An Efficient Temporal Modeling Approach for Speech Emotion Recognition by Mapping Varied Duration Sentences into Fixed Number of Chunks

Wei-Cheng Lin and Carlos Busso

Multimodal Signal Processing (MSP) lab, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

wei-cheng.lin@utdallas.edu, busso@utdallas.edu

Abstract

Speech emotion recognition (SER) plays an important role in multiple fields such as healthcare, *human-computer interaction* (HCI), and security and defense. Emotional labels are often annotated at the sentence-level (i.e., one label per sentence), resulting in a sequence-to-one recognition problem. Traditionally, studies have relied on statistical descriptions, which are computed over time from *low level descriptors* (LLDs), creating a fixed dimension sentence-level feature representation regardless of the duration of the sentence. However sentence-level features lack temporal information, which limits the performance of SER systems. Recently, new deep learning architectures have been proposed to model temporal data. An important question is how to extract emotion-relevant features with temporal information. This study proposes a novel data processing approach that extracts a fixed number of small chunks over sentences of different durations by changing the overlap between these chunks. The approach is flexible, providing an ideal framework to combine gated network or attention mechanisms with *long short-term memory* (LSTM) networks. Our experimental results based on the MSP-Podcast dataset demonstrate that the proposed method not only significantly improves recognition accuracy over alternative temporal-based models relying on LSTM, but also leads to computational efficiency.

Index Terms: speech emotion recognition, attention mechanism, long short-term memory, chunk-level segmentation

1. Introduction

Detecting human emotion state via speech is helpful in multiple fields such as *human-computer interaction* (HCI) [1] or healthcare [2, 3]. Therefore, *speech emotion recognition* (SER) has become a popular research area. Most existing corpora are labeled at the sentence-level, where one global label is assigned per sentence [4–6]. With these labels, SER is often formulated as a sequence-to-one problem (i.e., mapping sequence of frames into a single label). Traditional methods deal with this problem by using *high level descriptors* (HLDs) such as mean, minimum and variance, estimated from *low level descriptors* (LLDs) extracted from speech (e.g., fundamental frequency, *Mel frequency cepstral coefficients*(MFCCs), energy). This approach generates a fixed dimension vector for a sentence, regardless of its duration, reformulating the problem as a static one-to-one machine learning problem. However, HLDs are not able to reflect the dynamic temporal information in the expression of emotion, leading to limited performance for SER systems. Therefore, recent studies have explored methods that can directly build SER systems from frame-based features, without relying on predefined functionals used in HLDs.

Deep learning approaches for SER systems have recently led to state of the art performance [7]. Different architectures exploring temporal information such as *recurrent neural network* (RNN), *convolution neural network* (CNN) or hybrid neural network (CNN-LSTM) have shown state-of-the-art performance by deriving features directly from LLDs or raw waveform [8–12]. These temporal models can be divided into two main categories for dealing with sentences with different lengths. The first category formulates the sequence-to-one task as a sequence-to-sequence task by relying on methods such as the *connectionist temporal classification* (CTC) loss [8, 9] and Markov chain-based approaches [13]. These approaches aim to create latent frame-by-frame variables for the emotional labels, which are then used to train the model. Although this method can effectively get rid of non-emotional frames, the emotional frames are labeled with the same class as the label assigned to the entire sentence. Moreover, a big issue when using RNN-based models in practical applications is the computational resources needed for sequences with long duration [14]. The second category uses deep learning models to extract sentence-level feature representations directly from the data, avoiding extracting predefined HLDs. After the feature representation is extracted, the problem is formulated as a one-to-one task [10, 11, 15]. These methodologies jointly learn feature extraction and build the SER models, resulting in better performance compared to traditional HLD representation. Nevertheless, they have some limitations such as requiring sentences with fixed length [11, 12], or using temporal-based pooling before the output layer [10, 15]. These approaches either truncate a sentence, append zeros or ignore primitive temporal information.

