Semisupervised Autoencoders for Speech Emotion Recognition

Jun Deng, Xinzhou Xu, Zixing Zhang, *Member, IEEE*, Sascha Frühholz,
and Björn Schuller, *Senior Member, IEEE*

***Abstract*—Despite the widespread use of supervised learning methods for speech emotion recognition, they are severely restricted due to the lack of sufficient amount of labelled speech data for the training. Considering the wide availability of unlabelled speech data, therefore, this paper proposes semisupervised autoencoders to improve speech emotion recognition. The aim is to reap the benefit from the combination of the labelled data and unlabelled data. The proposed model extends a popular unsupervised autoencoder by carefully adjoining a supervised learning objective. We extensively evaluate the proposed model on the INTERSPEECH 2009 Emotion Challenge database and other four public databases in different scenarios. Experimental results demonstrate that the proposed model achieves state-of-the-art performance with a very small number of labelled data on the challenge task and other tasks, and significantly outperforms other alternative methods.**

***Index Terms*—Autoencoders, speech emotion recognition, semisupervised learning.**

I. Introduction

SPEECH emotion recognition is of vital importance in many real-world applications, such as human-computer interaction and computer-dedicated human communication [1]–[5]. Since its dawn, there has been little doubt that speech emotion recognition is based on supervised learning. Early studies showed that a wide diversity of supervised learning classifiers are competent to build good speech emotion recognition systems, which include Hidden Markov Models (HMMs) [3], [6], Gaussian Mixture Models (GMMs) [7], Support Vector Machines(SVMs) [8], and the likes. Much of the more recent speech emotion recognition research rested on supervised learning methods as well. For example, with the big success of deep learning in speech recognition and image processing [9], [10], a large number of current efforts have been made to leverage deep neural networks for speech emotion recognition [11]–[14].

Despite their widespread use in the speech emotion recognition community, supervised learning methods are restricted by requiring that sufficient amount of labelled speech data for the task of interest are provided at hand. Acquiring a lot of labelled speech data is notoriously difficult since labelling data requires experts' knowledge, which has proven to be prohibitively expensive and time consuming in large quantity. When labelling an emotional corpus, even worse, there is no certain ground truth but a subjective ambiguous "gold standard" because different human raters may interpret the emotional state of the same speech in different ways. In contrast, with the availability of vast amounts of speech data obtained from the Internet, unlabelled speech data are relatively inexpensive and plenty. As a consequence, semi-supervised learning [15], which uses a small number of labelled data in conjunction with an additional set of unlabelled data to improve the generalisation of the learned model, has lately attracted increasing attention [16]–[21] in the community.

Recently, leveraging prior knowledge by means of unsupervised learning had played a key role in the early stage of deep learning [22], [23], which is currently the most active area in the machine learning community. In 2006, Hinton and Salakhutdinov initialised multiple-layered feedforward neural networks by prior information exploited by multiple unsupervised Restricted Boltzmann Machines (RBM) [22]. This is known as the greedy layer-wise unsupervised pre-training algorithm. Since then, a large variety of unsupervised and semi-supervised learning methods have emerged in diverse applications of machine learning. Most of these methods entail training unsupervised learning models with unlabelled data, such as *deep Boltzmann machines* (e. g., [24], [25]) and *autoencoders* (e. g., [26]). In this way, training a deep classifier becomes easy because of the target data distribution explicitly learnt by unsupervised learning models.

However, an underlying problem with combining such unsupervised learning models with supervised learning is that

This work was supported in part by the BMBF IKT2020-Grant under Grant 16SV7213 (EmotAsS), and in part by the European Communitys Seventh Framework Programme through the ERC Starting under Grant 338164 (iHEARu). The work of X. Xu was supported by the Research Grants from the China Scholarship Council. The work of S. Frühholz was supported by the Swiss National Science Foundation (SNSF PP00P1_157409/1).

J. Deng and Z. Zhang are with the Chair of Complex and Intelligent Systems, University of Passau, Passau 94032, Germany (e-mail: jun.deng@tum.de; zixing.zhang@tum.de).

X. Xu is with the Machine Intelligence and Signal Processing Group, MMK, Technische Universität München, Munich 80333, Germany, and also with the Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 210018, China (e-mail: xinzhou.xu@tum.de).

S. Frühholz is with the Institute of Psychology and Center for Integrative Human Physiology (ZIHP), University of Zurich, Zurich 8006, Switzerland, and also with the Neuroscience Center Zurich, University of Zurich and ETH Zurich, Zurich 8008, Switzerland (e-mail: sascha.fruehholz@psychologie.uzh.ch).

B. Schuller is with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg 86159, Germany, and also with the GLAM, Department of Computing, Imperial College London, London SW7 2AZ, U.K. (e-mail: schuller@ieee.org).

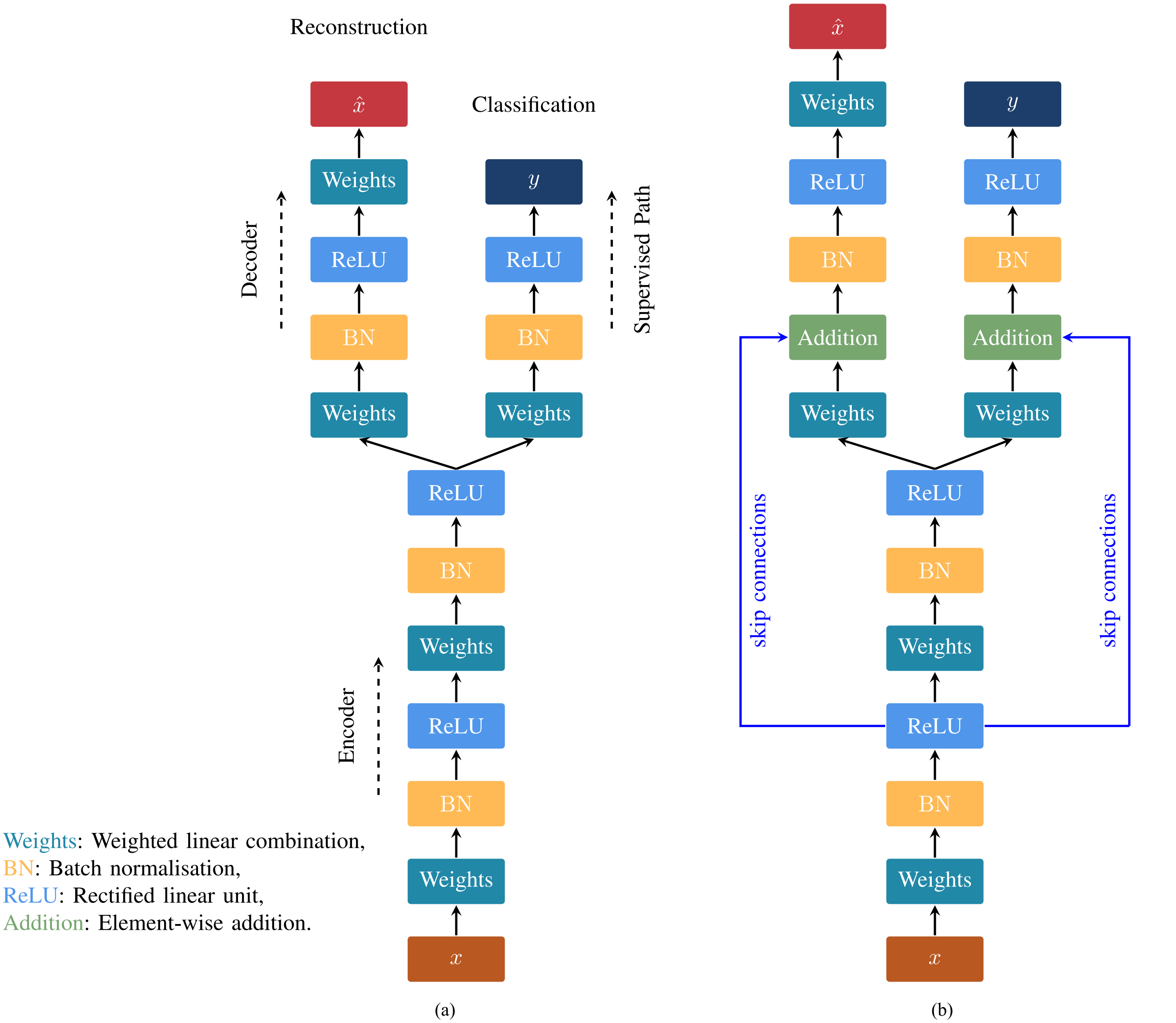


Fig. 1. Illustration of our proposed semi-supervised autoencoders. (a) Architecture of semi-supervised autoencoders (**SS-AE**). (b) Architecture of semi-supervised autoencoders with identity skip connections (**SS-AE-Skip**), where some layer of the encoder is directly connected to one layer of the decoder and supervised paths.

these unsupervised learning models aim to retain all the information that is needed to perfectly reconstruct the input examples, whereas supervised learning preserves only important information that is useful to predict the class label, and drops redundant information. In this case, there is a potential conflict of interest between the unsupervised and supervised learning.

