trend ($\downarrow$), or equal value ($=$). For example, entry (1,2) compares the first annotator of sentence one (4), with the second annotator of sentence two (2). Since $4 > 2$, the trend is positive ($\uparrow$). The last step is to count how many comparisons resulted in a positive, negative, or equal trend. If the proportion of positive or negative trends is higher than a given threshold τ, we establish a preference between the sentences. For the example in Figure 1, the positive trends in the preference matrix include 12 out of 15 cells (80%). If the threshold is $\tau = 60\%$, we would consider sentence one to be preferred over sentence two (e.g., more positive emotional content). If a clear trend cannot be established, we do not use this pair of sentences for training our models.

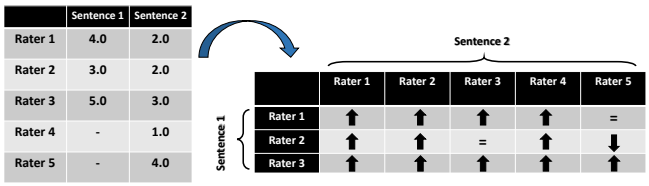


Fig. 1. QA-based relative labels for sentence-level annotations. In the example, there are 12 preferences for sentence one, 1 preferences for sentence two and 2 draws.

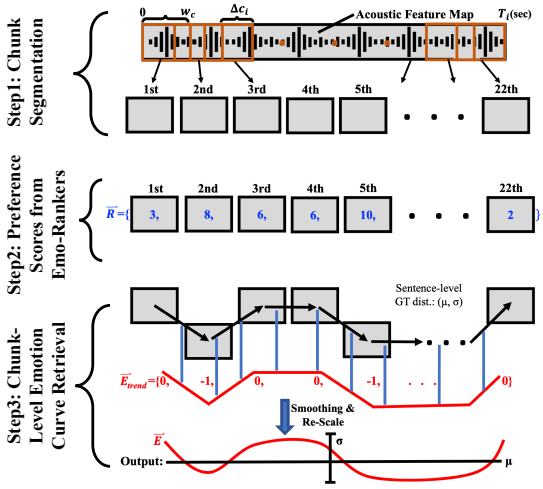


Fig. 2. A visualized example of the three steps to retrieve chunk-level emotional curve for a sentence. The retrieved curves are then treated as the target sequence to train the sequence-to-sequence SER model.

Since the defined preference labels are binary (i.e., sentence i is preferred over sentence j, or sentence j is preferred over sentence i), we formulate this problem as a binary classification task, where the input data are the features from sentence i (x_i) and sentence j (x_j). We use the *binary cross entropy* (BCE) as the loss function for the prediction output of $sigmoid(\Phi(x_i) - \Phi(x_j))$, where $\Phi(\cdot)$ is a function implemented with a *neural network* (NN) that preserves the preference between sentences. The chunk-level rankers are built relying on this approach, where the inputs correspond to the features of the chunks, rather than the entire sentence. We assume that the preference labels of the chunks are the same as the preference labels of the sentences. In this study, $\Phi(\cdot)$ is implemented with two consecutive LSTM layers. The final time step output of the last LSTM layer is considered the hidden representation of the input data of the chunk. This representation is further processed with two *fully connected* (FC) layers using the *rectified linear unit* (ReLU) as the activation function. We impose dropout regularization on both LSTM layers with a rate equal to $p = 0.5$. The hidden nodes of all the layers are set to 130, which is the same dimension as our input acoustic features (Sec. IV-D).

C. Chunk-level Continuous Emotional Curves Retrieval

We use the trained chunk-level rankers $\Phi(\cdot)$ to retrieve continuous emotional curves, relying exclusively on the sentence-level annotations. Figure 2 illustrates the proposed approach. The key idea is to identify emotional fluctuations within the sentence using preference learning. The proposed approach has three steps: create rankings across chunks, assign relative trends to create curves, and normalize the curve to match the second-order statistics of the emotional scores provided by annotators of the sentence.

The first step creates the ranks for the chunks. It starts by creating pairwise comparisons between chunks. If we have nC chunks, this step creates $nC(nC-1)/2$ ranking results. Then, we count how many times a chunk was preferred over other chunks, sorting the chunks in descending order according to this value. We denote by Rank-<*number*> the chunk-level ranking order. Rank-1 is given to the chunk that is the most preferred over the rest of the chunks. Rank-nC is given to the chunk that is the least preferred over the other chunks. If two chunks have the same number of preferences, we assign the same rank. We denote this rank as $\vec{R}$. For example, Figure 2 shows a sequence where the rank is $\vec{R} = [3, 8, 6, 6, 10, \ldots, 2]$, indicating that the first chunk in the sentence is the third most preferred chunk, and the second chunk is the eighth most preferred chunk within the sentence. In this example, the third and fourth chunks in the sentence have the same rank (i.e., sixth).

The second step is to assign relative trends to create the emotional curves. The chunk-based emotional curves are created with the trends across consecutive chunks in the rankings. The ranking sequence $\vec{R}$ provides up (1 unit) and down (1 unit) trends of the emotional content in a sentence with respect to a given emotional attribute. For example, Figure 2 shows that the rank of the first chunk (i.e., third) is higher than the rank of the second chunk (i.e., eight). Therefore, there is a negative trend from chunk one to chunk two. The value of the curve for the second chunk is *one unit* lower than the value of the curve for the first chunk. This process continues by comparing neighboring chunks. If two consecutive chunks have the same rank, we keep the curve constant. After this procedure, we can obtain the emotional trend sequence $\vec{E}_{trend}$ (see Fig. 2).