This paper proposes a novel and flexible data processing approach to model temporal acoustic information that addresses some of the key limitations of current temporal formulations in SER. The idea of the approach is to split sentences of different durations into a fixed number of small chunks by adjusting the overlap between chunks. This chunking procedure obtains a fixed number of data chunks, regardless of the duration of the sentence. It does not require dropping frames or appending zeros, preserving the complete temporal information of the original input sequence. The chunks are independently processed by a *long short-term memory* (LSTM) network with shared parameters. A key advantage of the approach is the flexibility offered by this formulation to combine the fixed number of feature representations created by the LSTMs for each of the chunks. This study proposes to combine these representations with a mean pooling layer (*NonAtten* network), a gated network (*GateVec* network) or an attention mechanism (*AttenVec* network). Finally, we obtain a sentence-level attention feature representation to generate emotion predictions via a fully connected output layer. Another key advantage of this approach is that it can significantly reduce the number of time steps by splitting the data sequence into small chunks, improving the computational efficiency of the architecture.

We evaluate the proposed models on the MSP-Podcast dataset [4], formulating the problem as a regression task to predict emotional attributes. Evaluating the models using *concor-*

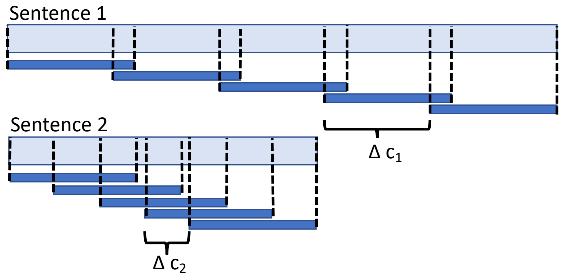

Figure 1: *Proposed chunk-based segmentation to split sentences of different durations into C chunks with fixed duration (w_c). We achieve this goal by adjusting the chunk step size (Δc_i).*

dance correlation coefficient (CCC), the experimental results demonstrate that the *AttenVec* method achieves the best balance between accuracy and computational efficiency. It achieves competitive performances for arousal ($CCC = 0.695$), dominance ($CCC = 0.613$) and valence ($CCC = 0.307$), which are better than the results obtained by a LSTM-based baseline. All the alternative methods implemented with our proposed chunk-level segmentation not only significantly improve the prediction accuracy, but also the computational efficiency.

2. Related Work

Various studies have utilized the concept of chunk or segment-level feature for SER. Han *et al.* [16] formed their segment-level feature by stacking the neighboring LLD frames to train a *deep neural network* (DNN). They train a DNN over the chunk, where the outputs of the model were used to estimate the probability of each emotion for that frame. This approach created curves with the probabilities for the emotions. Finally, they estimated statistics over these curves, which were used as the sentence-level feature representation of a static classifier implemented with *extreme learning machine* (ELM). Tzinis and Potamianos [17] employed HLD to represent segment-wise global features from LLDs, which obtains better performance compare to LLDs under a LSTM model. Tarantino *et al.* [18] built a self-attention model by setting different step size of input data chunks. They found that smaller step size (i.e., more overlap between chunks) can increase the discrimination of the feature representation in the network. These studies set the step size of the chunks as a fixed number. However, our proposed approach generates data chunks by varying the step size of the chunk as a function of the duration of the utterances. This key distinction leads our models to achieve clear improvements.

In our study, one of the proposed approaches for fusing the chunks is through attention models, which has been widely used in SER [19–22]. The most common way to apply this method is by building an attention model using frame-level features. Typically, the inputs of the attention model are activations of intermediate representation layers in the network. This approach produces attention weights per frame, which are used to obtain a final attention vector to recognize emotions [20–22]. In contrast to previous studies, our formulation constructs attention model at the chunk-level, reducing the computational cost since the number of chunks is fixed.

Another feature of our models is that we recognize arousal, valence and dominance with a single model using *multitask learning* (MTL). MTL allows a model to learn a common feature representation (i.e., a shared layer) by solving multiple related tasks, which increases the generalization of the models. MTL has been implemented in SER by considering several tasks, including multiple emotional attribute prediction [23,24], gender classification [25], and primary and secondary categorical emotion classification [26]. Since previous studies have consistently showed better performance by using multitask learning, we implement our model with MTL.

