Curriculum Learning for Speech Emotion Recognition from Crowdsourced Labels

Reza Lotfian, *Student Member, IEEE,* and Carlos Busso, *Senior Member, IEEE*

***Abstract*—This study introduces a method to design a curriculum for machine-learning to maximize the efficiency during the training process of *deep neural networks* (DNNs) for speech emotion recognition. Previous studies in other machine-learning problems have shown the benefits of training a classifier following a curriculum where samples are gradually presented in increasing level of difficulty. For speech emotion recognition, the challenge is to establish a natural order of difficulty in the training set to create the curriculum. We address this problem by assuming that ambiguous samples for humans are also ambiguous for computers. Speech samples are often annotated by multiple evaluators to account for differences in emotion perception across individuals. While some sentences with clear emotional content are consistently annotated, sentences with more ambiguous emotional content present important disagreement between individual evaluations. We propose to use the disagreement between evaluators as a measure of difficulty for the classification task. We propose metrics that quantify the inter-evaluation agreement to define the curriculum for regression problems and binary and multi-class classification problems. The experimental results consistently show that relying on a curriculum based on agreement between human judgments leads to statistically significant improvements over baselines trained without a curriculum.**

***Index Terms*—Curriculum learning, speech emotion recognition, inter-evaluator agreement**

I. Introduction

EMOTION recognition can play an important role in advanced *human computer interaction* (HCI) applications that are more aware of the users. Recently, there has been a number of studies showing the feasibility of emotion recognition in different applications, opening novel research directions. For example, a robot capable of recognizing and expressing facial expressions can more effectively communicate and interact with humans [1]. Emotion recognition systems have also been used in designing interactive games [2], [3] and tutoring systems [4], [5] to teach social interactions to children with autism spectrum disorders [6].

While previous studies have made important advances, emotion recognition from speech still faces many challenges. The most important challenges are generalizing the models across datasets [7]–[10], accounting for individual differences [11]–[14], dealing with environmental conditions [15] and recognizing subtle expressions [16], [17].

Recently deep learning techniques have enabled advances in many areas of speech processing. A barrier of using deep learning in *speech emotion recognition* (SER) is the lack of large training datasets. While tasks such as *automatic speech recognition* (ASR) have access to datasets with several hours of speech samples collected from multiple speakers, most affective datasets contain only tens or hundreds of subjects displaying a small range of emotions. The limited resources make it extremely difficult to train models that generalize well within and across datasets. Therefore, it is important to utilize the limited available training resources in the most efficient way.

Inspired by language learning in children, Elman [18] demonstrated that a neural network is able to better learn a task when the training data is sequentially presented from simple to complex. When the order is random, the network may fail to converge to an optimal set of parameters. This training approach is referred to as curriculum learning and has been successfully used in word representations in natural language processing [19], working memory based problem solving tasks [20], controlling robots [21], [22], and executing computer programs [23]. More recently, Bengio et al. [24] demonstrated that curriculum learning results in better generalization and faster learning on synthetic vision and word representation learning tasks. Interestingly, this approach has not been used on SER. We hypothesize that curriculum learning can be effectively employed to deal with the limited training resources available in this area. We also hypothesize that using a curriculum leads to models that learn more general emotion cues before learning more subtle speaker dependent cues, improving the generalization of the models. For these reasons, this work explores curriculum learning for SER. Emotion recognition has attributes which make it a good candidate for applying curriculum learning training schemes. The main evidences are the complex nature of the problem and the way we learn to perceive emotions, which happens gradually from infantry to adulthood.

This study consistently demonstrates that prioritizing less difficult samples at the beginning of the training process of a *deep neural network* (DNN) leads to significant improvements for SER systems. We suggest different methods to detect challenging examples which are only used in the more advanced phases of the training process. By adjusting the learning rate of the model in decreasing order, we reduce the influence of difficult samples in setting the parameters of the system. The key hypothesis to build the proposed curriculum is that ambiguous samples for humans are also ambiguous for computers. We have observed in our previous work that classifiers trained and tested with samples consistently annotated by human evaluators, achieved higher accuracies than classifiers training with more ambiguous samples [25]. Han et al. [26] observed that emotion prediction can be improved by estimating at the same

R. Lotfian and C. Busso are with the Erik Jonsson School of Electrical & Computer Engineering, The University of Texas at Dallas, Richardson TX 75080.
E-mail: reza.lotfian@utdallas.edu, busso@utdallas.edu

time the inter-rater agreement. These observations suggest that the level of difficulty for each sentence can be estimated by considering the inter-evaluators' agreement. The level of uncertainty in perceptual evaluations can be used as a measure of difficulty, since an easy task for human is often an easy task for computer. We propose two different implementations for this idea using individual annotations available for each sentence. We rely on the reliability of the annotations since SER is a problem known to have high level of uncertainty for both human and machine. The third metric measures the reliability of automatic methods, which is used as a baseline. We exhaustively evaluate this approach by considering three different formulations for SER. For attribute-based emotional descriptors (e.g., arousal, valence and dominance), we consider regression models to predict the label values. We also consider binary classification after dichotomizing the classes into low versus high values (e.g., low valence versus high valence). For categorical emotions (e.g., happiness, sadness, and anger), we consider a classification problem. We observe improvement in the performance in all three problems when the training process follows the right curriculum, compared to the approach of training with all the samples in one pass. We observe that the best curriculum is based on the item difficulty derived with the minimax conditional entropy criterion suggested by Zhou et al. [27], [28].

The main contributions of this work are two folds. First, to the best of our knowledge, this is the first study on SER using curriculum learning. Second, we propose a novel method to estimate the difficulty of the training examples using individual evaluations. We derived different metrics to quantify the disagreement between evaluators, proposing alternative curriculums for each of the machine-learning formulations considered in this study. The use of curriculum learning leads to consistent performance improvements, demonstrating the benefits of the proposed approach.

The rest of this paper is structured as follows. Section II reviews previous studies on SER that are related to this study. The section also discusses the implementations of curriculum learning in other machine learning applications. Section III describes our proposed method to apply curriculum learning to SER. Section IV presents the database, features and implementation details used in the evaluation. Section V includes experiments that consistently show the benefit of using the proposed curriculums for this task. Section VI concludes this study, giving new directions for future research in this area.

II. Background

This section reviews studies on SER and the process of annotating emotional labels, which provide the information needed to design the proposed curriculum learning framework (Sec. II-A). The section also discusses the general idea of curriculum learning and how it has been applied to different problems in previous studies (Sec. II-B). The section further reviews methods to estimate the difficulty of the tasks without knowing the ground truth for the tasks using minmax conditional entropy framework (Sec. II-C). We apply this method to estimate the difficulty of recognizing emotions conveyed in each speech sample.

A. Speech Emotion Recognition

The problem of detecting human emotions in speech, like any supervised machine-learning problem, relies on training examples. Early attempts for collecting emotional databases relied on actors to collect expressive behaviors with predefined emotional category [29]–[31]. This approach led to satisfactory results when testing under the same database, but failed to perform well under naturalistic situations [32]. In everyday spontaneous conversations, emotions are subtly expressed, creating expressive displays that are not well represented by portrayed behaviors provided by actors, especially if the recordings correspond to read sentences [32], [33]. Therefore, the classifiers trained with acted corpora do not generalize well in real life situations. To address this issue, researchers are collecting speech samples under more naturalistic conditions, either by indirectly eliciting target emotions [25], [34] or recording spontaneous speech [35], [36]. In these scenarios, the emotional content of the speech samples are not determined, and emotional labels need to be assigned using perceptual evaluations [37], [38].

