

# Survey of Deep Representation Learning for Speech Emotion Recognition

Siddique Latif<sup>1</sup>, Rajib Rana<sup>2</sup>, Sara Khalifa<sup>3</sup>, Raja Jurdak<sup>4</sup>, *Senior Member, IEEE*,  
Junaid Qadir<sup>5</sup>, *Senior Member, IEEE*, and Björn Schuller<sup>6</sup>, *Fellow, IEEE*

**Abstract**—Traditionally, speech emotion recognition (SER) research has relied on manually handcrafted acoustic features using feature engineering. However, the design of handcrafted features for complex SER tasks requires significant manual effort, which impedes generalisability and slows the pace of innovation. This has motivated the adoption of representation learning techniques that can automatically learn an intermediate representation of the input signal without any manual feature engineering. Representation learning has led to improved SER performance and enabled rapid innovation. Its effectiveness has further increased with advances in deep learning (DL), which has facilitated *deep representation learning* where hierarchical representations are automatically learned in a data-driven manner. This article presents the first comprehensive survey on the important topic of deep representation learning for SER. We highlight various techniques, related challenges and identify important future areas of research. Our survey bridges the gap in the literature since existing surveys either focus on SER with hand-engineered features or representation learning in the general setting without focusing on SER.

**Index Terms**—Speech emotion recognition, multi task learning, representation learning, domain adaptation, unsupervised learning

## 1 INTRODUCTION

SPEECH is a natural mode of communication among humans. It conveys affective information about emotional expression through explicit (linguistic) and implicit (paralinguistic) cues. Studies report that linguistic messages are rather unreliable means to predict and analyse human affective behaviour [1] because linguistic content is language-dependent, and the generalisation of emotions for multiple languages is very difficult to achieve. People often choose different words to express emotion, making it hard to anticipate a speaker's word choice and the associated affective expressions. The paralinguistic content of speech,

on the other hand, provides an immense body of acoustic features that can be used to encode the emotional state of the speaker. These acoustic features are reliable indicators of basic emotions and have been explored by different machine learning (ML) [2], [3], [4] as well as deep learning (DL) models [5], [6], [7], [8] for speech emotion recognition (SER).

Traditionally, the efficiency of ML algorithms in SER has been critically dependent on the quality of hand-crafted acoustic features. Consequently, feature engineering, which focuses on creating features from raw speech, has been an important part of SER research for a long time. Deep representation learning encompasses DL techniques to learn representations of input data, usually through the non-linear transformation of the input data. Researchers have evaluated different DL models for representation learning in SER. Fig. 1 presents an overview of deep representation learning in SER, showing that DL models can learn emotionally salient representations from raw speech as well as from acoustic features. These deep models can be trained in different ways, including supervised, unsupervised, semi-supervised, and transfer learning techniques to learn emotional representation from speech. This review covers all these deep representation learning techniques for SER proposed to date.

We compare our paper with the existing surveys on deep learning or representation learning for SER in Table 1. The comparison shown demonstrates the uniqueness of our paper. The article by Bengio *et al.* [9] focus on the geometrical connections between representation learning, manifold learning, and density estimation. Since this article was published in 2013, it predated the development of modern generative models, and the discussions focus mostly on traditional techniques such as principal component analysis (PCA) [14],

- Siddique Latif is with the University of Southern Queensland (USQ), Springfield, 4300, QLD, and also with the Distributed Sensing Systems Group, Data61–CSIRO, Pullenvale, QLD, 4069, Australia. E-mail: siddique.latif@usq.edu.au.
- Rajib Rana is with the University of Southern Queensland (USQ), Springfield 4300, QLD, Australia. E-mail: rajib.rana@usq.edu.au.
- Sara Khalifa is with the Distributed Sensing Systems Group, Data61–CSIRO, Pullenvale, QLD 4069, Australia, and with the University of New South Wales, Sydney, NSW 2052, Australia, and also with the University of Queensland, St Lucia, QLD 4072, Australia. E-mail: sara.khalifa@data61.csiro.au.
- Raja Jurdak is with the Distributed Systems and Applied Data Sciences, Queensland University of Technology, Brisbane 4000, Queensland, Australia. E-mail: r.jurdak@qut.edu.au.
- Junaid Qadir is with Department of Computer Science and Engineering, College of Engineering, Qatar University. E-mail: jqadir@qu.edu.qa.
- Björn Schuller is with the GLAM – The Group on Language, Audio, and Music, Imperial College London, London SW7 2BX, U.K., and also with the Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany. E-mail: bjoern.schuller@imperial.ac.uk.

Manuscript received 10 May 2021; revised 16 Sept. 2021; accepted 19 Sept. 2021.

Date of publication 21 Sept. 2021; date of current version 31 May 2023.

(Corresponding author: Siddique Latif.)

Recommended for acceptance by F. Metze.

Digital Object Identifier no. 10.1109/TAFFC.2021.3114365

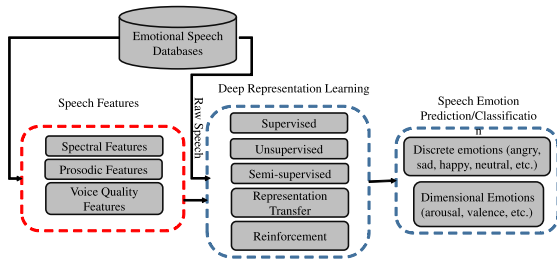


Fig. 1. An overview of deep representation learning-based SER system.

restricted Boltzmann machines (RBMs), and autoencoders (AEs). Research on representation learning has since evolved significantly with generative models like variational autoencoders (VAEs) [15], and generative adversarial networks (GANs) [16] demonstrating superior performance in representation learning compared to autoencoders and other classical methods [17]. Furthermore, we find surveys that focus on different emotional datasets and handcrafted features and classification networks do not focus on deep representation learning for SER.

We consider multiple databases to find the relevant literature, including IEEE Xplore, Springer, Elsevier, and Google Scholar. We searched articles using related keywords: “representation learning”, “feature learning”, and “feature extraction”. Many studies were also found in the bibliographies of reviewed papers and were included. In general, studies in this review use DL for representation or feature learning/extraction on publicly available datasets.

The major contribution of our paper is that we cover a comprehensive survey that bridges the gaps in the existing literature. More specifically, we focus on (1) the importance of deep representation learning for SER; (2) the popular DL models and their representation learning abilities; and (3) the various representation learning techniques used for SER in the literature. We further highlight the challenges of deep representation learning in SER and conclude this paper by discussing the findings of our review and by identifying future research directions.

The remainder of the paper is organised as follows. We provide a discussion on the relevant background and concepts in Section 2. The use of deep representation learning for SER is discussed in Section 3. The challenges of deep representation learning for SER are discussed in Section 4. Discussions and future directions follows in Section 5. Finally, the paper is concluded in Section 6.

## 2 BACKGROUND AND CONCEPTS

Representation learning has become a rich research discipline in the ML community. In SER, representation learning can use raw speech as well as speech features to learn emotionally discriminating representations for emotion classification or prediction (as highlighted in Fig. 1). This section briefly discusses various important concepts related to representation learning in SER.

### 2.1 Representation Learning Versus Feature Engineering

The manual design of a conversion of the speech signal into meaningful information and a reasonably limited number of attributes using domain knowledge is called (speech) feature engineering. In SER, feature engineering and designing ML models for classification or prediction are often considered separate problems. Most of the actual SER research has focused on feature engineering or the design of pre-processing data transformation pipelines to craft emotional representations that support ML algorithms. Although feature engineering techniques can help improve the SER performance, the downside is that these techniques are labour-intensive and time-consuming. For decades, Mel frequency cepstral coefficients (MFCCs) [18] has been used as the principal set of features for SER and other speech analysis tasks. The four steps involved in the extraction of MFCCs are: (1) computation of the Fourier transform, (2) projection of the powers of the spectrum onto the Mel scale, (3) taking the logarithm of the Mel frequencies, and (4) applying discrete cosine transformation (DCT) or other suited transformations for compressed representations. It is found that the

TABLE 1  
Comparison of Our Paper With That of the Existing Surveys

Paper	Representation Learning	Focus				Details
		Speech Emotion Recognition			Deep Learning	
		Datasets	Hand Engineered Feature	Deep Representation Learning		
Bengio et al. [9] 2013	✓	✗	✗	✗	✓	This paper reviews the work in the area of unsupervised feature learning and deep learning. It also covers advancements in probabilistic models and autoencoders. It does not include recent models like VAE and GANs.
Zhong et al. [10] 2016	✓	✗	✗	✗	✓	This paper covers the history of data representation learning from traditional to recent DL methods. Challenges for deep representation learning, recent advancement, and future directions are not covered.
Basu et al. [11] 2017	✗	✓	✓	✗	✓	This paper focuses on the challenges of choosing emotional corpora and identification of different hand crafted features for classification model.
Swain et al. [12] 2018	✗	✓	✓	✗	✓	This paper reviews the literature on various databases, handcrafted features, and classifiers for the SER system.
Akccay et al. [13] 2020	✗	✓	✓	✗	✓	This paper focuses on emotional corpora, preprocessing techniques for handcrafted features, supporting modalities and emotion classifiers.
Our paper	✓	✓	✓	✓	✓	Our paper covers deep different representation learning techniques from emotional speech in comparison traditional methods and handcrafted features, covers poplar emotional corpora, DL models, discusses different challenges, highlights future directions.

TABLE 2  
Comparing Feature Engineering and Representation Learning

	Automated	Human Independence	Generalisability
Feature Engineering	×	×	×
Representation Learning	✓	✓	✓

last step loses information and destroys spatial relations; therefore, it is usually omitted, which results in the LogMel spectrum, a popular feature used by the speech community. It is also the most popular feature to train DL networks in the speech domain. Minimalist feature sets like GeMAPs and eGeMAPs [19] are also widely used (e. g., [20], [21]) as benchmarks. They are designed/engineered to (a) index affective physiological changes in voice production, and (b) achieve automatic extractability [19].

On the other hand, representation learning is the technique of learning representations, usually through the automatic transformation of the input data. It comes under the header of DL or feature learning. The key goal of representation learning is yielding abstract and useful representations for ML tasks such as classification and prediction. We compare feature engineering with representation learning in Table 2. The comparisons show that representation learning is a less time consuming automatic process and requires minimal human domain knowledge to produce better results than the hand-engineered features. Also, unlike feature engineering, representation learning does not require extra efforts to design features for a new task and have more generalisation ability.

## 2.2 Traditional Versus Deep Representations Learning Techniques

In the field of representation learning, the algorithms are generally categorised into two classes: shallow and deep [22]. Shallow learning algorithms are also considered as traditional methods. They aim to learn transformations of data by extracting useful information. One of the oldest shallow learning algorithms is PCA [14], which has been studied extensively over the last century. Similar to PCA, linear discriminant analysis (LDA) [23] is another shallow learning algorithm. Unlike PCA, LDA is a supervised method that requires class labels to maximise class separability. Other linear feature learning methods include canonical correlation analysis (CCA) [24], multi-dimensional scaling (MDS) [25], and independent component analysis (ICA) [26].

Many methods for nonlinear feature reduction are also proposed to discover the non-linear hidden structure from the high dimensional data [27] including locally linear embedding (LLE) [28], non-negative sparse coding [29], isometric feature mapping (Isomap) [30], t-distributed stochastic neighbour embedding (t-SNE) [31], and neural networks (NNs) [32]. The kernel PCA (KPCA) [33], and generalised discriminant analysis (GDA) [34] are non-linear versions of PCA and LDA, respectively.

The shortcoming of shallow representation learning is that such representations contain only a small number of non-linear operations and are unable to accurately model complex, high-dimensional, and noisy real-world data (such as emotional speech) [35]. The shallow feature learning algorithms have, however, dominated representation

TABLE 3  
Characteristics of DL Model for Representation Learning

Model	Characteristics
DNNs	Good for learning a hierarchy of representations. They can learn invariant and discriminative representations. Features learnt by DNNs are more generalised compared to traditional methods.
CNNs	Good for learning both low-level as well as high-level representation from emotional speech.
RNNs	Good for sequential modelling. They can learn temporal structures from speech suitable for emotion classification.
AEs	Powerful unsupervised representation learning models that encode the emotional speech data in sparse and compressed representations.
VAEs	Stochastic variational inference and learning model. Popular in learning disentangled emotional representations from speech.
GANs	A Game-theoretical framework that is useful for data generation and is robust to overfitting. They can learn disentangled representations that are very suitable for SER.

