# ATTENTION-AUGMENTED END-TO-END MULTI-TASK LEARNING FOR EMOTION PREDICTION FROM SPEECH

*Zixing Zhang*$^{\star 1,2}$, *Bingwen Wu*$^{\star 3}$, *Björn Schuller*$^{1,2,4}$

[1]GLAM – Group on Language, Audio & Music, Imperial College London, UK
[2] audEERING GmbH, Germany
[3] Department of Computing, Imperial College London, UK
[4] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing,
University of Augsburg, Germany
{zixing.zhang|bingwen.wu17}@imperial.ac.uk

## ABSTRACT

Despite the increasing research interest in end-to-end learning systems for speech emotion recognition, conventional systems either suffer from the overfitting due in part to the limited training data, or do not explicitly consider the different contributions of automatically learnt representations for a specific task. In this contribution, we propose a novel end-to-end framework which is enhanced by learning other auxiliary tasks and an attention mechanism. That is, we jointly train an end-to-end network with several different but related emotion prediction tasks, i. e., arousal, valence, and dominance predictions, to extract more robust representations shared among various tasks than traditional systems with the hope that it is able to relieve the overfitting problem. Meanwhile, an attention layer is implemented on top of the layers for each task, with the aim to capture the contribution distribution of different segment parts for each individual task. To evaluate the effectiveness of the proposed system, we conducted a set of experiments on the widely used database IEMOCAP. The empirical results show that the proposed systems significantly outperform corresponding baseline systems.

***Index Terms***— Speech emotion prediction, end-to-end, attention mechanism, multi-task learning

## 1. INTRODUCTION

Automatic speech emotion prediction endows machines with the capability of natural and empathic communication with humans, which is considered to be essential to sustain long-term human–machine interactions. In spite of remarkable advances over the past decades [1], the extraction of representative features associated with emotions remains an open challenge. The conventional approaches normally extract a variety of acoustic descriptors, such as pitch and energy, on the frame level in the first place. Then, mostly they derive the super-segmental features via applying some mathematical functionals (e. g., mean and maximum) [2, 3], or counting the normalised occurrence frequency of certain frame-level acoustic feature units [4].

However, these approaches have several disadvantages. All these approaches largely require acoustic experts and psychologists to manually design the features. Only the feature attributes that explicitly showed high correlation with emotion, normally through extensive and carefully prepared experiments, will be selected [2], which is quite time-consuming and exhausting. Moreover, the effectiveness of selected features still heavily depends on the implemented pattern recognition model [2], resulting in their lower generality. In this regard, end-to-end learning has emerged as a promising alternative [5–8]. It aims to *automatically* explore the most salient representations related to the task of interest by using neural networks to jointly train the representation extraction process and the pattern recognition process, wiping away the brute-force feature designing procedure.

Since the inception [9], a number of end-to-end learning frameworks have been quickly and widely applied to various speech-related tasks, for example, speech recognition [10, 11], speaker recognition [12], and speech synthesiser [13]. As to speech emotion prediction, the first end-to-end work was shown in [5], where the authors intended to extract implicit representations directly from digital raw signals by using one-dimensional Convolutional Neural Networks (CNNs) followed by Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) for learning a sequential pattern. Due to its great success, this end-to-end framework has been extended to deal with other modalities for emotion recognition, such as video and electroencephalogram signals [6, 8]. Likewise, other similar end-to-end frameworks were further proposed and investigated for speech emotion prediction. For example, in [8], the conventional CNNs were replaced by time-delay neural networks whereas the LSTM-RNNs were kept following behind.

Nevertheless, these proposed frameworks easily suffer from overfitting [5], leading to severe performance degradation when the data mismatch increases between the training and evaluation phases. This is because of not only the limited size of training data, but also further concerns, such as the task-specific training [14]. In this regard, in this paper, we propose to integrate multi-task learning (MTL) into the end-to-end framework, which intends to jointly train several different, but related tasks simultaneously. By doing this, it is assumed that the more tasks are learnt simultaneously, the more common representations shared by all of the tasks and the less chance of overfitting on the original task will be gained [15].

