Semi-Supervised Speech Emotion Recognition with Ladder Networks

Srinivas Parthasarathy, *Student Member, IEEE,* Carlos Busso, *Senior Member, IEEE,*

***Abstract*—*Speech emotion recognition* (SER) systems find applications in various fields such as healthcare, education, and security and defense. A major drawback of these systems is their lack of generalization across different conditions. For example, systems that show superior performance on certain databases show poor performance when tested on other corpora. This problem can be solved by training models on large amounts of labeled data from the target domain, which is expensive and time-consuming. Another approach is to increase the generalization of the models. An effective way to achieve this goal is by regularizing the models through *multitask learning* (MTL), where auxiliary tasks are learned along with the primary task. These methods often require the use of labeled data which is computationally expensive to collect for emotion recognition (gender, speaker identity, age or other emotional descriptors). This study proposes the use of ladder networks for emotion recognition, which utilizes an unsupervised auxiliary task. The primary task is a regression problem to predict emotional attributes. The auxiliary task is the reconstruction of intermediate feature representations using a denoising autoencoder. This auxiliary task does not require labels so it is possible to train the framework in a semi-supervised fashion with abundant unlabeled data from the target domain. This study shows that the proposed approach creates a powerful framework for SER, achieving superior performance than fully supervised *single-task learning* (STL) and MTL baselines. We implement the approach with sentence-level or frame-level features, demonstrating the flexibility of our approach. Additionally, the generalization of the ladder networks is evaluated in cross-corpus settings using sentence-level features, obtaining important improvements. Compared to the STL baselines, the proposed approach achieves relative gains in *concordance correlation coefficient* (CCC) between 3.0% and 3.5% for within corpus evaluations, and between 16.1% and 74.1% for cross corpus evaluations, highlighting the power of the architecture.**

***Index Terms*—Semi-supervised emotion recognition, ladder networks, speech emotion recognition.**

I. Introduction

Recognizing emotions is a key feature needed to build socially aware systems. Therefore, it is an important part of *human computer interaction* (HCI). Emotion recognition can play an important role in various fields such as healthcare (mood profiles) [1], education (tutoring) [2] and security and defense (surveillance) [3]. *Speech emotion recognition* (SER) have enormous potential given the ubiquity of speech-based devices. However, it is important that SER models generalize well across different conditions and settings showing robust performance.

Conventionally, emotion recognition systems are trained with supervised learning solutions. The generalization of the models is often emphasized by training on a variety of samples with diverse labels [4]. The state-of-the-art models for standard computer vision tasks utilize thousands of labeled samples. Similarly, *automatic speech recognition* (ASR) systems are trained on several hundred hours of data with transcriptions. Generally, labels for emotion recognition tasks are collected with perceptual evaluations from multiple evaluators. The raters annotate samples by listening or watching to the stimulus. This evaluation procedure is cognitively intense and expensive [5]. Therefore, standard benchmark datasets for SER have limited number of sentences with emotional labels, often collected from a limited number of evaluators. This limitation severely affects the generalization of the systems.

An alternative approach to increase the generalization of the models is by building robust models. An effective approach to achieve this goal is with *multitask learning* (MTL) [6], where relevant auxiliary tasks are simultaneously solved along with the primary task. By solving relevant auxiliary tasks, the models are regularized by finding more general high-level feature representations that are still discriminative for the primary task. Multitask learning has been successfully used for emotion recognition tasks [7]–[10]. While these MTL methods have achieved promising results, most of the proposed solutions have focused on MTL problems that utilize supervised auxiliary tasks. Examples include gender recognition [10], [11], speaker information [10], other emotional attributes [8], [9] and secondary emotions [7]. This approach requires the use of meta labels which further limits the training of the models. In many scenarios, it is possible to collect large amount of data without labels from the target domain. It is important to build models that can effectively utilize unsupervised auxiliary tasks to regularize the model, leveraging these unlabeled recordings. This study explores this idea with ladder networks, building upon our previous work [12]. The ladder network architecture is a framework that combines supervised tasks with unsupervised auxiliary tasks. These auxiliary tasks correspond to the reconstruction of feature representations at various layers in a *deep neural network* (DNN). Since, this reconstruction is completely unsupervised, this framework has clear advantages: (a) it improves the regularization of the model through auxiliary tasks without the need for extra labels, and (b) it increases the generalization of the model by utilizing unlabeled data from the target domain. The uniqueness of the framework is the skip connections between the corresponding

S. Parthasarathy was with the Erik Jonsson School of Engineering & Computer Science, The University of Texas at Dallas, TX 75080 USA (e-mail: sxp120931@utdallas.edu).

C. Busso is with the Erik Jonsson School of Engineering & Computer Science, The University of Texas at Dallas, TX 75080 USA (e-mail: busso@utdallas.edu).

Manuscript received May 7, 2019; revised xxxx xx, xxxx.

encoder and decoder layers, which reduce the load on the encoder layers to carry information for decoding the layers. With this approach, the higher layers of the encoder learn discriminative representations for the supervised task.

This study provides a comprehensive analysis of auxiliary tasks for speech emotion recognition on the MSP-Podcast corpus [13]. The study focuses on regression problems, where the primary task is to predict the arousal, valence and dominance scores. The proposed implementation uses *high-level descriptors* (HLDs), computed at the speaking turn-level as feature inputs. The evaluation compares the performance of the proposed ladder network framework for emotion recognition against *single-task learning* (STL) and MTL baselines. The experimental results show the benefits of the framework, obtaining state-of-the-art performance on this speech emotional corpus. This study also examines the performance of the proposed architectures for cross-corpus experiments, where the models are trained on the MSP-Podcast corpus and tested on other popular databases for SER tasks (USC-IEMOCAP and MSP-IMPROV corpora). The proposed architectures achieve significant improvements in the cross-corpus experiments, leading to models that generalize better to unseen conditions. Finally, this study replicates the proposed architecture for two low-level feature inputs: (a) dynamic *low-level descriptors* (LLDs), and (b) *Mel-frequency band* (MFB) energies. The model with ladder networks achieve significant improvements over the baselines in most cases.

With respect to previous studies, including our own work [12], this study includes the following contributions:

- We train ladder networks in a semi-supervised fashion, where the reconstruction of intermediate layers and the prediction of emotional attributes are jointly optimized. We utilize unlabeled data for the reconstruction loss. This formulation extends our previous work that trained ladder networks exclusively on fully labeled data [12].
- We demonstrate that the proposed ladder network architecture can be trained with features extracted at the sentence-level (high-level descriptors), or at the frame-level (low-level descriptors), facilitating end-to-end training of the models.
- We provide a comprehensive analysis of training ladder networks for speech emotion recognition, showing its capability in within corpus evaluations, and cross-corpus evaluations, where we observe significant performance gains.

The rest of the paper is organized as follows. Section II reviews studies on research areas that are relevant to this work. Section III presents the proposed architecture that exploits unsupervised auxiliary tasks to regularize the network. Section IV gives details on the experimental setup including the databases and features used in this study. Section V presents the exhaustive experimental evaluations, showing the benefits of the proposed architecture. Finally, Section VI provides the concluding remarks, discussing potential areas of improvements.

II. Background

This study uses the emotional attributes arousal, valence and dominance to describe emotions. SER systems for these problems are often built to recognize individual emotional attributes. Most frameworks are trained in a supervised fashion with labeled data. Given the limited size of most speech emotional databases, these supervised emotion recognition frameworks are commonly trained with a few hours of labeled data. Using unlabeled data is an interesting method to increase the generalization of the models to a new domain.

A. Semi-Supervised Learning

Previous studies for semi-supervised learning have considered the *inductive learning* technique, where a classifier is first trained on the labeled samples. The trained classifier is then used on the unlabeled set to obtain predictions. The training set is then augmented with samples having highly confident predictions. The classifier is retrained with this augmented training set. This process is iterated a fixed number of times after which the performance often saturates. Zhang et al. [14] used this inductive learning procedure for SER to leverage unlabeled data. They enhanced their supervised learning approach with this method, obtaining better predictions on labeled data [15], [16]. Cohen et al. [17], [18] proposed a similar strategy for facial expressions using probabilistic Bayesian classifiers.

Another approach for semi-supervised learning is the *co-training* or *multi-view learning* procedure [19]. In this method, the classifiers are trained on distinct feature partitions (views). The different classifiers are used for predictions on the unlabeled data, augmenting the training set with samples that are consistently recognized by the classifiers. Mahdhaoui and Chetouani [20] proposed multi-view training for SER using different sets of acoustic features. Similarly, Zhang et al. [21] utilized co-training along with *active learning* where they only annotated emotional labels for samples where the predictions were made with low confidence by the multi-view classifiers. Liu et al. [22] proposed multi-view learning for SER, where they used temporal and statistical acoustic features. Studies have also considered multi-view learning by incorporating multiple modalities [11], [15], [16].

