Automatic Assessment of Depression From Speech via a Hierarchical Attention Transfer Network and Attention Autoencoders

Ziping Zhao, Zhongtian Bao, Zixing Zhang, *Member, IEEE*, Jun Deng, Nicholas Cummins, *Member, IEEE*, Haishuai Wang, *Member, IEEE*, Jianhua Tao, *Member, IEEE*, and Björn Schuller, *Fellow, IEEE*

***Abstract*—Early interventions in mental health conditions such as Major Depressive Disorder (MDD) are critical to improved health outcomes, as they can help reduce the burden of the disease. As the efficient diagnosis of depression severity is therefore highly desirable, the use of behavioural cues such as speech characteristics in diagnosis is attracting increasing interest in the field of quantitative mental health research. However, despite the widespread use of machine learning methods in the depression analysis community, the lack of adequate labelled data has become a bottleneck preventing the broader application of techniques such as deep learning. Accordingly, we herein describe a deep learning approach that combines unsupervised learning, knowledge transfer and hierarchical attention for the task of speech-based depression severity measurement. Our novel approach, a Hierarchical Attention Transfer Network (HATN), uses hierarchical attention autoencoders to learn attention from a source task, followed by speech recognition, and then transfers this knowledge into a depression analysis system. Experiments based on the depression sub-challenge dataset of the Audio/Visual Emotion Challenge (AVEC) 2017 demonstrate the effectiveness of our proposed model. On the test set, our technique outperformed other speech-based systems presented in the literature, achieving a Root Mean Square Error (RMSE) of 5.51 and a Mean Absolute Error (MAE) of 4.20 on a Patient Health Questionnaire (PHQ)-8 scale [0, 24]. To the best of our knowledge, these scores represent the best-known speech results on the AVEC 2017 depression corpus to date.**

***Index Terms*—Depression, attention transfer, hierarchical attention, monotonic attention.**

I. Introduction

AS PART of an effort to assist clinicians in diagnosing depression more efficiently, automatic detection and monitoring of depression from speech signals attracted considerable research attention in recent years [1]. The clinical potential of depression analysis has motivated the creation of the *Depression Recognition Sub-Challenge* (DSC) of the *Audio/Visual Emotion Challenge and Workshop* (AVEC 2013 [2], AVEC 2014 [3], AVEC 2016 [4], AVEC 2017 [5]) and the *Bipolar Disorder Sub-challenge* (BDS) of AVEC 2018 [6]. These challenges provide a common platform within which to explore the efficacy of the application domains of depression recognition.

Various machine learning approaches have been proposed as a part of these challenges. Most recently, deep neural networks, and *Convolutional Neural Networks* (CNNs) in particular, have been shown to produce state-of-the-art performances in speech-based depression analysis [7]–[10]. For example, in [8], Yang *et al.* proposed a multi-modal fusion framework composed of CNNs and DNN. Based on a combination of audio, text and visual information, this system achieved an RMSE of 5.97 and an MAE of 5.16 on the test set of the AVEC 2017 depression corpus. In another work, also by Yang *et al.* [9], the final *Patient Health Questionnaire* (PHQ)-8 prediction score (an RMSE of 5.40 and an MAE of 4.36 on the AVEC 2017 test set) was achieved by fusing the outputs of the four systems, across all modalities, via multivariate linear regression. However, despite the promising results acheived to date with CNNs, other contemporary approaches, most notably *Recurrent Neural Networks* (RNNs), remain understudied in this context. In principle, RNNs should be particularly effective, as they are capable of modelling the sequential structure of speech; these networks have demonstrated

This work was supported in part by the National Natural Science Foundation of China under Grant 61702370, in part by the National Science Fund for Distinguished Young Scholars under Grant 61425017, in part by the Key Program of the National Natural Science Foundation of China under Grant 61831022, in part by the Key Program of the Natural Science Foundation of Tianjin under Grant 18JCZDJC36300, in part by the Open Projects Program of the National Laboratory of Pattern Recognition and the Senior Visiting Scholar Program of Tianjin Normal University, and in part by the Innovative Medicines Initiative 2 Joint Undertaking under Grant 115902, which receives support from the European Union's Horizon 2020 research and innovation program and EFPIA.

Z. Zhao and Z. Bao are with the College of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387, China (e-mail: ztianjin@126.com; 1942952862@qq.com).

Z. Zhang is with the Department of Computing, Imperial College London, SW7 2AZ London, U.K. (e-mail: zixing.zhang@tum.de).

J. Deng is with Agile Robots AG, 82205 Gilching, Germany (e-mail: jun.deng@tum.de).

N. Cummins and B. Schuller are with the Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany (e-mail: nicholas.cummins@ieee.org; schuller@ieee.org).

H. Wang is with the Department of Computer Science, Fairfield University, Fairfield, CT 06824 USA (e-mail: haishuai_wang@hms.harvard.edu).

J. Tao is with the National Laboratory of Pattern Recognition of Institute of Automation of Chinese Academy of Sciences, The Center for Excellence in Brain Science and Intelligence Technology (CEBSIT), School of Artificial Intelligence of University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: jhtao@nlpr.ia.ac.cn).

state-of-the-art results in many related speech-based tasks [11], [12].

One potential reason why RNN-based approaches remain underexplored in depression analysis could be the structure of the associated databases. In depression analysis tasks, a typical operating scenario involves longer audiovisual files – it is normal for these to exceed 15 minutes in length – with only one target label. It is well-established that conventional RNN-based approaches struggle under such learning conditions. The use of attention mechanisms, specifically hierarchical attention mechanisms [13], [14], can aid RNN-based approaches in such circumstances. Hierarchical attention approaches have achieved state-of-the-art performance in various *Natural Language Processing* (NLP) document-based classification tasks [13]–[15]. However, the inclusion of attention mechanisms increases the number of learnable parameters in the associated models; this increase does not fit with smaller (in terms of the number of unique samples) depression corpora [1]. The attention mechanism can, however, be used in combination with transfer learning paradigms [16]–[19].

While the sparseness of training data constitutes a well-known bottleneck for speech-based depression analysis [20], a range of efficient machine learning-based labelling techniques exist that can be used to leverage both labelled and unlabeled data in order to improve model performance [21]. One such technique is *Semi-Supervised Learning* (SSL), which has also been demonstrated to have good regularisation and optimisation properties [22]. A standard method for realising SSL is to perform the training in two phases: unsupervised pre-training, followed by supervised fine-tuning [21].

To enable deep learning via such an approach, the unsupervised pretraining often involves training a variant of an autoencoder, such as a denoising autoencoder. By reconstructing inputs with respect to a given loss function, autoencoders act as a feature learning method. Accordingly, their use makes it easier to train a deep classifier, as the target data distribution is explicitly learnt in the unsupervised learning model [23]. Autoencoders therefore play a fundamental role in building deep architectures for transfer learning and other tasks [24]. Previous works have shown that autoencoder models are capable of learning meaningful, abstract representations and can thus achieve better classification results, as in [25], [26]. Moreover, recent works have demonstrated the benefits of including attention – both flat and hierarchical – in autoencoder networks [27], [28].

Motivated by the above observations, we herein propose a novel depression estimation framework that combines a carefully designed hierarchical attention transfer mechanism and hierarchical attention autoencoders into a unified framework, which is in turn used to aid the training of an attention-enhanced RNN framework for depression severity analysis. Encouraged by the recent success of attention transfer mechanisms and unsupervised learning, we propose a *Hierarchical Attention Transfer Network* (HATN) for speech-based depression severity assessment. We further explore the contribution of both the hierarchical attention mechanism and the teacher-student framework to attention transfer for this task; more specifically, we use hierarchical attention autoencoders to transfer knowledge from the speech recognition task to our depression detection task.

There are several advantages to our proposed model. Firstly, it provides a hierarchical attention transfer mechanism; this framework automatically transfers attention from speech recognition at the frame level while simultaneously providing improved interpretability as to what knowledge should be transferred. Our goal is to improve the learning of a student network, given a teacher network trained on a similar task. The proposed hierarchical structure of our model is based on the observation that the clinical interviews in the AVEC 2017 depression dataset [5] (i.e., our training and evaluation corpora) have a clear hierarchical structure: namely, each clinical interview is composed of multiple sentences, while each of these sentences is composed of multiple feature frames. This hierarchical structure allows the network to explicitly model the contribution of each frame in a particular sentence towards the target clinical depression score, as well as modelling the task-specific context at semantically higher levels (such as at the sentence or interview level) [13].

Secondly, this work also specifically targets improvements in depression analysis utilising a semi-supervised labelling paradigm. It uses autoencoders to discover the intrinsic knowledge in unlabelled training samples, as well as a small number of labelled training samples to allow the autoencoders to learn the required latent feature representation. Moreover, a speech-based depression analysis system integrated with the proposed structure not only reduces the need for a large number of labelled training examples, but also endows the system with the ability to distil essential knowledge from the unlabelled data into the supervised learning. To the best of our knowledge, this is the first time that such a study has been conducted for depression severity measurement.

II. Related Work

To date, speech-based automated depression analysis has primarily been performed using conventional – rather than deep learning – machine learning methods, in combination with hand-crafted feature engineering [1], [20]. Many novel approaches have been proposed that have demonstrated increasingly good performance over the past few years [1], [29]–[32]. Models commonly used in such studies include *Gaussian Mixture Models* (GMM), *Support Vector Machines* (SVM) and *Relevance Vector Machines* (RVM) [1]. Studies comparing the suitability of different models have generally concluded that no one model is superior [33], [34]. One particularly interesting approach is the Gaussian staircase approach [35], [36], in which each GMM comprises an ensemble of Gaussian classifiers designed to model the ordinal nature of clinical depression scales [37].

With respect to hand-crafted audio features, researchers have found that depressed subjects are more likely to exhibit a low dynamic range of the fundamental frequency, a slow speaking rate, a slightly shorter speaking duration, and a relatively monotone delivery [1]. Several works have leveraged knowledge-based features that are specifically designed to capture these characteristics or related effects [36], [38]–[40]. Large supra-segmental feature spaces, in combination

with an SVM regressor, were used to set the baseline scores in the AVEC 2013 and 2014 challenges [2], [3]. The audio feature set consisted of 2268 features: namely, 32 energy- and spectral-related *low-level descriptors* (LLD) $\times$ 42 functionals, six voicing-related LLD $\times$ 32 functionals, 32 delta coefficients of energy/spectral LLD $\times$ 19 functionals, six delta coefficients of voicing-related LLD $\times$ 19 functionals, and 10 voiced/unvoiced durational features. The feature set was extracted using the OPENSMILE toolkit [41]. Similarly, for the audio baseline features of both AVEC 2016 [4] and AVEC 2017 [5], prosodic, voice quality, and spectral features were extracted by the COVAREP toolkit [42].

As already discussed in the introduction, a small number of studies have begun to explore the application of deep neural networks, CNNs in particular, for the task of depression analysis [7]–[10], [43], [44]. Furthermore, while the benefits of RNNs have yet to be fully established for depression analysis from speech, the advantage of RNNs can be seen in the related field of *Speech Emotion Recognition* (SER) [11], [12], [45]–[47].

Recently, attention mechanisms have become widely adopted among the deep learning community. In the context of deep learning, attention mechanisms are a family of algorithms that enable a network to dynamically select subsets of input attributes given a particular context (input-output pair) setting. The overall goal of applying attention is, of course, to improve decision accuracy. Attention has been successfully applied in tasks including speech recognition [48], NLP [49], [50], and speech emotion recognition [12], [47], [51]. Additionally, hierarchical attention networks have been shown to be superior to non-hierarchical networks in a range of tasks, such as NLP [13], [15]; this is due to their ability to leverage more than one level of attention in a network with the aim of capturing hierarchical structures contained in the data being modelled [13]. To the best of our knowledge, attention mechanisms have never before been employed for the detection of depression from speech.

When compared to conventional machine learning methods, however, deep learning is very strongly dependent on massive amounts of training data [52]. Moreover, an absence of sufficient training data has become an inescapable problem for depression analysis [1], [20]. Accordingly, autoencoders, which can be used as an unsupervised learning model [53], represent a type of deep learning paradigm that could be well-suited for use under such circumstances. Previous research efforts have used autoencoders in related speech tasks, such as emotion recognition systems [23], [54]–[57]. Again, to the best of our knowledge, no previous work has employed autoencoders to analyse depression severity from speech.

Another technique used to improve network performance when only a limited amount of labelled training data is available - namely, knowledge transfer - has been widely employed in various settings [58], [59]. Very recently, attention maps have been studied as a knowledge transfer mechanism [16]. It has been demonstrated that training smaller 'student' networks to mimic the attention maps of larger, higher-performing 'teacher' network architectures can lead to considerable performance gains in these smaller networks. In [17], the authors explored the use of attention for cross-domain knowledge transfer from online images to videos. Similarly, Zhuo *et al.* [18] developed an attention transfer process for convolutional domain adaptation. However, the work in both of these papers was based on CNNs; to the best of our knowledge, to date, no attention transfer process for RNNs has yet been designed.

From the literature, we can see that recent works present strong evidence for the value added by attention transfer and autoencoders. Accordingly, our approach utilises a combination of these two existing ideas for speech-based depression severity measurement. To the best of our knowledge, no existing work has yet combined these two methods for such a task.

III. Proposed Methodology

In this section, we first present an overview of the proposed model for 'cross-task' depression severity measurement. We then introduce the technical details of the model.

A. An Overview of the Proposed Model

The present work incorporates two key tasks: (i) *speech recognition*, and (ii) *depression recognition from speech* (Fig. 1). Our goal is to improve the performance on the target task (i.e., depression recognition) by leveraging the spatial attention maps from the classifier in the source task (i.e., speech recognition). Given the scarcity of our target data, we learn this mapping on resource-rich tasks where high-quality attention can be obtained during training.

In this regard, our proposed model comprises four key components. The first, the teacher network, is an attention-based encoder-decoder network, trained for the speech recognition task, which learns the initial attention maps. The second component, the fundamental component of the model, is the *attention transfer mechanism*. This component is used to train a (shallower) student network for the task of depression recognition, such that it mimics the attention maps of the teacher network [16]. In the third component, hierarchical attention autoencoders are used in an unsupervised manner to generate a rich set of feature representations upon which the related supervised tasks can be built. In the final component of our hybrid model, the depression recognition module, we use a hierarchical attention neural network, which consists of frame-level and sentence-level attention mechanisms.

In the proposed model, the standard bidirectional long short-term memory recurrent neural networks (BLSTM) are used to sequentially process each frame in the input. Further we explore the benefits of two different attention mechanisms, *Standard Soft* and *Monotonic*, which are introduced in the next two sub-sections.

B. Standard Soft Attention

Standard soft attention mechanisms are used to select relevant encoded hidden vectors via attention weights (an informative sequence of weights) during the decoding phase [50]. At each timestep i, the attention weights $\alpha_{i,j}$ are produced by normalising the scalar values $e_{i,j}$ across the memory using a softmax function:

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^{T} \exp(e_{i,k})}. \quad (1)$$

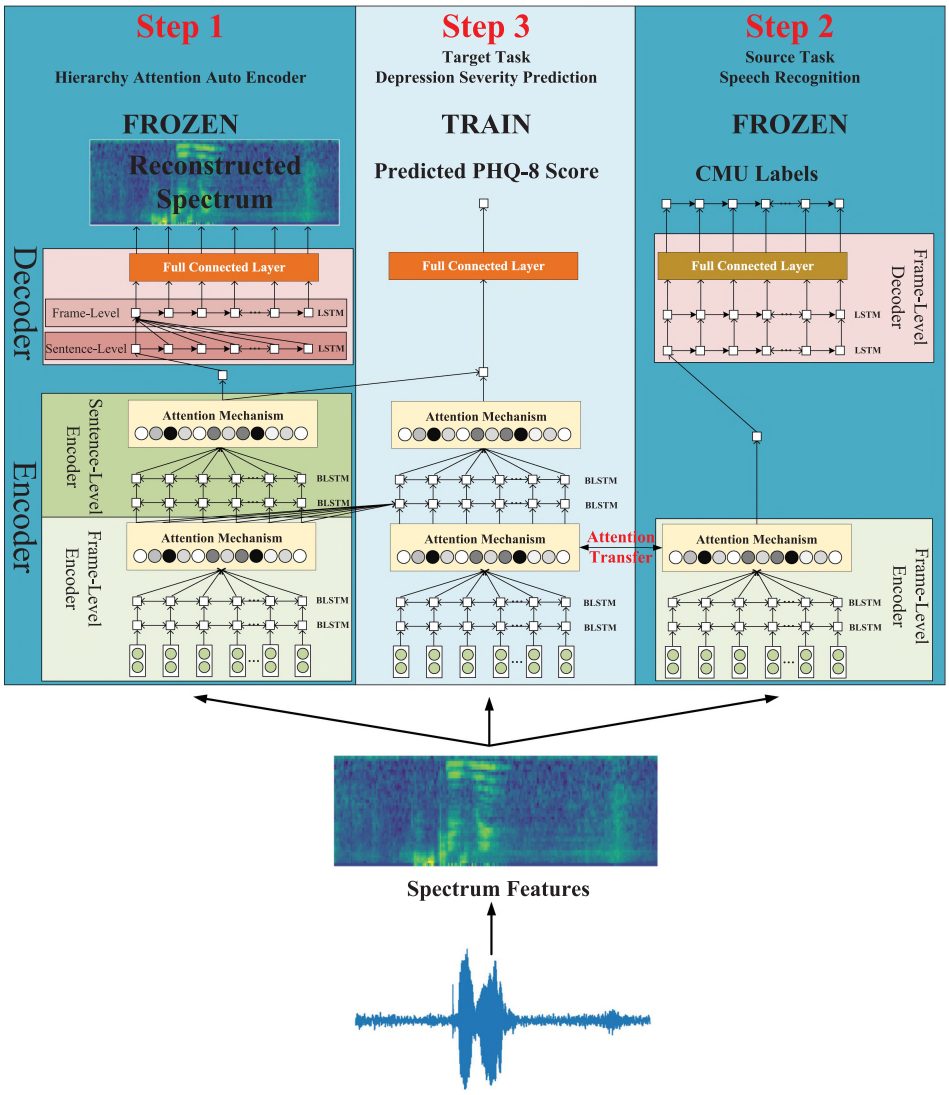


Fig. 1. An overview of our proposed model. First, we trained a hierarchical attention autoencoder network that works in a bottom-up manner: an encoder applies the attention mechanism at the frame level, which is then conveyed to a sentence-level encoder to form the hidden vectors of the context. These vectors are further processed by the sentence-level attention, which produces a latent representation of the clinical interview. The learnt representation of the clinical interview is then fed into a decoder to enable reconstruction of the input sequence, after which the representations of both the sentence and the clinical interview are utilised as input and fed into the BLSTM to predict the overall clinical depression score. Once the weights of the hidden layer of the autoencoders are trained, its parameters are frozen; the model then learns its attentions through a speech recognition task, and these are then transformed into the hierarchical depression detection system.

$e_{i,j}$ is an alignment scoring mechanism used to determine how well the inputs around position j and the output at position i match. It is computed via:

$$e_{i,j} = a(s_{i-1}, h_j), \tag{2}$$

in which s_i denotes the decoder's state, h_j indicates the j-th entry of the hidden state sequence $\mathbf{h} = \{h_1, \ldots, h_T\}$, and $a(\cdot)$ is a learnable deterministic 'energy function'. Typically, a single-layer neural network using a tanh nonlinearity is utilised as $a(\cdot)$; however, other functions (such as a simple dot product between s_{i-1} and h_j) have been used as well [50]. Note that, we use tanh as the non-linear activation function in the presented work.

The output of the attention layer, denoted as c_i, is the weighted average of the encoder hidden state sequence h, which is defined as follows:

$$c_i = \sum_{j=1}^{T} \alpha_{i,j} h_j. \tag{3}$$

Finally, the decoder state is updated to s_i based on s_{i-1}, c_i, and the decoder outputs y_i:

$$s_i = f(s_{i-1}, y_{i-1}, c_i), \tag{4}$$

$$y_i = g\left(s_i, c_i\right), \tag{5}$$

where $f(\cdot)$ is BLSTM in our work, while $g(\cdot)$ is a learnable nonlinear function that maps the decoder state to the output space.

In this work, unless otherwise stated, soft attention refers to the global attention approach. In global attention all the hidden states of the encoder are considered in order to derive the context vector c_i. We also test the advantage of applying local attention mechanisms [60]. This approach selectively focuses on a small window of context and is differentiable. Local attention, therefore, has the advantage of avoiding the expensive computation incurred in the global soft attention.

C. Monotonic Attention

Monotonic attention mechanisms address two of the main limitations associated with soft attention: namely, quadratic-time complexity and the lack of options for online decoding [61], [62]. By adaptively splitting into smaller chunks on which attention can be be computed, monotonic attention mechanisms enable linear and online decoding [61], [62]. Moreover, works in the *Automatic Speech Recognition* (ASR) literature demonstrate that encoder-decoder-based ASR with monotonic attention can achieve additional and considerable performance improvements, as well as reducing the associated computational complexity, relative to a comparable system utilising a standard global attention architecture [63].

The monotonic attention process can be described as follows: at timestep i, the attention mechanism begins to inspect memory entries (the hidden state sequences), starting from the memory index where it left off at the previous output timestep, herein referred to as t_{i-1}, where t_i is the index of the memory entry chosen at output timestep i (for convenience, we let $t_0 = 1$). It then computes an unnormalised energy scalar $e_{i,j}$ for $j \in \{t_{i-1}, t_{i-1}+1, t_{i-1}+2 \ldots\}$ via Equation (2).

Monotonic attention can, therefore, be interpreted as the probability of choosing memory element j at output timestep i. The selection probabilities $p_{i,j}$ are produced by passing these energy values through a logistic sigmoid function:

$$p_{i,j} = \sigma(e_{i,j}) \in (0, 1) \tag{6}$$

A boolean attend/don't attend decision $z_{i,j} \in \{0, 1\}$ is then sampled from a Bernoulli random variable parameterised by $p_{i,j}$:

$$z_{i,j} \sim \text{Bernoulli}(p_{i,j}), \tag{7}$$

namely,

$$P(z_{i,j} = 1) = p_{i,j},\ P(z_{i,j} = 0) = 1 - p_{i,j}. \tag{8}$$

As soon as $z_{i,j} = 1$, the process stops, and the attention context vector c_i is set as h_{t_i}. Each $z_{i,j}$ can be seen as representing a discrete choice as to whether to attend a new item from the memory ($z_{i,j} = 0$) or produce an output ($z_{i,j} = 1$). This process is repeated for the subsequent output time steps, always beginning at t_{i-1} – i.e., the memory index identified in the previous step. Note that if $z_{i,j} = 0$ for all $j \in \{t_{i-1}, t_{i-1}+1, \ldots, T\}$, then c_i is set to be a vector of zeros.

The energy function used for hard monotonic alignments is as follows:

$$a(s_{i-1}, h_j) = \frac{g v^T \tanh(W_s s_{i-1} + W_h h_j + b)}{\|v\|} + r, \tag{9}$$

where g, r are the learnable scalars, while $W_h \in \mathbb{R}^{d \times dim(h_j)}$, $W_s \in \mathbb{R}^{d \times dim(s_{i-1})}$, $b \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$ are the learnable parameters and d denotes the hidden dimensionality of the energy function.

As the monotonic attention process involves sampling and hard assignment, models utilising this technique cannot be trained via backpropagation. As in [62], we compute the context vector c_i as the probability distribution over the memory induced by the attention process:

$$\alpha_{i,j} = p_{i,j} \left((1 - p_{i,j-1}) \frac{\alpha_{i,j-1}}{p_{i,j-1}} + \alpha_{i-1,j} \right). \tag{10}$$

The output of the attention layer c_i can also be computed as in Eq. (3). Further details regarding monotonic attention are provided in [61].

D. Hierarchical Attention Autoencoders

Our hierarchical attention autoencoders rely on the attention mechanism to aid the reconstruct the input sequences. We test three separate hierarchical BLSTM model incorporating either (i) standard global soft attention; (ii) local soft attention; or (iii), monotonic attention. The hierarchical attention autoencoder itself works in a bottom-up manner applying two levels of attention. First, an encoder and attention are applied at the frame level. The resulting output is then passed to a sentence-level encoder to form context-based hidden vectors, which are then processed by the sentence-level attention mechanism, producing a latent hierarchical representation of the clinical interview. Subsequently, this representation is fed into a decoder in order to reconstruct the input sequence. In the remainder of this section, we describe the key details of our proposed hierarchical attention autoencoders.

1) Attention Encoder: Noting that a clinical interview consists of a set of sentences, each of which consists of a set of speech frames, the encoder has a hierarchical structure which matches this structure. Specifically, attention mechanisms are incorporated at two different levels of the encoder. First, we use a BLSTM with attention to aid the selection of informative frames at frame-level. Next, we apply another BLSTM with attention over the frame-level representations to learn the associations between representations and aid the selection of informative representations at sentence-level.

Given the input spectrogram x, we produce hidden representations using the following equations:

$$e_s{}' = Attention_{frame}(Encoder_{frame}(x)W_s + b_s), \tag{11}$$

$$e_s = LayerNorm(e_s{}'), \tag{12}$$

$$e_c = Attention_{sentence}(Encoder_{sentence}(e_s)W_c + b_c), \tag{13}$$

where W_s, b_s, W_c and b_c are learnable parameters, while $LayerNorm$ denotes the layer normalisation and *ReLU* is used as the activation function. In Equation (11), we first obtain the representation vectors $e_s{}'$ at the sentence level using one layer of BLSTMs with attention. Another layer of BLSTM with attention is then placed on top of all sentences to enable learning

of the representations of the entire clinical interview e_c using equation (13).

2) Decoder: Given the encoder representations e_c, the decoder is responsible for regenerating the input sequence. We first use a hierarchical LSTM decoder to generate the sentence-level representations d_s:

$$d_s = Decoder_{sentence}(e_c). \tag{14}$$

Subsequently, based on the representation of each sentence, a frame-level LSTM decoder is utilised to generate the input sequence:

$$\bar{x} = Decoder_{frame}(d_s). \tag{15}$$

E. Hierarchical Attention Transfer Network

Many recent works on attention transfer have generally focused on computer-vision-related tasks, with the developed spatial attention maps being designed for CNNs [16], [18]. Within these approaches, the activation maps, for both the source and target domains in a specific convolutional layer, are first calculated via an Lp-norm pooling on all convolutional response channels. Domain discrepancy minimisation is then performed on the second-order correlation statistics of the attention maps [18]. Inspired by this approach, we develop an attention transfer process designed for BLSTM.

1) Activation-Based Attention Model: In this section, we explain the method used to define the spatial attention map, along with the way in which we transfer attention information from a teacher to a student network.

We first consider a BLSTM layer and its corresponding activation tensor $A \in \mathbb{R}^{C\times H\times W}$, which consists of C ($C = 1$ for BLSTM) channels with spatial dimensions $H \times W$, as well as a mapping function F that takes the above BLSTM layer activations A (3D tensor) as input and outputs. Using F, a spatial attention map can be defined as follows:

$$F : \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^{H\times W}. \tag{16}$$

Because the absolute value of a hidden neuron activation indicates that neuron's importance w. r. t. the specific input, we observe that can construct a spatial attention map by computing the statistics of these absolute values across the channel dimension. More specifically, we consider the following spatial attention mappings:

$$(F(A))_{i,j} = \sum_{k=1}^{C} |A_{k,i,j}|^p, \tag{17}$$

where $i \in \{1, 2, \ldots, H\}$ and $j \in \{1, 2, \ldots, W\}$ are spatial indexes.

In attention transfer, given the spatial attention maps of a teacher network, the goal is to train a student network that will not only make correct predictions, but will also have attention maps that are similar to those of the teacher network.

Without loss of generality, we therefore assume that transfer losses are placed between the student and teacher attention maps of the same spatial resolution; if required, however, attention maps can be interpolated to match their shapes. We can then define the following total loss:

$$L_T = L_D + W_{AT} \times L_{AT}, \tag{18}$$

where L_D denotes the loss of the depression recognition task, W_{AT} denotes the weight of the attention transfer, and L_{AT} denotes the loss of the attention transfer, which can be computed as follows:

$$L_{AT} = \sum_{j\in\mathcal{I}} \left\| Q_D^j - Q_S^j \right\|_1, \tag{19}$$

here, $\mathcal{I}$ denotes the indices of the attention map, while Q_D^j and Q_S^j represent the j-th pair of the attention map of the depression recognition and speech recognition tasks respectively. As can be seen, during attention transfer we make use of $l1$-normalized attention maps.

2) Hierarchical Attention Model: Our attention model has a hierarchical structure in which two levels of attention mechanisms are applied (at the frame and sentence level respectively), enabling our model to attend differentially to more and less important content when constructing the depression analysis. Mathematically, we represent a clinical interview with m sentences $\{S_1, S_2, \ldots, S_m\}$, the i-th sentence is S_i, which in turn consists of l_i frames as $S_i = f_1^i f_2^i \ldots f_{l_i}^i$; f_t^i is the t-th frame in S_i, $t \in [0, l_i]$. In the following, we present how to build the depression representation progressively from frame vectors hierarchically.

a) Frame level: Each frame, as previously discussed, has a different influence on the representation of the whole sentence. It is therefore necessary to qualify the contributions of each frame and learn its unique representation. Fortunately, attention mechanisms can be used to highlight the relative importance of different parts of the input sequence by assigning weights to the encoding vectors. Therefore, we introduce frame-level attention to weight frames of each sentence and output a weighted sum of all the frames' information.

First, we obtain the hidden state for the t-th frame in sentence S_i by concatenating the forward LSTM and the backward LSTM outputs reads from f_1^i to $f_{l_i}^i$:

$$h_t^i = \overrightarrow{h_t^i} \,||\, \overleftarrow{h_t^i}, \tag{20}$$

where $||$ denotes concatenation; moreover, $\overrightarrow{h_t^i}$ and $\overleftarrow{h_t^i}$ are the hidden states for the forward and backward LSTM, respectively, of the t-th frame of the i-th sentence.

We then use an attention layer in order to identify the most informative frames in each sentence and enforce their contribution to the final sentence representation. The sentence vector v_i, which is the vector representation of the i-th sentence, is computed as the weighted sum of all frame hidden states h_t^i:

$$\alpha_t^i = Attention_{frame}(h_t^i), \tag{21}$$

$$v_i = \sum_t \alpha_t^i h_t^i. \tag{22}$$

b) Sentence level: Under the assumption that contextual sentences will not contribute equally to the semantic meaning of a clinical interview, we introduce sentence-level attention to

weight each sentence of each clinical interview and output a weighted sum of all sentences' embedding vectors to make up the representation of the clinical interview of each patient.

The final representation of each clinical interview is obtained analogously: given the sentence representation v_1 to v_m, the concatenation of the bidirectional LSTM is obtained as follows:

$$h_i = \overrightarrow{h_i} \,||\, \overleftarrow{h_i}, \tag{23}$$

with the general form of sentence-level attention weights

$$\alpha_i = Attention_{sentence}(h_i). \tag{24}$$

We aggregate them by computing the weighted sum of all the sentences' representations, thereby obtaining the final representation e of the clinical interview, which is formulated as follows:

$$e = \sum_i \alpha_i h_i. \tag{25}$$

IV. Experiments and Results

A. Datasets

The AVEC2017 series depression detection task utilised the *Distress Analysis Interview Corpus – Wizard of Oz* (DAIC-WOZ) [64] database, containing 189 segments of clinical interviews designed to support the diagnosis of conditions such as depression. The recorded clinical interviews were split into a training set (comprising 107 segments), a development set (35 segments), and a test set (47 segments). For each segment, the database includes audio and video features along with the transcript of an interview, ranging between 7–33 minutes in length, conducted by an animated virtual interviewer called Ellie, which was controlled by a human interviewer in another room. For each interview in the training and development sets, the exact *Patient Health Questionnaire* (PHQ)-8 depression index score [65] is given for each participant in the DAIC-WOZ corpus. The PHQ-8 is an ordinal scale, in the range [0, 24], which reflects depression severity. The score is obtained from answers given to eight questions that reflect the core criteria used to diagnose clinical depression, as laid out in the *Diagnostic and Statistical Manual of Mental Disorders - Fourth Edition* (DSM-IV) [66]. Answers to individual PHQ-8 items are given within the ordinal range [0, 4]. The final score is then obtained by summing the eight individual scores. For the training and development partitions, the average depression severity was $M = 6.67$ ($SD = 5.75$) out of a maximum score of 24.

During our analysis, the audio file was processed so as to isolate the participant's voice only. We split the individual DAIC-WOZ recordings into individual participant turns based on the transcriptions provided. A total of. 32 401 turns extracted from aligned textual transcriptions are provided with the dataset. The division of the participants is presented in Table I.

TABLE I
Gender and Depression Status Information of the Participants in the Training, Development and Test Sets of the DAIC-WOZ Database. Note That the Total Length (hh:mm) of Each Partition is Also Given

Gender	class	Train	Dev.	Test
Male	N-Dep.*	55	12	18
	Dep.*	8	4	5
Female	N-Dep.*	32	16	20
	Dep.*	13	3	4
Total Number of participants		107	35	47
Total Number of participant turns		16 906	6 679	8 816
Total hh:mm		12:52	4:52	6:40

*All participants were assigned into one of two classes, *depressed* (Dep.) or *non-depressed* (N-Dep.), based on the PHQ-8 scores.

B. Features

In this paper, we used the spectrogram extraction process described in [67]. The spectrogram was constructed using the output of a 40-dimensional mel-scale log filter bank. These features were computed over frames 25 ms in length with a 10 ms stride and normalised to be in the range [0, 1].

C. Experimental Setup and Evaluation Metrics

1) Model Parameters: In our work, all models were implemented using the *TensorFlow*[1] framework. The favourable training epoch was set to 100 due to the restrictions imposed by computational costs and time expenses. We first trained a hierarchical attention autoencoder network. For the encoder, a two-layer BLSTM containing 128 single-memory-cell LSTM memory blocks in the forward and backward hidden layers was utilised both at the frame level and sentence level. For the decoder, the representation of the clinical interview was decoded by the two-layer LSTM, which contained 256 single-memory-cell LSTM memory blocks. The RMSProp optimiser was used to train our autoencoders, employing a fixed learning rate of 10^{-4}. Once the weights of the hidden layer of the autoencoders were trained, the parameters of the autoencoders were frozen.

In order to implement attention transfer, we pre-trained on a source task (speech recognition) and acquired the attention maps gained after solving the speech recognition task in advance. For the training of the speech recognition model, we explored the use of a monotonic attention-based encoder-decoder model and utilised the AVEC 2017 depression dataset as the database. In our work, we reduced the number of states of the origin transcription by leveraging the CMU pronouncing dictionary [68] in the speech recognition task. For the pre-training of the speech recognition task, moreover, the forward and backward hidden layers of the BLSTM network had 128 blocks each. The learning rate was also set to 10^{-4}, and the parameters were frozen during training.

Once the above two training steps were complete, we began to train the depression recognition model. For training, we also used a two-layer BLSTM consisting of 128 single-memory-cell LSTM memory blocks in the forward and backward hidden layers; the learning rate was again set to 10^{-4}. Finally, the outputs of the fully connected layers could be regarded as the final predicted PHQ-8 score in the range [0, 24].

2) Evaluation Metrics: As depression severity prediction is a regression task, the accuracy metric for the challenge was the

[1][Online]. Available: https://www.tensorflow.org

TABLE II
A COMPARISON OF ROOT MEAN SQUARE ERROR (RMSE) AND MEAN AVERAGE ERROR (MAE) SCORES FROM BOTH KEY RESULTS IN THE LITERATURE AND FROM EXPERIMENTS ON THE PROPOSED HIERARCHICAL MODEL. THE SCORES ARE ACHIEVED ON BOTH THE DEV(ELOPMENT) AND TEST SETS OF THE DAIC-WOZ CORPUS

Methods	Dev.		Test	
	RMSE	MAE	RMSE	MAE
Previously reported methods				
AVEC 2017 Audio Baseline [5]	6.74	5.36	7.78	5.72
AVEC 2017 Audio-Video Baseline [5]	6.62	5.52	7.05	5.66
DCNN-DNN [8]*	4.65	3.98	5.97	5.16
Multivariate regression model [9]*	3.09	2.48	5.40	4.36
Proposed hierarchical attention models (AT denotes attention transfer)				
Hie. global soft attention	5.26	3.67	6.54	5.03
Hie. local soft attention (classic)	5.20	3.59	6.43	4.99
Hie. local soft attention (monotonic)	4.87	3.02	6.14	4.76
Hie. global soft attention w/AT	4.14	3.65	5.96	4.78
Hie. local soft attention (classic) w/AT	4.07	3.56	5.81	4.76
Hie. local soft attention (monotonic) w/AT	**3.85**	**2.99**	**5.66**	**4.28**
Proposed hierarchical autoencoders (HAE denotes hierarchical autoencoders)				
Vanilla HAE	5.16	3.97	6.20	5.01
HAE (global soft attention)	4.74	3.69	5.93	4.78
HAE (local soft attention-classic)	5.14	3.67	5.92	4.55
HAE (local soft attention-monotonic)	**4.67**	**3.36**	**5.72**	**4.47**
Combination of HATN with hierarchical autoencoders				
HATN & HAE (global soft attention)	4.55	3.58	5.77	4.53
HATN & HAE (local soft attention-classic)	4.16	3.02	5.72	4.66
HATN & HAE (local soft attention-monotonic)	**3.68**	**2.87**	**5.51**	**4.20**

*Note that these two studies are multi-modal works in which audio, video and text streams were utilised.

