



# Discriminative Adversarial Learning for Speaker Independent Emotion Recognition

L.L Chamara Kasun<sup>1</sup>, Chung-Soo Ahn<sup>2</sup>, Jagath C. Rajapakse<sup>2</sup>, Zhiping Lin<sup>1</sup>, Guang-Bin Huang<sup>3</sup>

<sup>1</sup> School of Electrical and Electronic Engineering, Nanyang Technological University, 50, Nanyang Avenue, Singapore

<sup>2</sup> School of Computer Science and Engineering, Nanyang Technological University, 50, Nanyang Avenue, Singapore

<sup>3</sup> #18-06/07/08, Vision Exchange Venture Drive, Singapore

{chamarakasun, csahn, asjagath, ezplin}@ntu.edu.sg, gbhuang@mindpointeye.com

## Abstract

Traditional adversarial learning (AL) algorithms learn a speaker independent embedding from low level audio features. This paper introduces discriminative adversarial learning (DAL) which learn a discriminative speaker independent embedding from low level audio features such as mel frequency cepstral coefficients (MFCC) and high level audio features such as Interspeech Para-linguistics Challenge 2010. To this end, DAL jointly minimize triplet and cross-entropy losses with gradient reversal strategy for speaker independent emotion recognition (SIER). Triplet loss reduce intra-class and increase the inter-class embedding distance to improve the discriminativeness of the embedding while the cross-entropy loss determine the emotion or speaker class of the embedding and gradient reversal learn speaker independent embedding for SIER. Experiments on Emo-DB and RAVDESS datasets show that DAL outperform other traditional adversarial learning (AL) algorithms. **Index Terms:** adversarial learning, deep learning, speaker independent emotion recognition, speech emotion recognition.

## 1. Introduction

Speech emotion recognition (SER) aims at recognition of speakers' emotion from speech signals. Performances of typical SER methods are dependent on whether the speaker is present or absent in the training set. Ideally, SER algorithms are to perform irrespective of whether speaker is new to the algorithm. In this paper, we focus on Speaker-Invariant Emotion Recognition (SIER) where the training and testing data come from different speakers. The aim of SIER is to create a model for speech data spoken from a set of speakers and apply the model to predict emotions from speech from a different set of speakers. SIER is more challenging than SER as SIER algorithms learn a speaker invariant representation.

Typically for SIER tasks, the training data is collected on an environment (source domain) different from the environment (target domain) where the predictions are performed. In SIER tasks source and target domain distributions are different and learn representations that are invariant to domain. In order to address the issue of variability between training and testing data, adversarial learning approaches, inspired by Generative Adversarial Network (GAN) [1], have been proposed for SIER. One approach is to use GAN to generate synthetic data to augment the training set [2, 3, 4]. The models trained with the augmented training data have shown to perform better than models trained with only original data. Alternatively, AL can be used to resolve the disparity between training and test-

ing environments or domains [5]. Domain adversarial networks using AL was trained with source data and to predict on target data which is from another source different from the source data [6, 7]. Adversarial learning has also been investigated in removing speaker variability in speech by learning speaker invariant representations [8, 9].

AL algorithms extract features from speech and are then followed by classification of speech signals into emotions. AL algorithms use low level features such as mel frequency cepstral coefficients (MFCC) features to learn a speaker independent embedding and classify emotion. AL algorithms using both low level features and high level features such as those specified by Interspeech Para-linguistics Challenge 2010 (IS10) [10] has not been investigated. However, using both low and high level features has been investigated for non-AL algorithms and shown state-of-the-art accuracy in speech recognition [11].

SIER algorithms using both low and high level features increase the variability of the embedding as the number of input data features increase and reduce the generalization capability. Variability of the embedding can be reduced by triplet [12] or center loss [13]. Triplet loss reduce the intra-class and increase the inter-class distance [14], while center loss reduce the intra-class distance [15]. Both triplet and center loss improves the discriminativeness of an embedding as these losses learn embedding points further away from the class boundary and improves the confidence of determining the class. Learning discriminative embedding has been investigated in SER tasks but has not been investigated in SIER tasks using an AL network.