When annotators are asked to answer questions about their perceived emotion, they often provide conflicting answers. There are several reasons for the disagreement, including presence of subtle emotions, presence of multiple valid emotions (e.g., sarcasm), evaluators' reliability, and inadequate questionnaires with options that do not properly describe the perceived emotion. Recent studies have even leveraged the disagreement between evaluators to train more robust classifiers using soft labels [39], [40]. In our previous work, we demonstrated that classifiers trained with more ambiguous samples achieved lower classification performance that classifiers trained with samples that are consistently evaluated by raters [25]. This result suggested that ambiguous samples for evaluators are also ambiguous samples for speech emotion classifiers. Lee et al. [41], [42] proposed to learn categorical emotions by using a hierarchical framework, where easy problems are solved before resolving more difficult problems. They obtained improved performance over classifiers without this hierarchical approach. These studies suggest that curriculum learning can be an appropriate framework for SER using more effectively the limited size of existing emotional databases.

B. Curriculum Learning

Most supervised machine learning methods take a one-pass learning approach, where all the training data is used to build the models. Studies have argued that this is not the best approach, suggesting that a classifier should learn basic patterns from clear examples, leaving harder examples for later training stages [18]. The idea of learning simple patterns before complex patterns is usually called curriculum learning. This idea is implemented during the training process by first presenting examples that can be easily recognized, after determining the difficulty level of the samples.

Bengio et al. [24] studied different approaches to implement this idea, showing that a curriculum that introduces the training data from easy to hard can lead to better local minima when

training a classifier with a non-convex criterion, which is commonly the case when training DNNs. This training approach results in better generalization and speed-up the convergence. Defining the training objective by giving higher weight to the easier samples is equivalent to solving a smoothed version of the target criterion. The easiest optimization problem $C_0(\theta)$ shares the same parameters (θ) as the target problem $C_1(\theta)$, capturing its coarse structure. The parameter λ for the intermediate problems of the curriculum, C_λ, has to monotonically increase based on the difficulty of the samples from 0 in the easiest problem (C_0) to 1 in the target problem (C_1). A training example z receives a weight to manipulate the difficulty of a problem. Designing a curriculum requires to adjust the weight parameter $W_\lambda(z)$ such that the difficulty of C_λ monotonically increases with λ. Bengio et al. [24] considered discrete numbers of λ and also discrete weights (0 or 1) for the training examples (z) at each step. As the value of λ increases by changing $W_\lambda(z) = 0$ to $W_{\lambda+\epsilon}(z) = 1$, they gradually introduced more difficult samples to the training pool, therefore, increasing the overall difficulty of the problem. They showed the effectiveness of this approach on different problems by training DNNs, comparing the results with a framework using the common one-pass learning approach.

If a natural order to quantify the difficulty of the task is available, the design of the curriculum becomes straightforward. In shape recognition of objects, Bengio et al. [24] suggested to start with basic shapes. They increased the difficulty of the task by varying the object position, size, and orientation. They also changed the gray levels of the foreground and background in the image. In language modeling, they predicted the most likely word given the previous words in a sentence in grammatically correct English. For this problem, they controlled the difficulty of the training samples by controlling for the vocabulary size [24]. When quantifying the difficulty of the training set is not clear, studies have proposed alternative methods. An intuitive approach for supervised task is to train a classifier using the conventional one-pass learning approach, evaluating the models on the training set. A sample is considered difficult if it falls on the incorrect side of the classifier's hyperplane [43]. The distance of the sample to the hyperplane can also be informative, where samples located close to the hyperplane in the feature space are considered more difficult examples than samples that are far from the hyperplane. Gui et al. [44] use this method as a baseline to build a curriculum for facial expression analysis. They compared this method to a curriculum built by relying on the intensity of the expressed emotions. In language model, this model-driven approach to build the curriculum can be implemented by assigning the difficulty level based on whether the model is able to predict the next word [24]. These methods are vulnerable to over-fitting, since the difficulty of the samples is obtained by testing the training examples with a classifier trained on the same data.

There are few studies that have considered curriculum learning in speech tasks. Braun et al. [45] used curriculum learning to improve the noise robustness of the automatic speech recognition by gradually adding samples with higher *signal-to-noise ratio* (SNR). Ranjan and Hansen [46] used curriculum learning for speaker recognition. They created the curriculum by adding audio from different channels according to their noise level. In their applications, the measure of difficulty is explicit, since higher SNR directly increases the difficulty of the speech tasks. As an alternative to curriculum learning, Zhang et al. [47] exploited the difficulty to boost the learning process in SER. They used the reconstruction error of the features. This information is then provided to another learning stage, which is expected to focus on more difficult regions. Their proposed approach requires time-continuous annotation of the input samples. Our approach creates the curriculum based on the uncertainty in the annotations, which is a more appropriate approach for SER.

Instead of relying on the performance of previously trained classifiers to extract the difficulty information, this study proposes to rely on the labels provided by human annotators by measuring their inter-evaluator agreement. The motivation behind this framework is that, in recognizing emotions, humans clearly outperform existing artificial intelligence solutions, and, therefore, it is reasonable to assume that a curriculum based on a metric of reliability across evaluators may provide a suitable criterion for this task. The challenge is that, at the individual level, we do not have access to the confidence level of human annotators, unless we explicitly ask them to annotate their certainty level in their perceptual judgments. Therefore, we rely on the implementation of the minmax conditional entropy method proposed by Zhou et al. [27], [28] to indirectly estimate the level of difficulty of a sample, creating an appealing curriculum for SER.

C. Minmax Method for Crowdsourced Labels

The proposed criteria to define curriculum is based on the minmax method. Recognizing emotions from speech is a subjective task where different evaluators often disagree on the perceived emotion [48]. Therefore, researchers tend to rely on ratings provided by evaluations with different reliability, aggregating the results to achieve consensus labels for the train set. If the reliability of all the raters are identical, the straightforward approach to find consensus judgment from multiple dissident annotations is to find the average for numerical labels or use the majority vote rule for categorical labels. In case the individual's reliability is known, we can rely on weighted majority consensus by giving higher weights to the annotations from reliable raters. In practice, the reliability of an individual worker is unknown. This problem is particularly important for evaluations conducted using crowdsourcing. When ratings are collected through crowdsourcing evaluations, some workers are accurate, while others are less reliable. Some raters are only interested in the payments, limiting their effort as much as possible, providing very noisy annotations.

To correct the bias of unreliable raters in the overall decision, Dawid and Skene [49] introduced an *expectation maximization* (EM) algorithm to simultaneously estimate the bias of annotators and the label classes in an iterative process. A key part of the Dawid and Skene method is to estimate a latent probabilistic confusion matrix for generating labels for each worker. The off-diagonal elements of the matrix

represent the probabilities that the worker mislabels an item from one class as another. The diagonal elements correspond to his/her accuracy for each class. The assumption is that the performance of a worker characterized by his/her confusion matrix stays the same no matter which tasks they are assigned. This assumption is not accurate in many labeling tasks, where some items are more difficult to label than others. For example, a worker is more likely to mislabel a difficult item than an easy one. Moreover, an item may be easily mislabeled with another class due to the ambiguity in the sample. Therefore, a fair evaluation of the raters should separately compensate for the difficulty of each task. Zhou et al. [28] developed a minimax conditional entropy method to jointly infer the task difficulty, raters' bias, and the labels of the samples. Our work uses the difficulty measure found in this *minmax entropy* (ME) approach to design the curriculum for SER. This method is explained in more details in Section III-C3.