Root Mean Square Error (RMSE) and *Mean Absolute Error* (MAE) given by:

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2} \tag{26}$$

$$MAE = \frac{1}{N}\sum_{n=1}^{N}|\hat{y}_n - y_n|, \tag{27}$$

where $\hat{y}_n$ denotes the predicted value, y_n, the actual score, and N the total number of test instances.

D. Results and Discussion

1) Experiments on the Hierarchical Attention Transfer Network: Here, we conducted experiments aimed at verifying the efficiency of our proposed hierarchical attention transfer network. In order to provide supervision for attention generation, an attention-based encoder-decoder model was implemented for the speech recognition task; subsequently, a WER of 12.9% was obtained on the test set of the DAIC-WOZ corpus, while a WER of 8.9% was obtained on the development set.

We observed that the best MAE (4.28) and RMSE (5.66) on the test set, as well as the best MAE (2.99) and RMSE (3.85) on the development set of the DAIC-WOZ corpus were achieved by our hierarchical monotonic attention model in combination with the attention transfer mechanism (cf. Table II). In Table II, the state-of-the-art models utilised for comparison purposes include AVEC 2017 baseline methods, two previous multi-modal methods that previously achieved good performance on the AVEC 2017 dataset and in which audio, video, and text streams were utilised [8], [9]; several hierarchical attention models that do not employ an attention transfer mechanism are also compared with our proposed hierarchical attention transfer network.

For both the development and test sets of the DAIC-WOZ corpus, our hierarchical attention models outperformed the AVEC 2017 Audio-Video baseline, regardless of whether or not an attention transfer mechanism was adopted. Still more importantly, the best MAE and RMSE achieved by our hierarchical monotonic attention model with attention transfer were higher than those achieved by the multi-modal work [8] on both the development set and the test set; however, the best results for MAE and RMSE are a little lower than those of another multi-modal work presented in [9] on the development and test sets, since the results we obtained on this experimental corpus are speech-based only.

As for the attention transfer mechanism introduced in this work, the performance of the hierarchical attention model without attention transfer is observed to be lower than that of the model using the attention transfer mechanism. Therefore, we can conclude that incorporating an attention transfer mechanism into a hierarchical attention model such as ours can be considered an effective solution that is better suited for depression analysis; this in turn validates our hypothesis that learning to mimic the attention maps of the teacher model can be helpful.

Furthermore, we can observe that although the performance of the hierarchical global soft attention model, or classic local soft attention model, is inferior to that of the hierarchical monotonic attention model with or without attention transfer, it still surpasses that of the AVEC 2017 baseline model and also outperforms the multi-modal work proposed in [8] in terms of MAE on both the development set and the test set. However, the performance of the hierarchical global soft attention model or classic local soft attention model is lower than that of model [9] with or without attention transfer. The main reason for this is that the model utilised in [9] is more complicated than that presented in [8]; moreover, the final PHQ-8 score in [9] is obtained using a multivariate regression model from the initial predictions of depressed and non-depressed DCNN-DNN models, as well as the depression classification results. This observation demonstrates the effectiveness of the hierarchical attention strategy when applied to the task of depression classification.

2) Experiments on Hierarchical Attention Autoencoders: In this paper, we propose and develop hierarchical attention autoencoders and apply them to analysing depression using speech. To evaluate the effectiveness of the clinical interview representations, we conducted extensive experiments. We also compared the performance of our hierarchical attention autoencoders with the traditional hierarchical BLSTM autoencoders without attention, hereafter referred to as 'vanilla hierarchical autoencoders,' to investigate the benefits of using attention in our hierarchical autoencoders for depression analysis. For comparison, the vanilla hierarchical autoencoders, along with three hierarchical attention autoencoders, are also included, in addition to the AVEC 2017 baseline methods, hierarchical attention models, and two previous methods [8], [9].

We observed that the best MAE (4.47) and RMSE (5.72) on the test set and the best MAE (3.36) and RMSE (4.67) on the development set of the DAIC-WOZ corpus were achieved when monotonic attention was used in our hierarchical attention autoencoders (cf. Table II).

For both the development and test sets of the DAIC-WOZ corpus, our proposed hierarchical attention autoencoders outperformed the AVEC 2017 Audio-Video baseline, regardless of which kind of attention strategy was used by our attention autoencoders.

Furthermore, the MAE results achieved by our hierarchical attention autoencoders are even better than those obtained by the multi-modal work [8] on the test set; however, our results are a little lower than those achieved by the multi-modal work presented in [9] for both MAE and RMSE.

Moreover, as shown in Table II, the vanilla hierarchical autoencoders did not perform as well as our hierarchical attention autoencoders, although they still surpassed the AVEC 2017 Audio-Video baseline.

In order to investigate the impact of different attention strategies on the performance of the hierarchical attention autoencoders, we further conducted a comparative study in which three separate attention mechanisms – namely, standard global soft attention, local soft attention, and monotonic attention - were experimentally evaluated in the experiment. From the corresponding results shown in Table II, it can be observed that the model with monotonic attention yields the best results; this is consistent with the results of the experiments on the Hierarchical Attention Transfer Network.

In order to verify the efficiency and effectiveness of our autoencoders, we further conducted experiments to compare the performance of our hierarchical attention autoencoders with that of the hierarchical attention model that does not use autoencoders and attention transfer. As can be seen from Table II, regardless of which kind of attention strategy was adopted in our hierarchical attention autoencoders, our proposed method outperformed the purely hierarchical attention models on both the test and development sets. Therefore, hierarchical attention autoencoder such as our can be considered an effective solution that is better suited for depression analysis.

3) Experiment on the Combination of the Hierarchical Attention Transfer Network With Hierarchical Attention Autoencoders: The effectiveness of our hybrid framework can be highlighted through comparison with other key results obtained on the DAIC-WOZ corpus in the literature (Table II). It can be observed that the best MAE (4.20) and RMSE (5.56) on the test set, as well as the best MAE (3.87) and RMSE (2.85) on the development set of the DAIC-WOZ corpus were achieved by leveraging monotonic attention in our proposed hybrid network. Our best results, including MAE and RMSE, are consistently superior to those obtained by the AVEC 2017 Audio-Video baseline and the multi-modal work [8] on both the test set and development set; however, they are slightly lower than those achieved by the multi-modal work presented in [9] on the development set. It can therefore be concluded that using monotonic attention yields the best results, a result that remains consistent across the three experiments.

Furthermore, we observed that the performance of the combined HATN and hierarchical attention autoencoders is superior to that of either of these two methods used alone. This validates our hypothesis that combining attention transfer and hierarchical attention autoencoders results in additional improvement, and is thus a better-suited solution for the task of depression analysis.

V. Conclusion

In this paper, we proposed a novel hierarchical attention-based model that combines unsupervised learning and attention transfer to assess depression severity using speech. Our three core contributions can be summarised as follows. Firstly, we develop and propose an attention transfer process that transfers attentions in order to measure the depression severity in both frame and sentence levels across tasks. Secondly, we propose a novel hierarchical attention autoencoder paradigm that applies attention mechanisms to train hierarchical autoencoders capable of generating representations of depressed speech in an unsupervised manner; Thirdly, through extensive experiments, we demonstrate that the proposed hybrid model achieves the best-known speech results on the AVEC 2017 depression corpus to date.