In this paper, we propose a discriminative adversarial learning (DAL) network that learn discriminative embeddings independent of speaker characteristics for SIER task. The proposed DAL network consist of a MFCC feature encoder that learns discriminative speaker-invariant embedding, IS10 feature encoder that learns high level feature embedding, an emotion classifier to predict emotion labels, and a speaker classifier that helps remove speaker variability. Our work investigate learning discriminative speaker-invariant embeddings using MFCC and IS10 features together.

## 2. Related Works

SIER tasks typically have a limited number of training data due to the difficulty of collecting and labeling speech emotion data. Hence, GAN [1] approaches have been introduced to SIER tasks to generate data for augmentation. GAN architectures model speech emotion data as a Gaussian distribution to generate data. However, due to the high variability of speech

emotion data, it cannot be modeled as Gaussian distribution and leads to poor performance [3]. Hence, speech emotion data is modeled as coming from the mixed distribution of the original speech emotion data and random Gaussian distribution by adding speech emotion data with a random Gaussian vector and generate speech emotion data. Another approach is to use representation learning approaches such as autoencoders to learn a low dimensional manifold and generate speech emotion data by modeling the low dimensional manifold as a Gaussian distribution [4]. This approach has been shown of capable of generating speech emotion belonging to a specific category such as anger or happy.

AL has also been attempted to resolve the issue of variability of data for SER [5]. Initial AL networks aims to remove the variability caused by different spoken languages such as German and English. As algorithms trained on German speech emotion data performed poorly on English speech emotion data [6]. Traditional AL networks consists of an encoder which learns a language-invariant embedding, emotion classifier to predict emotion and language classifier with gradient reversal to remove variability of the language. Learning a language invariant embedding has been extended to learning a speaker invariant embedding by changing the language classifier to speaker classifier with gradient reversal [8, 9].

Speech emotion data has high variance as speakers with different age, cultures tend to represent emotion differently. Hence, SIER models can have poor generalization due to the high variance of speech emotion data. One approach to resolve this is to learn a discriminative embedding by adding extra penalty such as triplet [12] or center loss [13]. These triplet and center losses optimizes the embedding in such a way that each embedding point moves away from the class boundaries by decreasing the intra-class and increasing inter-class distances. Triplet loss based approach use a two stage minimization, where the network is minimized by triplet loss followed by cross-entropy loss [12]. In contrast, center loss based approach jointly minimizes cross-entropy and center losses [13].

### 3. Methodology

Our approach the discriminative adversarial learning (DAL) is illustrated in figure 1. The proposed DAL network consist of four components: (i) MFCC encoder made out of 2D convolutional neural network (CNN) and bi-directional gated recurrent unit (biGRU) layers, which generate discriminative speaker invariant embedding; (ii) IS10 encoder made out of two fully connected layers, which generate high level feature embedding; (iii) an emotion classifier that predict emotional labels; and (iv) a speaker classifier with gradient reversal, which remove speaker variability from the representation used for emotion classification.

#### 3.1. MFCC Encoder

MFCC encoder takes MFCC features of input speech emotion data  $x_m = (x_m(t))$  where  $x_m(t)$  denote the features extracted at time  $t$  with a length of  $T$ ,  $x_m \in \mathbb{R}^{T \times n_m}$  and  $n_m$  denote the number of MFCC features. The encoder consist of a one or more 2D CNN layers which learn filters of short-term frequency and short-term temporal information followed by one biGRU layer which learn long-term temporal information among the learned filters. 2D CNN layer consist of (i) a 2D convolutional layer; (ii) a batch normalization layer; and (iii) a max pooling layer.