The third step is to smooth and normalize the $\vec{E}_{trend}$ curve to match the second-order statistics of the sentence-level label provided by annotators (Fig. 2). The curve is firstly smoothed with a moving average filter with window size N. We assume that a sentence is annotated by several annotators. We estimate the mean (μ), and standard deviation (σ) of these annotations, which are used to center the curves and scale their amplitude. We re-scale the smoothed sequence to maintain the sentence-level second-order statistic of the annotations (i.e., μ and σ). Equation 3 shows the re-scaling formula. The formula matches the mean of the chunk-based curve to μ, and scales the values such that the standard deviation is equal to σ. We can regard the procedure as a post-processing step on the retrieved $\vec{E}_{trend}$ curve. Finally, the post-processed vector $\vec{E} = \{e_1, e_2, \ldots, e_{nC}\}$ represents the chunk-level local continuous emotional sequence for a sentence.

$$\vec{E} = \mu + (\vec{E}_{trend} - \mu_{trend}) \times \frac{\sigma}{\sigma_{trend}} \tag{3}$$

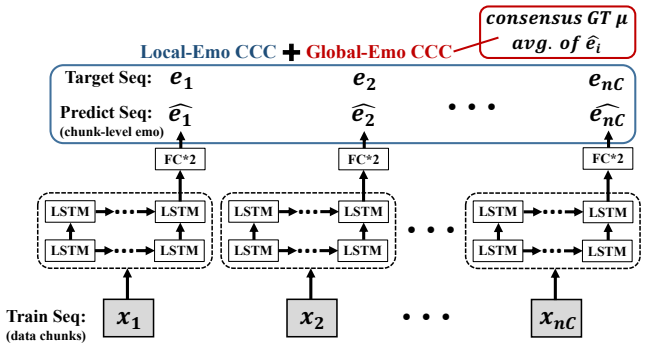


Fig. 3. The sequence-to-sequence model proposed in this study, where the input of data chunks are trained with a synchronized chunk-level emotion sequence. The loss function optimizes both local CCC and global CCC to capture the chunk-level and sentence-level emotions, respectively.

D. Sequence-to-sequence SER Model

The last step in our formulation is to use the chunk-level local continuous emotional sequence to train a sequence-to-sequence model. We treat $\vec{E}$ as the target sequence training with its corresponding data chunks $\{x_1, x_2, \ldots, x_{nC}\}$, resulting in a time-synchronized sequence-to-sequence SER model. Figure 3 describes the proposed network. In this study, the model architecture of the sequence-to-sequence NN model $\Psi(\cdot)$ is the same as $\Phi(\cdot)$, which we described in Section III-B. The important difference is the loss function. Equation 4 shows the loss function of the $\Psi(\cdot)$ model, which corresponds to a combination of local and global *concordance correlation coefficient* (CCC) losses. The local CCC loss is computed based on the prediction output of each chunk, while the global CCC loss is calculated by the average result of these chunks. As a result, the model $\Psi(\cdot)$ is optimized to capture both the chunk-level and sentence-level emotions. During the testing stage, we only evaluate the performance using the global CCC result (i.e., sentence-level prediction), since we do not have the *ground-truth* (GT) label for the chunk-level sequence.

$$Loss = \frac{1}{2} \times [(1 - CCC)_{local} + (1 - CCC)_{global})] \quad (4)$$

IV. Experimental Setting

This section describes the experimental setting used in this study.

A. The MSP-Podcast Corpus

This study mainly relies on the MSP-Podcast corpus [4], which consists of spontaneous speech segments that are rich in emotional content. These speech segments are collected from various online audio-sharing websites under Creative Commons licenses. The recordings include diverse topics such as sports, politics, and entertainment. The podcasts are segmented into speaking turns with a duration between 2.75 and 11 seconds and then processed to identify clean audio without music, noise, or overlapping speech. Then, candidate segments are selected relying on the retrieval-based framework proposed by Mariooryad et al. [52]. This approach uses machine learning SER models to identify speaking turns with target emotional content. The study in Lotfian and Busso [4] presents the details of this protocol.

The MSP-Podcast corpus is annotated using a crowdsource-based protocol inspired by the work of Burmania et al. [53] that tracks, in real-time, the performance of workers completing the evaluation. The perceptual evaluation includes categorical and attribute-based emotional annotations which are annotated at the sentence level after listening to the sentence. This study uses emotional attributes, which include arousal (calm to active), valence (negative to positive), and dominance (weak to strong). Each speech segment is labeled by at least five annotators to obtain robust consensus results. We use version 1.6 of the corpus, which is split into the train (34,280 speech turns), development (5,958 speech turns) and test (10,124 speech turns) sets. The partitions are manually defined to minimize cases where recordings from a speaker are included in different sets. We use the train set to build our experiments, maximizing performance on the development set. The final SER model is evaluated on the test set.

B. The MSP-Conversation Corpus

We consider the MSP-Conversation corpus [10], which complements the recordings of the MSP-Podcast corpus. Instead of focusing on isolated speaking turns, the corpus includes longer conversations with durations between 10 and 20 minutes. The recordings are sourced from the same podcasts used to collect the MSP-Podcast dataset. The annotations for the MSP-Conversation corpus correspond to time-continuous emotional traces of valence, arousal, and dominance. The perceptual annotations are conducted by evaluators hired by our laboratory. Each conversation is annotated by at least five evaluators. We use release 1.0 of this corpus. Martinez-Lucas et al. [10] provides more details about this corpus.

The overlap between the recordings in the MSP-Conversation and MSP-Podcast corpora provides a key resource for our study since we have time-continuous annotations for a set of speaking turns in the MSP-Podcast corpus. This feature is ideal to validate how good the retrieved chunk-level continuous emotional sequences in our proposed framework are. The collection of the MSP-Podcast and MSP-Conversation corpora are ongoing efforts in our laboratory. At the time we conducted this study, the number of speaking turns in the MSP-Podcast overlapping with recordings annotated with time-continuous annotations in the MSP-Conversation corpus is 2,884. Our evaluation uses these segments. Notice that the MSP-Conversation corpus is only used in this study for evaluating the retrieved chunk-level emotional curves. We do not utilize any label from this corpus to build our SER model.

C. The IEMOCAP and MSP-IMPROV databases

We validate the generalization ability of our proposed approach on the IEMOCAP [54] and MSP-IMPROV [3] datasets. Both corpora are widely used in SER studies. These two datasets are designed to elicit spontaneous emotional

expression with dyadic interactions between actors. In total, there are 10,039 audio clips included in the IEMOCAP corpus, and 8,438 audio clips included in the MSP-IMPROV corpus. Both of the datasets provide attribute-based annotations for arousal, valence, and dominance. Similar to the MSP-Podcast corpus, the labels correspond to sentence-level annotations.

D. Acoustic Features

To train our models, we use the *low-level descriptors* (LLDs) used in the proposed set for the Interspeech 2013 computational paralinguistic challenge [55]. The LLDs are frame-based features such as fundamental frequency, energy, and *Mel-frequency cepstral coefficients* (MFCCs). The LLDs are extracted using the OpenSmile toolkit [56]. In total, the set includes 130 frame-based acoustic features, which are normalized by subtracting the mean and dividing by the standard deviation. These parameters are estimated over the training set. The normalized LLDs feature map for each speaking turn is the input of our framework. Notice that we do not use *high-level descriptors* (HLDs), which are commonly extracted from LLDs (e.g., mean of the fundamental frequency).