Here DNN represents fully connected deep networks.

learning until the successful training of deep models for representation learning of data reported by Hinton and Salakhutdinov in 2006 [36]. This work was quickly uptaken by other researchers [37], [38], which led to a large number of deep models suitable for deep representation learning. In DL models (e. g., feed-forward neural networks), all hidden layers except the last layer (i.e., softmax classifier) learn representations, which often leads to much better performance compared to the hand-designed representations [39].

Studies show that deep architectures can learn more complex relationships that greatly help improve performance [22]. The non-linearity in deep models help to learn more robust representations when multiple layers/modules are stacked atop one another. Such robust representation in lower dimensions can be easily transmitted to the communication network for a wide range of real-time applications and services [40].

## 2.3 Deep Learning Models for Representation Learning

This section covers DL models, including feed-forward neural networks (FNNs), autoencoders, and generative models, which have been widely used for emotional representation learning in SER research. We highlight the characteristics of these models in terms of their emotional representation learning abilities in Table 3.

### 2.3.1 Deep Neural Networks (DNNs)

Historically, the idea of fully connected DNNs<sup>1</sup> is an extension of ideas emerging from artificial neural networks (ANNs) [41]. Multilayer perceptrons (MLPs) [42] with multiple hidden layers are indeed a good example of deep architectures. DNNs consist of multiple layers, including an input layer, hidden layers, and an output layer, of processing units called “neurons”. These neurons in each layer are densely connected with the neurons of the adjacent layers.

1. We use DNNs for fully connected deep neural networks.

Each layer of a DNN performs representation learning based on the input provided to it. A well trained DNN learns a hierarchy of distributed representations [43]. Increasing the depth of DNNs enables the learning of a deep hierarchy of representations at different levels of abstraction. DNN with highway connectivity [44] is a good example of such deep architecture. Higher levels of abstract representations generally offer invariance to local changes of the input [9] and are helpful in designing SER systems.

*Convolutional neural networks (CNNs)* [45] are a specialised kind of deep architecture for processing data having a grid-like topology. Examples include image data that have 2D grid pixels and time-series data (i. e., 1D grid). CNNs introduce convolutional, and pooling layers into the structure of DNNs, which take into account the spatial representations of the data and make the network more efficient by introducing sparse interactions, parameter sharing, and equivariant representations [39]. There are many variants of CNNs; however, ResNet [46], and DenseNet [47] are especially popular in SER due to their complex emotional representation learning ability.

In contrast to fully connected DNNs, the training process of CNNs is more straightforward due to fewer parameters [48]. CNNs are powerful at extracting low-level representations at the initial layers, and high-level features (textures and semantics) in the higher layers [49]. The convolution layer in CNNs acts as a data-driven filterbank that is able to capture emotional representations from speech, which are more generalised, discriminative, and contextual [49], [50].

*Recurrent neural networks (RNNs)* [51], [52] introduce recurrent connections within layers to enable parameters sharing across time. They create a memory in the network by using the information from all previous inputs. This makes RNNs have stronger representational memory compared to hidden Markov models (HMMs), whose discrete hidden states bound their memory [53]. Simple RNNs usually fail to model the long-term temporal contingencies due to the vanishing gradient problem. Multiple specialised RNN architectures, including long short-term memory (LSTM) [54] and gated recurrent units (GRUs) [53] address this problem using a gating mechanism to add and forget the information selectively. Bidirectional RNNs [55] also model both past and future contexts by passing the input sequence through two separate recurrent hidden layers. RNNs introduce recurrent connections to allow parameters to be shared across time, making them powerful in learning temporal dynamics from sequential data, e. g., speech. In SER, temporal dynamics modelling using RNNs-Connectionist Temporal Classification (CTC) [56] based models shows improved results.

### 2.3.2 Autoencoders (AEs)

The idea of an autoencoding network [57] is to learn a mapping from high-dimensional data to a lower-dimensional feature space such that the input observations can be approximately reconstructed from the lower-dimensional representation. A function  $f_\theta$  called the encoder maps the input vector  $x$  into feature/representation vector  $h = f_\theta(x)$ . The decoder network is responsible to map a feature vector  $h$  to reconstruct the input vector  $\hat{x} = g_\theta(h)$ . The decoder network parameterises the decoder function  $g_\theta$ . Overall, the

parameters are optimised by minimising the following cost function:

$$\mathcal{L}(x, g_\theta(f_\theta(x))) = \|x - \hat{x}\|_2^2. \quad (1)$$

The set of parameters  $\theta$  of the encoder and decoder networks are simultaneously learned by incurring a minimal reconstruction error. To capture useful representations  $h$ , the cost function of Equation (1) is usually optimised with an additional constraint to prevent the AE from learning the useless identity function having zero reconstruction error. One way of learning useful feature representations  $h$  is to regularise the autoencoder by imposing constraints to have a low dimensional feature size. In this way, the AE is forced to learn the salient representations of data from a high dimensional space to a low dimensional feature space. Below we discuss some other autoencoding networks.

*Sparse autoencoders (AEs)* can discover a useful feature representation with the size larger than the input vector  $x$  [58]. This is done using the sparsity regularisation [38]. Sparseness plays a key role in learning a more meaningful representation of input data [59]. It has been found that sparse AEs are simple to train and can learn better representation compared to denoising autoencoders (DAE) and RBMs [60]. In particular, sparse encoders can learn useful information and attributes from emotional speech, which can facilitate better classification performance [61].

*Denoising autoencoders (DAEs)* are considered as a stochastic version of the basic AE. They are trained to reconstruct a clean input from its corrupted version [62]. The objective function of a DAE is given by:

$$\mathcal{L}(x, g_\theta(f_\theta(\tilde{x}))), \quad (2)$$

where  $\tilde{x}$  is the corrupted version of  $x$ , which is done via stochastic mapping  $\tilde{x} \sim q_D(\tilde{x}|x)$ . During training, DAEs minimise the same reconstruction loss between a clean  $x$  and its reconstruction from  $h$ . The difference is that  $h$  is learnt by applying a deterministic mapping  $f_\theta$  to a corrupted input  $\tilde{x}$ . It thus learns higher level feature representations that are robust to input corruption. Therefore, DAEs are suitable for learning emotional representations from noisy speech [63].

### 2.3.3 Deep Generative Models

*Deep belief network (DBN)* [64] is a powerful probabilistic generative model that consists of multiple layers of stochastic latent variables, where each layer is a restricted Boltzmann machine (RBM) [65]. Boltzmann machine (BM) is a bipartite graph in which visible units are connected to hidden units using undirected connections with weights. A BM is restricted in the sense that there are no hidden-hidden and visible connections. During the training phase, an RBM uses Markov chain Monte Carlo (MCMC)-based algorithms [64] to maximise the log-likelihood of the training data. RBMs are very effective at approximating any distribution. However, training RBMs based on MCMC computes the gradient of the log-likelihood, which poses a significant learning problem [66]. In recent years, generative models like GANs and VAEs have been proposed that can be trained via direct back-propagation and avoid the

TABLE 4  
Review of Different SER Databases

Corpus Name	Language	Speakers	Mode	Type	Emotions	Duration (approx)	Public Access
EMODB [78])	German	10 speakers (5 males, 5 females)	audio	stimulated	anger, boredom, disgust, fear, happiness, sadness, neutral	< 1 hour	yes
MSP-IMPROV [79]	English	12 actors (6 males and 6 females)	audio, video	stimulated	anger, happiness, sadness, neutral	18 hour	yes
MSP-Podcast [80]	English	60 speakers (30 females, 30 males)	audio	naturalistic	arousal, valence, dominance	27 hours	yes
SEMAINE [81]	English	150 participants	audio, video	induced	5 affective dimensions (i. e., valence, activation, power, anticipation/expectation, intensity)	6.2 hours	yes
IEMOCAP [82]	English	5 females, 5 males	audio, video	stimulated	neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excited and other)	12 hours	yes
EMOVO [83]	Italian	6 speakers (3 males, 3 females)	audio	stimulated	disgust, happiness, fear, anger, surprise, sadness, neutral	< 2 hours	yes
RECOLA [84]	French	46 speakers (19 males, 27 females)	audio, video	natural	five social behaviours (dominance, agreement, performance, engagement, rapport); arousal and valence	3.5 hours	Yes
CMU-MOSEI [85]	English	single speaker	multimodal	natural	anger, anxious, disgust, happiness, neutral, sadness, surprise and fear	65 hours	Yes

difficulties of MCMC-based training. We discuss GANs and VAEs in more detail next.

*Generative adversarial networks (GANs)* [16] use adversarial training to directly shape the output distribution of the network via back-propagation. They include two neural networks—a generator,  $G$ , and a discriminator,  $D$ , which play a min-max adversarial game defined by the following optimisation problem:

$$\min_G \max_D \mathbb{E}_x [\log (D(x))] + \mathbb{E}_z [\log (1 - D(G(z)))]. \quad (3)$$

The generator,  $G$ , maps the latent vectors,  $z$ , drawn from some known prior,  $p_z$  (e. g., Gaussian), to fake data points,  $G(z)$ . The discriminator,  $D$ , is tasked to differentiate between generated samples (fake),  $G(z)$ , and real data samples,  $x \in p_{\text{data}}$ . Overall, GAN is trained to achieve a generator network that maximally confuses the discriminator into believing that the samples it generates come from the data distribution.

Benefiting from the flexibility of GAN's framework, adversarial training methodology has been successfully leveraged to many traditional tasks, including unsupervised representation learning. For example, DCGAN [17] uses the intermediate features from the discriminator as the representations of the input images. On the other hand, the input of the generator, i.e., noises, can be viewed as the representations of the output images.

*Variational autoencoders (VAEs)* are probabilistic models that use a stochastic encoder for modelling the posterior distribution  $q(z|x)$ , and a generative network (decoder) that models the conditional Log-likelihood  $\log p(x|z)$ . Both of these networks are jointly trained to maximise the following variational lower bound on the data loglikelihood:

$$\log p(x) > \mathbb{E}_{q(z|x)} \log p(x|z) - \text{KL}(q(z|x)||p(z)). \quad (4)$$

The first term is the standard reconstruction term of an AE, and the second term is the KL divergence between the prior

$p(z)$  and the posterior distribution  $q(z|x)$ . The second term acts as a regularisation term, and without it, the model is simply a standard autoencoder. In contrast to standard AEs, VAEs learn the probability distribution parameters from the input in a latent space by making the latent distribution close to a 'prior' distribution. Due to these characteristics, VAEs are becoming very popular in learning emotional representation from speech [67]. Recently, various variants of VAEs are proposed in the literature, which include  $\beta$ -VAE [68], InfoVAE [69], FactorVAE [70], and many more [71]. All these VAEs are very powerful in learning disentangled, and hierarchical representations and are also popular in clustering multi-category structures of data [71].

## 2.4 Emotional Corpus

Emotional databases can be divided into three types: simulated, elicited (induced), and natural. Simulated emotional speech databases consist of recordings collected from experienced and trained actors or artists. Induced emotional speech databases are collected by generating an emotional situation artificially by involving the speaker in the emotional dialogue or conversational setting. The speaker's reactions towards the emotional situation are potentially recorded without their knowledge after ethical approval. Natural emotional databases are produced by recording emotions from real-world applications such as call centres [72], or patient and doctors conversations, among others. Natural emotional databases may not contain all emotions and also have copyright and privacy issues.

