An Interpretable Deep Mutual Information Curriculum Metric for a Robust and Generalized Speech Emotion Recognition System

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

***Abstract*—It is difficult to achieve robust and well-generalized models for tasks involving subjective concepts such as emotion. It is inevitable to deal with noisy labels, given the ambiguous nature of human perception. Methodologies relying on *semi-supervised learning* (SSL) and curriculum learning have been proposed to enhance the generalization of the models. This study proposes a novel *deep mutual information* (DeepMI) metric, built with the SSL pre-trained DeepEmoCluster framework to establish the difficulty of samples. The DeepMI metric quantifies the relationship between the acoustic patterns and emotional attributes (e.g., arousal, valence, and dominance). The DeepMI metric provides a better curriculum, achieving state-of-the-art performance that is higher than results obtained with existing curriculum metrics for *speech emotion recognition* (SER). We evaluate the proposed method with three emotional datasets in matched and mismatched testing conditions. The experimental evaluations systematically show that a model trained with the DeepMI metric not only obtains competitive generalization performances, but also maintains convergence stability. Furthermore, the extracted DeepMI values are highly interpretable, reflecting information ranks of the training samples.**

***Index Terms*—Curriculum learning, speech emotion recognition, mutual information, clustering, modeling methodologies.**

I. Introduction

CLASSIFICATION problems with ambiguous labels (i.e., noisy labels) are important but challenging tasks. In subjective recognition tasks such as *speech emotion recognition* (SER) [1], [2], humans may disagree on the emotional label despite listening to the exact same audio clips [3], [4], [5]. Therefore, emotional labels are often obtained with perceptual evaluations from multiple annotators. Different approaches have emerged to deal with multiple annotations describing the differences in human emotional perception. One approach aims to directly derive emotional relevant information by normalizing the annotations to create distributions of the emotional content (i.e., soft-label) [6], [7], [8], [9], [10], [11]. This approach introduces the uncertainty of the labels during the training procedure by considering secondary emotional content. Another more conventional approach is to use consensus labels (i.e., the wisdom of the crowd). Once the number of annotations per sample is large enough (typically more than five [12]), we can derive reliable consensus labels by computing their mean value or finding the majority class [13]. In fact, most SER studies apply this scheme to build their recognition models. However, a consensus label does not eliminate the presence of uncertainty embedded in the samples and ignores concurrency of emotions [14], which may lead to unstable and inferior local optimal convergence, resulting in significant degradation of accuracy [15]. Besides noisy labels, speech signals convey dynamic complex information such as speaker traits, channel effects and background interferences. These factors greatly increase the difficulty of building a robust and well generalized SER system.

A practical alternative to handling uncertainty in SER is to take full advantage of the source domain data for better model convergence. In particular, the use of curriculum learning is an appealing modeling methodology to achieve a robust and well generalized model [16]. Curriculum learning is a training strategy that identifies the difficulty of the training samples to define the order that the data is presented during training. A model using curriculum learning progressively optimizes the model from easy to difficult samples. This training procedure resembles the cognitive learning process of humans [17] and results in improved generalization and faster convergence [16]. Curriculum learning has been widely applied to various tasks, showing its effectiveness in areas such as machine translation [18], facial expression recognition [19] and audiovisual learning [20]. The core problem of curriculum learning is how to define a meaningful metric to quantify the difficulty of a sample. Difficulty metrics can typically be divided into two main categories: metrics based on the data or the label characteristics (e.g., high SNRs refer to the easy samples), and metrics based on a pre-trained model's prediction errors. A recent SER study by Lotfian and Busso [21] proposed to utilize a label-driven metric by quantifying the inter-evaluation agreement between annotators, under the premise that samples that are ambiguous for humans are more difficult for SER systems. With the growing amount of available training samples, we argue that a pure model-driven metric could be superior to a label-driven metric,

Received 22 January 2024; revised 2 July 2024; accepted 9 November 2024. Date of publication 27 November 2024; date of current version 9 December 2024. This work was supported by the National Science Foundation (NSF) under Grant CNS-2016719. The associate editor coordinating the review of this article and approving it for publication was Dr. Panayiotis Georgiou. *(Corresponding author: Carlos Busso.)*

Wei-Cheng Lin and Kusha Sridhar are with the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: wei-cheng.lin@utdallas.edu; kusha.sridhar@utdallas.edu).

Carlos Busso is with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: busso@cmu.edu).

Github: https://github.com/winston-lin-wei-cheng/DeepMI-Curriculum-Metric.

Digital Object Identifier 10.1109/TASLP.2024.3507562

since it implicitly inherits useful discriminative information from the pre-trained model.

This study introduces the *deep mutual information* (DeepMI) metric, which is a novel model-driven curriculum metric for attribute-based SER. This metric is extracted from a pre-trained semi-supervised DeepEmoCluster model [22]. The computation of the DeepMI metric mainly consists of two parts. First, we construct a DeepEmoCluster model, which contains a feature extractor, an emotional discriminator, and a cluster classifier. The cluster classifier recognizes pseudo-labels assigned to the training samples created with the K-means algorithm. This component is jointly optimized with an emotional discriminator (i.e., arousal, valence, or dominance regressor) to capture emotional information in the clustering process. Since the pseudo-labeling procedure is totally unsupervised, the model can incorporate a large amount of unlabeled data into the training process, forming a powerful *semi-supervised learning* (SSL) framework. Second, we consider the ground-truth consensus emotional labels Y and the two prediction outputs from the pre-trained DeepEmoCluster model (i.e., the emotional attribute score Y', and the clustering class X). With these variables, we construct their joint probabilities and obtain the *mutual informations* (MIs) $I(Y';Y)$ and $I(X;Y)$. The DeepMI metric is finally defined as the weighted combination of these two MI terms. $I(Y';Y)$ quantifies the model prediction errors, indicating training difficulties from the optimization perspective, and $I(X;Y)$ quantifies the relation between the emotions and acoustic patterns through the data-driven clusters, showing the corresponding agreements of emotional expressiveness in speech.

Our experimental results based on the MSP-Podcast corpus [23] demonstrate that the proposed DeepMI metric achieves *state-of-the-art* (SOTA) performances for curriculum learning approaches in SER. We consistently find an improved generalization ability and robust model performance for within corpus condition (i.e., MSP-Podcast corpus [23]) and cross-corpora evaluations (i.e., IEMOCAP [24] and MSP-IMPROV [25] corpora). These results show that the DeepMI metric is a better measurement of the difficulty level of a sample, creating a better curriculum to train SER models. Our analysis indicates that the DeepMI metric is highly interpretable. It explicitly shows the ambiguity level of the emotion targets in the training set. We found that the DeepMI curriculum model learns the data from extreme to neutral values of the emotional attribute. This result is interesting since samples with a neutral value for the emotional attributes are often more uncertain in SER systems [26]. The main contributions of this study are:

- We propose a novel model-driven DeepMI metric based on the DeepEmoCluster model, which achieves SOTA performances among other existing curriculum learning approaches in the SER field.
- The proposed DeepMI metric not only offers competitive recognition performances, but also provides interpretable insights of the training data.

The rest of the paper is organized as follows. Section II discusses the research background and related work. Section III presents the proposed methodology, providing detailed explanation of the DeepMI metric. Section IV gives the experimental setup, emotional corpora, acoustic features, baseline models, and implementation details used to train and test our approach. Section V provides the experimental results, including comprehensive interpretations of the proposed metric. Finally, Section VI presents the concluding remarks and future directions of this study.

II. Related Work

The DeepMI metric relies on SSL to construct reliable neural representations for complex acoustic patterns by using large amounts of unlabeled data. The approach is used to build a curriculum to train the SER models. This section discusses efforts to improve robustness and generalization of SER solutions (Section II-A). Then, it presents related studies on two key areas for our paper: SSL (Section II-B) and curriculum learning (Section II-C). We also summarize the major differences between the DeepMI strategy and previous studies (Section II-D).

A. Robustness and Generalization

Robustness of a recognition model is critical to avoid fluctuations to trivial variations of the input. Most studies in SER concentrate on signal-based variations. For instance, Triantafyllopoulos et al. [27] adopted an independent *speech enhancement* (SE) system prior to the main SER model as a pre-processing step, aiming to construct a robust SER under different *signal-to-noise ratios* (SNRs). Lin and Busso [28] proposed a complete chunk-level attention SER model to obtain a robust temporal model that can handle sentences of different durations. Another important but less discussed direction is a model-based variation, which is related to the convergence of the network. It has been found that a SER model can be easily overfitted during training [29], leading to drastic differences in performance for different initializations and poor model generalization. Therefore, researchers have developed various approaches to prevent overfitting a SER model such as increasing model regularization with multitask learning [29], data augmentation [30] or dropout layers [31]. These methods are often required to empirically fine-tune the network structure.

Generalization can be regarded as a wider scope of model robustness, requiring not only that the model is robust against within-corpus conditions (i.e., source domain), but also in cross-corpus conditions (i.e., different target domains). Typically, the core concept is to utilize partial information of the target domain to extract meaningful domain transformation functions between the source and the target domains [32]. Deng et al. [33] proposed an unsupervised domain adaptation framework based on a deep neural autoencoder architecture. The model learns the prior knowledge from unlabeled target domain data to achieve better cross-corpora recognition performances. Another domain adversarial approach aims to reduce the mismatches between source and target domains [34], [35]. By attaching an auxiliary domain classifier with a gradient reversal layer, the model is encouraged to learn a common representation that aligns the source and the target distributions [34]. However, these approaches require some prior knowledge of the target domain to construct the recognition model (i.e., part of the labeled or unlabeled data

from the target domain). This requirement is not feasible if the target domain is unavailable or unknown during the training stage.