This study is more closely related to the recent advances in deep learning that combine supervised and unsupervised learning. Similar to our work, Deng et al. [23] proposed combining an autoencoder and a classifier for SER. Their framework is based on a discriminative *Restricted Boltzmann Machine* (RBM), which considers unlabeled samples as an extra *garbage* class in the classification problem. Huang et al. [24] proposed learning affect sensitive features using a semi-supervised implementation of a *convolutional neural network* (CNN) for SER. In this study, general features are learned using an unsupervised CNN architecture, and then these features are fine-tuned for affect recognition. Similarly, Mao et al. [25] trained a CNN to learn salient features for SER. The CNNs were first trained on unlabeled samples using a sparse autoencoder and a reconstruction penalization. The invariant features were then used as inputs for learning affect sensitive feature representations. Our work follows these studies, further extending semi-supervised SER. Our study differs from previous studies by effectively training an

autoencoder and a regressor together, such that the auxiliary task of reconstructing the input feature vector and intermediate feature representations helps the primary supervised learning task. Jointly training the autoencoder (auxiliary task) and the regression problem (primary task) is an important contribution leading to more discriminative SER models.

B. Auxiliary Tasks and Multitask Learning

There are multiple studies that have analyzed the regularizing benefits of auxiliary tasks for SER. Xia and Liu [9] combined the learning of emotional categories and emotional attributes. The primary task was the classification of emotional categories. The secondary task was either classification or regression of emotional attributes. Parthasarathy and Busso [8] proposed to jointly predict arousal, valence and dominance scores using a MTL framework, where recognizing one of the attributes was the primary task and recognizing the other two attributes were the secondary tasks. The MTL framework learned the inherent correlation between the various emotional attributes leading to improvements over STL. Similarly, Chang and Scherer [26] used arousal prediction as an auxiliary task for a valence classifier. Chen et al. [27] showed similar improvements in performance for the prediction of time-continuous emotional attributes. Their system jointly predicted arousal and valence scores, obtaining the best performance for the affect sub-task in the *audio/visual emotion challenge* (AVEC) in 2017 [28]. Le et al. [29] also used a MTL framework for time-continuous attribute recognition. Their framework trained classifiers by discretizing attribute scores into discrete classes using the *k-means* algorithm with $k \in \{4, 6, 8, 10\}$. The different classifiers were then learned together as multiple auxiliary tasks using MTL framework. (e.g., learning together a four-class problem and a six-class problem). Similarly, Lotfian and Busso [7] showed improvements for categorical emotion recognition by using a MTL framework for learning the dominant emotion (primary task) and secondary emotions also conveyed in the sentence (auxiliary task).

Previous studies have also considered using other auxiliary tasks to improve SER. Kim et al. [30] used gender and naturalness recognition as auxiliary tasks for emotion recognition. The naturalness task consisted of a binary classifier that determines whether the sentences were natural or acted recordings across different databases. Tao and Liu [10] used gender recognition and speaker identification as auxiliary tasks for classifying emotional categories on the USC-IEMOCAP corpus. Similarly, Zhao et al. [16] transferred age and gender attributes as auxiliary tasks to predict emotion attributes. Abdelwahab and Busso [31] used an auxiliary task for cross-corpus SER. The auxiliary task learned common representation between the source and target domains using a *domain adversarial neural network* (DANN).

C. Ladder Networks

The idea of ladder networks was first proposed by Valpola [32]. This work showed the benefits of using lateral shortcut connections to aid deep unsupervised learning. Rasmus et al. [33], [34] further extended this idea to support supervised learning. Classification and regression tasks were added to the unsupervised reconstruction of inputs through a denoising autoencoder. Finally, Pezeshki et al. [35] studied the various components that affected the ladder network, noting that lateral connections between encoder and decoder and the addition of noise at every layer of the network greatly contributed to the improved performance of this framework.

D. Sentence-Level and Frame-Level Features

Conventionally, SER problems are formulated using sentence-level features over short speech segments. Previous studies often rely on statistics estimated over LLDs, where popular examples include the feature sets proposed for the paralinguistic challenges at Interspeech [36], [37]. An alternative approach is to directly use a sequence of features extracted at the low-level over short segments (e.g., 40 ms). We refer to these features as low-level features or frame-level features. Cummins et al. [38] borrowed successful CNN architectures from the computer vision domain by treating speech spectrograms as images. Mao et al. [25] performed SER in a two step approach using a CNN architecture on low-level features. The first step learned features from unlabeled data and a sparse autoencoder. These features were then used for the recognition task. Trigeorgis et al. [39] proposed a CNN architecture to perform end-to-end SER that took raw speech waveforms as inputs. Neumann and Vu [40] proposed an attention based convolutional neural network for emotion recognition. Yang and Hirschberg [41] predicted arousal and valence using CNNs trained on spectrogram inputs. Likewise, Aldeneh and Provost [42] proposed to train 1-D CNNs on mel-filter bank energies to capture regional saliency for emotion recognition. Following these previous works, our study also examines the effect of our system using CNNs on low-level features, demonstrating that the proposed ladder network architecture can also be implemented with these features (Sections III-D and V-C).

E. Relation to Prior Work

This study presents important contributions with respect to previous studies, including our previous work. The use of ladder network for SER is appealing since the auxiliary task is unsupervised, so we can use data from the target domain without labels. This feature of the proposed approach is a key distinction between our work and most MTL studies, which use supervised auxiliary tasks. When compared to the work of Parthasarathy and Busso [12], this study (1) implements the ladder networks in a semi-supervised fashion instead of a supervised fashion, (2) demonstrates that the proposed architecture can be implemented with different features both at the sentence level, and the frame-level, and (3) evaluates the proposed architecture with extensive within-corpus and cross-corpus evaluations.

The closest study to our paper is the work of Huang et al. [43], which was simultaneously developed with our preliminary study [12]. They also proposed ladder networks for SER tasks. A key distinction between this study and our paper is that Huang et al. [43] only used ladder network to

learn feature representations, where the final classifier was a separate *support vector machine* (SVM). This two-step process is equivalent to training an autoencoder followed by a classifier. Instead, our proposed formulation jointly optimizes the unsupervised reconstruction and the supervised regression task in a single step. The proposed ladder network provides the final predictions for the emotional attribute without any additional regressor, which (1) creates better feature representations that are discriminative for the target task, and (2) allows our formulation to be trained as an end-to-end system (Sec. V-C).

III. Proposed Methodology

A. Motivation

As stated in Section II, data with emotional labels are limited. Furthermore, data from the source domain (train set) is not guaranteed to have the same distribution as the target domain (test set). Therefore, most supervised frameworks trained on one corpus do not generalize well when tested across different tasks and corpora. Therefore, there is a fundamental need to regularize architectures such that they generalize across different tasks. This study aims to increase the generalization of SER models with (1) unsupervised auxiliary tasks, (2) and unlabeled data from the target domain. Our motivations are based on solving unsupervised auxiliary tasks, which aid the primary emotion recognition task. First, we want to fully utilize available labeled data which is expensive to annotate. To this extent, we build a MTL framework (Section III-C) where we jointly learn the dependencies between multiple emotional attributes. The MTL framework regularizes our architecture, but it still demands the utilization of expensive data labeled with emotional information. While labeling audio data for emotion is expensive, unlabeled data is more easily available. The amount of unlabeled data is often greater than the amount of labeled data. The unlabeled data from the target domain can be used to reduce the gap between the source and target domains. We propose to use ladder networks (Section III-B) to effectively leverage unlabeled data. Collectively, the MTL approach combined with the ladder network creates a semi-supervised architecture that effectively generalizes to new domains. This study shows that these powerful representations created by our model can be used across emotional corpora to achieve state-of-the-art SER performance.

B. Ladder Network for Speech Emotion Recognition

Ladder networks, at their core, combine an unsupervised auxiliary task with a supervised classifier or regressor. Using an autoencoder for supervised tasks is not new. Traditionally, the autoencoder is trained separately from the supervised task, where its goal is to learn features representations that are useful for reconstructing the input. However, the information needed to reconstruct the input does not necessarily create a discriminative representation for the classification or regression task. Therefore, it is important to combine the training of the autoencoder with the supervised task, which is a key feature of the ladder networks. Figure 1 illustrates a conceptual ladder network. A noisy version of the encoder is created by adding noise at every layer of the encoder. The goal of the

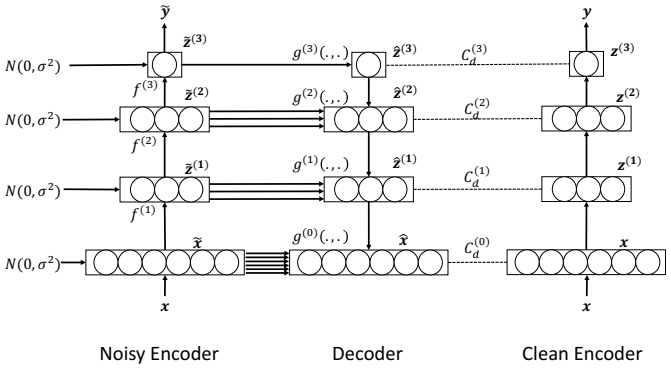


Fig. 1. Ladder network architecture using auxiliary tasks for emotion attribute prediction. The network has a noisy encoder, decoder and clear encoder used for inferences. The ladder connections connect the noisy encoder with the decoder.

autoencoder is to reconstruct the feature representations at the input and intermediate layers. The core concept of the autoencoder in the ladder network involves skip connections between corresponding encoder and decoder layers. Effectively, these skip connections provide a shortcut between the decoder and encoder, bypassing higher layers of the encoder. Therefore, the top layers of the encoder can learn representations better suited towards the primary discriminative task. This is a fundamental difference with simple autoencoders. Note that the ladder network combines the supervised task with an unsupervised auxiliary task. Therefore, the true benefit of the architecture is when it is used in a semi-supervised fashion. The rest of this section explains in detail the encoder and decoder of the ladder network.

