A Dual-Complementary Acoustic Embedding Network Learned from Raw Waveform for Speech Emotion Recognition

Tzu-Yun Huang
Electrical Engineering
National Tsing Hua University
Taiwan
twchris.0213@gmail.com

Jeng-Lin Li
Electrical Engineering
National Tsing Hua University
Taiwan
cllee@gapp.nthu.edu.tw

Chun-Min Chang
Electrical Engineering
National Tsing Hua University
Taiwan
cmchang@gapp.nthu.edu.tw

Chi-Chun Lee
Electrical Engineering
National Tsing Hua University
Taiwan
cclee@ee.nthu.edu.tw

***Abstract*—Speech emotion recognition (SER) technology has recently become a trend in a broader field and has achieved remarkable recognition performances using deep learning technique. However, the recognition performances obtained using end-to-end learning directly from raw audio waveform still hardly exceed those based on hand-crafted acoustic descriptors. Instead of solely rely on raw waveform or acoustic descriptors for SER, we propose an acoustic space augmentation network, termed as Dual-Complementary Acoustic Embedding Network (DCaEN), that combines knowledge-based features with raw waveform embedding learned with a novel complementary constraint. DCaEN includes representations from eGeMAPS acoustic feature and raw waveform by specifying a negative cosine distance loss to explicitly constrain the raw waveform embedding to be different from eGeMAPS. Our experimental results demonstrate an improved emotion discriminative power on the IEMOCAP database, which achieves 59.31% in a four class emotion recognition. Our analysis also demonstrates that the learned raw waveform embedding of DCaEN converges close to reverse mirroring of the original eGeMAPS space.**

***Index Terms*—speech emotion recognition, raw waveform, end-to-end learning, acoustic space augmentation**

I. Introduction

Speech emotion recognition (SER) is a key technology in driving next-generation decision support solutions and improving human-machine interface designs beyond current capabilities [1], [2]. Utilization of SER has already been shown in applications such as robotics [3], [4], marketing [5] and medical care [6]. The increasing proliferation of research in deploying SER technologies across domains inevitably require further algorithmic development in improving its modeling power especially given the current surge and successful use of end-to-end deep learning architecture.

Traditionally, extracting hand-crafted acoustic features, i.e., computing a variety of knowledge-inspired low-level acoustic descriptors (LLDs), provides the most effective way to achieve high performing speech emotion recognition accuracy [7]–[9]. Recently, with deep networks, researches have demonstrated that promising recognition accuracy could be achieved by modeling the speech spectrograms without extracting hand-crafted LLDs [10], [11]. Even further, end-to-end learning architecture provides an opportunity to handle raw time-domain acoustic data directly, e.g., Sarma et al. proposed a sophisticated procedure in dealing with 1-D raw waveform, including components of data perturbation, convolution layer designed using Network-in-Network (NIN) and TDNN-LSTM-attention temporal model [12].

However, performing SER directly from time-domain audio data remains challenging. The accuracies obtained in most cases do not surpass using spectrogram or acoustics LLDs. Several research works have instead enhanced the discriminative capacity of the recognition framework by augmenting the acoustic feature space using representations learned from both raw waveform and spectrogram/LLDs. For example, Yang et al. utilized convolutional neural network (CNN) on raw waveform and spectrogram separately and leverage their complementary information by integrating both with a bidirectional long short-term memory neural network (BLSTM) [13]; Lakomkin et al. proposed a progressive network augmenting pre-trained spectrogram representation with mel-frequency cepstral coefficients (MFCCs) and pitch to benefit the robustness of cross-domain SER tasks [14]. These augmentation methods do not, however, explicitly learn to leverage the complementary information exists between different types of acoustic inputs while learning their associated representations. In this work, we propose to learn from the raw waveform directly to derive representations *beyond* knowledge-based features as a method in augmenting the feature space used to improve SER.

