ADVERSARIALLY-ENRICHED ACOUSTIC CODE VECTOR LEARNED FROM OUT-OF-CONTEXT AFFECTIVE CORPUS FOR ROBUST EMOTION RECOGNITION

Chun-Min Chang, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

ABSTRACT

Advancement in speech emotion recognition technology has brought tremendous potential in designing human-centered applications across a wide range of scenarios. However, due to the difficulty in obtaining large-scale labeled emotion corpus for every application domains, most of the existing databases are collected within disparate and limited contexts. This contextualization often undermines the variability in the emotional acoustic manifestation due to the limitation in the amount of labeled data that can be collected for each particular context. This, hence, creates a robustness issue across emotional scenarios. In this work, we propose to learn an enhanced acoustic code vector for in-context emotion database through adversarially learning from large out-of-context emotion corpus to obtain robust emotion recognition. We demonstrate that our framework can obtain improved recognition accuracy using low dimensional representations on two different databases, and it maintains its modeling power even when given very limited in-context training samples.

Index Terms— behavioral signal processing (BSP), adversarial network, emotion recognition, cross corpus learning

1. INTRODUCTION

Affective computing has taken major steps in the past decade with algorithmic advancements finding its way to integrate with modern commercial applications, e.g., natural human-computer interface [1], health care [2], and marketing. While many research has focused on studying different non-verbal modalities, e.g., facial landmarks, action units, and physiological signals [3], speech continues to be the most information rich and accessible message exchange medium for human. A number of survey papers has indicated several key acoustic cues would carry important emotion information [4, 5, 6, 7], but the variability of emotion modulation in these cues remain highly variant. Given that the major variability in emotional acoustic manifestation is in the context, e.g., recording conditions [8], interaction types [9], application domains [10], etc, most of the real life emotion applications and/or corpus collected are often highly contextualized.

These contextualization processes in applications result in disparate emotions corpus collected for each scenario that can be limited in scale due to the expensive data collection process. Obtaining robust speech emotion recognition across domains is thus challenging. Several past works have relied on the psychological theory of universal emotion perception [11, 12] in deriving transferable algorithms between corpora, e.g., Bezooijen et al. ultilized three different languages (Dutch, Taiwanese, and Japanese) to identify Dutch vocal expressions of emotion [13] and Schuller et al. attempted to construct an universal emotion recognizer for multiple languages simultaneously through representation normalization [14]. Most of these works have not yet shown their significant effectiveness likely due to the high heterogeneity exists between corpora. The limited data size problem further makes the true variability of emotion information not well captured, which creates a robustness issue. Hence, instead of learning to transfer, we have previously proposed to perform multi-view integration by leveraging out-of-context emotion corpus to improve robust recognition of in-context data [15, 16].

In this work, we propose a novel adversarial network architecture to learn an emotion-enriched acoustic vector for in-context emotion data (limited in scale) by leveraging out-of-context (larger in scale) emotion corpora. The use of adversarial loss has been shown to be a promising methodology for learning the underlying generative model. Few works have utilized adversarial network to perform emotion recognition showing that valence dimension could be further improved [17]. Most consider adversarial learning essentially as data augmentation method for the in-context database only. Our proposed network instead generates the enriched acoustic vector learning through out-of-context emotion from larger emotion corpus using adversarial mechanism.

We evaluate our framework on two in-context databases, i.e., the USC CreativeIT database [18] and the VAM database [19], by leveraging out-of-context databases, i.e., the USC IEMOCAP [20] and the NNIME [21]. Our experiments show that our proposed adversarially-enriched vector can obtain improved emotion recognition accuracy with low dimensional representation in both databases. Importantly, our analysis demonstrates that our framework retains its robustness even when trained with much reduced in-context data.