Moreover, when modelling the automated learnt representative sequence, the representations at each time point are normally equally considered [5, 6]. This process largely ignores the different importance of the parts within one unit of analysis with respect to different

Both authors contributed equally to this work. This work was supported by the TransAtlantic Platform "Digging into Data" collaboration grant (ACLEW: Analysing Child Language Experiences Around The World), with the support of the UK's Economic & Social Research Council through the research Grant No. HJ-253479.

emotions. For instance, the work done in [16] has shown that the short silence periods within an utterance often have little relevance with emotions. To this end, we further propose to implement an attention mechanism to the end-to-end frameworks, hoping that it can automatically learn the most interesting parts of an utterance containing strong characteristics relating to the given emotions.

Therefore, the main contribution of this paper pertains to the proposal of a novel end-to-end framework, which is augmented with an attention mechanism and jointly trained with multiple auxiliary tasks, for speech emotion prediction. To the best of our knowledge, this is the first time to investigate such an end-to-end framework in the context of speech processing.

## 2. RELATED WORK

For speech emotion prediction, MTL has been frequently utilised. Eyben et al. [17] firstly proposed to jointly train five different emotional dimensions for continuous emotion recognition. The experimental results have clearly indicated that the MTL model remarkably outperforms single-task-based models. Following this work, Han et al. [18] combined the emotion prediction with an annotation uncertainty as joint tasks to be learnt together. Xia and Liu [19] suggested incorporating the losses from both the categorical and the dimensional emotion recognition to optimise the neural networks. Zhang et al. [20] investigated MTL in a cross-corpus scenario, where many auxiliary tasks, such as corpus, domain, and gender distinctions, were considered to be optimised along with emotion recognition. Other similar works have also been done in [21]. However, most of these studies have focused on the usage of hand-crafted features.

As to the attention mechanism, Mirsamadi [16] firstly integrated an attention layer within Deep Neural Networks (DNNs), resulting in a significant performance improvement for speech emotion prediction. Similarly, Zhao et al. [22] implemented an attention layer right after the RNNs to extract the most interesting acoustic parts in the continuum. Apart from the RNNs and DNNs, the attention layer was also integrated with CNNs [23, 24]. All these works, nevertheless, were conducted under the usage of traditional hand-crafted features, and have not explicitly investigated the differences of attention in an MTL framework.

## 3. ATTENTION-AUGMENTED END-TO-END MULTI-TASK LEARNING

Figure 1 illustrates the proposed end-to-end framework for speech emotion prediction, which can be considered as an extension of a basic end-to-end system, augmented with attention and MTL strategies. In the following subsections, we comprehensively describe the framework.

### 3.1. Single-Task End-to-End Framework

Despite several existing end-to-end frameworks for speech emotion recognition, we retained the basic network structure in our previous work [5, 6]. This is due to its effectiveness being well demonstrated in continuous emotion recognition, and its widespread usage in many other computational paralinguistic tasks [25].

The basic single-task-based end-to-end system generally consists of a feature extraction modelling and a sequence modelling. More specifically, the feature extraction modelling mainly consists of two one-dimensional convolutional layers each followed by an element-wise rectified linear non-linearity (ReLU) $max(0, x)$ and by additional max pooling layer. The reason behind the usage of CNNs mainly lies in their well-known capability of feature extraction not only in image processing, but also in speech processing [5]. Particularly, one dropout layer is employed to increase the model generalisation. In contrast to the feature extraction modelling, the sequence modelling employs two recurrent layers equipped with Gated Recurrent Units (GRUs), due to their effectiveness in modelling temporal patterns and less complexity in comparison with LSTMs [26].