To address this problem, recently, an emerging area of deep semi-supervised learning has attracted growing interest [19], [27], [28]. For example, the semi-supervised variational autoencoders [28] method and the ladder network [19] were proposed, which both obtained very impressive results on several image classification benchmarks using just hundreds of labelled training examples. One of the main attractions of using deep semi-supervised learning for speech emotion recognition is the fact that it supports to simultaneously deploy both unsupervised and supervised learning. It also facilitates the construction of deep structures. This is an important benefit because deep structures always represent many complex functions more concisely than common shallow models (e. g., SVMs and GMMs) [29].

Encouraged by the recent success of deep semi-supervised learning, we propose *semi-supervised autoencoders* for speech emotion recognition. The idea of the semi-supervised autoencoders method is to combine both the generative and the discriminate perspective. In the end, speech emotion recognition integrated with the proposed semi-supervised leaning would not only reduce the dependence on the great quantity of labelled training examples, but would also be endowed with an ability to distil essential knowledge from unlabelled data into the supervised learning. To the best of our knowledge, this is the first work on deep semi-supervised learning in speech emotion recognition.

To learn from labelled and unlabelled data, the Semi-Supervised Autoencoder (SS-AE), illustrated in Fig. 1, extends a popular unsupervised deep *denoising autoencoder* (DAE) [26] by adjoining the supervised learning objective of a deep feedforward network. As the supervised classifier learns from given labelled data, in our method, this classifier is also enforced to predict all unlabelled data as a "dustbin" class, which leads to explicitly aiding the supervised learning that incorporates prior information from unlabelled examples. This accomplishes by just appending an extra class to the supervised task. Guided with the knowledge of the supervised task, further, the unsupervised objective comes down to providing representations that are important for classification. Consequently, this present approach constructs a joint objective function that must be optimised to ensure that the reconstruction error of the unsupervised objective, as well as the predictive error measured by the supervised objective, are minimised on both the labelled and unlabelled data. In addition, to address the problem of exploding and vanishing gradients in deep neural networks [30], we propose a variant of SS-AE that introduces *skip connections* from the lower layer to the upper one. Such a variant, called *SS-AE-Skip*, is illustrated in Fig. 1(b). Because of the use of identity skip connections, SS-AE-Skip can have paths along which information can flow smoothly across various layers during the training. Finally, we demonstrate the effectiveness and efficiency of SS-AE and SS-AE-Skip through extensive experimental evaluation on the INTERSPEECH 2009 Emotion Challenge database and other four public speech emotional databases in different scenarios.

In addition to the motivation provided above, the core contributions of this paper can be summarised as follows:

1) Most existing methods in speech emotion recognition are restricted by requiring that sufficient amount of labelled speech emotional data are provided at hand. In this paper, we exhibit, for the first time ever, our proposed semi-supervised learning method for speech emotion recognition, which can reach state-of-the-art accuracy with only a few labelled examples.
2) Unlike the previous autoencoder-based approaches in speech emotion recognition, which often perform unsupervised feature learning and then train a classifier, we propose the self-contained SS-AE framework. A dedicated integration of a supervised path and autoencoders is presented to extend the horizon beyond the current limit of unsupervised learning autoencoders. In this way, SS-AE is not only just a powerful feature extractor, but is also a competitive semi-supervised classifier.
3) We compare our semi-supervised learning approach with other prominent semi-supervised learning methods as well as supervised learning methods for speech emotion recognition. We subject this method to thorough evaluation on the INTERSPEECH 2009 Emotion Challenge database and other four speech emotional databases. Extensive experimental results show our semi-supervised learning method outperforms the counterparts.

The remainder of this paper is organised as follows. Section II first discusses related work. We then present the proposed semi-supervised learning methods in Section III. In Section IV, we briefly introduce the selected real-world classification task for exemplification of effectiveness, including five chosen speech emotional databases and acoustic features used. Sections V and VI demonstrate experiments on the five corpora. Finally, a general discuss is provided in Section VII before we draw a conclusion and point out promising future work in Section VIII.

II. Related Work

Speech emotion recognition has greatly benefited from the success of deep learning. For example, in [13], the authors proposed convolutional recurrent neural networks to enhance feature extraction from emotional speech data, which shows an improvement in performance when compared to traditional supervised learning methods. In [14], the authors obtained impressive performance by proposing the deep convolutional neural networks based framework that directly inferred emotional states from the raw speech waveform, instead of from the hand-crafted features. However, there is little work on deep semi-supervised learning for speech emotion recognition. In this work, we bridge the gap by exploiting semi-supervised autoencoders for speech emotion recognition. In particular, we show for the first time how deep semi-supervised learning methods can be brought to considerably advance speech emotion recognition.

Most existing semi-supervised learning approaches, such as self-training [17], [31]–[33] or co-training [34], [35], generally start with training a weak supervised learning classifier with a small training set and then iteratively retrain the classifier on self-labelled examples predicted by the current weak classier. Obviously, such methods are vulnerable to the *training bias* problem, that is, unlabelled data may be predicted erroneously, which will degrade the model in the next iteration. To overcome this major issue, one usually needs to introduce some scheme for only selecting self-labelled examples that meet defined requirements. Much recent research found that integrating active learning and these mentioned semi-supervised learning approaches would effectively overcome the problem and embrace the benefit of both (e. g., cooperative learning [17]). Nevertheless, these heuristic choices may still yield unreliable predictions. Hence, unlike these approaches dependent on a self-labelling process, our present model advances autoencoders considerably in a way that the learning process is guided by the knowledge from unlabelled and labelled data. This allows our semi-supervised learning algorithms to facilitate relating data and labels, and in turn improve classification.

Graph-based approaches make the smoothness assumption: labelled and unlabelled examples are nodes connected by a graph, where undirected edges reflect similarity between examples. Consequently, connected nodes tend to have the same label. In other words, label information 'propagates' from labelled examples to unlabelled examples via graph edges. In this context, the discriminant function is encouraged to vary smoothly with respect to the graph [36], [37]. Instead of resting on the label information propagation, our present methods assign a single dustbin class for unlabelled data, leading to a united supervised objective function for unlabelled and labelled data. Furthermore, graph-based approaches rely much more on the discriminative quality of the features that are given as input, whereas our proposed autoencoders can learn useful non-linear features themselves using their hidden units.

Autoencoders, which are often an unsupervised learning model, have been highly successful in addressing the distribution mismatch issue in speech emotion recognition [16], [38]. Most of the existing approaches in the field tend to use autoencoders to discover common feature representations across different domains in an unsupervised way and then feed them as input to a discriminative classifier (e. g., SVMs). In addition, they are also usually used to provide salient representation, leading to notable improvement for speech emotion recognition. However, all of the above autoencoder-based approaches perform feature learning and classifier learning in separate phases [12], [39]. In this case, the unsupervised feature learning is taught towards minimising the reconstruction error, rather than towards minimising the classification error. In contrast, our proposed method is the first self-sufficient deep semi-supervised structure for speech emotion recognition. That is, there is direct interaction between feature learning and classifier learning. In our method, all the parameters of feature learning and classifier learning are jointly learned by directly minimising both the classification error and the reconstruction error.

A recent work proposed a deep semi-supervised learning algorithm using variational autoencoders that perform efficient approximate inference and learning with generative models [28], whereas our work that is a just simple feed-forward neural network for semi-supervised learning is easy to be trained via the *backpropagation* procedure.

Our semi-supervised learning approach is related to the discriminative RBM [40], [41] in terms of incorporating a discriminative component to training. However, the discriminative RBM worked only with a negative log-likelihood objective for the unlabelled data whereas our autoencoders proposed here can distil simultaneously the information from the unlabelled and labelled data using a novel joint objective function presented in Section III, which thus leads to greater expressive and discriminative power.

III. Semi-Supervised Autoencoders

In this section, we describe the Semi-Supervised Autoencoders (SS-AEs) architecture. Let us consider a dataset with N labelled examples $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ and M unlabelled examples $\{x_{N+1}, x_{N+2}, \ldots, x_{N+M}\}$, where $y \in \{1, 2, \ldots, K\}$, and K is the total number of labels. The objective is to learn a function $P(y|x)$ from both the labelled and unlabelled data. In the SS-AE model, this function is a deep denoising autoencoder that consists of multiple hidden layers and includes a stochastic corruption process applied to the input. Further, the SS-AE model particularly assigns the M unlabelled examples to the pseudo-class $K+1$, resulting in a supervised path with shared parameters, which is responsible for the classification (see Fig. 1). In this end, the objective function is a weighted sum of the supervised cross entropy loss and the unsupervised mean square error loss for the labelled and unlabelled data.