E. Experimental Settings

This section describes the parameters used to implement our proposed framework. For the segmentation procedure, we set the desired chunk window length w_c to 1 second, which was the value used by Lin and Busso [20], [21]. We have noticed that emotion recognition can be effectively implemented even with a small window of 500 milliseconds [57], so this chunk duration is long enough to learn meaningful emotional information. We set the value of T_{max} to 11 seconds, which is the maximum duration of the sentences in the MSP-Podcast corpus (Sec. IV-A). With these parameters, we set C=11, following Equation 1. The factor n in Equation 2 is set to 2 to increase the overlaps between chunks and obtain more dense curves. After fixing these parameters, we can split a sentence into 22 chunks. This dynamic chunk segmentation process creates 22 chunks (nC) with a fixed duration of one second (w_c), regardless of the duration of the sentence T_i.

We need pairs of sentences to train the rankers (Sec. III-B). If the size of the corpus is L, there are $L \cdot (L-1)/2$ potential pairs that can be used to train and evaluate the models. In the MSP-Podcast corpus, there are hundreds of millions of relative labels that can be defined. It is expensive to utilize all pairs to build the model. Therefore, we randomly sample 100K preference pairs per epoch for both the training and validation processes. We observe that this number is sufficient to obtain robust emotional rankers. The threshold to establish preference (τ) is set to 60% (Sec. III-B). Finally, we use a window size of N=3 for the moving average filter that smoothes the generated emotional curves.

The models are implemented in Pytorch with a single NVIDIA GeForce RTX 2080 Ti GPU. The detailed model architecture is described in Section III-B. We use the Adam optimizer [58] with a batch size of 128 to train the emotion rankers ($\Phi(\cdot)$, Sec. III-B) and the sequence-to-sequence SER model ($\Psi(\cdot)$, Sec. III-D). The models $\Phi(\cdot)$ and $\Psi(\cdot)$ have the same architecture. The learning rate is fixed at 0.001. We do not observe any underfitting or overfitting issues during the training process. The models are able to converge within a maximum of 30 epochs. We optimized performance on the development set, saving the best models with an early stopping criterion based on the development loss. The performance of the final model is evaluated on the test set, reporting the average CCC results after running 10 experiment trials with different network initializations. We utilize this implementation strategy to conduct statistical analyses of the results using a two-tailed t-test based on the 10 trials. We define statistical significance at p-value $= 0.05$.

F. Baseline Methods

We design four baseline models to compare our proposed framework: *Seq2One*, *MeanInfo*, *RandomInfo* and *RandomInfo-smooth*. For a fair comparison, all models are trained and evaluated as described in Section IV-E.

We refer to the first baseline as *Seq2One*. For this approach, we do not split the sentence data into small chunks. Instead, the model directly predicts the sentence-level emotion using the feature map of the entire sentence. We use zero padding to reach a fixed length of 11 secs (i.e., max segment duration of the corpus) for the batch training. The loss function does not have the local CCC component (Eq. 4), since there are no local emotions provided during the training process.

We refer to the second baseline as *MeanInfo*. All the chunks share the same label assigned to the sentence μ, which follows the similar modeling concept used in Han et al. [41]. This method assumes that emotional information is uniformly distributed across a sentence (i.e., $\vec{E} = \{\mu, \mu, \ldots, \mu\}$).

We refer to the third baseline as *RandomInfo*. This baseline assigns local emotional information by considering the mean and standard deviation in the labels provided by the annotators. Each score e_i in $\vec{E}$ is independently sampled from the label distribution $N(\mu, \sigma^2)$. We re-sample the vector $\vec{E}$ in every epoch. This baseline model considers the differences in perception across evaluators by using μ and σ to create the curves. However, the emotional curves are not conditioned on the input acoustic features.

We refer to the fourth baseline as *RandomInfo-smooth*. This baseline is based on the *RandomInfo* curves, where we perform two additional post-processing steps. First, we smooth the sequence using the same approach used to smooth our predicted curves (Sec. III-C). Second, we correct variations of the sequence created by the smoothing step by re-scaling the created curves to again match the mean and standard deviation of the sentence-level annotations.

V. Experimental Results

A. The Performance of Emotion Rankers

The first evaluation in this study is assessing the performance of the rankers, which is a critical component in our formulation. We use Spearman's and Kendall's Tau correlation coefficient to evaluate the performance of the predicted rank orders. Both metrics measure the correlation between two ranked sequences by considering the number of concordant

TABLE II
EMOTIONAL RANKING PERFORMANCES OF THE TRAINED RANKERS. THE RESULTS ARE CALCULATED BASED ON THE AVERAGE PERFORMANCE OF 10 RESULTS, WHERE EACH RESULT IS A RANDOMLY SELECTED SUB-SET OF THE TEST SET.

	Aro.	Val.	Dom.
Kendall's Tau [22]	0.5810	0.1930	0.5050
Kendall's Tau	0.5215	0.1628	0.4519
Spearman's corr.	0.7040	0.2377	0.6225

and discordant pairs. We train the emotion rankers at the chunk-level, predicting preference between chunks. We aggregate the chunk-level predictions to derive sentence-level results. We achieve this goal by estimating the most predicted preference across chunk comparisons from both sentences. First, we split each of the two sentences into 22 chunks. Then, we compare the corresponding chunks creating 22 comparisons. We count the number of preferences predicted for each sentence (out of the 22 comparisons). The sentence with the higher number is considered as the preferred sentence, and its *preference score* increases by one. Every sentence is compared to the rest of the sentences (i.e., One-vs-Other) using the same process. Therefore, the *preference score* for each sentence ranges from 0 to $|V|$-1, where $|V|$ is the size of the test set. For instance, the sentence with the *preference score* equal to $|V|$-1 for valence is the predicted sentence with the most positive emotion in the test set (i.e., rank one). The ground-truth ranking order is estimated by sorting the consensus sentence-level labels (i.e., μ).

As we stated in the previous section, the number of compared pairs depends on the size of the test set, which is computationally expensive to predict if we consider all possible comparisons. We randomly sample 10% of the data from the test set to run the ranking predictions (i.e., subset containing 1,012 sentences). We repeat the random selection process 10 times, presenting the average prediction of the results.