Different emotional datasets are available; however, in this work, we only present the details of the most popular ones in Table 4, which are being utilised for emotional representation learning. In [73], [74], the authors provide further details on speech emotional databases. These emotional datasets are annotated using either categorical [75] or dimensional [76] emotion models. A categorical emotion model considers emotions as discrete classes, whereas a



TABLE 5  
Summary of Deep Representation Learning Techniques Used for SER in Different Studies

Paper (Year)	Technique	Corpus	Input	Model (C/P)	Performance
Cairong et al. [7] (2014)	Supervised Representation Learning	ABC corpus [95]	Spectrogram	DBN (C)	52.2 % (ARR)
Kim et al. [94] (2019)		IEMOCAP	acoustic +lexical features	DNN (C)	61.4 % (UAR)
Fayek et al. [93] (2017)		IEMOCAP	Mel scale features	DNN (C)	58.78 % (UAR)
Bertero et al. [96] (2017)		TED talks [97]	Raw speech	CNN (C)	66.1 % (AAC)
Aldeneh et al. [51] (2017)		IEMOCAP	Mel filterbanks	CNN (C)	61.8 % (UAR)
Lee et al. [8] (2015)		MSP-IMPROV			52.6 % (UAR)
Latif et al. [50] (2019)		IEMOCAP	Acoustic features	BLSTM (C)	63.89 % (UA)
		IEMOCAP MSP-IMPROV	Raw Speech	CNN-LSTM	60.23 % (UAR)
Ghosh et al. [98] (2016)		IEMOCAP	Spectrogram	DAE-BLSTM (C)	51.86% (UA)
Xia et al. [99] (2016)		IEMOCAP	IS10 [100]	DAE-SVM (C)	63.1% (UAR)
Paraskevopoulos et al. [101] (2019)		IEMOCAP	IS10	AE-SVM (C)	57.8 % (UA)
Latif et al. [68] (2018)		IEMOCAP	LogMel features	VAE-LSTM (C)	55.42% (UA)
Eskimez et al. [102] (2018)		IEMOCAP	MFCCs +LLDs and derivatives	AAE-DNN (C)	48.18 % (UAR)
Latif et al. [103] (2020)		IEMOCAP	IS10	GAN-DNN (C)	60.51 % (UAR)
Parthasarathy et al. [104] (2018)	Semi-supervised Representation Learning	MSP-Podcast	IS13 [105]	Ladder networks (P)	aro (0.803 CCC) val (0.458 CCC) dom( 0.746)
Parthasarathy et al. [106] (2020)		MSP-Podcast	IS13	Ladder networks (P)	aro (0.770 CCC) val (0.301 CCC) dom( 0.700)
Tao et al. [107] (2019)		IEMOCAP	IS09 [108]	Ladder networks (C)	59.7 % (UAR)
Latif et al. [109] (2020)		IEMOCAP MSP-IMPROV	Spectrogram	AAE-CNN (C)	66.7% (UA) 60.3%(UA)
Chang et al. [110] (2017)		IEMOCAP	Spectrogram	GAN-DNN (C)	48.88% (UA)
Xia et al. [111] (2017)	Multi-task Representation Learning	IEMOCAP SEMAINE	IS10	DBN (C)	60.5 % (UA) 35.9% (UA)
Lottian et al. [112] (2018)		MSP-Podcast	eGeMAPS	DNN (C)	66.8 % (UA)
Nediyanchath et al. [113] (2020)		IEMOCAP	LogMel	DNN (C)	70.1% (UA)
Tao et al. [114] (2018)		IEMOCAP	IS10	LSTM (C)	55.3% (WA)
Parthasarathy et al. [115] (2017)	Domain Adaptive Representation Learning	GeWEC [116]	IS09	Universum AE (C)	63.3 % (UAR)
Abdelwahab et al. [117] (2018)		IEMOCAP MSP-IMPROV	IS13	DANN (P)	aro (.489 CCC) val (.215 CCC) dom (.401 CCC)
Paraskevopoulos et al. [101] (2019)		IEMOCAP MSP-IMPROV	IS09	DNN (C)	56.2% (UAR) 44.1% (UAR)
Shukla et al. [118] (2021)	Self Supervised Representation Learning	IEMOCAP	Spectrogram	CNN-LSTM(C)	0.615 % (F1)
Siriwardhana et al. [119] (2020)		IEMOCAP	Wav2Vec [120]	Transformer (C)	74.7 (AAC)
Lakomkin et al. [121] (2021)	DRL for Representation Learning	IEMOCAP	MFCC	GRU (C)	84.9% AAC
Chen et al. [122] (2019)		CMU-MOSEI	MFCCs HOG [123]	Multimodal-LSTM (C)	76.5% (AAC)

C: classification, P: prediction, ARR: Average Recognition Rate, AAC: Average Accuracy, IS09-IS13: Interspeech 2009-2013 paralinguistics challenge feature set, LLDs: Low-Level Descriptors, act: activation, val: valence, dom: dominance, aro: arousal.

dimensional emotion model defines emotions as two or more dimensional space characterised mostly by arousal and valence and next frequently dominance.

## 2.5 Evaluation Metrics

In SER, the effectiveness of a deep emotional representation is evaluated by performing classification or regression using these representations as input. For classification, SER systems use a classification score or accuracy as a metric. However, as data is often imbalanced across the classes, in naturalistic emotion corpora, the accuracy is usually used as so-called unweighted accuracy (UA) or unweighted average recall (UAR), representing the average recall across classes, unweighted by the number of instances per classes. This has been introduced by the first challenge in the field—the Interspeech 2009 Emotion Challenge [85] and has since been used by other challenges across the field. SER systems that use deep representation for emotional attributes such as

arousal and valence or dominance prediction commonly optimise regression-based models using the mean squared error (MSE) and concordance correlation coefficient (CCC) as objective functions [86].

## 3 DEEP REPRESENTATION LEARNING IN SER

In this section, we review the existing literature on deep representation learning techniques for SER. The readers are referred to Table 5 for a summary of the reviewed studies in this paper. We present studies while describing their use of corpus, input features, models, and performance in Table 5. Studies are clustered into five major groups depending on the DL techniques employed for representation learning:

### 3.1 Supervised Representation Learning

In supervised representation learning, features are learnt from data samples using their labels. In SER, supervised

representation learning methods are widely used to improve performance. In [6], the authors use a DBN for emotional representation learning from speech and achieve 7% higher classification accuracy (86.5%) on the BUAA emotional corpus compared to the classical hand-engineered features. To improve the SER performance, Cairong *et al.* [7] fuse the classical features with emotional representation learnt by a DBN. They show that the fusion of deep emotional representations learnt by a DBN with classical features can improve SER by 8.8%. A similar fusion of a DBN representation with hand-engineered features was performed in [87] to improve SER performance in noisy conditions. Experiments were performed on the EMODB dataset, and the results show that the proposed approach improves the performance by 5.48%. In [88], the authors perform experiments on multiple datasets and show that DBNs can learn more powerful and effective discriminative long-range features that help improve SER performance. Similar to DBNs, researchers also explored DNNs with multiple fully connected hidden layers for emotional representation learning.

*Deep neural networks* (DNNs) are popular in learning high-level discriminative emotional representations. Hen *et al.* [89] use DNNs for high-level emotional representation learning from raw speech. They construct an utterance-level representation from a segment level probability distribution produced by a DNN and use extreme learning machines (ELMs) to perform emotion classification on these utterance level representations. They evaluate the proposed framework on the IEMOCAP data and show that the proposed approach effectively captures emotional representation and leads to 20% classification improvement. In [90], the authors attempt to learn a discriminative emotional representation in compressed size to facilitate fast classification. The results show that DNNs can capture a hidden emotional representation that leads to significant improvement in SER performance. Various other studies [91], [92], [93] also explore DNNs for emotional representation from speech; however, the improved ability of RNNs for better modelling of long-range emotional context shift the research towards using RNNs in state-of-the-art SER systems.

*Recurrent neural networks* (RNNs) with gated architectures are specialised to model a long-range of contexts. Emotions in human speech are contextually embedded; therefore, the context capturing abilities of RNNs such as LSTM and GRU were explored in SER research by various studies. Lee *et al.* [8] use RNNs to learn high-level temporal dynamics of emotional representation from speech. They adopt bidirectional long short-term memory (BLSTM) network and achieve 12% improved results compared to the DNN-ELM [89]. In [123], the authors evaluate different BLSTM-RNN architectures for emotional representation learning from speech. They use the IEMOCAP corpus for evaluation and found that RNNs can learn both emotionally relevant short-term frame-level acoustic representation and compact utterance-level emotional representation of frame-level features. They report that BLSTMs outperform DNNs and SVMs trained on hand-engineered features. Ghosh *et al.* [97] evaluate the representation learning from spectrogram and glottal flow signals using DAE-BLSTM models. They report that their proposed framework can generate highly

discriminative representations that produce comparable emotion classification results to state-of-the-art approaches.

*Convolutional neural networks* (CNNs) are also popular for emotional representation learning in SER [5], [124], [125], [126], [127]. They can learn more generalised features from speech compared to DNNs, and other feature-based approaches [49]. In [95], the authors explore CNN for speech emotion detection. They found that CNN filters capture emotions related to the fundamental frequency, which helps create discriminative features for SER. Feature representations learnt by deep CNNs show robustness against noisy situations [128], [129]. Therefore, studies in SER also use deep CNN architectures such as ResNet and DenseNet for SER in noisy environment [130], [131]. The research on emotional representation learning was further advanced by the use of combined CNN and RNN architectures. Various studies used CNN-LSTM [49], [86], [132], [133], where they used CNNs for feature extraction and LSTM (or GRU) networks for modelling long term dependencies. Based on the results, these studies showed that CNN-RNNs is a better choice in contrast to using CNN or LSTM individually. In order to learn the spatial relationships in spectrograms, the authors in [134] used capsule networks (CapsNets). They designed a sequential capsule structure to obtain utterance-level emotional representations and evaluated the proposed model on the IEMOCAP dataset. They compared the results with baseline CNN-LSTM and showed that the proposed CapsNets are able to produce improved results in SER compared to the CNN-LSTM.

In supervised representation learning, *attention-based networks* have recently become very popular. Attention layers in DL models help the networks to focus on important emotional representations in the input speech. Researchers have attempted various attention mechanisms including self-attention [135], local attention [123], multi-hop attention [136], and many other variants [112], [137]. In these studies, the authors show that attention mechanisms enable networks to focus on affect-salient components and extract emotional representations from speech sentences, which help to improve the performance of the system.

Despite the promising results, the success of supervised training is limited by the requisite of labels. It is important to note that creating and labelling these datasets is very expensive in terms of time and resources. To tackle these issues, unsupervised learning has been used to learn representations from unlabelled data. We discuss unsupervised representation learning of speech emotion in the next subsection.

### 3.2 Unsupervised Representation Learning

Unsupervised representation learning facilitates the analysis of input data without corresponding labels and aims to learn the underlying inherent structure or distribution of the data. Real-life data (such as speech, image, or text) have extremely rich structures. Algorithms trained in an unsupervised way aim to learn the underlying structure of the data rather than learning any particular tasks, e.g., classification, prediction etc. In speech analysis, unsupervised representation learning can exploit the unlimited amount of unlabelled corpora to learn good intermediate feature representations, which can then be used to improve the performance of

supervised SER, where availability of labelled data is limited [138].

*Autoencoders* (AEs) are mostly utilised for unsupervised emotional representation learning from speech in SER. AEs can learn high-level semantic contents that are invariant to confounding low-level details (pitch contour or background noise) in speech [139]. In [140], the authors explored Denoising Autoencoders (DAEs) for emotional representation learning on IEMOCAP data. They empirically showed that a representation captured by the bottleneck layer of AEs are highly discriminative in separating the emotions and help to achieve comparable results to that of using hand-engineered features (such as voice quality features and MFCCs). In [97], Ghosh *et al.* use stacked DAEs for learning frame-level emotional representations from the spectrogram of speech and glottal flow signals. They evaluated the proposed framework on IEMOCAP data and found that a stacked DAE can learn highly discriminative features that help to achieve state-of-the-art results (54.56% UA). Huang *et al.* [141] evaluated different unsupervised representation learning algorithms including K-means clustering, the sparse AE, and sparse RBMs for SER. They explored the effect of the content window size and the number of hidden nodes on the performance. They found that a larger content window and more hidden units produce better results. To extract robust features, Xia *et al.* [63] used DAEs for SER. The authors empirically showed that DAEs can extract more robust feature representations and significantly outperformed using the static features for SER. Another work [98] utilised a modified DAE to model gender information to learn more robust emotional representations. The authors evaluated the proposed model on IEMOCAP data achieving improved results compared to the DAE used in [63] and hand-engineered features. Models like RBMs and DBNs can also learn high-level feature representations [142] and auditory-like sub-band filters [143] from speech, which can help improve the performance compared to hand-engineered features when used in SER for unsupervised representation learning [6], [144]. In [145], the authors utilise unsupervised learnt representations from unlabelled data to improve SER. They integrate an AE with a CNN-based emotion classifier to improve SER performance for within-corpus and cross-corpus settings. Generative models have further advanced the performance of unsupervised representation learning in SER.

*Generative models* including VAEs, GANs, and adversarial autoencoders (AAE), are becoming very popular in emotional representation learning due to their exceptional performance in learning representation and generating new data samples. In [67], the authors explore VAE architectures for latent representations of speech emotion. They perform extensive experiments on the IEMOCAP dataset and show that VAEs can learn discriminative emotional attributes suitable for improving classification than standard AEs. Eskimez *et al.* [101] evaluate various unsupervised autoencoding networks including DAE, VAE, AAEs, and adversarial variational Bayes (AVB) for emotional representation learning. They find that these unsupervised methods can capture the intrinsic structures of speech emotion that help to achieve improved results compared to SVMs and CNNs. Latif *et al.* [102] use GANs for representation learning as

well as generating synthetic representations. They modify the GAN and use the mixup [146] technique to augment a GAN in SER. They achieve 61.05% and 46.60% accuracy on within-corpus and cross-corpus settings, respectively. These studies show that generative models can learn better emotional representations and generate synthetic data samples that can be used to train the classifier along with real samples. This can lead to improvements in SER performance even in low resource conditions.

Despite all these successes, the performance of unsupervised representation learning techniques is not as good as that of the supervised methods [108]. Semi-supervised representation learning techniques alleviate this problem by simultaneously utilising both labelled and unlabelled data. We discuss the semi-supervised representation learning technique in the next subsection.

### 3.3 Semi-Supervised Representation Learning

The success in DL has predominately been possible due to key factors like advanced algorithms, processing hardware, open sharing of codes and papers, and most importantly, the availability of large-scale labelled datasets (e. g., ImageNet) pre-trained networks. However, a large labelled database or pre-trained network for every problem like SER is not always available [20], [88], [147]. It is very difficult, expensive, and time-consuming to annotate speech emotional data as it requires manual expert human efforts [108]. Semi-supervised representation learning attempts to solve this problem by utilising the feature representations from large unlabelled data, jointly with the labelled data, to build better classifiers. It reduces human efforts and provides higher accuracy in contrast to unsupervised representation learning; therefore, semi-supervised models are of great interest both in theory and practice [148]. Huang *et al.* [149] use a CNN in a semi-supervised way for capturing affect-salient representations. They evaluate their model on four publicly available datasets and find that a semi-supervised CNN learned salience, orthogonal, and discriminative representations for SER. These representations help to achieve superior performance compared to results using well-established hand-engineered features. Deng *et al.* [150] propose a semi-supervised model by combining an AE and a classifier. They consider samples from unlabelled data as an extra garbage class in the classification problem. They evaluate the proposed architecture on five publicly available datasets and show that features learnt by a semi-supervised AE improve SER performance compared to an unsupervised AE. Parthasarathy *et al.* [151] utilise semi-supervised AAE to disentangle the discrete emotion distribution and show that the proposed learning approach performs better compared to a fully supervised method. In [108], the authors train an AAE by utilising the additional unlabelled emotional data to improve SER performance. They perform evaluations on IEMOCAP and MSP-IMPROV and show that additional data help to learn more generalised representations that perform better than various supervised and unsupervised methods.

*Ladder network*-based semi-supervised methods are very popular in SER. A ladder network is an unsupervised DAE that is trained along with a supervised classification or regression task. It can learn more generalised representations



suitable for SER compared to the standard methods. Parthasarathy *et al.* [103] use a ladder network with skip connections between encoder and decoder networks for emotional representation learning. They evaluate the proposed model on the MSP-Podcast dataset and find that a semi-supervised ladder network can learn more powerful representations that facilitate better performance in predictions of emotional attributes than conventional DAE. In another study [105], the authors show that a ladder network can generate powerful and generalised representations that help to achieve relative gains in concordance correlation coefficient (CCC) of 3.0% to 3.5% for within-corpus, and 16.1% to 74.1% for cross-corpus settings using the MSP-Podcast, IEMOCAP, and MSP-IMPROV datasets. In [152], the authors utilise a semi-supervised ladder network to generate a robust feature representation by simultaneously minimising the sum of supervised classification and unsupervised cost functions. The features generated by a ladder network are used as an emotional representation for classification with an SVM. They perform evaluations on the IEMOCAP corpus and show that the proposed framework achieves 2.6% improved performance than a DAE, and 5.3% higher than the static acoustic features. Tao *et al.* [106] also utilise a ladder network for emotional representation generation and conducted experiments on the IEMOCAP dataset, and achieve improved classification performance with a small number of labelled samples compared to DAE, VAE, and hand-engineered features.

Generative adversarial networks (GANs) were also explored for semi-supervised representation learning in SER. Chang *et al.* [109] performed emotional representation learning from speech using a convolutional GAN. They utilise 100 hours of unlabelled data and show that the proposed model derives automatic discriminative representations learning to improve the SER performance. They perform classification on emotional valence using a discrete 5-point scale and 3-point scale and achieve an accuracy of 43.88% and 49.80%, respectively. Sahu *et al.* [153] use a conditional GAN for modelling feature representations and generating new data samples. They performed evaluations on the IEMOCAP and MSP-IMPROV datasets and showed that synthetic feature vectors can help improve SER performance in different settings. In another study [154], the authors use a GAN to generate a high dimensional synthetic feature representation using lower-dimensional feature vectors and apply synthetic feature representations to augment the training data. Based on the within-corpus and cross-corpus evaluations, they find that synthetic data can help to improve performance. In [155], the authors present a semi-supervised adversarial model to facilitate knowledge transfer from videos to the audio domain, hence, improving SER performance. They show that the proposed model can outperform a baseline supervised method on the CREMA-D and RAVDESS datasets. Zhao *et al.* [156] present robust semi-supervised GANs to address the issue of labelled data unavailability. They evaluate the model on four publicly available datasets for capturing underlying emotional representation knowledge from both labelled and unlabelled data. They demonstrate that the proposed methods are superior to the state-of-the-art supervised and semi-supervised models. Sahu *et al.* [157] evaluate semi-supervised AAEs in SER for encoding emotional representation in a

compressed form and generating the synthetic data samples. They perform experiments on IEMOCAP and observed that an AAE can encode emotional representation in compressed form without losing emotional class discriminability and can generate synthetic samples that augment the training data to improve the SER performance.

Semi-supervised representation learning is mainly used in SER to circumvent the lack of sufficient labelled training data by utilising unlabelled data. These studies show that semi-supervised representation learning helps to learn generalised representations by including the unlabelled data into the training pipeline, which leads to performance improvements. Another way of using unlabelled data is the representation transfer learning that we discuss next.

### 3.4 Representation Transfer Learning

*Transfer learning (TL)* involves methods that utilise any knowledge resources (i. e., data, model, representations, labels, etc.) to increase the model learning and generalisation for the target task [158]. The idea behind TL is “Learning to Learn”, which specifies that learning from scratch (*tabula rasa learning*) is often limited, and experience should be used for deeper understanding [159]. Representation transfer learning involves using representations learnt on any large scale data and can be beneficially utilised for the target task. It encompasses different approaches, including multitask learning (MTL), domain adaptation, knowledge transfer, covariance, self-supervised learning, etc. This subsection covers domain adaptive, multitask, and self-supervised representation learning, which is a very popular representation transfer learning approach to improve SER’s performance.

#### 3.4.1 Deep Domain Adaptive Representation Learning

*Deep domain adaptive representation learning* is a sub-field of TL, and it has emerged to address the problem of domain shift. SER systems can achieve better results when evaluated on test data having a distribution similar to the training set. However, the performance of SER systems is degraded by the mismatch in training and testing data distributions. These differences become more significant with the training and test data of different languages, leading to failing SER systems to function. To build more robust systems for SER applications, domain adaptive representation learning techniques are usually applied to explicitly minimise the difference between the training (source) and testing (target) domains.

In SER, various domain adaptive representation learning methods are evaluated to enable the system to learn representations that can be used to perform emotion identification across different corpora and different languages. Deng *et al.* [160] use a DAE with shared hidden layers to learn common representations for different emotional datasets. The proposed model can minimise the discrepancy between different datasets and increase the emotion classification accuracy compared to other feature domain adaptation methods. In [161], the authors introduce shared hidden-layer AE to learn common feature representations shared across the source and target data to reduce the discrepancy in them. They perform evaluations on three publicly available corpora and

demonstrate that the proposed method significantly improves the emotion classification accuracy compared to a DAE. In another study [114], the authors use a Universum AE for unsupervised domain adaptation to improve cross-corpus SER. They performed evaluations on four publicly available datasets and showed that the Universum AE has the strong representational capability to discover common structures among the source and target speech data.

*Domain adversarial neural networks* (DANNs) [162] are becoming popular in learning domain adaptive representation learning. The authors in [116] suggest an adversarial domain network for cross-corpus emotional attributes' prediction. They focus on capturing common representations between the train and test domains by applying a gradient reversal layer (GRL) which propagates back the gradient produced by the domain classifier to the shared layers. They observe that the proposed model can learn domain invariant representations to improve the primary regression task. Xiao *et al.* [163] present an adversarial network for class-Aligned and generalised domain-invariant representations learning. They also consider GRL to facilitate shared representations among source and target domains. They evaluate the proposed architecture against cross-corpus settings and achieve improved results compared to AE-based models and DANNs. In [164], the authors evaluate a DANN against cross-lingual SER. They use GRL with a language classifier, which helps the model to learn language-independent emotional representations. Experiments with the IEMOCAP and RECOLA datasets show that their proposed method achieves 3.91% improved accuracy than the baseline system (naive cross-lingual SER) for the arousal and valence classification tasks.

Some studies also use different adversarial networks for domain adaptive representation learning for cross-corpus and cross-language SER. Zhou *et al.* [165] investigate a class-wise domain adaptation method using adversarial training to address cross-corpus mismatch issues and show that adversarial training is useful when the model is to be trained on a target language with minimal labels. The authors perform evaluations on the EMODB, and FAU-AIBO datasets [166] and show that the proposed architecture learns generalised representations that minimise the domain shift between positive and negative emotion classes. Gideon *et al.* [167] introduce an adversarial discriminative domain generalisation method that follows a "meet in the middle" approach for cross-corpus emotion recognition. The proposed model improves the cross-corpus generalisation by iteratively moving the learnt representations for each dataset closer. They perform evaluations on the IEMOCAP, MSP-IMPROV, and PRIORI emotion datasets [168] and find that the proposed model consistently converges and generates more generalised representations for cross-corpus SER, even when no target labelled data is used. In [20], the authors utilise a GAN-based model in an unsupervised way to learn language invariant features and evaluate the model over four different language datasets. They significantly improve the performance of SER across different languages using language invariant representations.

### 3.4.2 Multi-Task Representation Learning

*Multi-task learning* (MTL) has led to successes in different applications of ML, from NLP [169] and speech analysis

[170] to computer vision [171]. MTL aims to optimise more than one loss function in contrast to single-task learning (STL) and uses auxiliary tasks to improve the main task of interest [172]. Multitask representation learning (MTRL) can improve the performance of the main task by capturing underlying relevant factors from the auxiliary tasks [9], [173]. In this way, representations learnt in the MTL scenario become more generalised, which helps improve the performance. In SER, the speech also contains multi-dimensional information about the message, speaker, and gender that can be used as auxiliary tasks to improve performance without external speech data.

For SER, studies use emotional attributes (e. g., arousal and valence) as auxiliary tasks to improve the performance of the system. Xia *et al.* [110] apply a DBN-based MTL model that uses dimensional emotions as auxiliary tasks to improve the performance of categorical emotion as a major task. Their results indicate that learning shared representations for different tasks acts as complementary information to SER systems and helps to improve performance. In [174], Kim *et al.* presented an MTRL framework that utilises gender and naturalness as auxiliary tasks. They evaluate the proposed model with within-corpus and cross-corpus settings on five publicly available datasets and find that MTL improves the generalisation in SER by learning more generalised representations when compared to the state-of-the-art STL methods. Further, Parthasarathy *et al.* [175] proposed an MTRL framework for joint prediction of emotional attributes including arousal, valence, and dominance by exploiting their interdependencies. Their results indicate that the proposed model learns shared representations that maximise the performance of the regression models. They base their experiments on three datasets for evaluations and demonstrate that the proposed MTRL model gains a concordance correlation coefficient (CCC) as high as 4.7% for within-corpus and 14.0% for cross-corpora experiments compared to STL. In [111], the authors introduce an MTRL framework for jointly learning primary and secondary emotions. They perform evaluations on the MSP-Podcast database and show that the proposed MTRL model can leverage the extra information about the secondary emotions and leads to relative improvements of 7.9% in F1-score for an 8-class emotion classification task.