III. Methodology

This section discusses the motivation for using curriculum learning in SER problems (Sec. III-A). It introduces three formulation for SER used to evaluate the proposed curriculum policies (Sec. III-B). The section also describes the alternative policies for generating the difficulty measure from crowd-sourcing labels to build the curriculum (Sec. III-C).

A. Motivation

A candidate problem for curriculum learning should rely on non-convex optimization [24]. Emotion recognition is a good candidate because of its complex nature [50]. A person takes years to master the essential skills to recognize emotion [51]. Infants start with limited capabilities to recognize emotions, developing with time more sophisticated representations of the structure of emotions [52]. Due to the gradual increase of expertise in emotion recognition, psychologists measure the abilities to perceive affects as an important indicator of human emotional intelligence at different ages [53]. This step-by-step nature of the process of acquiring the capability to perceive emotions suggests that curriculum learning can be an effective method for training speech emotion classifiers.

The first step to establish a curriculum for learning emotions by machines is to quantify the difficulty of the training examples. If the utterances are not explicitly annotated with the difficulty of the emotional content conveyed on the sentences, the difficulty has to be indirectly estimated. In many tasks, the true labels exist (e.g., transcriptions in speech, presence or absence of an object in an image), so the difficulty can be set according to the proportion of raters providing a wrong answer. Such a ground-truth is not available in spontaneous emotional sentences, as the perception of emotion varies across listeners. During the annotation of categorical emotions, evaluators are commonly asked to choose the most relevant emotional classes for a given speech sample. However, some sentences may convey more than a single emotional class (e.g., frustration and anger) [39]. For attribute-based annotations, evaluators may also disagree when assigning absolute scores to the emotional content of the sentences. This work exploits different policies to build a curriculum for the application of SER. We compare this approach with the one-pass method where all the training samples are equally treated (uniform policy) without curriculum. The results confirm our hypothesis, giving statistically significant differences in the classification results.

B. Formulation of Machine Learning Problems

Before we introduce the proposed curriculum policies, it is important to define the machine-learning problems considered in this study, as their implementation are slightly different across problems. We define three machine-learning problems to evaluate the role of curriculum learning in SER: regression prediction of emotional attributes, binary classification of emotional attributes, and classification of emotional categories.

1) Regression of Dimensional Emotions: The first task is to predict the emotional dimension scores for the attributes arousal (calm versus active), valence (negative versus positive) and dominance (weak versus strong), using the training set (one regressor per emotional attribute). Each sentence is annotated by multiple annotators, where the gold-standard is the average across the annotations at the utterance level. We use the *concordance correlation coefficient* (CCC) metric to measure the performance of the regressors on the testing set. CCC is a metric of agreement between two interval variables by considering the Pearson correlation and difference between them [54]. We employ this metric to compare the true and predicted emotional attribute values. This metric has been previously used in related studies [55], including the *audio/visual emotion challenge* (AVEC) in 2015 [56].

2) Binary Classification of Dimensional Emotions: The second problem is a binary classification task where we split the samples into two classes based on their attribute values (e.g., low valence versus high valence). We use their median value such that the classes are balanced. For the test set, we use the same threshold derived from the training set so the classes are not necessarily balanced (Section IV-A introduces the actual partitions used in the evaluation). Although this formulation introduces artificial classes (see discussion on Mariooryad and Busso about dichotomizing continuous labels [57]), this approach is commonly used in studies on emotion recognition [58]–[60]. The performance of the classification problem is measured using the F1-score metric, which combines the mean precision and recall rates for low and high classes.

3) Classification of Categorical Emotions: The third problem corresponds to a classification task, where we recognize the categorical class assigned to each sample. The labels for training and testing the classifiers are assigned using the majority vote rule by considering all individual annotations provided during the evaluation (e.g., multiple evaluators are asked to annotate each sentence). Samples without agreement due to tie between two or more classes are not used for training or testing the system (i.e., the samples are removed from the corresponding sets). The classes are not balanced and each class is equally important for our problem. Therefore, we separately estimate the precision and recall rates for each of the classes. We use the average precision and recall rates

across emotional classes to estimate the F1-score. Therefore, this metric equally weights each class

C. Curriculum Policies for Emotional Speech

The task difficulty for each sample depends on the subjective evaluations conducted by multiple annotators and the machine-learning tasks. This section focuses on explaining the proposed curriculum policies for the formulations for SER described in Section III-B. We suggest three general criteria to design the curriculum for training the classifiers: Criterion 1 is a basic curriculum that relies on the results of a pre-trained model. Criterion 2 creates the curriculum by considering the disagreement between evaluators, assuming that all the evaluators have similar skills to complete the task. Criterion 3 also considers the disagreement between evaluators, taking into account the reliability of the annotators (e.g., annotators are not equally good to complete the task).

1) Criterion 1: Error of Predicted Label: The first criterion uses pre-trained models to determine the difficulty order in the curriculum. First, we train a classifier using all the samples available for training. The models are tested on the same training set. Then, we compare the predicted class with the actual label, defining specific rules to quantify the difficulty of sample i for each machine learning problem, denoted by d_i.

For regression problems, the difficulty of a sample is defined using Equation 1 by estimating the distance between the predicted value (y'_i) and the ground-truth (y_i). The easiest training samples have $d_i = 0$, where the pre-trained regression model successfully predicts the actual value without error.

$$d_i = |y_i - y'_i| \tag{1}$$

For binary (emotional attributes) and multi-class (categorical emotions) tasks, we define the difficulty of a sentence based on the classification results and the confidence of the classifier. For training sample i with true label y_i and feature vector $\mathbf{x_i}$, the pre-trained classifier predict the label y'_i. For the samples that are correctly classified (i.e., $y_i = y'_i$), the difficulty is defined as the confidence of the classifier on the predicted classes (with negative sign). For samples that are incorrectly classified (i.e., $y_i \neq y'_i$), we consider the confidence in the incorrect class as the difficulty metric (Eq. 2).

$$d_i = \begin{cases} -P(y_i = y'_i|\mathbf{x_i}), \text{if } y_i = y'_i \\ P(y_i = y'_i|\mathbf{x_i}), \text{if } y_i \neq y'_i \end{cases} \tag{2}$$

With this measure, an easy sample has a smaller d_i than a more difficult sample. The level of confidence P is inferred from the classifier's output. For a DNN with softmax function in the output layer, the maximal neuronal response of the softmax layer can be used as a measure of confidence [61].

The metric d_i for regression (Eq. 1) and classification (Eq. 2) considers the consensus label, which is generated by aggregating the answers from all the evaluators. The variations between individual evaluations have no effect on the difficulty measure, and, therefore, the curriculum.

2) Criterion 2: Disagreement Between Annotators: The second criterion relies on finding the level of disagreement between annotators for each sentence. Intuitively, annotators will have higher agreement on samples with clear emotional content (i.e., easy samples), and lower agreement for more emotionally ambiguous samples (i.e., difficult examples). For regression problem, a metric of disagreement across evaluators is the variance of the scores provided to a sample. Therefore, d_i is defined as:

$$d_i = \frac{1}{N_i}\sum_{j=1}^{N}(y_{ij} - \bar{y_i})^2 \tag{3}$$

where y_{ij} is the label assigned by annotator j for sample i, $\bar{y_i}$ is the average across all the annotations available for sample i, and N_i is the number of annotations available for sample i.

For categorical emotions, d_i quantifies how popular is the emotional class with more votes by finding the ratio between annotators selecting that class and the total number of annotators:

$$d_i = \frac{1}{N_i}\sum_{j=1}^{N}[y_{ij} = \hat{y_i}] \tag{4}$$

where $\hat{y_i}$ is the consensus label for sample i using the majority vote rule. In the binary problems for attribute-based annotations, the dichotomized labels are obtained by splitting the average scores using median split. We also use Equation 4 to estimate d_i.

3) Criterion 3: Minmax Conditional Entropy Inference: The third criterion for curriculum learning considers the disagreement across evaluators by modeling the level of expertise of the raters. The previous criterion assumes that all the annotators are equally reliable (Sec. III-C2). In reality, annotators have different level of expertise. For example, some raters are unfamiliar with the concept of emotional attributes, easily distracted, or inconsistent with their judgments. Therefore, we should consider the expertise of the raters to better determine whether the disagreement in the emotional labels is due to poor raters or the difficulty of the samples. A fair assessment process should jointly estimate the task difficulty and annotator's skills. The *item response theory* (IRT) [62] is a method to model the probability of a correct answer to a given item by a person with a specific ability level. It uses latent characterization of individuals and items as predictors of observed responses. This model relies on item discrimination, item difficulty, and the probability that an individual with very low ability correctly answers a question. Since the true labels of the items are not available in our case, the most likely correct response needs to be estimated in addition to the item difficulty and the ability of the workers.

To address this problem, Zhou et al. [27], [28] employed a minmax conditional entropy method subject to constraints to encode the observation for ordinal [27] and categorical [28] labels. In their formulation, the objective is to aggregate crowdsourced labels for a set of items annotated by a group of workers. The input is the observed label $\tilde{y}_{ij}$ which is the label selected for item i by annotator j. The goal is to

estimate the unobserved true labels y_i from the noisy workers' labels (i.e., estimating the probability of the item belonging to each class $Q(Y_i = c)$ given the set of observed labels $\tilde{y}_{ij}$). $P(\tilde{Y}_{ij} = k|Y_i = c)$ denotes the probability that worker j annotates the item i with label k while the true label is c. They proposed to jointly estimate the distributions of P and Q by minimizing the entropy of the observed workers' labels conditioned on the true labels.

$$\min_{Q} \max_{P} H(\tilde{Y}|Y) \tag{5}$$

The study solved this problem by converting this formulation into the dual form. They introduced two Lagrange multipliers $\sigma_j(c,k)$ and $\tau_i(c,k)$. Intuitively, $\sigma_j(c,k)$ is the measure of ability of worker j and $\tau_i(c,k)$ is the intrinsic difficulty of item i. The variable $[\tau_i]$ is a confusion matrix of item i, where its $(c,k) - th$ entry measures how likely item i in class c is labeled as class k by a randomly chosen worker [27], [28].

The minmax formulation by Zhou et al. [27], [28] provides a principled framework to estimate the difficulty of each sentence. We propose to use the difficulty measure $[\tau_i]$ to design the curriculum. The measure of difficulty d_i is estimated with the ratio between the trace of $[\tau_i]$ and the sum of all the elements in the matrix.

$$d_i = \frac{\sum_{k} \tau_i(k,k)}{\sum_{c} \sum_{k} \tau_i(c,k)} \tag{6}$$

The regularized minmax conditional entropy formulation proposed by Zhou et al. [27], [28] finds the item confusion matrix $[\tau_i]$ for both ordinal [27] and categorical [28] labels. After estimating the corresponding $[\tau_i]$, we find the difficulty metric using Equation 6 for regression, binary classification and multi-class classification problems.

IV. Experimental Evaluation

This section describes the experiments conducted to assess the performance of SER using the proposed curriculum learning schemes. This section introduces the database and the feature set used in the experiments (Sec. IV-A). The section also describes acoustic features (Sec. IV-B) and architecture of the classifier used in the evaluation (Sec. IV-C).

A. The MSP-Podcast Database

This study relies on the MSP-Podcast corpus collected at the University of Texas at Dallas [63]. The database includes a large set of speech segments from podcast recordings available in audio sharing websites. The podcasts are selected from various topics including politics, sports, talk shows, and movies, including a broad range of emotions. The podcasts are segmented into speech turns using a speaker diarization tool. We implement an automatic process that selects only speech segments with one speaker, without noise (SNR value less than 20dB), background music, or phone quality audio (4 KHz cut-off frequency). The duration of the obtained segments from the audio file are between 2.75s and 11s (including pauses). Following the ideas described in Mariooryad et al. [64], we use different machine learning formulations to retrieve segments from the pool of available segments conveying target emotion behaviors. Then, these segments are annotated with emotional labels. The data collection process is an ongoing effort, where the current study uses version 1.0 of the corpus. This set includes 20,045 speech segments (34 hrs, 15 min). We have manually annotated the speaker identity of 244 speakers (16,026 segments). We use segments from 50 speakers as our test set (6,069 segments), and data from 15 speakers as our development set (2,226 segments). We use the development set to optimize the hyper-parameters of the network and training process. The train set includes the rest of the corpus (11,750 segments). This data partition attempts to create speaker independent datasets for train, test, and development sets.

The speech segments are annotated with emotional labels using an improved version of the crowdsourcing method introduced by Burmania et al. [38]. Within the perceptual evaluation, the raters are asked to choose emotional attributes (arousal, valence and dominance) using a seven-point Likert scale. Then, they select the primary emotion from anger, sadness, happiness, surprised, fear, disgust, contempt, and neutral. They can also choose other if none of the previous labels are suitable. For attribute-based annotations, we measure the inter-evaluator agreement using the Krippendorffs alpha coefficient since the labels are interval and not categorical. For arousal, valence and dominance, the inter-evaluator agreements are $\alpha_{aro} = 0.431$, $\alpha_{val} = 0.447$ and $\alpha_{dom} = 0.391$, respectively. For the primary emotions, the inter-evaluator agreement is measured using the Fleiss kappa, which is $\kappa = 0.218$. The current version of the corpus is not balanced across emotional classes, with many happy and neutral sentences and very few fear sentences. Therefore, we remove the label *fear* from the target emotion classes and consider it as if the raters had selected *other* as the perceived emotion. While not used in this study, the annotation also includes secondary emotions where annotators choose all the emotions relevant to the speech segment from the list provided in the questionnaire. The list of secondary emotions includes all the primary emotions, in addition to amused, frustrated, depressed, concerned, disappointed, excited, confused, and annoyed. Each speech segment is annotated by at least five annotators.

B. Features

All the emotion recognition problems are implemented with the feature set proposed for the *computational paralinguistics challenge* (ComParE) at Interspeech 2013 [65]. The feature set includes *low-level descriptors* (LLDs) such as fundamental frequency and *Mel-frequency cepstral coefficients* (MFCCs). These frame-based features are used to estimate statistics over a speech segment (e.g., mean of the fundamental frequency). The approach creates a 6,373 dimensional feature vector per speech segment, regardless of its length. The features are extracted with the open-source OpenSMILE toolkit [66]. The work by Eyben [67] provides more information about this toolkit and ComParE feature set.