Encoder: The encoder consists of a *multilayer perceptron* (MLP). A zero-mean Gaussian noise with variance σ^2 is added to each layer of the MLP ($N(0, \sigma^2)$ in Fig. 1). The decoder is constructed to denoise the noisy latent representations $\tilde{\mathbf{z}}$ at every layer. Therefore, a clean copy of the encoder path is built to get the targets $\mathbf{z}$ for reconstruction (*clean encoder* in Fig. 1). Since the architecture reconstructs intermediate layers, $\hat{\mathbf{z}}$, a trivial solution to minimize the cost is $\hat{\mathbf{z}} = \mathbf{z} =$ constant. To avoid this trivial solution, intermediate layers are normalized using batch normalization. Batch normalization is performed on all layers except the input layer. The scaling and bias values are learned as trainable parameters before applying the activation. Besides encoding the representations for reconstruction, the final layer of the encoder, $\tilde{\mathbf{z}}^{(\mathbf{L})}$, is used for training the supervised regression task, which in our case is the prediction of emotional attributes. The noisy representation $\tilde{\mathbf{z}}$ further regularizes the network. The clean representations $\mathbf{z}$ are used during inference.

Decoder: Similar to the encoder, the decoder of the ladder network is a MLP (*decoder* in Fig. 1). The layers of the decoder network mirrors the layers of the encoder. The decoder is constructed to denoise the noisy representations of the encoder. The denoising process combines top down information from the decoder ($\hat{\mathbf{z}}^{(\mathbf{l+1})}$) with lateral information from the corresponding encoder layer ($\tilde{\mathbf{z}}^{(\mathbf{l})}$). With the lateral connections, the network passes the information needed for denoising the latent representations, bypassing the top layers of the encoder,

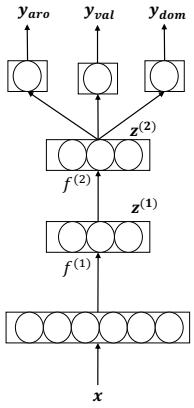


Fig. 2. *Multitask learning* (MTL) architecture to jointly predict arousal, valence and dominance [8]. The ladder network architecture can be implemented with MTL, combining supervised and unsupervised auxiliary tasks.

which can, instead, provide abstract discriminative information for the supervised regression task. As a result, an unsupervised auxiliary cost is added without sacrificing the performance of the architecture for the supervised task. Different denoising functions, $\mathbf{g}(\cdot)$, can model different probability distributions of the latent variables [33], [35]. Previous studies have shown that a single layer MLP combining top decoder layers and lateral encoder layers works the best for most tasks. Our preliminary experiments concluded that the same observation also holds for SER tasks. The denoising function $\mathbf{g}(\cdot)$ takes as input $\mathbf{u}, \tilde{\mathbf{z}}$ and $\mathbf{u} \odot \mathbf{z}$ (layer abbreviations are dropped for clarity), where $\mathbf{u}$ is a batch normalized projection of the decoder layer above $\hat{\mathbf{z}}^{(\mathbf{l+1})}$, $\tilde{\mathbf{z}}$ is the corresponding noisy representation, and $\mathbf{u} \odot \mathbf{z}$ is a element wise product between the decoder and encoder elements. The element wise multiplication assumes that the latent variables are conditionally independent and modulates the encoder representation with the previous decoder layer ($\hat{\mathbf{z}}^{(\mathbf{l+1})}$).

The overall loss for the ladder network is given by

$$C_{Ladder} = C_c + \sum_l \lambda_l C_d^{(l)} \tag{1}$$

where $C_d^{(l)}$ is the reconstruction loss at layer l and λ_l is a hyper-parameter that weighs the reconstruction loss at that layer. The supervised loss for predicting the emotional attributes, C_c, is added when labeled samples are available. Section IV-D gives the implementation and experimental setup used to train the regressor using the proposed ladder network framework.

C. Multitask Learning for Emotion Recognition

While the ladder network architecture makes efficient use of unlabeled samples to regularize the models, the generalization of the models can also be achieved by better utilizing labeled samples. For the prediction of emotional attributes, one appealing method is to jointly learn multiple emotional attributes. This procedure can be effectively done through MTL with shared and attribute-dependent layers [8]. Figure 2 illustrates a MTL network with shared hidden layers that jointly predicts arousal, valence and dominance scores. The overall loss for the MTL architecture is given by

$$C_{MTL} = \alpha C_{aro} + \beta C_{val} + (1 - \alpha - \beta) C_{dom}, \tag{2}$$

where C_{aro}, C_{val} and C_{dom} are individual losses for the prediction of arousal, valence and dominance, respectively. These losses are multiplied by the hyper-parameters α and β, respectively, with $\alpha, \beta \in [0, 1]$ and $\alpha + \beta \leq 1$. Particular solutions of this formulation are the STL frameworks for arousal ($\alpha = 1$, $\beta = 0$), valence ($\alpha = 0$, $\beta = 1$) and dominance ($\alpha = 0$, $\beta = 0$).

An interesting extension of the proposed ladder network formulation for SER is combining the unsupervised and supervised auxiliary losses. We achieve this goal by replacing C_c in Equation 1 with C_{MTL} from Equation 2. In Section V, we evaluate the implementation of the ladder network with STL and MTL.

$$C_{Lad+MTL} = \alpha C_{aro} + \beta C_{val} + (1 - \alpha - \beta) C_{dom} + \sum_l \lambda_l C_d^{(l)} \tag{3}$$

D. Extension of the Architecture for Low-Level Features

The evaluation section mostly considers the proposed ladder networks implemented with sentence-level features, where a feature vector with fixed duration is obtained regardless of the duration of the segment (Sec. IV-B). In Section V-C, we also present results with low-level descriptors to illustrate the flexibility of the proposed architecture for SER. Toward this goal, we present an extension of our architecture using CNNs, which are used to learn discriminative features facilitating end-to-end training.

Most previous frameworks designed for low-level features rely on either low-level features (e.g., MFB) or audio waveforms to learn discriminative features for the task at hand. Such methods enable end-to-end learning, where the features and the classification or regression tasks are jointly learned during training. Following this formulation, this study explores the use of the proposed ladder networks with low-level features. We consider two alternative low-level features: (1) using the LLDs of the ComParE feature set (65D vector – see Sec. IV-B), and (2) MFB energies. Similar to previous studies, we use n=40 bands for the MFB [42]. These models are compared with systems trained with HLDs.

Figure 3 shows the proposed CNN-based architecture for low-level features. The input to the CNN is a 65D $\times T$ matrix (ComParE LLD) or a 40D $\times T$(MFB) matrix, where T is the time dimension. The CNN architecture consists of four convolutional layers followed by two *fully connected* (FC) layers and a linear output layer. The convolutional layers perform 1D convolutions along the time axis with the low-level features as the inputs for the first convolutional layer. We use a 1D max pooling layer after every convolutional layer to sequentially reduce the dimension of the time axis. We flatten the outputs from the final convolutional layer before passing them to the FC layers. While the downstream convolutional

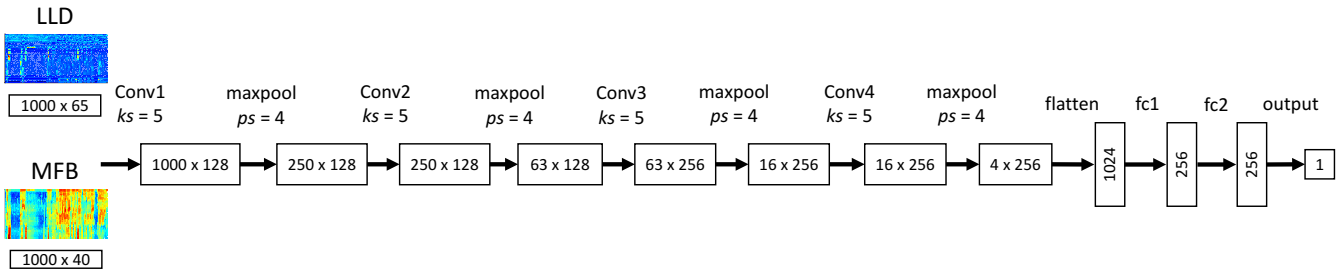


Fig. 3. CNN architecture to predict emotional attributes with low-level features (LLDs or MFB). The architecture contains 4 blocks of 1D convolutional layer followed by a 1D maxpooling layer. After flattering the last convolution layer, the network includes two fully connected layers and the output prediction layer (*ks*: kernel size, *ps*: pool size and *fc*: fully connected).

layers can deal with variable length sequences, the upstream FC layers require a fixed length input. Therefore, we fix T at 1000, which corresponds to 10 seconds of speech (100 fps). We use this value, since most speech segments in the different datasets used in this study are less than 10 seconds. Segments with duration greater than 10 seconds are truncated. Sentences with duration less than 10 seconds are padded with zeros.