B. Semi-Supervised Learning Approach in SER

Instead of continuously collecting a time-consuming and costly labeled data set, a large speech emotion corpus which consists of unlabeled data is much cheaper for building a reliable SER model. The major usage of SSL in SER is to leverage this large set of unlabeled data for extracting complementary information. SSL can capture diverse acoustic patterns across different speakers, noises, or microphone settings to form a better representation of emotional speech [29], [36], [37]. Studies with SSL proposed in SER include an emotional discriminator with a reconstruction-based network such as *autoencoder* (AE) [38], [39] or *variational autoencoder* (VAE) [9], [40]. Instead of attaching a single discriminator at the bottleneck layer, Chang et al. [41] and Latif et al. [42] extended the architecture to utilize different multitask discriminators such as other speech attributes (e.g., gender and speakers) and fake or real adversarial detectors. Parthasarathy et al. [29] and Huang et al. [43] adopted the ladder networks to introduce random noise perturbations between the encoder and decoder layers, imposing additional consistency regularization to obtain noise-invariant intermediate representations. All of these proposed methods can learn powerful and discriminative bottleneck representations, and improve the generalization of the models.

Another strategy is to employ inductive SSL approaches such as co-training [44] or pseudo-labeling [45]. The goal is to assign low entropy pseudo labels on the unlabeled data and force a discriminator to learn their relationship. The main drawback of self-training is error accumulation by mislabeled samples, which could lead to worse recognition results. Zhang et al. [46] proposed an enhanced SSL approach to alleviate this problem. The core idea was to keep re-evaluating the previously selected data (i.e., the unlabeled data that have high confidence pseudo labels) by the following retrained classifier. The approach can correct mislabeled data in future iterations with an improved model. Likewise, Lin et al. [22] presented the DeepEmoCluster framework, which utilizes an unsupervised cluster classifier to leverage unlabeled data. The classifier recognizes the clustering assignment labels created by the K-means algorithm. The model is able to fully exploit information from unlabeled data without potentially undermining original discrimination ability by mislabeled samples. Section III-A provides more details about this framework.

C. Curriculum Learning in Speech

Various speech-related studies have successfully applied curriculum learning methodology to obtain remarkable results. Studies have used the SNRs of the recordings as the direct difficulty metric for the curriculum learning to build a robust system against noise [47]. In the field of speech enhancement, Gao et al. [48], [49] proposed a *progressive learning* (PL) approach to train the model. The concept of PL is to learn the *reverse* order of a curriculum metric, guiding the model from the difficult samples (i.e. low SNR) to the easy samples (i.e., high SNR). Their results demonstrated the improved enhancement results in low SNR environments. Braun et al. [50] and Ranjan et al. [51] also introduced similar curriculum learning approaches to obtain improved recognition performance over baselines in *automatic speech recognition* (ASR) and speaker recognition tasks, respectively. Marchi et al. [52] presented another interesting curriculum metric designed for a multimodal speaker verification system. The text-dependent tasks (i.e., including a desired keyword) were regarded as easy samples. Then, they gradually relaxed the constraint of having a given spoken content to construct a text-independent task. These tasks were regarded as the difficult samples, completing the curriculum learning.

A growing number of studies have begun to use curriculum learning in the area of SER. The first study that introduced this training technique in SER was Lotfian and Busso [21]. This study quantified the disagreement-level of emotional labels between evaluators to define the level of difficulty for the curriculum. Some studies have explicitly incorporated the difficulty indicator as a multitask learning target during the training process, resulting in a difficulty awareness model [54]. More recently, Zhou et al. [55] defined a curriculum metric using the difficulty nature of the dataset. They treated recordings from a balanced and acted speech emotional corpus as the easy samples, and recordings from a spontaneous, large scale, in-the-wild dataset as the difficult samples. They used this strategy for cross-corpora SER modeling. In conversation emotion recognition, Yang et al. [56] proposed a hybrid curriculum, which consists of the *conversation-level* (CC) and *utterance-level* (UC) curriculum. CC defines the difficulty metric based on the *emotion shift* frequency observed within a conversation. UC computes the utterance-level emotion label similarity (i.e., cosine similarity) as the difficulty measurement.

D. Relation to Prior Work

In this study, we propose a novel MI-based metric which considers not only the sample difficulties in the training process (i.e., model prediction errors), but also the connection between acoustic patterns and emotions. This study builds upon our previous studies that introduced the semi-supervised DeepEmoCluster framework [22]. The focus of Lin et al. [22] is the model architecture and training strategy of the DeepEmoCluster formulation. In contrast, the aim of this study is to utilize the DeepEmoCluster framework for extracting the novel DeepMI curriculum metric. Understanding what samples are "hard" or "easy" for building the curriculum is not easy, especially when dealing with subjective concepts such as emotions. The proposed method using the DeepEmoCluster framework provides a principled approach to achieving this goal, leading to clear improvements in SER performance.

The approach is also related to the study by Lotfian and Busso [21], which defined the curriculum based on emotional perceptual differences (i.e., disagreement between different annotators). In contrast, the DeepMI approach quantifies the agreement level between the acoustic and predicted emotional space

TABLE I
SUMMARY OF THE MAIN DIFFERENCES BETWEEN THE DEEPMI STRATEGY AND OTHER EXISTING CURRICULUM METRICS USED FOR SER TASKS

Curriculum Approach	Interpretability	Model Performance	Worker-level Labels	Extra Complexity
Label-driven: [21]	High	Not considered	Required	Not required
Model-driven: [53]	Low	Considered	Not required	Required
DeepMI (prop.):	High	Considered	Not required	Required

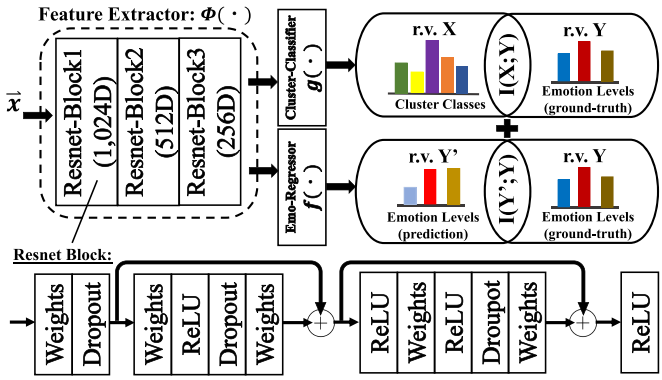

Fig. 1. The process of extracting the DeepMI curriculum metric using the training set. First, the DeepEmoCluster framework is pre-trained, including the feature extractor $\Phi(\cdot)$, the cluster classifier $g(\cdot)$, and the emotional regressor $f(\cdot)$. Then, the emotional label, the recognized cluster and predicted emotional attributes are used to compute joint probabilities. Finally, we derive mutual information measures to estimate the DeepMI metric.

from a SER model using mutual information units. This novel formulation does not require knowledge of individual emotional annotations used to derive consensus labels, which are not available for most of the SER corpora. Therefore, the proposed approach has a wider range of applications. In addition, DeepMI uses MI as the metric which has an appealing physical interpretation that is not available in other conventional mathematical formulations such as model prediction errors [53].

Table I summarizes the features of the DeepMI framework, comparing it to other existing curriculum approaches.

III. PROPOSED METHODOLOGY

Fig. 1 visualizes the system flow and model architecture for extracting the DeepMI metric. There are three main steps for building the proposed curriculum SER model, which utilizes the DeepMI metric. The first step is to construct the DeepEmoCluster model using the complete training set (Section III-A). The second step is to compute the DeepMI metric for the training set with the pre-trained DeepEmoCluster model (Section III-B). The third step is to define the curriculum learning strategy for the DeepMI metric (Section III-C).

A. The Semi-Supervised DeepEmoCluster Model

The proposed approach relies on the SSL framework proposed in our previous study [22], which we referred to as DeepEmoCluster. This section describes this approach. The DeepEmoCluster model consists of three modules as shown in Fig. 1: (1) the *feature extractor*, $\Phi(\cdot)$, which maps the input acoustic feature representations to hidden neural embeddings to extract task-relevant information, (2) the *cluster classifier*, $g(\cdot)$, which classifies pseudo class labels generated by the K-means clustering based on these embeddings, and (3) the *emotional regressor*, $f(\cdot)$, which predicts an emotional attribute (i.e., arousal, valence or dominance).

Algorithm 1: The Complete Training Procedure for the Semi-Supervised DeepEmoCluster Model.

