Hierarchical Component-attention Based Speaker Turn Embedding for Emotion Recognition

Shuo Liu
Chair of Embedded Intelligence for Health Care and Wellbeing
University of Augsburg
Augsburg, Germany
shuo.liu@informatik.uni-augsburg.de

Jinlong Jiao
College of Computer and Information Engineering
Tianjin Normal University
Tianjin, China
jiaojinlong@stu.tjnu.edu.cn

Ziping Zhao
College of Computer and Information Engineering
Tianjin Normal University
Tianjin, China
zhaoziping@tjnu.edu.cn

Judith Dineley
Chair of Embedded Intelligence for Health Care and Wellbeing
University of Augsburg
Augsburg, Germany
judith.dineley@informatik.uni-augsburg.de

Nicholas Cummins
Chair of Embedded Intelligence for Health Care and Wellbeing
University of Augsburg
Augsburg, Germany
nicholas.cummins@ieee.org

Björn Schuller
GLAM – Group on Language, Audio & Music
Imperial College London
London, UK
bjoern.schuller@imperial.ac.uk

***Abstract*—Traditional discrete-time *Speech Emotion Recognition* (SER) modelling techniques typically assume that an entire speaker chunk or turn is indicative of its corresponding label. An alternative approach is to assume emotional saliency varies over the course of a speaker turn and use modelling techniques capable of identifying and utilising the most emotionally salient segments, such as those with higher emotional intensity. This strategy has the potential to improve the accuracy of SER systems. Towards this goal, we developed a novel hierarchical recurrent neural network model that produces turn level embeddings for SER. Specifically, we apply two levels of attention to learn to identify salient emotional words in a turn as well as the more informative frames within these words. In a set of experiments on the *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) database, we demonstrate that component-attention is more effective within our hierarchical framework than both standard soft-attention and conventional local-attention. Our best network, a hierarchical component-attention network with an attention scope of seven, achieved an *Unweighted Average Recall* (UAR) of 65.0 % and a *Weighted Average Recall* (WAR) of 66.1 %, outperforming other baseline attention approaches on the IEMOCAP database.**

***Index Terms*—Hierarchical attention network, Speech emotion recognition, Component-attention, Turn embedding**

I. Introduction

Speech emotion recognition (SER), whose purpose is to identify the emotional state of an individual from their speech, continues to be a popular topic for researchers in *human-computer interaction* (HCI) [1]–[3] and beyond. It exploits the rich emotional content of speech, which has previously been demonstrated by research in psychology and affective computing [4]–[6]. Discrete SER tasks use machine learning to label utterances, turns, or chunks of speech with a single emotional label, such as happy, sad, or angry. Traditional approaches do this by feeding 'handcrafted' audio features from the speech signal into a suitable machine learning algorithm [7]–[9]. However, designing features that reflect the emotional content of speech requires extensive research [10]. Furthermore, features of this kind that have been developed to date arguably lack the specificity required for emotion modelling.

Recently, neural network models have managed to outperform classic signal processing approaches in many speech-related applications, including *automatic speech recognition* (ASR) [11], [12], speech enhancement [13], [14], and speech generation [15]. *Recurrent neural networks* (RNNs), which are capable of capturing time-dependencies in sequential data, can code speech into its high-level representations and have shown promising results in many SER tasks [16]–[19].

A particular issue in discrete SER tasks is that each particular chunk of speech has a single label. It is often assumed that all input data in a given speaker turn or chunk is indicative of its corresponding label, but this is not always the case [20], [21]. Recent research has demonstrated that attention mechanisms can enable machine learning models, in particular RNNs, to focus on salient sections of their input sequence. This approach has achieved overwhelming success in *neural language processing* (NLP), especially for *neural machine translation* (NMT) [22]–[24]. Moreover, the promise of attention in identifying emotionally salient speech segments, with the aim of improving SER performance, has been demonstrated [25].

