ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447616

GRADIENT-BASED DIMENSIONALITY REDUCTION FOR SPEECH EMOTION RECOGNITION USING DEEP NETWORKS

*Hongxuan Wang, Prahlad Vadakkepat**

Department of Electrical and Computer Engineering,
National University of Singapore, Singapore

ABSTRACT

This paper introduces a gradient-based approach for reducing the dimensionality of acoustic features, tailored for supervised deep learning models used in speech emotion recognition (SER). This method allows us to pinpoint the crucial acoustic features that the network heavily relies on, enabling us to simplify and retrain the network accordingly. It significantly boosts testing speed, making real-time SER systems suitable for embedded systems with resource constraints in speech processing units. The proposed method is evaluated on four convolutional neural network (CNN)-based deep learning models, and one of the best results demonstrates a 56.96% reduction in test time, albeit with a minor 3.81% drop in test accuracy. The method is compared with three mainstream dimensionality reduction techniques across various dimensions, consistently outperforming them in most scenarios. A Python implementation of the method is available at `https://github.com/hxwangnus/Grad-based-Dim-Red-for-SER.git`.

***Index Terms*—** speech emotion recognition, dimensionality reduction, deep neural networks

1. INTRODUCTION

With the advancement of human-machine interaction technology, we've witnessed the emergence of various dialogue systems in our daily lives, such as Alexa, Siri, and Cortana [1]. Ensuring machines can perceive human emotions is crucial for making these human-machine interactions feel natural [2], and one practical approach towards achieving this is through Speech Emotion Recognition (SER) systems. Past studies have demonstrated the value of SER systems across several domains [3, 4, 5].

In pursuit of improved SER performance, researchers have explored diverse acoustic features and classification methodologies [6]. For instance, Eyben et al. [7] introduced the Geneva Minimalistic Acoustic Parameter Set (GeMAPS), initially comprising 62 features but later extended to 88 for further investigations. Researchers have also incorporated various salient, adieu, and spectral features to enhance recognition accuracy in different SER systems [8, 9, 10]. As for classification methods, linear classifiers such as Bayesian Networks, Maximum Likelihood Principle, and Support Vector Machines (SVM), as well as nonlinear classifiers like Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM), have been considered [11]. In 2017, Badshah et al. [12] proposed using deep Convolutional Neural Networks (CNN) to learn spectrograms for SER. Their network featured three convolutional layers and three fully connected layers, achieving an average accuracy of 84.3% on the Berlin dataset. Unlike traditional classifiers, deep learning models don't suffer from increasing model size with larger training datasets, making them more suitable for embedded systems. In 2021, Han et al. [13] introduced the attention mechanism to CNN, integrating Resnet-18 [14], a four-layer CNN, and four multi-head Transformer encoders into a parallel network, achieving an 80.89% accuracy on the RAVDESS dataset [15]. The trend has also shifted towards multimodal deep learning models combining video and speech for training emotion recognition systems [16, 17], as well as developing specialized models for cross-language emotion recognition [18].

Despite the continuous improvement in SER system accuracy through deep learning techniques, this progress has introduced certain challenges. Many studies rely on empirical selections of features and models, lacking a systematic approach. Complex features and models often result in better convergence but at the cost of increased execution time and memory usage. To address the efficiency of intricate deep learning models, Samek et al. [19] delved into visualizing deep neural networks to understand and enhance trained models. In computer vision, Zeiler et al. [20] visualized feature gradients to identify and leverage important pixels for improving model performance. Bertero et al. [21] acknowledged the need for similar investigations in emotion detection tasks, providing insights into model activations in time and frequency domains, albeit without substantial improvements in model performance.

To enhance the efficiency of SER models, we introduce a novel gradient-based dimensionality reduction method designed for acoustic features. The key contributions of our work can be summarized as follows:

- This method enables the systematic selection of pivotal speech emotional features for outcome prediction of SER, and subsequently facilitates dimensionality reduction. We evaluate this method on four CNN-based deep learning models, demonstrating a significant reduction in recognition time without sacrificing much accuracy.
- We conduct a comparative analysis, pitting our proposed approach against three widely recognized mainstream dimensionality reduction methods, across various dimensions. Our method consistently outperforms these alternatives in most cases, reinforcing its efficacy as a valuable tool in enhancing the efficiency of SER models.

2. DIMENSIONALITY REDUCTION METHOD

Fig. 1 depicts the flowchart illustrating our gradient-based dimensionality reduction method. The circles labeled with serial numbers 1 to 4 correspond to four distinct steps in the process. In the initial step, we compute the activated feature maps for each target emotion

*Corresponding Author.