Initialize:
$\Phi(\cdot)$: the feature extractor
$g(\cdot)$: the cluster classifier
$f(\cdot)$: the emotional regressor
$\mathbf{x}_{iU}$: i-th batch data for the unlabeled set $\mathbb{U}$
$\mathbf{x}_{iL}$: i-th batch training data for the labeled set $\mathbb{L}$
$\mathbf{y}_{emo,iL}$: i-th batch training targets for set $\mathbb{L}$
Main:
for *epoch = 1, 2, ..., N* **do**
  % pseudo-labeling for $\mathbb{U}$
  $\mathbf{h}_{iU} = \Phi(\mathbf{x}_{iU})$, $\forall$ i
  $\mathbf{y}_{cltr,iU} \leftarrow Kmeans(\mathbf{h}_{iU})$, $\forall$ i
  % Stage I: self-supervised path
  for *i = 1, 2, ... I* **do**
    $\hat{\mathbf{y}}_{cltr,iU} = g(\Phi(\mathbf{x}_{iU}))$
    $\mathcal{J}$ = CE($\hat{\mathbf{y}}_{cltr,iU}$, $\mathbf{y}_{cltr,iU}$)
    update $\Phi(\cdot)$ *and* $g(\cdot)$ *by* $\mathcal{J}$
  end
  % pseudo-labeling for $\mathbb{L}$
  $\mathbf{h}_{iL} = \Phi(\mathbf{x}_{iL})$, $\forall$ i
  $\mathbf{y}_{cltr,iL} \leftarrow Kmeans(\mathbf{h}_{iL})$, $\forall$ i
  % Stage II: jointly optimized path
  for *i = 1, 2, ... I* **do**
    $\hat{\mathbf{y}}_{cltr,iL}$, $\hat{\mathbf{y}}_{emo,iL} = g(\Phi(\mathbf{x}_{iL}))$, $f(\Phi(\mathbf{x}_{iL}))$
    $\mathcal{J}$ = CE($\hat{\mathbf{y}}_{cltr,iL}$, $\mathbf{y}_{cltr,iL}$) +
      [1 - CCC($\hat{\mathbf{y}}_{emo,iL}$, $\mathbf{y}_{emo,iL}$)]
    update $\Phi(\cdot)$, $g(\cdot)$ *and* $f(\cdot)$ *by* $\mathcal{J}$
  end
end

The DeepEmoCluster model is trained with a labeled set ($\mathbb{L}$) and an unlabeled set ($\mathbb{U}$). Algorithm 1 shows the detailed steps of the semi-supervised DeepEmoCluster training procedure, where each training epoch contains two stages. The first stage in each training epoch is to run the K-means clustering algorithm for neural embeddings $\mathbf{h}_{iU}$ of the unlabeled set $\mathbb{U}$. After this step, the cluster assignments of samples (i.e., $\mathbf{y}_{cltr,iU}$) are treated as training targets of the cluster classifier $g(\cdot)$. We compute the *cross entropy* (CE) loss that is used to update $\Phi(\cdot)$ and $g(\cdot)$. The second stage does the same pseudo-labeling procedure on the labeled set $\mathbb{L}$ to generate the cluster class labels $\mathbf{y}_{cltr,iL}$.

Since we also have the emotional labels $\mathbf{y}_{emo,iL}$ (where $emo \in \{aro, dom, val\}$), we can jointly optimize $\Phi(\cdot)$, $g(\cdot)$ and $f(\cdot)$. The overall cost function $\mathcal{J}$ (1) is the direct combination of the *concordance correlation coefficient* (CCC) and CE losses,

$$\mathcal{J} = \mathcal{J}_{emo} + \mathcal{J}_{CE} \tag{1}$$

where $\mathcal{J}_{emo} = 1 - CCC$ is the loss term for the emotional regressor. Notice that here we can introduce an additional weighting factor in $\mathcal{J}$ to control which task should be emphasized. The two-stage training strategy makes the DeepEmoCluster model to explore useful prior knowledge from the unlabeled set, before bringing the second discriminative stage for constructing meaningful cluster representations using an auxiliary emotional regressor.

A difference in our implementation of the DeepEmoCluster is the feature extraction module. The original implementation of the DeepEmoCluster used an end-to-end framework using the chunk-based segmentation presented in Lin and Busso [28]. Instead, we extract a feature representation by processing the *high level descriptor* (HLDs) described in Section IV-B. This study uses the *residual neural network* (ResNet) [57] architecture as the backbone for the feature extractor $\Phi(\cdot)$. Since our input feature is a vector (i.e., the 6,373 sentence-level representation described in Section IV-B) rather than a 2D feature map, we replace the original convolutional weighted layers with fully connected layers without pooling operations. The detailed network structure is illustrated in Fig. 1. It contains three standard ResNet blocks, where each block has two residual connections implemented with dropout layers (rate=0.3). The settings of the hidden nodes are formed to be an encoder-like architecture, mapping a 6,373D input vector into a hidden 256D output vector. We implement the cluster classifier $g(\cdot)$ and the emotional regressor $f(\cdot)$ with two fully connected layers implemented with the *rectified linear unit* (ReLU). The outputs are dependent on their corresponding losses to compute, where the softmax activation is applied for the cluster classifier, and a linear activation is used for the emotion score regressor.

B. The DeepMI Curriculum Metric

In this step, we use the pre-trained SSL DeepEmoCluster model to perform predictions on the labeled training set $\mathbb{L}$. Therefore, every sample in the training set has two predicted outputs (i.e., the cluster class and the emotional score) and its ground-truth emotional label. Then, we discretize the ground-truth and the predicted continuous emotional scores into E equally-spaced levels, regarding them as two discrete random variables: Y for the ground-truth label and Y' for the predicted label. Similarly, the cluster class output is directly treated as another discrete random variable X. The variable X is already discrete, with a total of K classes, where K is the number of clusters created by the K-means algorithm. We can construct a joint *probability mass function* (PMF) and its corresponding marginal PMFs by simply counting the number of joint combinations of the state of the variables, and by normalizing the resulting matrices by the size of the training set. We estimate $P(X, Y)$ and $P(Y', Y)$ to calculate the proposed DeepMI metric. Typically, the number of clusters K will be greater than the number of emotional levels, E. Then, we apply the definition of the *mutual information* (MI) in (2) to compute $I(X;Y)$ and $I(Y';Y)$. Since the speech cluster X is directly obtained from data using the DeepEmoCluster framework, it captures patterns observed on the acoustic features. $I(X;Y)$ captures the emotional dependency on the clusters created with the acoustic features. Therefore, $I(X;Y)$ indicates the mutual information between the acoustic features and emotional labels through the clusters. For the $I(Y';Y)$ term, it quantifies the discrimination ability of the pre-trained model (i.e., the prediction error). Finally, the DeepMI curriculum metric is defined in (3), which is a weighted combination of these two MI terms controlled by the hyper-parameter α.

$$I(X = x; Y = y) = P(x, y) \log \left(\frac{P(x, y)}{P(x)P(y)} \right) \tag{2}$$

$$DeepMI = \alpha I(Y';Y) + (1 - \alpha) I(X;Y) \tag{3}$$

Notice that the complete formula of MI includes the summation over all X and Y ranges. However, here we compute the MI value for each training sample, resulting in a DeepMI sequence of the train set. These instance-based MI values can be regarded as small components of the full MI, which might contain negative and positive values for different training samples. If we sum up all the MI components over the entire training set (since we define our random variables based on it), we obtain a positive value satisfying the non-negativity property of MI. Similar use of this instance-wise MI computation can also be found in the *natural language processing* (NLP) field for sentiment detection [58].

C. Proposed Curriculum Learning Schedule

Curriculum learning optimizes a recognition model step-by-step by gradually adding data bins into the training process according to their defined curriculum. Conventionally, we will start with the easiest data bin using a larger learning rate. Then, we progressively combine the existing data with the data from the bin with the next difficulty level, using a smaller learning rate to achieve hierarchical convergence. The key aspect is how to define the difficulty metric, which determines the sequential order in which the data is introduced. A good curriculum metric can effectively lead the model to find better local optima, resulting in well generalized recognition performances across matched or mismatched conditions.

For our curriculum settings, we split the original training set into 10 data bins based on the sorting order provided by the DeepMI metric. Larger values of DeepMI indicate that the samples are easy. Indeed, a larger value for $I(X;Y)$ indicates a stronger relation between the emotions and the acoustic clusters (e.g., extreme emotions tend to have a stronger acoustic pattern). Likewise, a larger value for $I(Y';Y)$ indicates a strong correlation between the predicted and ground truth labels (i.e., it refers to the model prediction errors). Notice that the $I(X;Y)$ term plays the most critical role in the DeepMI metric, since the variable X can leverage additional information from the unlabeled set to obtain extra prior knowledge of the data distribution, such

as higher diversity of the speaker traits, channel conditions and acoustic events.

We do not use the pre-trained SSL DeepEmoCluster model to evaluate our metric. Instead, we independently train a simple emotional regressor $\Theta(\cdot)$ following the curriculum learning scheme to evaluate the effectiveness of the DeepMI metric. The input of the model is the 6,373D feature vector mentioned in Section IV-B. The model structure of $\Theta(\cdot)$ contains three fully connected layers implemented with 512 nodes and with ReLU activation. We use dropout with a rate $p = 0.3$ to increase the regularization of the network. The output layer is a linear activation, and the loss function is implemented to minimize the term 1-CCC. We add a new data bin (i.e., the next level of difficulty) into the existing bins for every five training epochs. Every time we incorporate new data, we reduce the learning rate by half. Hence, the learning rate for the final stage will be 2^{-9} of the starting learning rate used for the first bin.

IV. Experimental Settings

A. Emotional Databases

The main corpus we utilize to build the recognition models is the MSP-Podcast corpus [23]. We also use the IEMOCAP [24] and the MSP-IMPROV [25] corpora as two additional test sets for the cross-corpora evaluations (i.e., both datasets are not considered during training).

1) The MSP-Podcast Corpus: The primary corpus used in this study is the MSP-Podcast corpus [23], which is the largest naturalistic speech emotion dataset consisting of emotionally rich spontaneous speech recordings gathered from podcasts from various audio-sharing websites. The podcast recordings are segmented using speaker diarization. We also use automatic algorithms for *signal-to-noise ratio* (SNR) estimation and music detection, creating a speech repository with speaking turns with durations between 2.75 s and 11 s without excessive background noise, music or overlapped speech. We use different machine learning techniques to retrieve emotionally rich segments to balance the emotional content in the corpus by following a retrieval-based strategy to collect data. The segments are annotated using *Amazon Mechanical Turk* (AMT). We follow a crowdsourcing protocol inspired by the method discussed in Burmania et al. [59]. The annotation includes primary and secondary emotions, and emotional attributes. This study relies on the emotional attributes arousal (calm versus active), valence (negative versus positive) and dominance (weak versus strong), which are annotated with *self-assessment manikins* (SAMs) on a seven point Likert scale. The ground truth attribute label for each speech segment is obtained by averaging the scores across the five or more annotators.