Hierarchical attention networks (HANs) exploit more than one level of attention in a network with the aim of capturing hierarchical structures present in the data being modelled [26]. They have been shown to be superior to non-hierarchical networks in a range of tasks, e. g., in NLP, including document classification [26], document summarisation [27], sentiment analysis and sentiment detection [28]. The attention mech-

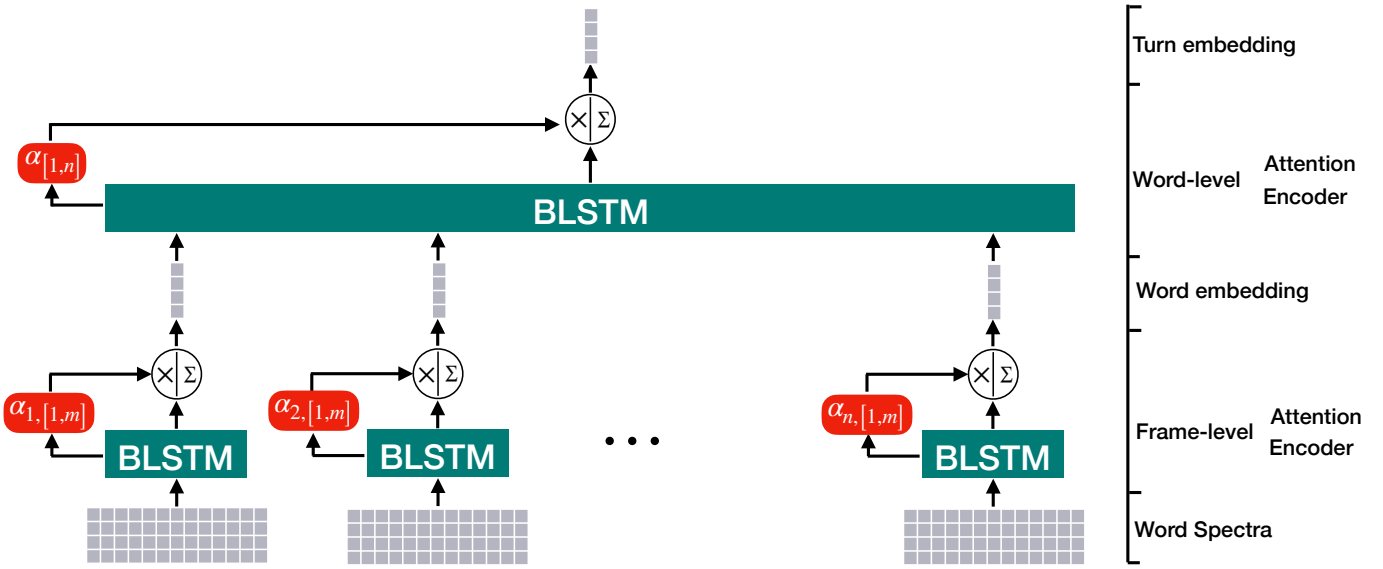


Fig. 1. An overview of the proposed two-level hierarchical attention network. The frame-level encoder processes input word spectra via a bidirectional long short-term memory (LSTM), followed by the frame-level attention layer to produce word embeddings. The word-level encoder processes the learnt word embeddings, which is followed by the word-level attention layer, which produces turn embeddings.

anisms used in these works identified the most informative sentences and the most important words in a given sentence with respect to the networks' learning objectives. However, the use of HANs for speech-related tasks, in discrete SER in particular, has not yet been fully explored.

Since a speaker turn or chunk consists of multiple words and each word corresponds to multiple speech frames, a two-level HAN architecture is expected to be more effective than a single-level attentive RNN [25]. Using this arrangement, the first level can learn *word embeddings* from the corresponding frames of *low-level descriptors* (LLDs). Based on the learnt word embeddings, the second level can code the speaker turn into *turn embeddings*.

Within a hierarchical approach, several attention mechanisms exist that can be exploited. Conventional attention mechanisms, such as standard soft-attention and local-attention, as used in sequence modelling, learn to assign attention weights across time [29]. They take no account of any informative distribution across the high-level features learnt by the RNN at each time-step.

To address this issue, we built a hierarchical component-attention based RNN for SER, referred to herein as *HiCAN*. To the best of our knowledge, this is the first time this approach has been applied to a discrete SER task. The rest of this paper is laid out as follows: Section II describes the structure of HiCAN, Section III describes the experiments used to test the approach and an accompanying evaluation, and Section IV details conclusions based on our results.