IV. Experimental Setup

A. Datasets

This study uses multiple datasets for the different experiments in Section V. The primary corpus is the MSP-Podcast (Version 1.2) [13], used for all the within corpus experiments (Sec. V-A). The MSP-Podcast contains speech collected from online downloadable audio shows, covering various topics such as politics, sports, entertainment, and motivation talks. Therefore, they contain naturalistic speech spanning the emotional spectrum observed during natural conversations. We use a diarization toolkit which identifies segments from distinct speakers. The podcast conversations are sequentially analyzed by automatic algorithms to remove music, silence portions and noisy recordings. We also remove segments with overlapped speech. The selected segments contain a single speaker with duration between 2.75s and 11s. To balance the emotional content of the corpus, we retrieve samples that we believe are emotional following the idea proposed in Mariooryad et al. [44]. Overall, the corpus contains 50 hours of speech (29,440 speaking turns), which were annotated with emotional labels using Amazon Mechanical Turk. The perceptual evaluation used a modified version of the crowdsource-based protocol presented in Burmania et al. [45] to track in real-time the performance of the annotators. The data was annotated for both categorical emotions as well as emotional attributes. This study focuses on the emotional attributes. Each speaking turn was annotated on a scale from one to seven by at least five raters for arousal (1 - very calm, 7 - very active), valence (1 - very negative, 7 - very positive) and dominance (1 - very weak, 7 - very strong). We manually identified speaking turns belonging to 346 speakers in the MSP-Podcast database. The test set contains data from 50 speakers (7,341 speaking turns). The development set contains data from 20 speakers (3,753 speaking turns). The training set has the remaining labeled speaking turns (18,346 segments). This data partition aims to create speaker independent sets for the train, development and testing sets. Besides the labeled data, the MSP-Podcast also contains more than 300 hours of unlabeled data (175,196 segments), corresponding to the pool of clean segments identified from the podcasts, which have not been annotated. This study uses these segments to train the ladder networks (Sec. III-B) in a semi-supervised fashion (within corpus evaluation). Section V-A presents the results of the experiments conducted on the MSP-Podcast corpus.

Besides the MSP-Podcast corpus, we use two other databases for cross corpora evaluations (Sec. V-B). The first database is the USC-IEMOCAP corpus [46], which contains interactions between pairs of actors improvising scenarios. The database contains 10,527 speaking turns from 10 actors appearing in five dyadic sessions. The speech segments were annotated for arousal, valence and dominance by two raters on a five-Likert scale. More information about this corpus is provided in Busso et al. [47]. We also use the MSP-IMPROV corpus [48], which contains interactions between pairs of actors improvising scenarios. In addition to the improvised scenarios, the dataset also contains the interactions between the actors during the breaks, resulting in more naturalistic data. The MSP-IMPROV corpus was annotated with emotional labels using Amazon Mechanical Turk using the approach proposed by Burmania et al. [45]. Each sentence was annotated for arousal, valence and dominance by five or more raters using a five-Likert scale. More information about this corpus is provided in Busso et al. [48].

B. Acoustic Features

This study predominantly uses the acoustic features introduced for the paralinguistic challenge at Interspeech 2013 [37]. These features, which are referred to as the ComParE feature set, are extracted in a two-step procedure. First, LLDs are extracted over 20 millisecond frames (100 fps). These LLDs include loudness, *mel-frequency cepstral coefficients* (MFCCs), fundamental frequency (F0), spectral flux, spectral slope, jitter and shimmer. Second, segment-level features are calculated over the LLDs, leading to a fixed dimensional feature vector. These statistics are referred to as *high-level descriptors* (HLDs) and include various functionals such as the arithmetic and geometric means, standard deviations, peak to peak distances and rise and fall times. The ComParE feature set contains 130 LLDs (65 LLDs + 65 delta) and 6,373 HLDs. Most databases are annotated at the segment-level with a

single annotation capturing the emotional content of the entire segment. Since speech segments have variable lengths, most emotion recognition algorithms have to deal with variable length inputs. The HLDs alleviate this problem by creating a fixed dimension input regardless of the length of the sequence. Previous studies have shown the benefits of HLDs for SER tasks [8], [49].

C. System Description

This study uses two baselines and different implementations of the proposed approach to analyze the performance of the ladder network architecture. All regression systems are trained on the train set, optimizing their performances on the development set. The best system per condition in the development set is then evaluated on the test set, where we report the results.

The study uses two baselines to compare the performance of the proposed architecture. The first baseline uses the STL framework, which is the conventional method for the regression of emotional attributes. The STL framework considers only one of the emotional attribute at a time, creating separate models for arousal, valence and dominance. This approach is referred to as *STL*. The second baseline uses the MTL framework proposed by Parthasarathy and Busso [8] (Section III-C). This system jointly predicts all three emotional attributes, but it only uses supervised auxiliary tasks without the ladder network. It is expected that the MTL systems should provide a stronger baseline compared to the STL systems, since they use supervised auxiliary tasks. This approach is referred to as *MTL*.

The ladder network architecture, denoted with *Lad*, is studied using four implementations grouped into two settings. The first setting only uses the labeled portion of the corpus. The ladder network is implemented as a supervised problem. We denote this setting by adding the term *L* to the name of the system. The second setting uses the entire corpus containing the labeled and unlabeled portions of the corpus. The ladder network is implemented as a semi-supervised problem. For training with the unlabeled set, we alternate between a mini-batch of unlabeled samples and a mini-batch of labeled samples. During training, the losses are not in the same scale when labeled and unlabeled data are alternatively considered. When we present labeled data in the batch, the regression loss is added leading to a total cost that is about twice as high as the reconstruction loss considered with only unlabeled data. As the models are trained, the regression loss for predicting the attributes eventually reduces, leading to equivalent losses for both conditions. Our preliminary experiments indicated that the alternate use of unlabeled and labeled data worked better than learning the models by combining the labeled and unlabeled data. This result indicates that the learning scheme itself regularizes the model as is the case when MTL architectures are used instead of the STL architectures. We denote this setting that uses unlabeled data by adding the term *UL* to the name of the system. For both settings, we implement the ladder network with either STL (Eq. 1) or MTL (Eq. 3). We denote the corresponding implementation by adding the term *STL* or *MTL* to the name of the system. For example, the ladder network trained with labeled and unlabeled data using STL is denoted as *Lad + UL + STL*. We expect that combining MTL with the ladder network should result in improved performance as we use both supervised and unsupervised auxiliary tasks to aid our primary task of predicting emotional attributes.

D. System Architecture

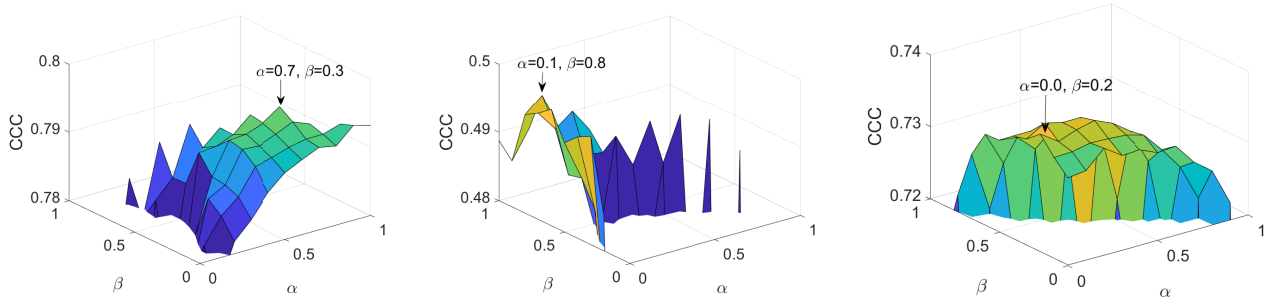
The baselines and the proposed ladder network models are implemented with feed forward dense networks using sentence-level features as inputs. The dense networks contain two hidden layers with 256 nodes in each layer. The activation of the neurons in each layer corresponds to the *rectified linear unit* (ReLU). The input to the dense network is a 6,373D feature vector containing the HLDs for a speaking turn (Sec. IV-B). The output is the predicted value of the emotional attribute. The features and labels are normalized using the z-normalization with the mean and standard deviation calculated over the train set. The models are trained with a learning rate of $5e-5$ for 100 epochs. The model with the best performance on the development set across epochs is evaluated on the test set.