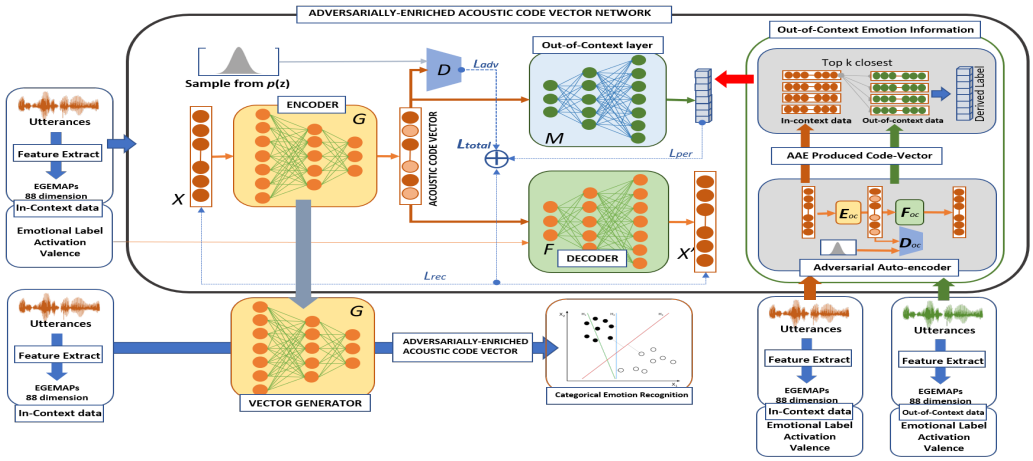

Fig. 1. Figure shows our proposed framework we proposed. An emotion-enriched acoustic vector for in-context emotion data is learned adversarially by leveraging out-of-context emotion corpora, then SVM is used to perform final classification.

2. RESEARCH METHODOLOGY

2.1. Emotion Databases

In this work, we utilize two different in-context (limited in scale) databases as our main emotion recognition evaluation corpus, and two out-of-context (much larger in scale) emotion corpora to be used in the training of our adversarially-enriched code vector. We will briefly describe each database.

2.1.1. In-Context Databases: CreativeIT and VAM

We use two different in-context emotion corpus, the USC Creative IT database and the VERA AM MITTAG (VAM) database. The USC CreativeIT database is a public available emotion corpus designed within the *context* of an established theatrical acting technique to carry out affective dyadic interactions. This database includes a total of 16 actors (8 male, 8 female) forming 8 pairs to engage in 3 to 5 minutes long improvisations. Each interaction is rated by 3 raters using continuous-in-time annotation on attributes of valence, activation, and dominance (the scale ranges between 1 to -1). There are a total of 2162 utterances in the database, we further binarize the average values of rated valence and activation of an utterance (class 1: [-1:0] and class 2: (0:1]). In summary, a total of 2162 utterances each with a binary label indicating high versus low for activation and valence are used as our target data for the USC CreativeIT.

The VAM corpus consists of recordings in the *context* of a German talk show. Each show consists of several multi-party (2 to 5 persons) dialogs, and 70% of speakers collected are 35 or younger at that time. The annotation includes attributes of activation, valence, and dominance on the segmented sentences. We also binarize the emotion annotations into high versus low (class 1: [-1:0] and class 2: (0:1]). The VAM corpus includes a total of 947 samples with each being labeled with a binarized valence and activation score used as another target in-context database in this work.

2.1.2. Out-of-Context Databases: NNIME and IEMOCAP

In this work, we use two different out-of-context emotion corpus, the NNIME and the USC IEMOCAP database. The NNIME is a new multimodal Mandarin Chinese affective interaction corpus. The NNIME database contains recordings of 44 subjects engaged in spontaneous dyadic spoken interactions with each lasts approximately 3-minute long. The database is annotated by 4 naive annotators on attributes of valence and activation. There are a total of 6509 utterances segmented in the database. We average the 4 naive annotator's rating and binarize it into high versus low (binary class: [-1,0] and (0,1]) as the labels used in this work

The USC IEMOCAP database is a well-known audio-visual English emotional database. The database consists of 5 dyadic sessions with a total of 10 actors (5 males and 5 females) grouping in pairs to engage in dyadic face-to-face interactions. There is approximately 12 hours of data segmented into utterance (a total of 6905 sentences). We average the rated activation and valence labels for each utterance over raters and binarize the values into high versus low (binary class: [1,3) and [3,5]) to be used in this work.

2.2. Adversarially-Enriched Acoustic Vector

Fig. 1 displays our framework in deriving adversarially-enriched acoustic vector by leveraging out-of-context corpus. We will describe acoustic features and adversarial learning from out-of-context emotion corpus in the following.