II. Proposed Methodology

Our hierarchical component-attention network consists of two encoding levels to produce word embedding and turn embedding. Each level comprises a bidirectional *long short-term memory* (LSTM) layer and a component-attention layer. We split each input turn into individual words as described in Section III-C. Using the spectrogram of a turn, we identify and isolate the frames belonging to each word within the turn; these frames are referred to herein as *word spectra*. The first level of our network then processes the word spectra as the input and produces word embeddings. The second level learns a turn embedding based on the learnt word embeddings as the context representation of the turn.

Component-attention mechanisms are applied for both levels, referred to as *frame-level attention* and *word-level attention*, to place attention weights across both the time and frequency dimensions of the spectra. The framework of the network is shown in Figure 1. In the following section, we provide an overview of the hierarchical attention network and derive the component-attention mechanism step by step.

A. Overview of HAN

Suppose a turn (T) consists of n words, and each word contains m frames. Let $T = (w_1, w_2, ...w_n)$, where $w_i (1 \leq i \leq n)$ denotes the ith word, and $w_i = (f_{i,1}, f_{i,2}, ...f_{i,m})$, where $f_{i,t} (1 \leq t \leq m)$ is a d-dimensional vector representing the tth frame in the ith word. The input word w_i is encoded by concatenating the forward LSTM and the backward LSTM outputs reads from $f_{i,1}$ to $f_{i,m}$:

$$\overrightarrow{h_{i,t}} = \overrightarrow{LSTM}(f_{i,t}, \overrightarrow{h_{i,t-1}}) \tag{1}$$

$$\overleftarrow{h_{i,t}} = \overleftarrow{LSTM}(f_{i,t}, \overleftarrow{h_{i,t-1}}) \tag{2}$$

$$h_{i,t} = \left[\overrightarrow{h_{i,t}}; \overleftarrow{h_{i,t}}\right]. \tag{3}$$

An attention mechanism takes the hidden state as input and yields the frame-level attention weights:

$$\alpha_{i,t} = Att\left(h_{i,t}\right). \tag{4}$$

We then obtain the word embedding as the attention weighted sum of hidden states:

$$w_i = \sum_t \alpha_{i,t} h_{i,t}. \tag{5}$$

Turn embeddings are obtained analogously given the word embeddings w_1 to w_n; the concatenation of the bidirectional LSTM is obtained by

$$\overrightarrow{h_i} = \overrightarrow{LSTM}(w_i, \overrightarrow{h_{i-1}}) \tag{6}$$

$$\overleftarrow{h_i} = \overleftarrow{LSTM}(w_i, \overleftarrow{h_{i-1}}) \tag{7}$$

$$h_i = \left[\overrightarrow{h_i}; \overleftarrow{h_i}\right], \tag{8}$$

with the general form of word-level attention weights

$$\alpha_i = Att\left(h_i\right). \tag{9}$$

The turn embedding s is a high level representation of the turn and is formulated as

$$s = \sum_i \alpha_i h_i. \tag{10}$$

Finally, the turn embedding is projected onto the discrete emotion categories through a fully-connected layer.

As the choice of attention mechanism can influence SER performance, we investigated standard soft-attention, local-attention, and component-attention, as described in the following sections.

B. Standard Soft-attention

As introduced in [23] for an RNN encoder-decoder sequence-to-sequence model, attention weights are determined by the compatibility between *key*, the encoder hidden states, and *query*, the previous hidden state of the decoder. As discrete SER is a sequence-to-one modelling task, the query is set to be an external trainable context vector e_q. The alignment score between the encoder hidden states and the query determine the attention weights. If we denote the complete output of a bidirectional LSTM as

$$H = \left(h_1, h_2, ...h_n\right). \tag{11}$$

The alignment score and attention weights can be calculated as

$$e_{[1,n]} = \tanh(WH + b), \tag{12}$$

$$\alpha_{[1,n]} = \frac{\exp(e_{[1,n]}^T e_q)}{\sum_s \exp(e_s^T e_q)}, \tag{13}$$

where we use the subscripts in $x_{[a,b]}$ *THERE IS NO x OR a IN THESE EQUATIONS?* to denote $[x_a, x_{a+1}, ...x_b]$. W and b are trainable parameters, and e_q is the trainable query vector with the same dimension as $e_i (1 \leq i \leq n)$.

Note, when e_q is not employed as a query as in [27], [30], but as a normal trainable vector, the above attention mechanism is also regarded as a kind of self-attention, because computation of the attention weights only depends on the encoder hidden states. This self-attention is different from the one defined in [22], which also requires no external context information to train attention weights, but conducts connections between two source positions directly.