A. The Encoder

Formally, SS-AEs inherit an autoencoder architecture, which consist of an *encoder*, a *decoder*, and a *supervised path*. Given an input x, the encoder that non-linearly transforms the input into a new representation via a multi-layer feedforward neural network is defined as follows:

$$h_e^L = f(z_e^L), \tag{1}$$

where

$$z_e^L = W_e^L h_e^{L-1} + b_e^L, \tag{2}$$

$$h_e^l = f\left(W_e^l h_e^{l-1} + b_e^l\right), \text{and } 2 \leq l \leq L-1, \tag{3}$$

$$h_e^1 = x. \tag{4}$$

The matrix W^l and the vector b^l are referred to as *weights* and *bias* respectively, which are adaptive parameters. In the neural network, they are used to take a *weighted linear combination* of the inputs. The function $f(\cdot)$ is a differentiable and nonlinear *activation function*, such as the Rectifier Linear Unit (ReLU) [42] in this work. L represents the number of layers.

B. The Decoder and the Unsupervised Objective Function

The decoder maps the hidden representation h_e^L from the encoder back to a reconstruction $\hat{x}$ of the original input

$$\hat{x} = h_d^L = z_d^L, \tag{5}$$

where

$$z_d^L = W_d^L h_d^{L-1} + b_d^L, \tag{6}$$

$$h_d^l = f\left(W_d^l h_d^{l-1} + b_d^l\right), \text{and } 2 \leq l \leq L-1, \tag{7}$$

$$h_d^1 = h_e^L. \tag{8}$$

In order to encourage the encoder and decoder to generate underlying feature representations, the inputs are stochastically disturbed via a corrupting function. Such autoencoders are known as DAEs [26].

With the encoder and decoder, the unsupervised objective function, which is a measure of the reconstruction error, is thus computed on the labelled and unlabelled data

$$\mathcal{L}^u = \frac{1}{2} \sum_{i=1}^{N+M} |x_i - \hat{x}_i|^2. \tag{9}$$

C. The Supervised Path and the Supervised Objective Function

In addition, the SS-AE has the supervised path which takes the hidden representation h_e^L and then learns the link between it and its label. The output of the supervised path is computed as follows:

$$\hat{p} = h_s^{L_s} = f(z_s^{L_s}), \tag{10}$$

where

$$z_s^{L_s} = W_s^{L_s} h_s^{L_s-1} + b_s^{L_s}, \tag{11}$$

$$h_s^l = f\left(W_s^l h_s^{l-1} + b_s^l\right), \text{and } 2 \leq l \leq L_s - 1, \tag{12}$$

$$h_s^1 = h_e^L. \tag{13}$$

Here, L_s represents the number of layers in the supervised path. The attractive feature of such autoencoders built in this way is that they consider supervising information for learning predictive latent features, which are good for classification.

For semi-supervised learning, the supervised path also has a part to play in incorporating information from unlabelled data into learning how to recognise emotions. To this end, we further propose to specially describe the unlabelled data as a dustbin class. That is, the M unlabelled examples have an identical label, $\left\{(x_{N+1}, y^\star_{N+1}), (x_{N+2}, y^\star_{N+2}), \ldots, (x_{N+M}, y^\star_{N+M})\right\}$, where $y^\star = K + 1$. In this case, the cross entropy objective function, which is normally used as the loss function for classification tasks in neural networks, is computed over all labelled and unlabelled data

$$\mathcal{L}^s = -\sum_{i=1}^{N+M} \log\left(\frac{\exp\left(\hat{p}_i[y_i]\right)}{\sum_k^{K+1} \exp\left(\hat{p}_i[k]\right)}\right) \tag{14}$$

$$= \sum_{i=1}^{N+M} \left(-\hat{p}_i[y_i] + \log\left(\sum_k^{K+1} \exp\left(\hat{p}_i[k]\right)\right)\right), \tag{15}$$

where y_i is the truth label for the i-th input example, and $\hat{p}_i[k]$ is the k-th element of the vector of class scores $\hat{p}_i$.

The introduction of a dustbin class for the unlabelled data is of importance because it acts as a regulariser during training to improve generalisation and reduce overfitting by preventing the network from assigning full probability to each labelled training example.

D. Joint Objective Function

The objective function in SS-AE ends up a joint function between the reconstruction error and the cross entropy loss,

$$\mathcal{L} = \mathcal{L}^s + C\mathcal{L}^u, \tag{16}$$

where $C \geq 0$ is a term controlling the trade-off between the unsupervised and supervised objectives.

For evaluation, we pick out the index of the highest probable element in the output score vector as the prediction for a test input

$$\tilde{y} = \mathop{\arg\max}_{1 \leq k \leq K} \hat{p}[k], \tag{17}$$

where $\hat{p}$ is the output of the supervised path. Note that the score from the dustbin node is just ignored in the evaluation process.

Introducing a dustbin class to unlabelled data in the SS-AEs has two properties, which are crucial for recognition. First, using a single dustbin class for unlabelled data forces all inputs to contain supervising information, allowing us to trick a classifier into learning from all given data for the problem. Second, because speech data are typically characterised by high degrees of variation, an extra supervised learning task to recognise the dustbin class encourages the neural network classifier to be insensitive to rich variations among the unlabelled and labelled speech emotional data.

Although the SS-AE structure excluding the decoder corresponds to a special case of a deep supervised neural network, which is able to deal with a semi-supervised learning task, the decoder plays a key role in SS-AEs. In speech emotion recognition, we are always faced with a small amount of data, which easily leads to over-fitting in regular deep neural networks and heavily limits the capacity due to the lack of a large size of training data to train deep neural networks well. SS-AEs alleviate these problems by including the reconstruction process into regulating the supervised learning, which significantly improves the model capacity to learn good representations.

As shown in Fig. 1, we apply *Batch Normalisation* before applying a nonlinear activation function for all hidden layers in this work so as to address the *internal covariance shift* issue [43], leading to the fast learning.

E. Semi-Supervised Autoencoders with Skip Connections

More recently, using identity skip connections has been found helpful for easing optimisation in learning very deep feedforward neural networks, such as the highway networks [44] and the residual networks [45], since the use of identity skip connections can mitigate the problem of exploding and vanishing gradients [30]. These skip connections between internal layers of neural networks can allow the information to flow more freely in both forward and backward passes. Motivated by these works, we extend SS-AEs by including skip connections, called SS-AE-Skip. As shown in Fig. 1(b), some layer of the encoder is directly connected to one layer of the decoder and supervised path. Mathematically, the outputs of these layers are defined as follows

$$h_d^l = h_e^l + z_d^l, \tag{18}$$

$$h_s^l = h_e^l + z_s^l. \tag{19}$$

These skip connections in SS-AE-Skip are appealing because unlike SS-AEs which take an effort to maintain the information flow, the SS-AE-Skip method fully focuses on discovering expressive features which are relevant for the task at hand.

IV. Experiments

A. Selected Task and Data

We first perform the INTERSPEECH 2009 Emotion Challenge five-class task [46] using our proposed semi-supervised learning method. It is based on the spontaneous *FAU Aibo Emotion Corpus* (AEC), where each utterance is assigned one of five class labels: *Anger*, *Emphatic*, *Neutral*, *Positive*, and*Rest*. In total, the training and test sets have 9 959 and 8 257 chunks. Table I briefs on FAU AEC. For the experiments to follow, we always evaluate the emotion recognition model on the test set of the AEC as was used in the challenge. Further, we choose Unweighted Average Recall (UAR) to measure the performance as was the competition measure in the challenge. The AEC dataset often serves as a benchmark dataset in speech emo-

TABLE I
SUMMARY OF THE FIVE CHOSEN DATABASES

Corpus	Age	Language	Content	Type	# Emotion	# All	h:mm	#m	#f	Rec	Rate (kHz)
AEC	children	German	variable	natural	5	9 959 / 8 257	9:20	21	30	normal	16
ABC	adults	German	variable	acted	6	430	1:15	4	4	studio	16
EMO	adults	German	fixed	acted	7	494	0:22	5	5	studio	16
SUSAS	adults	English	fixed	natural	4	3 593	1:01	4	3	noisy	8
GeWEC	adults	French	fixed	acted	4	1 200	0:14	2	2	studio	16

Content fixed/variable (spoken text). Type of material (acted/natural). Number of emotion categories (# Emotion). Overall number of turns (# All) – for AEC divided into official training and test set by "/". Total audio time (h:mm). Number of female (# f) and male (# m) subjects. Recording (Rec) conditions (studio/normal/noisy).

tion recognition. The baseline system in [46] achieved a 38.2% UAR with a standard feature set plus linear-kernel SVMs. The best performance system in the 2009 challenge achieved a 41.7% UAR [47]. The best known result, 45.6% UAR was achieved by using hidden Markov models with deep belief networks [11]. More recently, a Convolutional LSTM net was proposed to enhance feature extraction, resulting in an average of 39.7% UAR [13]. Based on contractive autoencoders, a semi-supervised feature learning framework was proposed in which a 40.2 % UAR was achieved [12]. However, this work first made use of autoencoders to generate good features and then employed SVMs to perform classification.