Table II shows the performance of the rankers. We also display the performances reported on the original QA-ranker study [22] as a reference. Note that the results cannot be directly compared, since the acoustic features (LLDs vs HLDs), modeling approach (chunk-level vs sentence-level), and test set size (10,124 vs 7,181) are different. Table II shows that our ranker formulation achieves a certain level of accuracy that is slightly lower than the results reported in Parthasarathy and Busso [22]. Since our goal is to utilize the rankers as an intermediate model to generate the chunk-level emotion sequence, we consider that the performances are sufficient to detect local emotional changes within a sentence.

B. Sentence-level Emotion Recognition

This study aims to exploit the local emotional information to facilitate the sentence-level prediction of emotional attributes. Since our proposed framework adopts rankers to retrieve local emotional curves, we denote our method as *RankerInfo*. We compare this model with the baselines *Seq2One*, *MeanInfo*, *RandomInfo* and *RandomInfo-smooth* (Sec. IV-F).

TABLE III
SENTENCE-LEVEL CCC PREDICTION RESULTS FOR THE BASELINE MODELS AND OUR PROPOSED RANKERINFO APPROACH. THE RESULTS ARE AVERAGED OVER 10 TRIALS ON THE TEST SET OF THE MSP-PODCAST CORPUS. THE SYMBOLS $*$, $\star$, † AND ‡ INDICATE THAT THE IMPROVEMENTS IN THE PERFORMANCES OVER THE *Seq2One*, *MeanInfo*, *RandomInfo* AND *RandomInfo-smooth* BASELINES ARE STATISTICALLY SIGNIFICANT, RESPECTIVELY (TWO-TAILED T-TEST, p-VALUE < 0.05).

Approach	Aro.	Val.	Dom.
Seq2One	0.7075	0.2346	0.6300
MeanInfo	0.6968	0.2714	0.6123
RandomInfo	0.7028	0.2750	0.6191
RandomInfo-smooth	0.7019	0.2740	0.6158
RankerInfo (prop.)	$\mathbf{0.7120}^{**\dagger\ddagger}$	$\mathbf{0.3125}^{**\dagger\ddagger}$	$\mathbf{0.6324}^{\star\dagger\ddagger}$

All the models are trained with the same experimental setting described in Section IV-E. The only difference is the target local emotional sequence $\vec{E} = \{e_1, e_2, \ldots, e_{nC}\}$, noting that the *Seq2One* approach does not rely on local emotions. The evaluation of the model performance is solely based on sentence-level CCC results, since we do not have ground-truth labels for the target local emotions.

Table III presents the sentence-level prediction results obtained using the entire test set. The table shows that our proposed *RankerInfo* method significantly outperforms other baseline approaches for all emotional attributes. The sequence-to-sequence modeling approaches obtain better valence performance than the *Seq2One* method. The CCC value increases from 0.2346 to 0.3125 for this emotional attribute. It is common to observe that SER models struggle to capture long-term temporal dependencies when we directly feed the feature map of the entire sentence as input for a sequence-to-one task [21]. This result shows that our sequence-to-sequence modeling strategy can be a good way to resolve long-term temporal modeling issues, since it forces the models to track short-term (i.e., data chunks) emotions that are aligned with the global sentence-level emotion. However, the assignment of short-term local emotions for a sentence during training is critical. Incorrect information could mislead the models, leading to degraded performances. Specifically, we can see that the dominance results for the MeanInfo (0.6123), RandomInfo (0.6191), and RandomInfo-smooth (0.6158) approaches decrease from the Seq2One model (0.6300). In contrast, the proposed *RankerInfo* method significantly improves performance over all the sequence-to-sequence baseline models. This result suggests that the retrieved curves need to consider the emotional information conveyed in the acoustic features. Our method utilizes rankers trained with acoustic features to track relative changes in emotions over time, producing reasonable emotional curves that are useful for training sequence-to-sequence models.

C. Comparison With Other SOTA Performances

Since our proposed formulation is very novel, to the best of our knowledge, there is no other existing study that aims to achieve a similar goal (i.e., retrieving continuous emotion curves solely based on sentence-level labels as the complementary training target to facilitate a sequence-to-sequence

SER regression model). As shown in Table I, the conventional sequence-to-sequence SER approach requires human-annotated time-continuous labels to train the model, which we do not require in our task setup. Another relevant direction is to leverage the CTC loss to infer hidden emotion trends from sentence-level labels to facilitate the model training. However, this method is only applicable to an emotion classification task. Therefore, both cases are not ideal to compare with the proposed method. We decide to focus on sequence-to-one methods that aim to predict the emotional label assigned to the entire sentence. We list the results of existing SOTA methods that have used the same release of the MSP-Podcast corpus (v1.6) and have reported performance using the CCC evaluation metric. Notice that these results are based on different research topics (e.g., contrastive learning, teacher distillation, multitask learning, etc.), acoustic features (e.g., LLDs, Mel-spectrogram, deep features), and model complexity/architectures (e.g., Transformers, LSTM, CNN), which might result in an unfair comparison between each other. Table IV shows the comparison results using traditional acoustic features to construct their SER models, such as LLDs. The table shows that the proposed approach outperforms most cases achieving competitive recognition performances compared to other existing SOTA results. This comparison demonstrates the effectiveness of our method. Notice that the approach in Sridhar and Busso [59] requires access to unlabeled data from the target test set during the training stage, and imposes increased computational cost (i.e., 100 repetitions of Monte Carlo predictions, using five ensemble teacher-student models) to achieve better results.

Recently, studies have successfully used pretrained self-supervised deep feature representation such as wav2vec [60], and HuBERT [61] pretrained on a large-scale corpus to obtain better SER performances [62]–[64]. Since these pretrained *deep features* (DFs) involve extra hundreds of hours of external resources to pretrain the models that have highly complex architectures (e.g., 24-layers Transformers), a direct comparison with our approach is not fair. Therefore, we reimplement our approach by replacing the original LLDs features with the *wav2vec2-large* DF (1024D, extracted with the Huggingface library [65]). Table V compares the results with three other approaches that have reported CCC performance on the MSP-Podcast corpus (v1.6) using DFs. The table shows that our proposed approach obtains competitive performance compared to other SOTA methods. One potential reason for our results to be slightly lower than the method proposed in Mitra et al. [64] is the use of *multi-task learning* (MTL), which requires additional emotion categorical labels during the model training. Therefore, we also provide the results of our approach implemented with MTL modeling scheme (i.e., *RankerInfo-MTL*). The MTL adds a multiclass classification task to recognize the primary emotions, in addition to jointly predicting the three emotional attributes. This method obtains further CCC improvements. In summary, our proposed approach either achieves competitive or better performances than other existing SOTA frameworks without relying on additional information (e.g., emotion class labels or unlabeled testing data), and without compromising too much the computational cost. Notice that our approach is also flexible and can be implemented with other SOTA methods. We just need to reimplement the model $\Psi(\cdot)$. The retrieved curves can then be used to formulate the problem as a sequence-to-sequence task.