Specifically, we propose a Dual-Complementary Acoustic Embedding Network (DCaEN) to derive an augmented acoustic feature space, i.e., concatenation of hand-crafted knowledge-based feature network (eGeMAPS [15]) with complementary representation learned from the raw waveform. The waveform representation is learned using a novel negative cosine similarity loss to explicitly extract embedding that captures information *beyond* eGeMAPS. We evaluate our framework on a large scale emotion corpus, the USC IEMOCAP database [16]. Our proposed DCaEN achieves 59.31% in a 4-class emotion classification task, which improves 6.49% and

978-1-7281-3888-6/19/$31.00

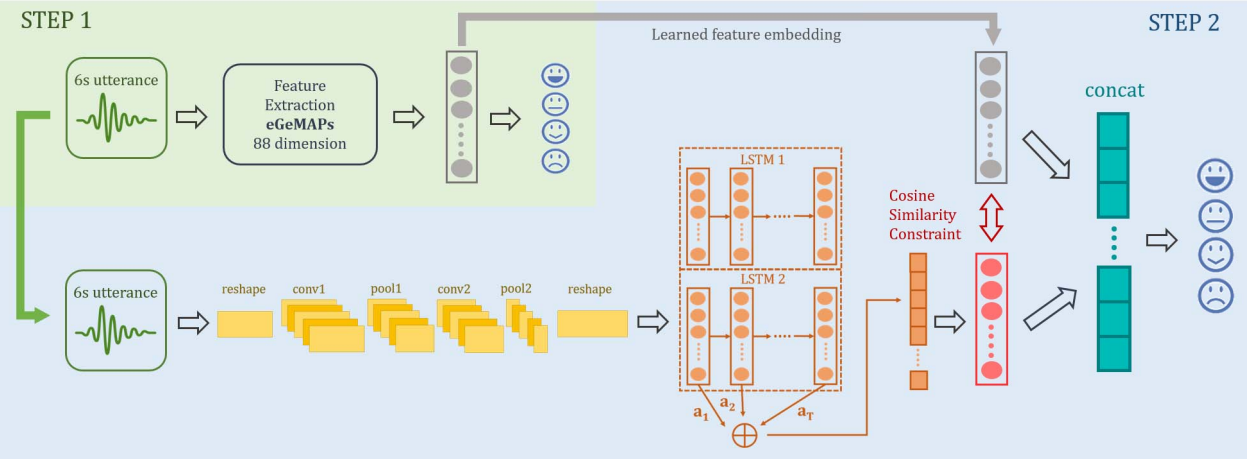


Fig. 1. Illustration of our framework: Dual Complementary Acoustic Embedding Model. The model is divided into two stages: first, embeddings learned from *Feature Network* with eGeMAPS is the input; second, we use an end-to-end architecture to learn complementary embedding from raw waveform with cosine similarity constraint. Finally, these representations are concatenated to perform final SER.

1.95% relative compared to using raw waveform or eGeMAPS feature set solely. We further provide an analysis to visualize how the learned raw waveform embedding changes as a function on the levels of negative cosine similarity constraint placed on the architecture and its corresponding recognition accuracy. The rest of the paper is organized as follows: Section 2 introduce the emotion database, features and our proposed DCaEN model. In Section 3, we describe the experimental setup, the recognition and visualization results. Finally, in Section 4 we conclude with future works.

II. RESEARCH METHODOLOGY

A. Emotion Database

We use the USC Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [16] to evaluate our proposed framework. The database consists of 5 sessions acted by 10 different actors including 5 males and 5 females. The database has approximately 12 hours of data segmented manually into utterances. Each utterance is annotated by at least three annotators on 10 categorical labels.

In our experiment, we follow the exact same experimental setting used in a previous work [17], i.s., using 4 emotion classes as the targeted labels: sadness, happiness (include excitement), anger and neutral with a total of 5531 utterances. There are 1103 utterances for angry, 1636 for happy (includes excitement), 1084 for sad, and 1708 for neutral.