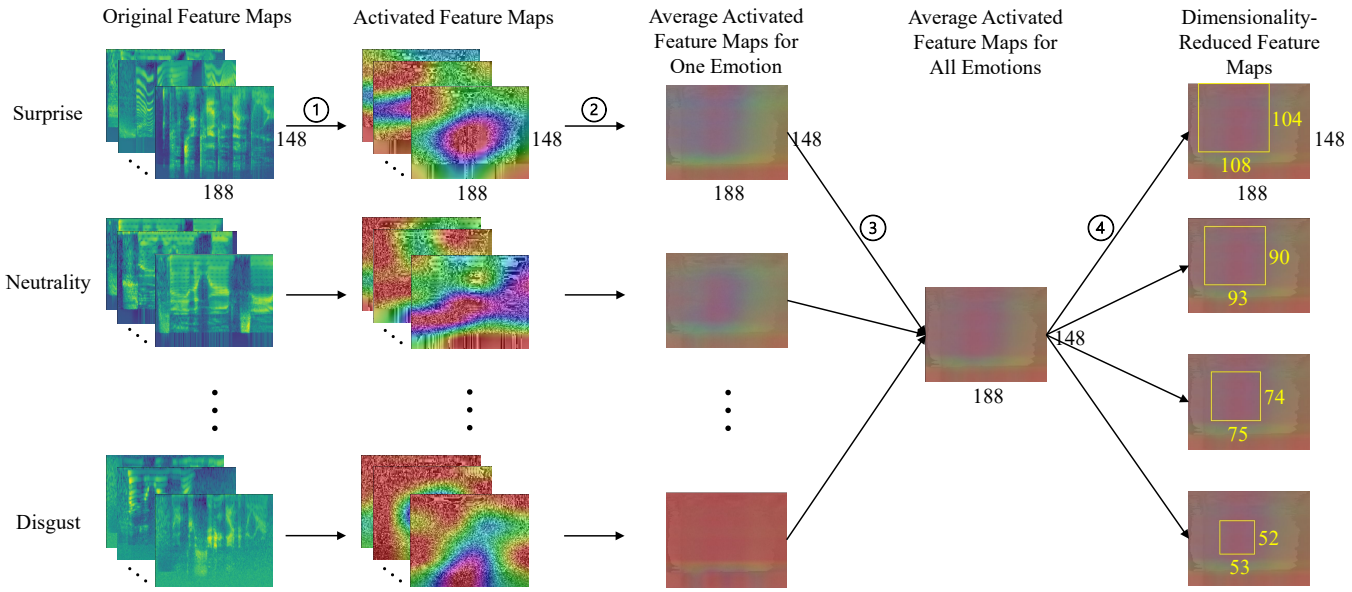

Fig. 1: Flowchart of the proposed gradient-based dimensionality reduction method.

using the original feature maps, relying on the gradient of the prediction loss. However, the activation of a single speech segment alone fails to capture the overall activation of the target emotion. Consequently, in the second step, we compute the average of all activated feature maps related to the target emotion, resulting in an average activated feature map for that specific emotion, denoted as $\bar{F}_t$. $\bar{F}_t$ offers valuable insights into how the model responds to the particular emotion in terms of feature activation.

Moving to Step 3, we extend this averaging procedure to encompass all target emotions, creating an average activated feature map that encompasses all emotions, denoted as $\bar{F}_{all}$. $\bar{F}_{all}$ encapsulates the model's comprehensive activation response throughout the entire SER task.

Finally, in the fourth and final step, we pinpoint the region within $\bar{F}_{all}$ with the highest numerical sum, signifying the portion that significantly influences the model's prediction outcomes. This selected area then serves as the new input feature map, effectively achieving dimensionality reduction.

2.1. Acoustic features activation (step 1 - 3 in Fig. 1)

In this paper, we employ CNN-based deep learning models for SER. The number of output neurons in the final fully connected layer of the SER model matches the count of emotion categories present in the training dataset. To elaborate further, let's consider $z = [z_1, z_2, ..., z_n]$ be the neuron outputs, where n signifies the number of emotion categories within the training dataset. In Step 1 of our method, we calculate the activated feature maps for a specific target emotion (e_t, where t falls within the range $[1, n]$), based on the gradients derived from the original feature maps.