TABLE IV
COMPARISON WITH SOTA RESULTS USING TRADITIONAL ACOUSTIC FEATURES AS INPUT FOR TRAINING THE SER MODEL. THE TABLE LISTS OTHER APPROACHES THAT HAVE REPORTED CCC PERFORMANCE ON THE MSP-PODCAST CORPUS (V1.6)

Approach	Aro.	Val.	Dom.
MeanInfo [41]	0.6968	0.2714	0.6123
Lin [66]	0.6611	0.1756	0.5490
Lin [20]	0.6947	0.3072	0.6132
Sridhar [59]	0.7345	0.3230	0.6652
RankerInfo (prop.)	0.7120	0.3125	0.6324

TABLE V
COMPARISON WITH SOTA RESULTS USING PRETRAINED DEEP FEATURES AS INPUT FOR TRAINING THE SER MODEL. THE TABLE LISTS OTHER APPROACHES THAT HAVE REPORTED CCC PERFORMANCE ON THE MSP-PODCAST CORPUS (V1.6)

Approach	Aro.	Val.	Dom.
Li [63]	0.7060	0.3770	0.6390
Mitra [64]- WAV2VEC2.0	0.7300	0.4600	0.6500
Mitra [64]- Distillation	0.7300	0.3700	0.6500
RankerInfo (prop.)	0.7288	0.4471	0.6218
RankerInfo-MTL	0.7418	0.4615	0.6328

D. Comparison with Time-Continuous Traces

As we mentioned in Section IV-B, the MSP-Conversation corpus has conversations that overlap with 2,884 speaking turns in the MSP-Podcast corpus. Since the MSP-Conversation corpus was annotated with time-continuous traces, we can directly compare our chunk-based emotional curves with traces created by human evaluators. The first step is to establish the *ground-truth* (GT) of the chunk-level target sequence using the continuous annotations provided in the MSP-Conversation corpus. The time-continuous traces are annotated with a sampling rate of 60Hz. We utilize temporal pooling to collapse the time-continuous traces into chunk-level results. This step considers the average of the values of the time-continuous trace over the duration of the chunk as the ground truth. Every sentence in the corpus is labeled by at least five raters. We calculate the consensus (i.e., mean) result across all annotators by considering the reaction lag of the evaluators during time-continuous annotations [23], [24] (i.e., time that takes an evaluator to identify, judge, and annotate an emotional stimulus). Studies have shown that this reaction lag can be between three to six seconds [23], [67]. The results in Mariooryad and Busso [23] showed that using a fixed reaction lag is a good approximation, achieving similar results as more complicated compensation methods. Therefore, we compute six conditions with different reaction lags: $D \in \{0s, 1s, 2s, 3s, 4s, 5s\}$. The values of the labels in the MSP-Conversation corpus are between -100 and 100, which is different from the range of values in the MSP-Podcast corpus (i.e., from one to seven). Although the label

TABLE VI
SPEARMAN'S CORRELATION RESULTS BETWEEN RETRIEVED AND GROUND-TRUTH (GT) CURVES. THESE RESULTS ARE CALCULATED BASED ON THE 2,884 OVERLAPPED SEGMENTS BETWEEN THE MSP-PODCAST AND MSP-CONVERSATION CORPUS.

	Aro.	Val.	Dom.
BestLagGT vs RandomInfo	0.2595	0.2504	0.2572
BestLagGT vs RandomInfo-smooth	0.3996	0.3815	0.4018
BestLagGT vs RankerInfo (prop.)	**0.5885**	**0.5447**	**0.5688**

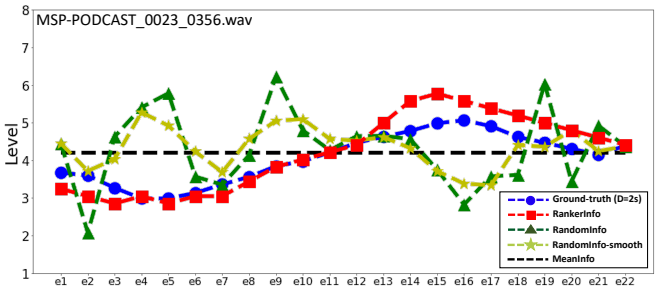


Fig. 4. Example of retrieved curves for valence based on different approaches. The figure also provides the ground-truth curve from the time-continuous annotations provided in the MSP-Conversation corpus.

scale does not affect the evaluation of emotional trends, we re-scale the chunk-level continuous GTs to fit the scale in the MSP-Podcast corpus.

After we obtained the six consensus chunk-level continuous GTs for each sentence, we compare them to the retrieved curves generated by the *RankerInfo*, *RandomInfo*, and *RandomInfo-smooth* approaches. We use Spearman's correlation coefficient as the evaluation metric. We compare the retrieved chunk curves with six versions (i.e., different reaction lags) of the chunk-level continuous GTs, reporting the one with the highest correlation for the sentence. Table VI shows the results. We can see that the correlations between *RankerInfo* and the chunk-level continuous GTs are above 0.5 (i.e., $\rho > 0.5$) for all the emotional attributes. These results suggest that our proposed method can effectively retrieve meaningful local emotion trends from sentence-level labels. In contrast, we observe a low correlation for the curves created with the *RandomInfo* and *RandomInfo-smooth* approaches, demonstrating that they cannot reflect dynamic emotional changes within a sentence.

Figure 4 shows a representative example of the retrieved curves by different approaches for a sentence (valence). As a reference, we also include the time-continuous emotional trace estimated using a reaction lag of two seconds. We observe that the curve generated by the proposed *RankerInfo* method effectively captures local emotional trends as the patterns are quite similar to the time-continuous emotional trace. The chunk-based emotional curve is also smooth, resembling the natural smoothness of the human ground-truth trace. In contrast, the curve generated by the *RandomInfo* approach presents abrupt changes over time. Even if we smooth the curves, they still result in unreliable emotional curves (see *RandomInfo-smooth*).