B. Acoustic Feature Sets

Our proposed Dual Complementary Acoustic Embedding Network (DCaEN) uses two different acoustic inputs to derive the augmented acoustic representation, i.e., hand-crafted acoustic features and raw waveform. In terms of the hand-crafted features, we use the eGeMAPS feature set [15]. We also perform preprocessing on the raw waveform. Both will be described in the following.

1) Hand-Crafted Features: eGeMAPS: We compute the functional extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [15], which is an effective acoustic parameter set used in multiple prior speech emotion recognition works (e.g., [7]–[9]). The eGeMAPS includes 88 dimensions of acoustic features per utterance resulting from implementing the arithmetic mean, the coefficient of variation, 20th, 50th, 80th percentile on pitch and loudness LLDs, the mean and the standard deviation of the slope of rising/falling signal parts on the energy, spectral, and frequency. These LLDs are extracted using 20ms frame size with 10ms step using the openSMILE toolkit [18].

2) Raw Waveform Preprocessing: We define a fixed segment length for each utterance used in derving embedding from raw waveform. The length is set as 6 seconds. If an utterance is longer than 6 seconds, we take the first 6 seconds worth of data; otherwise, we pad zeros for those utterances that are shorter than 6 seconds. At 16 kHz sampling rate, each utterance corresponds to a 96000-dimensional input vector. We further divide this input vector into 150 sequential units with a frame size of 640.

C. Dual Complementary Acoustic Embedding Network

Our proposed model, DCaEN, is illustrated in Fig. 1. It includes two sub-structures: a *Feature Network* operates on eGeMAPS and a *Raw Waveform Complementary Network* learns complementary embedding directly from raw waveform. The training process consists of two stages. Firstly, we train the *Feature Network* using hand-crafted feature set with a neural network architecture. Secondly, we train the *Raw Waveform Complementary Network* using a novel constraint of negative cosine similarity between the output of fully-connected layer in the *Feature Network* and the output of the fully-connected layer after the LSTM layer in the *Raw Waveform Network*. Lastly, the two learned representations are concatenated for final four-class emotion classifications. The details are de-

scribed in the following sections and the model parameters are illustrated in III-A.

1) Feature Network (Stage 1): In the first stage, we train an emotion recognizer on the hand-crafted feature set with fully connected layers followed by a softmax layer using categorical cross-entropy loss:

$$L_{ce}(p, q) = -\sum_{x} p(x) \log q(x) \tag{1}$$

where $p(x)$ is the target label distribution and $q(x)$ corresponds to the predictions.The embedding layer from the feature network learning can be seen as capturing relevant emotion information in these hand-crafted features that characterize the known psychoacoustics knowledge from feature engineering experts. The embedding is then frozen (not updated again) to be used as a basis for further complementary feature learning from raw waveform (section II-C2).

2) Raw Waveform Complementary Network (Stage 2): With the extracted embedding from the learned *Feature Network*, we use an end-to-end deep learning architecture to model the raw waveform for emotion recognition task [19]. It consists of two stacked convolution layers and two stacked (long short-term memory) LSTM layers followed by a fully-connected layer and softmax dense layer for prediction. We additionally employ attention mechanism in each LSTM layer for flexible time dependency modeling. In order to learn the complementary representation, we constrain the output of the fully connected layer in the *Raw Waveform Complementary Network* and the corresponding output of the *Feature Network* using a negative cosine similarity (Equation 2), i.e., we expect the cosine similarity to be as small (as close to -1) as possible when learning the complementary embedding from raw wavform. The cosine complementary loss is shown below, where x_1 and x_2 are different representations.

$$L_{cos}(x_1, x_2) = \frac{x_1 \cdot x_2}{\|x_1\| \, \|x_2\|} \tag{2}$$

The representation derived from the *Raw Waveform Complementary Network* can be seen as emotionally-relevant embedding though resides in an opposite space as the embedding space learned from the *Feature Network* operated on the hand-crafted feature set.