Secondly, we investigate effectiveness of the proposed method with a whispered speech database. Specifically, the *Geneva Whispered Emotion Corpus* (GeWEC) is used to provide normal phonated/whispered paired utterances [48]. Two male and two female professional French-speaking actors in Geneva were recruited to speak eight predefined French pseudo-words (e. g., *"belam"* and *"molen"*) with a given emotional state in both normal and whispered speech modes as in the GEMEPS-corpus that was used in the INTERSPEECH 2013 Computational Paralinguistics Challenge [49]. Speech was expressed in four emotional states: *angry*, *fear*, *happiness*, and *neutral*. The actors were requested to express each word in all four emotional states five times. The utterances were labelled based on the state they should be expressed in, i. e., one emotion label was assigned to each utterance. As a result, GeWEC consists of 1 280 instances in total. In the upcoming experiments, the whispered speech is used for training while the normal speech mode data are used for testing.

In addition, three further publicly available and popular databases, namely the *Airplane Behavior Corpus* (ABC) [50], the *Berlin EMOtional speech database* (EMO) [51], and the *Speech Under Simulated and Actual Stress* (SUSAS) set [52] are chosen as unlabelled training sets. ABC was introduced for the special application of automatic public transport surveillance about passenger emotions. Eight German-speaking subjects actively participated in the recording. In total, there are 431 utterances which were annotated using a closed set of emotion categories, including *neutral*, *tired*, *aggressive*, *cheerful*, *intoxicated*, and *nervous*. EMO contains 494 German utterances expressed in one of seven emotions: *anger*, *boredom*, *disgust*, *fear*, *happiness*, *neutral*, and *sadness*. SUSAS was a first reference for spontaneous recordings. The 3 593 actual stress speech samples are used for the upcoming evaluation in this work, which were recorded in subject motion fear and stress tasks.

TABLE II
OVERVIEW OF THE STANDARDISED FEATURE SET PROVIDED BY THE INTERSPEECH 2009 EMOTION CHALLENGE

LLDs (16 × 2)	Functionals (12)
(Δ) ZCR	mean
(Δ) RMS Energy	standard deviation
(Δ) F0	kurtosis, skewness
(Δ) HNR	extremes: value, rel, position, range
(Δ) MFCC 1–12	linear regression: offset, slope, MSE

As the above outlined, five publicly available emotion corpora were selected to evaluate the effectiveness of the proposed methods. Table I summarises the five selected databases and shows the existing difference between them. Our work differs from the previous work for speech emotion recognition in that we use a small set of labelled data and a set of unlabelled data to do a speech emotion recognition task. To the best of our knowledge, it is the first work to show that the semi-supervised learning algorithm with just a few labelled examples will achieve performance as competitive as others in speech emotion recognition.

B. Acoustic Features

To keep in line with the INTERSPEECH 2009 Emotion Challenge [46], we decided to use its standardised feature set of 12 functionals applied to 2×16 acoustic Low-Level Descriptors (LLDs) including their first order delta regression coefficients as shown in Table II. In detail, the 16 LLDs are zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) frame energy, fundamental frequency (normalised to 500 Hz), harmonics-to-noise ratio (HNR) by autocorrelation function, and Mel-frequency cepstral coefficient (MFCC) 1–12. Then, 12 functionals – mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and ranges as well as two linear regression coefficients with their mean square error (MSE) – are applied on the chunk level. Thus, the total feature vector per chunk contains $16 \times 2 \times 12 = 384$ attributes. To ensure reproducibility as well, the open source openSMILE toolkit was used with the pre-defined challenge configuration [53], [54].

TABLE III
AVERAGE UAR WITH STANDARD DEVIATION OVER TEN TRIALS ON THE AEC TEST SET WITH 100, 200, 500, AND 1000 LABELLED TRAINING EXAMPLES

	# of labelled examples from AEC				
	100	200	500	1 000	All
Supervised methods:					
DNN	$33.5_{\pm 2.1}$	$34.9_{\pm 1.8}$	$36.8_{\pm 1.9}$	$38.6_{\pm 1.3}$	
SVM	$32.8_{\pm 1.9}$	$33.8_{\pm 2.4}$	$36.0_{\pm 1.6}$	$37.6_{\pm 1.5}$	
Semi-Supervised methods:					
Self-training+SVM	$32.1_{\pm 2.0}$	$34.1_{\pm 2.1}$	$36.9_{\pm 1.2}$	$39.6_{\pm 0.6}$	
Label propagation+SVM	$29.9_{\pm 2.2}$	$32.7_{\pm 1.7}$	$33.9_{\pm 1.3}$	$35.6_{\pm 1.3}$	
Label spreading+SVM	$30.6_{\pm 2.1}$	$33.7_{\pm 1.5}$	$35.0_{\pm 1.4}$	$36.4_{\pm 1.2}$	
DAE+SVM	$34.7_{\pm 1.3}$	$35.3_{\pm 2.2}$	$38.9_{\pm 1.1}$	$40.3_{\pm 0.9}$	
Previously reported methods:					
Challenge baseline [46]					38.2
Convolutional LSTM [13]					$39.7_{\pm 0.2}$
Semi-sup. contractive autoencoders [12]					40.2
Bayesian logistic regression [55]					41.6
GMM [47]					41.7
uLSIF [56]					42.7
Ranking SVM [57]					44.8
Deep belief networks [11]					45.6
Our proposed methods:					
SS-AE	$36.6_{\pm 1.6}$ (39.9)	$38.4_{\pm 2.1}$ (42.4)	$40.1_{\pm 1.6}$ (42.6)	$41.5_{\pm 1.1}$ (43.2)	
SS-AE-Skip	$36.5_{\pm 1.8}$ (40.1)	$38.5_{\pm 2.5}$ (43.1)	$41.1_{\pm 1.3}$ (43.1)	$41.8_{\pm 1.0}$ (43.6)	

Unlabelled data are from the AEC training set. We compare our proposed methods with previously reported AEC test UARs and other semi-supervised learning (Semi-sup.) methods. Best UAR in parentheses.

C. Experimental Setup and Evaluation Metrics

In the neural network learning process, we applied the Adam optimisation algorithm [58] with maximum 100 epochs to optimise the parameters. For training the SS-AE neural networks, we inject Gaussian noise with a variance of 0.3 to generate the corrupted input. We used grid search to search over the learning rate $\{0.1, 0.01, 0.001, 0.0001\}$ and the number of hidden nodes $\{128, 256, 512, 1\,024\}$. The number of hidden layers for the encoder and decoder is set to two while the supervised path has only one hidden layer. Each hidden layer has the same hidden nodes. The hyper-parameter C in the objective function is set as to one in order to reduce the effort of parameter search. Input and target features are standardised to zero mean and unit variance on the training set.

We evaluate the performance by UAR, which is often used as the officially-recommended measure for speech emotion recognition. It equals the sum of recalls per class divided by the number of classes, and better reflects overall accuracy in the given case of presence of class imbalance. Besides, significance tests are conducted by computing a one-sided z-test.

V. EXPERIMENTAL RESULTS ON AEC

A. Emotion Recognition with Unlabelled In-Domain Data

First, we pay attention to four semi-supervised emotion recognition tasks with 100, 200, 500, and 1000 labelled examples for the Emotion Challenge. Labelled examples are chosen randomly from the officially training set but the number of examples in each class is balanced. In order to determine the hyper-parameters, we held out 1 000 training examples for validation. Each experiment is repeated ten times with different seeds and different selections of labelled examples to reduce singularity effects. Unlabelled examples are the rest AEC training examples, which is from same domain as the test set.