TABLE VII
LAG-1 AUTOCORRELATION RESULTS OF THE RETRIEVED CURVES BY DIFFERENT APPROACHES. WE ALSO SHOW THE CORRESPONDING RESULTS FROM THE GROUND-TRUTH (GT) CURVES AS A REFERENCE TO COMPARE WITH. THESE RESULTS ARE CALCULATED BASED ON THE 2,884 OVERLAPPED SEGMENTS BETWEEN THE MSP-PODCAST AND MSP-CONVERSATION CORPUS.

Approach	Aro.	Val.	Dom.
GT-curves	0.9046	0.9160	0.9079
MeanInfo	1.0000	1.0000	1.0000
RandomInfo	-0.0440	-0.0423	-0.0493
RandomInfo-smooth	0.5650	0.5592	0.5594
RankerInfo (prop.)	**0.8942**	**0.8824**	**0.8919**

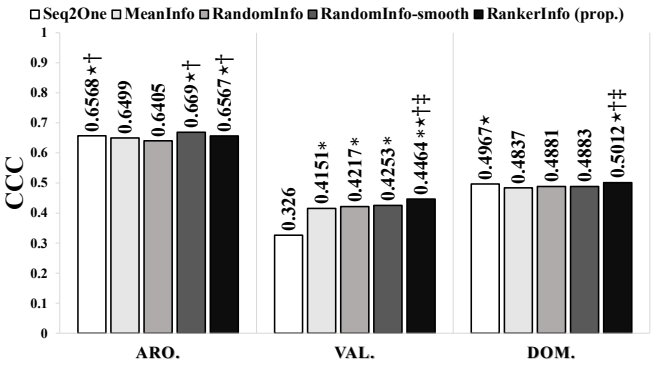
E. Time-Dependency of Retrieved Curves

We expect that the emotional content during a conversation should be related to the emotional content conveyed in previous segments [68]. Therefore, we should expect that the retrieved curves inherit this time-dependency characteristic presenting a high level of correlation between consecutive segments. We evaluate this property of the retrieved curves by using the lag-1 autocorrelation function, which is defined as the Pearson correlation between the curve and its lag-1 shifted signal. This analysis is computed at the chunk level. Therefore, this metric computes the relationship between the emotional values of consecutive chunks. We also estimate the lag-1 autocorrelation function over the chunk-level continuous GT labels as a reference (i.e., *GT-curves*). We report the results on the set of 2,884 speaking turns in the MSP-Podcast corpus that overlap with recordings in the MSP-Conversation corpus.

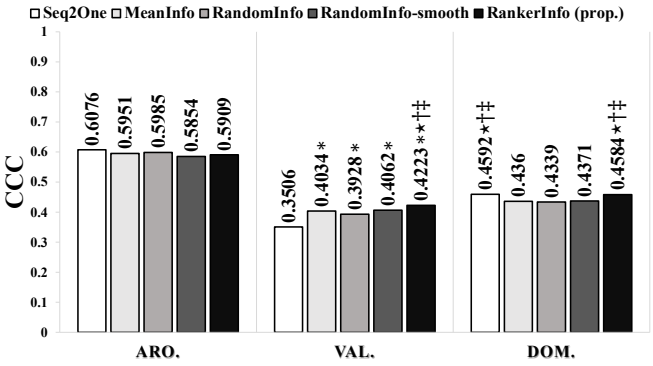
Table VII shows the results. We observe that chunk-based curves generated with the *RandomInfo* approach do not have the expected time consistency property, since their values are independently produced by sampling from the sentence-level label distribution without any conditions on the acoustic features. Table VII reports the results for the *RandomInfo-smooth* method since this baseline smooths the curves (i.e., moving average filter), which presumably improves the time consistency of the curves. However, the table shows that the results of this approach are still far from the ones observed in human annotations. In contrast, the curves generated by the proposed *RankerInfo* method are constrained on the input acoustic feature, where these chunks are partially overlapped. As a result, the values of the curves are more correlated, achieving high time dependency across nearby chunks. The correlation levels are very similar to the correlation values observed for the GT-curves.

F. Validation of Proposed Approach in Other Corpora

This section validates the benefits of our proposed approach with evaluations on the IEMOCAP [54] and MSP-IMPROV [3] datasets (Sec. IV-C). We only consider the within-corpus scenario, separately evaluating each dataset with its own defined train/development/test partitions. Therefore, the emotion rankers are trained per dataset to produce the chunk-level emotion curves for training the subsequent sequence-to-sequence



(a) Results on the IEMOCAP corpus



(b) Results on the MSP-IMPROV corpus

Fig. 5. Generalization CCC performances of our proposed approach based on the IEMOCAP and MSP-IMPROV datasets. The results are averaged over cross-validation folds designed per corpus. We use symbols to indicate whether the improvements are statistically significant over a given model: $*$ for *Seq2One*, $\star$ for *MeanInfo*, † for *RandomInfo* and ‡ for *RandomInfo-smooth* (two-tailed t-test, p-value < 0.05).

SER model. Since the collection of the IEMOCAP and MSP-IMPROV datasets did not impose a strict duration range for the audio clips, we artificially define the range from 1 to 17 seconds (i.e., $T_{\max}$=17 secs) and discard sentences outside of this range. This setting covers over 97% of the data for both datasets. Similar to Section IV-E, we set the chunk window size to w_c=1 sec (i.e., C=17), and the factor n=2. Therefore, each sentence is split into a fixed number of 34 chunks. The datasets do not provide predefined partitions, so we implement the evaluation using a speaker-independent *cross-validation* (CV) setting. The IEMOCAP has 5 dyadic sessions (i.e., each session involves 2 speakers) and the MSP-IMPROV has 6 dyadic sessions. Therefore, we perform a leave-one-session-out CV strategy, where one dyadic session is used for the test set, one dyadic session is used for the development set, and the rest of the corpus is used for the train set, resulting in a total of 5-folds and 6-folds CV for the IEMOCAP and MSP-IMPROV, respectively. All the model and training settings are the same as the ones described in Section IV-E, including the acoustic features (i.e., LLDs) and the baselines (Sec. IV-F). The reported CCC performances are the average of different CV-fold test results, which are also used to conduct statistical t-test. We assert statistically significant when p-value < 0.05.

