INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Reconciliation of Multiple Corpora for Speech Emotion Recognition by Multiple Classifiers with an Adversarial Corpus Discriminator

Zhi Zhu, Yoshinao Sato

Fairy Devices Inc., Japan
{zhu, sato}@fairydevices.jp

Abstract

Research on affective computing has achieved remarkable success with the development of deep learning. One of the major difficulties in emotion recognition is inconsistent criteria for emotion categorization between multiple corpora. Most previous studies using multiple corpora discard or merge a part of their emotion classes. This prescription causes catastrophic information loss with respect to emotion categorization. Furthermore, the influences of corpus-specific factors other than emotions, such as languages, speech registers, and recording environments, should be eliminated to fully utilize multiple corpora. In this paper, we address the challenge of reconciling multiple emotion corpora by learning a corpus-independent emotion encoding disentangled from all the remaining factors without causing catastrophic information loss. For this purpose, we propose a model that consists of a shared emotion encoder, multiple emotion classifiers, and an adversarial corpus discriminator. This model is trained with multi-task learning harnessed by adversarial learning. We conducted speech emotion classification experiments with our method on two corpora, namely, EmoDB and CREMA-D. The results demonstrate that our method achieves higher accuracies than mono-corpus models. In addition, it is indicated that the proposed method suppresses corpus-dependent factors other than emotions in the embedding space.

Index Terms: speech emotion recognition, multiple corpora, adversarial learning

1. Introduction

In recent years, research on affective computing has achieved tremendous success with the rapid development of deep learning. Among other issues, a wide range of neural network models has been proposed for speech emotion recognition (SER) [1, 2, 3, 4, 5, 6, 7].

One of the most significant challenges in this research field is the resolution of discrepancies in the criteria for emotion categorization between multiple corpora. An emotion corpus has its own purpose and defines relevant emotion classes accordingly, ignoring unnecessary discriminations and categories of emotional expressions. When we look at existing emotion corpora, each corpus postulates a different number of different emotion classes, as scrutinized in [8]. Even if multiple corpora share an emotion label, emotional expressions belonging to that class are not necessarily equivalent. Therefore, it is not appropriate to fully equate any pair of emotion categories across corpora. Each emotion class of each corpus needs to be considered as unique, without any changes, even if the labels assigned to them are the same. In other words, the methods to categorize emotion are intrinsically corpus-dependent, and there exists no corpus-independent universal standard. Consequently, the definitions of emotion categories are inconsistent between multiple corpora. These discrepancies in criteria prevent researchers from fully utilizing multiple emotion corpora. Hence, we need to resolve the incoherence of emotion categorization between multiple emotion corpora to advance research beyond the limitations of a single corpus.

Furthermore, we need to overcome the influences of corpus-specific factors other than emotional expressions, such as language, speech register, and recording environment. These factors are intrinsically entangled with emotional expressions; thus, they can hinder research on emotion recognition using multiple corpora. Therefore, we need a method to learn an emotion encoding disentangled from all the remaining factors to fully utilize multiple corpora.

Most previous studies using multiple emotion corpora redefined new classes after determining the correspondence of classes between corpora more or less subjectively. For example, some studies discarded emotion classes with non-shared labels [9, 10]. In other studies, emotion classes were merged into a limited number of common classes, such as positive, negative, and neutral [11, 12]. Yet other studies used adversarial learning for domain aggregation, and these methods suppose that all domains share the same objective classes [11, 13]. These coarse-graining prescriptions change the boundaries of emotional expressions and exclude parts of categories, thus resulting in catastrophic information loss. In contrast, multiple emotion classifiers trained by multi-task learning, each of which corresponds to each corpus, do not sacrifice information regarding emotion categorization as investigated in [14]. However, corpus-specific factors are not eliminated using multiple emotion classifiers alone.

The purpose of this paper is to reconcile multiple emotion corpora by learning a corpus-independent emotion encoding that is disentangled from all the remaining corpus-specific factors without causing catastrophic information loss. For this purpose, we propose a model that consists of a shared emotion encoder, multiple emotion classifiers, and an adversarial corpus discriminator, which is harnessed by adversarial learning. The proposed method is the first attempt to apply multi-task learning harnessed by adversarial corpus discrimination to emotion recognition on multiple corpora. Our method applies to a wide range of research on emotion recognition, not only in audio but also in other modalities, including a large-scale study using a number of corpora, a comparative study of emotional expressions between corpora, and a study of domain aggregations for emotion recognition. Furthermore, the proposed method is valuable for various applications of emotion recognition "in the wild." In practical applications of SER, emotion categories that differ from those of a common corpus are postulated according to the purpose. Additionally, in most cases, only a small amount of data is available. Nevertheless, the proposed system can perform efficiently even in such cases.

http://dx.doi.org/10.21437/Interspeech.2020-1618

2. Methodology

2.1. Model architecture

We propose a neural network model that consists of a shared emotion encoder, multiple corpus-dependent emotion classifiers, and an adversarial corpus discriminator. Figure 1 illustrates the architecture of our model. As the encoder part is shared between corpora, input audio data is projected into the same embedding space, regardless of which corpus it originates from. Each emotion classifier in the output layer corresponds to a specific corpus. This structure enables us to deal with all emotion classes of all corpora as different ones without discarding or merging any of them. In addition to these components, the proposed model is equipped with a corpus discriminator in the output layer. The entire model is trained with multi-task learning for emotion classifiers harnessed by adversarial learning for the corpus discriminator, as detailed in Section 2.2. Thanks to adversarial learning, any remaining factors other than emotional expressions are annihilated during the encoding. With this mechanism, the proposed method can learn corpus-independent encoding of emotional expressions without causing catastrophic loss of information.

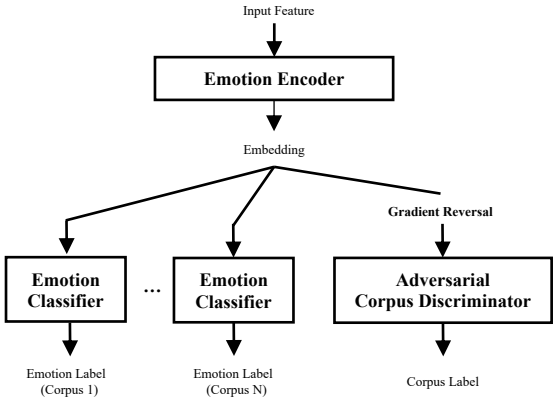

Figure 1: *Model architecture*

The proposed method differs from those investigated in previous studies as follows. If we get rid of the adversarial corpus discriminator, the structure of our model becomes equivalent to the one investigated in [14]. We refer to this as a multiple emotion classifier model. Such a model cannot be harnessed by adversarial learning. Therefore, corpus-specific factors other than emotional expressions inevitably remain in the encoding. Aside from this, adversarial learning was applied to SER in another previous study [11], which investigated a model with a single emotion classifier. The use of a single classifier alone is possible by merging all emotion classes into two, namely positive and negative. Therefore, some of the intrinsic information about the emotions of multiple corpora is inevitably lost. In addition, the purpose of this previous study was domain adaptation, which is different from our aim: to learn corpus-independent emotion embedding disentangled from all remaining corpus-specific factors without causing catastrophic information loss. These differences explain the reason why combinations of multiple emotion classifiers and an adversarial corpus discriminator are essential for our purpose.

2.2. Adversarial learning

Multi-task learning acts as a kind of regularizer in training neural network models and improves their generalization abilities [14, 9]. However, this learning method alone is not enough to work successfully work with multiple corpora. In an embedding space, corpus-specific information remains entangled with emotional expression information. As a result, utterances that originate from different corpora are projected in this space far from each other. In order to learn corpus-independent encoding of emotions, we harness the power of adversarial learning for corpus discrimination.

Before describing an adversarial learning procedure in detail, we introduce some notations. The total loss function $\mathcal{L}$ is given by

$$\mathcal{L} = \mathcal{L}_{\mathrm{emo}} + \mathcal{L}_{\mathrm{cor}}, \tag{1}$$

where $\mathcal{L}_{\mathrm{emo}}$ and $\mathcal{L}_{\mathrm{cor}}$ represent loss functions for emotion classification and corpus discrimination, respectively. $\mathcal{L}_{\mathrm{emo}}$ is defined by the sum of the loss functions for emotion classification of each corpus $\mathcal{L}^{(i)}_{\mathrm{emo}}$:

$$\mathcal{L}_{\mathrm{emo}} = \sum_{i=1}^{N} \mathcal{L}^{(i)}_{\mathrm{emo}}. \tag{2}$$

Here, N denotes the number of corpora. An adversarial loss for corpus discrimination $\mathcal{L}_{\mathrm{adv}}$ is defined by

$$\mathcal{L}_{\mathrm{adv}} = -\mathcal{L}_{\mathrm{cor}}. \tag{3}$$

Besides, we use two different learning rates: ϵ and ϵ_{adv}.

During an adversarial learning procedure, a network is trained at each batch in the following way:

- Step 1: Update the emotion encoder and the emotion classifiers to minimize $\mathcal{L}_{\mathrm{emo}}$.
- Step 2: Update the corpus discriminator to minimize $\mathcal{L}_{\mathrm{cor}}$.
- Step 3: Update the emotion encoder to minimize $\mathcal{L}_{\mathrm{adv}}$, and hence, to maximize $\mathcal{L}_{\mathrm{cor}}$.

As a learning rate, we use ϵ at step 1 and step 2, whereas we use ϵ_{adv} at step 3. Note that at step 2, while the corpus discriminator is updated on the basis of $\mathcal{L}_{\mathrm{cor}}$, the emotion encoder is fixed. At step 2, in contrast, the emotion encoder is updated on the basis of $\mathcal{L}_{\mathrm{adv}}$, whereas the corpus discriminator is fixed. Another important point is that when an error in corpus discrimination propagates backwards to update the emotion encoder at step 3, the sign of the gradient of $\mathcal{L}_{\mathrm{cor}}$ is reversed. This means the encoder learns to eliminate factors relevant to corpus discrimination. As a result, utterances of speech accompanied by similar emotional expressions are projected close to each other in the embedding space, regardless of which corpus they originated from.

An adversarial game is played during the learning procedure. Firstly, the whole network is trained so that it can classify emotions, as well as discriminate between corpora. Then, the encoder part is re-trained, so that it eliminates corpus-dependent information from the embedding space. The goal of adversarial learning is to classify emotion classes and annihilate corpus-specific factors other than emotions. The relative magnitude of the two learning rates controls which of the emotion classifiers and the corpus discriminator will prevail in adversarial games. Note that our aim is not to achieve the chance-level performance in corpus discrimination because multiple corpora have intrinsic differences in emotional expressions.

3. Experiments

3.1. Feature and model details

Firstly, we split all utterances into segments with a length of 3 s and a shift of 1 s. Zero-padding was applied when necessary. Each segment was assigned the emotion label of the original utterance. In our experiments, we dealt with these segments as data samples. We used a log mel-spectrogram as the input features. After down-sampling to 16 kHz, a 40-dimensional log mel-spectrogram was calculated with a window size of 25 ms and a window shift of 10 ms. We applied z-score normalization to the input features.

We used an attention-based convolutional recurrent neural network (ACRNN) as the emotion encoder. ACRNN has been shown to be efficient for SER [4, 5, 7, 6]. Table 1 shows a detailed structure of the network used in our experiments. The emotion classifier and corpus discriminator used in our experiments consist of three fully connected layers with 128 units, followed by a single softmax layer, which output an emotion and a corpus label, respectively.

Table 1: *Structure of ACRNN emotion encoder*

Layer	Filter		Output Size
	No.	Size	
Input			300×40
Convolution	128	5×3	$300 \times 40 \times 128$
Max Pooling		2×4	$150 \times 10 \times 128$
Convolution	256	5×3	$150 \times 10 \times 256$
Max Pooling		1×5	$150 \times 2 \times 256$
Reshape			150×512
TDFC	768		150×768
BLSTM	128		150×256
Attention			8×256
Flatten			2048
Fully Connected	128		128
Dropout			128

TDFC: Time Distributed Fully Connected
BLSTM: Bidirectional Long Short-Term Memory

3.2. Corpora

We used two emotional speech corpora: the Berlin emotional speech database (EmoDB) [15], and the crowd-sourced emotional multimodal actors dataset (CREMA-D) [16]. EmoDB is a German emotional speech corpus that includes 535 utterances. Ten professional actors read scripts in seven different emotional states: neutral, happiness, sadness, anger, disgust, fear, and boredom. CREMA-D is an audio-visual corpus collected to explore human emotional expressions and perceptions. This corpus includes 7442 utterances in 2-party acted dialogs by 91 actors in six emotional states: neutral, happiness, sadness, angry, disgust, and fear. We used all utterances from these two corpora without merging or discarding any emotion classes.

3.3. Experimental setup

We compared the proposed model with the following models: a mono-corpus model, a multiple emotion classifier model, and a multiple emotion classifier model with a corpus discriminator. A mono-corpus model has a single emotion classifier and is trained on a single corpus. We examined mono-corpus models for EmoDB and CREMA-D. A multiple emotion classifier model has two emotion classifiers and is trained on both corpora. The last model is equivalent to our model, except that adversarial learning has not been used. All models have an encoder with the same structure.

For each model, we performed 10-fold leave one speaker group out (LOSGO) cross-validation ten times and calculated average weighted accuracy (WA) and unweighted accuracy (UA). Specifically, we divided the speakers into ten groups, so that each group contained almost the same number of speakers. All samples were grouped into three sets based on the speaker groups: eight groups for training, another group for validation, and the last group for evaluation. The training set was shuffled randomly and divided into mini-batches of 64 samples. The validation set was used to choose the optimal epoch measured by the mean accuracy. Finally, the test set was used to evaluate the performance of the optimal model. Under multi-corpus conditions, each speaker group included one speaker group of EmoDB and one speaker group of CREMA-D.

We used the cross-entropy loss function. We set the initial value of the learning rate ϵ to 10^{-4} with a decay rate of 0.91 until 10^{-5}. The learning rate ϵ_{adv} was fixed to 5^{-8}. The training process was terminated when the accuracy of the validation set had not improved in the last 500 epochs.

4. Results

Table 2 shows WA and UA averaged over ten cross-validations. The proposed method achieved higher performance than both mono-corpus models. Moreover, the performance of our method was the best of all models, measured by the mean accuracy over two corpora. Note that multiple emotion classifier models, with or without a corpus discriminator, were better than mono-corpus models. The improvements in performance for EmoDB, which is of a smaller size, were more substantial than for CREMA-D, which is of a larger size. This indicates that a multiple emotion classifier model is effective even when only a few utterances in a target domain are available.

In addition, we analyzed the confusion patterns of our method to clarify how adversarial learning affected the encoding. We refer to a sample, whose posterior probability of correct corpus classification is lower than 0.8 as a corpus-ambiguous one, and all others as corpus-definite ones. Figure 2 illustrates how a sample of one corpus was classified by the emotion classifier of another corpus. As for EmoDB, disgust was the most corpus-ambiguous class, whose utterances tended to be classified as anger according to the criteria of CREMA-D. In contrast, for CREMA-D, anger was determined to be the most corpus-ambiguous class, whose utterances tended to be classified as anger or disgust according to the criteria of EmoDB. These results suggest that anger in CREMA-D and disgust in EmoDB are relatively similar emotional expressions. The ratio of corpus-definite samples r_{cor} was lower than 1.0, as shown in Table 2. This means that the emotional expressions of the two corpora partially overlapped. In a preliminary experiment, a lower r_{cor} did not necessarily result in better accuracy.

Furthermore, we explored the emotion embedding space of the proposed method using uniform manifold approximation and projection (UMAP) [17]. Figure 3 illustrates the distribution of data in the embedding space at a certain fold of the cross-validation. The efficiency of adversarial learning can be confirmed from the distributions of EmoDB and CREMA-D, which are relatively close in the embedding space. This result indicates that an adversarial corpus discriminator suppresses the influences of corpus-specific factors other than emotional expressions.

Table 2: *Results of experiments.* r_{cor} *represents the ratio of corpus-definite samples.*

	EmoDB			CREMA-D			mean	
	WA	UA	r_{cor}	WA	UA	r_{cor}	WA	UA
mono-corpus EmoDB	0.753	0.743						
mono-corpus CREMA-D				0.787	0.729			
multiple emotion classifier	0.758	0.748		**0.796**	**0.734**		0.777	0.741
+ corpus discriminator	0.769	0.760	0.998	0.789	0.728	1.000	0.779	0.744
+ adversarial [proposed]	**0.779**	**0.772**	0.911	0.795	0.731	0.989	**0.787**	**0.752**

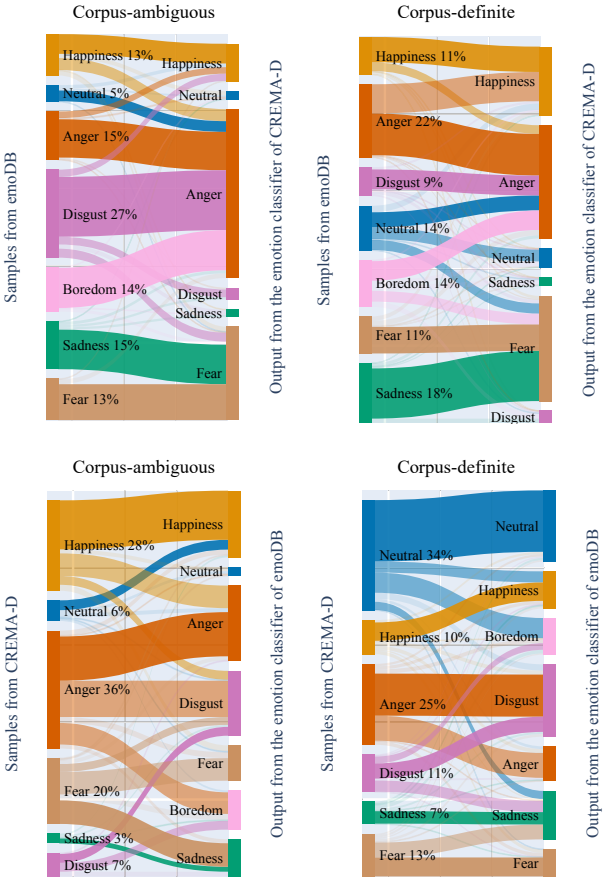

Figure 2: *Confusion patterns of corpus-ambiguous and corpus-definite samples*

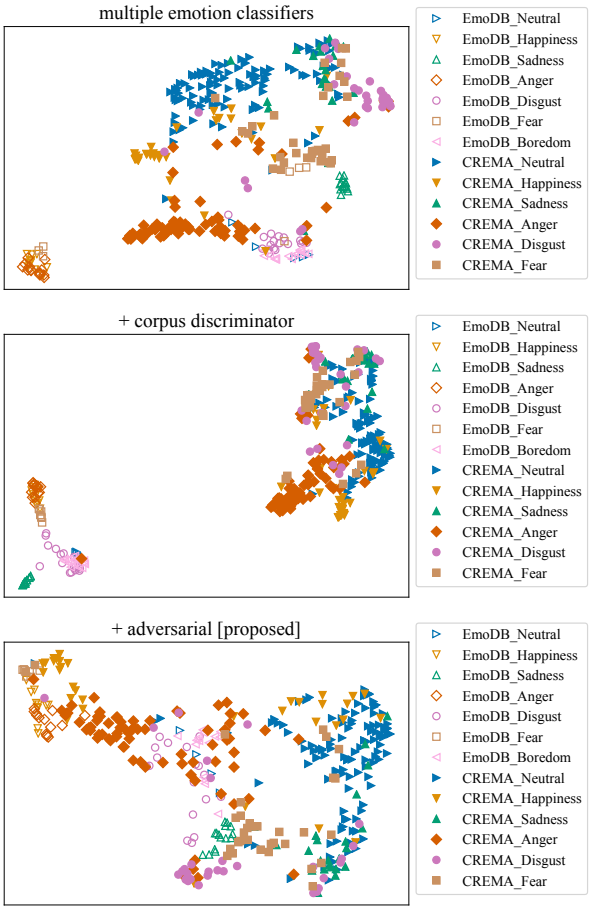

Figure 3: *Distributions in embedding spaces*

5. Conclusion

In this paper, we investigated a method for the reconciliation of multiple emotion corpora, which present incoherent criteria for emotion categorization, without causing catastrophic information loss by learning a corpus-independent encoding of emotions disentangled from all the remaining corpus-specific factors. For this purpose, we proposed a neural network model for SER that consists of a shared emotion encoder, multiple corpus-dependent emotion classifiers, and an adversarial corpus discriminator. To evaluate our method, we conducted speech emotion classification experiments using two corpora: EmoDB and CREMA-D. The results demonstrate that the proposed method achieves better performance than mono-corpus models trained on each corpus. Furthermore, the performance of our model as measured by the mean accuracy on the two corpora was the best of all the models with which it was compared. In addition to these analyses, we explored confusion patterns across corpora and distributions in the emotion embedding spaces. The results indicate that an adversarial corpus discriminator suppresses the influences of corpus-specific factors other than emotional expressions. Future research directions include investigation of various other emotion corpora, evaluation of the extent to which an encoder suppresses corpus-specific factors quantitatively, introduction of a fiercer adversarial game. Cross-corpus speech emotion recognition using the proposed method is another direction for future research.

6. References

[1] 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.

[2] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, Proceedings*, 2018, pp. 3087–3091.

[3] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP 2016 – 42th International Conference on Acoustics, Speech, and Signal Processing, March 20-25, Shanghai, China, Proceedings*, 2016, pp. 5200–5204.