Table III collects the average UARs with standard deviation over ten trials for SS-AE and SS-AE-Skip and other models for comparison. Furthermore, the best UAR over ten trials is present in the table as well. In Table III, these comparison models include two supervised baseline methods, four semi-supervised learning methods, and eight previous methods that reported performance on the AEC benchmark database. A supervised baseline, **DNN**, consisted only of the encoder and the supervised path. Another one, **SVM**, which is similar to the challenge baseline, is a linear SVM trained using exclusively the small amount of labelled data. In addition, we compare our proposed methods with four classic semi-supervised learning algorithms, including **self-training**, **label propagation** [59], **label spreading** [60], and **DAE** [26] in combination with SVMs, for the four semi-supervised tasks. For fair comparison, we follow the same experimental validation procedure. Note that the eight previously reported methods achieved good performance using all the AEC training data.

As can be seen from Table III, both SS-AE and SS-AE-Skip outperform the two supervised methods and the classical semi-supervised learning methods in terms of the average UAR by a large margin. It is worth noting that even when using 100 labelled examples, our proposed models reach a 40.1% UAR and surprisingly surpass the challenge baseline that uses 9 959 labelled examples for training. Further, the best UAR achieved by our proposed methods with only 1 000 labelled examples is 43.6%, which is comparable with the best known UAR (45.6%) obtained by [11] and exceeds the UAR obtained by other modern models (e. g., [12], [47], [56]). In this

TABLE IV
CLASSIFICATION PERFORMANCE OF THE **SS-AE** AND **SS-AE-SKIP** FEATURES + SVM ON THE SPEECH EMOTION RECOGNITION BENCHMARK DATABASE AEC

Feature Type	# of labelled examples from AEC			
	100	200	500	1 000
Sparse-AE+SVM	32.8	35.5	38.8	39.8
WTO-AE+SVM	34.8	38.6	40.5	40.6
SS-AE/SS-AE-Skip+SVM	41.1/41.1	42.1/42.5	43.6/42.8	44.3/44.3

Two representative autoencoders (e. g., Sparse Autoencoders (Sparse-AE) and Winner-Take-All Autoencoders (WTO-AE) [61]) are considered for comparison. The best UAR over ten trials is presented.

scenario, our proposed methods have a statistical significance test at the 0.01 level, when compared to the best performance system in the challenge (i. e., GMM [47]). Although the SS-AE method almost performs as well as the SS-AE-Skip method, it seems that skip connections indeed provide a beneficial effect on the information flow in SS-AE-Skip, resulting in better performance than SS-AE particularly for the very small labelled sets.

In addition to the four semi-supervised tasks mentioned above, we continue to investigate the impact of increasing the number of labelled examples on the proposed method. Here, we increase the number of labelled examples to either 1 500 or 2 000. With 1 500 labelled training examples available, SS-AE and SS-AE-Skip reach an average UAR of 42.3% and 42.5%, respectively, which consistently provides a noticeable boost in performance. For the 2 000 labelled training examples setting, SS-AE and SS-AE-Skip obtain an average UAR of 42.7% and 42.8%, respectively. These results suggest that, our proposed methods profits greatly from the labelled training examples while the performance improvement gradually levels off as the increase in the labelled training examples.

B. Feature Learning

As mentioned earlier (see Section II), a major advantage of autoencoders is the ability to learn useful non-linear features. Hence, here, we evaluate the quality of features learnt by the proposed autoencoders by training a separate linear classifier (i.e., SVM) on top of them. Table IV shows the classification performance of SS-AE and SS-AE-Skip where only N labels are available. We compare our method with the performance of two representative autoencoders (e. g., Sparse Autoencoders (Sparse-AE) and Winner-Take-All Autoencoders (WTO-AE) [61]), where the feature learning process is only governed by an unsupervised learning objective. For fair comparison, the two autoencoders are trained on the whole training dataset, but the SVM is trained only on the N labelled examples. We can see that, the existing unsupervised autoencoders fail to extract good acoustic representations from the unlabelled data, whereas our SS-AE method can learn salient features and thus achieve a better classification. This strongly supports that the supervised path (see Section III-C) is of direct benefit to improve the feature learning ability of our proposed autoencoders.

TABLE V
IMPACTS OF THE VARIANT OF AUTOENCODERS ON OUR PROPOSED METHODS (**SS-AE/SS-AE-SKIP**)

Feature Type	# of labelled examples from AEC			
	100	200	500	1 000
Sparse-AEs-Based	37.2/37.7	39.6/40.2	40.7/42.0	42.0/43.2
WTO-AEs-Based	32.1/37.9	34.3/40.2	35.1/41.4	38.8/42.8
DAE-Based (default)	39.9/40.1	42.4/43.1	42.6/43.1	43.2/43.6

The best UAR over ten trials on AEC is presented.

TABLE VI
EFFECTS OF THE VARIANT OF STANDARD ACOUSTIC FEATURES AS INPUT TO OUR PROPOSED METHODS (**SS-AE/SS-AE-SKIP**)

Feature Type	# of labelled examples from AEC			
	100	200	500	1 000
IS09 EC (default, 394)	39.9/40.1	**42.4/43.1**	42.6/43.1	43.2/43.6
IS10 LOI (1 582)	39.6/40.0	41.2/43.0	42.3/43.0	43.5/43.7
IS11 SS (4 368)	40.0/40.2	41.9/41.1	**43.2/43.6**	**44.5**/44.0
IS12 ST (6 125)	39.9/**40.4**	40.9/42.4	**43.2**/42.8	44.0/43.9
IS13 EMO (6 373)	**40.9**/40.0	41.7/**43.1**	43.0/43.1	43.9/**44.1**

The best UAR over ten trials on AEC is presented. The number of features is given in parentheses.

C. Impacts of the Variant of Autoencoders

In order to assess the impact of the choice of autoencoders on the semi-supervised performance of SS-AE and SS-AE-Skip, we further consider Sparse-AE and WTO-AE [61] in place of DAE. Table V shows the performance change in autoencoders of SS-AE and SS-AE-Skip on the AEC database. We see all the alternate methods still retain the noticeable performance on the speech emotion recognition benchmark database when very few labelled data are available. In the meanwhile, it is worth noting that, the DAE-based framework generally outperforms other alternatives. For this reason, the DAE-based algorithm is used for the further experiments.

D. Impacts of the Variant of Acoustic Features

It is well known that acoustic features have a big effect on speech emotion recognition systems. Generally speaking, a good acoustic representation can facilitate model learning. Therefore, here, we assess the impact of different acoustic features as input to our proposed method. In addition to the default acoustic features from the INTERSPEECH 2009 Emotion Challenge (referred to as IS09 EC), four alternative feature sets, including Level of Interest in 2010 [62] (referred to as IS10 LOI), Speakers States in 2011 [63] (referred to as IS11 SS), Speaker Traits in 2012 [64] (referred to as IS12 ST), Emotion in 2013 [49] (referred to as IS13 EMO), are used for comparison. Table VI presents the effects of the variant of acoustic features on the performance of our proposed methods for the four semi-supervised tasks.

As can be seen from Table VI, the variant of acoustic features indeed have an effect on the classification performance. For example, SS-AE obtains a 44.5% UAR using the IS11 SS feature set when 1 000 labelled examples are available, which is slightly higher than a 43.2% UAR obtained by our default feature set using IS09 EC. However, the observed effect is not too pronounced. One possible explanation is that our proposed autoencoders naturally learn useful representations from the inputs; in turn, the learning process discovers the intrinsic attributes necessary to solve the emotion recognition. In order to keep in line with the INTERSPEECH 2009 Emotion Challenge, we stick to the IS09 EC feature set for the following experiments.

E. Emotion Recognition With Unlabelled Out-of-Domain Data

In Section V-A, we have evaluated our proposed methods in *matched conditions* where all training data and test data come from one corpus. Here, we further evaluate our proposed methods in two cross-corpus settings based on the absence (*mismatched*) or presence (*semi-matched*) of in-domain data in unlabelled training data, i. e., whether unlabelled training data include some data from the same domain as the test set. In the two cross-corpus settings, the domain mismatch issue immediately emerges, which tends to considerably degrade the performance of conventional methods [8]. Here, we focus exclusively on our proposed methods for the four semi-supervised learning tasks presented in Section V-A on the AEC dataset. To avoid expensive computation for hyper-parameter optimisation by grid search, the best architecture previously found in each semi-supervised task is borrowed to be trained in cross-corpus settings.

Let us first consider the mismatched setting where unlabelled data are only chosen from ABC, EMO, and SUSAS, or the combinations of them while labelled data are chosen from the AEC training set. It appears that all these unlabelled data are significantly different from the AEC dataset, as shown in Table I. Table VII presents the experimental results for SS-AE and SS-AE-Skip. It is observed that different selection of unlabelled data has a strong impact on the recognition performance of our proposed methods. Also, as expected, the disparity between the labelled training data and unlabelled training data causes a decrease in UAR obtained by SS-AE and SS-AE-Skip, the decrease is negligible.