C. Local-attention

Unlike standard soft-attention, which places attention on all hidden states to derive a context vector, local-attention only places attention on a small subset of the hidden states. Shrinking the attention scope enhances the attention to focus on local information [29]. For discrete SER, local attention has the potential to focus on particular regions of a speech signal that are more emotionally salient. Local attention is implemented using a sliding window, with the attention weight being inferred from the current target state and the source states within the window. Specifically,

$$e_{[t-\tau,t+\tau]} = \tanh(WH_{[t-\tau,t+\tau]} + b) \tag{14}$$

$$\tilde{\alpha}_{[t-\tau,t+\tau]} = \frac{\exp(e_{[t-\tau,t+\tau]}^T e_q)}{\sum_{t=-\tau}^{\tau} \exp(e_{[t-\tau,t+\tau]}^T e_q)}, \tag{15}$$

where the centre value $\tilde{\alpha_t}$ is the alignment score for time-step t and the other 2τ values are ignored.

After acquiring the alignment scores for all time-steps by shifting sliding window, softmax is applied again to keep the sum of attention values equal to 1,

$$\alpha_{[1,n]} = \frac{\exp(\tilde{\alpha}_{[1,n]})}{\sum_s \exp(\tilde{\alpha}_s)}. \tag{16}$$

Note that e_q in Equation (15) is different from that described for standard soft-attention in the sense of scope view. The local-attention e_q is trained to be the external query vector, which has the same scale as the local-attention scope.

D. Component-attention

Similar to [31], the component-attention model proposed here considers a small subsequence of hidden features rather than the entire sequence, as is the case for local-attention. In addition, instead of applying the same attention weight to all features extracted from one frame, we allocate attention weights to each spatial feature component of the frame. This strategy results in an attention vector being generated for each frame, rather than just a single attention value.

The weights of each component in the component-attention mechanism are computed for each frame as follows:

$$e_{[t-\tau,t+\tau],f} = \tanh(WH_{[t-\tau,t+\tau],f} + b), \tag{17}$$

where f denotes the fth feature component. We then normalise these weights across the time-step axis:

$$\tilde{\alpha}_{[t-\tau,t+\tau],f} = \frac{\exp(e_{[t-\tau,t+\tau],f}^T)}{\sum_{t=-\tau}^{\tau} \exp(e_{[t-\tau,t+\tau],f}^T)}, \tag{18}$$

where $\tilde{\alpha}_{t,f}$ is the centre alignment vector and the other alignment vectors are neglected. Finally, all the resulting alignment

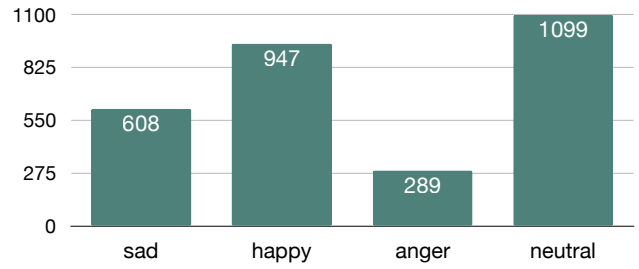


Fig. 2. Class distribution of the improvised speech instances in IEMOCAP

vectors are normalised across the time-step axis a second time as

$$\alpha_{[1,n]} = \frac{\exp(\tilde{\alpha}_{[1,n],f})}{\sum_s \exp(\tilde{\alpha}_{s,f})}. \tag{19}$$

Meanwhile, since features are individually treated in component-attention, the final output of the attention mechanism s can be computed similar to Equations (5) and (10):

$$s = \sum_t \alpha_{t,f} \times h_{t,f}. \tag{20}$$

The component-attention mechanism enables our system to learn individual attentions for all components in the LSTM hidden states. The component-attention is therefore more detailed than the attention learnt by the soft and local approaches, which only focus on the importance of a hidden state as a whole. This extra characteristic should enhance the accuracy of our HiCAN system as it has the versatility to automatically emphasise which components from which hidden states are more informative.

III. Experiments and Results

We conducted a series of experiments to evaluate the effectiveness of our proposed model. Its performance was compared with state-of-the-art baseline approaches in an SER task using the *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) database [32].