To delve into the process, let's introduce the variable m representing the number of original audio clips associated with the target emotion. Each audio clip undergoes processing through a CNN, and their feature maps are forwarded through the network, with subsequent loss propagation via back-propagation techniques [22]. If we denote the number of input channels for a particular CNN layer as p and the number of output channels as q, this layer will generate a total of q feature maps. The gradient of the feature map within the k^{th} output channel, designated as G^k, is computed during back-propagation by taking partial derivatives of the feature map A^k with respect to z_t:

$$G^k = \frac{\partial z_t}{\partial A^k}, \tag{1}$$

where the index k falls within the range $[1, q]$. Typically, in CNN architectures, shallow layers are specialized in capturing fine-grained details, while deeper layers are geared towards capturing broader, global information [23]. To maximize the incorporation of global information, we calculate gradients using the deepest convolutional layer within the CNN, specifically referring to the last conv2D block within the CNN module.

To further enhance our capacity to capture global information while conserving computational resources, we implement global average pooling on G^k across its height and width, which are denoted by the indices h and w:

$$s_t^k = \frac{1}{HW} \sum_h \sum_w G_{hw}^k, \tag{2}$$

where we define $s_t = [s_t^1, s_t^2, ..., s_t^q]$, with s_t^k representing the average sensitivity between the target emotion e_t and the k^{th} feature map [24].

The two-dimensional activated feature map for the i^{th} audio clip within the target emotion is expressed as follows:

$$F_t^i = \sum_{k=1}^{q} s_t^k \times A^k. \tag{3}$$

We proceed to compute the activated feature maps for all audio clips within the target emotion e_t. Moving on to Step 2, these maps are then averaged to obtain the average activated feature map specifically tailored to the target emotion:

$$\bar{F}_t = \frac{1}{m} \sum_{i=1}^{m} F_t^i. \tag{4}$$

Continuing to Step 3, we take the average of $\bar{F}_t$ across all emotion categories, resulting in the creation of the average activated feature map encompassing all emotions:

$$\bar{F}_{all} = \frac{1}{n}\sum_{t=1}^{n}\bar{F}_t. \quad (5)$$

2.2. Feature selection (step 4 in Fig. 1)

While the initial three steps primarily focus on acquiring global information, the fourth step shifts its emphasis towards local information to prevent loss. Here, we employ convolution operations to pinpoint acoustic features exhibiting strong activation, characterized by large values within $\bar{F}_{all}$, which are crucial for the models to make accurate predictions.

The dimensionality-reduced feature map is configured as a rectangle with a specified height (x) and width (y). To achieve this, we define a convolution kernel as a rectangle with dimensions matching the chosen height and width (x and y), with each element in the kernel set to 1. The convolution process takes place between this kernel and $\bar{F}_{all}$, utilizing a unit stride. As a result, a new matrix is generated post-convolution, and the coordinates of the maximum value within this matrix, denoted as (a^*, b^*), represent the upper-left boundary coordinates of the desired dimensionality-reduced feature map.

If we consider an element in $\bar{F}_{all}$ situated at row a and column b with a value of f_{ab}, then the coordinates of the optimal point, (a^*, b^*), are calculated as follows:

$$(a^*, b^*) = argmax_{a,b}\sum_{a}^{a+x}\sum_{b}^{b+y} f_{ab} \times 1. \quad (6)$$

Following the completion of the fourth step, the original feature maps undergo dimension reduction for testing purposes. These dimension-reduced feature maps are depicted as yellow boxes in Fig. 1, with their corresponding dimensions indicated alongside. The reason to select a continuous local area rather than a feature map composed of discrete points is not only due to a heightened focus on local information in the fourth step, but also because speech is continuous time-series data and the horizontal axis in the Mel-spectrogram represents time, making the selection of a continuous area more logical than discrete points. Additionally, in the "Activated Feature Maps" depicted in Fig. 1, the purple regions indicate features with the highest activation values, most of which are continuous areas rather than discrete points.

3. EXPERIMENTS

3.1. Emotional speech dataset preparation

To enhance the generalization capability of our proposed method across different datasets, we created a composite dataset by amalgamating three widely used speech-emotion datasets: RAVDESS [15], SAVEE [25], and TESS [26]. These datasets exhibit variations in audio lengths and starting frames. Specifically, the RAVDESS dataset comprises 1440 audio clips contributed by 24 actors (12 females and 12 males). The SAVEE dataset includes 480 audio clips from 4 male actors, while the TESS Dataset contains 2800 audio clips from 2 female actresses. In total, our combined dataset encompasses 4720 audio clips, encompassing eight distinct emotions: calmness, happiness, sadness, anger, fear, surprise, disgust, and neutrality.

We utilized the "librosa" library for each audio clip to eliminate any silent segments. