

Multitask learning from augmented auxiliary data for improving speech emotion recognition

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Latif, Siddique, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn W. Schuller. 2023. "Multitask learning from augmented auxiliary data for improving speech emotion recognition." *IEEE Transactions on Affective Computing* 14 (4): 3164–76.
<https://doi.org/10.1109/taffc.2022.3221749>.



Multitask Learning From Augmented Auxiliary Data for Improving Speech Emotion Recognition

Siddique Latif^{ID}, Rajib Rana^{ID}, Sara Khalifa^{ID},
Raja Jurdak^{ID}, *Senior Member, IEEE*, and Björn W. Schuller^{ID}, *Fellow, IEEE*

Abstract—Despite the recent progress in speech emotion recognition (SER), state-of-the-art systems lack generalisation across different conditions. A key underlying reason for poor generalisation is the scarcity of emotion datasets, which is a significant roadblock to designing robust machine learning (ML) models. Recent works in SER focus on utilising multitask learning (MTL) methods to improve generalisation by learning shared representations. However, most of these studies propose MTL solutions with the requirement of meta labels for auxiliary tasks, which limits the training of SER systems. This paper proposes an MTL framework (MTL-AUG) that learns generalised representations from augmented data. We utilise augmentation-type classification and unsupervised reconstruction as auxiliary tasks, which allow training SER systems on augmented data without requiring any meta labels for auxiliary tasks. The semi-supervised nature of MTL-AUG allows for the exploitation of the abundant unlabelled data to further boost the performance of SER. We comprehensively evaluate the proposed framework in the following settings: (1) within corpus, (2) cross-corpus and cross-language, (3) noisy speech, (4) and adversarial attacks. Our evaluations using the widely used IEMOCAP, MSP-IMPROV, and EMOB datasets show improved results compared to existing state-of-the-art methods.

Index Terms—Speech emotion recognition, multi task learning, representation learning

1 INTRODUCTION

SPEECH Emotion Recognition (SER) is an emerging area of research. Speech contains information about human emotions, which can be utilised by machine learning (ML) systems for automatic detection redefining human-computer interactions. SER can help improve the quality of customer service by tracking customer-agent reactions. In healthcare, SER can be used for diagnosis and monitoring of affective behaviours [1], [2]. Service delivery in transport [3], forensics [4], education [5], media [6] can be improved by utilising SER.

Human emotion modelling is quite complex due to its dependency on many factors including speaker [7], gender [8], age [9], culture [10], and dialect [11]. Researchers have explored various ML techniques, including hidden Markov

models, support vector machines, and deep neural networks (DNNs) for SER, wherein DNNs have improved performance compared to the classical ML techniques. Deep belief networks (DBN) [12], convolutional neural networks (CNN) [13], and recurrent neural network (RNNs) have been successful in modelling emotions in speech and widely explored in SER [14], [15], [16], [17]. In particular, RNN architectures like short term memory (LSTM) networks [18] or bidirectional LSTM (BLSTM) combined with CNNs are a popular choice in SER for capturing emotional attributes and have been explored by many researchers [19], [20]. Studies [15], [21] show that the CNN-LSTM can learn better emotional features for SER compared to using CNN or LSTM individually. This work presents a unique semi-supervised configuration using CNN-BLSTM with attention mechanisms. We utilise an attention mechanism in our emotion classifier to combine the important emotional information extracted from the overall utterance and improve emotion classification performance.

Literature shows, SER models lack generalisation due to the single task-specific training and perform poorly when the data mismatch increases between the training and testing phases [22], [23]. Typically, generalisation of deep learning models is improved by training them on diverse data. For example, state-of-the-art models in computer vision are trained on thousands of labelled samples, and automatic speech recognition systems are trained on thousands of hours of transcribed data [24], [25]. In contrast, SER corpora are relatively small, and the creation of emotional corpora is a time consuming and expensive task [22], [26] as emotion is subjective, and several annotators are usually required, which often have to repeatedly go through the speech material to annotate, e. g., affective dimension by affective dimension. To obtain data volume, most existing studies in

- Siddique Latif is with the University of Southern Queensland (UniSQ), Springfield 4300, Australia, and also with Distributed Sensing Systems Group, Data61, CSIRO Australia, Pullenvale, QLD 4069, Australia. E-mail: siddique.latif@usq.edu.au.
- Rajib Rana is with the University of Southern Queensland (UniSQ), Springfield 4300, Australia. E-mail: rajib.rana@usq.edu.au.
- Sara Khalifa is with Distributed Sensing Systems Group, Data61, CSIRO Australia, Pullenvale, QLD 4069, Australia. E-mail: sara.khalifa@data61.csiro.au.
- Raja Jurdak is with Trusted Network Lab, and Applied Data Sciences, Queensland University of Technology (QUT), Brisbane, QLD 4000, Australia. E-mail: r.jurdak@qut.edu.au.
- Björn W. Schuller is with GLAM – The Group on Language, Audio, and Music, Imperial College London, SW7 2BX London, 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.

(Corresponding author: Siddique Latif.)

Digital Object Identifier no. 10.1109/TAFCC.2022.3221749

SER attempt to train models on multiple corpora [10], [27]. However, standard benchmark datasets are also very limited, which creates tremendous barriers to achieving generalisation in SER systems [17].

An alternative technique to improve the generalisation of DL models is multitask learning (MTL) [28], which simultaneously solves the multiple relevant auxiliary tasks along with the primary task. MTL can use different aspects of the same data or get data supplement from the secondary tasks. In this way, models can be better regularised by capturing shared and essential high-level representations, leading to an improved generalisation of the system. MTL has been successfully used in SER by achieving promising performance. However, most of these MTL techniques present *supervised* auxiliary tasks, which require accurate annotations just like the primary emotion recognition tasks. Examples include emotional attributes (i.e., arousal, valence, and dominance) prediction [29], [30], gender identification [22], [31], [32], speaker recognition [22], [33], and secondary emotion learning [34]. The MTL methods with any of the above auxiliary tasks need accurate meta labels that limit the SER models' training. In some scenarios, larger data can be utilised for auxiliary tasks like speaker and gender identification [22]; however, collecting speaker and gender labels is also time- and labour-intensive. This also makes the model's performance speaker-dependent in some cases. Moreover, a generalised representation for SER containing speaker and gender information might be used maliciously without the user's consent by an eavesdropping adversary [35], [36].

In this paper, we propose a semi-supervised MTL framework that learns from augmented data—we call it MTL-AUG. It primarily classifies emotions and utilises *data augmentation-type classification* and *unsupervised reconstruction* as auxiliary tasks to learn generalised representations. We use types of augmentation as labels for *data augmentation for classification* as an auxiliary task. In this way, these auxiliary tasks do not require meta labelling performed by experts. Our idea is inspired by ConvNets, which learn image classification features by predicting the 2D image rotation that is applied to the input image [37]. Such geometric transformation cannot be applied to the speech signal. Therefore, we propose to use speech-based augmentation types that enable multitask training to learn a generalised representation without requiring meta labels. We apply temporal, frequency, and mixup related augmentations to the input speech. This allows the model to learn temporal and frequency related variations applied to the input data through augmentation-type classification as an auxiliary task. Learning the temporal and frequency variations in the data helps the MTL model to improve SER performance. Our second auxiliary task of unsupervised reconstruction acts as a regulariser and improves the quality of learnt representations. Overall, both auxiliary tasks enable the proposed framework to effectively utilise the augmented and unlabelled data to improve the generalisation of the SER system.

Most of previous MTL studies [22], [26], [29], [30], [32], [33], [38] evaluate the proposed models in within-corpus SER, and very few studies perform cross-corpus and cross-language SER. Moreover, none of these studies performs evaluations in noisy and adversarial attack settings. This is mainly due to the complexity of mismatch conditions in

noisy and adversarial attacks. To show the advantage of our proposed MTL framework, we rigorously evaluate it against noisy and adversarial conditions. For evaluation, we use three widely used emotional databases: The interactive emotional dyadic motion capture (IEMOCAP) [39] database, MSP-IMPROV [40], and the EMODB data. We compare our framework's performance with multiple recent studies and baseline CNN-BLSTM implementations. The comparative results in within-corpus, cross-corpus, cross-language, noisy and adversarial settings show that the proposed MTL-AUG framework achieves considerably improved performance, which attests to the strong generalisation power of the proposed MTL-AUG framework.

2 RELATED WORK

2.1 Multi-Task Learning for SER

Multitask learning (MTL) [28] aims to improve the generalisation of models by learning the similarities and differences among the given tasks from the training data. It has been successful to produce shared representation by simultaneously modelling multiple related tasks. The conventional single task learning technique ignores the information of related tasks and can increase the risk of overfitting [23]. In contrast, MTL acts as a regulariser to reduce the risk of overfitting by introducing an inductive bias. Several MTL approaches [41], [42], [43] have been exploited in computer vision to address various problems with significantly improved results. The speech community also explored MTL approaches to improve the performance of the tasks, including automatic speech recognition [44], speaker identification [45], and also emotion classification [46].

Eyben et al. [48] were the first to explore MTL in SER. They empirically found that multi-task training of models help improve performance in contrast to single-task training. Xia et al. [29] presented a DBN based MTL model for SER and utilised activation and valence labels as an auxiliary task. They demonstrated that the performance of SER for categorical emotion could be enhanced using activation and valence label information as auxiliary tasks. Parthasarathy and Busso [30] presented a DNN-based MTL model that jointly learns the arousal, dominance, and valence from a given utterance. The authors found that joint training of the model with multiple emotional attributes enhances the performance compared to training with single attribute information. Ma et al. [49] used a multitask attention-based DNN model for SER and showed that a high performance could be achieved by optimising the model for joint classification of categorical emotions along with valence and activation labels classification. Similarly, Lotfian et al. [34] utilised a DNN based framework for modelling primary and secondary emotions. Based on the results, the authors showed that the performance of the primary classification task (categorical emotions) is enhanced by utilising the information of secondary emotions and emotional classes perceived by the evaluators.

Another way to implement MTL in SER is to use speaker and gender identification as auxiliary tasks. Multiple studies have explored this phenomenon to improve SER performance. In [38], the authors presented an LSTM-based MTL framework that uses speaker and gender classification as

auxiliary tasks to improve the performance of the main task, emotion classification. In another study [22], the authors proposed an MTL framework that uses speaker and gender recognition as auxiliary tasks and used other speech corpora with speaker and gender labels and injected this data into the model. They showed that the performance could be significantly improved. Kim et al. [47] utilised gender and naturalness (natural or acted corpus) recognition as auxiliary tasks and evaluated the model using different corpora. They found that a performance gain can be achieved using gender or naturalness classification as auxiliary tasks. Other recent studies also utilised [8], [32] gender-aware MTL SER models and found that emotion classification can be improved with additional gender label information.

Previous studies on MTL demonstrate that the use of auxiliary tasks helps improve SER performance compared with STL. However, these approaches either use information about emotional attributes (activation, valence, etc.) or non-emotional attributes (speaker, gender, etc.) that are not widely available in real-life. Also, labelling speech data with such meta-information is a cumbersome and expensive process. Some studies [22], [26] exploit the unsupervised reconstruction as auxiliary tasks; however, they also require additional labels for emotional attributes [26], and gender and speaker labels in [22] for their MTL frameworks.

In contrast to previous studies, we propose an MTL framework that improves the performance without requiring such meta labels by annotators. We propose using data transformation (or augmentation)-type recognition and unsupervised feature reconstruction as auxiliary tasks. This allows us to utilise the type of augmentation applied to the input data as labels for the auxiliary task to train the proposed MTL framework.

2.2 Data Augmentation in SER

Data augmentation techniques are widely being used as a training trick in deep learning to improve the network generalisation ability. The main limitation of data augmentations is that it enhances the data bias, if the original data has biases. This data bias leads to a suboptimal performance. Data augmentation techniques have been used to generate additional training data for SER. For example, studies [20], [50] show that the speed perturbation [51] data augmentation technique can improve the performance of an SER system by generating copies of each utterance with different speed effects. The mixup [52] technique augments an SER system by generating the synthetic sample as a linear combination of the original sample. In SER, Latif et al. [15] augment the SER system with mixup to achieve robustness against noisy conditions. They showed that augmentation techniques make the training data diverse and help improve performance. A new method of data augmentation is SpecAugment [53] and was proposed for automatic speech recognition, which is directly applied to the feature inputs of a neural network. In [54], the authors utilised the SpecAugment technique to augment their SER system with the duplicate samples by a factor of two. The authors highlighted that the data augmentation improves the robustness of the model by providing diverse training samples. Other studies [20], [50],

[55] also achieve improved performance by exploiting data augmentation techniques to increase the training data. However, these studies only utilised the data augmentation in single-task learning to increase the training samples. In this paper, we propose to use data augmentation-type recognition as our auxiliary task in our proposed multitask learning framework. We hypothesise that multitask learning models are able to understand the concept of emotions while recognising the transformation performed on the input signal.

2.3 SER Robust to Adversarial Attacks and Noise

In SER, it is essential to achieve robustness against perturbation/noise added to the input samples. However, very few studies focus on evaluating SER systems' robustness against noisy conditions and adversarial attacks. Huang et al. [56] used a CNN-LSTM model for robust SER. They found that CNN demonstrates a certain degree of noise robustness. In [57], the authors utilised deep residual networks for speech enhancement to remove noise from speech while preserving emotions for SER. Some other studies [58], [59] also explored different noise removal frameworks for SER in noisy environments instead of achieving robustness by learning generalised representation. Based on the findings of data augmentation techniques to improve robustness [60], [61], a recent study [15] evaluated the regularising effect of data augmentation to improve the robustness of SER. They show that data augmentation helps to improve the robustness of SER against noise and adversarial attacks. However, no study has evaluated data augmentation in MTL scenarios to learn generalised representation to improve robustness in SER.

2.4 Summary

We summarise the differences between our work and the existing literature in Table 1.

- 1) While some studies used reconstruction as an auxiliary task, no studies used augmentation-type classification as the auxiliary task.
- 2) None of the studies evaluated their models' generalisation ability against noisy conditions and adversarial attacks.
- 3) Most of the studies evaluated their model within-corpus settings by using training and testing data from the same corpus. Only a few studies evaluated the generalisation of proposed models in cross-corpus and even less in cross-language settings.

3 METHODOLOGY

The proposed MTL-AUG framework uses the *augmentation-type classification* and *unsupervised reconstruction* as auxiliary tasks to learn generalised emotional representations. Before we describe our framework, we briefly introduce speech data augmentation, especially the techniques used for this work.

3.1 Speech Data Augmentation

We use augmentation to introduce variability and volume in the data. Speech signals can be augmented/transformed

TABLE 1
Summary of a Comparative Analysis of Our Paper With That of the Existing Literature

Paper/Author (Year)	Label dependent auxiliary tasks	Label independent auxiliary tasks			Evaluations			
		Reconstruction	Augmentation- type classification	within- corpus	Cross- corpus	Cross- language	Noisy conditions	Adversarial attacks
Prthasarathy and Busso [30] (2017)	emotional attributes prediction	✗	✗	✓	✓	✗	✗	✗
Xia et al. [29] (2017)	emotional attributes prediction	✗	✗	✓	✓	✗	✗	✗
Kim et al. [47] (2017)	emotional attributes prediction +gender identification	✗	✗	✓	✓	✓	✗	✗
Lotfian et al. [34] (2018)	emotional attributes classification	✗	✗	✓	✗	✗	✗	✗
Tao et al. [38] (2018)	speaker classification +gender classification	✗	✗	✓	✗	✗	✗	✗
Li et al. [32] (2019)	gender identification	✗	✗	✓	✗	✗	✗	✗
Prthasarathy and Busso [26] (2020)	emotional attributes prediction	✓	✗	✓	✓	✗	✗	✗
Latif et al. [22] (2020)	speaker classification +gender classification	✓	✗	✓	✓	✗	✗	✗
Peri et al. [33] (2021)	speaker identification	✗	✗	✓	✗	✗	✗	✗
Our Paper (2022)	None	✓	✓	✓	✓	✓	✓	✓

using different techniques. We use the following three techniques: (1) speed perturbation [51], (2) mixup [52], and (3) SpecAugment [53]. We select these augmentation techniques due to their popularity in the speech domain and particularly effectiveness in SER supported by previous studies [15], [50], [54].

Speed perturbation. is a very popular and widely used audio augmentation technique that produces a warped time signal. Given a speech signal $x(t)$, time warping is performed by a factor α to produce the signal $x(\alpha t)$. In this way, speed perturbation changes the duration of a given speech signal. It can be applied directly on raw speech as we use in this paper.

SpecAugment. is used as a simple data augmentation method for Automatic Speech Recognition (ASR). It acts on the log-Mel spectrogram directly with a negligible amount of additional computational cost [53]. In SpecAugment, training data can be augmented using spectro-temporal modifications to the original spectrograms by applying frequency and time masks. In frequency masking, a mask of size f is chosen from a uniform distribution (0 to F) and consecutive log-Mel frequency channels $[f_0, f_0+f)$ are masked, where f_0 is chosen from $[0, v-f)$ and v represents the number of Mel-frequency channels. In the time masking, a mask size of t is chosen from a uniform distribution from 0 to T , and the consecutive time steps $[t_0, t_0+t)$ are masked in time – here, t_0 is chosen from $[0, \tau-t)$ and τ represents log-Mel spectrogram time steps.

Mixup. generates an augmented sample and its label by randomly mixing two inputs and their corresponding labels. This regularises the neural network to favour simple linear behaviour in-between training samples. It constructs augmented training examples as follows:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \quad (1)$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j, \quad (2)$$

where (x_i, y_i) and (x_j, y_j) are randomly selected two examples from training data, and $\lambda \in [0, 1]$. Mixup can be applied on the features as well as on the raw speech [52]. We use mixup on Mel-spectrograms.

Data augmented using the above three techniques are fed to the proposed MTL-AUG framework to learn temporal, frequency, and mixup related changes applied to the data through the augmentation-type classification as auxiliary task. Note that, in SER, it is always important to capture spectro-temporal dynamics to accurately identify speech emotions [15], [62], [63]. In our proposed framework, we model spectro-temporal and augmentation related dynamics through auxiliary tasks in an MTL setting, which helps improve the performance of the primary emotion classification task. We will explain our proposed MTL-AUG framework next.

3.2 MTL-AUG Framework

Fig. 1 describes the proposed semi-supervised MTL architecture. Overall, the framework has four subnetworks: (1) encoder E , (2) decoder D , (3) emotion classifier C_E , and (4) augmentation-type classifier C_A . The proposed model is trained with MTL loss:

$$\mathcal{L}_{\text{MT}} = \mathcal{L}_{\text{pri}} + \lambda_1 \mathcal{L}_{\text{aux}}, \quad (3)$$

where \mathcal{L}_{pri} and \mathcal{L}_{aux} represent the primary and auxiliary tasks, respectively. λ_1 is a hyper-parameter trading off primary and auxiliary tasks.

Our primary task is optimised with an emotion classifier C_E that takes the encoded representation (Z) by the encoder (E) network to perform an emotion classification. It uses BLSTM layers for contextual modelling and an attention layer to combine the most salient features given to a dense layer for discriminative feature representation before classification. For a given output sequence h_i , utterance level

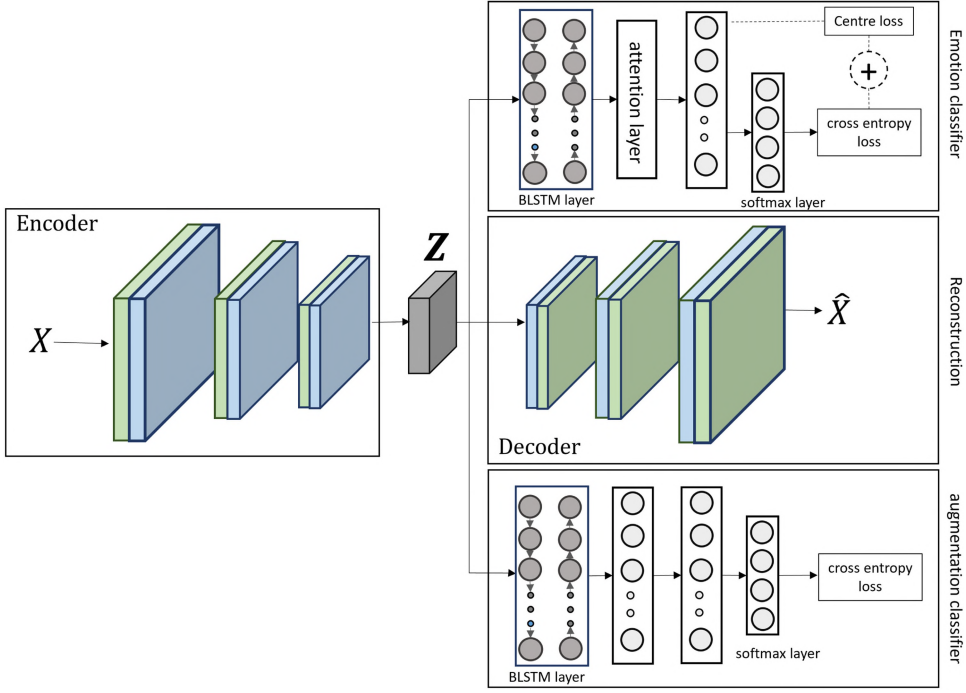


Fig. 1. Illustration of our proposed multitask framework for SER, which uses augmentation-type classification and reconstruction as auxiliary task to achieve better performance on the primary emotion classification task.

important features are computed by the attention layer using:

$$R_{\text{attentive}} = \sum_i \alpha_i h_i, \quad (4)$$

where α_i represents the attention weights that can be computed as follows:

$$\alpha_i = \frac{\exp W^T h_i}{\sum_j \exp W^T h_j}, \quad (5)$$

where W is a trainable parameter. The output attentive representation $R_{\text{attentive}}$ computed by the attention layer is fed to the dense layer for emotion classification. Our intuition of using the attention layer for SER is that the emotional content is distributed over the speech utterances. The attention layer weighs information extracted from different pieces of utterance and combines them into a weighted sum that helps produce better emotion classification performance [16]. The emotion classifier (C_E) is optimised using the sum of cross-entropy and centre loss functions:

$$\mathcal{L}_{\text{pri}} = \mathcal{L}_S + \lambda_2 \mathcal{L}_C, \quad (6)$$

where \mathcal{L}_S and \mathcal{L}_C represent softmax cross-entropy loss and centre loss, respectively. λ_2 is the trade-off parameter between these two losses. The use of centre loss helps to minimise intra-class variations while maintaining separation between features of different classes by pulling them closer to their correspondence centres. The centre loss function can be defined as:

$$\mathcal{L}_C = \frac{1}{2} \sum_{i=1}^m \|f(x_i) - c_{y_i}\|_2^2, \quad (7)$$

where $f(x_i)$ represents the deep features extracted from the last hidden layer and $c_{y_i} \in \mathbb{R}$ denotes y_i^{th} class centre of the deep features.

The secondary tasks in our framework are augmentation-type classification and reconstruction of the input speech features. In the reconstruction auxiliary task, the encoder and decoder networks minimise the reconstruction loss. The objective function for the autoencoder is:

$$\mathcal{L}_{\mathcal{AE}}(x, D_\theta(E_\theta(x))) = \|X - \hat{X}\|_2^2. \quad (8)$$

The other auxiliary task is to classify the transformation applied to the input. For this, we use classifier C_A that takes the encoder E output ($Z = E_\theta(x)$) and performs classification. We created augmented data by applying speed perturbation on raw speech, and the SpecAugment and mixup techniques to the Mel-spectrogram of emotional data samples. In augmentation-type classification, we also consider samples with no augmentation as one class. Thus, classifier C_A is trained on the four-way classification task of recognising one of the four classes (i. e., speech perturbation, SpecAugment, mixup, and no augmentation).

The proposed framework is trained in a semi-supervised way as it uses both unsupervised and supervised learning [64]. For the input X , the encoder network creates the latent code, which is an unsupervised process. The latent code is then used by the classifiers (C_A, C_E) with labels conforming to supervised learning. Note here that when using additional auxiliary data with no labels for emotion, the loss functions for augmentation-type classification and the autoencoding network are only calculated to update the encoder network.

4 EXPERIMENTAL SETUP

4.1 Datasets

To evaluate the performance of our MTL-AUG model, we use three different datasets: IEMOCAP, EMODB, and MSP-IMPROV, which are commonly used for speech emotion classification research [65], [66]. Both, the IEMOCAP and the MSP-IMPROV datasets are collected by simulating naturalistic dyadic interactions among professional actors and have similar labelling schemes. EMODB contains audio samples in the German language, and we use it for cross-language evaluations. In order to use additional data for auxiliary tasks, we use the LibriSpeech [67] dataset.

4.1.1 IEMOCAP

This is a multimodal database containing 12 hours of recorded data [39]. The recordings were collected during dyadic interactions from 10 professional actors (five males and five females). Dyadic interactions allowed the actors to perform spontaneous emotion in contrast to reading text with prototypical emotions [68]. Each interaction is around five minutes long and segmented into smaller utterances of sentences. Each sentence is annotated by the participant and three annotators for categorical labels. Finally, an utterance is assigned a label if at least three annotators assigned the same label. Overall, this corpus contains nine emotions including angry, disgust, fearful, frustrated, sad, happy, excited, surprised, and neutral. Similar to prior studies [14], [20], [22], we use utterances of four categorical emotions, including angry, happy, neutral, and sad in this study by merging “happy” and “excited” as one emotion class “happy”. The final dataset includes 5531 utterances (1103 angry, 1708 neutral, 1084 sad, and 1636 happy).

4.1.2 MSP-IMPROV

This corpus is a multimodal emotional database recorded from 12 actors performing dyadic interactions [40], similar to IEMOCAP [39]. The utterances in MSP-IMPROV are grouped into six sessions, and each session has recordings of one male, and one female actor. The scenarios were carefully designed to promote naturalness while maintaining control over lexical and emotional contents. The emotional labels were collected through perceptual evaluations using crowdsourcing [69]. The utterances in this corpus are annotated in four categorical emotions: angry, happy, neutral, and sad. To be consistent with previous studies [20], [70], we use all utterances with four emotions: anger (792), sad (885), neutral (3477), and happy (2644).

4.1.3 EMODB

EMODB [71] is a popular and most widely used publicly available emotional dataset in the German Language. This corpus was recorded by the Institute of Communication Science, Technical University Berlin. EMODB contains audio recordings of seven emotions recorded by ten professional speakers in 10 German sentences. In this work, we select four basic emotions: angry, sad, neutral, and happy, to perform categorical cross-language SER as executed in [72]. Overall, we use 420 utterances for angry (127), sad (143), neutral (79), and happy (71) emotions.

4.1.4 LibriSpeech

The LibriSpeech dataset [67] contains 1 000 hours of English read speech from 2 484 speakers. This corpus is derived from audiobooks and is commonly used for automatic speaker and speech recognition tasks [73], [74]. The training portion of LibriSpeech is divided into three subsets, with an approximate recording time of 100, 360, and 500 hours. Here, we choose the subset that contains 100 hours of recordings and use it as additional unlabelled data. These recordings span over 251 speakers.

4.1.5 DEMAND

We select the Diverse Environments Multichannel Acoustic Noise Database (DEMAND) dataset [75] as a source of our noise signal. DEMAND contains audio recordings of various real-world noises recorded in various indoor and outdoor settings. In our experiments, we select noise recordings with 16 kHz sampling rate to match with that of the audio recording of the speech emotion datasets.

4.2 Features and Augmentation-Types

We represent the speech utterances in log-Mel spectrograms, which is a popular 2D feature representation widely used for speech-related tasks, including SER. We apply overlapping Hamming windows with a size of 40 ms and with a 10 ms window shift. The height of the log-Mel spectrogram is 128. We set the length of utterances to 7.5 s. Longer utterances are cut at 7.5 s, and smaller utterances are padded with zeros. We select the length of the utterances based on validation results and previous studies [20], [32]. We remove the silence from the start and end of utterances.

As outlined above, we apply three augmentation-types, including speed perturbation, mixup, and SpecAugment. For the speed perturbation, we create two copies of each training utterance by applying the speed effect at 0.9 and 1.1. We apply speed perturbation on the raw speech using the Sox¹ audio manipulation tool, while we apply mixup and SpecAugment on the Mel spectrogram. We utilise each augmentation technique to increase the size of training by a factor of 2. Overall, the three augmentations increase training data by factor 4.

4.3 Hyperparameters

For all the experiments, we use the Adam optimiser with default parameters. We start training models with a learning rate of 0.0001 and calculate the validation accuracy at the end of each epoch. If the validation accuracy does not improve after five consecutive epochs, we halve the learning rate and restore the model to the best epoch. This process continues until the learning rate reaches below 0.00001. We apply a rectified linear unit (ReLU) as a non-linear activation function type, as it gave us better performance than leaky ReLU and hyperbolic tangent during validation.

Our *baseline model* consists of the convolutional encoder network and Bidirectional LSTM (BLSTM)-based classification network. CNN layers in the encoder network produce the high-level feature representations. We use a larger kernel size for the first convolutional layer and reduce the

1. <http://sox.sourceforge.net>

kernel size in the remaining layers, as suggested by previous studies [76], [77]. Feature representations learnt by the encoder network are given to the BLSTM layer with 128 LSTM units for emotional context modelling. After the BLSTM layer, we apply an attention layer to aggregate the emotional content distributed over the different parts of the given utterance. The attentive features are fed to the fully connected layer with 128 hidden units to produce emotionally discriminative features for a softmax layer. The softmax layer uses the crossentropy loss to produce the posterior class probabilities by enabling the network to learn separable features. In addition, we also exploit the centre loss to reduce the features' intra-class variation to improve the classification performance.

In contrast to the baseline model, our MTL-AUG model contains two additional components: the decoder and augmentation-type classifier. The decoder network is used to reconstruct the input log-Mel spectrograms back from the encoded output by the encoder network. It has a similar architecture to the encoder, replacing convolutional layers with the transposed convolutional layers. The augmentation-type classifier takes the encoded representation and uses a BLSTM based classifier to classify different augmentation-types. We use one BLSTM layer with 256 LSTM units and two fully connected layers with 128 hidden units for auxiliary task classification. In addition, we use a dropout layer with a dropout rate of 0.3 between two dense layers. We decide on the dropout rate based on validation experiments. We conduct statistical analysis of results using a two-tailed t-test over 15 subsets in the test data. We randomly split the test data into 15 small subsets. Statistical significance is defined at $p = 0.05$.

For augmentation selection, we use the validation data during experimentation. In addition to the speed perturbation, mixup, and SpecAugment, we explored noise addition and pitch changing as potential augmentation techniques. However, we achieved an improved result on the validation set using speed perturbation, mixup, and SpecAugment.

5 EXPERIMENTS AND RESULTS

All the experiments are performed in a speaker-independent manner. In particular, we follow a easily reproducible leave-one-speaker-out cross-validation scheme commonly used in the literature [14], [22]. For cross-language SER, we follow [47], [72] and use IEMOCAP and EMODB for a four-class emotion classification task. We use LibriSpeech as additional unlabelled data; results are presented in this section as "MTL-AUG (additional data)". For all the experiments, we repeated each experiment ten times and calculated the mean and standard deviation. Results are presented using the unweighted average recall rate (UAR), a widely accepted metric in the field.

5.1 Within Corpus Experiments

For the within-corpus setting, we compare the performance of the proposed model with the baseline. We also extend our evaluation by comparing the results with different single-task learning (STL) and multi-task learning approaches [22], [46], [78] in Table 2. Our proposed MTL-AUG achieves better results than the baseline CNN-BLSTM architecture,

TABLE 2
Comparison of Results (UAR %) of Our Proposed MTL-AUG Framework With Those of Recent MTL Studies

Model	IEMOCAPMSP_IMPROV	
Graph Convolution Network (STL) [46]	62.2	55.42
3D Convolution Network (STL) [46]	62.7	55.7
DBN (MTL) [46]	62.2	-
Attentive CNN (MTL) [78]	60.15	-
CNN (MTL) [22]	65.6±2.0	59.5±2.4
Semi-supervised AAE (MTL) [22]	66.7±1.4	60.3±1.1
CNN-BLSTM _(baseline) (STL)	64.3±1.9	57.2±2.1
CNN-BLSTM _(baseline) \$STL + augmentations)	65.1±1.8	58.5±1.7
MTL-AUG	68.1±1.5*	61.4± 0.9*
MTL-AUG (additional data)	68.7±1.3*	62.1± 1.2*

MTL-AUG (additional data) represents when additional unlabelled data from LibriSpeech is used. An asterisk denotes statistical significant results ($p = 0.05$).

and other STL and MTL approaches. Some studies [46], [78] use dimensional emotion prediction as a secondary task to improve the classification of categorical emotions. They use additional information labels annotated by experts for dimensional emotions to perform an auxiliary task in their MTL frameworks. In another MTL study, [22], speaker and gender identification are used as secondary tasks for shared generalised representation learning with multitasking semi-supervised adversarial autoencoder (SS-AAE). The authors also exploit the additional unlabelled data for the auxiliary task to boost the primary emotion classification task. However, this model also requires additional labels for speaker and gender and cannot exploit unlabelled data without this meta information. In contrast, we can utilise any speech data in the system without requiring information about the speaker and gender. In Table 2, we present these results with MTL-AUG (additional data) that performs augmentation-type classification and reconstruction as the auxiliary tasks on the additional speech from LibriSpeech to learn generalised representations. As our proposed auxiliary tasks do not require additional annotation by experts, it makes the MTL training more practical, yet better performing than the existing studies. We also present class-wise performance in Table 3.

5.2 Cross-Corpus and Cross-Language Evaluations

5.2.1 Cross-Corpus

In this experiment, we perform a cross-corpus analysis to verify the generalisability of the proposed framework. We trained models on IEMOCAP, and testing is performed on the MSP-IMPROV data. We choose IEMOCAP as training data, since it is more balanced than other corpora. The other reason to select this scheme is for comparison with existing studies, which decided for a similar training [22], [78], [79]. We select 30 % of the MSP-IMPROV data for parameter selection and 70 % as testing data. The training and testing data are randomly selected.

We compare our results with different studies in Table 4. In [78], the authors utilise the representations learnt from unlabelled data and feed it to an attention-based multitask CNN classifier. They show that the classifier's performance

TABLE 3
Class-Wise Performance and F1 Scores for IEMOCAP and MSP-IMPROV Using MTL-AUG With Additional Data

true label	IEMOCAP				F1	MSP-IMPROV				F1
	predicted label					predicted label				
	neutral	happiness	sadness	anger		neutral	happiness	sadness	anger	
neutral	64.5	19.3	10.1	6.1	62.7	59.7	19.5	13.8	7.0	57.8
happiness	15.8	65.2	11.5	8.5	63.9	9.2	68.6	7.7	14.5	66.9
sadness	12.6	7.5	73.2	6.7	71.6	22.7	11.2	56.8	9.3	56.4
anger	8.1	13.2	6.9	71.8	67.2	10.3	15.6	10.8	63.3	61.9

can be improved by using the representations from unlabelled data. In [79], the authors use the synthetic data generated by a generative adversarial network (GAN) to augment the emotional classifier. They show that augmentation can improve the generalisation that leads to performance improvement. A recent study [22] utilised a semi-supervised AAE in an MTL setting to improve the generalisation of SER systems. They use supervised auxiliary tasks, including speaker and gender identification. The authors show that the generalisation of SER systems can be improved by learning the speaker and gender information from the data. In contrast, our proposed MTL-AUG framework learns the generalised representations from the augmented data by learning augmentation-types changes applied to the data. These generalised representations help achieve improved results for cross-corpus SER.

5.2.2 Cross-Language

We also evaluate the MTL-AUG setup on cross-language SER. For this experiment – as outlined above – we use the IEMOCAP and EMODB corpora. We compare the results with [47] for cross-language SER, where the authors used a multitask LSTM model with gender and naturalness as auxiliary tasks. The results of the comparison are presented in Table 5. We train the models on IEMOCAP (English), and EMODB (German) is used for validation and testing for four class emotion classification. Similar to the cross-corpus experiments, we also achieve improved results for cross-language SER.

5.3 Evaluation of Robustness to Noise

In this experiment, we evaluate the proposed model in noisy conditions. We compare our results with a recent study [15] that applies a deep architecture to learn a robust representation and exploits a combination of mixup and speed perturbation

data augmentation techniques to achieve improved generalisation. We consider the same settings chosen in [15] and train the model on clean data and evaluate on noisy samples. For a fair comparison with [15], we select the same three signal-to-noise ratio (SNR) values [0, 10, 20] and select six noises, including kitchen, park, cafeteria, station, traffic, and bubble noise [80]. These noises are randomly added to the testing data at three SNR values [0, 10, 20]. We also implemented models used [56], [81] for robust SER to extend our comparison scope. In [56], authors use attentive CNN-BLSTM model for robust SER. Similarly, authors in [81] use attention based CNN model to perform noise robust SER. Results on the IEMOCAP data are compared with [15], [56], [81] and the baseline in Table 6.

In contrast to the deep networks used in [15], [56], [81] and baseline, we achieve better results. This shows that the proposed MTL approach enables the MTL-AUG to learn generalised representations, which help achieve robustness to perform SER in noisy conditions. Both “baseline (+augmentation)” and the deep DenseNet used in [15] are trained in STL setting exploiting the augmented data. We show in Table 7 that training the STL model with augmented data helps improve robustness against noisy conditions; however, these models do not have access to the latent information available in the augmented data. We use this extra information in our proposed MTL-AUG model, where we perform augmentation-type classification as an auxiliary task to exploit the augmented data in the MTL setting.

5.4 Adversarial Attacks

In adversarial settings, we choose two adversarial attacks, including the Fast Gradient Sign Method (FGSM) [82] and

TABLE 5
Cross-Language Evaluation Results (UAR %) for Emotion Recognition

TABLE 4
Cross-Corpus Evaluation Results for Emotion Recognition

Model	UAR (%)
Attentive CNN (MTL) [78]	45.7
Conditional-GAN (STL) [79]	45.4
Semi-supervised AAE (MTL) [22]	46.4±0.32
CNN-BLSTM (STL) _(baseline)	45.4±0.83
CNN-BLSTM (STL) _(baseline) (+ augmentations)	46.2±1.3
MTL-AUG	47.2±0.41*
MTL-AUG (additional data)	48.1±0.30*

An asterisk denotes statistical significant results ($p = 0.05$).

Model	IEMOCAP (English) to EMODB (German)	EMODB (German) to IEMOCAP (English)
MTL-LSTM [47]	43.4±1.8	39.1±1.6
CNN-BLSTM (STL) _(baseline)	42.1±1.9	38.4±1.8
CNN-BLSTM (STL) _(baseline) (+ augmentations)	43.6±1.5	39.5±1.7
MTL-AUG	45.7±1.3*	42.1±1.6
MTL-AUG (additional data)	46.8±1.4*	41.5±1.6*

An asterisk denotes statistical significant results ($p = 0.05$).

TABLE 6
Comparing the Proposed Model Against Noisy Condition With State-of-the-Art Architectures

Model	UAR (%)		
	0 dB	10	20
DenseNet (STL) (+augmentations) [15]	33.8± 1.2	41.7±1.5	43.1 ±1.1
CNN-BLSTM +attention (STL) [56]	34.1±1.2	39.8±1.5	41.6 ±1.5
CNN +attention (STL) [81]	33.4±1.8	39.5±1.9	41.2 ±1.6
CNN-BLSTM (STL) _(baseline)	33.5±1.5	39.5±1.6	41.7 ±1.5
CNN-BLSTM (STL) _(baseline) (+ augmentations)	35.2±1.3	41.5±1.5	42.9 ±1.6
MTL-AUG	37.8±1.0*	43.1±1.4	44.8±1.3*
MTL-AUG (additional data)	40.0±1.2*	44.3±1.3*	46.2±1.4*

An asterisk denotes statistical significant results ($p = 0.05$).

the Basic Iterative Method (BIM) [83] to evaluate the robustness of MTL-AUG. FGSM generates adversarial samples by adding a scaled perturbation in the direction of the gradient of the loss function. The BIM attack builds upon the FGSM attack by applying it multiple times iteratively with small ϵ instead of applying the adversarial noise in a single step. We apply these two attacks with the perturbation factor $\epsilon = 0.08$, and the performance is reported in Table 7. We compare our results with that of [15], where the authors consider the same adversarial attacks. In addition, we also use the implementation of robust models use in [56], [81] for evaluation against adversarial attacks. Comparisons show that we achieve better performance than these existing studies.

In [15], the authors develop a deep architecture to learn a robust representation. In addition, they utilise speed perturbation and mixup augmentation in the STL setting to achieve generalisation. In contrast, we select augmentation-type classification as an auxiliary task in the MTL scenario. This facilitates generalisation in the network by learning the common representations for both primary and auxiliary tasks.

5.5 Selection of Data Augmentation

In this experiment, we evaluate the model using different schemes in the auxiliary task of augmentation-type classification. We start with single augmentation and perform binary classification (augmented or not augmented) in the auxiliary task using different data augmentation techniques. Results are plotted in Fig. 2, which highlight that the performance of the MTL model with a single augmentation-type in the augmentation-type classifier is poorer than using multiple augmentation-types classification. This shows that giving the model more diverse augmented data helps to learn generalised representations compared to learning to classify single data augmentation.

5.6 Size of Labelled Data

In this experiment, we change the amount of labelled data for training the models, and the results are compared with a semi-supervised AAE (SS-AAE) [22]. We present the outcomes on IEMOCAP and MSP-IMPROV in Fig. 3. We plot the results with different percentages of labelled training data. The proposed framework improves the SER performance considerably against the baseline CNN-BLSTM. We

also compare the results with SS-AAE [22] on the SER performance. Results are plotted in Fig. 3, where the red dot shows the performance achieved by SS-AAE [22] using 100 % of source data along with the unlabelled data of Libri-Speech. We achieve similar performance using 80-86 % of labelled training data as highlighted by a dotted blue line. This shows that the proposed MTL-AUG effectively learns the emotional representation from augmented data to improve the performance while reducing the required labelled data.

5.7 Ablation Experiments

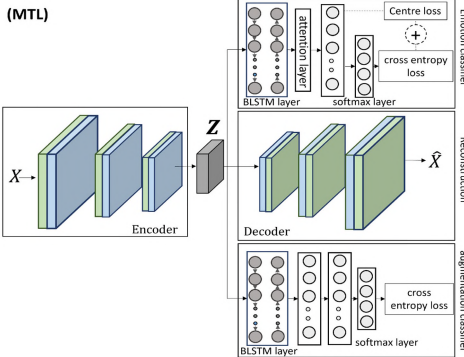
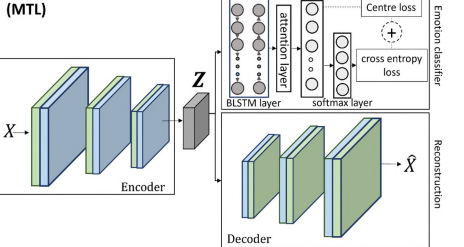
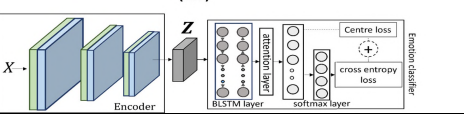
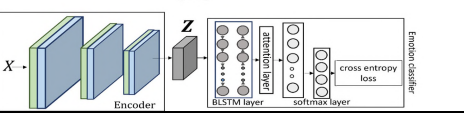
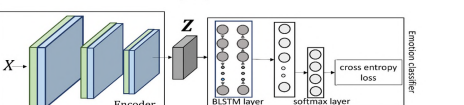
In this experiment, we validate the necessity and effectiveness of each module integrated with our proposed framework. Results are presented in Table 8. This experiment starts with the proposed framework containing all components, including the attention layer, centre loss, auxiliary augmentation-type classifier, and reconstruction decoder. We remove the auxiliary augmentation-type classifier and reconstruction decoder in models 2 and 3. We keep removing different components until we obtain a simple CNN-BLSTM (model 5) classifier without the attention, centre loss, augmentation-type classifier, and reconstruction decoder. We use model 4 as baseline classifier in other Sections 5.1, 5.2, 5.3, 5.4, 5.5, and 5.6. There is a considerable drop in UAR (%) when one or more modules are removed from the proposed framework. When an STL CNN-BLSTM classifier (module 5) is used, we see a considerable performance drop for both within and cross-corpus SER. This

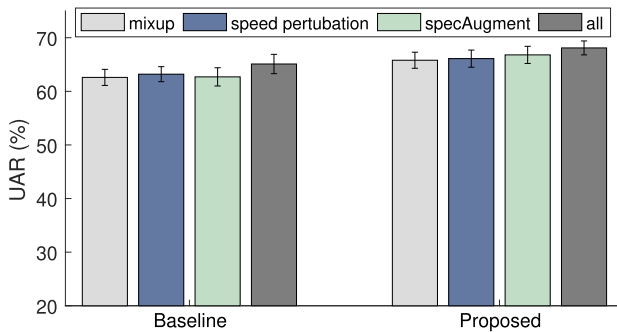
TABLE 7
Performance (UAR %) Comparison Against Adversarial Attacks Using Different Models

Model	Adversarial Attacks	
	FGSM	BIM
DenseNet (STL) (+augmentations) [15]	44.0± 1.1	36.4±1.3
CNN-BLSTM +attention (STL) [56]	43.8± 1.5	37.2±1.5
CNN +attention (STL) [81]	42.7± 1.7	36.7±1.4
CNN-BLSTM (STL) _(baseline)	42.5±1.5	35.8± 1.6
CNN-BLSTM (STL) _(baseline) \$+ augmentations)	44.6±1.4	37.8±1.4
MTL-AUG	46.2±1.2*	39.1±1.4*
MTL-AUG (additional data)	47.5±1.0*	40.6±1.2*

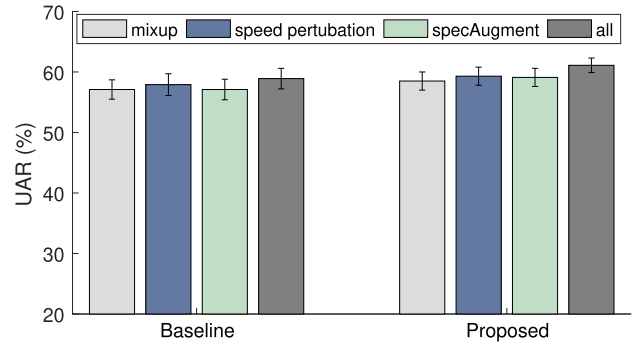
MTL-AUG (additional data) represents when additional unlabelled data is used.