The database is split into train, test and development sets with the goal of minimizing the speaker overlap in the sets. We use version 1.8 of the MSP-Podcast corpus in this study. The development set has 7,800 samples from 40 speakers, the test set has 15,326 samples from 50 speakers, and the train set has 44,879 samples from the rest of the speakers (more than 1,000 speakers). There are around 600,000 speech segments that are not yet annotated with emotional labels. We use a portion of these recordings as the unlabeled set.

2) The IEMOCAP Corpus: The USC-IEMOCAP corpus [24] is an audiovisual corpus, where we only utilize the audio modality for this study. The corpus consists of dyadic interactions from 10 actors in improvised scenarios. The database contains 10,039 speaking turns. All the segments are annotated for arousal, valence and dominance by at least two raters using a five point Likert scale. The ground-truth for the emotional label for each sentence is obtained by averaging the scores across different annotators.

3) The MSP-IMPROV Corpus: We also use the MSP-IMPROV corpus [25], which is a multimodal emotional database that contains interactions between pairs of actors engaged in improvised scenarios. The dataset also contains the interactions between the actors during the breaks, making it more naturalistic. The corpus uses a novel elicitation scheme, where two actors in an improvised scenario interact, leading one of them to utter a target sentence. For each of the target sentences, four emotional scenarios were created to contextualize the sentence while eliciting happy, angry, sad and neutral reactions, respectively. This corpus consists of 8,438 speaking turns recorded from 12 actors (over 9 hours). The corpus was annotated with emotional categories and emotional attributes (arousal, valence and dominance) using the protocol described in Burmania et al. [59]. The ground-truth emotional attribute label assigned to each utterance is the average across the scores provided by the annotators, which is linearly mapped between -3 and 3.

B. Acoustic Features

This study uses the Interspeech 2013 *computational paralinguistics challenge* (ComParE) [60] feature set, extracted with the OpenSmile [61] toolkit. For each speaking turn, the toolkit firstly extracts frame-level *low level descriptors* (LLDs) such as fundamental frequency (f0), energy and *Mel-frequency cepstral coefficients* (MFCCs) using 32ms windows with 16ms overlap between windows. Then, the toolkit applies various statistical functions computed over the temporal dimension for each LLD (e.g., mean and standard deviation of the f0). These functions are also called *high level descriptors* (HLDs). In total, we obtain a 6,373 dimensions sentence-level feature vector, regardless of the duration of the speaking turn. We also perform the z-normalization on these features by estimating the mean and the standard deviation over the train set.

C. Implementation Details

Each emotional attribute is regarded as an independent regression task, where we build separate models for arousal, valence and dominance. The train and development sets are only coming from the MSP-Podcast corpus. We use three test sets in this study: the test set of the MSP-Podcast, the IEMOCAP and the MSP-IMPROV datasets. With the test set of the MSP-Podcast corpus, we show the recognition performances under matched conditions (e.g., spontaneous emotions and podcast recordings). With the IEMOCAP and the MSP-IMPROV corpora, we show

the recognition performances with mismatched conditions (e.g., acted emotions and laboratory recordings). The scales of the annotated emotional attributes are different across databases (e.g., the range for the attributes in the IEMOCAP is from 1 to 5, and in the MSP-Podcast is from 1 to 7), we train our model with the z-normalized emotional targets where the normalization parameters are calculated over the training set. Therefore, the model prediction outputs are normalized emotional scores without a mismatch with the emotional labels of the evaluation corpora.

For building the pre-trained semi-supervised DeepEmoCluster model, we randomly select 90K samples from the unlabeled pool (see Section IV-A1) to form the unlabeled set. Notice that we do not include any sample neither from the IEMOCAP nor the MSP-IMPROV corpora as data for the unlabeled set. The DeepMI metric has two important hyper-parameters: 1) the number of the K-means clusters K, which directly determines the diversity of the acoustic pattern space, and 2) the weighting factor α, which defines the importance of different relations between the variables X, Y and Y' ((3)). We fine-tune these two hyper-parameters based on the development set performances. For the K-means clusters K, we set it to 50 and discretize the continuous emotional scores into E=6 levels. Section V-D presents experimental results to demonstrate the impact of changing the number of clusters on the model recognition performances. For the DeepMI weighting factor α, we set to 0, 0 and 0.5 for arousal, dominance and valence, respectively. Interestingly, we found that for arousal and dominance, we can directly rely on $I(X;Y)$, since their acoustic patterns are highly related to the emotional labels. However, valence is a more challenging attribute compared to other emotions [31], [62], [63], demanding complementary information from the model's prediction error $I(Y';Y)$ to obtain better recognition results. Section V-E demonstrates the performance as we change the value of α.

To train the DeepEmoCluster model, we use the *stochastic gradient descent* (SGD) optimizer with a learning rate set to 0.001 for the first training stage (i.e., the self-supervised path in Algorithm 1). For the second stage (i.e., the jointly optimized path in Algorithm 1), we use the Adam optimizer, setting the learning rate equal to 0.0001. The models are trained with a batch size of 512 samples, using 200 epochs. For the model $\Theta(\cdot)$, which is the network used to evaluate our proposed curriculum learning formulation, we use the Adam optimizer where the learning rate starts at 0.001 and then follows the learning rate decay scheme mentioned in Section III-C to train the model. We use a batch size of 512, training the model for 70 epochs. This training procedure is sufficient to converge for all the models. We save the best models with an early stopping criterion based on the development loss, and report the accuracy of the model predictions in the three testing sets in terms of CCC. The CCC values are the average results after running 10 experiment trials with different network initializations. This implementation strategy allows us to conduct statistical analysis using a two-tailed t-test over the 10 trials. We define statistical significance at p-value $=$ 0.05. All the models are implemented in PyTorch under a single NVIDIA GeForce RTX 2080 Ti GPU environment.

D. Baselines for Curriculum Metrics

We consider three alternative curriculum strategies as baselines to compare our method: pretrained [53], disagreement, and minmax entropy [64]. We also consider training the models without a curriculum or defining the curriculum at random to serve as ablation baselines for verifying the effectiveness of curriculum learning.

- *W/O Curriculum:* This approach does not consider any curriculum, which follows the conventional training strategy to train the model with all the data for every epoch.
- *Random:* This approach creates a curriculum with random assignments to form the data bins (i.e., simply adding subsets of the training data to train the model). It serves as an ablation result for evaluating the role of having a meaningful curriculum strategy.
- *PreTrained [53]:* The approach first pre-trained an emotion recognition model based on the given training data (i.e., the model architecture only consists of the feature extractor $\Phi(\cdot)$ and the emotional regressor $f(\cdot)$, without the cluster module $g(\cdot)$ in Fig. 1). Then, the model is utilized to make predictions on the same training set. We compare point-wise distances between the predicted results with the actual ground-truths (i.e., L1 prediction errors). The difficulty metric is determined by the distances, assuming that more difficult samples have higher errors. We can simply switch the structure of $\Phi(\cdot)$ for model-agnostic PreTrained results. In this study, we use two implementations, one with RestNet and the other with *long-short term memory* (LSTM). PreTrained-ResNet denotes the approach that uses the same ResNet structure for $\Phi(\cdot)$ as we used in the DeepEmoCluster framework (Section III-A). This model provides a fair comparison, since it fixes the model complexity in $\Phi(\cdot)$. The PreTrained-LSTM baseline implements the structure for $\Phi(\cdot)$ using a LSTM, where the model input receives frame-level features (i.e., LLDs) instead of HLDs. Specifically, the model tracks the frame-level temporal dynamics with two stacked LSTM layers. Then, the last time-step output is passed through the emotional regressor $f(\cdot)$ to predict the sentence-level emotional attributes. This setting provides a complementary model-agnostic architecture to compare the results.
- *Disagreement:* This method quantifies the disagreement level between annotators for each sentence by computing the standard deviation among their annotated results. The higher the standard deviation, the more difficult the sample is.
- *Minmax Entropy [21]:* This approach is inspired by the *item response theory* (IRT) [65]. This method considers the expertise of annotators to determine whether the disagreement between labels is because of poor labelers or difficult samples. Zhou et al. [64], [66] modeled this problem using a minimax conditional entropy formulation, where the goal is to estimate the probability of a sample belonging to a particular class (unobserved true label of a sample) by taking both the sample difficulty and annotator expertise into account. If the observed label is $\tilde{y}_{i,j}$, which is the label

given by annotator j to a sample i. The probability that an annotator j labels a sample i with label q while the true label is e is given by, $P(\tilde{Y}_{i,j} = q|Y_i = e)$. $Q(Y_i = e)$ is the probability that a sample i has a true label e. The observed (P) and true (Q) label distributions are jointly estimated using a minimax entropy criterion shown in (4)

$$\min_{Q} \max_{P} H(\tilde{Y}|Y) \tag{4}$$

This formulation is used to estimate the difficulty of a sample by calculating the sample confusion matrix. The elements of the confusion matrix measure how likely a sample i is labeled as q by a randomly chosen annotator while its true label is e. Finally, the difficulty is calculated by taking the ratio between the trace of the matrix and the sum of its elements.

For fair comparison, all curriculum training settings and model structures follow the same strategy mentioned in Sections III-C and IV-C. The only difference is how we define the difficulty metric for determining the sequential order to create the curriculum learning.