The main contributions of this paper is a novel chunk-based temporal modeling method that can map varied length data into fixed number of data chunks. This novel formulation is flexible, allowing different feature extraction methods and different combinations of the chunk-level representation. Our proposed chunk-based temporal modeling not only increases the accuracy of our predictions, but also reduces the complexity of model.

3. Proposed Methodology

3.1. Chunk-Based Segmentation

The core idea of our chunk-based segmentation method is to split a varied duration sentence into a fixed number of data chunks that have the same fixed duration. We achieve this goal by changing the step size (i.e., overlap between chunks). First, we need to set our desired length for the chunk window w_c. This variable should be big enough so the models can estimate reliable emotional information from the chunk, and small enough to be able to process short sentences. We estimate the maximum sentence duration $T_{\max} = \max\{T_1, T_2, \ldots, T_i, \ldots, T_N\}$, where T_i denotes the duration of sentence i. We use $T_{\max}$ to estimate a fixed number of chunks C, according to:

$$C = \left\lceil \frac{T_{\max}}{w_c} \right\rceil. \tag{1}$$

Since the overlap between chunks for longer sentences will be limited with this approach, we could increase the value of C by, for example, multiplying this value by an integer n (e.g., nC). The step size of the chunks Δc_i for sentence i is given by Equation 2. This equation shows that as we increase C (i.e., the number of chunks), Δc_i decreases, resulting in more overlap between chunks.

$$\Delta c_i = \frac{T_i - w_c}{C - 1} \tag{2}$$

Figure 1 visualizes the proposed approach for two sentences with different durations. The key difference between them is the chunk step size Δc_i (the overlapping area between chunks). By adjusting the chunk step size, this approach is able to split different duration sentences into a fixed number of chunks C that have the same duration w_c. Section 4.2 describes the actual implementation of the approach, including the values for w_c.

A key advantage of formulating SER problems with the proposed chunk-based segmentation is the simplification in modeling temporal information. Two important steps are (1) extracting feature representation from speech, and (2) combining chunk-based feature representations. The next two subsections describe these steps.

3.2. Extracting Feature Representation

Each chunk has a fixed length so extracting a feature representation is straightforward (w_c). As shown in Figure 2, This study restricts the analysis to LSTMs, although several alternative methods can be used (e.g., CNNs, estimation of HLDs per chunk). We extract the feature set proposed for the Interspeech 2013 computational paralinguistics challenge [27] using the OpenSmile toolkit [28]. The extracted LLDs include spectral, prosodic and energy-based acoustic features such as the fundamental frequency (f0), energy, and MFCCs. In total, the set includes 130 frame-based acoustic features, which are normalized by subtracting the mean and dividing by the standard deviation (these parameters are estimated over the training

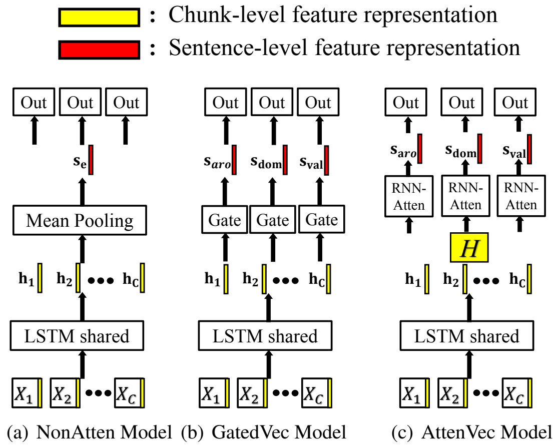

Figure 2: *Multitask frameworks using chunk-level segmentation. The chunk-level representation is combined with either mean pooling (NonAtten), gated mechanism (GatedVec) or attention model (AttenVec).*

set). Notice that we do not extract HLDs. The normalized input LLDs (X) are first split into data chunks $\{X_1, X_2, \ldots, X_C\}$ where $X_i \in \mathbb{R}^{m\times d}$. The dimension m is the number of frames per chunk and d is the dimension of the feature vector (i.e., $d = 130$). Then, we feed these data chunks into two consecutive LSTM shared layers with b hidden nodes. We exploit the final time step of the output as the representation vector for each chunk, denoted to as $\{\boldsymbol{h_1}, \boldsymbol{h_2}, \ldots, \boldsymbol{h_C}\}$ where $\boldsymbol{h_i} \in \mathbb{R}^{1\times b}$.