2.2.1. Acoustic Features

We first extract 88 dimensional eGeMAPS acoustic features using the OpenSmile toolbox for every utterance due to its demonstrated robustness in characterizing acoustic emotion information across databases [22]

2.2.2. Adversarial Representation Learning

We use adversarial auto-encoder (AAE) as our core representation learning approach for utterances of our in-context database. As depicted in Fig. 1, $G : X \rightarrow C$ is the encoder portion to extract the representation, and $F : C \rightarrow X'$ is the decoder part to project back to the original feature space. C is the learned latent acoustic code vector. D is the discriminator to discriminate C, latent vector. Instead of learning

Table 1. *Exp 1: Summary of our proposed adversarially-enriched acoustic vector for emotion classification*

In-Context: Creative IT							
	$baseline_{fs}$	$baseline_{fs}$	AAE	*out-of-context:* IEMOCAP		*out-of-context:* NNIME	
	88 Dimension	*88 Dimension*	*64 Dimension*	*64 Dimension*	*10 Dimension*	*64 Dimension*	*10 Dimension*
Act.	62.9 (Ds=88)	64.6 (Ds=36)	64.1 (Ds=16)	65.7 (Ds=26)	***66.1*** (Ds=**4**)	66.0(D=45)	***66.1*** (Ds=8)
Val.	52.3 (Ds=88)	53.4(Ds=36)	52.7 (Ds=16)	55.3 (Ds=26)	54.3 (Ds=4)	***55.7***(Ds=45)	***55.0*** (Ds=**1**)
In-Context: VAM							
	$baseline_{fs}$	$baseline_{fs}$	AAE	*out-of-context:* IEMOCAP		*out-of-context:* NNIME	
	88 Dimension	*88 Dimension*	*64 Dimension*	*64 Dimension*	*10 Dimension*	*64 Dimension*	*10 Dimension*
Act.	72.8 (Ds=88)	***76.9*** (Ds=**9**)	74.7 (Ds=38)	75.5 (Ds=26)	***75.2*** (Ds=4)	76.1(Ds=45)	***76.1*** (Ds=7)
Val.	52.2 (Ds=88)	52.9(Ds=78)	56.6 (Ds=38)	***63.0*** (Ds=**7**)	58.4 (Ds=2)	60.9(Ds=7)	59.3 (Ds=1)

vanilla AAE, we add an additional condition to ensure the representation can integrate the emotion labels of the in-context database. The conditional constraint is added to the decoder, $F : (C, Y) \rightarrow X'$, where Y indicates the in-context emotion label. The in-context emotional acoustic code vector can then be learned using the following modified reconstruction loss:

$$L_{rec}(G, F) = \arg\min_{G,F}[\|X' - X\|]^2$$

where $X' = F(G(X), Y)$ denoted as the reconstructed data. The latent vector is further constrained using a Gaussian distribution, Z which $p(z) = N(z|0, I)$ making the adversarial loss to be in the following form:

$$L_{adv}(G, D, X, Z) = \min_G \max_D E_{z\sim p_z}[log(D(z))] + E_{x\sim p_{data}(x)}[log(1 - D(G(x)))]$$

2.2.3. Enriched Code Vector from Out-of-Context Data

In order to learn an enriched code vector from out-of-context, we adopt a similar concept as our previous approach [15, 16], i.e., by jointly learning the emotion label derived from out-of-context data as additional auxiliary label for the in-context sample. For every sample j of the in-context data X_i and out-of-context data X_o, we first map them into a latent vector, denoted as E_i and E_o, using conventional adversarial autoencoder [23].

For every j-th sample in E_i, denoted as E_i^j, we identify K-nearest samples from the out-of-context dataset by computing cosine similarity between E_i^j to all samples in E_o. Each of k-th identified E_o has an associated label in the out-of-context Y_{io}^k dataset. With this information, we modify the conditional AAE architecture to integrate this auxiliary out-of-context emotional information for every j-th sample for the in-context database as an additional loss:

$$L_{per}(X_i) = \|M(G(x)) - Y_{io}^K\|^2$$

The objective function includes three different loss defined:

$$L_{total}(G, F, D, M) = L_{rec} + L_{per} + L_{adv}$$

2.3. Emotion Classifier

After learning the complete out-of-context enriched conditional adversarial autoencoder structure, we take $G : X \rightarrow C$ as the latent representation extractor to derive features to be used in training a support vector machine (SVM) for final emotion classification.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental Setup

In this work, we conduct two different experiments:

- Exp1: Recognition experiments for the two in-context databases.
- Exp2: Reduced in-context labeled samples for recognition experiment.

All evaluation is done via leave-one-speaker-out cross validation, and the accuracy is measured in unweighted average recall (UAR). Univariate feature selection is carried out based on ANOVA-F test, the adversarial learning framework is done in every fold of training set. The parameters of the adversarial network are listed below: the learning rate, and the number of epoch is set to be 0.0005, $100 \sim 300$ respectively. All the models are 3-layer DNN architecture, activation function of all the layer use are leaky_relu.