C. Implementation of Machine-Learning Frameworks

The evaluation of the curriculum learning uses a DNN with the same architecture across the three machine-learning problems considered in this study (Sec. III-B). The only difference is the output layer, which is implemented according to the problem. The architecture is a fully connected feed forward neural network with two hidden layers, each of them implemented with 1,024 nodes. In our previous work, we have studied different configurations for deep models for SER problems, increasing the number of layers or number of nodes per layer [68]. The model with two layers with 1,024 nodes provided very competitive results. The activation function corresponds to a *rectified linear unit* (ReLU). The input corresponds to the 6,373 dimensional segment-label feature vector described in Section IV-B. The output layer is added on top of the second hidden layer. The output layer is the identity activation function for the regression problems, and the softmax layer for binary and multi-class classification tasks. The regression network is trained by maximizing the CCC. The binary and multi-class networks are trained to minimize the cross-entropy cost function. The softmax layer gives a vector with a score for each emotional class. The evaluation uses Keras with TensorFlow as backend to implement and train the models. We rely on *adaptive moment estimation* (ADAM) [69] for the optimization of the parameters of the network. We use mini-batches of size 256.

Based on the measure of difficulty, we divide the training set into five bins, where the first bin contains the easiest samples. The training process starts with the easiest bin and continue by adding more difficult bins to the training set. After adding each bin to the training set, the network is trained until the performance on the development set stops increasing for three consecutive epochs. We find the optimal learning rate for each bin by maximizing the performance on the development set. In our search, we consider the following values for the learning rates: 1×10^{-1}, 5×10^{-2}, 1×10^{-2}, 5×10^{-3}, 1×10^{-3}, 5×10^{-4},1×10^{-4}, 5×10^{-5}, 1×10^{-5}, 5×10^{-6}, and 1×10^{-6}. We start with bin 1, finding the optimal learning rate by considering only the performance using the training samples on bin 1. The learning rate for bin 1 does not change during the rest of the search. Next, we add the training samples of bin 2, finding the optimal learning rate for bin 1 and 2. This process is repeated until we find the optimal learning rates for all the five sets. Although this search does not guarantee an optimal solution, we choose this method since it can be efficiently implemented. We separately repeat this process for every machine learning problem, including every curriculum considered in this paper. In most cases, we observe that the optimal combination of learning rates with this approach monotonically decreases from the first set (only bin 1) to the last training set (all bins). For example for the categorical emotion classification task, with Criterion 3, the optimal learning rates are: 1×10^{-3} (bin 1), 1×10^{-3} (bins 1-2), 5×10^{-4} (bins 1-3), 1×10^{-4} (bins 1-4), and 5×10^{-5} (bins 1-5).

We consider two baseline frameworks to evaluate the proposed curriculum learning approaches. The first approach does not consider any curriculum, training the models with all the data. The learning rate and number of training epochs for this framework is independently optimized for each problem using the development set. The second baseline uses a curriculum where the bins are created at random. This baseline is also implemented by optimizing its learning rates for this task on the development set.

TABLE I
RESULTS OF REGRESSION MODELS FOR AROUSAL (*Aro*), VALENCE (*Val*), AND DOMINANCE (*Dom*).THE ASTERISK [*] AND CIRCLE [°] INDICATE THE APPROACH OUTPERFORMS THE BASELINES *w/o curriculum* AND *with random curriculum*, RESPECTIVELY. WE ASSERT SIGNIFICANCE AT p-VALUE≤ 0.05.

	Aro. [CCC]	Val. [CCC]	Dom. [CCC]
w/o curriculum	0.724	0.298	0.690
With random curriculum	0.729	0.293	0.686
Criterion 1-Error of predicted label	0.725	0.313°	0.694
Criterion 2-Disagreement between annotators	0.730*	0.320*°	0.696°
Criterion 3-Minmax entropy	0.745*°	0.325*°	0.705*°

V. RESULTS

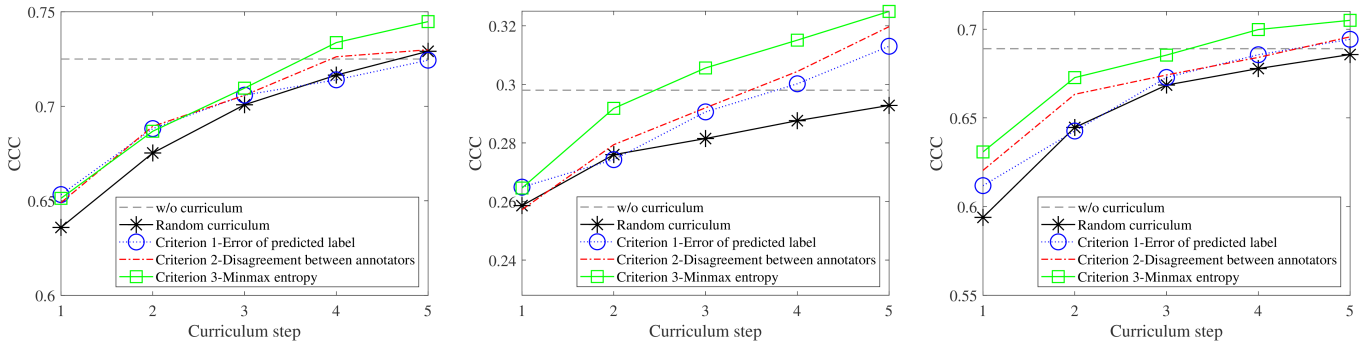
This section describes the results obtained on the three machine-learning problems for emotion recognition (regression, binary classification, and multi-class classification). We train the DNNs over 10 trials, using different random initializations. We report the average results, evaluating the performance metric using the one-tail, population mean t-test over the 10 trials. We assert statistical significance at p-value$\leq$ 0.05. We denote with an asterisk (*) when a model trained with curriculum learning is statistically better than the baseline model trained without a curriculum, and with a circle (°) when a model trained with curriculum learning is statistically better than the baseline trained with a random curriculum (see Tables I, II and III).

A. Regression of Emotional Attributes

The first experimental evaluation demonstrates the role of curriculum learning on predicting emotional attributes with regression models. The performance is computed with CCC.

Table I shows the average CCC values for arousal, valence and dominance across the 10 trials. The results show that the condition with the highest CCC values corresponds to the regression models trained with curriculum learning using the criterion 3 (i.e., minmax entropy). The results are statistically better than both baseline methods. The table also shows that randomly selecting the training bins does not lead to significant improvements. Criterion 1 is less effective than criteria 2 and 3. These results demonstrate the importance of quantifying the disagreement between evaluators to assess the difficulty of the samples. We can effectively achieve this goal by considering individual evaluations assigned to the samples.

Figure 1 shows the CCC values for the regression models for emotional attributes by following different curriculum policies. The figure shows that the performance increases as we effectively add more difficult samples in the training set. The criterion 3 based on the minmax entropy framework



(a) Arousal (b) Valence (c) Dominance

Fig. 1. Intermediate results for regression models. The results are obtained on the test set at different steps of the training process using curriculum learning. More difficult samples are added at each step. The dashed lines indicate the performance of regression models trained without curriculum learning.

TABLE II
RESULTS OF BINARY CLASSIFICATION FOR AROUSAL (*Aro*), VALENCE (*Val*), AND DOMINANCE (*Dom*).THE ASTERISK [*] AND CIRCLE [°] INDICATE THE APPROACH OUTPERFORMS THE BASELINES *w/o curriculum* AND *with random curriculum*, RESPECTIVELY. WE ASSERT SIGNIFICANCE AT p-VALUE≤ 0.05.