3.3. Combining Chunk-Based Feature Representations

Having a fixed number of chunks per sentence regardless of its duration simplifies the aggregation of temporal information across different chunks to form a sentence-level feature representation. Our formulation is flexible, where several approaches can be used. This study explores three alternative methods illustrated in Figure 2.

NonAtten Model – Fig. 2(a): After we obtained the chunk-level representations $\{\boldsymbol{h_1}, \boldsymbol{h_2}, \ldots, \boldsymbol{h_C}\}$, we directly average these vectors to obtain the sentence-level representation.

GatedVec Model – Fig. 2(b): The gated mechanism [29] allows the model to control the information flow from different channels. Equation 3 shows this operation, which consists of a sigmoid *neural network* (NN) layer (W_e, b_e) and a pointwise multiplication operation. By concatenating the gate model after the LSTM shared layer, we can produce the gating weights $g_{i,e}$ (scalar) for each emotional attribute e, where $e \in \{aro, dom, val\}$. This approach obtains the sentence-level representation vector $\boldsymbol{s_e}$ with equation 4.

$$g_{i,e} = \sigma(W_e \cdot \boldsymbol{h_i} + b_e) \tag{3}$$

$$\boldsymbol{s_e} = \sum_{i=1}^{C} g_{i,e}\boldsymbol{h_i} \tag{4}$$

AttenVec Model – Fig. 2(c): We first stack $\{\boldsymbol{h_1}, \boldsymbol{h_2}, ..., \boldsymbol{h_C}\}$ into a chunk-level feature map $H \in \mathbb{R}^{C\times b}$ and feed H into a vanilla RNN attention model. Then, we train the attention weights $\alpha_{t,e}$ for each emotional attribute e (i.e., different attention model for different emotional attributes) by using the *general* function from Luong *et al.* [30]. We use these attention scores to multiply the corresponding time step's hidden state $\{\overline{\boldsymbol{h}}_{\boldsymbol{1,e}}, \overline{\boldsymbol{h}}_{\boldsymbol{2,e}}, \ldots, \overline{\boldsymbol{h}}_{\boldsymbol{C,e}}\}$, where $\overline{\boldsymbol{h}}_{\boldsymbol{t,e}} \in \mathbb{R}^{1\times q}$. The dimension q is the number of nodes in the RNN attention model. This approach results in the context vector $\boldsymbol{c_e}$ (Eq. 5). Finally, we concatenate the vector $\boldsymbol{c_e}$ with the last hidden state $\overline{\boldsymbol{h}}_{\boldsymbol{C,e}}$, passing through a NN layer (W_e) with the $tanh$ activation function to obtain a sentence-level feature representation $\boldsymbol{s_e}$ (Eq. 6). Since the time steps in the RNN layer is fixed to C (i.e., attention on chunks rather than attention on all the input frames), our attention model is very computationally efficient.

$$\boldsymbol{c_e} = \sum_{t=1}^{C} \alpha_{t,e}\overline{\boldsymbol{h}}_{\boldsymbol{t,e}} \tag{5}$$

$$\boldsymbol{s_e} = tanh(W_e[\boldsymbol{c_e}; \overline{\boldsymbol{h}}_{\boldsymbol{C,e}}]) \tag{6}$$

For the three models in Figure 2, we feed the sentence-level feature representation $\boldsymbol{s_e}$ into their corresponding emotional attribute output layer, which is formed by two fully connected layers. The outputs of the multitask frameworks are the predictions for arousal, valence and dominance. We do not fine-tune the hyperparameters for the multitask model in this work. Our loss function L_{total} is just a direct summation of the different task losses: $L_{total} = L_{aro} + L_{dom} + L_{val}$.