TABLE 8
Results (UAR %) for Within-Corpus and Cross-Corpus Settings Using Different Configurations of the Proposed Model

Model	Configuration	Auxiliary tasks		Centre loss	Attention	Within corpus UAR(%)		Cross-corpus UAR(%)
		Augmentation-type classifier	Reconstruction			IEMOCAP	MSP-IMPROV	
1	(MTL) 	✓	✓	✓	✓	68.1 ± 1.5	62.1 ± 1.2	47.2 ± 0.41
2	(MTL) 	X	✓	✓	✓	66.7 ± 1.5	60.5 ± 1.4	46.2 ± 0.81
3	(STL) 	X	X	✓	✓	65.1 ± 1.7	59.0 ± 1.8	45.8 ± 1.0
4	(STL) 	X	X	X	✓	64.3 ± 1.9	58.2 ± 2.1	45.4 ± 1.2
5	(STL) 	X	X	X	X	62.8 ± 2.1	56.5 ± 1.9	43.6 ± 1.5



(a) IEMOCAP



(b) MSP-IMPROV

Fig. 2. Results using single augmentation in the auxiliary task of augmentation-type classification versus all. Results are statistical significant results ($p = 0.05$).

shows that the STL CNN-BLSTM cannot learn better generalisation compared to the MTL framework using the augmentation-type classifier, the reconstruction decoder, or both as auxiliary tasks. This shows that auxiliary tasks promote

generalised representations in the network by learning the shared representations. Overall, these ablation experiments show that all the proposed model components are chosen carefully to improve the SER performance effectively.

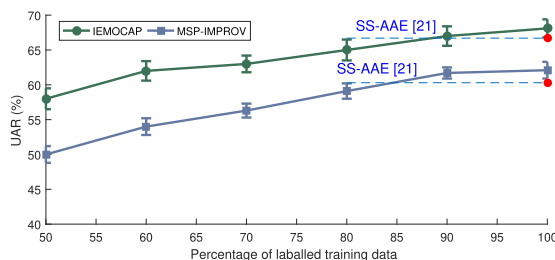


Fig. 3. Results for SER (UAR %) with changing the amount of labelled training data in IEMOCAP and MSP-IMPROV. Results are statistical significant results ($p = 0.05$).

6 CONCLUSIONS AND OUTLOOK

This contribution addressed the open challenge of improving the generalisation of speech emotion recognition (SER) with novel auxiliary tasks that do not require any additional labels for training a multi-task learning (MTL) model. We proposed augmentation-type classification and reconstruction as auxiliary tasks that minimise the required labelled data by effectively utilising the information available in the augmented data and facilitating the utilisation of unlabelled data in a semi-supervised way. The key highlights are as follows:

- The multi-task model offers improved within-corpus, cross-corpus, and cross-language emotion classification. It also shows improved generalisation against noisy speech and adversarial attacks. This is due to the proposed auxiliary tasks that helps the model learn shared representations from augmented data.
- Considerable improvements in results were found when additional unlabelled data was incorporated into the proposed MTL semi-supervised framework. This helped the model to regulate the generalised representations by learning from unlabelled data.
- We were able to reduce the amount of labelled training data by more than 10 % while achieving a similar performance reported by a recent related study [22] using 100 % training data.

Future work includes exploring multi-model auxiliary tasks to improve the primary task of speech emotion recognition by learning generalised representation. The current work is only use acted or elicited emotions by actors, however, in our future works, we aim to utilise the use of spontaneous data.

REFERENCES

- [1] 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, Jul. 2020.
- [2] R. Rana et al., "Automated screening for distress: A perspective for the future," *Eur. J. Cancer Care*, vol. 28, no. 4, 2019, Art. no. e13033.
- [3] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, and R. W. Picard, "Driver emotion recognition for intelligent vehicles: A survey," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–30, 2020.
- [4] L. Tavi, "Prosodic cues of speech under stress: Phonetic exploration of finnish emergency calls," PhD dissertation, Diss. Edu., Humanities, and Theology, Univ. Eastern Finland Philos. Fac. Sch. Humanities Pub. Univ. Eastern Finland, Finland, 2020.
- [5] E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, and N. B. Hussin, "Affective computing in education: A systematic review and future research," *Comput. Edu.*, vol. 142, 2019, Art. no. 103649.

- [6] B. Vanderplaetse and S. Dupont, "Improved soccer action spotting using both audio and video streams," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 896–897.
- [7] Z. Aldeneh and E. M. Provost, "You're not you when you're angry: Robust emotion features emerge by recognizing speakers," *IEEE Trans. Affective Comput.*, to be published, doi: [10.1109/TAFFC.2021.3086050](https://doi.org/10.1109/TAFFC.2021.3086050).
- [8] 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.
- [9] Z.-Q. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 5150–5154.
- [10] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross lingual speech emotion recognition: Urdu vs. Western languages," in *Proc. Int. Conf. Front. Inf. Technol.*, 2018, pp. 88–93.
- [11] P. Laukka, D. Neiberg, and H. A. Elfénbein, "Evidence for cultural dialects in vocal emotion expression: Acoustic classification within and across five nations," *Emotion*, vol. 14, no. 3, 2014, Art. no. 445.
- [12] 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.
- [13] Y. LeCun et al., "Handwritten digit recognition with a back-propagation network," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 1989, pp. 396–404.
- [14] 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.
- [15] 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," *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 2327–2331.
- [16] 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," *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1263–1267.
- [17] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Trans. Affect. Comput.*, to be published, doi: [10.1109/TAFFC.2021.3114365](https://doi.org/10.1109/TAFFC.2021.3114365).
- [18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [19] O. Atila and A. Şengür, "Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition," *Appl. Acoust.*, vol. 182, 2021, Art. no. 108260.
- [20] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 3920–3924.
- [21] 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.
- [22] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 992–1004, Second Quarter 2022.
- [23] Z. Zhang, B. Wu, and B. Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 6705–6709.
- [24] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Deep representation learning in speech processing: Challenges, recent advances, and future trends," 2020, *arXiv:2001.00378*.
- [25] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: A survey," *Multimedia Tools Appl.*, vol. 80, no. 6, pp. 9411–9457, 2021.
- [26] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2697–2709, 2020.
- [27] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 257–261.
- [28] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