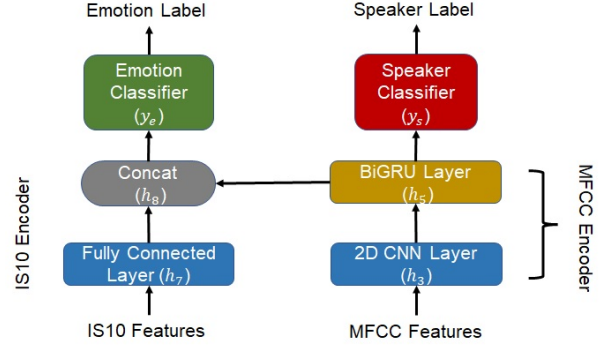


Figure 1: *Proposed discriminative adversarial learning (DAL) network architecture. The MFCC encoder consist of 2D CNN layers and a bidirectional gated recurrent unit (biGRU) layer and IS10 encoder consist of two fully connected neural network layers with ReLU activation function, dropout and batch normalization. Emotion and speaker classifier consist of two hidden layer fully connected neural network layers with Relu activation function, dropout and batch normalization followed by a softmax layer.*

Let  $w_1$  denote the filter weight of the 2D convolutional layer and the output  $h_1$  is given by:

$$h_1 = x_m \otimes w_1 \quad (1)$$

where  $\otimes$  denote the convolution operation. The convolution layer output is processed by a batch-normalization layer, which is processed by dropout and exponential linear units (eLU). The batch-normalization layer output  $h_2$  is given by:

$$h_2 = \text{eLU}(\text{dropout}(w_2 \cdot \text{BN}(h_1) + b_2)) \quad (2)$$

where  $\cdot$  denote the element-wise multiplication, BN denote batch-normalization function normalize the data by subtracting the mean and dividing by the standard deviation over the batch.  $w_2$  and  $b_2$  denote learnable parameters. `dropouts` and `eLU` denote dropout operation and eLU activation function. Max-pooling layer output  $h_3$  is given by:

$$h_3 = \text{pool}(h_2) \quad (3)$$

where `pool` denote a 2D max pooling operation which take maximum value. Let  $h_3 \in \mathbb{R}^{T \times n_3}$  where  $n_3$  is the 2D CNN output features.

The biGRU layer learn the long-term temporal relationship among the 2D CNN filters. The biGRU output  $h_4(t)$  is given by:

$$h_4(t) = \text{biGRU}(h_3(t), h_4(t-1)) \quad (4)$$

where `biGRU` are two gated recurrent units (GRU) which process the input in the forward and backward direction,  $h_4(t) \in \mathbb{R}^{n_4}$  and  $n_4$  is the number of hidden neurons in biGRU. The output of the stats pooling layer  $h_5 \in \mathbb{R}^{4 \times n_4}$  is given by:

$$h_5 = \text{stats.pool}(h_4) \quad (5)$$

where `stats.pool` function calculate the mean and standard deviation of  $h_4$  along the time dimension and concatenates along the feature dimension.

### 3.2. IS10 Encoder

IS10 encoder uses IS10 features of input speech emotion data generated by the opensmile package [16]  $x_i \in \mathbb{R}^{n_i}$  and  $n_i$  denote the number of IS10 features, which is 1582. IS10 encoder consist of two fully connected layers and each layer perform batch normalization followed by drop out and ReLU activation function. IS10 encoder calculate high level feature embedding as:

$$\begin{aligned} h_6 &= \text{ReLU}(\text{dropout}(W_6\text{BN}(V_6x_i + c_6) + b_6)) \\ h_7 &= \text{ReLU}(\text{dropout}(W_7\text{BN}(V_7h_6 + c_7) + b_7)) \end{aligned} \quad (6)$$

### 3.3. Emotion and Speaker Classifiers

Emotion and speaker classifiers consist of two fully connected layers and each layer perform batch-normalization and dropout, and is processed with ReLU activation function. Output layer is a softmax layer.