Table 1: *Performance in CCC achieved by our proposed models, which are compared with LSTM-based models (i.e., LSTM(130) and LSTM(260)).*

Model	Aro [CCC]	Dom [CCC]	Val [CCC]
LSTM(130)	0.6520	0.5711	0.2031
LSTM(260)	0.6875	0.6045	0.2847
NonAtten	0.6781	0.6019	**0.2925**
GatedVec	0.6747	0.5944	**0.3199**
AttenVec	**0.6947**	**0.6132**	**0.3072**

Table 2: *The efficiency of the models in terms of number of parameters, Mega FLOPs, time cost for training and time cost for online processing.*

Model	# of Par. [10^6]	MFLOPs [MFLOPS]	Train [sec/epoch]	Online [ms/uttr]
LSTM(130)	0.323	5.67	437.1	547.5
LSTM(260)	1.052	18.49	439.4	598.1
NonAtten	0.323	**0.49**	**74.9**	**42.2**
GatedVec	0.324	**0.49**	246.6	**44.6**
AttenVec	0.577	**1.50**	353.1	**45.6**

4. Experimental Results

4.1. Resources

We utilize the version 1.6 of the MSP-Podcast corpus [4] to build and evaluate our proposed approach. The dataset consists of spontaneous, emotional-rich speech segments collected from various online audio-sharing websites. Following the ideas presented in Mariooryad *et al.* [31], we process the segments to identify clean audio, from a single speaker, without music in the background. The dataset provides both categorical and attribute-based emotional annotations, which are labeled by at least five annotators for each speech segment using a crowdsourcing approach [32]. We build our multitask model for arousal (calm versus active), valence (negative versus positive) and dominance (weak versus strong), where the ground truth label is the average of the scores across annotators. The version 1.6 of the corpus is split into train (34,280 speech turns), development (5,958 speech turns) and test (10,124 speech turns) partitions. The partitions are defined to reduce cases where data from one subject is included in more than one set. The readers are referred to Lotfian and Busso [4] for more details.

Table 3: *Different duration sets of emotion prediction CCC results for the testing set. The Short set contained sentences which duration less than 5 seconds, the Long set for duration greater than 8 seconds and remained for the Middle set.*

Short($\leq 5sec$)			
	Act-CCC	**Dom-CCC**	**Val-CCC**
LSTM(130)	0.6636	0.5812	0.2389
NonAtten	0.6761	0.6077	0.3129
GatedVec	0.6621	0.5865	0.3263
AttenVec	0.7003	0.6192	0.3363
Middle($5 \sim 8sec$)			
	Act-CCC	**Dom-CCC**	**Val-CCC**
LSTM(130)	0.6484	0.5642	0.1735
NonAtten	0.6779	0.6071	0.2839
GatedVec	0.6807	0.6042	0.3279
AttenVec	0.6880	0.6129	0.2978
Long($\geq 8sec$)			
	Act-CCC	**Dom-CCC**	**Val-CCC**
LSTM(130)	0.6314	0.5559	0.1737
NonAtten	0.6811	0.5822	0.2331
GatedVec	0.6933	0.6030	0.2835
AttenVec	0.6912	0.5989	0.2539

4.2. Experimental Settings

We implement our approach using chunks of 1 sec ($w_c = 1$). We have found that emotion can be estimated even with 0.5 secs speech segments [33], so 1 sec is a reasonable value. Since the duration of the sentences is between 2.75 and 11 secs, $T_{\max}$ is 11 secs. The number of chunks is 11 ($C = 11$) with these parameters (Eq. 1). The window analysis for the LLDs is 32ms with a step size of 16ms (50% overlap). Therefore, the number of frames within one chunk is $m = 62$. The value for the step size for the chunks (Δc_i) depends on the duration of the sentence (T_i) according to Equation 2. For example, if $T_i = 6$ secs, then $\Delta c_i = 0.5$ secs. For the network settings, we fixed the number of nodes in the layer matching the dimensions of the input (i.e., $d = b = q = 130$). We use dropout with $p = 0.5$ for the LSTM layers. We use batch normalization after the shared LSTM layers. We use Adam optimizer with a batch size of 128. The cost function optimizes the *concordance correlation coefficient* (CCC). We also report the accuracy of our prediction in terms of CCC. The models are implemented in Keras.