For the architectures of the STL and MTL baselines, we include a dropout of $p = 0.5$ between the input and the first hidden layer, and between the first and second hidden layers. This setting provides the best regularization on the development set.

The hyper-parameters for the MTL methods are optimized using the development set using a grid search approach with a step size is 0.1 for both α and β. These parameters are separately optimized for each emotional attribute. Therefore, we have three systems, one for each attribute, with different combination for α and β. This search approach is independently conducted for the *MTL*, *Lad + L + MTL* and *Lad + UL + MTL* networks. Figure 4 shows the performance of the *MTL* models on the development set as a function of α and β. The figure highlights the values for α and β that produce the best results with the *Lad + L + MTL* network (e.g., $\alpha = 0.7$ and $\beta = 0.3$ for arousal).

For the ladder network, we only use dropout between the input layer and the first hidden layer, following our previous work [12]. The dropout is set to $p = 0.1$. Since the ladder network is also regularized by unsupervised auxiliary tasks, reducing the influence of dropout led to better performance on the development set. For the noisy encoder (Fig. 1), we add a Gaussian noise with variance σ^2 = 0.3 to the encoder. The hyper-parameter for the reconstruction loss is set to $\lambda_l = 1$ (Eqs. 1 and 3). A preliminary search on the development set showed no significant difference between $\lambda_l = 1$, $\lambda_l = 0.1$ and $\lambda_l = 10$. We do not optimize the value of λ_l to reduce the computational resources needed to train the system, acknowledging that better results may be possible by conducting an exhaustive search for this parameter over the development set. The *mean squared error* (MSE) function is used to measure the reconstruction loss.

All our models are trained and evaluated using the *concordance correlation coefficient* (CCC). The CCC maximizes the



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

Fig. 4. Finding the optimal hyper-parameters on the development set for the MTL networks using a grid search with step size is 0.1 for both α and β. The figures show the optimal values for the *Lad + L + MTL* network using sentence-level features (HLDs).

TABLE I
WITHIN-CORPUS EVALUATION ON THE MSP-PODCAST CORPUS. THE RESULTS CORRESPOND TO THE CCC VALUES ACHIEVED BY DIFFERENT IMPLEMENTATIONS OF THE LADDER NETWORK ARCHITECTURE ON THE DEVELOPMENT AND TEST SETS. (● INDICATES THAT ONE MODEL IS SIGNIFICANTLY BETTER THAN THE STL BASELINE; ∗ INDICATES THAT ONE MODEL IS SIGNIFICANTLY BETTER THAN THE MTL BASELINE).

Task	Development		
	Arousal	Valence	Dominance
[43] Lad + L + STL + SVR	0.768	0.355	0.696
STL	0.773	0.491	0.713
MTL	0.782	**0.509**	0.726
Lad + L + STL	0.793$^{\bullet *}$	0.489	0.732$^{\bullet}$
Lad + L + MTL	**0.795**$^{\bullet *}$	0.497	**0.736**$^{\bullet}$
Lad + UL + STL	0.792$^{\bullet *}$	0.489	0.733$^{\bullet}$
Lad + UL + MTL	0.792$^{\bullet *}$	0.489	0.733$^{\bullet}$
	Test		
	Arousal	Valence	Dominance
[43] Lad + L + STL + SVR	0.739	0.202	0.650
STL	0.743	**0.312**	0.670
MTL	0.745	0.293	0.671
Lad + L + STL	0.765$^{\bullet *}$	0.303	0.678
Lad + L + MTL	0.763$^{\bullet *}$	0.293	0.690$^{\bullet *}$
Lad + UL + STL	**0.770**$^{\bullet *}$	0.301	**0.700**$^{\bullet *}$
Lad + UL + MTL	**0.770**$^{\bullet *}$	0.301	**0.700**$^{\bullet *}$

Pearson's correlation between the true and predicted values, while minimizing the difference between their means. Previous studies have shown the benefits of training with CCC as the objective function over the MSE [12], [39], [50]. All neural networks in this study are trained using the NADAM optimizer [51].

V. EXPERIMENTAL RESULTS

A. Within Corpus Results

The experimental evaluation in this section analyzes the power of the proposed ladder network systems for within corpus experiments in the MSP-Podcast corpus.

The unlabeled data comes from segments that are not part of the train, development, or test sets. These recordings are from all the podcasts from where we extracted the segments in the corpus, but have not been retrieved for annotation (Sec. IV-A). Since the segments from the test set were also retrieved from these podcasts, we expect the model to see data that are more similar to the test set by including these unlabeled segments, reducing the mismatch between train and test sets. Notice that

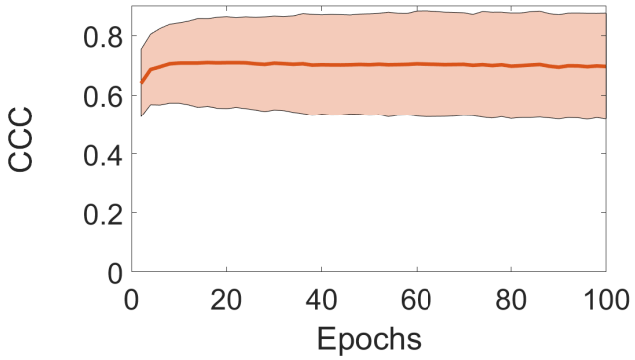
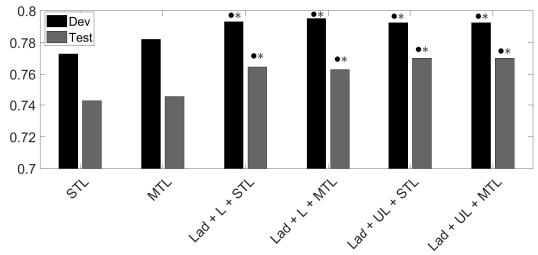


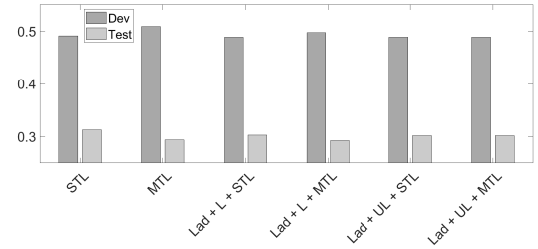
Fig. 5. Performance of the proposed approach for arousal in terms of CCC for multiple initializations of the *Ladder + L + MTL* model. The performance is presented as a function of the number of epochs. The figure illustrates the convergence of the model.

the unlabeled samples do not have emotional labels so it is not a problem to use them during training, even if they were recorded by the same speakers in the test set. The systems are trained and tested on the MSP-Podcast corpus using the ComParE feature sets (Section IV-B). We analyze the performance in terms of CCC for arousal, valence and dominance. In this section, we report and compare the performance of our models on the development and test sets to evaluate the generalization of our approach (Table I). The development set includes the best performance, per model, across epochs obtained on this set. We compare the CCC scores of the proposed models against the baselines, asserting whether the differences in performance are statistically significant using the Fisher Z-transformation test (one-tailed z-test, p-value$<$0.05).

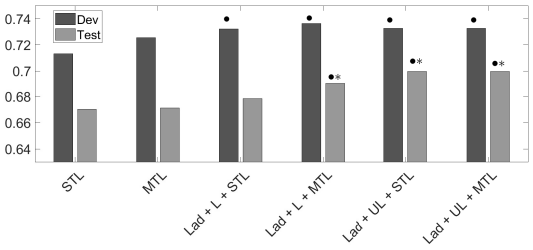
Before we start with the evaluation, we empirically study the convergence of the proposed networks. Using the multiple systems trained to identify the optimal parameters for α and β in Equation 3, we estimate the mean and standard distribution of the CCC values for the emotional attributes as a function of the epochs. This analysis is conducted on the development set. Figure 5 shows an example for arousal using the *Ladder + L + MTL* model. The figure shows that the mean value of the predictions is very stable. The results also show that the standard deviation tends to increase, indicating that after some epochs the models may start to overfit. The results for valence



(a) Arousal



(b) Valence



(c) Dominance

Fig. 6. Within-corpus evaluation on the MSP-Podcast corpus using sentence-level features (HLDs). The figures report the CCC values obtain in the development and test sets (• indicates that one model is significantly better than the STL baseline; $*$ indicates that one model is significantly better than the MTL baseline).

and dominance are similar to the results shown in Figure 5, where the CCC values quickly converge to stable values.