Finally, to leverage the discriminative power of both hand-crafted features network embedding and the complementary representation from raw waveform, we concatenate these two representations and optimize the network with a joint loss (Equation 3),

$$L = wL_{cos} - (1 - w)L_{ce} \tag{3}$$

where w is a weighting term. The joint loss is the weighted sum of the categorical cross-entropy loss of emotion categories and the cosine complementary loss. The model jointly optimizes these two losses to obtain an augmented acoustic embedding to perform the SER task on IEMOCAP.

III. EXPERIMENTAL SETUP AND RESULT

In section III-A, we will describe different comparison models used in our recognition tasks and the network configuration along with parameter choices. In section III-B and section III-C, we will report the accuracy comparison of the recognition experiments and also the visualization analysis on the learned complementary embedding.

A. Experiment Setting

In our experiment, we use leave-one-session-out 5-fold cross validation scheme where 4 sessions are used for training and 1 complete session for test evaluation. In all the results, we use unweighted average recall (UAR) as the metric for evaluation.

1) Comparison Models: We compare with the following models and report results in section III-B:

- **Raw**: *Raw Waveform Complimentary Network* without using the cosine constraint
- **Ftr_eg**: *Feature Network* with eGeMAPS features introduced in section II-B
- **Ftr_em**: *Feature Network* with emobase2010 features [20]
- **E_nC**: Dual network architecture with the same structure as our proposed DCaEN but learning without the cosine constraint
- **E_Ceg**: Dual complementary acoustic embedding network but with cosine similarity constrain placed directly on the eGeMAPS without learning through network embedding from the *Feature Network*
- **E_uF**: Dual complementary acoustic embedding network with the *Feature Network* weights updated together when training *Raw Waveform Complementary Network* in stage 2 (i.e., non-frozen hand-crafted feature embedding)
- **E_C0**: Dual complementary acoustic embedding network with cosine similarity constraint set close to 0
- **R_C**: Raw waveform network with cosine similarity constraint but without concatenate the representation learned from the hand-crafted feature set.

Specifically, we compare eGeMAPS feature set introduced in section II-B1 with the emobase2010 [20], which contains 1582 dimensions computed based on 34 low-level descriptors in constructing the input feature set for *Feature Network*. We also compare the effectiveness of different levels of consine constraints to further validate the complementary learning strategy in our proposed network for emotion recognition.

2) Network Configuration: In all of our network optimization, we use the Adam optimizer with learning rate 10^{-3} and train 20 epochs with mini-batch size 32. Each dimension of the hand-crafted feature is z-normalized with respect to individual speaker. The *Feature Network* uses a 64 nodes fully-connected layer with dropout probability 0.5.

In the *Raw Waveform Complementary Network*, each frame is convolved with 40 filters of kernel size 20 to extract features from high sampling rate signal, and then downsample to 8 kHz by pooling each filter output with a pool size 2. To extract long-term characteristics of the speech and roughness of the

TABLE I
COMPARED MODELS ON EMOTION RECOGNITION RESULTS. DCAEN IS OUR PROPOSED DUAL COMPLEMENTARY ACOUSTIC EMBEDDING NETWORK. THE BEST UAR OBTAINED IS IN BOLD: 59.31%

Models	Raw	Ftr_eg	Ftr_em	E_nC	E_Ceg	E_uF	E_C0	R_C	DCaEN
Sad	72.05%	64.11%	66.70%	66.14%	71.22%	68.63%	64.21%	71.22%	68.45%
Happy	27.81%	52.87%	51.89%	49.69%	54.16%	49.39%	49.82%	25.31%	50.12%
Angry	55.85%	54.67%	59.11%	54.67%	47.96%	54.49%	54.49%	57.75%	58.30%
Neutral	55.56%	57.79%	52.93%	57.44%	52.93%	56.03%	59.66%	60.25%	60.36%
UAR	52.82%	57.36%	57.66%	56.99%	56.57%	57.14%	57.04%	53.63%	**59.31%**

speech signal, we convolve the pooled frame using kernel size 40 in each filter followed by max-pooling layer across the channel domain with a pool size of 10. Then, we use 2 stacked LSTM each with 256 cells with attention to learn the sequential dependency. Afterwards, a fully connected layer with 64 nodes to derive the final embedding.