Other auxiliary tasks that researchers consider in MTL SER are speaker and gender recognition to improve the accuracy of a system compared to STL [176]. In [112], Nediyanath *et al.* utilise a multi-head attention-based MTRL framework with gender classification as an auxiliary information source. They find that gender-specific representations influence the emotion characteristics in speech and achieved 70.1% for UA—that is 5.3% higher than the state-of-the-art reported accuracy in SER for four emotion classes at the time. Tao *et al.* [113] utilise a variant of a multitask LSTM for learning contextual representations with speaker and gender as auxiliary tasks. The proposed model learns a shared representation for multitasks, which help to achieve a 5.5% relatively higher accuracy than the 'standard' LSTM on the IEMOCAP dataset. Next, Latif *et al.* [108] introduce an MTRL framework that uses auxiliary tasks for which data is abundantly available and find that utilisation of this additional data for auxiliary tasks can improve the main

task of emotion classification with limited available labelled data. They apply AAE to learn powerful and discriminative representations with gender identification and speaker recognition as the auxiliary tasks. Evaluations performed on IEMO-CAP, and MSP-IMPROV show that the proposed model can generate generalised and discriminative representations that help to achieve results better than the state-of-the-art comparable studies, with a supervised single- and multitask CNN, and single- and multitask semi-supervised AEs.

MTRL is an effective approach to learning a shared representation that leads to no major increase in computational power while improving the system's recognition accuracy and decreasing the chance of overfitting [108]. However, MTL implies the preparation of labels for considered auxiliary tasks, which is expensive and time-consuming. Recently, self-supervised representation learning is emerging as a solution to utilise representations learnt from unlabelled data to supervise a downstream task. We discuss self-supervised representation learning next.

### 3.4.3 Self-Supervised Representation Learning

*Self-supervised representation learning* [177] is a new paradigm in ML, which is a form of unsupervised learning, where the data provides the supervision. The self-supervised task, also known as the pretext task, uses the unlabelled data to guide downstream tasks. Self-supervised representation learning utilises both labelled and unlabelled data. However, unlabelled data do not need to belong to the same class labels or generative distribution as the labelled data. Such a loose restriction on unlabelled data in self-supervised learning significantly simplifies learning from a large volume of unlabelled data.

Self-supervised representation learning-based models are getting tremendous interest in vision [178], NLP [179], and speech recognition [180], however, emotional representation learning with self-supervising needs exploration in SER. We find a recent study [181] that presents a visual data-guided self-supervised framework for speech representation learning. The authors evaluate the proposed model in SER and automatic speech recognition (ASR) and achieve state-of-the-art results for emotion recognition and competitive results for speech recognition. In [182], the authors propose a multitask self-supervised method for shared speech representation learning, where a single neural encoder is followed by multiple workers that jointly solve different self-supervised tasks. They achieve improved results for speaker, phonemes, and emotional cues identification. Further, transformers [183] are becoming very popular in applying a self-supervised multi-modal representation to improve SER [184]. In [185], the authors exploit transformer-based self-supervised representation learning to improve multi-modal emotion recognition. They report that fine-tuning the transformer from the masked language modelling task improve emotion recognition performance by 3% on the CMU-MOSE dataset. These studies indicate the potential of self-supervised representation learning, which need to be further explored in SER.

## 3.5 DRL for Representation Learning

*Deep reinforcement learning (DRL)* is a combination of DL and reinforcement learning (RL) principles to create efficient

and autonomous systems that can learn by interacting with their environment. RL follows the principle of behaviourist psychology, where an agent learns to take actions in an environment and tries to maximise the accumulated reward over its lifetime. RL has been repeatedly successful in solving various problems [186]; however, previous methods were inherently limited to low-dimensional problems and lacked scalability. The advancements in DL have accelerated the progress in RL and gave rise to various algorithms to solve high-dimensional complex problems [187].

Recently, DRL is also gaining interest in the speech community, and researchers have proposed multiple approaches to model different speech problems [188]. Some of the popular RL-based solutions include dialog modelling and optimisation [189], [190], speech recognition [191], and speech enhancement [192]. In SER, researchers also use DRL algorithms for emotion modelling in speech [120], [193]. However, the problem of emotional representation learning for improving SER is not explored using DRL.

## 4 CHALLENGES OF DEEP REPRESENTATION LEARNING IN SER

In this section, we present the major challenges of using deep representation learning in SER.

### 4.1 Training Complexity

Training DL models for representation learning is not straightforward. It applies highly non-linear functions to the input signal to learn abstract representations. Learning representations associated with input manifolds requires intense and difficult training to unfold and distort complicated input manifolds [9]. Speech signals have complex manifolds [194] that inherently embody information related to the message as well as the speaker's gender, age, health status, personality, friendliness, mood, and emotion. These types of information are entangled together [40], and training DL models for disentanglement of emotional representations from other attributes (in a latent space) is a difficult task and has been a long-standing goal in SER.

The training of unsupervised representation learning models in SER is much more difficult than that of supervised ones. Due to relatively small emotional speech datasets, unsupervised representation learning methods do not guarantee to learn a useful representation and can potentially ignore emotional attributes [195]. In contrast, unsupervised representation learning from larger audio or multi-modal datasets can be potentially used to improve the SER performance. However, the performance improvement is not significant despite increasing the complexity of the system by using hundreds of hours of audio data [145]. Therefore, semi-supervised representation learning models are considered as an alternative, as they utilise both labelled and unlabelled data (as discussed in Section 3.3). However, blind training of DL models for semi-supervised representation learning may not necessarily improve the performance over supervised learning [196]. Empirical evidence suggests that unlabelled data only help in certain favourable situations when there is a link between the marginal data distribution and the target function [197]. In fact, noisy and biased unlabelled data can even lead to worse performance

[198]. Therefore, it is required to manually select the learning parameters and regularise the semi-supervised model to learn a generalised representation from both, labelled and unlabelled data that help improve the performance compared to the supervised techniques.

Literature shows that GANs have a strong ability to model data distribution and learn discriminative representations. However, they are difficult to train on available emotional corpora, as they face convergence issues [154]. To address this, researchers use various techniques, including conditional architectures [154], [199], unlabelled data [200], and data augmentation techniques [102] with substantial room for further improvement for effective training of GANs in learning emotion representations.

## 4.2 Lack of Emotional Speech Data

Deep representation learning models aim to identify potentially useful and ultimately understandable patterns. This demands not just plenty of data, but diverse data that capture all the directions of variation in the data [201]. For learning a good representation, data must be accurately labelled and unbiased. Most of the SER corpora are designed in laboratories, which may have bias and the recorded corpus on acted emotions may not represent real-life human emotions. This can lead the algorithms to exhibit erroneous behaviour [202].

The quality of speech emotional data can also be poor due to various other reasons. For example, different background noises and music can corrupt speech data. Similarly, the noise of microphones or recording devices can also contaminate the speech signal. Although studies use ‘noise injection’ techniques to avoid overfitting, this only works for moderately high signal-to-noise ratios [203]. Due to the current emphasis on emotion recognition in the wild, this has become ever so important to recognise speech emotion from noisy data. DAEs [62] can learn a representation of data with noise, imputation AE [204] can learn a representation from incomplete data, and non-local AE [205] can learn reliable representations from corrupted data. These models are very popular for noise invariant emotional representation learning; however, their performance still needs further improvement.

In SER, the design and use of the existing emotional speech database primarily depend on the research goals. For example, the emotions can be classified as soothing, and prohibition [206]; or joy and anger [207]. The number and type of emotions contained in a database define the emotional classification task. In most cases, the corpora are purpose-driven, developed by professional actors, and do not naturally incorporate and simulate emotions. Most importantly, these corpora are annotated by human raters as an ‘outer emotion’, which can be highly different from the ‘inner emotion’ of an individual. Representations learnt from such laboratory designed datasets cannot be generalised to real-life natural emotions. To eventuate real-world applications of deep emotional representation learning-based SER systems, there is still a need for generic emotional speech corpora by using standard ground truth, which captures all the human emotions. Moreover, these emotional speech databases should be standardised and available for the research community.

## 4.3 Corpus and Lingual Variance

Deep representation learning-based SER systems have achieved improved results when evaluated using similar training and testing data. However, the performance of these systems drops significantly if the test samples deviate from the distribution of the training data. Learning emotional representations that are invariant to speakers, language, etc., are difficult to achieve. The representations learnt from one corpus tend not to work well on other corpora with different recording conditions.

In the past few years, researchers have achieved competitive performance by learning speaker invariant emotional representations [208], [209]. However, language and corpus invariant representation learning is still a very challenging task. Although emotions are considered language invariant, the performance of SER systems degrades when tested across different language emotional corpora [20]. Representations learnt by a few shot learning can be a solution for adapting SER systems, which needs a few samples of target language data. Compared to the number of spoken languages globally, we have speech corpora covering a few languages only. Even though there are more than 5 000 spoken languages in the world, 389 languages alone account for 94 % of the world’s population.<sup>2</sup> However, speech corpora are missing even for all of these 389 languages, which makes cross-language speech emotion research more challenging. The variation, imbalance, diversity, and dynamics in speech and language corpora present hurdles to designing generalised representation learning algorithms. Recent studies are focusing on representation learning for languages with the very small size of emotional datasets [20], but a fully satisfactory solution has not yet emerged.

## 4.4 Privacy and Robustness Issues

When people use SER services, they usually grant complete access to their speech recordings or transmit features through the network. It can cause a leak of user’s information such as gender, ethnicity, and emotional state and can be used for unintended purposes [40], [210]. Similarly, the users’ recordings can also be edited or used to create a fake speech that the user never spoke, or the voiceprints can be used to deceive voice-authentication systems. In healthcare applications of SER, there are also risks of users’ personal private information leakage [211]. Various other privacy-related issues that arise while using speech-based services have been discussed in [212]. It is desirable in SER applications that there are suitable provisions for ensuring that there is no unauthorised and undisclosed eavesdropping and violation of privacy.

Privacy-preserved representation learning can alleviate this problem but is a relatively unexplored research topic. Recently, researchers have started to utilise privacy-preserving representation learning models to protect speaker identity [213], gender identity [210], and language information [40]. This motivates the exploitation of deep representation learning models on devices, or edge servers [211]. In this way, robust representation from speech in smaller dimensions can be learned and transmitted to the network for real-life applications [40]. However, model or feature

2. <https://www.ethnologue.com/statistics>

**TABLE 6**  
Summary of Challenges, Gaps, and Future Directions of Deep Representation Learning in SER

Challenges	Solutions explored in Literature	Existing Gaps	Future Directions
Training complexity	Static representation learning methods	Lack of exploration	DRL-based methods need to be explored for emotional representation learning
Limited size labelled emotional data.	Unlabelled data using Unsupervised representation learning techniques.	Low performance.	Investigation of self-supervised representation learning methods.
Corpus and Lingual Variance	Domain adaptive representation learning methods.	Performance not comparable with baseline results.	Investigation of multi-modal representation to improve SER.
Privacy and Robustness Issues.	Privacy preserving representations.	Performance drops using these representations.	Research is required on effective defence mechanisms against adversarial attacks.

sizes should be optimised for on-device utilisation for feature extraction. To preserve users' privacy, federated learning [214] is another technique, explored in [215] for SER, where the training of a shared global model is performed using multiple participating computing devices. These participating devices collaboratively learn a shared model without revealing their local data and avoid privacy infringement.

Recent studies on adversarial examples pose enormous challenges for robust representation learning from speech by showing the susceptibility of deep models to adversarial examples having imperceptible perturbations [216]. Some popular adversarial attacks include the fast gradient sign method (FGSM) [217], Jacobian-based saliency map attack (JSMA) [218], and DeepFool [219]. They compute the perturbation noise based on the gradient of the targeted output. SER systems are also vulnerable to these attacks [130], [220]. The success of adversarial attacks against SER systems shows that the representations learnt by underlying DL models are not robust [221]. Immunity against such adversarial perturbations, which could mislead SER classifiers, can be achieved by training a DL model to generate an invariant representation to such transformations. This has been explored in the image domain [222] but needs to be explored for emotional representation learning from speech. In SER, emotional representations learnt by very deep architectures are found robust against adversarial attacks [130]. However, further research is required to tackle the challenge of adversarial attacks by exploring what DL models capture from the input speech data and how adversarial examples can be defined as a combination of previously learnt representations without any knowledge of adversaries [223].

## 5 DISCUSSION AND FUTURE DIRECTIONS

This section highlights the gaps and provides future research pointers for different aspects of deep representation learning in SER. For quick insights, the readers are referred to Table 6 that presents the summary of challenges, solutions presented in the literature, existing gaps, and future directions.