On the development set, Table I shows that the best performing systems for ladder network architectures are significantly better than the STL baseline for arousal and dominance. For these emotional attributes, the best performance is achieved by the ladder network implemented with MTL with only labeled data.

The results on the test set are very consistent with the trends observed in the development set, demonstrating the generalization of the models (Table I). However, the CCC values are lower for the test set compared to the development set (also shown in Fig. 6). The difference can be explained since the development and test sets are different. Also, the results on the development set consist of the best performance obtained during training. In contrast, the results on the test set are CCC values observed with the best configuration of the models evaluated on data that have not been seen before. For arousal, the results of the ladder network frameworks are statistically significantly better than the results achieved by both baseline methods. For dominance, the ladder network architectures trained with labeled and unlabeled data lead to statistically significant improvements over both baseline frameworks. The frameworks trained with unlabeled data give the best performance for both arousal and dominance. Under this setting, the ladder network truly utilizes the abundant unlabeled data and generalizes to unseen data. Table I shows that for within corpus evaluations, the baseline methods achieve better results for valence. We will show in Section V-B that this is not the case for cross corpus evaluations, where our proposed ladder network architectures achieve better performance than the baseline methods for all the emotional attributes.

Figure 6 shows the CCC results for the development and test sets for each of the methods, to visualize the general trends in the results. The statistically significant improvements over the baseline methods are denoted with symbols on top of the bars. Overall, we achieve relative gains of 3.0% for arousal, and 3.5% for dominance using the proposed architectures over the STL method. The performance of the models is lower for valence, following the general patterns reported in previous studies that have shown the difficulty of predicting valence from acoustic cues [50], [52]–[56].

As discussed in Section II-E, the closest study to our work is the approach presented in Huang et al. [43]. This approach uses a two-step process for classification/regression. Similar to autoencoders, the ladder networks in these architectures are used to learn features, which are then used as inputs of a classifier/regressor for the emotion recognition task. We compare our approach with this network, which we refer to as *[43] Lad+L+STL+SVR*. We follow as close as possible the implementation provided by the authors. First, the features are learnt with the ladder network using a single task learning approach (e.g., *Lad+STL*). The feature representation learned by this system is used as input of a separate regressor. Since the task in Huang et al. [43] was a classification problem, they used SVM. Since our task is a regression problem, we use *support vector regressor* (SVR). Following the parameters presented in Huang et al. [43], we used a *radial basis function* (RBF) kernel with the regularization parameter equals to $C = 1.0$, and the tolerance margin equals to 0.1. Table I lists the results, which show that the strategy presented in this paper, where the feature representation and the regression problem are jointly learned leads to significant improvements over the approach presented by Huang et al. [43]. Therefore, the rest of the study will focus on comparisons with the baselines described in Section IV-C.

B. Cross Corpus Results

This study also explores the generalization of the proposed ladder network with cross-corpus experiments. Specifically, we train the models on the MSP-Podcast corpus maximizing performance on its development set. The models are then tested on either the USC-IEMOCAP corpus or MSP-IMPROV corpus. We compare the results with within corpus evaluations using the STL framework, where the models are trained and tested with data from the same corpus (*Within-corpus (WC) Baseline* in Table II). For the within corpus evaluation, the USC-IEMOCAP and MSP-IMPROV corpora are divided into speaker independent partitions. The results are reported across

TABLE II
CROSS-CORPUS EVALUATION WHERE THE MODELS ARE TRAINED ON THE MSP-PODCAST CORPUS AND TESTED ON EITHER THE USC-IEMOCAP OR THE MSP-IMPROV CORPORA. THE TABLE REPORTS THE AVERAGE CCC VALUES ACROSS FOLDS AND THE STANDARD DEVIATION. *WC Baseline* CORRESPONDS TO THE WITHIN-CORPUS BASELINE. ($\bullet$ INDICATES THAT ONE MODEL IS SIGNIFICANTLY BETTER THAN THE STL BASELINE; $*$ INDICATES THAT ONE MODEL IS SIGNIFICANTLY BETTER THAN THE MTL BASELINE).

Task	IEMOCAP		
	Arousal	Valence	Dominance
STL	0.560 ± 0.122	0.135 ± 0.070	0.378 ± 0.103
MTL	0.584 ± 0.078	0.144 ± 0.067	0.370 ± 0.097
Lad + L + STL	$0.590 \pm 0.074^{\bullet *}$	$0.154 \pm 0.052^{\bullet}$	$0.391 \pm 0.107^{\bullet *}$
Lad + L + MTL	$0.589 \pm 0.065^{\bullet}$	0.141 ± 0.056	$0.408 \pm 0.103^{\bullet *}$
Lad + UL + STL	$0.603 \pm 0.043^{\bullet *}$	0.092 ± 0.071	$0.476 \pm 0.076^{\bullet *}$
Lad + UL + MTL	$0.623 \pm 0.036^{\bullet *}$	$0.235 \pm 0.056^{\bullet *}$	$0.441 \pm 0.086^{\bullet *}$
WC Baseline	0.661 ± 0.051	0.487 ± 0.044	0.512 ± 0.055
	MSP-IMPROV		
	Arousal	Valence	Dominance
STL	0.471 ± 0.112	0.235 ± 0.078	0.440 ± 0.134
MTL	0.442 ± 0.116	0.231 ± 0.082	0.449 ± 0.128
Lad + L + STL	$0.490 \pm 0.108^{*}$	$0.287 \pm 0.075^{\bullet *}$	0.436 ± 0.130
Lad + L + MTL	$0.480 \pm 0.107^{*}$	$0.293 \pm 0.073^{\bullet *}$	$0.464 \pm 0.123^{\bullet *}$
Lad + UL + STL	$0.547 \pm 0.094^{\bullet *}$	$0.349 \pm 0.087^{\bullet *}$	$0.463 \pm 0.096^{\bullet *}$
Lad + UL + MTL	$0.547 \pm 0.094^{\bullet *}$	$0.328 \pm 0.091^{\bullet *}$	$0.463 \pm 0.096^{\bullet *}$
WC Baseline	0.599 ± 0.112	0.408 ± 0.090	0.471 ± 0.098

all the test partitions. For consistency, the results for the ladder networks are also estimated for each partition, reporting the average across folds.

We train the ladder network architectures introduced in Section IV-C using labeled and unlabeled data. For the labeled setting, we use samples only from the MSP-Podcast corpus. For the unlabeled setting (*UL* in Table II), we assume we have access to the samples from the target corpus. We include the target corpus for the unsupervised reconstruction using the autoencoder. The use of unlabeled data from the target domain (test set) is not a problem since we do not require the emotional labels, which is the strength of our semi-supervised approach. Since the target domain is used for the unsupervised reconstruction loss, we can guarantee that the distribution of the unlabeled data is exactly the same as the test set, reducing the train-test mismatch. Since the emotional attributes in the MSP-Podcast and the target corpora are annotated on different scales, we transform the attribute scores of the MSP-Podcast corpus to match the scales of the target corpora using an affine transformation. We report the mean and standard deviation over all the test partitions. We compare the CCC values obtained by the ladder networks with the results from the baselines, testing their significance with the one-tailed, matched-paired t-test asserting significance at p-value<0.05. Table II describes the results for the cross-corpus experiments. Figures 7 (USC-IEMOCAP) and 8 (MSP-IMPROV) illustrate the mean performance across test partitions.

First, we discuss the results for the USC-IEMOCAP database. Under the fully labeled setting, the ladder network systems achieve significant improvements over the STL baseline for arousal and dominance. Additionally, the systems significantly improve the performance for dominance compared to the MTL baseline. For valence, we achieve significant gains over the STL baseline with the *Ladder + L + STL* model. With unlabeled data from the USC-IEMOCAP corpus (UL setting), we obtain significant gain over the baselines. The systems perform significantly better than the baselines for all three emotional attributes. We observe relative gains up to 11.3% for arousal, 74.1% for valence, and 25.9% for dominance over the STL baseline (Fig. 7). The CCC values for these systems are closer to the results obtained by the within-corpus baseline. The significant gains reported in this section show the potential of the ladder network architecture, especially when unlabeled data from the target corpus is available.