A. Data Description

IEMOCAP is a corpus comprising audio-visual data and accompanying transcriptions from recordings of paired actors in five dyadic sessions [32]. Emotional responses were elicited from the actors through the use of scripts and improvisation. In our experiments, we only used the improvised speech in order to reduce the potentially confounding effect of semantic information disturbance. In a limitation of the improvised speech dataset, however, the distribution of these instances across the five annotated emotion classes is heavily unbalanced. We therefore incorporated the excited turns into the happy class, as in [33], resulting in four emotion categories for training and evaluation: angry, happy, sad, and neutral (Figure 2).

B. Baselines

We used two baseline systems from the literature and two of our own systems to assess the performance of our model. The first of the four, described in [25], was a single-level (frame-level) local-attention network based on an RNN architecture, which the authors tested on the IEMOCAP dataset for SER. The second baseline was a hierarchical soft-attention network, described in [34]. For comparability, we took the results from their audio-only model. The third and fourth baselines were hierarchical attention networks with standard soft-attention (HiSAN) and local-attention (HiLAN), respectively.

C. Experimental Setup

We used several pre-processing steps to prepare the speech data for input to our model, such that all inputs had a uniform structure, consisted of the same number of words and each word comprised the same number of frames. Firstly, we divided the recordings of the speaker turns into individual words, according to the word boundaries provided by IEMOCAP. The boundaries were estimated using a forced-alignment technique. The log magnitude spectrum was then extracted from each of the word samples by applying a short-time Fourier transform (STFT), using a 25 ms Hanning window shifted by 10 ms. The sampling frequency of the audio files was 16 kHz, therefore each frame consisted of 400 samples, and its resulting feature vector comprised 201 frequencies. To make all words and sentences the same length, we zero-padded all sequences to the overall maximum length of the words from all speaker turns.

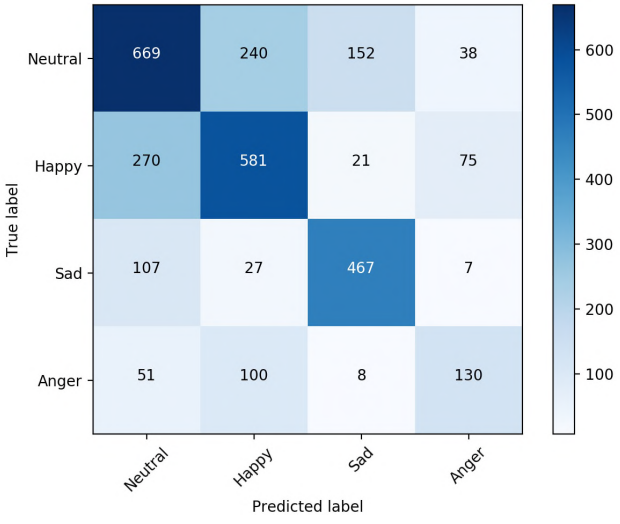
The two bidirectional LSTMs used in our model contained 600 hidden units, 300 for each direction. Additionally, we applied $l2$-normalisation when implementing LSTM to reduce the potential issue of over-fitting. All trainable parameters in the model were initialised by truncated normal initialisers. In the local-attention and component-attention models, the hyperparameter τ was tuned with three different settings, 3, 5 and 7, which led to attention scopes of 7, 11 and 15 ($2 \times \tau + 1$, including the centre frame and its left and right-side τ frames).

During the training phase, the batch size was set to 32, and the network parameters were optimised by minimising cross-entropy loss between the predicted labels and the ground-truth labels, using the Adam optimiser with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a fixed learning rate of 0.01.