### 5.1 Input Features

In recent years, the trend of using hand-engineered acoustic features has been progressively changing. Deep representation learning is gaining popularity as a viable alternative to

learn directly from raw speech or features requiring fewer processing steps. Researchers achieved promising results using CNNs and CNN-RNNs to learn low-level speech representations from raw waveforms, allowing a network to capture important emotional characteristics better. However, the proper design of the feature extraction block is crucial to achieving this goal [49]. However, raw speech as input to deep models requires enormous data to achieve competitive performance. Researchers use data augmentation techniques to meet the data requirement [49], [224]. Log-Mel features and spectrograms are considered popular choices to alleviate this problem as they need less processing, fewer data samples, and training to achieve state-of-the-art classification performance compared to setups where raw audio is used. Table 5 shows that different hand-engineered features are more popular compared to the raw speech as input. However, it has been shown in recent studies [21], [50] that deep representation learning techniques can extract discriminative representations and a particular choice of input features is not as important as the model architecture. Therefore, future research is required to design deep architectures that have minimal human knowledge to learn generalised representations across emotions, languages, and corpora.

### 5.2 Models

A summary of various deep representation learning techniques is presented in Table 5. Studies using supervised representation learning methods typically focus on learning discriminative and robust representations. Models like CNNs, LSTM/GRU RNNs, and CNN-LSTM/GRU-RNNs are widely used for learning salient emotional representations from raw speech. The reason for their popularity is that CNN layers act as data-driven filterbanks that can model spectral envelope of raw speech, and LSTM/GRU-RNNs can model contextual information. Therefore, most of the studies on raw speech either use CNN, LSTM/GRU-RNNs, or their combination for SER. Among these models, LSTM/GRU-RNN-based architectures are mostly applied due to their ability to capture temporal context. However, RNNs use computationally expensive back-propagation through time (BPTT) [225] to learn temporal dependencies by sequentially processing the speech signal. Recently, Transformers solve this issue by utilising the self-attention mechanism for learning temporal correlations from the



TABLE 7  
Comparing Attributes of Different Deep Representation Learning Techniques

Attributes	Supervised Representation Learning	Unsupervised Representation Learning	Semi-Supervised Representation Learning	Representation Transfer Learning
Fully labelled Data Compulsory	✓	×	×	×
Unlabelled Data	×	✓	✓	✓
Partial Labelled Data	×	×	✓	✓
Predict Label/Future	✓	×	✓	✓
Accuracy	High	Low	High	High
Exploration	×	×	×	×

sequential data [183]. This makes Transformers capable of capturing more temporal contexts with less computation complexity. Various studies in ASR [226], [227], [228] and speech synthesis [229], [230] highlight the contextual representation learning ability and computational efficiency of Transformers. Emotions in speech are also contextually dependent. Therefore, Transformers need to be explored in SER.

Following the success of generative models, recent papers are mostly focused on utilising their distribution learning and generation abilities. Models like VAEs, AAAs, and GANs are becoming popular choices in SER due to their representation learning and feature generating abilities. An interesting utilisation of GANs is generating synthetic data, which can be utilised for SER to solve the data scarcity problem. Learning a deep representation from synthetic data can help improve the performance, and researchers have validated the effective use of synthetic data for SER [102], [157], [200]. However, creating ‘accurate’ synthetic speech or features in different emotions using the available limited sized emotional corpora is a difficult task, and generative models like GANs face convergence issues [154]. Therefore, further research is needed to explore effective training methods for generative models.

It is still an open research question which deep model is superior for emotional representation learning to improve SER performance. It is very hard to answer this question from this literature search since different studies achieved state-of-the-art results in different settings with different models. This mainly depends on how effectively a particular deep architecture is designed, pre-trained, and tuned for deep representation learning.

### 5.3 Training Technique

We present a comparison in Table 7 on the training techniques of deep representation learning models based on different attributes to highlight trade-offs. Supervised representation learning models are very popular in SER due to better performance. However, the unavailability of labelled data is a real bottleneck. These models require fully labelled data as highlighted in Table 7. Unsupervised representation learning can alleviate this problem by learning emotional structures and patterns without requiring any labels, which can help to improve the performance of the emotion classification task [145]. Autoencoding networks are widely used in SER for unsupervised feature learning from speech; however, SER performance with unsupervised models is always inferior to that of supervised models (as highlighted in Tables 5 and 6). Self-supervised representation learning is gaining interest in vision, NLP, and speech recognition. It enables better

utilisation of unlabelled data by learning high-level representations that can be used for different downstream tasks. Research is required to explore the utilisation of representations learnt from multi-modal data by self-supervision on the downstream task of SER.

Semi-supervised representation learning models are widely used in SER because they can exploit both labelled and unlabelled data to learn more. Studies utilise generalised representations from both labelled and unlabelled data to improve the performance of SER. The popular models include GANs [109], AE-based models [108], [150] and other discriminative architectures [231], [232]. However, there are still opportunities in semi-supervised representation learning for improving the performance of SER by concomitantly reducing the annotation burden of emotional corpora. Most importantly, further research is required to provide theoretical guidelines on the number of labelled versus unlabelled samples required to build semi-supervised SER systems for practical applications.

MTRL methods are very popular in SER, where researchers utilise additional label information available (e. g., speaker or gender) in speech as auxiliary tasks to learn more generalised representations that help to improve the SER performance [108]. However, extra efforts are required to prepare the labels for the auxiliary task in MTRL approaches, which is an expensive and time-consuming task. Another problem that hinders MTRL is the temporal differences among tasks. For instance, the modelling of speaker recognition requires different temporal information than phoneme recognition [233]. Ideally, memory-based deep neural networks, e. g., LSTM or GRU cells, can help address this issue.

Good representation disentangles the underlying explanatory factors of variation. However, it is an open research question what kind of training framework can potentially learn disentangled representations from input data. As highlighted in Table 7, all static representation learning methods do not involve exploration. Reinforcement learning (RL), on the other hand, facilitates the idea of exploration while learning by interacting with the environment. A good representation can be learned if RL is used to disentangle factors of variation by interacting with the environment. This will lead to faster convergence, in contrast to blindly attempting to solve given problems. Thomas *et al.* [234] recently validate this idea, where the authors use RL to disentangle the independently controllable factors of variation by using a specific objective function. The authors show that the agent can disentangle these aspects of the environment without any extrinsic reward. This is an important finding that will act as the key to further research

in this direction. Some other studies [235], [236], [237], [238] also use RL-based approaches to learn representations and have achieved considerable success. However, such RL-based approaches need to be further explored for SER.

## 6 CONCLUSION

This article focuses on providing a comprehensive review of speech emotional representation learning using deep learning approaches. In speech emotion recognition (SER), the use of representation learning is very important, and there is ongoing research on this topic in which different models and methods are being explored to disentangle speech attributes suitable for emotion detection and identification. The highlights of this survey are as follows:

- Most SER corpora are developed in laboratories by professional actors and annotated by human raters as outer emotions that might be significantly different from inner ones. Future efforts are required to motivate the collection of natural emotional corpora by defining standardised data collection protocols with a special focus on protecting speakers' personal information. It will encourage participants to enrol in data collection.
- The SER research community is increasingly shifting its focus to designing systems using raw speech or input features that have minimal human knowledge dependency.
- LSTM/GRU-RNNs combined with CNNs are very popular and suitable for capturing emotional attributes in a supervised way. In unsupervised representation learning, Denoising Autoencoders (DAEs) and Variational Autoencoders (VAEs) are widely deployed architectures in SER, and GAN-based models also gaining attention. Further research on utilising these models in a semi-supervised way is required to learn representation from unlabelled and labelled data.
- Emotions are context-dependent, and Transformers can better capture temporal contexts compared to RNNs. This encourages SER researchers to utilise Transformers in their studies.
- Static deep representation learning methods are very popular; however, they lack exploration. In contrast, Deep Reinforcement Learning (DRL) facilitate exploration by interacting with the environment and learn better representation. This calls SER researchers to utilise DRL to disentangle the emotional attributes to achieve better performance.
- Privacy-preserving representation learning in SER is very important to explore. This can help achieve robustness against adversarial attacks and ensure no unauthorised access to users' personal information while using SER services.
- Domain adaptation solutions based on adversarial neural networks are widely used for cross-corpus and cross-language emotion recognition; however, performance is not comparable to baseline. Multi-modal self-supervised domain adaptive representation learning models can potentially improve performance.

We believe this article has the potential to become a definitive guide to researchers and practitioners interested in deep representation learning for SER. With all these changes and advancements in place, we look forward to an exciting era of SER starting to enable artificial intelligence (AI) to sense our emotions better. We are curious whether, in the longer run, deep representation learning will be the standard paradigm in SER. If so, we are currently changing a paradigm moving away from signal processing and expert-crafted features into a highly data-driven era—with all its advantages, challenges, and risks.

## REFERENCES

- [1] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychol. Bull.*, vol. 111, no. 2, pp. 256–274, 1992.
- [2] S. Kuchibhotla, H. Vankayalapati, R. Vaddi, and K. R. Anne, "A comparative analysis of classifiers in emotion recognition through acoustic features," *Int. J. Speech Technol.*, vol. 17, no. 4, pp. 401–408, 2014.
- [3] A. Konar and A. Chakraborty, *Emotion Recognition: A Pattern Analysis Approach*. Hoboken, NJ, USA: Wiley, 2014.
- [4] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
- [5] 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.
- [6] C. Huang, W. Gong, W. Fu, and D. Feng, "A research of speech emotion recognition based on deep belief network and SVM," *Math. Problems Eng.*, vol. 2014, no. 1, pp. 1–7, 2014.
- [7] Z. Cairong, Z. Xinran, Z. Cheng, and Z. Li, "A novel DBN feature fusion model for cross-corpus speech emotion recognition," *J. Elect. Comput. Eng.*, vol. 2016, pp. 1–11, 2016.
- [8] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. Int. Speech Commun. Assoc.*, 2015, pp. 1537–1540.
- [9] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
- [10] G. Zhong, L.-N. Wang, X. Ling, and J. Dong, "An overview on data representation learning: From traditional feature learning to recent deep learning," *J. Finance Data Sci.*, vol. 2, no. 4, pp. 265–278, 2016.
- [11] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, "A review on emotion recognition using speech," in *Proc. Int. Conf. Inventive Commun. Computat. Technol.*, 2017, pp. 109–114.
- [12] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: A review," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 93–120, 2018.
- [13] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, 2020.
- [14] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901.
- [15] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *Int. Conf. Learn. Representations*, pp. 5223–5234, 2014.
- [16] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [17] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *Int. Conf. Learn. Representations*, 2016.
- [18] S. Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1986, pp. 1991–1994.
- [19] F. Eyben et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2016.

- [20] S. Latif, J. Qadir, and M. Bilal, "Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact.*, 2019, pp. 732–737.
- [21] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1263–1267.
- [22] G. Zhong, X. Ling, and L.-N. Wang, "From shallow feature learning to deep learning: Benefits from the width and depth of deep architectures," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discov.*, vol. 9, no. 1, 2019, Art. no. e1255.
- [23] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
- [24] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
- [25] I. Borg and P. Groenen, "Modern multidimensional scaling: Theory and applications," *J. Educ. Meas.*, vol. 40, no. 3, pp. 277–280, 2003.
- [26] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4–5, pp. 411–430, 2000.
- [27] F. Lee, R. Scherer, R. Leeb, A. Schlögl, H. Bischof, and G. Pfurtscheller, *Feature Mapping Using PCA, Locally Linear Embedding and Isometric Feature Mapping for EEG-Based Brain Computer Interface*. Princeton, NJ, USA: Citeseer, 2004.
- [28] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
- [29] T. Virtanen *et al.*, "Active-set newton algorithm for non-negative sparse coding of audio," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 3092–3096.
- [30] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
- [31] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [32] Y. LeCun *et al.*, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
- [33] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.
- [34] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computat.*, vol. 12, no. 10, pp. 2385–2404, 2000.
- [35] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognit. Lett.*, vol. 42, pp. 11–24, 2014.
- [36] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [37] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 153–160.
- [38] C. Poulthney *et al.*, "Efficient learning of sparse representations with an energy-based model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1137–1144.
- [39] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: <http://www.deeplearningbook.org>
- [40] H. S. Ali, F. ul Hassan, S. Latif, H. U. Manzoor, and J. Qadir, "Privacy enhanced speech emotion communication using deep learning aided edge computing," in *Proc. IEEE Int. Conf. Commun. Workshops*, 2021, pp. 1–5.
- [41] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory*. London, U.K.: Psychol. Press, 1949.
- [42] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychol. Rev.*, vol. 65, no. 6, pp. 386–408, 1958.
- [43] Y. Bengio *et al.*, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
- [44] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, vol. 2, pp. 2377–2385.
- [45] Y. LeCun *et al.*, "Handwritten digit recognition with a back-propagation network," *Adv. Neural Inf. Process. Syst.*, vol. 2, pp. 396–404, 1989.
- [46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [47] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.
- [48] Y. LeCun *et al.*, "Convolutional networks for images, speech, and time series," *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA, USA: MIT Press, 1995, pp. 255–258.
- [49] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modeling of speech emotion from raw speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 3920–3924.
- [50] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 2741–2745.
- [51] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [52] M. I. Jordan, "Serial order: A parallel distributed processing approach," in *Advances in Psychology*. New York, NY, USA: Elsevier, 1997, vol. 121, pp. 471–495.
- [53] K. Cho, B. van Merriënboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2014, pp. 1724–1734.
- [54] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [55] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [56] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.
- [57] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length and helmholtz free energy," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 3–10.
- [58] M. Usman, S. Latif, and J. Qadir, "Using deep autoencoders for facial expression recognition," in *Proc. 13th Int. Conf. Emerg. Technol.*, 2017, pp. 1–6.
- [59] A. Ng *et al.*, Sparse autoencoder. CS294A Lecture Notes, 2011. [Online]. Available: <https://web.stanford.edu/class/cs294a/sparseAutoencoder.pdf>
- [60] A. Makhzani and B. Frey, "K-sparse autoencoders," 2013, *arXiv:1312.5663*.
- [61] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, 2013, pp. 511–516.
- [62] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1096–1103.
- [63] R. Xia and Y. Liu, "Using denoising autoencoder for emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 2886–2889.
- [64] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [65] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for Boltzmann machines," *Cognit. Sci.*, vol. 9, no. 1, pp. 147–169, 1985.
- [66] Y. Bengio, E. Laufer, G. Alain, and J. Yosinski, "Deep generative stochastic networks trainable by backprop," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 226–234.
- [67] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 3107–3111.
- [68] I. Higgins *et al.*, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–13.