Our baseline model is directly trained with the entire input LLD sequence. We zero pad the feature vector until matching the maximum frame length of the dataset for batch training. We implement the LSTMs with either 130 nodes (*LSTM(130)*) or 260 nodes (*LSTM(260)*) per layer. The feature representation to estimate the output layers is the last frame of the LSTM layer.

4.3. Results and Analysis

Table 1 summarizes the recognition performance in terms of CCC. Table 2 shows the model efficiency in terms of the number of parameters, megaFLOPs, time cost for training (i.e., average in seconds per training epoch), and time cost for online processing (i.e., average in millisecond per utterance during inference). The key difference between the *NonAtten* and *LSTM(130)* models is the use of the chunk-based segmentation. Table 1 shows clear improvements in CCC for the *NonAtten* model compared to the *LSTM(130)* model. The results are particularly clear for valence improving the predictions from $CCC = 0.2031$ to $CCC = 0.2925$. The results demonstrate that the *LSTM(130)* model has poor capacity and low efficiency. It has the same number of parameters as the *NonAtten* model, but it requires 5.67 MFLOPs and considerable time cost to capture temporal cues for long sequences. This result shows the performance and efficiency advantages of our chunk-based segmentation.

We can further improve the recognition accuracy by adding gate (*GatedVec*) or attention (*AttenVec*) models instead of applying a mean pooling layer after the LSTM shared layer. The *GatedVec* model improves the prediction of valence with a minimal increase in the number of parameters (from 323K to 324K), and without sacrificing computation cost (0.49 megaFLOPs). This solution is suitable for memory-limited or efficiency-oriented systems. The *AttenVec* model achieves the best CCC performance for arousal ($CCC = 0.6947$) and dominance ($CCC = 0.6132$), and a very competitive performance for valence ($CCC = 0.3072$). Valence is an attribute that is particularly challenging to predict with acoustic features [34]. The model has a reasonable increase in the number of parameters (from 323K to 577K parameters) and computation efficiency (from 0.49 to 1.50 MFLOP). These results show that the *AttenVec* model provides the best tradeoff between complexity and accuracy. We also increase the model complexity of our baseline LSTM model by doubling the number of its nodes (i.e., *LSTM(260)*. Tables 1 and 2 show that this model improves the recognition accuracy, but at a very expensive cost: approximately 3 times the number of parameters and MFLOPs of the *LSTM(130)* model. This model also increases the time cost for training and online processing. Even with these extra costs, the *LSTM(260)* model is still less accurate than the *AttenVec* model.

We evaluate the performance of the systems as we increase the duration of the sentence, which will result in less overlap between the chunks. For this analysis, we split the testing set into short ($<$5sec), middle (5-8sec) and long($>$8sec) sentences. The test set has 4,280 short, 3,684 middle and 2,160 long sentences. Table 3 shows the results. The performance of the *LSTM(130)* model degrades as we increase the duration of the sentence for all three emotional attributes, showing poor accuracy for long sequences. In contrast, we can observe that the proposed chunk-based segmentation models systematically improve the performance for different duration of the data, especially for middle and long sequences. These results demonstrate that temporal modeling based on smaller chunks can be useful to aggregate long-term temporal information, leading to robust prediction accuracy, regardless of the duration of the sentences.

5. Conclusions

This study proposed a novel segmentation approach that splits a sentence into a fixed number of chunks, which have the same duration. By changing the step size between chunks, we can process sentences of different durations. The experimental evaluation showed the benefits in efficiency and accuracy by using the proposed chunk-level temporal modeling methodology. This simple concept, which can be easily implemented, offers the flexibility to explore different feature representation (fixed size of the chunk) and different fusions for chunk-based representation (fixed number of chunks). This solution also facilitates parallel processing for GPU. We expect that this approach can be also effective in other sequence-to-one problems, beyond the field of affective computing.

Our future directions are: to (i) validate the proposed chunk-level temporal modeling on multiple datasets and different sequence-to-one tasks (e.g. age detection), (ii) explore other solutions for feature extraction such as CNN and DNN, and (iii) analyze insights derived from chunk-level attention weights to understand better the non-uniform externalization of emotion.