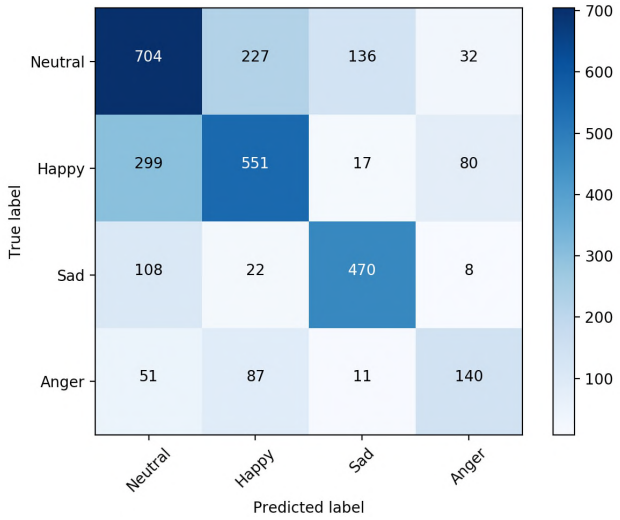
We adopted *leave-one-session-out cross-validation* (LOSOCV) to train and evaluate our model, which is the standard training and evaluation method used on IEMOCAP. For each round of cross-validation, four of the total five sessions in IEMOCAP were used for training. Using the remaining session, the turns from one speaker was served as development set for determining the network hyper-parameters, and the remaining turns were used as the test set.

D. Results and Discussion

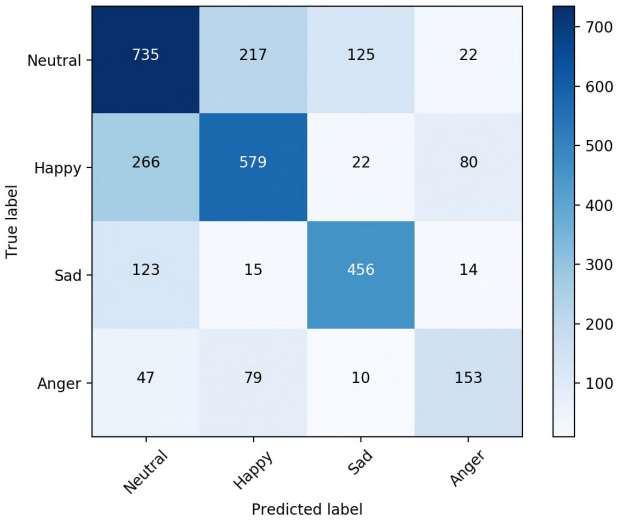
We used the *unweighted average recall* (UAR) as an evaluation metric in this work, as it is suitable for assessing



(a) HiSAN



(b) HiLAN



(c) HiCAN

Fig. 3. Confusion matrix results of hierarchical attention network with different attention strategy

TABLE I
COMPARISON OF UNWEIGHTED AVERAGE RECALL (UAR) AND WEIGHTED AVERAGE RECALL (WAR) EVALUATION METRICS. HISAN, HILAN, AND HICAN ARE HIERARCHICAL ATTENTION NETWORKS WITH STANDARD SOFT-ATTENTION, LOCAL-ATTENTION AND COMPONENT-ATTENTION, RESPECTIVELY.

Network	Attention Scope	UAR	WAR
1-level RNN+Attention [25]	global	61.8%	56.3%
Baseline HiSAN [34]	global	-	62.5%
HiSAN	global	62.0%	62.6%
HiLAN	7	61.6%	62.9%
	11	62.3%	64.4%
	15	62.5%	63.4%
HiCAN	7	**65.0%**	**66.1%**
	11	64.3%	65.3%
	15	64.5%	63.9%

performance in unweighted distributions. For a fair comparison with the HiSAN baseline, the *weighted average recall* (WAR) was also calculated.

The hierarchical attention networks largely outperformed the single-level attentive RNN, which achieved a 61.8% UAR and a 56.3% WAR (Table I). Comparing UAR values, the HiLAN with the two higher attention scope settings outperformed the HiSAN implemented in [34] and this work. Comparing WAR values, the best performing HiLAN, which had an attention scope of 11, achieved an improvement of 1.8 percentage points over HiSAN. HiCAN performed best, achieving optimal performance of a UAR of 65.0% and a WAR of 66.1% when its attention scope was reduced to 7. This demonstrated a significant improvement over the baseline systems ($p < 0.05$) in a one-tailed z-test.

The classification results of hierarchical attention networks with different attention strategies and their best scope settings are depicted in the confusion matrices in Fig. 3. Compared to classic soft attention, the hierarchical attention network with local attention enhances the recognition of 'neutral' emotion. However, with the increase of correct classification of 'neutral', more 'happy' turns were misclassified into 'neutral', leading to a worse classification rate of 'happy'. HiCAN achieved the best classification accuracy for 'neutral' and 'anger' emotions, and performed slightly worse than HiSAN for 'happy' and 'sad'. In addition, the 'sad' turns are the most accurately classified, achieving an accuracy higher than the average of the four emotion classes.