The two concatenated embedding layers from the *Feature Network* and the *Raw Waveform Complementary Network* lead to a representation with a total of 128 feature dimension for each utterance sample. The weight of loss function (equation 3) is specified as $w = 0.6$ using grid-search.

TABLE II
THE RESULTS OF DIFFERENT THRESHOLD ON COSINE SIMILARITY. WITH NUMBER OF THRESHOLD CLOSE TO -1 WHICH MEANS NEGATIVE CORRELATION, THE UAR IS HIGHER.

Threshold	UAR
-0.5	57.49%
-0.6	57.82%
-0.7	58.21%
-0.8	58.57%
Learnable	59.31%

B. Recognition Results

All of our comparison results are reported in Table I. For the baseline models, *Raw* (52.82%) obtains lower accuracy than *Ftr_eg* (57.36%) reinforcing past knowlege that the knowledge-based acoustic features possess better discriminative power and the challenges in modeling raw waveform for speech emotion recognition. When examining *Ftr_em*, which includes a much higher dimension, the performance obtained is not significantly better than using just 88 dimenional eGeMAPS. The emobase2010 feature set likely encompasses non-emotionally-meaningful redundancy as compared to eGeMAPS for the IEMOCAP dataset.

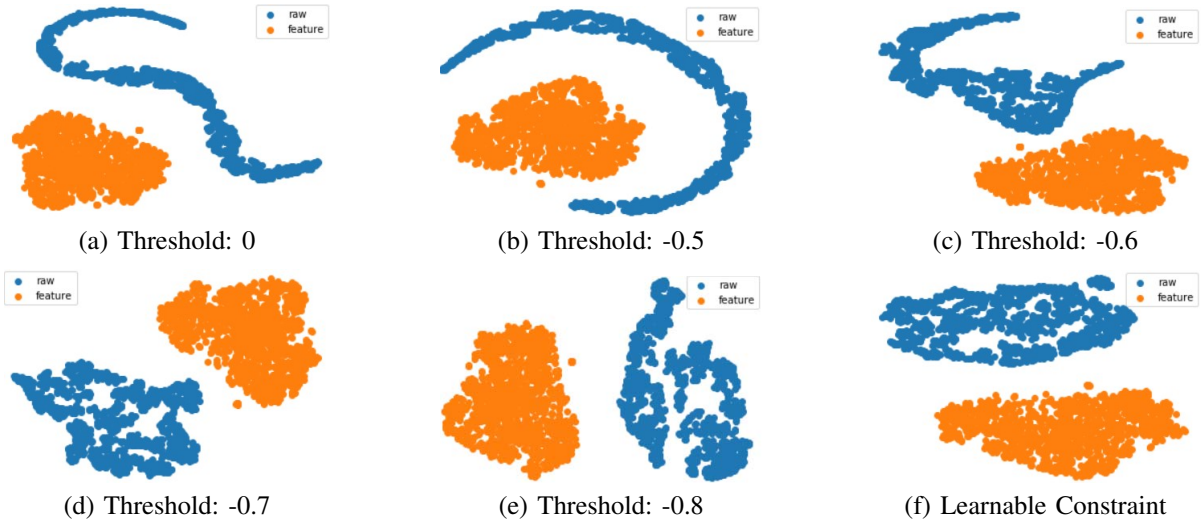
Our proposed DCaEN model significantly outperforms both *Raw* and *Ftr_eg* as shown in Table I (6.49% and 1.95% relative improvement respectively). We compare it with the *E_nC* to examine the importance of using our proposed use of cosine dissimilarity constraint. *E_nC* can be imagined as a naive concatenation of both hand-crafted feature embedding and raw waveform embedding. The accuracy obtained using *E_nC* shows barely no improvement over *Ftr_eg*, likely due to the fact without using a proper constraint, these two representations (hand-crafted feature set and raw time waveform) could be capturing redundant acoustic characteristics. Furthermore, our method outperforms accuracy obtained using the *E_Ceg*. *E_Ceg* places the constraint directly on the original eGeMAPS 88 dimensions without first learning an emotional discriminatory embedding with the *Feature Network*, which further degrades the quality of the complementary learning in the *Raw Waveform Complementary Network*.