- [69] S. Zhao, J. Song, and S. Ermon, "InfoVAE: Balancing learning and inference in variational autoencoders," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 5885–5892, Jul. 2019.
- [70] H. Kim and A. Mnih, "Disentangling by factorising," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2649–2658.
- [71] M. Tschannen, O. Bachem, and M. Lucic, "Recent advances in autoencoder-based representation learning," *CoRR*, vol. abs/1812.05069, 2018, *arXiv:1812.05069*. [Online]. Available: <http://arxiv.org/abs/1812.05069>
- [72] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.
- [73] D. Ververidis and C. Kotropoulos, "A state of the art review on emotional speech databases," in *Proc. 1st Richmedia Conf.*, 2003, pp. 109–119.
- [74] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.
- [75] P. Ekman, "Expression and the nature of emotion," *Approaches to Emotion*, vol. 3, New York, NY, USA: Taylor Francis, 1984, pp. 19–344.
- [76] J. A. Russell, "A circumplex model of affect," *J. Pers. Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.
- [77] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2005, vol. 5, pp. 1517–1520.
- [78] C. Busso, S. Parthasarathy, A. Burmanian, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan.–Mar. 2017.
- [79] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, Oct.–Dec. 2019.
- [80] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5–17, Jan.–Mar. 2012.
- [81] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, 2008, Art. no. 335.
- [82] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO corpus: An Italian emotional speech database," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2014, pp. 3501–3504.
- [83] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.
- [84] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.
- [85] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, no. 9–10, pp. 1062–1087, Nov./Dec. 2011.
- [86] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5089–5093.
- [87] Y. Huang, K. Tian, A. Wu, and G. Zhang, "Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition," *J. Ambient Intell. Humanized Comput.*, vol. 10, no. 5, pp. 1787–1798, 2019.
- [88] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 257–261.
- [89] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. Fifteenth Annu. Conf. Int. Speech Commun. Assoc.*, 2014, 223–227.
- [90] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 5688–5691.
- [91] L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang, "Speech emotion recognition based on DNN-decision tree SVM model," *Speech Commun.*, vol. 115, pp. 29–37, 2019.
- [92] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Netw.*, vol. 92, pp. 60–68, 2017.
- [93] E. Kim and J. W. Shin, "DNN-based emotion recognition based on bottleneck acoustic features and lexical features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6720–6724.
- [94] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2007, pp. II733–II736.
- [95] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5115–5119.
- [96] D. Bertero, F. B. Siddique, and P. Fung, "Towards a corpus of speech emotion for interactive dialog systems," in *Proc. Conf. Oriental Chapter Internat. Committee Coordination Standardization Speech Databases Assessment Techn.*, 2016, pp. 241–246.
- [97] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 3603–3607.
- [98] R. Xia, J. Deng, B. Schuller, and Y. Liu, "Modeling gender information for emotion recognition using denoising autoencoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 990–994.
- [99] B. Schuller *et al.*, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 2794–2797.
- [100] G. Paraskevopoulos, E. Tzinis, N. Ellinas, T. Giannakopoulos, and A. Potamianos, "Unsupervised low-rank representations for speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 939–943.
- [101] S. E. Eskimez, Z. Duan, and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5099–5103.
- [102] S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Augmenting generative adversarial networks for speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, pp. 521–525, 2020.
- [103] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. Conf.*, pp. 3698–3702, 2018.
- [104] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 148–152.
- [105] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2697–2709, Sep. 14, 2020, doi: [10.1109/TASLP.2020.3023632](https://doi.org/10.1109/TASLP.2020.3023632).
- [106] J.-H. Tao, J. Huang, Y. Li, Z. Lian, and M.-Y. Niu, "Semi-supervised ladder networks for speech emotion recognition," *Int. J. Autom. Comput.*, vol. 16, no. 4, pp. 437–448, 2019.
- [107] S. Bjorn, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. Conf.*, 2009, pp. 312–315.
- [108] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Trans. Affect. Comput.*, to be published, doi: [10.1109/TAFC.2020.2983669](https://doi.org/10.1109/TAFC.2020.2983669).
- [109] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 2746–2750.
- [110] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 3–14, Jan.–Mar. 2017.
- [111] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multi-task learning," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 951–955.
- [112] A. Nediyanachath, P. Paramasivam, and P. Yenigalla, "Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7179–7183.



- [113] F. Tao and G. Liu, "Advanced LSTM: A study about better time dependency modelling in emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2906–2910.
- [114] J. Deng, X. Xu, Z. Zhang, S. Fröhholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 500–504, Apr. 2017.
- [115] J. Deng, X. Xu, Z. Zhang, S. Fröhholz, D. Grandjean, and B. Schuller, "Fisher kernels on phase-based features for speech emotion recognition," in *Dialogues with Social Robots*. Berlin, Germany: Springer, 2017, pp. 195–203.
- [116] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 12, pp. 2423–2435, Dec. 2018.
- [117] A. Shukla, S. Petridis, and M. Pantic, "Does visual self-supervision improve learning of speech representations for emotion recognition," *IEEE Trans. Affect. Comput.*, 2021.
- [118] S. Siriwardhana, T. Kaluarachchi, M. Billinghamurst, and S. Nanyakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176 274–176 285, 2020.
- [119] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 3465–3469.
- [120] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, "EmoRL: Continuous acoustic emotion classification using deep reinforcement learning," in *Proc. IEEE Int. Conf. Robot. Automa.*, 2018, pp. 4445–4450.
- [121] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2017, pp. 163–171.
- [122] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1491–1498.
- [123] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 2227–2231.
- [124] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. Int. Conf. Platform Technol. Serv.*, 2017, pp. 1–5.
- [125] W. Zheng, J. Yu, and Y. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2015, pp. 827–831.
- [126] J. Liu, W. Han, H. Ruan, X. Chen, D. Jiang, and H. Li, "Learning salient features for speech emotion recognition using CNN," in *Proc. 1st Asian Conf. Affect. Comput. Intell. Interact.*, 2018, pp. 1–5.
- [127] J. Kim, K. P. Truong, G. Englebienne, and V. Evers, "Learning spectro-temporal features with 3D CNNs for speech emotion recognition," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact.*, 2017, pp. 383–388.
- [128] L. Zhu, R. Deng, M. Maire, Z. Deng, G. Mori, and P. Tan, "Sparsely aggregated convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 186–201.
- [129] M. Guo, Y. Yang, R. Xu, Z. Liu, and D. Lin, "When NAS meets robustness: In search of robust architectures against adversarial attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 628–637.
- [130] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 2327–2331.
- [131] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. W. Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 1691–1695.
- [132] G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5200–5204.
- [133] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Emotion identification from raw speech signals using DNNs," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 3097–3101.
- [134] X. Wu et al., "Speech emotion recognition using capsule networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6695–6699.
- [135] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2578–2582.
- [136] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 2822–2826.
- [137] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6685–6689.
- [138] S. Latif, "Deep representation learning for improving speech emotion recognition," *Doctoral Consortium, Interspeech*, vol. 2020, 2020. [Online]. Available: <http://www.isca-students.org/sacweb/index.php/resources?id=45>
- [139] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using WaveNet autoencoders," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2041–2053, Dec. 2019.
- [140] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Learning representations of affect from speech," 2105, *arXiv:1511.04747*.
- [141] Z.-W. Huang, W.-T. Xue, and Q.-R. Mao, "Speech emotion recognition with unsupervised feature learning," *Front. Inf. Technol. Electron. Eng.*, vol. 16, no. 5, pp. 358–366, 2015.
- [142] G. Dahl, A.-R. Mohamed, G. E. Hinton et al., "Phone recognition with the mean-covariance restricted Boltzmann machine," in *Proc. Neural Inf. Process. Syst.*, 2010, pp. 469–477.
- [143] H. B. Sailor and H. A. Patil, "Unsupervised deep auditory model using stack of convolutional RBMs for speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 3379–3383.
- [144] L. Li et al., "Hybrid deep neural network-hidden Markov model (DNN-HMM) based speech emotion recognition," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, 2013, pp. 312–317.
- [145] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 7390–7394.
- [146] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [147] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Cross corpus speech emotion classification: an effective transfer learning technique," 2018, *arXiv:1801.06353*.
- [148] X. J. Zhu, "Semi-supervised learning literature survey," *Dep. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep.*, 2005.
- [149] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 801–804.
- [150] J. Deng, X. Xu, Z. Zhang, S. Fröhholz, and B. Schuller, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 1, pp. 31–43, Jan. 2018.
- [151] S. Parthasarathy, V. Rozgic, M. Sun, and C. Wang, "Improving emotion classification through variational inference of latent variables," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 7410–7414.
- [152] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and J. Yi, "Speech emotion recognition using semi-supervised learning with ladder networks," in *Proc. 1st Asian Conf. Affect. Comput. Intell. Interact.*, 2018, pp. 1–5.
- [153] S. Sahu, R. Gupta, and C. Espy-Wilson, "Modeling feature representations for affective speech using generative adversarial networks," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2020.2998118.
- [154] S. Sahu, R. Gupta, and C. Espy-Wilson, "On enhancing speech emotion recognition using generative adversarial networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, pp. 3693–3697, 2018.
- [155] G. He, X. Liu, F. Fan, and J. You, "Image2Audio: Facilitating semi-supervised audio emotion recognition with facial expression image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 3978–3983.
- [156] H. Zhao, Y. Xiao, and Z. Zhang, "Robust semisupervised generative adversarial networks for speech emotion recognition via distribution smoothness," *IEEE Access*, vol. 8, pp. 106889–106900, 2020.