Figures 5a and 5b show the summary of the results for the IEMOCAP and MSP-IMPROV corpora, respectively. The figures show that our approach consistently achieves the best recognition performances (except in some cases for arousal) when compared to the baselines on different SER datasets. This result is particularly clear for valence. The gains in performance are aligned with the results obtained in Table III, which demonstrates the effectiveness and generalization ability of the proposed method.

TABLE VIII
MEMORY AND COMPUTATIONAL COMPLEXITY ANALYSIS OF THE PROPOSED APPROACH DURING TRAINING AND INFERENCE. Φ AND Ψ REPRESENT THE SER MODELS FOR RANKERS AND EMOTIONAL PREDICTIONS, RESPECTIVELY.

Model Training Stage				
Approach	Model Size		Computational Cost	
	symbol	# para.	symbol	MACs
Baselines (Sec. IV-F)	$\mathcal{S}(\Psi)$	290K	$nC \cdot \mathcal{Q}(\Psi)$	0.37G
RankerInfo (prop.)	$\mathcal{S}(\Psi)$ + $\mathcal{S}(\Phi)$	580K	$nC \cdot \mathcal{Q}(\Psi)$ + $\mathcal{Q}(\Phi)$	8.23G
Model Inference Stage				
Approach	Model Size		Computational Cost	
	symbol	# para.	symbol	MACs
Baselines (Sec. IV-F)	$\mathcal{S}(\Psi)$	290K	$nC \cdot \mathcal{Q}(\Psi)$	0.37G
RankerInfo (prop.)	$\mathcal{S}(\Psi)$	290K	$nC \cdot \mathcal{Q}(\Psi)$	0.37G

G. Memory and Computational Complexity Analysis

This section provides an analysis of the memory and computational complexity of the proposed approach. Table VIII summarizes the values of the proposed model and *MeanInfo* and *RandomInfo* baselines. Since both of these baselines have the same architecture, we refer to them as "baselines" in Table VIII. $\mathcal{S}(\cdot)$ denotes the total number of parameters of the model, and $\mathcal{Q}(\cdot)$ denotes the computational operation, measured with *multiply–accumulate operation* (MACs). Table VIII shows that the main difference between our approach and the baselines is the need for extra emotional rankers $\Phi(\cdot)$. This block adds an extra cost during training to the computational complexity required by the sequence-to-sequence SER model $\Psi(\cdot)$. Since $\Psi(\cdot)$ produces chunk-level predictions, the full computational cost of the sentence-level emotion recognition system includes the terms $nC \cdot \mathcal{Q}(\Psi)$ to account for the nC chunks in the sentence. We set the emotion rankers to have the same architecture as the SER model (Sec. III-D). Therefore, the computational cost of $\mathcal{Q}(\Phi)$ is equivalent to $2\mathcal{Q}(\Psi) \cdot [nC(nC-1)/2]$. The $2\mathcal{Q}(\Psi)$ term represents the ranking cost since the rankers require a pair of inputs to generate the ranking result (i.e., doubled the cost). The $nC(nC-1)/2$ term indicates the total number of chunk-level comparisons needed to obtain the full ranking within a sentence (Sec. III-C).

For our LSTM model settings, $\mathcal{S}(\Psi)$=$\mathcal{S}(\Phi)$=290K parameters, nC=22, and $\mathcal{Q}(\Psi)$=0.017G MACs. As a clear limitation, our approach demands higher computational resources to train the model. However, it does not compromise extra complexity during the inference stage, since we do not need the emotion rankers to produce the final sentence-level recognition results.

VI. CONCLUSIONS

This study proposed a ranker-based framework to retrieve local continuous emotional curves from sentence-level labels.

The approach segments the sentence into chunks, comparing their emotional content with emotional rankers. The order generated by the rankers is used to establish positive and negative trends that are used to generate emotional curves. This approach facilitates a sequence-to-sequence formulation for SER tasks, leveraging local emotional changes within a sentence. The SER model that is trained with these retrieved chunk-based curves achieves the best sentence-level performance compared to baseline models. The performance gains are validated with three corpora, confirming the benefits of the proposed approach. Further analysis shows that the chunk-based emotional curves can successfully reflect dynamic local emotional changes observed in time-continuous annotations by human labelers. The retrieved chunk-based curves also follow the time consistency property observed in the time-continuous traces. Our novel formulation enables a sentence-level SER model to effectively leverage the non-uniform emotional distribution within a sentence.

A limitation of the approach is the increased computational complexity required to train the models since the approach requires the emotional rankers in addition to the emotional regression model. Since the rankers are not needed during inference, however, this problem is only present while training the models. Another limitation is that errors introduced by the retrieved curve may impair the learning of the emotions. The emotional rankers need to predict the underlying emotional content within the sentence with enough performance to avoid this problem.

There are different research directions to extend this approach. For example, we can modify our current two-stage framework into an end-to-end formulation. Our approach first uses the pretrained emotion rankers to generate the emotional curves. Then, it uses the chunk-based curves as labels to train the sequence-to-sequence SER model. These two steps can be combined by treating local emotions as a hidden variable output of the jointly trained rankers. Another research direction is to explore the nonuniform externalization of emotion using chunk-based curves. Analyzing emotional salient regions identified by our approach can lead to new findings on how we express and perceive emotions. Finally, our proposed approach can be effectively used to apply formulations designed for time-continuous traces to databases that are annotated with sentence-level annotations. Given the limited number of databases annotated with time-continuous annotations, using the retrieved chunk-based models can open new and exciting research opportunities.

ACKNOWLEDGMENT

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

REFERENCES

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

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

[3] 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.

[4] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

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

[6] C. Busso and S. Narayanan, "Joint analysis of the emotional fingerprint in the face and speech: A single subject study," in *International Workshop on Multimedia Signal Processing (MMSP 2007)*, Chania, Crete, Greece, October 2007, pp. 43–47.

[7] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007, pp. 2225–2228.

[8] H. Wang, A. Li, and Q. Fang, "F0 contour of prosodic word in happy speech of Mandarin," in *Affective Computing and Intelligent Interaction (ACII 2005)*, ser. Lecture Notes in Computer Science, J. Tao, T. Tan, and R. Picard, Eds. Beijing, China: Springer Berlin Heidelberg, October 2005, vol. 3784, pp. 433–440.

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

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

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

[12] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 1537–1540.