Since our DCaEN tends to minimize the cosine similarity, i.e., making the distance closer to -1, we further compare to the results obtained with orthogonal cosine constraint, i.e., distance lower bounded by 0 (*E_C0*). *E_C0* does not improve performances in this recognition tasks further demonstrating that while mining additional information from raw waveform, making the embeddings *dissimilar* to the hand-crafted feature embedding is not sufficient. There needs to be a stronger constraint pushing the learned embedding space to the opposite direction. The simple dissimilar criterion may lead to a convergence point that carries emotionally-irrelevant information. Furthermore, the sequential training strategy is important in our proposed network architecture, i.e., the learned knowledge-derived feature embedding from the stage 1 *Feature Network* needs to be frozen when feeding into the stage 2 *Raw Waveform Complementary Network*. This effect is evident when comparing the results obtained using *E_uF* and our DCaEN.

Additionally, *R_C* uses embedding only from the *Raw Waveform Complementary Network*. This model outperforms *Raw* slightly indicating that the intregration of the joint cosine similarity loss from expert-knowledge features is beneficial in searching for a better hidden dimensions directly from the raw waveform. Lastly, our framework improves upon previous algorithm [14] on using two different acoustic input on the same set of IEMOCAP database, which reports a UAR of 58% as compared to our method of 59.13%.

C. Analysis on Levels of Complementary Constraint

In this part, we provide an analysis on the learned embedding from the *Raw Waveform Complementary Network* by visualizing it on a 2D space with different levels of cosine distance constraint. In order to understand the accuracy obtained as a function on the levels of the constraint, we train



(a) Threshold: 0 (b) Threshold: -0.5 (c) Threshold: -0.6

(d) Threshold: -0.7 (e) Threshold: -0.8 (f) Learnable Constraint

Fig. 2. Visualization of the raw waveform embedding and eGeMAPS feature embedding on 2-D space with different constraint thresholds. As the thresholds tends toward -1, the two representations become similiar though in a reverse direction.

our proposed *DCaEN* with a modified threshold loss function,

$$L_{thres} = w\,|L_{cos} - threshold| - (1-w)L_{ce} \qquad (4)$$

which can limit the lower-bound values of cosine distance loss to the specified threshold. Table II lists the UARs obtained as different values of threshold placed on the constraint. A quick observation can be seen is that the UARs improves as the threshold tending toward more negative values. Meanwhile, we use t-distributed Stochastic Neighbor Embedding (t-SNE) to project the embedding onto a 2D space for visualization on both the learned embedding using raw waveform and the *Feature Network* embeddings from the eGeMAPS features.

From Fig 2, we can observe that the raw waveform embedding and hand-crafted feature embedding form a separated and yet opposite pattern as the threshold gets closer to -1. Specifically, if we compare between threshold set to 0 and closer to -1, the learned raw representation at threshold 0 shows a weary shape, and as the threshold becomes more negative, the representation learned from raw waveform seems to converge to a similar shape as the hand-crafted feature embedding just at a 180 degree mirroring reverse. The improvement in the recognition could potentially be attributed to the fact that each of these spaces includes relevant emotional discriminatory information but reside in a different subspace, hence, by concatenating the two representations, it provides an augmented modeling power.