Emotion classifier predict the emotion labels  $y_e$  from the stats pooling layer output  $h_5$  and IS10 Encoder output  $h_7$  as:

$$\begin{aligned} h_8 &= \text{concat}(h_5, h_7) \\ h_9 &= \text{ReLU}(\text{dropout}(W_9\text{BN}(V_9h_8 + c_9) + b_9)) \\ h_{10} &= \text{ReLU}(\text{dropout}(W_{10}\text{BN}(V_{10}h_9 + c_{10}) + b_{10})) \\ y_e &= \text{softmax}(W_{11}h_{10} + b_{11}) \end{aligned} \quad (7)$$

where `concat` function concatenate features along the feature dimension. Speaker classifier predict speaker labels  $y_s$  from the stats pooling layer output  $h_5$  as:

$$\begin{aligned} h_{12} &= \text{relu}(\text{dropout}(W_{12}\text{BN}(V_{12}h_5 + c_{12}) + b_{12})) \\ h_{13} &= \text{relu}(\text{dropout}(W_{13}\text{BN}(V_{13}h_{12} + c_{13}) + b_{13})) \\ y_s &= \text{softmax}(W_{14}h_{13} + b_{14}) \end{aligned} \quad (8)$$

### 3.4. Optimization

We minimize the cross-entropy loss of the emotion classifier as:

$$J_e = -E_{x_m}[d_e \log(y_e)] \quad (9)$$

where  $d_e$  is the emotion labels and  $E_{x_m}$  is the expectation over data  $x_m$ . We minimize the emotion triplet loss  $J_{te}$  of the emotion classifier as:

$$J_{te} = E_{x_m} [ \|h_{10_{pos}} - h_{10}\|_2^2 - \eta \|h_{10_{neg}} - h_{10}\|_2^2 ] \quad (10)$$

where  $h_{10_{pos}}$  and  $h_{10_{neg}}$  are the  $h_{10}$  embedding of positive and negative samples respectively. The cross-entropy loss  $J_s$  of speaker classifier is given by:

$$J_s = -E_{x_m}[d_s \log(y_s)] \quad (11)$$

where  $d_s$  are the speaker identity labels. We minimize the speaker triplet loss  $J_{ts}$  of the speaker classifier as:

$$J_{ts} = E_{x_m} [ \|h_{13_{pos}} - h_{13}\|_2^2 - \eta \|h_{13_{neg}} - h_{13}\|_2^2 ] \quad (12)$$

where  $h_{13_{pos}}$  and  $h_{13_{neg}}$  are the  $h_{13}$  embedding of positive and negative samples respectively.

DAL is trained by minimizing the losses of emotion classifier  $J_e$  and  $J_{te}$  and maximizing the losses of speaker classifier  $J_s$  and  $J_{ts}$ . Maximizing speaker classifier losses is archived by multiplying the gradients with a negative scalar  $\lambda$  and updating the weights in the reverse direction during backpropagation which generate discriminative speaker-invariant embedding.

The overall loss that is minimized during learning is given by:

$$J = J_e + \beta J_{te} - \lambda(J_s + \beta J_{ts}) \quad (13)$$

where  $\lambda = \frac{2}{1 + \exp(-\gamma p)} - 1$ ,  $\gamma$  is a positive annealing hyper parameter, and  $p$  is the percentage of training. As gradient reversal strategy introduce instability in the early stages of training, gradient update factor of the speaker classifier  $\lambda$  gradually increase from 0 to at most 1 as the training progress [5]. The rate of changing gradient update factor  $\lambda$  during training can be adjusted by  $\gamma$ , where a higher  $\gamma$  value will result in a higher rate of change.  $\beta$  is the importance factor of the triplet losses  $J_{te}$  and  $J_{ts}$ .

## 4. Experiments and Results

In this section, we investigate the efficacy of the DAL network encoder on two benchmark datasets for SIER: Emo-DB [17] and RAVDESS datasets [18]. The performance of the our network is compared with two existing methods that uses AL for SIER: (i) 1D time dilated neural network (TDNN) [8]; and (ii) 1D CNN [9].

All the experiments were carried out on server with Xeon W-2295 clocked at 3.0 GHz, 256 GB RAM and four Nvidia 2080Ti 12 GB graphic cards running Ubuntu 18.04. The scripts were written in Python using Pytorch package.