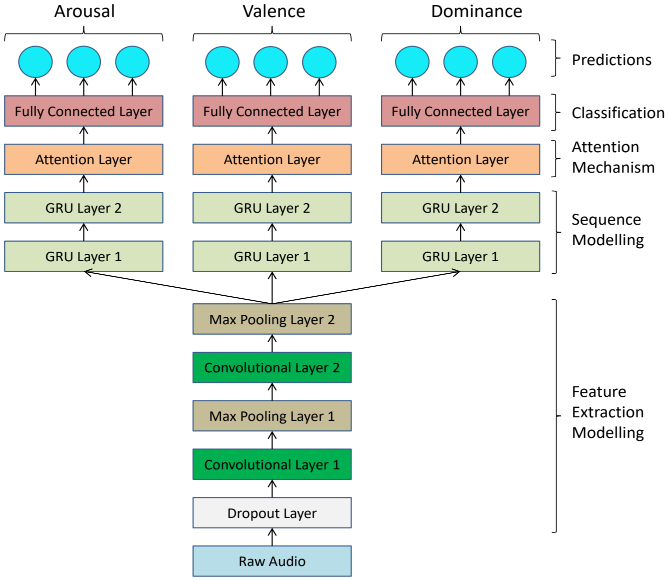

**Fig. 1**. The framework of the attention-augmented end-to-end multi-task learning for speech emotion prediction.

Given an utterance in form of raw audio signals $s(t)$, it is firstly split into several sequential segments $\{\mathbf{s}_1, \dots, \mathbf{s}_L\}$ ($L$ indicates the number of the obtained segments given an utterance) with a sliding fixed-length window. Then, each segment is successively fed into the feature extraction modelling ($f_c$) so that input raw signals of each segment are transformed into one vector (i. e., representation). That is, given a segment $\mathbf{s}_i$, one obtains

$$\mathbf{v}_i = f_c(\mathbf{s}_i), \tag{1}$$

which is assumed to well represent the temporal speech patterns. After that, the extracted representation $\mathbf{v}_i$ is further successively fed into the sequence modelling ($f_r$), i. e.,

$$\mathbf{h}_i = f_r(\mathbf{v}_i), \tag{2}$$

leading to a new output sequence from the last recurrent layer, i. e., $\{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_L\}$. Normally, only the last output of the sequence $\mathbf{h}_L$ is used and fed into the final softmax layer for emotion prediction.

### 3.2. Weighted Pooling with Attention

The principle of the attention mechanism originally stemmed from the characteristic of human perception. That is, humans normally focus attention selectively on parts of the visual or auditory space to acquire information when and where it is needed, and combine information from different fixations over time to build up an internal representation of the scene [27]. Nowadays, the attention mechanism has been widely used in image processing and natural language processing [28].

By far, a variety of attention mechanisms have been investigated in machine learning. According to whether the calculation of attention requires to access positions across sequences, they can be generally categorised into inter-attention and intra-attention mechanisms. Intra-attention, also known as self-attention, is often used to compute a representation of a sequence by leveraging different importance of the parts in a sequence. In this contribution, we employed an intra-attention layer following the last recurrent layer as illustrated in the top of Fig. 1.

Mathematically, given an output sequence of the last recurrent layer $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_L\}$, the attention layer tries to deliver an utterance-level representation for such a sequence with the following equation

$$\mathbf{r} = \sum_{i=1}^{L} \alpha_i \mathbf{h}_i, \tag{3}$$

where $\alpha_i$ stands for the weight of the output $\mathbf{h}_i$ at the $i$-th frame. From this equation, it can be seen that the attention layer can be considered as a weighted-average pooling layer, in comparison with the traditional maximum, average, or the last output pooling strategies. Finally, the obtained new representation $\mathbf{r}$ is fed into one fully connected layer for emotion prediction.

Therefore, the calculation of weight $\alpha_i$ becomes the central problem of the attention layer. Specifically, $\alpha_i$ is computed by

$$\alpha_i = \frac{\exp(\mathbf{w}^T \mathbf{h}_i)}{\sum_{t=1}^{L} \exp(\mathbf{w}^T \mathbf{h}_t)}, \tag{4}$$

where $\mathbf{w}$ is the learnable parameter vector, and the inner product between $\mathbf{w}$ and $\mathbf{h}_i$ is interpreted as a score for the contribution of the frame $t$. Besides, in this equation, a softmax function is applied leading to the sum of the weight distribution to be a unity.

### 3.3. Joint Training with Auxiliary Tasks

As discussed in Section 1, end-to-end learning frameworks normally require massive training data, which, however, are largely absent in the context of emotion recognition, resulting in a severe overfitting problem. To improve the model generalisation, in this paper we endeavour to seek help from other auxiliary tasks through MTL. The underlying idea is that the model learns more tasks simultaneously will contribute to more robust learnt representations that capture all of the tasks [14].

Figure 1 illustrates the structure of the proposed MTL. From the technical view of point, MTL is a process of learning multiple tasks *concurrently*. Typically, there is one main task and one or more auxiliary tasks. By attempting to model the auxiliary tasks together with the main task, the model learns shared information among tasks, which may be beneficial to learning the main task. In this paper, the tasks refer to three-dimensional emotions, i. e., arousal, valence, and dominance predictions.