In Exp1, we generate latent vectors of two different size (64 dimensions and 10 dimensions). Result of each indicates training a SVM after performing univariate feature selection. They are compared with three different baselines. $baseline$ indicates training a SVM directly on the 88 dimensional eGeMAPs, and $baseline_{fs}$ indicates training a SVM after performing univariate feature selection, and AAE indicates that the acoustic vector are first learned using AAE on the in-context emotion data only and then trained with SVM.

In Exp2, our aim is to assess the robustness of our framework where there is limited labeled in-context data. Specifically, we present accuracy obtained for our framework as we reduce the number of in-context labeled data.

3.2. Experimental Results and Analysis

3.2.1. Exp 1: Recognition Experiments

Table 1 lists a summary of our emotion recognition results experiments of in-context database for our proposed framework. D indicates the number of latent dimension used. In the Creative IT database, the best result are obtained by leveraging the NNIME as the out-of-context emotion database. Our method achieves UAR of 66.1% and 55% for activation and valence respectively, which is 3.2 % and 2.7% relative

Table 2. *Exp2: Summary of reduced sample recognition experiments*

	In-Context: Creative IT						**In-Context: VAM**					
	$baseline_{fs}$						$baseline_{fs}$					
Sample#	**full(100%)**		500 (25%)		200(10%)		**full(100%)**		500(50%)		100(10%)	
Act.	64.6		61.5		60.1		76.9		75.5		72.5	
Val.	53.4		53.9		53.3		52.9		52.0		50.0	
	64 Dimension			*10 Dimension*			*64 Dimension*			*10 Dimension*		
IEMOCAP	**full**	500	200	**full**	500	200	**full**	500	100	**full**	500	100
Act.	65.7	66.5	64.7	66.1	66.2	65.8	75.5	75.6	72.5	75.2	75.9	73.8
Val.	55.3	61.1	56.0	54.3	57.5	55.0	63.0	62.6	57.8	58.4	60.8	57.5
NNIME	**full**	500	200	**full**	500	200	**full**	500	100	**full**	500	100
Act.	66.0	66.7	64.8	66.1	66.2	64.0	76.1	76.6	73.7	76.1	76.3	73.7
Val.	55.7	60.6	57.1	55.0	58.0	55.3	60.9	63.7	58.0	59.3	60.3	57.8

improvement over the baseline model. Importantly, the required feature dimensions in this case are only 8 and 1 respectively to obtain the best recognition improvement. This result shows the ability of our proposed structure in compactly represent emotionally-relevant information. Furthermore, for VAM database, we observe comparable performance in the activation dimension with as little as 7 dimensions. It is likely due to the fact that baseline method already perform quite well on the activation dimensions (76.9%). In the valence dimension, by using IEMOCAP as the out-of-context database, we obtain an accuracy of 63.3%, which is 9.8% better over the baseline model, with only 7 dimensions as well.

In general, our proposed acoustic vector contain more emotionally-relevant representation power by leveraging out-of-context databases, e.g., comparing between enriched vectors and AAE or baseline methods. Furthermore, we consistently observe that our adversarially learned code vector only need very few dimensions (< 10) to obtain the best discriminative power.

3.2.2. Exp 2: Reduced In-Context Labeled Samples

In Exp 2, we reduce the total number of available labeled data samples in both learning the adversarially-enriched acoustic code vector and the training of the SVM classifier to assess the robustness of our framework. Our goal is to understand whether it is possible that our method can retain *high-level* emotion information in the acoustic vector even under severe lack-of-data condition.

Table 2 lists a summary of our emotion recognition experiments for the Exp 2. We test samples number in the range of 500, 200, and 100 (50% to 10%). From the Table, while all of the methods suffer loss of accuracy as we decrease the number of available samples. Our method is more robust against this severe training condition. For example, in the Creative IT, an obvious significant drop in baseline activation accuracy occurs when using 500 samples (from 64.6% to 60.1%); however our proposed method still maintain its activation recognition rate at close to 66%. In the VAM corpus, this effect is also evident. When reducing the sample number to 100 (about 10% of the original VAM corpus), the accuracy drops from 76.9% and 52.9% to 72.5% and 50% in activation and valence, respectively. While our accuracy for activation trained with 100 samples is also comparable at around 73%, the valence, however, maintains its accuracy at close to 58%.