6. Acknowledgements

Study supported by NSF (CNS-1823166; IIS-1453781).

7. References

[1] S. Rautaray and A. Agrawal, “Vision based hand gesture recognition for human computer interaction: a survey,” *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, January 2015.

[2] 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 Transactions on Affective Computing*, vol. 7, no. 1, pp. 59–73, January-March 2015.

[3] J. Gideon, E. Mower Provost, and M. McInnis, “Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 2359–2363.

[4] R. Lotfian and C. Busso, “Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings,” *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[5] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[6] M. Grimm, K. Kroschel, and S. Narayanan, “The Vera am Mittag German audio-visual emotional speech database,” in *IEEE International Conference on Multimedia and Expo (ICME 2008)*, Hannover, Germany, June 2008, pp. 865–868.

[7] M. Abdelwahab and C. Busso, “Study of dense network approaches for speech emotion recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5084–5088.

[8] V. Chernykh, G. Sterling, and P. Prihodko, “Emotion recognition from speech with recurrent neural networks,” *ArXiv e-prints (arXiv:1701.08071)*, pp. 1–16, January 2017.

[9] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, “Attention-enhanced connectionist temporal classification for discrete speech emotion recognition,” in *Interspeech 2019*, Graz, Austria, September 2019, pp. 206–210.

[10] Z. Aldeneh and E. Mower Provost, “Using regional saliency for speech emotion recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 2741–2745.

[11] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5200–5204.

[12] W. Q. Zheng, J. S. Yu, and Y. X. Zou, “An experimental study of speech emotion recognition based on deep convolutional neural networks,” in *International Conference on Affective Computing and Intelligent Interaction (ACII 2015)*, Xi’an, China, September 2015, pp. 827–831.

[13] J. Lee and I. Tashev, “High-level feature representation using recurrent neural network for speech emotion recognition,” in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 1537–1540.

[14] S. Bai, J. Z. Kolter, and V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” *ArXiv e-prints (arXiv:1803.01271)*, pp. 1–10, March 2018.

[15] Z. Wang and I. Tashev, “Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5150–5154.

[16] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in *Interspeech 2014*, Singapore, September 2014, pp. 223–227.

[17] E. Tzinis and A. Potamianos, “Segment-based speech emotion recognition using recurrent neural networks,” in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 190–195.

[18] L. Tarantino, P. Garner, and A. Lazaridis, “Self-attention for speech emotion recognition,” in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2578–2582.

[19] 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, October 2018.

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

[21] C.-W. Huang and S. Narayanan, “Attention assisted discovery of sub-utterance structure in speech emotion recognition,” in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 1387–1391.

[22] P.-W. Hsiao and C.-P. Chen, “Effective attention mechanism in dynamic models for speech emotion recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, Canada, April 2018, pp. 2526–2530.

[23] S. Parthasarathy and C. Busso, “Jointly predicting arousal, valence and dominance with multi-task learning,” in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1103–1107.

[24] ——, “Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes,” in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3698–3702.

[25] B. Zhang, E. Mower Provost, and G. Essi, “Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5805–5809.

[26] R. Lotfian and C. Busso, “Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning,” in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 951–955.

[27] B. Schuller and et al., “The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism,” in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[28] F. Eyben, M. Wöllmer, and B. Schuller, “OpenSMILE: the Munich versatile and fast open-source audio feature extractor,” in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[29] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, November 1997.

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

[31] S. Mariooryad, R. Lotfian, and C. Busso, “Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora,” in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[32] A. Burmania, S. Parthasarathy, and C. Busso, “Increasing the reliability of crowdsourcing evaluations using online quality assessment,” *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[33] J. Arias, C. Busso, and N. Yoma, “Energy and F0 contour modeling with functional data analysis for emotional speech detection,” in *Interspeech 2013*, Lyon, France, August 2013, pp. 2871–2875.

[34] K. Sridhar, S. Parthasarathy, and C. Busso, “Role of regularization in the prediction of valence from speech,” in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 941–945.