Mathematically, the objective function in MTL can be formulated as:

$$\mathcal{J}(\boldsymbol{\theta}_0) = \sum_{m=1}^{M} w_m L_m(\mathbf{x}, y_m, [\boldsymbol{\theta}_m; \boldsymbol{\theta}_c]) + \lambda R(\boldsymbol{\theta}_0), \tag{5}$$

where $M$ denotes the number of tasks and $L_m(\cdot)$ represents the loss function of the task $m$, which is weighted by $w_m$. The weights $w_m$ are optimised by a random search. $\boldsymbol{\theta}_c$ and $\boldsymbol{\theta}_m$ represent, respectively, the shared and task-specific model parameters with respect to the task $m$, whereas $\boldsymbol{\theta}_0$ indicates all shared and task-specific network parameters, and $\lambda$ is a hyper-parameter that controls the importance of the regularisation term $R(\boldsymbol{\theta}_0)$ (i. e., L2 in our case). In the network training process, the network is optimised by minimising the objective function $\mathcal{J}(\boldsymbol{\theta}_0)$.

## 4. EXPERIMENTS AND RESULTS

In this section, we implement and evaluate our approach for emotion classification on an emotion database.

### 4.1. Selected Database

To validate the proposed paradigm, we used the widely used Interactive Emotional dyadic MOtion CAPture (IEMOCAP) database, which contains approximately 12 hours of audio-visual recordings from five pairs of experienced actors [29]. For each improvised interaction between two actors, they communicated with each other in scenarios where specific emotions were elicited. The recordings were then segmented into utterances and further annotated in all three-dimensional aspects, i. e., activation, valence, and dominance, on a five-point scale by at least two different annotators.

For our experiments, only the audio recordings were utilised. Following the work of [30], we further discretised the five-point scale into three levels (classes) – low level contains ratings in the range [1,2], middle level contains ratings in the range (2,4), and high level contains ratings in the range [4,5] [30]. We divided the dataset into three speaker independent partitions, i. e., 6 319 for the training set (session 1-3), 1 811 for the development set (session 4), and 1 819 for the test set (session 5). All the recordings were sampled with 16 kHz.

### 4.2. Implementation Details

Before feeding the raw speech signal into the network, we applied an online standardisation to the development and test sets by using the mean and standard deviation information from the training set. The raw speech signals were then split into sub-segments with a fixed-size window of 40 ms at a step size of 10 ms. Given the 16 kHz sampling rate of raw signals, the network input vector is of dimension 640 for each sub-segment. For the first and second convolutional layers, we used 40 filters with the size of 40, resulting in 40 feature maps after each layer. For the first max pooling layer, we took a kernel with the size of two in a zero-padding strategy, leading to feature maps with a dimension of 320; whereas, for the second max pooling layer, we used a cross-channel max pooling with the pool size of 10, yielding to four feature maps with the dimension of 320. Finally, the obtained feature map is expanded and concatenated as a long vector with 1 280 dimensions, which is the extracted representation for each sub-segment. To improve the model generalisation, we set the keep probability of the dropout layer to be 0.9. For the sequence modelling, we employed 128 nodes per GRU hidden layer. The training of the proposed framework was conducted using the Adam optimisation algorithm with a learning rate of 0.0001. Note that all these network and training hyper-parameters were optimised on the development.

To evaluate the model performance, we utilised the frequently used metric Unweighted Average Recall (UAR), i. e., the sum of classwise recall divided by the number of classes, for emotion recognition.

### 4.3. Results and Discussion

To compare the performance of the proposed approach with other traditional speech emotion prediction systems, we conducted two experiments with hand-engineered acoustic features. That is, we

**Table 1**. Performance comparison (UAR: unweighted average recall) between the proposed attention-augmented end-to-end multi-task learning system with other baseline systems as well as other traditional recognition models on the development and the test partitions for activation, valence, and dominance predictions. OS: openSMILE features; SVMs: support vector machines; RNNs: recurrent neural networks; STL: single-task learning; MTL: multi-task learning; e2e: end-to-end learning; att.: attention. The sign of '$\star$' indicates statistic significance (one-tailed $z$-test, $p < .05$) of performance improvement of the proposed systems in comparison with the baseline system (i. e., e2e STL).