### 4.1. Datasets and feature extraction

Emo-DB dataset contain 535 speech emotion collected from 10 German speaking speakers performing 7 emotions. These emotions are anger, boredom, disgust, fear, happy, sad and neutral. We performed 5-fold leave-two-speakers-out cross-validation and testing, where in each fold, 6 speakers were selected for training, 2 speakers for validation and 2 speakers for testing. The ratio of male to female speakers in training, validation, and testing was set to 1:1.

RAVDESS dataset contain 1440 speech emotion collected from 24 professional actors with American English accent performing 8 emotions. These emotions are anger, calm, surprised, fear, happy, sad, neutral and disgust. We performed 12-fold leave-two-speakers-out cross-validation and testing, where in each fold 20 speakers were selected for training, 2 for validation and 2 for testing. The ratio of male to female speakers in training, validation and testing was set to 1:1.

Speech samples was trimmed to remove silence, filtered to remove background noise and padded with zeros so that each sample had the same length as the longest speech sample in the dataset. 40 MFCC features of the speech emotion samples were generated with frame size of 2048 frames, hop size of 512 frames, and Hann Window of 2048 frames. We also generated IS10 features with a length of 1582 using the opensmile package [16] from the preprocessed speech emotion samples.

### 4.2. Parameter initialization

Grid search method was used for selecting the hyper-parameters of DAL network which produced the best cross-validation accuracy. We chose  $\gamma$  from [1.25, 2.5, 3.33] and the number of layers from 1 to 4 in 2D CNN with biGRU encoder. We used convolutional filter size from [5 × 5, 2 × 2], number of convolutional filters [128, 64], number of hidden neurons in GRU [256, 128] and the number of hidden neurons in the fully connected layer [128, 64]. We set the convolutional stride to 2 and the max pooling size 2 × 2 and stride pooling stride of 2. We

Table 1: Comparison of leave-two-speaker-out CV accuracy (%) of AL algorithms for SIER.

Algorithm	Features	Emo-DB	RAVDESS
TDNN[8]	MFCC	61.4±8.7	52.4±10.9
1D CNN[9]	LogMFB	44.2±4.7	30.5±4.6
DAL	MFCC	77.1±5.5	58.1±9.2
DAL	MFCC, IS10	<b>81.6± 4.7</b>	<b>61.7±9.2</b>

Table 2: Comparison of leave-two-speaker-out CV accuracy (%) of AL+center algorithms for SIER.

Algorithm	Features	Emo-DB	RAVDESS
AL	MFCC, IS10	77.2±9.7	59±8.4
AL+center	MFCC	75±6.5	58.4±7.9
AL+center	MFCC, IS10	78.8±5.6	58.1±8
DAL	MFCC, IS10	<b>81.6± 4.7</b>	<b>61.7±9.2</b>

chose  $\beta$  from  $[1, 1e-1, 1e-2, 1e-3, 1e-4, 1e-5]$  and the best DAL hyper-parameters for Emo-DB and RAVDESS dataset is  $\gamma = 2.5, \beta = 1e - 2$ , filter size  $2 \times 2$ , convolutional filters 64, biGRU hidden neurons 256.

#### 4.3. Comparison with AL algorithms

Table 1 shows that for Emo-DB and RAVDESS datasets proposed DAL network outperform TDNN and 1D CNN AL networks by at least 9.3%. Furthermore, for Emo-DB and RAVDESS datasets DAL network using MFCC features only outperform TDNN and 1D CNN AL networks by at least 5.7%. This shows that proposed DAL networks discriminative speaker-invariant embedding is beneficial for SIER.