The results presented in Table 2 is quite encouraging. Most of the contextualized emotion corpus usually starts off as a small database in its scale, by adversarially learn an enriched code vector by leveraging existing large-scale out-of-context emotion corpus, it *adds* relevant emotion information to the acoustic representation that are not present in the limited in-context database.

4. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose a novel adversarial network architecture to learn an emotion-enriched acoustic vector for in-context emotion data by leveraging out-of-context emotion corpora. Our experiments demonstrate an improved recognition accuracy in two in-context database by integrating two different out-of-context emotion corpus. We observe that few dimensions of representations is sufficient to train the emotion recognizer. Additional experiments demonstrate that our framework is robust to training scenarios where only limited data is available. This work presents one of the first works in using adversarial mechanism in improving the robustness of emotion recognition by integrating information from larger out-of-domain corpus.

There are several future directions. The first is to examine the effect on the characteristics of out-of-context database to understand whether the types of interactions, the language of the database, and the size of the available samples would have an impact on the in-context emotion recognition. Further, the framework now requires emotion labels from the out-of-context database, which limits the scale and the availability of the database that could be utilized. We would further investigate approach in deriving unsupervised weak emotion labels, e.g., [24], to robustly integrate yet another diverse view of emotion perception in enhancing emotionally-relevant acoustic representation.

5. REFERENCES

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

[2] Min Chen, Yin Zhang, Yong Li, Mohammad Mehedi Hassan, and Atif Alamri, "Aiwac: Affective interaction through wearable computing and cloud technology," *IEEE Wireless Communications*, vol. 22, no. 1, pp. 20–27, 2015.

[3] Wei-Long Zheng and Bao-Liang Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

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

[5] Rainer Banse and Klaus R Scherer, "Acoustic profiles in vocal emotion expression.," *Journal of personality and social psychology*, vol. 70, no. 3, pp. 614, 1996.

[6] Klaus R Scherer and James S Oshinsky, "Cue utilization in emotion attribution from auditory stimuli," *Motivation and emotion*, vol. 1, no. 4, pp. 331–346, 1977.

[7] Jo-Anne Bachorowski, "Vocal expression and perception of emotion," *Current directions in psychological science*, vol. 8, no. 2, pp. 53–57, 1999.

[8] Antonio Camurri, Ingrid Lagerlöf, and Gualtiero Volpe, "Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques," *International journal of human-computer studies*, vol. 59, no. 1-2, pp. 213–225, 2003.

[9] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[10] Debra L Roter, Richard M Frankel, Judith A Hall, and David Sluyter, "The expression of emotion through nonverbal behavior in medical visits," *Journal of general internal medicine*, vol. 21, no. 1, pp. 28–34, 2006.

[11] James A Russell, "Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies.," *Psychological bulletin*, vol. 115, no. 1, pp. 102, 1994.

[12] Paul Ekman, "Universals and cultural differences in facial expressions of emotion.," in *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

[13] Renée Van Bezooijen, Stanley A Otto, and Thomas A Heenan, "Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics," *Journal of Cross-Cultural Psychology*, vol. 14, no. 4, pp. 387–406, 1983.

[14] Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[15] Chun-Min Chang and Chi-Chun Lee, "Fusion of multiple emotion perspectives: Improving affect recognition through integrating cross-lingual emotion information," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5820–5824.

[16] Chun-Min Chang, Bo-Hao Su, Shih-Chen Lin, Jeng-Lin Li, and Chi-Chun Lee, "A bootstrapped multi-view weighted kernel fusion framework for cross-corpus integration of multimodal emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 2017, pp. 377–382.

[17] Mohammed Abdelwahab and Carlos Busso, "Domain adversarial for acoustic emotion recognition," *arXiv preprint arXiv:1804.07690*, 2018.

[18] Angeliki Metallinou, Chi-Chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan, "The usc creativeit database: A multimodal database of theatrical improvisation," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, p. 55, 2010.

[19] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan, "The vera am mittag german audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 865–868.

[20] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[21] Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee, "Nnime: The nthu-ntua chinese interactive multimodal emotion corpus," in *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 2017, pp. 292–298.

[22] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, 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.

[23] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[24] Chih-Chuan Lu, Jeng-Lin Li, and Chi-Chun Lee, "Learning an arousal-valence speech front-end network using media data in-the-wild for emotion recognition," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 99–105.