	Aro. [F1-score]	Val. [F1-score]	Dom. [F1-score]
w/o curriculum	0.778	0.592	0.685
With random curriculum	0.771	0.591	0.685
Criterion 1-Error of predicted label	0.785	0.606*°	0.684
Criterion 2-Disagreement between annotators	0.789*°	0.616*°	0.695*°
Criterion 3-Minmax entropy	0.791*°	0.616*°	0.696*°

achieves the highest performance for arousal, valence, and dominance. The performance for this approach is consistently better than a system trained without curriculum learning.

B. Binary Classification of Emotional Attributes

We also evaluate the role of curriculum learning on binary classification of emotional attributes (e.g., low arousal versus high arousal). The test set is divided into two classes for high and low values of a given attribute based on the median split obtained on the training set. This method makes the test set almost balanced.

Table II lists the average F1-score for arousal, valence, and dominance across the 10 trails. The table demonstrates that using curriculum learning with criteria 2 and 3 achieves statistically significant improvements over a model trained without curriculum learning or with randomly selected bins. However, the best performance is also obtained with criterion 3, which uses the policy based on the minmax entropy framework. Criterion 1 is only effective for valence, showing that quantifying the difficulty of the samples based on pre-trained models is not the best approach for SER.

Figure 2 shows the F1-score values for attribute based binary classification tasks following the three different curriculum policies. It also shows the performance of a model trained without a curriculum and with a curriculum using randomly selected bins. Similar to the results with regression problems, criterion 3 achieves the best performance for arousal, valence, and dominance, obtaining important improvements over the baselines.

TABLE III
PERFORMANCE OF CATEGORICAL EMOTION CLASSIFICATION. THE ASTERISK [*] AND CIRCLE [°] INDICATE THE APPROACH OUTPERFORMS THE BASELINES *w/o curriculum* AND *with random curriculum*, RESPECTIVELY. WE ASSERT SIGNIFICANCE AT p-VALUE≤ 0.05.

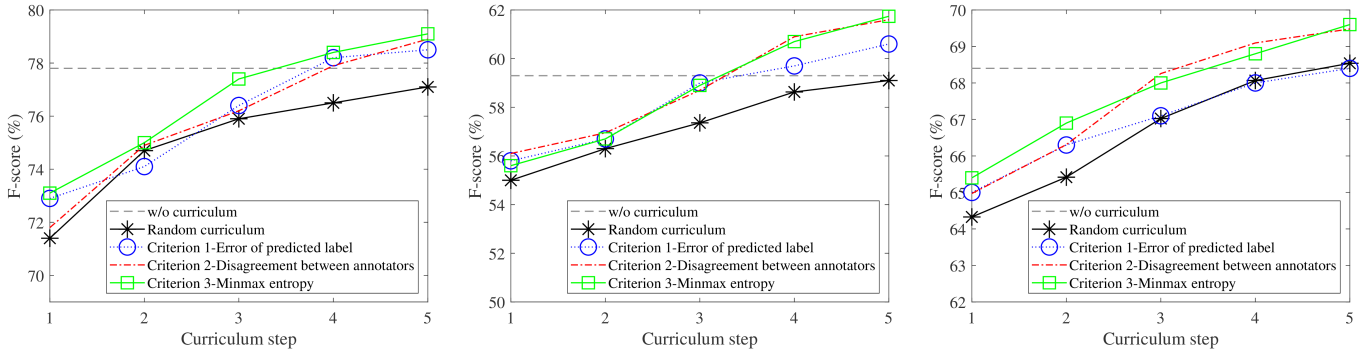
	F1-score [%]
w/o curriculum	39.7
With random curriculum	39.8
Criterion 1-Error of predicted label	40.8
Criterion 2-Disagreement between annotators	41.5*°
Criterion 3-Minmax entropy	42.1*°

C. Multi-class Categorical Emotion Classification

The third problem involves the classification of categorical emotions. The evaluation consider the following five classes: happiness, anger, sadness, disgust, and neutral state. The ground truth labels for the test set are generated by finding the majority vote between all the annotators. Samples that do not reach agreement with this rule are discarded from this evaluation.

Table III shows the average F1-score of the five class categorical emotion classification across the 10 trails. The results are consistent with the findings presented for regression and binary classification problems. Curriculum learning using policies that quantify the agreement between evaluators (criteria 2 and 3) provides statistically significant improvement over the baseline methods. For criterion 3, the use of curriculum learning leads to an absolute improvement of 2.4% over a model trained without curriculum learning. This gain corresponds to a 6% relative improvement over the baseline, which is obtained by just changing the order in which the samples are introduced during the training process.

Figure 3 shows the F1-score curve for the classification of categorical emotions by following different curriculum policies. It is clear that criterion 3 based on the minmax entropy framework also achieves the best performance for this task. The use of criteria 2 and 3 leads to statistically significant improvements over the baselines. We also observe that for criterion 3, the last bin does not improve the performance, suggesting that the aggregation and learning method used in this study cannot benefit from the information obtained from these more challenging sentences.



(a) Arousal (b) Valence (c) Dominance

Fig. 2. Intermediate results for binary classification. The results are obtained on the test set at different steps of the training process using curriculum learning. More difficult samples are added at each step. The dashed lines indicate the performance of binary classifiers trained without curriculum learning.

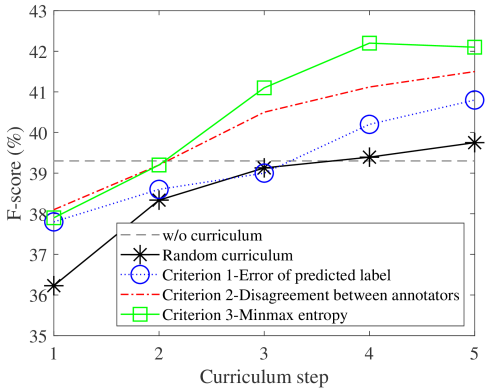


Fig. 3. Intermediate results for categorical emotions. The results are obtained on the test set at different steps of the training process using curriculum learning. More difficult samples are added at each step. The dashed line indicates the performance of a classifier trained without curriculum learning.