IV. CONCLUSIONS

We propose a Dual Complementary Acoustic Embedding Network (DCaEN) to perform speech emotion recognition. DCaEN includes two sub-structures: *Feature Network* that uses expert knowledge-driven acoustic parameters and *Raw Waveform Complementary Network* that uses raw waveform directly to learn an acoustic embedding. We propose to use an explicit cosine distance constraint with a sequential optimization strategy for our DCaEN. This method effectively learns emotionally-relevant information *beyond* conventional acoustic parameters directly from the raw waveform, and by concatenating both representations, we demonstrate that our framework can improve 4-class emotion recognition rates to 59.13% on the IEMOCAP dataset. For future works, we will immediately apply the same architecture on other large scale emotion corpus to validate the robustness of our DCaEN, and further, aside from constraining on eGeMAPS, we can explore other domain expert knowledge as auxiliary information in better achieving high-performing end-to-end speech emotion recognition.

REFERENCES

[1] Björn Schuller, Gerhard Rigoll, and Manfred Lang, "Hidden markov model-based speech emotion recognition," in *Multimedia and Expo*, 2003, vol. 1, pp. I–401.

[2] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[3] Jeong-Sik Park, Ji-Hwan Kim, and Yung-Hwan Oh, "Feature vector classification based speech emotion recognition for service robots," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 3, 2009.

[4] Cynthia Breazeal and Lijin Aryananda, "Recognition of affective communicative intent in robot-directed speech," *Autonomous robots*, vol. 12, no. 1, pp. 83–104, 2002.

[5] Kristin Byron, Sophia Terranova, and Stephen Nowicki, "Nonverbal emotion recognition and salespersons: Linking ability to perceived and actual success," *Journal of Applied Social Psychology*, vol. 37, no. 11, pp. 2600–2619, 2007.

[6] Alex Pentland, "Healthwear: medical technology becomes wearable," *Computer*, 2004.

[7] Michael Neumann and Ngoc Thang Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," arXiv preprint, 2017, arXiv:1706.00612.

[8] Jing Han, Zixing Zhang, Gil Keren, and Bjrn Schuller, "Emotion recognition in speech with latent discriminative representations learning," in *Acta Acustica united with Acustica*, 2018, vol. 104, pp. 737–740.

[9] Zakaria Aldeneh and Emily Mower Provost, "Using regional saliency for speech emotion recognition," in *ICASSP*, 2017, pp. 2741–2745.

[10] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *ICASSP*, 2017, pp. 2227–2231.

[11] WQ Zheng, JS Yu, and YX Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," in *2015 international conference on affective computing and intelligent interaction (ACII)*. IEEE, 2015, pp. 827–831.

[12] Mousmita Sarma, Pegah Ghahremani, Daniel Povey, and et al., "Emotion identification from raw speech signals using dnns," in *Interspeech 2018 - 19^{th} Annual Conference of the International Speech Communication Association*, 2018.

[13] Zixiaofan Yang and Julia Hirschberg, "Predicting arousal and valence from waveforms and spectrograms using deep neural networks," in *Proc. Interspeech 2018*, 2018, pp. 3092–3096.

[14] Egor Lakomkin, Cornelius Weber, Sven Magg, and Stefan Wermter, "Reusing neural speech representations for auditory emotion recognition," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, 2017, vol. 1.

[15] Florian Eyben, Klaus R Scherer, Böjrn W Schuller, and et al., "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.

[16] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, and et al., "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[17] Haytham M. Fayek, Margaret Lech, and Lawrence Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[18] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[19] Panagiotis Tzirakis, George Trigeorgis, Mihalis A. Nicolaou, and et al., "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, 2017.

[20] Florian Eyben, Martin Wöllmer, and Böjrn Schuller, "The opensmile book - opensmile: The munich versatile and fast open-source audio feature extractor," in *ACM Multimedia*, 2010.