IV. CONCLUSIONS

We have proposed a hierarchical component-attention network for SER. The hierarchical architecture models the structure of speaker turns, and the frame- and word-level component-attention enable the network to focus on the salient time steps and features for discrete SER. Our proposed model considerably outperformed the existing baseline models in experiments. However, a potential limitation of component-attention is increased computations due to the generation of more trainable weights and hyperparameters. Moreover, as this approach requires the accurate identification of word

boundaries using forced-alignment, its performance might be limited by the associated automatic speech recognition (ASR) system. In our future work, we will explore the hierarchical component-attention network based turn embedding for continuous emotion recognition.

ACKNOWLEDGMENT

The work presented in this article was substantially supported by the National Natural Science Foundation of China (Grant No. 61702370), the Key Program of the Natural Science Foundation of Tianjin (Grant No. 18JCZDJC36300), the Open Projects Program of the National Laboratory of Pattern Recognition, and the Senior Visiting Scholar Program of Tianjin Normal University. This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE).

REFERENCES

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

[3] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[4] H. Gunes and B. W. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[5] E. Kramer, "Elimination of verbal cues in judgments of emotion from voice," *The Journal of Abnormal and Social Psychology*, vol. 68, no. 4, pp. 390–396, 1964.

[6] G. Fairbanks and W. Pronovost, "Vocal pitch during simulated emotion," *Science*, vol. 88, pp. 382–383, 1938.

[7] B. W. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, China, 2003, pp. 401–404.

[8] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, pp. 572–587, 2011.

[9] C. N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015.

[10] F. Eyben, K. R. Scherer, B. W. 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, 2016.

[11] M. Wöllmer, Z. Zhang, F. Weninger, B. W. Schuller, and G. Rigoll, "Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSIP)*, Vancouver, Canada, 2013, pp. 6822–6826.

[12] A. Graves, A.-R. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6645–6649.

[13] G. Keren, J. Han, and B. W. Schuller, "Scaling speech enhancement in unseen environments with noise embeddings," in *Proc. CHiME Workshop on Speech Processing in Everyday Environments*, Hyderabad, India, 2018, pp. 25–29.

[14] Y. Xu, J. Du, L. Dai, and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[15] Z. Ling, S. Kang, H. Zen, A. Senior, M. Schuster, X. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, 2015.

[16] L. Tian, J. D. Moore, and C. Lai, "Emotion recognition in spontaneous and acted dialogues," in *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*, Xian, China, 2015, pp. 698–704.

[17] J. Han, Z. Zhang, G. Keren, and B. W. Schuller, "Emotion recognition in speech with latent discriminative representations learning," *Acta Acustica united with Acustica*, vol. 104, pp. 737–740, 2018.

[18] G. Keren, F. Ringeval, E. Marchi, and B. W. Schuller, "End-to-end learning for dimensional emotion recognition from physiological signals," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, Hong Kong, 2017, pp. 985–990.

[19] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore, 2014, pp. 223–227.

[20] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition," in *Proc. Seventh International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, 2002, pp. 873–876.

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

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, 2017, pp. 5998–6008.

[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

[24] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, 2015, pp. 379–389.

[25] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 2227–2231.

[26] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, San Diego, CA, 2016, pp. 1480–1489.

[27] K. Al Sabahi, Z. Zuping, and M. Nadher, "A hierarchical structured self-attentive model for extractive document summarization (HSSAS)," *IEEE Access*, vol. 6, pp. 24 205–24 212, 2018.

[28] 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. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 6680–6684.

[29] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, 2015, pp. 1412–1421.

[30] Z. Lin, M. Feng, C. Dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *Proc. International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[31] A. Das, J. Li, R. Zhao, and Y. Gong, "Advancing connectionist temporal classification with attention modeling," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSIP)*, Calgary, Alberta, Canada, 2018, pp. 4769–4773.

[32] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. M. Provost, 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, 2008.

[33] R. Xia and Y. Liu, "DBN-ivector framework for acoustic emotion recognition," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, San Francisco, USA, 2016, pp. 480–484.

[34] 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. Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, 2018, pp. 2225–2235.