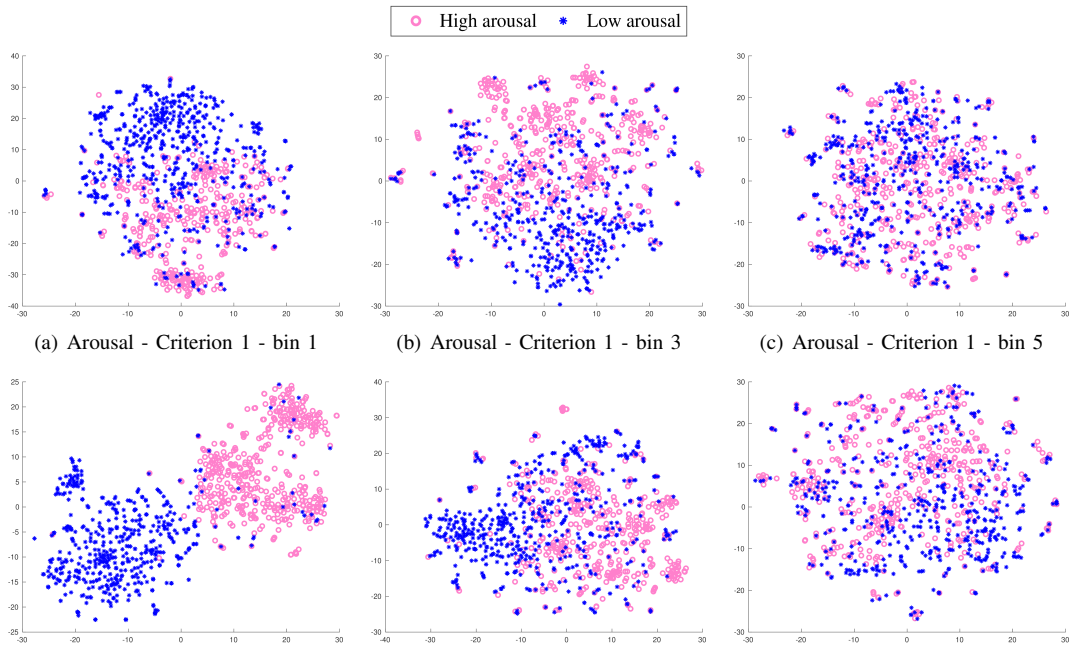
D. Analysis of Feature Representation

This section visualizes how the difficulty measures used to build the proposed curriculum is reflected on the features. This analysis is applicable to classification problems, where we only consider binary classification of emotional attributes. The analysis relies on the t-SNE method proposed by Van Der Maaten and Hinton [70], which is a useful tool to visualize high dimensional data. This toolkit is used to reduce the dimension of the features from 6,373 to 2. We expect that the difficulty measure used to build the curriculum is reflected in the feature representation, where the different classes in the first bin (i.e., the easiest samples) are better separated than the classes in other bins (i.e., more difficult samples). This feature visualization provides another venue to evaluate the proposed criteria for curriculum learning. The analysis considers criterion 1 (error of predicted error) and criterion 3 (minmax entropy), which provided the worse and best policies for curriculum learning in previous sections.

We report the results for arousal in the binary classification problem. Figure 4 shows the feature representation for arousal, where each point is associated with one sentence. The color of the points represents the category of the sentence (low arousal versus high arousal). Figures 4(a), 4(b), and 4(c) show the results when using criterion 1 for bin 1 (easy), bin 3 (medium difficulty) and bin 5 (hard), respectively. The classes for the easy samples have a clear separation. Figure 4(c) shows that the overlap between classes increases in bin 5, which has the most difficult examples. The correlation between task difficulty and ambiguity in the feature domain was expected since the difficulty is derived from the prediction of pre-trained classifiers, which directly rely on the discrimination of the features. This connection is not assumed in the minmax entropy framework (criterion 3), which relies on the disagreement between evaluators. Figures 4(d), 4(e), and 4(f) show the results using criterion 3 for bin 1 (easy), bin 3 (medium difficulty) and bin 5 (hard), respectively. It is interesting to observe similar trends, even when the difficulty metric is not derived from pre-trained classifiers.

VI. Conclusions

This study proposed the use of curriculum learning for speech emotion recognition. Since the policies to determine the difficulty of emotional sentences are not straightforward, as in other problems, the study explored different methods to design the curriculum. The study proposed to quantify the difficulty level of the sentences by relying on results from pre-trained models on the training set, or by considering inter-evaluator agreement metrics under the assumption that ambiguous sentences for human are also ambiguous for machines. The experimental evaluation considered three formulations of SER: regression of emotional attributes, binary classification of emotional attributes, and classification of emotional categories. The results demonstrated the benefits of using curriculum learning in SER, showing consistent improvements over baselines trained with randomly selected bins or without curriculum learning. The most successful policy for building the curriculum considers the agreement level of the annotations by estimating the expertise of the labelers. This approach relies on the minmax entropy framework that learns a latent variable describing the difficulty of the speech samples evaluated by the raters. The machine-learning models trained with this curriculum achieved significant improvements over the baseline methods.



(a) Arousal - Criterion 1 - bin 1 (b) Arousal - Criterion 1 - bin 3 (c) Arousal - Criterion 1 - bin 5

(d) Arousal - Criterion 3 - bin 1 (e) Arousal - Criterion 3 - bin 3 (f) Arousal - Criterion 3 - bin 5

Fig. 4. Visualization using the t-SNE toolkit of acoustic features for arousal in binary classification tasks for bin 1 (easy), bin 2 (medium difficulty), and bin 3 (hard). The top figures correspond to criterion 1 (i.e., error of predicted labels) and the bottom figures correspond to criterion 3 (i.e., minmax entropy).

As a future direction of the curriculum learning framework proposed in this study, we will use the difficulty measure to find training examples that negatively affect the performance of the models. Abdelwahab and Busso [71] showed that selecting a subset of the data for supervised adaptation of speech emotional models led to improvements over results obtained when the entire adaptation set was used. By removing these samples, we expect to increase the classification performance, since these examples can be too difficult to learn due to unreliable or incorrect labels. Likewise, there are important parameters that we have not investigated, including the optimum number of difficulty bins, and the optimum number of epochs. Adjusting these parameters may lead to further improvements in classification performance. Finally, we will explore whether curriculum learning is still effective as the size of the training set increases. The collection of the MSP-Podcast is an ongoing effort in our laboratory, which will allow us to evaluate the approach in the future with a larger training set.

ACKNOWLEDGMENT

This study was funded by the National Science Foundation (NSF) CAREER grant IIS-1453781.

REFERENCES

[1] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction," in *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW 2003)*, Madison, WI, USA, June 2003, pp. 1–6.

[2] M. Szwoch and W. Szwoch, "Emotion recognition for affect aware video games," in *Image Processing & Communications Challenges 6*, ser. Advances in Intelligent Systems and Computing, R. Choraś, Ed. Cham: Springer International Publishing, 2015, vol. 313, pp. 227–236.

[3] M. Obaid, C. Han, and M. Billinghurst, ""Feed the fish": an affect-aware game," in *Australasian Conference on Interactive Entertainment*, Brisbane, Australia, December 2008.

[4] D. Litman and K. Forbes-Riley, "Predicting student emotions in computer-human tutoring dialogues," in *ACM Association for Computational Linguistics (ACL 2004)*, Barcelona, Spain, July 2004, pp. 1–8.

[5] D. Litman and K. Forbes-Riley, "Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors," *Speech communication*, vol. 48, no. 5, pp. 559–590, May 2006.

[6] B. Abirached, Y. Zhang, and J.-H. Park, "Understanding user needs for serious games for teaching children with autism spectrum disorders emotions," in *World Conference on Educational Media and Technology (EdMedia 2012)*, Denver, CO, USA, June 2012, pp. 1054–1063.

[7] 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.

[8] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, July-Dec 2010.

[9] Z. Zhang, F. Weninger, M. Wollmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2011)*, Waikoloa, HI, USA, December 2011, pp. 523–528.

[10] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, December 2018.

[11] T. Vogt and E. André, "Improving automatic emotion recognition from speech via gender differentiation," in *International conference on Language Resources and Evaluation (LREC 2006)*, Genoa,Italy, May 2006, pp. 1123–1126.

[12] B. Schuller, R. Müller, M. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *9th European Conference on Speech Communication and Technology (Interspeech 2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 805–808.

[13] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative feature normalization scheme for automatic emotion detection from speech," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 386–397, October-December 2013.

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

[15] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 3593–3597.

[16] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, April 2003.

[17] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, Amsterdam, The Netherlands, September 2009, pp. 1–8.

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

[19] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," in *ACM Association for Computational Linguistics (ACL 2010)*, Uppsala, Sweden, July 2010, pp. 384–394.

[20] K. Krueger and P. Dayan, "Flexible shaping: How learning in small steps helps," *Cognition*, vol. 110, no. 3, pp. 380–394, March 2009.