Encouraged by the above good results, we further evaluate our methods under semi-matched conditions. Here, unlabelled training data come from a combination of in-domain and out-of-domain data. In this setting, unlabelled training data consist of a partition of the AEC training data and a mixed partition of ABC, EMO, and SUSAS. Fig. 2 depicts the results achieved by SS-AE and SS-AE-Skip under the semi-matched conditions as well as the results under matched conditions. Although under semi-matched conditions the whole training data are increasingly augmented by including other corpora when compared with matched conditions, such augmentations fail to result in the substantial performance improvement and sometimes even hurt the performance. One possible explanation is that simply augmenting unlabelled training data with other out-of-domain data for semi-supervised learning could cause a domain mismatch issue, which hurts classification performance. Regardless of the differences caused by the domain mismatch problem, however, we still observe that our proposed methods retain impressive performance particularly when 1 000 labelled examples are available. This indicates that our proposed methods are highly efficient in learning from data and can naturally prevent the harmful effects of the domain mismatch problem.

TABLE VII
BEST UAR OVER TEN TRIALS OBTAINED BY OUR PROPOSED METHODS (SS-AE AND SS-AE-SKIP) ON THE AEC TEST SET WITH LABELLED DATA FROM THE AEC TRAINING SET AND UNLABELLED DATA FROM ABC, EMO, AND SUSAS WITH SEVERAL COMBINATIONS

Unlabelled data			*# of labelled examples from AEC*			
ABC	EMO	SUSAS	100	200	500	1 000
SS-AE:						
+			36.1	38.0	39.2	41.2
	+		38.6	41.5	42.4	43.2
		+	35.2	36.4	37.6	39.7
+	+		38.6	42.2	42.2	43.3
+		+	35.4	38.0	40.4	40.5
	+	+	39.4	41.0	42.6	43.2
+	+	+	38.8	42.0	41.7	42.1
Mean			37.4	39.9	40.9	41.9
SS-AE-Skip:						
+			37.2	39.4	41.4	40.8
	+		39.7	41.6	43.3	42.9
		+	38.0	38.0	41.7	42.2
+	+		39.5	41.6	43.1	42.8
+		+	35.5	39.0	41.3	41.8
	+	+	39.2	40.7	42.9	43.6
+	+	+	38.5	41.1	42.6	42.7
Mean			38.2	40.2	42.3	42.4

VI. EXPERIMENTAL RESULTS ON GEWEC

By now, we have shown that our proposed method is applicable for general speech emotion problems, where all the training data are the normal phonated speech. In this section, we further exemplify our proposed semi-supervised learning methods to the domain mismatch problem introduced by GeWEC (cf. Section IV-A), which contains the normal phonated speech mode data and whispered speech data. In this setting, we train a model on the whispered speech data while testing on the normal speech data from GeWEC. This task is challenging because of the fundamental differences between normal phonated speech and whispered speech in vocal excitation and the vocal tract transfer function. As a result, it is implausible to train a good model with whispered speech for normal phonated speech. Previous work suggested that using transfer learning methods, such as **uLSIF** [56] and **DAE** [8], can reduce the existing differences among the training and test set.

In this work, we propose to use our semi-supervised learning to exploit other emotional corpora such as ABC or EMO in hope to compensate for the lack of the prior knowledge of normal phonated speech in the training data. In these experiments, three normal phonated speech mode databases, ABC, EMO, and SUSAS, are used as unlabelled data. Labelled training data are only from a subset of the GeWEC whispered data. Note that the remaining whispered data are excluded in the training phase. In total, we consider five semi-supervised learning tasks with 50, 100, 200, 500, 640 (all whispered data) labelled training

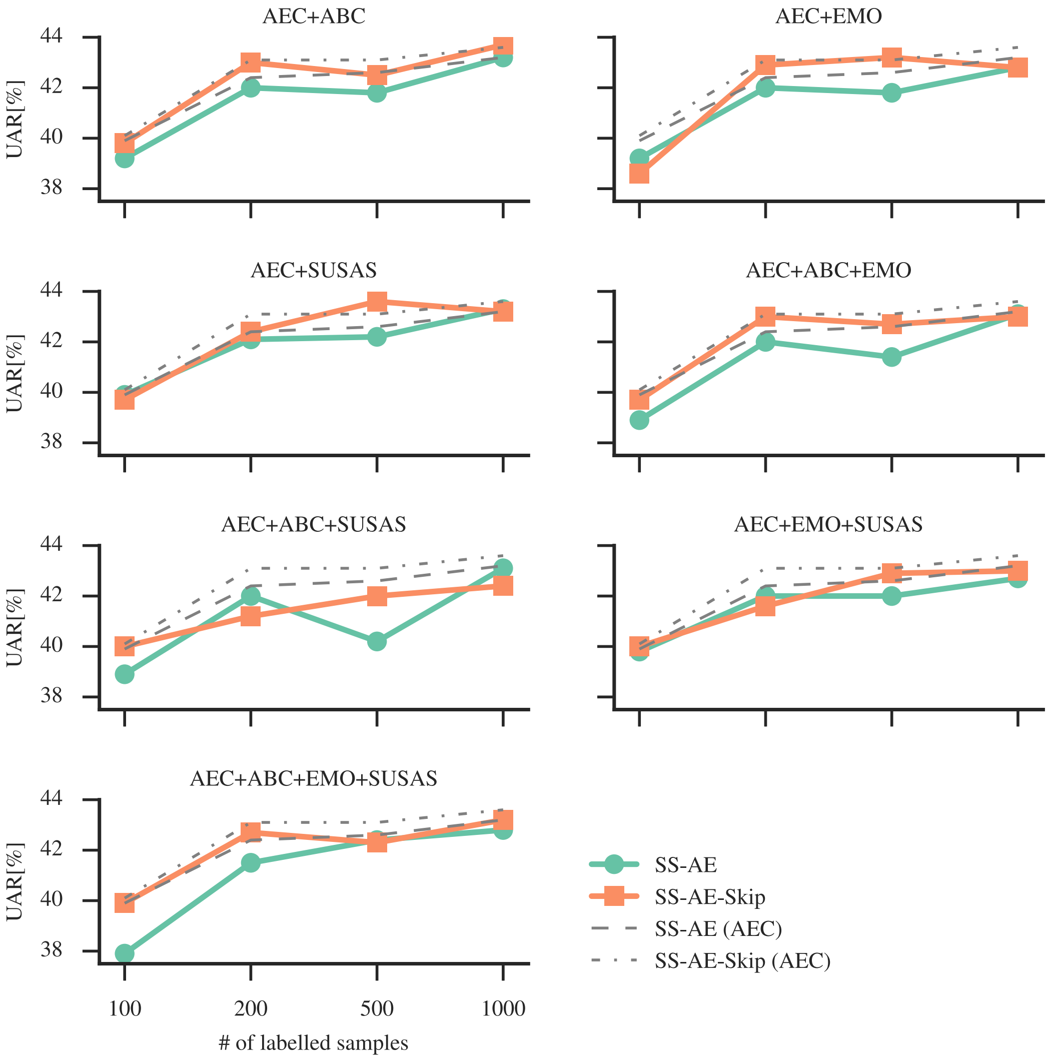


Fig. 2. Best UAR over ten trials obtained by our proposed methods under *semi-matched* and *matched conditions* with labelled data from the AEC training set. Semi-matched conditions: (**SS-AE** and **SS-AE-Skip**) on the AEC test set with unlabelled data from various combinations of the AEC training set, ABC, EMO, and SUSAS. Matched conditions: **SS-AE (AEC)** and **SS-AE-Skip (AEC)** are trained with unlabelled data from the AEC training set.

data. The experimental setup shown in Section IV-C is adopted. Since the GeWEC dataset is quite small, we apply five-fold validation to hyper-parameter tuning. Each experiment is repeated 10 times with different seeds for parameter installation and data selection.

In comparison with the state-of-the-art models, three modern methods, including a supervised leaning method, SVM, two transfer learning methods, DAE and uLSIF, are considered. In the experiments, they are trained on the whole labelled whispered examples while tested on the normal speech data.