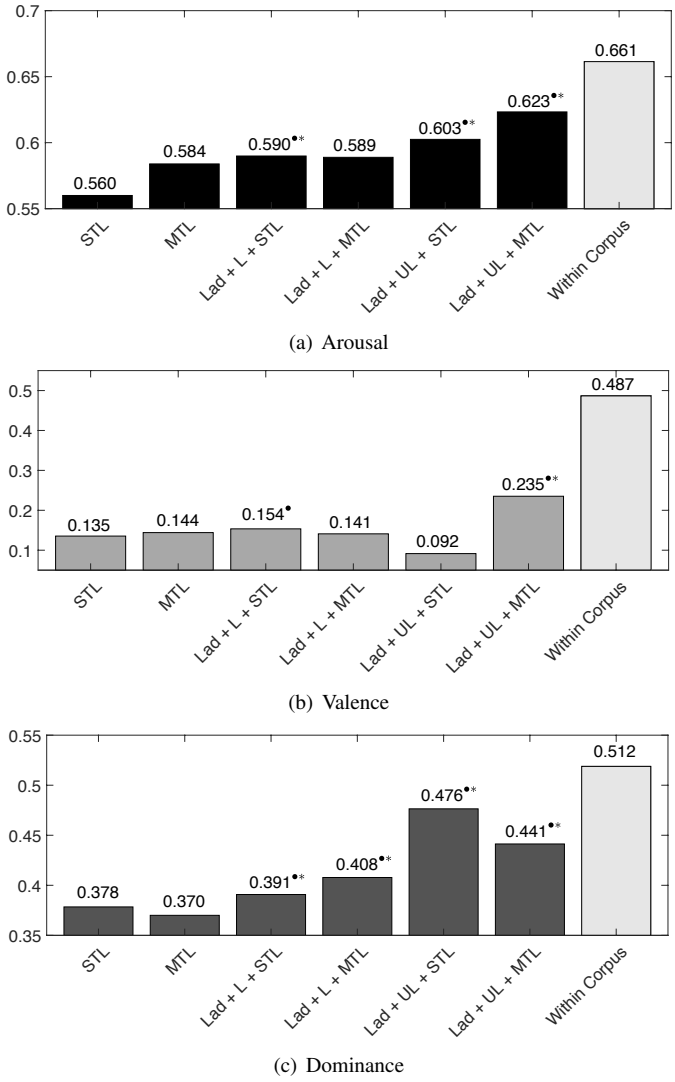
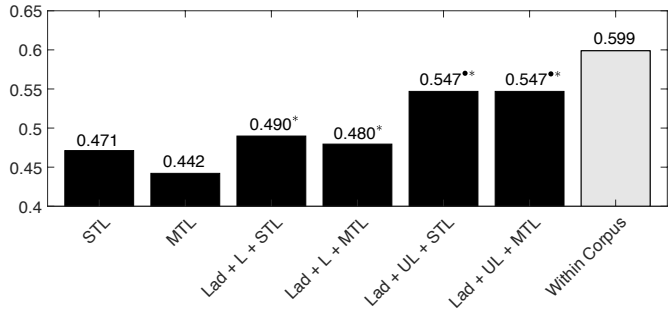
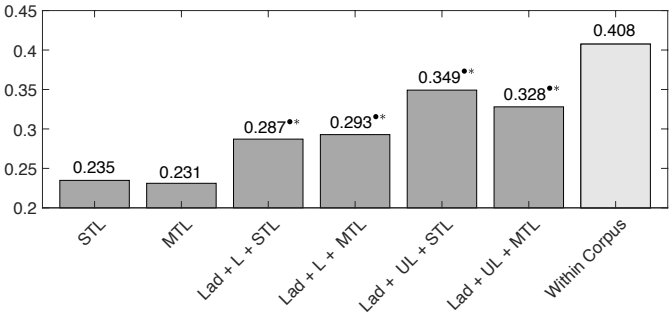


Fig. 7. Cross-corpus evaluation, when the models are tested on the USC-IEMOCAP corpus. The figure reports the average CCC values across folds ($\bullet$ indicates that one model is significantly better than the STL baseline; $*$ indicates that one model is significantly better than the MTL baseline).

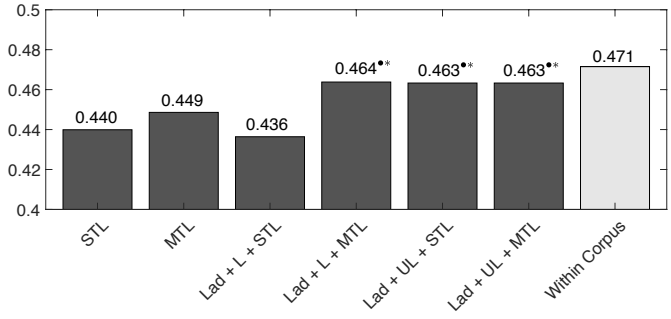
We observe similar results in the evaluation on the MSP-IMPROV database, where most of the architectures using ladder network achieve significant improvements in the CCC values over the STL and MTL baselines. For valence, the proposed architectures perform significantly better than both



(a) Arousal



(b) Valence



(c) Dominance

Fig. 8. Cross-corpus evaluation, when the models are tested on the MSP-IMPROV corpus. The figure reports the average CCC values across folds (• indicates that one model is significantly better than the STL baseline; $*$ indicates that one model is significantly better than the MTL baseline).

baselines. For arousal and dominance, the use of unlabeled data leads to statistically significant improvements over the STL and MTL baselines. Figure 8 shows that the inclusion of unlabeled data from the target corpus greatly improves the performance of the ladder network architectures, achieving CCC scores that are closer to the within-corpus baseline. Under this setting, the proposed systems are significantly better than the baselines for all three emotional attributes. Overall, the *Lad + UL +MTL* architecture achieves relative gains of 16.1% for arousal, 40% for valence, and 5.5% for dominance over the STL baseline. These results demonstrate the real benefits of the ladder network architecture, which generalizes better in cross corpus SER problems.

C. Results with Low-Level Features

This section evaluates the extension of the proposed approach for low-level features described in Section III-D. This analysis aims to show the flexibility of this approach, facilitating an end-to-end training. The analysis in this section includes only within corpus experiments on the MSP-Podcast corpus. All the parameters for the CNN architecture are optimized on the development set of the MSP-Podcast corpus. Training ladder networks with low-level features is computationally expensive. To ease this process, we impose two constraints on the ladder networks trained with low-level features. First, the reconstruction costs are implemented only on the two fully connected layers after the flattening layer (i.e., layers *fc1* and *fc2* in Fig. 3). This network is similar to the τ network suggested by Valpola et al. [32]. Second, we do not use the entire unlabeled portion of the corpus in every epoch. Instead, we use the same number of unlabeled and labeled samples for every epoch, randomly selecting 29,440 unlabeled samples in every epoch. The STL and MTL baselines are also implemented with CNNs.

TABLE III
EVALUATION OF LADDER NETWORK WITH LOW-LEVEL FEATURES. THE RESULTS CORRESPOND TO WITHIN-CORPUS EVALUATIONS USING THE MSP-PODCAST CORPUS. THE TABLE REPORTS CCC FOR DIFFERENT ARCHITECTURES USING CNNS TRAINED WITH EITHER LLDS OR MFB (• INDICATES THAT ONE MODEL IS SIGNIFICANTLY BETTER THAN THE STL BASELINE; $*$ INDICATES THAT ONE MODEL IS SIGNIFICANTLY BETTER THAN THE MTL BASELINE).

Task	LLD-CNN		
	Arousal	Valence	Dominance
STL	0.756	0.244	0.682
MTL	0.759	0.223	0.684
Lad+STL+L	$0.768^{\bullet}$	$0.274^{\bullet *}$	**0.687**
Lad+MTL+L	$0.769^{\bullet}$	$0.274^{\bullet *}$	0.681
Lad+STL+UL	$0.769^{\bullet}$	$\mathbf{0.279}^{\bullet *}$	**0.687**
Lad+MTL+UL	$\mathbf{0.771}^{\bullet *}$	0.269^{*}	0.685
	MFB-CNN		
	Arousal	Valence	Dominance
STL	0.733	0.204	**0.659**
MTL	0.738	$\mathbf{0.254}^{\bullet}$	**0.659**
Lad+STL+L	**0.744**	0.200	0.659
Lad+MTL+L	0.741	0.200	0.659
Lad+STL+UL	0.743	$0.232^{\bullet}$	0.655
Lad+MTL+UL	0.740	0.184	0.656

Table III shows the results for the different systems using the CNN-based architecture trained with either LLDs or MFB features. Similar to Section V-A, we evaluate the differences in CCC values using the Fisher Z-transformation (one-tailed z-test, p-value$<$0.05). When the CNNs are trained with LLDs, we observe that the ladder networks provide significant gains over the baseline for arousal (STL) and valence (STL, MTL). For valence, the proposed architectures provide relative gains up to 14.3% on the test set. For dominance, the models achieve similar performance to the baselines, where the differences are not statistically significant. When the CNNs are trained with MFB, we observe similar performance. We observe statistically significant improvements over the STL baseline only for valence using the *Lad+STL+UL* network. We expect that a better result can be achieved if the reconstruction loss is implemented to also include the convolutional layers.