| UAR [%] | arousal | | valence | | dominance | |
|---|---|---|---|---|---|---|
| methods | dev | test | dev | test | dev | test |
| OS+SVMs | 52.1 | 50.5 | 50.5 | 49.8 | 39.1 | 48.7 |
| OS+RNNs | 57.8 | 53.1 | 51.8 | 51.0 | 56.4 | 49.6 |
| e2e STL | 45.3 | 45.1 | 60.9 | 60.1 | 50.9 | 51.1 |
| e2e STL+att. | 46.5 | 45.4 | 61.4 | 60.7 | 51.4 | 52.6 |
| e2e MTL | 46.2 | 44.0 | 64.6 | 63.4 | 52.3 | **53.9** |
| e2e MTL+att. | **48.7**$^\star$ | **48.5**$^\star$ | **66.2**$^\star$ | **63.8**$^\star$ | **53.4** | 51.6 |

used our opensource toolkit openSMILE [2] to extract a minimalistic expert-knowledge based feature set [31], which contains 23 Low-Level Descriptors (LLDs). After that, we applied a set of statistical functionals to the LLDs, leading to 88 acoustic features (i. e., eGeMAPS) on the utterance level. As to the classifier, we utilised the sequence classifier of RNNs to model frame-level features; whereas utilised the static classifier of Support Vector Machines (SVMs) to model the utterance-level features. Both systems have been successfully and frequently utilised for speech emotion prediction [3, 31].

Table 1 shows the obtained results in terms of UAR from the proposed attention-augmented end-to-end MTL system, the related baseline systems, as well as the aforementioned other state-of-the-art systems. It is noted that the basic end-to-end (e2e) learning systems refer to the ones without attention and MTL learning strategies (see Section 3.1), and take the last output from the last recurrent layer for a final prediction. It can be seen that the e2e systems are competitive to the two state-of-of-the-art systems based on hand-crafted features for both the valence and dominance predictions but not for the arousal prediction. This generally confirms our previous findings [5, 6].

When integrating the attention strategy into the baseline systems (i. e., e2e STL), one can note that the system performance is generally improved on all three prediction tasks. These findings suggest that the attention mechanism does not only work in the conventional learning framework with hand-crafted features [3], but also in the proposed end-to-end framework. In parallel, when conducting the MTL strategy into the baseline systems (i. e., e2e STL), similar observations are made. That is, the end-to-end MTL systems are superior to the task-specific baseline systems in most cases. This conclusion implies that the MTL method can partially increase the generalisation of the extracted representations, i. e., the information learnt from other auxiliary tasks can benefit the task of interest, even in an end-to-end learning framework.

Moreover, the incorporation of attention and MTL strategies achieves the best performance in five out of six cases. For example, the obtained results for arousal and valence predictions are achieved at 48.5 % and 63.8 % UAR, which significantly (one-tailed $z$-test, $p < .05$) outperform the baseline results (i. e., 45.1 % and 60.9 % UAR) on the test set. This suggests that both attention mechanism and MTL can work in a complementary way.

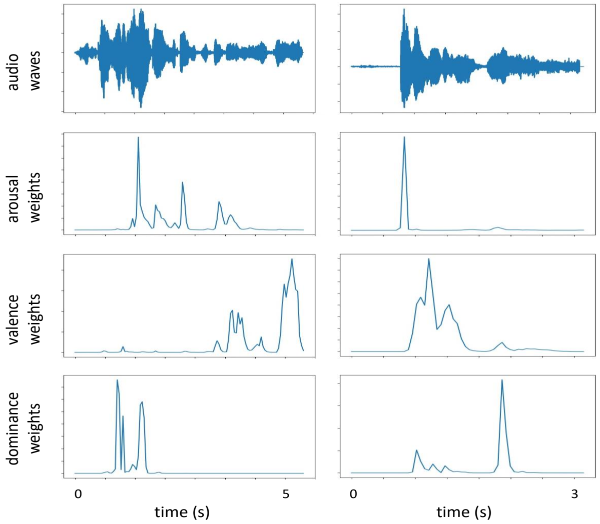

**Fig. 2**. Automatically learnt attention distribution for arousal (second row from top), valence (third row), and dominance (fourth row) predictions for two randomly selected audio wave files (first row).