Fig. 3 presents all the experimental results for ABC, EMO, and SUSAS with several combinations. The previous reported UAR obtained by using the Modified Group Delay (**MGD**) features with SVMs [48] is given Fig. 3 as well. As shown in Fig. 3, SVM and MGD achieve a 53.4% UAR and a 54.8% UAR while uLSIF and DAE achieve an average UAR of $49.2 \pm 0.1\%$ and $53.1 \pm 1.7\%$, respectively. On the other hand, our proposed semi-supervised learning methods, SS-AE and SS-AE-Skip, have comparable performance to other methods when only 50 labelled examples are available. This suggests that our proposed methods have the powerful capability to incorporate the prior knowledge of the unlabelled normal speech phonated data into learning, thus improving the recognition performance. Besides, SS-AE and SS-AE-Skip also consistently benefit from the increase in the number of the labelled data. When the whole whispered data (i. e., 640 examples) are used as labelled training data and the EMO data servers as unlabelled training data, SS-AE-Skip gets the best average UAR of $63.6 \pm 1.4\%$, which yields 8.8% absolute improvement when compared to MGD. This improvement has a high statistical significance at the 0.001 level. It can also be found that SS-AE-Skip generally outperforms SS-AE in performance, emphasising the benefit of skip connections once again.

VII. Discussion

Despite remarkable advances in speech emotion recognition, the ability of previous emotion recognition engines to deliver

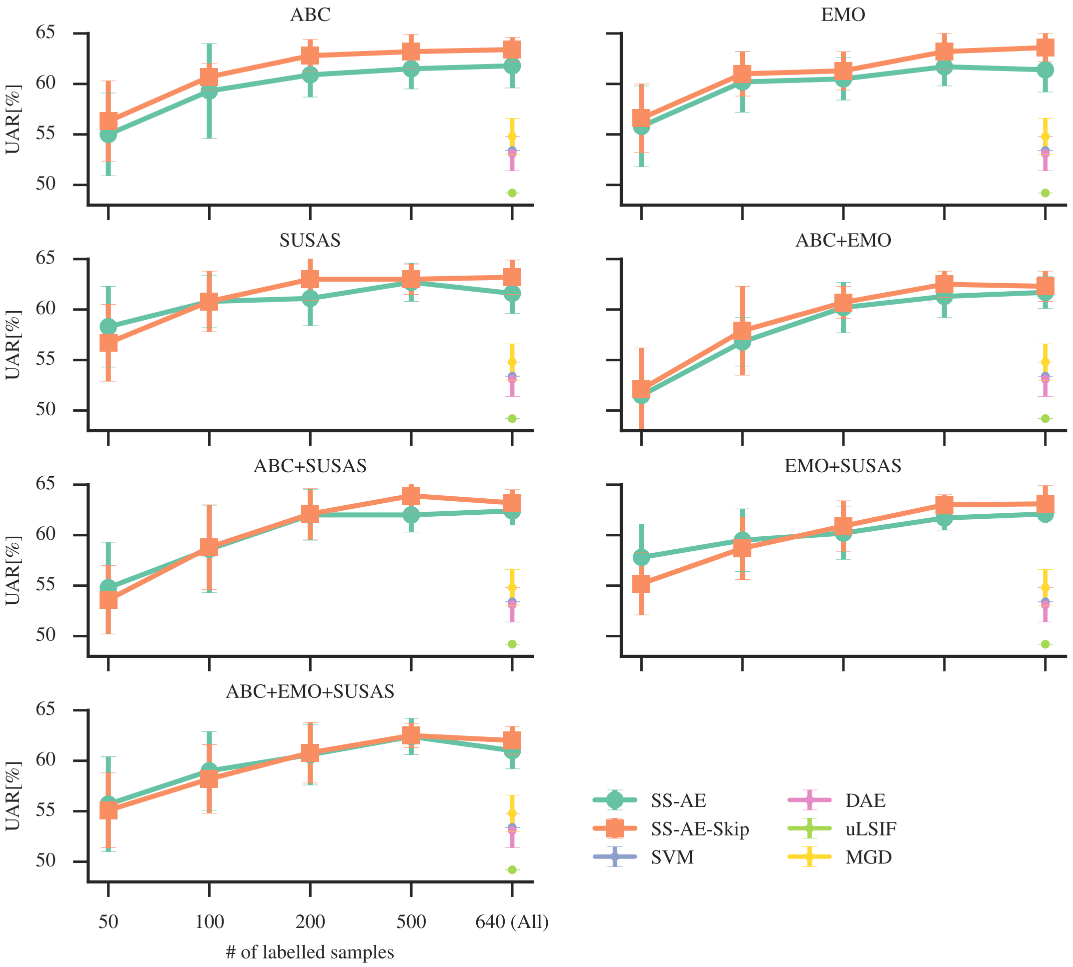


Fig. 3. Average UAR with standard deviation over ten trials obtained by our proposed methods (**SS-AE** and **SS-AE-Skip**) on the GeWEC normal speech phonated data. The training set used comprises labelled data from the GeWEC whispered speech data and unlabelled data from various combinations of ABC, EMO, and SUSAS. We compare our methods with results from the supervised learning method, **SVM**, two transfer learning methods, **DAE** and **uLSIF**, and the **MGD** method [48].

good performance comes at the cost of the large number of labelled speech data, even in state-of-the-art systems built from deep learning approaches. Acquiring a lot of labelled speech data is a tedious and time-consuming process that prevents speech emotion recognition from embracing the vast amount of data from the Internet. For this reason, we have striven to perform semi-supervised learning for speech emotion recognition, aiming to open up the possibility of leveraging unlabelled data. In Sections V and VI, we have shown that our proposed method using a few labelled data and unlabelled data achieve competitive results with state-of-the-art supervised learning approaches. Furthermore, we have extensively tested the applicability of our proposed deep semi-supervised learning framework in various situations, ranging from different number of labelled data (see Section V-A), through different speech data obtained from different devices and varied recording conditions (see Section V-E), and different base autoencoders (see Section V-C), to different acoustic features (see Section V-D). The present method performs favourably in each experiment, substantiating the claim that the proposed framework increases the robustness to complex variations among the unlabelled and labelled speech emotional data (see Section III-D).

VIII. Conclusions and Outlook

Unlike previous research focusing on unsupervised learning with autoencoders for speech emotion recognition, this paper focuses on semi-supervised learning with autoencoders. Specifically, we put the considerable emphasis on combining generative and discriminative training, by presenting semi-supervised learning algorithms tailored to settings where unlabelled data are available. The proposed methods have been systematically evaluated with five databases in various settings. The experimental results demonstrate that the proposed methods clearly improve recognition performance by learning the prior knowledge from unlabelled data in situations with a small number of labelled examples. Furthermore, the proposed methods can overcome the difficulties in mismatched settings and incorporate the knowledge from several different domains into the classifiers, eventually resulting in state-of-the-art performance. This indicates that the present model is capable of making good use of the combination of labelled and unlabelled data for speech emotion recognition.

More recently, the residual neural network showed that very deep architectures make the classifier advantageous to extract complex structure in image processing [45]. Thus a future line of research will be to construct very deep semi-supervised learning algorithms for speech emotion recognition. Other future work includes to study how to extend our proposed semi-supervised autoencoders to Recurrent Neural Networks (RNNs), such as Long Short-Term Memory RNNs [65]. This will benefit from the current powerful RNN model for effective sequence learning.

References

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

[2] J. B. Walther and K. P. D'Addario, "The impacts of emoticons on message interpretation in computer-mediated communication," *Social Sci. Comput. Rev.*, vol. 19, no. 3, pp. 324–347, 2001.

[3] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *Artif. Intell. Rev.*, pp. 1–23, 2012.

[4] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[5] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Hoboken, NJ, USA: Wiley, Nov. 2013.

[6] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, 2003.

[7] M. E. Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using Gaussian mixture vector autoregressive models," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, USA, 2007, pp. 957–960.

[8] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.,* vol. 21, no. 9, pp. 1068–1072, Sep. 2014.

[9] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.,* vol. 29, no. 6, pp. 82–97, Nov. 2012.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[11] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden Markov models with deep belief networks," in *Proc. Autom. Speech Recognit. Understanding*, Olomouc, Czech Republic, 2013, pp. 216–221.

[12] W. Xue, Z. Huang, X. Luo, and Q. Mao, "Learning speech emotion features by joint disentangling-discrimination," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, Xi'an, China, 2015, pp. 374–379.

[13] G. Keren and B. Schuller, "Convolutional RNN: An enhanced model for extracting features from sequential data," in *Proc. Int. Joint Conf. Neural Netw.*, Vancouver, BC, Canada, 2016, pp. 3412–3419.

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

[15] O. Chapelle, B. Schlkopf, and A. Zien, *Semi-Supervised Learning* (Adaptive computation and machine learning). Cambridge, MA, USA: MIT Press, Sep. 2006.

[16] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. Affect. Comput. Intell. Interaction*, Geneva, Switzerland, 2013, pp. 511–516.

[17] Z. Zhang, E. Coutinho, J. Deng, and B. W. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 115–126, Jan. 2015.