In our future work, we plan to further explore the use of hierarchical attention variational autoencoders for learning representations. Transferring attention from other speech tasks related to

depression assessment, such as emotion or mood detection, will also be explored.

Acknowledgment

The authors would like to thank Dr. J. Dineley for her proofreading work.

References

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

[2] M. Valstar *et al.*, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. 3rd ACM Int. Workshop Audio/Vis. Emotion Challenge*, Barcelona, Spain, 2013, pp. 3–10.

[3] M. Valstar *et al.*, "AVEC 2014: 3D dimensional affect and depression recognition challenge," in *Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge*, Orlando, FL, USA, 2014, pp. 3–10.

[4] M. Valstar *et al.*, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge*, Amsterdam, The Netherlands, 2016, pp. 3–10.

[5] F. Ringeval *et al.*, "AVEC 2017: Real-life depression, and affect recognition workshop and challenge," in *Proc. 7th Annu. Workshop Audio/Vis. Emotion Challenge*, Mountain View, CA, USA, 2017, pp. 3–9.

[6] F. Ringeval *et al.*, "Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proc. Audio/Vis. Emotion Challenge Workshop*, Seoul, South Korea, 2018, pp. 3–13.

[7] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *J. Biomed. Inform.*, vol. 83, pp. 103–111, May 2018.

[8] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in *Proc. 7th Annu. Workshop Audio/Vis. Emotion Challenge*, Mountain View, CA, USA, 2017, pp. 53–59.

[9] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. Jiang, "Hybrid depression classification and estimation from audio video and text information," in *Proc. 7th Annu. Workshop Audio/Vis. Emotion Challenge*, Mountain View, CA, USA, 2017, pp. 45–51.

[10] L. Yang, D. Jiang, and H. Sahli, "Integrating deep and shallow models for multi-modal depression analysis–hybrid architectures," *IEEE Trans. Affect. Comput.*, to be published.

[11] M. Wöllmer *et al.*, "Abandoning emotion classes – Towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. INTERSPEECH*, Brisbane, Australia, 2008, pp. 597–600.

[12] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. 42nd Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, USA, 2017, pp. 2227–2231.

[13] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. 15th Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, San Diego, CA, USA, 2016, pp. 1480–1489.

[14] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, Australia, 2018, pp. 2225–2235.

[15] L. Stappen, N. Cummins, E. Messner, H. Baumeister, J. Dineley, and B. W. Schuller, "Context modelling using hierarchical attention networks for sentiment and self-assessed emotion detection in spoken narratives," in *Proc. 44th IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brighton, U.K., 2019, pp. 6680–6684.

[16] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. 5th Int. Conf. Learn. Representations*, Toulon, France, 2017, p. 13.

[17] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Attention transfer from web images for video recognition," in *Proc. 25th ACM Int. Conf. Multimedia*, Mountain View, CA, USA, 2017, pp. 1–9.

[18] J. Zhuo, S. Wang, W. Zhang, and Q. Huang, "Deep unsupervised convolutional domain adaptation," in *Proc. 25th ACM Int. Conf. Multimedia*, Mountain View, CA, USA, 2017, pp. 261–269.

[19] J. Kim, S. Park, and N. Kwak, "Paraphrasing complex network: Network compression via factor transfer," in *Proc. 32nd Annu. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2018, pp. 2760–2769.

[20] N. Cummins, A. Baird, and B. W. Schuller, "The increasing impact of deep learning on speech analysis for health: Challenges and opportunities," *Methods, Special Issue Trans. Data Analytics Health Inform.*, vol. 151, pp. 41–54, 2018.

[21] Z. Zhang, N. Cummins, and B. Schuller, "Advanced data exploitation in speech analysis—An overview," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 107–129, Jul. 2017.

[22] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *J. Mach. Learn. Res.*, vol. 11, pp. 625–660, Feb. 2010.

[23] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Semi-supervised autoencoders for speech emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 1, pp. 31–43, Jan. 2018.

[24] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proc. ICML Workshop Unsupervised Transfer Learn.*, Bellevue, WA, USA, 2012, pp. 37–49.

[25] W. W. Ng, G. Zeng, J. Zhang, D. S. Yeung, and W. Pedrycz, "Dual autoencoders features for imbalance classification problem," *Pattern Recognit.*, vol. 60, pp. 875–889, Dec. 2016.

[26] H.-C. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach, "Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1930–1943, Jan. 2013.

[27] J. Li, M.-T. Luong, and D. Jurafsky, "A hierarchical neural autoencoder for paragraphs and documents," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process.*, Beijing, China, 2015, pp. 1106–1115.

[28] M. Zhang and Y. Wu, "An unsupervised model with attention autoencoders for question retrieval," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018, pp. 4978–4985.

[29] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 2997–3000.

[30] N. Cummins, B. Vlasenko, H. Sagha, and B. Schuller, "Enhancing speech-based depression detection through gender dependent vowel-level formant features," in *Proc. 16th Conf. Artif. Intell. Med.*, Vienna, Austria, 2017, pp. 209–214.

[31] H. Jiang *et al.*, "Investigation of different speech types and emotions for detecting depression using different classifiers," *Speech Commun.*, vol. 90, pp. 39–46, Jun. 2017.

[32] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 142–150, Apr. 2013.

[33] S. Alghowinem *et al.*, "A comparative study of different classifiers for detecting depression from spontaneous speech," in *Proc. 44th IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, Canada, 2013, pp. 8022–8026.

[34] V. Mitra *et al.*, "The SRI AVEC-2014 evaluation system," in *Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge*, Orlando, FL, USA,2014, pp. 93–101.

[35] J. R. Williamson *et al.*, "Detecting depression using vocal, facial and semantic communication cues," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge*, Amsterdam, The Netherlands, 2016, pp. 11–18.

[36] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge*, Orlando, FL, USA, 2014, pp. 65–72.

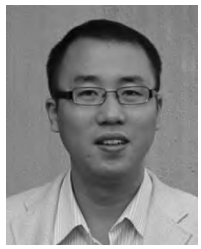
[37] N. Cummins, V. Sethu, J. Epps, J. R. Williamson, T. F. Quatieri, and J. Krajewski, "Generalized two-stage rank regression framework for depression score prediction from speech," *IEEE Trans. Affect. Comput.*, to be published.

[38] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski, "Analysis of acoustic space variability in speech affected by depression," *Speech Commun.*, vol. 75, pp. 27–49, 2015.

[39] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge*, 2016, pp. 43–50.

[40] S. Scherer, G. M. Lucas, J. Gratch, A. Skip Rizzo, and L. Morency, "Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 59–73, Jan.–Mar. 2016.