[13] X. Chen, W. Han, H. Ruan, J. Liu, H. Li, and D. Jiang, "Sequence-to-sequence modelling for categorical speech emotion recognition using recurrent neural network," in *First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia 2018)*, Beijing, China, May 2018, pp. 1–6.

[14] V. Chernykh, G. Sterling, and P. Prihodko, "Emotion recognition from speech with recurrent neural networks," *ArXiv e-prints (arXiv:1701.08071)*, pp. 1–16, January 2017.

[15] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Attention-enhanced connectionist temporal classification for discrete speech emotion recognition," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 206–210.

[16] J. Posner, J. Russell, and B. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, pp. 715–734, September 2005.

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

[18] W. Lin and C. Lee, "A thin-slice perception of emotion? an information theoretic-based framework to identify locally emotion-rich behavior segments for global affect recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5790–5794.

[19] H. K. Vydana, P. Vikash, T. Vamsi, K. P. Kumar, and A. K. Vuppala, "Detection of emotionally significant regions of speech for emotion recognition," in *Annual IEEE India Conference (INDICON 2015)*, New Delhi, India, December 2015, pp. 1–6.

[20] W.-C. Lin and C. Busso, "An efficient temporal modeling approach for speech emotion recognition by mapping varied duration sentences into

fixed number of chunks," in *Interspeech 2020*, Shanghai, China, October 2020, pp. 2322–2326.

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

[22] S. Parthasarathy and C. Busso, "Preference-learning with qualitative agreement for sentence level emotional annotations," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 252–256.

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

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

[25] C. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, October 2004, pp. 889–892.

[26] C. Busso and S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: a single subject study," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2331–2347, November 2007.

[27] R. Lotfian and C. Busso, "Lexical dependent emotion detection using synthetic speech reference," *IEEE Access*, vol. 7, no. 1, pp. 22 071–22 085, December 2019.

[28] S. Parthasarathy and C. Busso, "Defining emotionally salient regions using qualitative agreement method," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 3598–3602.

[29] R. Cowie and G. McKeown, "Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme," Belfast, Northern Ireland, UK, September 2010, SEMAINE Report D6b. [Online]. Available: http://semaine-project.eu

[30] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, October 2018.

[31] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, December 2014.

[32] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 2227–2231.

[33] D. Le, Z. Aldeneh, and E. Mower Provost, "Discretized continuous speech emotion recognition with multi-task deep recurrent neural network," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1108–1112.

[34] J. Huang, Y. Li, J. Tao, Z. Lian, and J. Yi, "End-to-end continuous emotion recognition from video using 3D ConvLSTM networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 6837–6841.

[35] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," in *International Conference on Electronic & Mechanical Engineering and Information Technology (EMEIT 2011)*, Harbin, China, August 2011, pp. 621–625.

[36] T. Seehapoch and S. Wongthanavasu, "Speech emotion recognition using support vector machines," in *International Conference on Knowledge and Smart Technology (KST 2013)*, Chonburi, Thailand, January-February 2013, pp. 86–91.

[37] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1845–1854, 2021.

[38] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA 2016)*, Jeju, Korea, December 2016, pp. 1–4.

[39] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions (IJSE)*, vol. 1, no. 1, pp. 68–99, January-June 2010.

[40] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *International Conference on Machine Learning (ICML 2006)*, Pittsburgh,PA, USA, June 2006, pp. 369–376.

[41] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech 2014*, Singapore, September 2014, pp. 223–227.

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

[43] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, May 2005.

[44] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The ambiguous world of emotion representation," *ArXiv e-prints (arXiv:1909.00360)*, pp. 1–19, May 2019.

[45] Y.-H. Yang and H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 762–774, May 2011.

[46] 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.

[47] S. Parthasarathy, R. Cowie, and C. Busso, "Using agreement on direction of change to build rank-based emotion classifiers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2108–2121, November 2016.

[48] 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.

[49] 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.

[50] ——, "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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

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

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

[57] J. Arias, C. Busso, and N. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," *Computer Speech and Language*, vol. 28, no. 1, pp. 278–294, January 2014.

[58] 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.

[59] K. Sridhar and C. Busso, "Ensemble of students taught by probabilistic teachers to improve speech emotion recognition," in *Interspeech 2020*, Shanghai, China, October 2020, pp. 516–520.

[60] 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.

[61] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

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

[63] M. Li, B. Yang, J. Levy, A. Stolcke, V. Rozgic, S. Matsoukas, C. Papayiannis, D. Bone, and C. Wang, "Contrastive unsupervised learning for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6329–6333.

[64] V. Mitra, H.-Y. S. Chien, V. Kowtha, J. Y. Cheng, and E. Azemi, "Speech emotion: Investigating model representations, multi-task learning and knowledge distillation," *arXiv preprint arXiv:2207.03334*, 2022.

[65] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

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

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

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

Wei-Cheng Lin (S'16) currently is a PhD student at the Electrical and Computer Engineering Department of The University of Texas at Dallas (UTD). He received his B.S. degree in communication engineering from the National Taiwan Ocean University (NTOU), Taiwan in 2014 and his M.S. degree in electrical engineering from the National Tsing Hua University (NTHU), Taiwan in 2016. His research interests are in human-centered behavioral signal processing (BSP), deep learning, and multimodal/speech signal processing. He is also a student member of the IEEE Signal Processing Society (SPS) and International Speech Communication Association (ISCA).

Carlos Busso (S'02-M'09-SM'13-F'23) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is an associate professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He was selected by the School of Engineering of Chile as the best electrical engineer who graduated in 2003 from Chilean universities. At USC, he received a provost doctoral fellowship from 2003 to 2005 and a fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) Laboratory [http://msp.utdallas.edu]. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. In 2015, his student received the third prize IEEE ITSS Best Dissertation Award (N. Li). He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He received the Best of IEEE Transactions on Affective Computing Paper Collection in 2021 (with R. Lotfian) and the Best Paper Award from IEEE Transactions on Affective Computing in 2022 (with Yannakakis and Cowie). He is the co-author of the winning paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interest is in human-centered multimodal machine intelligence and applications. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, nonverbal behaviors for conversational agents, in-vehicle active safety systems, and machine learning methods for multimodal processing. His work has direct implications in many practical domains, including national security, health care, entertainment, transportation systems, and education. He was the general chair of ACII 2017 and ICMI 2021. He is an IEEE Fellow. He is a member of ISCA, and AAAC, and a senior member of ACM.