[18] Z. Zhang, J. Deng, E. Marchi, and B. W. Schuller, "Active learning by label uncertainty for acoustic emotion recognition," in *Proc. Interspeech*, Lyon, France, 2013, pp. 2856–2860.

[19] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2015, pp. 3546–3554.

[20] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in *Proc. Int. Conf. Neural Inf. Process. Syst*, Montreal, QC, Canada, 2015, pp. 3061–3069.

[21] C.-L. Liu, W.-H. Hsaio, C.-H. Lee, T.-H. Chang, and T.-H. Kuo, "Semi-supervised text classification with universum learning," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 462–473, Feb. 2016.

[22] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[24] R. Salakhutdinov and G. E. Hinton, "Deep Boltzmann machines," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2009, pp. 448–455.

[25] I. Goodfellow, M. Mirza, A. Courville, and Y. Bengio, "Multi-prediction deep Boltzmann machines," in *Proc. Int. Conf. Neural Inf. Process. Syst*, Vancouver, BC, Canada, 2013, pp. 548–556.

[26] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. Int. Conf. Mach. Learn.*, Helsinki, Finland, 2008, pp. 1096–1103.

[27] J. Weston, F. Ratle, H. Mobahi, and R. Collobert, "Deep learning via semi-supervised embedding," in *Neural Networks: Tricks of the Trade*" 2nd ed. New York, NY, USA: Springer, 2012, pp. 639–655.

[28] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 3581–3589.

[29] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[30] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[31] J. Esparza, S. Scherer, and F. Schwenker, "Studying self-and active-training methods for multi-feature set emotion recognition," in *Partially Supervised Learning*. New York, NY, USA: Springer, 2011, pp. 19–31.

[32] Y. Zhang, E. Coutinho, B. Schuller, Z. Zhang, B. Schuller, and M. Adam, "On rater reliability and agreement based dynamic active learning," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, Xi'an, China, 2015, pp. 70–76.

[33] D. Le and E. M. Provost, "Data selection for acoustic emotion recognition: Analyzing and comparing utterance and sub-utterance selection strategies," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, Xi'an, China, 2015, pp. 146–152.

[34] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. Annu. Conf. Comput. Learn. Theory*, 1998, pp. 92–100.

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

[36] G. Druck, C. Pal, A. McCallum, and X. Zhu, "Semi-supervised classification with hybrid generative/discriminative methods," in *Proc. SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Jose, CA, USA, 2007, pp. 280–289.

[37] X. Zhu, A. B. Goldberg, and T. Khot, "Some new directions in graph-based semi-supervised learning," in *Proc. IEEE Int. Conf. Multimedia Expo.*, New York, NY, USA, 2009, pp. 1504–1507.

[38] J. Deng, Z. Zhang, and B. Schuller, "Linked source and target domain subspace feature transfer learning—Exemplified by speech emotion recognition," in *Proc. Int. Conf. Pattern Recognit.*, Stockholm, Sweden, 2014, pp. 761–766.

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

[40] H. Larochelle and Y. Bengio, "Classification using discriminative restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, Helsinki, Finland, 2008, pp. 536–543.

[41] H. Larochelle, M. I. Mandel, R. Pascanu, and Y. Bengio, "Learning algorithms for the classification restricted boltzmann machine," *J. Mach. Learn. Res.*, vol. 13, pp. 643–669, 2012.

[42] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, Haifa, Israel, 2010, pp. 807–814.

[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, Lille, France, 2015, pp. 448–456.

[44] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2015, pp. 2377–2385.

[45] J. Deng, X. Xu, Z. Zhang, S. Frühholz, D. Grandjean, and B. Schuller, "Fisher kernels on phase-based features for speech emotion recognition," in *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, K. Jokinen, and G. Wilcock, Eds. Singapore: Springer, 2017, pp. 195–203.

[46] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 312–315.

[47] M. Kockmann, L. Burget, and J. Černocký, "Brno university of technology system for Interspeech 2009 emotion challenge," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 348–351.

[48] J. Deng, X. Xu, Z. Zhang, S. Frühholz, D. Grandjean, and B. Schuller, "Fisher kernels on phase-based features for speech emotion recognition," in *Proc. Int. Workshop Spoken Dialogue Syst.*, Saariselkä, Finland, Jan. 2016.

[49] B. Schuller *et al.*, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*, Lyon, France, 2013, pp. 148–152.

[50] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Honolulu, HI, 2007, pp. 733–736.

[51] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 1517–1520.

[52] J. Hansen and S. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 1743–46.

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

[54] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. ACM Int. Conf. Multimedia*, Barcelona, Spain, 2013, pp. 835–838.

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

[56] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: Compensating for covariate shift," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1458–1468, Jul. 2013.

[57] H. Cao, R. Verma, and A. Nenkova, "Combining ranking and classification to improve emotion recognition in spontaneous speech," in *Proc. Interspeech*, Portland, OR, USA, 2012, pp. 358–361.

[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Pattern Recog*nit, San Diego, CA, USA, 2015.

[59] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Carnegie Mellon University, Pittsburgh, PA, USA, Tech. Rep. CMU-CALD-02-107, 2002.

[60] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2003, pp. 321–328.

[61] A. Makhzani and B. J. Frey, "Winner-take-all autoencoders," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2015, pp. 2791–2799.

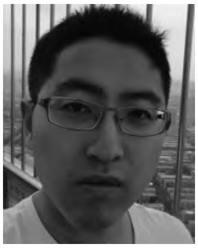
[62] B. Schuller *et al.*, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 2794–2797.

[63] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 3201–3204.

[64] B. Schuller *et al.*, "The INTERSPEECH 2012 speaker trait challenge," in *Proc. Interspeech*, Portland, OR, USA, 2012, pp. 254–257.

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

Jun Deng received the Bachelor's degree in electronic and information engineering from Harbin Engineering University, Harbin, China, and the Master's degree in information and communication engineering from Harbin Institute of Technology, Harbin, China, and the Doctoral degree for his study on Feature Transfer Learning for Speech Emotion Recognition, in electrical engineering and information technology from Technische Universität München, Munich, Germany, in 2009, 2011, and 2016, respectively. From 2015 to 2017, he was a Postdoctoral Researcher in the Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany. He is currently a Leader Researcher at audEERING, where he focuses on conducting fundamental research toward design and development of cutting edge technology for wide range of affective computing applications. His research interests include machine learning methods such as transfer learning and deep learning with an application preference to affective computing.



Xinzhou Xu received the Bachelor's degree from Nanjing University of Posts and Telecommunications, Nanjing, China, and the Master's degree from Southeast University, Nanjing, China, in 2009 and 2012, respectively. He is currently working toward the Ph.D. degree from Southeast University, Dhaka, Bangladesh, and he is also with the Machine Intelligence & Signal Processing Group, MMK, Technische Universität München Munich, Germany, and the Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany. His research interests include spoken signal processing, pattern recognition, machine learning, and affective computing.

Zixing Zhang received the Master's degree in physical electronics from Beijing University of Posts and Telecommunication, Beijing, China, 2010, and the Ph.D. degree in engineering from the Institute for Human-Machine Communication at Technische Universität München, Munich, Germany, 2015. He is currently a Postdoctoral Researcher at the University of Passau, Passau, Germany. Until now, he has authored more than fifty publications in peer-reviewed journals and conference proceedings. His research interests include in deep learning, semisupervised learning, active learning, and multitask learning, in the applications of computational paralinguistics (e. g., emotion recognition) and robust automatic speech recognition.

Sascha Frühholz received the Graduate degree in science of education in 2001 and in psychology in 2006, and the Ph.D. degree in the neural mechanisms of facial expressions from Bremen University, Bremen, Germany, in 2008. He is currently a SNSF Professor with the Department of Psychology, University of Zurich, Zurich, Switzerland. His current projects deal with dynamic connectivity patterns of local and remote brain regions during emotional prosody processing using high-resolution brain scans and specific connectivity modeling approaches for functional imaging data.

Björn Schuller (M'05–SM'15) received the Diploma in 1999, the Doctoral degree for the study on automatic speech and emotion recognition in 2006, and the Habilitation and Adjunct Teaching Professorship in the subject area of signal processing and machine intelligence in 2012, all in electrical engineering and information technology from Technical University of Munich, Munich, Germany. He is Reader of machine learning in the Department of Computing, Imperial College London, London, U.K., a Full Professor and head of the Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany, where he previously headed the Chair of Sensor Systems in 2013, and an Associate of the Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland. He is the President of the Association for the Advancement of Affective Computing, elected member of the IEEE Speech and Language Processing Technical Committee, and a member of the ACM and ISCA and (co-)authored five books and more than 600 publications in peer reviewed books, journals, and conference proceedings leading to more than 16 000 citations (h-index = 61).