[41] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia*, Barcelona, Spain, 2013, pp. 835–838.
[42] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarepa collaborative voice analysis repository for speech technologies," in *Proc. 39th IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, 2014, pp. 960–964.
[43] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge*, Amsterdam, The Netherlands, 2016, pp. 35–42.
[44] B. Sun, Y. Zhang, J. He, Y. Xiao, and R. Xiao, "An automatic diagnostic network using skew-robust adversarial discriminative domain adaptation to evaluate the severity of depression," *Comput. Methods Programs Biomed.*, vol. 173, pp. 185–195, May 2019.
[45] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.
[46] E. Tzinis and A. Potamianos, "Segment-based speech emotion recognition using recurrent neural networks," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact.*, San Antonio, TX, USA, 2017, pp. 190–195.
[47] C. W. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition," in *Proc. INTERSPEECH*, San Francisco, CA, USA, 2016, pp. 1387–1391.
[48] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. 29th Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2015, pp. 577–585.
[49] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," in *Proc. 29th Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2015, pp. 2773–2781.
[50] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, 2014, p. 15.
[51] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, "Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 272–276.
[52] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Proc. 27th Int. Conf. Artif. Neural Netw.*, Rhodes, Greece, 2018, pp. 270–279.
[53] Y. Bengio *et al.*, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, Jan. 2009.
[54] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3107–3111.
[55] 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.*, Geneva, Switzerland, 2013, pp. 511–516.
[56] 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, May 2014.
[57] J. Deng, X. Xu, Z. Zhang, S. Frühholz, 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.
[58] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 4133–4141.
[59] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *Proc. 3th Int. Conf. Learn. Representations*, San Diego, CA, USA, 2015, p. 13.
[60] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, 2015, pp. 1412–1421.
[61] C.-C. Chiu and C. A. Raffel, "Monotonic chunkwise attention," in *Proc. 6th Int. Conf. Learn. Representations*, Vancouver, BC, Canada, 2018, p. 16.
[62] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," in *Proc. 34th Int. Conf. Mach. Learn.*, Sydney, NSW, Australia, 2017, pp. 2837–2846.
[63] A. Tjandra, S. Sakti, and S. Nakamura, "Local monotonic attention mechanism for end-to-end speech and language processing," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.*, Taipei, Taiwan, Nov. 2017, pp. 431–440.
[64] J. Gratch *et al.*, "The distress analysis interview corpus of human and computer interviews," in *Proc. 9th Int. Conf. Lang. Resour. Eval.*, Reykjavik, Iceland, 2014, pp. 3123–3128.
[65] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *J. Affect. Disorders*, vol. 114, no. 1, pp. 163–173, Apr. 2009.
[66] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders: DSM-4*, 4th ed. Washington, DC, USA: American Psychiatric Press, 1994.
[67] Z. Zhao, Y. Zhao, Z. Bao, H. Wang, Z. Zhang, and C. Li, "Deep spectrum feature representations for speech emotion recognition," in *Proc. ASMMC-MMAC*. Seoul, South Korea, 2018, pp. 27–33.
[68] K. Lenzo, "The CMU pronouncing dictionary," 2007. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict



Ziping Zhao was born in Tianjin, China, in 1980. He received the B.Sc. and M.S. degrees in computer science from Tianjin Normal University, Tianjin, China, in 2002 and 2005, respectively, and the Ph.D. degree for his study on the automatic prediction of prosodic phrases from Nankai University, Tianjin, China, in 2008. In 2008, he started his teaching career with Tianjin Normal University. In 2010, he studied with the Key Laboratory of Trustworthy Computing, East China Normal University as a Visiting Scholar. In 2016, he became the Vice Dean of the College of Computer and Information Engineering, Tianjin Normal University. His research fields are affective computing and machine learning.

Zhongtian Bao was born in Ningbo, Zhejiang, China, in 1999. He received the undergraduate degree from Nanjing University, Nanjing, China, in 2017. Since 2018, he has been working toward the postgraduate degree from Tianjin Normal University, Tianjin, China. His research interests include speech emotion recognition and applications.

Zixing Zhang received the master's degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2010, and the Ph.D. degree in computer engineering from the Technical University of Munich, Munich, Germany, in 2015. Since 2017, he has been a Research Associate with the Department of Computing, Imperial College London, London, U.K. Before that, he was a Postdoctoral Researcher with the University of Passau, Germany, from 2015 to 2017. He has authored more than 80 publications in peer-reviewed books, journals, and conference proceedings to date. His research mainly focuses on deep learning technologies for the speaker-centered state (e.g., emotion) and health computing. He has organized special sessions, such as at the IEEE 7th ACII in 2017 and at the 43rd ICASSP in 2018, and a special issue in the IEEE TETCI in 2019. He also serves as a Reviewer for numerous leading-in-their fields journals and conferences, and as a Programme Committee Member and Area Chair for many international conferences.

Jun Deng received the bachelor's degree in electronic and information engineering from Harbin Engineering University, Harbin, China, in 2009, the master's degree in information and communication engineering from the Harbin Institute of Technology, Heilongjiang, China, in 2011, and the doctorate degree, for his study on Feature Transfer Learning for Speech Emotion Recognition, in electrical engineering and information technology from Technische Universität München, Munich, Germany, in 2016. He was a Postdoctoral Researcher and the Chair of Complex and Intelligent Systems with the University of Passau, Passau, Germany. He is currently the Head of Deep Learning, Agile Robots AG, Gilching, Germany. His interests are machine learning methods such as transfer learning and deep learning with an application preference to affective computing and robotics.

Nicholas Cummins received the undergraduate degree (with first class Hons.) from UNSW, Sydney, NSW, Australia, in 2011, and the Ph.D. degree in electrical engineering from UNSW, Sydney, NSW, Australia, in February 2016. He is currently working toward the Habilitation degree with the ZB.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, where he works on the Horizon 2020 projects DE-ENIGMA, RADAR-CNS, sustAGE, and TAPAS. He is also an External Researcher on the National Science Foundation of China Grant 31860285, Diagnosis of depression by speech signals. His Ph.D. investigated whether the voice can be used as an objective marker in the diagnosis and monitoring of clinical depression. He has authored or coauthored regularly in the field of depression detection since 2011; these papers have attracted significant attention and citations. His current research interests include areas of behavioral signal processing with a focus on the automatic multisensory analysis and understanding of different health states.

Haishuai Wang received the Ph.D. degree in computer science from the Center of Artificial Intelligence, University of Technology Sydney, Ultimo NSW, Australia. He is an Assistant Professor of computer science, Fairfield University, Fairfield, CT, USA. He is also a Visiting Assistant Professor of Biomedical Informatics, Harvard University, Cambridge, MA, USA. He did his postdoc training with Harvard University and Washington University in St. Louis. His research focuses on data mining, machine learning, and bioinformatics.

Jianhua Tao received the Ph.D. degree from Tsinghua University, Beijing, China, in 2001, and the M.S. degree from Nanjing University, Nanjing, China, in 1996. He is currently a Professor with NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He has authored or coauthored more than 80 papers on major journals and proceedings including the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. His current research interests include speech synthesis and coding methods, human computer interaction, multimedia information processing, and pattern recognition. He has received several awards from the important conferences, such as Eurospeech, NCMMSC, etc. He was the chair or program committee member for several major conferences, including ICPR, ACII, ICMI, ISCSLP, NCMMSC, etc. He also serves as the steering committee member for the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, an Associate Editor for *Journal on Multimodal User Interface* and *International Journal on Synthetic Emotions*, and the Deputy Editor-in-Chief for *Chinese Journal of Phonetics*.

Björn Schuller (M'06–SM'15–F'18) received the diploma in 1999, the doctoral 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 Technische Universität München, Munich, Germany. He is a tenured Full Professor heading the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, and a Professor of Artificial Intelligence heading GLAM the Group on Language, Audio & Music, Department of Computing, Imperial College London, London, U.K. Dr. Schuller was elected member of the IEEE Speech and Language Processing Technical Committee, an Editor in Chief for the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, the President-Emeritus of the AAAC, and a Senior Member of the ACM. He has (co-)authored five books and more than 800 publications in peer-reviewed books, journals, and conference proceedings, leading to more than 24 000 citations (h-index 73).