- [157] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, "Adversarial auto-encoders for speech based emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, pp. 1243–1247, 2017.
- [158] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proc. ICML Workshop Unsupervised Transfer Learn.*, 2012, pp. 17–36.
- [159] S. Thrun and L. Pratt, *Learning to Learn*. Berlin, Germany: Springer, 2012.
- [160] 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.
- [161] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 4818–4822.
- [162] J. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.
- [163] Y. Xiao, H. Zhao, and T. Li, "Learning class-aligned and generalized domain-invariant representations for speech emotion recognition," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 4, no. 4, pp. 480–489, Aug. 2020.
- [164] X. Cai, Z. Wu, K. Zhong, B. Su, D. Dai, and H. Meng, "Unsupervised cross-lingual speech emotion recognition using domain adversarial neural network," *Int. Symposium Chinese Spoken Lang. Process.*, 2021, pp. 1–5.
- [165] H. Zhou and K. Chen, "Transferable positive/negative speech emotion recognition via class-wise adversarial domain adaptation," in *Proc. IEEE Int. Conf. Speech Signal Process.*, 2019, pp. 3732–3736.
- [166] A. Batliner, S. Steidl, and E. Nöth, "Releasing a thoroughly annotated and processed spontaneous emotional database: The FAU Aibo emotion corpus," in *Proc. Satell. Workshop LREC*, 2008, pp. 28–31.
- [167] J. Gideon, M. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDog)," *IEEE Trans. Comput.*, to be published, doi: [10.1109/TAFFC.2019.2916092](https://doi.org/10.1109/TAFFC.2019.2916092).
- [168] S. Khorram, M. Jaiswal, J. Gideon, M. McInnis, and E.-M. Provost, "The priori emotion dataset: Linking mood to emotion detected in-the-wild," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 1903–1907.
- [169] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th ACM Int. Conf. Mach. Learn.*, 2008, pp. 160–167.
- [170] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Proc. IEEE Int. Conf. Speech Signal Process.*, 2013, pp. 8599–8603.
- [171] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Vis.*, 2015, pp. 1440–1448.
- [172] R. Caruana, "Learning to learn, chapter multitask learning," *Learning to Learn*. Norwell, MA, USA: Kluwer Acad. Pub., 1998, pp. 95–133.
- [173] M.-R. Bouguelia, S. Pashami, and S. Nowaczyk, "Multi-task representation learning," in *Proc. 30th Annu. Workshop Swedish Artif. Intell. Soc.*, 2017, pp. 53–59.
- [174] J. Kim, G. Englebiene, K. P. Truong, and V. Evers, "Towards speech emotion recognition 'in the wild' using aggregated corpora and deep multi-task learning," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.: Situated Interact.*, 2017, pp. 1113–1117.
- [175] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1103–1107.
- [176] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE Trans. Comput.*, vol. 10, no. 1, pp. 85–99, Jan.–Mar. 2019.
- [177] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 759–766.
- [178] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: [10.1109/TPAMI.2020.2992393](https://doi.org/10.1109/TPAMI.2020.2992393).
- [179] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for self-supervised learning of language representations," *CoRR*, vol. abs/1909.11942, 2019, *arXiv:1909.11942*. [Online]. Available: <http://arxiv.org/abs/1909.11942>.
- [180] A. Baevski, S. Schneider, and M. Auli, "VQ-WAV2VEC: Self-supervised learning of discrete speech representations," *CoRR*, vol. abs/1910.05453, 2019, *arXiv:1910.05453*. [Online]. Available: <http://arxiv.org/abs/1910.05453>.
- [181] A. Shukla, K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, "Visually guided self supervised learning of speech representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6299–6303.
- [182] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 161–165.
- [183] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
- [184] M. Macary, M. Tahon, Y. Estève, and A. Rousseau, "On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 373–380.
- [185] A. Khare, S. Parthasarathy, and S. Sundaram, "Self-supervised learning with cross-modal transformers for emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 381–388.
- [186] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1238–1274, 2013.
- [187] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.
- [188] S. Latif, H. Cuayáhuilitl, F. Pervez, F. Shamshad, H. S. Ali, and E. Cambria, "A survey on deep reinforcement learning for audio-based applications," 2021, *arXiv:2101.00240*.
- [189] H. Cuayáhuilitl, S. Renals, O. Lemon, and H. Shimodaira, "Reinforcement learning of dialogue strategies with hierarchical abstract machines," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2006, pp. 182–185.
- [190] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao, "Deep reinforcement learning for dialogue generation," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2016, pp. 1192–1202.
- [191] K.-F. Lee and S. Mahajan, "Corrective and reinforcement learning for speaker-independent continuous speech recognition," *Comput. Speech Lang.*, vol. 4, no. 3, pp. 231–245, 1990.
- [192] Y.-L. Shen, C.-Y. Huang, S.-S. Wang, Y. Tsao, H.-M. Wang, and T.-S. Chi, "Reinforcement learning based speech enhancement for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6750–6754.
- [193] J. Sangeetha and T. Jayasankar, "Emotion speech recognition based on adaptive fractional deep belief network and reinforcement learning," in *Cognitive Informatics and Soft Computing*. Berlin, Germany: Springer, 2019, pp. 165–174.
- [194] H. Li, B. Baucom, and P. Georgiou, "Unsupervised latent behavior manifold learning from acoustic features: Audio2behavior," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5620–5624.
- [195] F. Locatello *et al.*, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proc. Int. Conf. on Mach. Learn.*, 2019, pp. 4114–4124.
- [196] T. Yang and C. E. Priebe, "The effect of model misspecification on semi-supervised classification," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2093–2103, Oct. 2011.
- [197] A. Singh, R. Nowak, and J. Zhu, "Unlabeled data: Now it helps, now it doesn't," *Adv. Neural Inf. Process. Syst.*, vol. 21, pp. 1513–1520, 2008.
- [198] Y.-F. Li and D.-M. Liang, "Safe semi-supervised learning: A brief introduction," *Front. Comput. Sci.*, vol. 13, no. 4, pp. 669–676, 2019.
- [199] A. Chatziagapi *et al.*, "Data augmentation using GANs for speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 171–175.
- [200] F. Bao, M. Neumann, and N. T. Vu, "CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 2828–2832.

- [201] Z. Gong, P. Zhong, and W. Hu, "Diversity in machine learning," *IEEE Access*, vol. 7, pp. 64 323–64 350, 2019.
- [202] S. Latif, A. Qayyum, M. Usama, J. Qadir, A. Zwitter, and M. Shahzad, "Caveat emptor: The risks of using big data for human development," *IEEE Technol. Soc. Mag.*, vol. 38, no. 3, pp. 82–90, Sep. 2019.
- [203] S. Yin *et al.*, "Noisy training for deep neural networks in speech recognition," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 1, 2015, Art. no. 2.
- [204] F. Bu, Z. Chen, Q. Zhang, and X. Wang, "Incomplete big data clustering algorithm using feature selection and partial distance," in *Proc. 5th Int. Conf. Digit. Home*, 2014, pp. 263–266.
- [205] R. Wang and D. Tao, "Non-local auto-encoder with collaborative stabilization for image restoration," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2117–2129, May 2016.
- [206] C. Breazeal and L. Aryananda, "Recognition of affective communicative intent in robot-directed speech," *Auton. Robots*, vol. 12, no. 1, pp. 83–104, 2002.
- [207] N. Campbell, "Databases of emotional speech," in *Proc. ISCA Tut. Res. Workshop (ITRW) Speech Emotion*, 2000, pp. 34–38.
- [208] S. Asakawa, N. Minematsu, and K. Hirose, "Automatic recognition of connected vowels only using speaker-invariant representation of speech dynamics," in *Proc. 8th Annu. Conf. Int. Speech Commun. Assoc.*, 2007, pp. 890–893.
- [209] Y. Lu *et al.*, "Multitask learning in connectionist speech recognition," in *Proc. Australian Int. Conf. Speech Sci. Technol.*, 2004.
- [210] M. Jaiswal and E. M. Provost, "Privacy enhanced multimodal neural representations for emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 7985–7993.
- [211] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, "Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 342–356, 2021, doi: [10.1109/RBME.2020.3006860](https://doi.org/10.1109/RBME.2020.3006860).
- [212] M. A. Pathak, B. Raj, S. D. Rane, and P. Smaragdis, "Privacy-preserving speech processing: Cryptographic and string-matching frameworks show promise," *IEEE Signal Process. Mag.*, vol. 30, no. 2, pp. 62–74, Mar. 2013.
- [213] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-preserving adversarial representation learning in ASR: Reality or illusion?," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, pp. 3700–3704, 2019.
- [214] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, 2015, pp. 1310–1321.
- [215] S. Latif, S. Khalifa, R. Rana, and R. Jurdak, "Federated learning for speech emotion recognition applications," in *Proc. 19th ACM/IEEE Int. Conf. Process. Sensor Netw.*, 2020, pp. 341–342.
- [216] S. Latif, R. Rana, and J. Qadir, "Adversarial machine learning and speech emotion recognition: Utilizing generative adversarial networks for robustness," 2018, *arXiv:1811.11402*.
- [217] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *Int. Conf. Representations*, 2015.
- [218] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy*, 2016, pp. 372–387.
- [219] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2574–2582.
- [220] Z. Ren, A. Baird, J. Han, Z. Zhang, and B. Schuller, "Generating and protecting against adversarial attacks for deep speech-based emotion recognition models," in *Proc. IEEE Int. Conf. Speech Signal Process.*, 2020, pp. 7184–7188.
- [221] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14 410–14 430, 2018.
- [222] J. Chen, J. Konrad, and P. Ishwar, "A cyclically-trained adversarial network for invariant representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 3393–3402.
- [223] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, Sep. 2019.
- [224] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Improving emotion identification using phone posteriors in raw speech waveform based DNN," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, pp. 3925–3929, 2019.
- [225] P. J. Werbos, "Backpropagation through time: What it does and how to do it," *Proc. IEEE*, vol. 78, no. 10, pp. 1550–1560, Oct. 1990.
- [226] S. Karita *et al.*, "A comparative study on transformer vs RNN in speech applications," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 449–456.
- [227] A. Zeyer, P. Bahar, K. Irie, R. Schlüter, and H. Ney, "A comparison of transformer and LSTM encoder decoder models for ASR," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 8–15.
- [228] J. Li, Y. Wu, Y. Gaur, C. Wang, R. Zhao, and S. Liu, "On the comparison of popular end-to-end models for large scale speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 1–5.
- [229] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, 2019, pp. 6706–6713.
- [230] D. Lim, W. Jang, O. Gyeonghwan, H. Park, B. Kim, and J. Yoon, "JDI-T: Jointly trained duration informed transformer for text-to-speech without explicit alignment," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 4004–4008.
- [231] 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, 2018.
- [232] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 801–804.
- [233] G. Pironkov, S. Dupont, and T. Dutoit, "Multi-task learning for speech recognition: An overview," in *Proc. 29th Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn.*, 2016.
- [234] V. Thomas *et al.*, "Disentangling the independently controllable factors of variation by interacting with the world," *CoRR*, vol. abs/1802.09484, 2018, *arXiv:1802.09484*. [Online]. Available: <http://arxiv.org/abs/1802.09484>
- [235] C. Gelada, S. Kumar, J. Buckman, O. Nachum, and M. G. Belle-mare, "DeepMDP: Learning continuous latent space models for representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2170–2179.
- [236] H. Van Seijen, M. Fatemi, J. Romoff, R. Laroche, T. Barnes, and J. Tsang, "Hybrid reward architecture for reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5392–5402.
- [237] M. Jaderberg *et al.*, "Reinforcement learning with unsupervised auxiliary tasks," *Int. Conf. Learn. Representations*, 2017, pp. 1–17.
- [238] T. Zhang, M. Huang, and L. Zhao, "Learning structured representation for text classification via reinforcement learning," in *Proc. Thirty-Second AAAI Conf. Artif. Intell.*, 2018, pp. 6053–6060.



**Siddique Latif** is currently working toward the PhD degree with the University of Southern Queensland, Australia and Distributed Sensing Systems Research Group, Data61—CSIRO. His PhD work focuses on representation learning, using unlabelled, weakly-labelled, and partially-labelled multimodal data.



**Rajib Rana** is currently an experimental computer scientist, Advance Queensland research fellow, a senior lecturer with the University of Southern Queensland, and the director of IoT Health Research Program with the University of Southern Queensland. His current research focus on unsupervised representation learning.



**Sara Khalifa** received the PhD degree in computer science and engineering from the University of New South Wales, Sydney, NSW, Australia. She is currently a senior research scientist with the Distributed Sensing Systems Research Group, Data61—CSIRO. Her research interests include Internet of Things, smart wearables, energy harvesting, and pattern recognition.



**Björn Schuller** (Fellow, IEEE) is currently a professor of artificial intelligence with the Department of Computing, Imperial College London, U.K., where he heads GLAM — the Group on Language, Audio and Music. He is a full professor, and the head of the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing with the University of Augsburg, Germany, and the CEO of audEERING.



**Raja Jurdak** (Senior Member, IEEE) is currently a professor of distributed systems, the chair in applied data sciences with Queensland University of Technology, and the director of the Trusted Networks Lab. His research interests include trust, mobility and energy-efficiency in networks.

▷ **For more information on this or any other computing topic, please visit our Digital Library at [www.computer.org/csdl](http://www.computer.org/csdl).**



**Junaid Qadir** (Senior Member, IEEE) is currently a professor with the Qatar University in Doha, Qatar, and the Information Technology University of Punjab, Lahore, Pakistan. He directs the IHSAN Research Lab. His primary research interests include computer systems and networking and applied machine learning.