[4] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association, August 20–24, Stockholm, Sweden, Proceedings*, 2017, pp. 1098–1102.

[5] 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, 2018.

[6] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends," *IEEE Access*, vol. 8, pp. 16 560–16 572, 2020.

[7] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association, September 15–19, Graz, Austria, Proceedings*, 2019, pp. 2083–2087.

[8] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition," *International Journal of Speech Technology*, vol. 21, pp. 93–120, 2018.

[9] S. Lee, "The generalization effect for multilingual speech emotion recognition across heterogeneous languages," in *ICASSP 2019 – 45th International Conference on Acoustics, Speech, and Signal Processing, May 12-17, Brighton, UK, Proceedings*, 2019, pp. 5881–5885.

[10] S. Sahu, R. Gupta, G. Sivaraman, and C. Espy-Wilson, "Smoothing model predictions using adversarial training procedures for speech based emotion recognition," in *ICASSP 2018 – 44th International Conference on Acoustics, Speech, and Signal Processing, April 15-20, Calgary, Canada, Proceedings*, 2018, pp. 4934–4938.

[11] H. Zhou and K. Chen, "Transferable positive/negative speech emotion recognition via class-wise adversarial domain adaptation," in *ICASSP 2019 – 45th International Conference on Acoustics, Speech, and Signal Processing, May 12-17, Brighton, UK, Proceedings*, 2019, pp. 3732–3736.

[12] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-corpus classification of realistic emotions – some pilot experiments," in *LREC 2010 – 3th International Workshop on EMOTION: CORPORA FOR RESEARCH ON EMOTION AND AFFECT, May 23, Valetta, Malta, Proceedings*, 2010, pp. 77–82.

[13] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 26, no. 12, pp. 2423–2435, 2018.

[14] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *ICASSP 2017 – 43th International Conference on Acoustics, Speech, and Signal Processing, March 5-9, New Orleans, USA, Proceedings*, 2017, pp. 4990–4994.

[15] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *INTERSPEECH 2005 – 6th Annual Conference of the International Speech Communication Association, September 4-8, Lisbon, Portugal, Proceedings*, 2005, pp. 1517–1520.

[16] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affecting Computation*, vol. 5, no. 4, pp. 377–390, 2014.

[17] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.