[21] T. D. Sanger, "Neural network learning control of robot manipulators using gradually increasing task difficulty," *IEEE Transactions on Robotics and Automation*, vol. 10, no. 3, pp. 323–333, June 1994.

[22] A. Baranes and P. Oudeyer, "Active learning of inverse models with intrinsically motivated goal exploration in robots," *Robotics and Autonomous Systems*, vol. 61, no. 1, pp. 49–73, January 2013.

[23] W. Zaremba and I. Sutskever, "Learning to execute," *ArXiv e-prints (arXiv:1410.4615)*, October 2014.

[24] Y. Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *International Conference on Machine Learning (ICML 2009)*, Montreal, QC, Canada, June 2009, pp. 41–48.

[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 Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.

[26] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *ACM international conference on Multimedia (MM 2017)*, Mountain View, CA, USA, October 2017, pp. 890–897.

[27] D. Zhou, Q. Liu, J. Platt, and C. Meek, "Aggregating ordinal labels from crowds by minimax conditional entropy," in *International Conference on Machine Learning (ICML 2014)*, Beijing, China, June 2014, pp. 262–270.

[28] D. Zhou, Q. Liu, J. Platt, C. Meek, and N. Shah, "Regularized minimax conditional entropy for crowdsourcing," *ArXiv e-prints (arXiv:1503.07240)*, vol. abs/1503.07240, pp. 1–31, March 2015.

[29] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Proceedings of the Artificial Neural Networks in Engineering (ANNIE 1999)*, St. Louis, MO, November 1999.

[30] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *International Conference on Spoken Language (ICSLP 1996)*, vol. 3, Philadelphia, PA, USA, October 1996, pp. 1970–1973.

[31] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Sixth International Conference on Multimodal Interfaces ICMI 2004*. State College, PA: ACM Press, October 2004, pp. 205–211.

[32] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "Desperately seeking emotions or: actors, wizards and human beings," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, September 2000, pp. 195–200.

[33] 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.

[34] 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.

[35] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Interspeech - International Conference on Spoken Language (ICSLP)*, Pittsburgh, PA, USA, September 2006, pp. 801–804.

[36] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *IEEE International Conference on Multimedia and Expo (ICME 2008)*, Hannover, Germany, June 2008, pp. 865–868.

[37] 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.

[38] 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.

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

[40] H. Fayek, M. Lech, and L. Cavedon, "On the correlation and transferability of features between automatic speech recognition and speech emotion recognition," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 3618–3622.

[41] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, November-December 2011.

[42] ——, "Emotion recognition using a hierarchical binary decision tree approach," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 320–323.

[43] A. Graves, M. Bellemare, J. Menick, R. Munos, and K. Kavukcuoglu, "Automated curriculum learning for neural networks," in *International Conference on Machine Learning (ICML 2017)*, Sydney, Australia, August 2017, pp. 1–10.

[44] L. Gui, T. Baltrušaitis, and L. Morency, "Curriculum learning for facial expression recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*, Washington, DC, USA, May-June 2017, pp. 505–511.

[45] S. Braun, D. Neil, and S. Liu, "A curriculum learning method for improved noise robustness in automatic speech recognition," in *European Signal Processing Conference (EUSIPCO 2017)*, Kos island, Greece, August-September 2017, pp. 548–552.

[46] S. Ranjan and J. Hansen, "Curriculum learning based approaches for noise robust speaker recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 1, pp. 197–210, January 2018.

[47] Z. Zhang, J. Han, E. Coutinho, and B. W. Schuller, "Dynamic difficulty awareness training for continuous emotion prediction," *IEEE Transactions on Multimedia*, 2018.

[48] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, ""Of all things the measure is man" automatic classification of emotions and inter-labeler consistency," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, vol. 1, Philadelphia, PA, USA, March 2005, pp. 317–320.

[49] A. Dawid and A. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28, 1979.

[50] S. Marsella, J. Gratch, and P. Petta, "Computational models of emotion," in *A Blueprint for Affective Computing-A sourcebook and manual*, K. Scherer, T. Bänziger, and E. Roesch, Eds. New York, NY: Oxford University Press, November 2010, pp. 21–46.

[51] M. Zeidner, G.Matthews, R. Roberts, and C. MacCann, "Development of emotional intelligence: Towards a multi-level investment model," *Human development*, vol. 46, no. 2-3, pp. 69–96, March-June 2003.

[52] B. Volling, N. McElwain, P. Notaro, and C. Herrera, "Parents' emotional availability and infant emotional competence: Predictors of parent-infant attachment and emerging self-regulation," *Journal of Family Psychology*, vol. 16, no. 4, pp. 447–465, 2002.

[53] J. Mayer, D. Caruso, and P. Salovey, "Emotional intelligence meets traditional standards for an intelligence," *Intelligence*, vol. 27, no. 4, pp. 267–298, December 1999.

[54] F. Weninger, F. Ringeval, E. Marchi, and B. Schuller, "Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio," in *International Joint Conference on Artificial Intelligence (IJCAI 2016*, New York, NY, USA, July 2016, pp. 2196–2202.

[55] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional

emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, no. 15, pp. 22–30, November 2015.

[56] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AV$^+$EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *International Workshop on Audio/Visual Emotion Challenge (AVEC 2015)*, Brisbane, Australia, October 2015, pp. 3–8.

[57] S. Mariooryad and C. Busso, "The cost of dichotomizing continuous labels for binary classification problems: Deriving a Bayesian-optimal classifier," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 119–130, January-March 2017.

[58] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011- the first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction (ACII 2011)*, ser. Lecture Notes in Computer Science, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Memphis, TN, USA: Springer Berlin / Heidelberg, October 2011, vol. 6975/2011, pp. 415–424.

[59] M. Wöllmer, F. Eyben, B. Schuller, E. Douglas-Cowie, and R. Cowie, "Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 1595–1598.

[60] T. Rahman and C. Busso, "A personalized emotion recognition system using an unsupervised feature adaptation scheme," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012, pp. 5117–5120.

[61] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *In Advances in Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, December 2017, pp. 4878–4887.

[62] F. Lord, *Applications of item response theory to practical testing problems*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, July 1980.

[63] 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. To appear, 2018.

[64] 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.

[65] 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.

[66] 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.

[67] F. Eyben, *Real-time speech and music classification by large audio feature space extraction*, ser. Springer Theses. Springer, June 2017.

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

[69] 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.

[70] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, November 2008.

[71] M. Abdelwahab and C. Busso, "Incremental adaptation using active learning for acoustic emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5160–5164.

Reza Lotfian (SM'17) received his BS degree (2006) with high honors in Electrical Engineering from the Department of Electrical Engineering, Amirkabir University, Tehran, Iran, the MS degree (2010) in Electrical Engineering from the Sharif University (SUT), Tehran, Iran, and the PhD degree (2018) in electrical engineering from the University of Texas at Dallas (UTD). He is currently a research scientist at Cogito Corp at Boston, Massachusetts, USA. His research interest includes the area of speech signal processing, affective computing, human machine interaction, and machine learning.

Carlos Busso (S'02-M'09-SM'13) 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 graduated in 2003 across 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 is the co-author of the winner 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 system, and machine learning methods for multimodal processing. His work has direct implication in many practical domains, including national security, health care, entertainment, transportation systems, and education. He was the general chair of ACII 2017. He is a member of ISCA, AAAC, and ACM, and a senior member of the IEEE.