- [29] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Trans. Affective Comput.*, vol. 8, no. 1, pp. 3–14, First Quarter 2017.
- [30] 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.
- [31] D. H. Kim, M. K. Lee, D. Y. Choi, and B. C. Song, "Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 529–535.
- [32] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2803–2807.
- [33] R. Peri, S. Parthasarathy, C. Bradshaw, and S. Sundaram, "Disentanglement for audio-visual emotion recognition using multitask setup," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 6344–6348.
- [34] 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.
- [35] 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.
- [36] M. Jaiswal and E. M. Provost, "Privacy enhanced multimodal neural representations for emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 7985–7993.
- [37] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [38] F. Tao and G. Liu, "Advanced LSTM: A study about better time dependency modeling in emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 2906–2910.
- [39] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, 2008, Art. no. 335.
- [40] 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. Affective Comput.*, vol. 8, no. 1, pp. 67–80, First Quarter 2017.
- [41] S. Li, Z.-Q. Liu, and A. B. Chan, "Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2014, pp. 482–489.
- [42] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 94–108.
- [43] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.
- [44] M. Ravanelli et al., "Multi-task self-supervised learning for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 6989–6993.
- [45] N. Chen, Y. Qian, and K. Yu, "Multi-task learning for text-dependent speaker verification," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 185–189.
- [46] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Trans. Affective Comput.*, vol. 8, no. 1, pp. 3–14, First Quarter 2017.
- [47] 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.*, 2017, pp. 1113–1117.
- [48] F. Eyben, M. Wöllmer, and B. Schuller, "A multitask approach to continuous five-dimensional affect sensing in natural speech," *ACM Trans. Interactive Intell. Syst.*, vol. 2, no. 1, 2012, Art. no. 6.
- [49] F. Ma, W. Gu, W. Zhang, S. Ni, S.-L. Huang, and L. Zhang, "Speech emotion recognition via attention-based DNN from multi-task learning," in *Proc. 16th ACM Conf. Embedded Netw. Sensor Syst.*, 2018, pp. 363–364.
- [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] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3586–3589.
- [52] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [53] D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2613–2617.
- [54] A. Baird, S. Amiriparian, M. Milling, and B. W. Schuller, "Emotion recognition in public speaking scenarios utilising an LSTM-RNN approach with attention," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 397–402.
- [55] S. Latif, S. Khalifa, R. Rana, and R. Jurda, "Federated learning for speech emotion recognition applications," in *Proc. ACM/IEEE 19th Int. Conf. Inf. Process. Sensor Netw.*, 2020, pp. 341–342.
- [56] C.-W. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2017, pp. 583–588.
- [57] A. Triantafyllou, 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.
- [58] M. Pandharipande, R. Chakraborty, A. Panda, B. Das, and S. K. Kopparapu, "Front-end feature compensation for noise robust speech emotion recognition," in *Proc. 27th Eur. Signal Process. Conf.*, 2019, pp. 1–5.
- [59] L. Juszkievicz, "Improving noise robustness of speech emotion recognition system," in *Proc. 7th Int. Symp. Intell. Distrib. Comput.*, 2014, pp. 223–232.
- [60] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [61] T. Pang, K. Xu, and J. Zhu, "Mixup inference: Better exploiting mixup to defend adversarial attacks," in *Proc. Int. Conf. Learn. Representations*, 2019.
- [62] J. Kim, K. P. Truong, G. Englebiene, 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.
- [63] S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurda, and B. W. Schuller, "Augmenting generative adversarial networks for speech emotion recognition," 2020, *arXiv:2005.08447*.
- [64] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 1, pp. 31–43, Jan. 2018.
- [65] R. Lotfian and C. Busso, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 490–494.
- [66] Y. Kim and E. M. Provost, "Emotion spotting: Discovering regions of evidence in audio-visual emotion expressions," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, 2016, pp. 92–99.
- [67] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 5206–5210.
- [68] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, Fourth Quarter 2019.
- [69] A. Burmanian, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 374–388, Fourth Quarter 2016.
- [70] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," 2017, *arXiv:1706.03256*.
- [71] 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, pp. 1517–1520.
- [72] S. Latif, R. Rana, S. Khalifa, R. Jurda, and B. W. Schuller, "Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition," *IEEE Trans. Affect. Comput.*, to be published, doi: [10.1109/TAFFC.2022.3167013](https://doi.org/10.1109/TAFFC.2022.3167013).
- [73] A. Bérard, L. Besacier, A. C. Kocabiyyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 6224–6228.

- [74] H. Dubey, A. Sangwan, and J. H. Hansen, "Transfer learning using raw waveform SincNet for robust speaker diarization," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 6296–6300.
- [75] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proc. Meetings Acoust.*, 2013, Art. no. 035081.
- [76] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 7405–7409.
- [77] J. Gideon, M. G. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG)," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 1055–1068, Fourth Quarter 2021.
- [78] 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.
- [79] 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.*, 2018, pp. 3693–3697.
- [80] N. Krishnamurthy and J. H. L. Hansen, "Babble noise: Modeling, analysis, and applications," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1394–1407, Sep. 2009.
- [81] L. Wijayasingha and J. A. Stankovic, "Robustness to noise for speech emotion classification using CNNs and attention mechanisms," *Smart Health*, vol. 19, 2021, Art. no. 100165.
- [82] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, *arXiv:1412.6572*.
- [83] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. 5th Int. Conf. Learn. Representations*, 2017. [Online]. Available: <https://openreview.net/forum?id=HJGU3Rodl>



Siddique Latif received the bachelor's degree in electronic engineering from the International Islamic University, Islamabad, Pakistan, in 2014, sponsored by the National ICT Scholarship Program, and the MSc degree in electrical engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2018. Currently, he is working toward the PhD degree with the University of Southern Queensland (USQ), Australia, and the Distributed Sensing Systems Research Group, Data61—CSIRO.

Before that, he was working as a research associate with the IHSAN Lab, Information Technology University, Punjab, Lahore, Pakistan. His research work focuses on representation learning, using unlabelled, weakly-labelled, and partially-labelled multi-modal data.



Rajib Rana received the BSc degree in computer science and engineering from Khulna University, with the Prime Minister and President's Gold Medal for outstanding achievements, and the PhD degree in computer science and engineering from the University of New South Wales, Sydney, Australia, in 2011. He received his postdoctoral training with the Autonomous System Laboratory, CSIRO, before joining the University of Southern Queensland in 2015. He is currently a senior advance Queensland research fellow and professor of computer science with the University of Southern Queensland.

He is also the director of the IoT Health Research Program with the University of Southern Queensland, which capitalises on advancements in technology and sophisticated information and data processing to understand disease progression in chronic health conditions better and develop predictive algorithms for chronic diseases, such as mental illness and cancer. His current research interests include unsupervised representation learning, adversarial machine learning, re-enforcement learning, federated learning, emotional speech generation, and domain adaptation.



Sara Khalifa received the PhD degree in computer science and engineering from UNSW, Sydney, Australia. She is currently a senior research scientist with the Distributed Sensing Systems Research Group, Data61—CSIRO. She is also an honorary adjunct lecturer with the University of Queensland and conjoint lecturer with the University of New South Wales. Her research interests rotate around the broad aspects of mobile and ubiquitous computing, mobile sensing, and the Internet of Things (IoT). Her PhD dissertation

received the 2017 John Makepeace Bennett Award which is awarded by CORE (the Computing Research and Education Association of Australasia) to the best PhD dissertation of the year within Australia and New Zealand in the field of Computer Science. Her research has been recognised by multiple awards including 2017 NSW Mobility Innovation of the year, 2017 NSW R&D Innovation of the year, National Merit R&D Innovation of the year, and the Merit R&D Award at the Asia Pacific ICT Alliance (APICTA) awards, commonly known as the 'Oscar' of the ICT industry in the Asia Pacific, among others.



Raja Jurdak (Senior Member, IEEE) received the BE degree in computer and communications engineering from the American University of Beirut, in 2000, the MS degree in computer networks and distributed computing from the Electrical and Computer Engineering Department, University of California, Irvine, in 2001, and the PhD degree in information and computer science from the University of California, Irvine, in 2005. He is a professor of distributed systems and chair in applied data sciences with the Queensland University of Technology, and director

of the Trusted Networks Lab. He previously established and led the Distributed Sensing Systems Group, CSIRO's Data61. He also spent time as visiting academic with MIT and the Oxford University in 2011 and 2017. His research interests include trust, mobility, and energy-efficiency in networks. He has published more than 220 peer-reviewed publications, including two authored books. His publications have attracted over 10,200 citations, with an h-index of 46. He serves on the editorial board of the *Ad Hoc Networks*, *Nature Scientific Reports*, and on the organising and technical program committees of top international conferences, including Percom, ICBC, IPSN, WoWMoM, and ICDCS. He was TPC co-chair of ICBC in 2021. He is a conjoint professor with the University of New South Wales, and IEEE Computer Society Distinguished Visitor.



Björn W. Schuller (Fellow, IEEE) received the diploma, in 1999, the PhD degree for his study on automatic speech and emotion recognition, in 2006, and the habilitation and adjunct teaching professorship in the subject area of signal processing and machine intelligence, in 2012, all in electrical engineering and information technology from TUM in Munich/Germany. He is professor of artificial intelligence with the Department of Computing, Imperial College London/UK, where he heads GLAM – the Group on Language, Audio, & Music, full professor

and head of the chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg/Germany, and founding CEO/CSO of audEERING. He was previously full professor with the University of Passau/Germany. He is Golden Core member of the IEEE Computer Society, senior member of the ACM, fellow of the BCS, fellow of the ISCA, president-emeritus and fellow of the Association for the Advancement of Affective Computing (AAAC), and was elected member of the IEEE Speech and Language Processing Technical Committee. He (co-)authored five books and more than 1200 publications in peer-reviewed books, journals, and conference proceedings leading to more than overall 45 000 citations (h-index = 97). He was general chair of ACII 2019, co-program chair of Interspeech 2019 and ICMI 2019, repeated area chair of ICASSP, and former editor in chief of the *IEEE Transactions on Affective Computing* next to a multitude of further associate and guest editor roles and functions in Technical and Organizational Committees.