Finally, we also compare the overall trends of the models trained with sentence-level features (HLD) and low-level features (CNN-LLD, CNN-MFB). For arousal and dominance, we

observe similar performance for systems trained with either the HLDs (sentence-level features), or the CNN-LLD (low-level features). In contrast, the system trained with sentence-level features achieves better results for valence. Notice that models trained with HLDs are still very competitive over end-to-end systems trained with frame-level features [57]. The results are consistently lower when using MFB. MFB features only provide spectral information, while the LLDs and HLDs also provide prosodic and voice quality information, which are important cues for SER problems [58].

VI. Conclusions

This study proposed the use of ladder network in speech emotion recognition. The approach combines the unsupervised auxiliary task of reconstructing intermediate feature representations, with the primary task of predicting emotional attributes. The unsupervised nature of the auxiliary task eases the pressure on the expensive emotion labeling process by leveraging unlabeled data from the source domain. The unsupervised auxiliary task reconstructs the input and the intermediate feature representations through a denoising autoencoder. The ladder networks contain skip connections between the noisy encoder and the decoder, allowing the higher layers of the encoder to learn discriminative representations. Different implementations of the proposed system were evaluated in within corpus evaluations and cross-corpus evaluations. In the within-corpus evaluations, we analyzed the benefits of the proposed architectures over competitive STL and MTL baselines, showing significant improvements for arousal and dominance. In the cross-corpus evaluations, the models were trained on the MSP-Podcast corpus and evaluated on the USC-IEMOCAP and MSP-IMPROV corpora. The results indicated significant gains when using the proposed models, underlying the generalization power of the ladder networks. The improvements were particularly high when using unlabeled data from the target domain, exploiting all the benefits of the proposed architecture. Finally, the study analyzed the performance of the proposed architecture for different feature inputs. We showed that we can achieve similar performance with a CNN-based implementation trained on low-level features.

Based on the cross-corpus results, our future work will explore the use of the representations learned by the ladder networks as inputs for emotion recognition tasks in general. We will also explore the ladder network architecture for emotion recognition from other modalities such as video and image. The results in this study agree with the observations reported in previous studies that have shown the difficulty in predicting valence from acoustic cues [50], [52]–[56]. Our recent study has shown that acoustic cues for valence are highly speaker dependent, where the networks require higher regularization [52]. Our future research direction will use these findings to improve the ladder network architectures for predicting valence scores. Our future research direction will use these findings to improve the ladder network architectures for predicting valence scores. Finally, we aim to improve the performance of the ladder network architecture with low-level features, paying special attention to valence, extending the scope of our data-driven speech emotion recognition systems.

Acknowledgment

This study was funded by the National Science Foundation (NSF) under grant CNS-1823166 and CAREER IIS-1453781.

References

[1] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, July 2015.

[2] 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.

[3] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, vol. 50, no. 6, pp. 487–503, June 2008.

[4] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 4490–4494.

[5] 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.

[6] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, July 1997.

[7] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 951–955.

[8] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1103–1107.

[9] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, January-March 2017.

[10] F. Tao and G. Liu, "Advanced LSTM: a study about better time dependency modeling in emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 2906–2910.

[11] D. Kim, M. Lee, D. Y. Choi, and B. Song, "Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild," in *ACM International Conference on Multimodal Interaction (ICMI 2017)*, Glasgow, UK, November 2017, pp. 529–535.

[12] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3698–3702.

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

[14] 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.

[15] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schuller, "Enhanced semi-supervised learning for multimodal emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5185–5189.

[16] 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. 22 196–22 209, April 2018.

[17] I. Cohen, N. Sebe, F. Cozman, and T. Huang, "Semi-supervised learning for facial expression recognition," in *ACM SIGMM international workshop on Multimedia information retrieval (MIR 2003)*, Berkeley, CA, USA, November 2003, pp. 17–22.

[18] I. Cohen, N. Sebe, F. G. Gozman, M. C. Cirelo, and T. S. Huang, "Learning Bayesian network classifiers for facial expression recognition both labeled and unlabeled data," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, Madison, WI, USA, June 2003, pp. 1–7.

[19] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory (COLT 1998)*, Madison, WI, USA, July 1998, pp. 92–100.

[20] A. Mahdhaoui and M. Chetouani, "Emotional speech classification based on multi view characterization," in *International Conference on Pattern Recognition (ICPR 2010)*, Istanbul, Turkey, August 2010, pp. 4488–4491.

[21] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 115–126, January 2015.

[22] J. Liu, C. Chen, J. Bu, M. You, and J. Tao, "Speech emotion recognition using an enhanced co-training algorithm," in *IEEE International Conference on Multimedia and Expo (ICME 2007)*, Beijing, China, July 2007, pp. 999–1002.

[23] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 31–43, January 2018.

[24] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *ACM international conference on Multimedia (MM 2014)*, Orlando, FL, USA, November 2014, pp. 801–804.

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

[26] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 2746–2750.

[27] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Annual Workshop on Audio/Visual Emotion Challenge (AVEC 2017)*, Mountain View, California, USA, October 2017, pp. 19–26.

[28] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "AVEC 2017: Real-life depression, and affect recognition workshop and challenge," in *Annual Workshop on Audio/Visual Emotion Challenge (AVEC 2017)*, Mountain View, California, USA, October 2017, pp. 3–9.

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

[30] J. Kim, G. Englebienne, K. Truong, and V. Evers, "Towards speech emotion recognition "in the Wild" using aggregated corpora and deep multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1113–1117.

[31] 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.

[32] H. Valpola, "From neural PCA to deep unsupervised learning," in *Advances in Independent Component Analysis and Learning Machines*, E. Bingham, S. Kaski, J. Laaksonen, and J. Lampinen, Eds. London, UK: Academic Press, May 2015, pp. 143–171.

[33] A. Rasmusi, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in neural information processing systems (NIPS 2015)*, Montreal, Canada, December 2015, pp. 3546–3554.

[34] A. Rasmus, H. Valpola, and T. Raiko, "Lateral connections in denoising autoencoders support supervised learning," *CoRR*, vol. abs/1504.08215, pp. 1–5, April 2015. [Online]. Available: http://arxiv.org/abs/1504.08215

[35] M. Pezeshki, L. Fan, P. Brakel, A. Courville, and Y. Bengio, "Deconstructing the ladder network architecture," in *International Conference on Machine Learning (ICML 2016)*, New York, NY, USA, June 2016, pp. 2368–2376.

[36] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Interspeech 2009 - Eurospeech*, Brighton, UK, September 2009, pp. 312–315.

[37] 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.

[38] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *ACM international conference on Multimedia (MM 2017)*, Mountain View, CA, USA, October 2017, pp. 478–484.

[39] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5200–5204.

[40] M. Neumann and N. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1263–1267.

[41] Z. Yang and J. Hirschberg, "Predicting arousal and valence from waveforms and spectrograms using deep neural networks," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3092–3096.

[42] Z. Aldeneh and E. Mower Provost, "Using regional saliency for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 2741–2745.

[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 *Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia 2018)*, Beijing, China, May 2018, pp. 1–5.

[44] 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.

[45] 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.

[46] C. Busso and S. Narayanan, "Recording audio-visual emotional databases from actors: a closer look," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 17–22.

[47] 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.

[48] 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.

[49] M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic Bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 867–881, October 2010.

[50] 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, pp. 22–30, November 2015.

[51] T. Dozat, "Incorporating Nesterov momentum into Adam," in *Workshop track at International Conference on Learning Representations (ICLR 2015)*, San Juan, Puerto Rico, May 2015, pp. 1–4.

[52] 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.

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

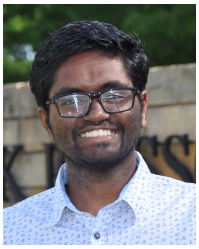
[54] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, October-November 2007.

[55] 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.

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

[57] W. Rao, Z. Lim, Q. Wang, C. Xu, X. Tian, E. Chng, and H. Li, "Investigation of fixed-dimensional speech representations for real-time speech emotion recognition system," in *International Conference on Orange Technologies (ICOT 2017)*, Singapore, Singapore, December 2017, pp. 197–200.

[58] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.



Srinivas Parthasarathy received his BS degree in degree in Electronics and Communication Engineering from College of Engineering Guindy, Anna University, Chennai, India (2012) and MS (2014) and PhD (2019) degrees in Electrical Engineering from the University of Texas at Dallas - UT Dallas. During the academic year 2011-2012, he attended as an exchange student The Royal Institute of Technology (KTH), Sweden. He is an applied scientist at Amazon Research. At UT Dallas, he was a member of the Multimodal Signal Processing (MSP) laboratory. He received the Ericsson Graduate Fellowship during 2013-2014. He has been a research intern at Amazon, Microsoft Research and Bosch Research and Training Center. His research interest includes affective computing, human machine interaction, machine learning and digital signal processing.

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, and AAAC, and a senior member of the IEEE and ACM.