#### 4.4. Comparison with center loss

We also replaced the triplet loss with center loss to determine the efficacy of center loss, as for SER algorithms it has been shown that center loss is more suitable than triplet loss to learn a discriminative embedding [13]. To this end we compare DAL with 'AL+center network' that has the same network architecture as DAL but use center loss instead of triplet loss to learn discriminative speaker invariant embedding. Table 2 shows that for Emo-DB and RAVDESS datasets DAL network outperform AL+center network by 3.6%. Furthermore, for RAVDESS dataset AL+center network perform in par to AL+center network using MFCC features only. While for Emo-DB dataset AL+center network outperform AL+center network using MFCC features only by 3.8%. This shows that center loss is not suitable learn discriminative speaker-invariant embedding from both low level and high level features together. Furthermore, we see that for Emo-DB and RAVDESS datasets DAL network outperform AL network (without triplet or center losses) with MFCC and IS10 features by 4.4% and 2.7% respectively.

### 5. Conclusions

Traditional approach is to jointly minimize center and cross-entropy losses to learn a discriminative embedding for SER. This paper shows that this traditional approach is not suitable for learning discriminative speaker invariant embedding for SIER and propose DAL which jointly minimizes triplet and cross-entropy losses and outperform other AL algorithms in Emo-DB and RAVDESS datasets. DAL network can learn a

discriminative speaker-invariant embedding with both MFCC and IS10 features together, while traditional approach can learn a discriminative speaker-invariant embedding using MFCC features only.

### 6. Acknowledgements

This work was supported by the Singapore National Research Foundation, Competitive Research Program, under Grant NRF-CRP18-2017-02.

### 7. References

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.
- [2] S. Sahu, R. Gupta, G. Sivaraman, C. Espy-Wilson, and W. AbdAlmageed, "Adversarial Auto-Encoders For Speech Based Emotion Recognition," in *Proceedings of INTERSPEECH*, 2017.
- [3] S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Augmenting Generative Adversarial Networks for Speech Emotion Recognition," in *Proceedings INTERSPEECH*, 2020, pp. 521–525.
- [4] S. E. Eskimez, D. Dimitriadis, R. Gmyr, and K. Kumanati, "GAN-Based Data Generation for Speech Emotion Recognition," in *Proceedings INTERSPEECH*, 2020, pp. 3446–3450.
- [5] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-Adversarial Training of Neural Networks," *Journal of Machine Learning Research*, vol. 17, no. 1, p. 2096–2030, Jan 2016.
- [6] M. Abdelwahab and C. Busso, "Domain Adversarial for Acoustic Emotion Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, 04 2018.
- [7] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, and G. Hofer, "Analysis of Deep Learning Architectures for Cross-Corpus Speech Emotion Recognition," in *Proceedings of INTERSPEECH*, 2019, pp. 1656–1660.
- [8] M. Tu, Y. Tang, J. Huang, X. He, and B. Zhou, "Towards adversarial learning of speaker-invariant representation for speech emotion recognition," 2019.
- [9] H. Li, M. Tu, J. Huang, S. Narayanan, and P. Georgiou, "Speaker-Invariant Affective Representation Learning via Adversarial Training," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7144–7148.
- [10] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proceedings of INTERSPEECH*, sep 2010.
- [11] S. Hizlisoy, S. Yildirim, and Z. Tufekci, "Music emotion recognition using convolutional long short term memory deep neural networks," *International Journal of Engineering Science and Technology*, vol. 24, no. 3, pp. 760–767, 2021.
- [12] J. Huang, Y. Li, J. Tao, and Z. Lian, "Speech Emotion Recognition from Variable-Length Inputs with Triplet Loss Function," in *Proceedings of INTERSPEECH*, 2018, pp. 3673–3677.
- [13] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning Discriminative Features from Spectrograms Using Center Loss for Speech Emotion Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 7405–7409.
- [14] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large Scale Online Learning of Image Similarity Through Ranking," *Journal of Machine Learning Research*, vol. 11, no. 36, pp. 1109–1135, 2010.
- [15] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition," in *European Conference on Computer Vision*, 2016, pp. 499–515.

- [16] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile," in *Proceedings of the international conference on Multimedia*. ACM Press, 2010.
- [17] F. Burkhardt, A. Paeschke, M. A. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proceedings of INTERSPEECH*, 2005, pp. 1517–1520.
- [18] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, 05 2018.