V. Experimental Results

This section presents the experimental results of the proposed curriculum metric, comparing the results with the alternative baselines. The analysis also discusses the model convergence stability, the generalization capability of the models, the impact of a semi-supervised pre-trained model and the sensitivity of the model for variations of the hyper-parameters. We also illustrate the emotional distribution of the data included in the bins during the curriculum learning procedure.

A. Curriculum Learning Results

Table II presents different recognition performances under matched (i.e., MSP-Podcast corpus) and mismatched (i.e., IEMOCAP and MSP-IMPROV corpora) conditions. We consider the model generalization ability by evaluating how good a trained SER model can perform with unseen data. The results on the IEMOCAP and MSP-IMPROV corpora provide insights about the generalization of the models. As mentioned in Section IV-C, we conduct statistical analysis across approaches. Each of the approaches is represented by a specific symbol showing right next to it (i.e., $*$, †, ‡, ♣, ♦, ♥ and ♠). Values in the table that are tagged with these symbols indicate that the approach is significantly better than the approach indicated by the corresponding symbol. This analysis helps us to visualize how the performance of one approach compares with the results of another. For example, the DeepMI strategy significantly outperforms the *W/O Curriculum* strategy in six out of the nine testing cases (i.e., 3 emotional attributes $\times$ 3 datasets), since the symbol $*$, associated with the *W/O Curriculum* strategy, appears six times in the DeepMI results. This result shows that the DeepMI strategy is better than the *W/O Curriculum* strategy.

There are some interesting observations from these results. First, the *Random* curriculum method does not provide any benefit compared to the normal training strategy (i.e., the *W/O Curriculum* method). This approach even degrades the recognition performances. This result emphasizes the critical role of having a well-designed curriculum. Only a carefully designed and meaningful difficulty indicator can result in an effective curriculum learning. Second, we observe that model-driven metrics (i.e., the *PreTrained-LSTM*, *PreTrained-ResNet* and DeepMI methods) generally obtain better performances than other approaches including the label-driven *Disagreement* and *Minmax Entropy* metrics. This result validates our argument that a model-driven metric can be more effective when we have sufficient amounts of training data, since the model can implicitly bring useful discriminative information into the curriculum metric. Third, our proposed DeepMI metric achieves the highest performances in most cases with either matched or mismatched conditions. Besides a clear performance gap over label-driven metrics, the DeepMI strategy significantly outperforms the *PreTrained-ResNet* strategy in four out of the nine cases (i.e., 44% symbol ♥ in DeepMI results). The differences are even clearer when our approach is compared with the *PreTrained-LSTM* method, where the DeepMI strategy significantly outperforms the *PreTrained-LSTM* strategy in six out of the nine cases (i.e., 67% symbol ♦ in DeepMI results). In contrast, there are no cases (i.e., 0%) in the table showing

TABLE II
Recognition Performance Using Different Curriculum Approaches for the Test Sets in Matched (MSP-Podcast) and Mismatched (IEMOCAP and MSP-IMPROV) Conditions

	Aro [CCC]	Val [CCC]	Dom [CCC]
	MSP-Podcast Test Set		
W/O Curriculum*	0.625$^{\dagger\ddagger\clubsuit}$	0.275$^{\dagger\ddagger\clubsuit}$	0.544$^{\dagger\ddagger\clubsuit}$
Random†	0.598	0.233	0.511
Disagreement‡	0.591	0.231	0.504
Minmax Entropy$^{\clubsuit}$	0.598	0.234	0.505
PreTrained-LSTM$^{\diamondsuit}$	0.628$^{\dagger\ddagger\clubsuit}$	0.283$^{\dagger\ddagger\clubsuit}$	0.554$^{*\dagger\ddagger\clubsuit}$
PreTrained-ResNet$^{\heartsuit}$	0.627$^{\dagger\ddagger\clubsuit}$	0.281$^{\dagger\ddagger\clubsuit}$	0.551$^{*\dagger\ddagger\clubsuit}$
DeepMI (Prop.)$^{\spadesuit}$	0.626$^{\dagger\ddagger\clubsuit}$	0.304$^{*\dagger\ddagger\clubsuit\diamondsuit\heartsuit}$	0.553$^{*\dagger\ddagger\clubsuit}$
	IEMOCAP		
W/O Curriculum*	0.448$^{\dagger\clubsuit\diamondsuit}$	0.204$^{\dagger\ddagger\clubsuit\diamondsuit\heartsuit}$	0.276
Random†	0.431$^{\clubsuit}$	0.161‡	0.287*
Disagreement‡	0.448$^{\dagger\clubsuit\diamondsuit}$	0.142	0.294$^{*\heartsuit}$
Minmax Entropy$^{\clubsuit}$	0.408	0.178†‡	0.319$^{*\dagger\ddagger\diamondsuit\heartsuit\spadesuit}$
PreTrained-LSTM$^{\diamondsuit}$	0.426$^{\clubsuit}$	0.194$^{\dagger\ddagger\clubsuit}$	0.294$^{*\heartsuit}$
PreTrained-ResNet$^{\heartsuit}$	0.446$^{\dagger\clubsuit\diamondsuit}$	0.193$^{\dagger\ddagger\clubsuit}$	0.278
DeepMI (Prop.)$^{\spadesuit}$	0.442$^{\dagger\clubsuit\diamondsuit}$	0.204$^{\dagger\ddagger\clubsuit\diamondsuit\heartsuit}$	0.292$^{*\heartsuit}$
	MSP-IMPROV		
W/O Curriculum*	0.522$^{\clubsuit}$	0.280$^{\dagger\ddagger\clubsuit\diamondsuit}$	0.360$^{\clubsuit}$
Random†	0.515$^{\clubsuit}$	0.242$^{\clubsuit}$	0.378$^{*\clubsuit}$
Disagreement‡	0.515$^{\clubsuit}$	0.245$^{\clubsuit}$	0.376$^{*\clubsuit}$
Minmax Entropy$^{\clubsuit}$	0.503	0.219	0.340
PreTrained-LSTM$^{\diamondsuit}$	0.526$^{\dagger\ddagger\clubsuit}$	0.268$^{\dagger\ddagger\clubsuit}$	0.377$^{*\clubsuit}$
PreTrained-ResNet$^{\heartsuit}$	0.547$^{*\dagger\ddagger\clubsuit\diamondsuit}$	0.316$^{*\dagger\ddagger\clubsuit\diamondsuit}$	0.380$^{*\clubsuit}$
DeepMI (Prop.)$^{\spadesuit}$	0.543$^{*\dagger\ddagger\clubsuit\diamondsuit}$	0.331$^{*\dagger\ddagger\clubsuit\diamondsuit\heartsuit}$	0.386$^{*\ddagger\clubsuit\diamondsuit}$

Each approach is represented by a specific symbol shown in the first column. Values tagged with a symbol indicate that the approach is significantly better than the approach indicated by the corresponding symbol (two-tailed t-test, p-value < 0.05).

TABLE III
THE DEFINED VAR VALUES TO ASSESS THE MODEL CONVERGENCE STABILITY ACROSS DIFFERENT NETWORK INITIALIZATIONS

	Aro [VAR]	Val [VAR]	Dom [VAR]
	MSP-Podcast Test Set		
W/O Curriculum	0.96%	4.72%	0.92%
Random	0.67%	4.29%	1.96%
Disagreement	0.51%	1.73%	0.40%
Minmax Entropy	0.33%	4.27%	0.40%
PreTrained-LSTM	0.16%	0.71%	0.00%
PreTrained-ResNet	0.00%	0.71%	0.00%
DeepMI (Prop.)	0.00%	1.32%	0.18%
	IEMOCAP		
W/O Curriculum	5.80%	20.58%	7.97%
Random	3.48%	16.77%	6.97%
Disagreement	1.56%	3.52%	1.02%
Minmax Entropy	1.96%	12.36%	2.82%
PreTrained-LSTM	1.64%	3.09%	1.36%
PreTrained-ResNet	0.45%	2.07%	0.72%
DeepMI (Prop.)	1.58%	2.45%	1.71%
	MSP-IMPROV		
W/O Curriculum	5.56%	13.57%	6.11%
Random	2.72%	16.94%	6.35%
Disagreement	1.17%	4.90%	1.06%
Minmax Entropy	1.59%	8.68%	2.65%
PreTrained-LSTM	0.95%	7.84%	0.80%
PreTrained-ResNet	0.55%	1.27%	0.79%
DeepMI (Prop.)	0.92%	1.51%	1.55%

A higher VAR value means that the model is more sensitive against different initializations.

TABLE IV
COMPARISON BETWEEN A SEMI-SUPERVISED VERSION (SSL) AND A FULLY SUPERVISED VERSION (FSL) OF THE DEEPEMOCLUSTER FRAMEWORK USED TO DERIVE THE DEEPMI METRIC

	Aro [CCC]	Val [CCC]	Dom [CCC]
	MSP-Podcast Test Set		
FSL (K=30)	0.624	0.283	0.548
SSL-45K (K=40)	0.626*	0.290*	0.548
SSL-90K (K=50)	**0.626***	**0.304***	**0.553***
	IEMOCAP		
FSL (K=30)	0.433	**0.211**	0.298
SSL-45K (K=40)	**0.445***	0.205	**0.315***
SSL-90K (K=50)	0.442*	0.204	0.292
	MSP-IMPROV		
FSL (K=30)	0.555	0.329	0.374
SSL-45K (K=40)	**0.557***	0.326	**0.403***
SSL-90K (K=50)	0.543	**0.331**	0.386*

The symbol * indicates that the SSL metric is statistically significantly better than the FSL metric (two-tailed t-test, p-value < 0.05).

that the *PreTrained* approaches lead to significantly better results than the proposed DeepMI strategy (i.e., no symbol ♠ in all the *PreTrained-ResNet* and *PreTrained-LSTM* results). Our approach leads to better performance for most of the cases or reaches a similar performance to these approaches. These results demonstrate the effectiveness of the proposed DeepMI strategy over the *PreTrained* methods.