To further investigate the effectiveness of the attention mechanism in the proposed end-to-end MTL framework, we illustrate the learnt attention across different tasks for two audio files in Fig. 2. Generally speaking, one can observe that the learnt attention weight distributions remarkably differ each task. That is, the arousal, valence, and dominance prediction tasks learnt their individual higher attention on the same segment parts, which matches our previous assumption in Section 1. Particularly, one can see that for arousal prediction the learnt attention weights (refer to the second row of Fig. 2) are highly correlated with the parts with high amplitude. Nevertheless, a similar observation is not made for valence prediction (refer to the third row of Fig. 2). In contrast, the parts with low speech amplitude often contribute more than the parts with high amplitude. This matches our previous knowledge that the valence prediction has limited relation to speech amplitude [2]. In addition, from the fourth row of Fig. 2, it can be seen that dominance prediction lays more attention on some parts with high amplitude.

## 5. CONCLUSION

With an end-to-end (e2e) learning framework, we, on the one hand, took a multi-task learning (MTL) strategy to improve the robustness of the learnt representations that are shared among several tasks. On the other hand, we integrated a self-attention layer on top of the layers for each prediction task, in order to distil more salient representations on the utterance level for a task of interest.

The experimental results obtained by performing experiments on the IEMOCAP database have shown that either the MTL-based or the attention-augmented e2e systems outperform the single-task-based e2e systems, which suggests the effectiveness of the proposed e2e learning framework. However, we also find that for arousal the introduced frameworks are inferior to the baseline systems with the classic functional-based features. This might be because the hand-crafted features are somewhat able to better represent the patterns for arousal due to its simplicity than for other tasks (i. e., valence), according to our prior knowledge.

## 6. REFERENCES

[1] Z. Zhang, J. Han, and B. Schuller, "Dynamic difficulty awareness training for continuous emotion prediction," *IEEE Transactions on Multimedia*, 2018, 13 pages, in print.

[2] F. Eyben, *Real-time speech and music classification by large audio feature space extraction*. Berlin, Germany: Springer, 2015.

[3] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. INTERSPEECH*, Singapore, 2014, pp. 223–227.

[4] J. Han, Z. Zhang, M. Schmitt, Z. Ren, F. Ringeval, and B. Schuller, "Bags in bag: Generating context-aware bags for tracking emotions from speech," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, 3082–3086.

[5] G. Trigeorgis, F. Ringeval, R. Bruckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a Deep Convolutional Recurrent Network," in *Proc. ICASSP*, Shanghai, China, 2016, pp. 5200–5204.

[6] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[7] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Emotion identification from raw speech signals using DNNs," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 3097–3101.

[8] J. Huang, Y. Li, J. Tao, Z. Lian, and J. Yi, "End-to-end continuous emotion recognition from video using 3D ConvLSTM networks," in *Proc. ICASSP*, Calgary, Canada, 2018, pp. 6837–6841.

[9] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. ICML*, Beijing, China, 2014, pp. 1764–1772.

[10] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proc. ASRU*, Scotsdale, AZ, 2015, pp. 167–174.

[11] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *Proc. ICML*, New York City, NY, 2016, 173-182.

[12] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. ICASSP*, Shanghai, China, 2016, pp. 5115–5119.

[13] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 4006–4010.

[14] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, June 2017.

[15] J. Baxter, "A bayesian/information theoretic model of learning to learn via multiple task sampling," *Machine learning*, vol. 28, no. 1, pp. 7–39, July 1997.

[16] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. ICASSP*, New Orleans, LA, 2017, pp. 2227–2231.

[17] F. Eyben, M. Wöllmer, and B. Schuller, "A multitask approach to continuous five-dimensional affect sensing in natural speech," *ACM Transactions on Interactive Intelligent Systems*, vol. 2, no. 1, pp. 1–29, Mar. 2012.

[18] J. Han, Z. Zhang, M. Schmitt, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proc. ACM MM*, Mountain View, CA, 2017, pp. 890–897.

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

[20] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE Transactions on Affective Computing*, 2017, 14 pages, in print.

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

[22] 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.

[23] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.

[24] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 3087–3091.

[25] Z. Zhang, J. Han, K. Qian, and B. Schuller, "Evolving learning for analysing mood-related infant vocalisation," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 142–146.

[26] Z. Zhang, D. Liu, J. Han, and B. Schuller, "Learning audio sequence representations for acoustic event classification," *arXiv preprint arXiv:1707.08729*, July 2017.

[27] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Proc. NIPS*, Montreal, Canada, 2014, pp. 2204–2212.

[28] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, San Diego, CA, 2015.

[29] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[30] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, Apr. 2012.

[31] F. Eyben, K. R. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, Apr. 2016.