B. Analysis of Convergence Stability

Besides the improved generalization performances, we find that models trained with the right metric to define the curriculum can obtain better model convergence stability. Specifically, if a model's recognition performance has drastic differences depending on the network initializations, we consider that the model has low robustness due to its unstable convergence. Since we repeatedly run our experiments for 10 trials with different initializations, we can compare the convergence stability of different approaches by computing the standard deviation of their prediction performances across these 10 trials. To have a fair comparison, (5) defines the *relative variation* (VAR) metric to represent model convergence stability. The standard deviation of the CCC is normalized by its corresponding mean CCC value to show the relative variation across trials. Therefore, a high value for VAR indicates that the model is more sensitive to the network initialization.

$$VAR = \left(\frac{std\, CCC}{mean\, CCC} \right) * 100\% \tag{5}$$

Table III lists the VAR results for different approaches. Conventional training methods without curriculum (*W/O Curriculum*) and curriculum learning with incorrect metric (*Random*) have low model robustness against different initialization, especially for the valence attribute under mismatched conditions (i.e., IEMOCAP and MSP-IMPROV sets). The worst case reaches more than 20% performance deviations across different trials. In contrast, the proposed DeepMI method consistently maintained robust recognition performances under matched or mismatched conditions, showing its model robustness and convergence stability. Interestingly, we observe a general trend for model-driven metrics (i.e., *PreTrained-LSTM*, *PreTrained-ResNet* and DeepMI) of having a stable model convergence, suggesting the extra benefit of model-driven approaches.

C. The Effectiveness of the SSL DeepEmoCluster

As we mentioned in Section II, we consider whether the success of the DeepMI metric is due to the semi-supervised DeepEmoCluster. In this section, we explicitly demonstrate the performance benefits obtained by using the SSL implementation. We directly compare our approach with the DeepMI metric obtained with a *fully supervised learning* (FSL) DeepEmoCluster. The FSL approach trains the DeepEmoCluster without relying on the unlabeled set (i.e., only performs the second stage in Algorithm 1 during the training process). Since we have fewer training samples, we reduce the number of clusters K from 50 to 30, which follows the suggestion of the original DeepEmoCluster paper [22]. We also include another intermediate result that trains the model using 45K unlabeled data and 40 clusters to provide further insights. These pre-trained models (i.e., SSL-45K, SSL-90K and FSL DeepEmoCluster) are then utilized to extract their corresponding DeepMI metric for constructing the curriculum model $\Theta(\cdot)$. Table IV reports the recognition performances. We find significant improvements in performances under matched conditions when using the SSL DeepEmoCluster (i.e., the MSP-Podcast test set). It suggests that increasing the amount of unlabeled data benefits

TABLE V
THE RECOGNITION PERFORMANCES ON THE MSP-PODCAST TEST SET AS A FUNCTION OF THE NUMBER OF CLUSTERS K IN THE K-MEANS ALGORITHM

	Aro [CCC]	Val [CCC]	Dom [CCC]
SSL-90K	MSP-Podcast Test Set		
K=10	0.621	0.272	0.546
K=30	0.620	0.277	0.551
K=50	**0.626**	**0.304**	**0.553**
K=70	0.623	0.272	0.553

The results are trained using the SSL DeepEmoCluster using 90K samples as the unlabeled set. The results are the average CCC across 10 trials.

in-domain generalization. We also observe improvement gains in some of the mismatched conditions. Generally, the SSL model achieves higher performance than the FSL model. This result reinforces the advantage of exploring the distribution of the acoustic features for the input data derived from the unlabeled set. This approach can effectively enhance the performance of the model [67], [68].

D. Number of K-Means Clusters

The parameter K (i.e., number of K-means clusters) plays an important role in the proposed DeepMI metric. It not only directly affects the encoding hidden representation outputs of the DeepEmoCluster model, but also changes the subsequent joint probability $P(X,Y)$. Table V presents the recognition results of the curriculum learning-based approach by training $\Theta(\cdot)$ with DeepMI metrics obtained with a different number of clusters. For simplicity, we only show the results on the test set of the MSP-Podcast corpus, since we can observe similar improvement trends if the model is evaluated in either matched or mismatched conditions. The performance peaks for the three emotional attributes are located at K=50. K is a fine-tuned parameter that influences the recognition accuracy, especially for valence. It is interesting that K=50 was also the optimal value observed on the development set, which was used to define this hyper-parameter. The inferior performances when K=10 indicate that we should increase the value of K when we have a large training corpus (i.e., labeled data plus the additional 90K unlabeled set). Having enough data for training can ensure the DeepEmoCluster model has sufficient hidden clusters to encode most acoustic conditions for obtaining better representations (e.g., speakers, channels or microphone settings).

E. The Impact of DeepMI Weighting Factor

As we mentioned in Section IV-C, the weighting factor α in (3) determines the importance of the acoustic cluster variable X. A lower value for α increases the importance of the term $I(X;Y)$. The variable X is the most critical part of the DeepMI metric, since it can leverage additional useful information from the unlabeled data. This section evaluates the sensitivity of α in our approach. Table VI shows the results, where α is bounded between 0 to 1. One-way *analysis of variance* (ANOVA) evaluations indicate that the differences are statistically significant across different values of α for the three databases and three attributes, except for dominance in the MSP-IMPROV corpus

TABLE VI
THE RECOGNITION PERFORMANCES ON THE MATCHED (MSP-PODCAST) AND MISMATCHED (IEMOCAP AND MSP-IMPROV) CONDITIONS AS A FUNCTION OF THE HYPER-PARAMETER α (THE WEIGHTING FACTOR OF THE DEEPMI METRIC IN (3))

	Aro [CCC]	Val [CCC]	Dom [CCC]
SSL-90K	MSP-Podcast Test Set		
α=*0.00*	<u>0.626</u>	0.298	<u>0.553</u>
α=*0.25*	0.618	0.309	0.547
α=*0.50*	0.621	<u>0.304</u>	0.546
α=*0.75*	0.618	0.302	0.548
α=*1.00*	0.620	0.305	0.544
SSL-90K	IEMOCAP		
α=*0.00*	<u>0.442</u>	0.195	<u>0.292</u>
α=*0.25*	0.392	0.187	0.276
α=*0.50*	0.384	<u>0.204</u>	0.269
α=*0.75*	0.372	0.206	0.267
α=*1.00*	0.384	0.210	0.283
SSL-90K	MSP-IMPROV		
α=*0.00*	<u>0.543</u>	0.319	<u>0.386</u>
α=*0.25*	0.531	0.313	0.387
α=*0.50*	0.518	<u>0.331</u>	0.378
α=*0.75*	0.514	0.335	0.379
α=*1.00*	0.531	0.351	0.388

These results are trained with the SSL DeepEmoCluster (K=50) using 90K samples as the unlabeled set. We report the average CCC across 10 trials.

(ANOVA-test, p-value < 0.05). The underlined values in Table VI correspond to the adopted settings with the best performance observed on the development set (i.e., the results shown in Table II). Interestingly, arousal and dominance achieve the best generalization performances without relying on the $I(Y';Y)$ term in (3) (i.e., α=0). This result demonstrates the benefit of leveraging unlabeled data, which might potentially provide extra acoustic patterns with respect to the unseen testing conditions, leading to improve the model's generalization. However, we find the DeepMI metric needs the additional discriminated information from the $I(Y';Y)$ term to obtain better performance for valence (i.e., $\alpha \geq$0.5). We hypothesize that this finding is due to the ambiguity in the relation between the acoustic patterns and valence, which has been identified as a more challenging prediction task using acoustic features than arousal or dominance [31], [63]. Therefore, it requires additional discriminate information.

F. Flexibility to Adopt in SOTA Approaches

Once the difficulty order in the samples of the training set has been established, curriculum learning can be applied to any SER framework since this strategy is a general training scheme. We evaluate this idea by training other existing SOTA SER approaches, bringing further improvements in the recognition performance. To demonstrate this point, we conduct experiments with other SOTA SER models [69], [70] that involve finetuning a pre-trained self-supervised learning model for downstream SER tasks. Specifically, we select WavLM-large [71] as the backbone pretrained self-supervised model. In our setup, we freeze the encoder and utilize the mean-pooled hidden output, then finetune only the emotion prediction head. Importantly, this SOTA SER model is fine-tuned with the same curriculum scheduler detailed

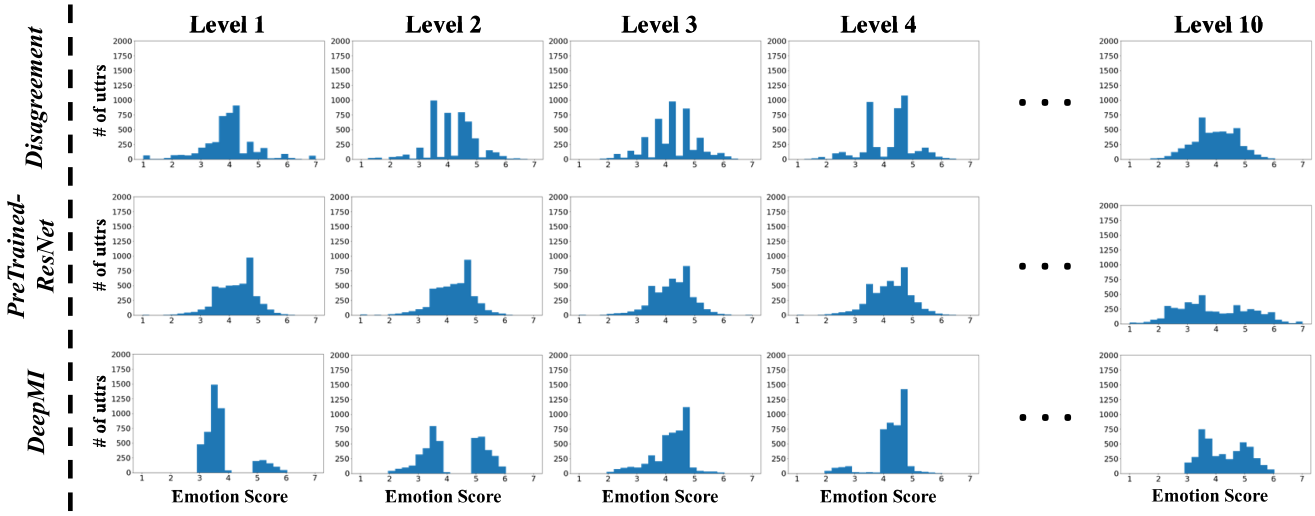

Fig. 2. The emotional distributions for valence of the data from the first four and last curriculum bins using the *Disagreement*, *PreTrained-ResNet* and DeepMI metrics. The plots only include the data introduced for each bin. The x-axes correspond to the ground truth emotional labels.

TABLE VII
THE SER PERFORMANCE WHEN APPLYING DEEPMI CURRICULUM LEARNING ON SOTA (WAVLM) SER APPROACHES

	Aro [VAR]	Val [VAR]	Dom [VAR]
	MSP-Podcast Test Set		
W/O Curriculum (WavLM)	0.674	0.469	**0.603**
DeepMI (WavLM)	**0.674**	**0.470**	0.601
	IEMOCAP		
W/O Curriculum (WavLM)	0.511	**0.401**	0.333
DeepMI (WavLM)	**0.541***	0.394	**0.379***
	MSP-IMPROV		
W/O Curriculum (WavLM)	0.545	**0.497**	0.397
DeepMI (WavLM)	**0.578***	0.470	**0.428***

The symbol * indicates that the proposed DeepMi approach is statistically significantly better than a model trained without curriculum (w/o curriculum) (10-trials, two-tailed t-test, p-value$<$ 0.05).

in Sections III-C and IV-C. This ablated approach allows us to clearly assess the impact of curriculum learning on the SER task. Similar to Table II, Table VII compares the strategies of training *W/O Curriculum* and with the DeepMI approach to show the validity of the proposed method.

Table VII shows that adopting curriculum learning with DeepMI metric for finetuning most of the time can significantly improve the model generalization performance over the plain finetuning strategy (see IEMOCAP and MSP-IMPROV, especially for arousal and dominance attributes). This consistent trend holds when using the traditional HLDs (Table II) or WavLM self-supervised deep features (Table VII), demonstrating the flexibility of the proposed approach.

G. Interpretation of the DeepMI Metric

Besides bringing performance gains, the proposed DeepMI metric is highly interpretable. One of the criticized drawbacks for model-driven curriculum metrics is their lack of interpretability. However, the DeepMI metric quantifies the curriculum into mutual information units, sorting information rankings between the acoustic patterns, predicted emotion and ground-truth emotional label. We can better understand our training corpus by observing the emotional patterns of the data with high mutual information.

To perform a curriculum learning scheme, we have to divide the original training set into various small data bins according to the predicted difficulties of the samples. Therefore, the most straightforward way to analyze the extracted curriculum metric is by comparing their corresponding ground-truth emotional target distributions among each training bin. Fig. 2 demonstrates the distributions for the first four and last bins of the valence attribute, where the horizontal axis corresponds to the ground truth labels. The figure considers only the new data entered at each stage, without accumulating the data from the bins. We only present the distribution for valence as an example since arousal and dominance show similar trend. Due to space consideration, we do not present all 10 levels in the figure. The size of MSP-Podcast train set is 44,879 utterances (see Section IV-A1) and is divided into 10 bins. Therefore, each bin has $\approx$4,488 utterances. Since the number of samples in each bins is consistent, the figures can be directly compared. We also report the distribution for the *Disagreement* and the *PreTrained-ResNet* metrics, since they are the most intuitive and competitive (in terms of recognition performances) approaches, respectively. Fig. 2 shows that the *Disagreement* and *PreTrained-ResNet* metrics do not reflect any specific pattern in the emotional distributions. They are generally a normal distribution centered around neutral values, mirroring the distribution of the entire data. We cannot find explicit explanation from these figures that accounts for the improvement in the recognition performances of the models. However, the DeepMI metric clearly shows distinctive patterns in the distributions. First, it identifies samples with extreme emotions (i.e., the two edge regions). Then, it gradually

includes the samples with neutral valence (i.e., the middle regions). This result suggests that the neutral areas are considered as more difficult samples for the model. This finding supports similar observations presented by Sridhar and Busso [26], who reported that the recognition model shows higher uncertainties (i.e., more ambiguous) in the prediction results for samples having neutral values for the emotional attributes. The model incrementally builds information knowledge from the easiest to the hardest samples, following the desired property of curriculum learning to improve recognition performances. Similar improvement results can be found in Lee et al. [72], where their model was designed to hierarchically recognize emotions based on empirical observations, focusing first on easy tasks (e.g., anger/happiness versus sadness), and leaving more difficult tasks for later stages (e.g., sadness versus neutral). Interestingly, the proposed DeepMI metric automatically defines this order directly from the data.

VI. Conclusions and Future Work

In this study, we introduced an advanced curriculum DeepMI metric for attribute-based SER tasks. The metric is extracted via a pre-trained semi-supervised DeepEmoCluster framework, forming a complete model-driven difficulty indicator for the training corpus to build the curriculum learning. Our evaluation results based on matched and mismatched testing conditions demonstrated that the proposed DeepMI results achieved SOTA recognition performances compared to other existing curriculum metrics in the SER field. We conclude that the success of the DeepMI metric is a combination of using semi-supervised learning and curriculum learning. The DeepEmoCluster model can utilize large amounts of unlabeled data to acquire robust hidden neural representations, resulting in a robust difficulty metric for performing curriculum learning. We also found that the convergence stability of DeepMI against different initializations is superior to other non-curriculum strategies or label-driven metrics. In addition to performance gains and robust model convergence, the DeepMI is highly interpretable, where we can directly observe the emotional patterns from the easiest to the hardest samples based on the DeepMI values.

One limitation of DeepMI is the requirement of the pre-trained DeepEmoCluster model, which inevitably introduces extra computational complexity, which is common to all other model-driven curriculum approaches [53]. An important future direction of this study is extending the DeepMI metric to a multimodality curriculum metric that also considers video, language and speech. Mutual information offers a flexible formulation to quantify relevant degree between two arbitrary types of variables once we can model them as discrete random variables. Since the DeepEmoCluster framework can be easily implemented with different modalities, we can always follow the same modeling strategy that we proposed in this study to obtain their corresponding hidden cluster representations to measure arbitrary combinations of information quantities from pair of variables, such as speech-language, speech-video and language-emotion. These cross modality metrics would provide comprehensive insights about human behavioral and emotional expressions.

References

[1] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 5084–5088.

[2] M. Abdelwahab and C. Busso, "Active learning for speech emotion recognition using deep neural network," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, Cambridge, U.K., Sep. 2019, pp. 441–447.

[3] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Commun.*, vol. 40, no. 1/2, pp. 5–32, Apr. 2003.

[4] B. Zhang, G. Essl, and E. Mower Provost, "Predicting the distribution of emotion perception: Capturing inter-rater variability," in *Proc. ACM Int. Conf. Multimodal Interaction*, Glasgow, U.K., Nov. 2017, pp. 51–59.

[5] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The ambiguous world of emotion representation," May 2019, *arXiv:1909.00360*.

[6] H.-C. Chou and C.-C. Lee, "Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Brighton, U.K., May 2019, pp. 5886–5890.

[7] A. Ando, S. Kobashikawa, H. Kamiyama, R. Masumura, Y. Ijima, and Y. Aono, "Soft-target training with ambiguous emotional utterances for DNN-based speech emotion classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 4964–4968.

[8] E. Mower, M. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotional profiles," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 5, pp. 1057–1070, Jul. 2011.

[9] K. Sridhar, W.-C. Lin, and C. Busso, "Generative approach using soft-labels to learn uncertainty in predicting emotional attributes," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, Nara, Japan, Sep./Oct. 2021, pp. 1–8.

[10] H. M. Fayek, M. Lech, and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *Proc. Int. Joint Conf. Neural Netw.*, Vancouver, BC, Canada, Jul. 2016, pp. 566–570.

[11] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, San Antonio, TX, USA, Oct. 2017, pp. 415–420.

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

[13] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng, "Cheap and fast–but is it good? Evaluating non-expert annotations for natural language tasks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Honolulu, HI, USA, Oct. 2008, pp. 254–263.

[14] H.-C. Chou, C.-C. Lee, and C. Busso, "Exploiting co-occurrence frequency of emotions in perceptual evaluations to train a speech emotion classifier," in *Proc. Interspeech*, Incheon, South Korea, Sep. 2022, pp. 161–165.

[15] B. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.

[16] Y. Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn.*, Montreal, QC, Canada, Jun. 2009, pp. 41–48.

[17] J. Elman, "Learning and development in neural networks: The importance of starting small," *Cognition*, vol. 48, no. 1, pp. 71–99, Jul. 1993.

[18] Y. Zhou, B. Yang, D. F. Wong, Y. Wan, and L. Chao, "Uncertainty-aware curriculum learning for neural machine translation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 6934–6944.

[19] L. Gui, T. Baltrušaitis, and L. Morency, "Curriculum learning for facial expression recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Washington, DC, USA, May/Jun. 2017, pp. 505–511.

[20] D. Hu, Z. Wang, H. Xiong, D. Wang, F. Nie, and D. Dou, "Curriculum audiovisual learning," Jan. 2020, *arXiv:2001.09414*.

[21] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 27, no. 4, pp. 815–826, Apr. 2019.

[22] W.-C. Lin, K. Sridhar, and C. Busso, "DeepEmoCluster: A semi-supervised framework for latent cluster representation of speech emotions," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Toronto, ON, Canada, Jun. 2021, pp. 7263–7267.

[23] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, Fourth Quarter 2019.

[24] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *J. Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[25] 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 Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, First Quarter 2017.

[26] K. Sridhar and C. Busso, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Barcelona, Spain, May 2020, pp. 8384–8388.

[27] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 1691–1695.

[28] W.-C. Lin and C. Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1215–1227, Second Quarter 2023.

[29] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2697–2709, 2020.

[30] F. Bao, M. Neumann, and N. Vu, "CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 2828–2832.

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

[32] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proc. Int. Conf. Mach. Learn.*, Bellevue, WA, USA, Jun./Jul. 2011, pp. 513–520.

[33] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 500–504, Apr. 2017.

[34] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 12, pp. 2423–2435, Dec. 2018.

[35] J. Gideon, M. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG)," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 1055–1068, Fourth Quarter 2021.

[36] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Learning representations of affect from speech," in *Proc. Workshop Track Int. Conf. Learn. Representations*, San Juan, Puerto Rico, May 2016, pp. 1–10.

[37] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Brighton, U.K., May 2019, pp. 7390–7394.

[38] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 1, pp. 31–43, Jan. 2018.

[39] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Florence, Italy, May 2014, pp. 4818–4822.

[40] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 3107–3111.

[41] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 2746–2750.

[42] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 992–1004, Second Quarter 2022.

[43] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and J. Yi, "Speech emotion recognition using semi-supervised learning with ladder networks," in *Proc. Asian Conf. Affect. Comput. Intell. Interaction*, Beijing, China, May 2018, pp. 1–5.

[44] Z. Zhang, J. Deng, and B. Schuller, "Co-training succeeds in computational paralinguistics," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 8505–8509.

[45] I. Triguero, S. García, and F. Herrera, "Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study," *Knowl. Inf. Syst.*, vol. 42, pp. 245–284, Feb. 2015.

[46] Z. Zhang, J. Han, J. Deng, X. Xu, F. Ringeval, and B. Schuller, "Leveraging unlabeled data for emotion recognition with enhanced collaborative semi-supervised learning," *IEEE Access*, vol. 6, pp. 22196–22209, 2018.

[47] P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe, "Curriculum learning: A survey," *Int. J. Comput. Vis.*, vol. 130, pp. 1526–1565, Jun. 2022.

[48] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "SNR-based progressive learning of deep neural network for speech enhancement," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 3713–3717.

[49] T. Gao, J. Du, L. R. Dai, and C. H. Lee, "Densely connected progressive learning for LSTM-based speech enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 5054–5058.

[50] S. Braun, D. Neil, and S. Liu, "A curriculum learning method for improved noise robustness in automatic speech recognition," in *Proc. Eur. Signal Process. Conf.*, Greece, Aug./Sep. 2017, pp. 548–552.

[51] S. Ranjan and J. Hansen, "Curriculum learning based approaches for noise robust speaker recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 1, pp. 197–210, Jan. 2018.

[52] E. Marchi et al., "Generalised discriminative transform via curriculum learning for speaker recognition," in *Proc. IEEE Int. Conf. Acoust. Speech aSignal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 5324–5328.

[53] A. Graves, M. Bellemare, J. Menick, R. Munos, and K. Kavukcuoglu, "Automated curriculum learning for neural networks," in *Proc. Int. Conf. Mach. Learn.*, Sydney, Australia, Aug. 2017, pp. 1–10.

[54] Z. Zhang, J. Han, E. Coutinho, and B. W. Schuller, "Dynamic difficulty awareness training for continuous emotion prediction," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1289–1301, May 2019.

[55] S. Zhou et al., "Inferring emotion from large-scale internet voice data: A semi-supervised curriculum augmentation based deep learning approach," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2021, pp. 6039–6047.

[56] L. Yang, Y. Shen, Y. Mao, and L. Cai, "Hybrid curriculum learning for emotion recognition in conversation," in *Proc. AAAI Conf. Artif. Intell.*, Feb./Mar. 2022, pp. 11595–11603.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun./Jul. 2016, pp. 770–778.

[58] F. Khan, U. Qamar, and S. Bashir, "SentiMI: Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection," *Appl. Soft Comput.*, vol. 39, pp. 140–153, Feb. 2016.

[59] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 374–388, 2016.

[60] B. Schuller et al., "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*, Lyon, France, Aug. 2013, pp. 148–152.

[61] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Int. Conf. Multimedia*, Florence, Italy, Oct. 2010, pp. 1459–1462.

[62] G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Shanghai, China, Mar. 2016, pp. 5200–5204.

[63] K. Sridhar and C. Busso, "Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 1959–1972, Fourth Quarter 2022.

[64] D. Zhou, Q. Liu, J. Platt, and C. Meek, "Aggregating ordinal labels from crowds by minimax conditional entropy," in *Proc. Int. Conf. Mach. Learn.*, Beijing, China, Jun. 2014, pp. 262–270.

[65] F. Lord, *Applications of Item Response Theory to Practical Testing Problems*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, Jul. 1980.

[66] D. Zhou, Q. Liu, J. Platt, C. Meek, and N. Shah, "Regularized minimax conditional entropy for crowdsourcing," Mar. 2015, *arXiv:1503.07240*.

[67] O. Chapelle, B. Schlkopf, and A. Zien, *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, Sep. 2006.

[68] V. Verma, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," in *Proc. Int. Joint Conf. Artif. Intell.*, Macao, China, Aug. 2019, pp. 3635–3641.

[69] L. Goncalves et al., "Odyssey 2024 - Speech emotion recognition challenge: Dataset, baseline framework, and results," in *Proc. Speaker Lang. Recognit. Workshop*, Quebec, Canada, Jun. 2024, pp. 247–254.

[70] J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10745–10759, Sep. 2023.
[71] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.
[72] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Commun.*, vol. 53, no. 9/10, pp. 1162–1171, Nov./Dec. 2011.

Wei-Cheng Lin (Member, IEEE) received the B.S. degree in communication engineering from the National Taiwan Ocean University (NTOU), Keelung, Taiwan, in 2014, the M.S. degree in electrical engineering from the National Tsing Hua University (NTHU), in 2016, and the Ph.D. degree in electrical engineering from the University of Texas at Dallas (UTD), Richardson, TX, USA, in 2023. He is currently a Full-Time Research Scientist with Bosch Research, Bosch Center for Artificial Intelligence, USA. His research interests include multimodal signal processing, speech and audio processing, deep learning, and affective computing. He is also a member of the IEEE Signal Processing Society (SPS), Association for the Advancement in Affective Computing (AAAC), and International Speech Communication Association (ISCA).

Kusha Sridhar (Student Member, IEEE) received the B.S. degree in electronics and communications engineering from PES University, Bangalore, Karnataka, India, in 2015, the M.S. degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 2017, and the Ph.D. degree in electrical engineering from the University of Texas at Dallas, Richardson, TX, USA, in 2021. He has worked as a Senior Machine Learning Engineer at Sony PlayStation from 2022 to 2024 and is currently a Staff Research Engineer at Hippocratic AI. His research interests include areas related to affective computing, focusing on emotion recognition from speech, multimodal LLMs, expressive speech synthesis and voice cloning, speech editing, machine learning, and speech signal processing.

Carlos Busso (Fellow, IEEE) received the B.S. and M.S. (with High Hons.) degrees in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, CA, USA, in 2008. He is currently a Professor with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA, where he is also the Director of the Multimodal Speech Processing (MSP) Laboratory. His research interests include human-centered multimodal machine intelligence and application, focusing on the broad areas of speech processing, affective computing, and machine learning methods for multimodal processing. He has worked on speech emotion recognition, multimodal behavior modeling for socially interactive agents, in-vehicle active safety systems, and robust multimodal speech processing. He was selected by the School of Engineering of Chile as the best electrical engineer who graduated in 2003 from Chilean Universities. He was the recipient of the NSF CAREER Award, ICMI Ten-Year Technical Impact Award in 2014, Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie), Best of IEEE Transactions on Affective Computing Paper Collection in 2021 (with R. Lotfian), Best Paper Award from IEEE Transactions on Affective Computing in 2022 (with Yannakakis and Cowie), Distinguished Alumni Award in the Mid-Career/Academia category by the Signal and Image Processing Institute (SIPI) at the University of Southern California in 2023, and 2023 ACM ICMI Community Service Award, and his Students were also the recipient of the Third Prize IEEE ITSS Best Dissertation Award (N. Li) in 2015, and the AAAC Student Dissertation Award (W.-C. Lin) in 2024. He is also an Associate Editor for IEEE TRANSACTIONS ON AFFECTIVE COMPUTING. He is also a Member of AAAC, Senior Member of ACM, and an ISCA Fellow.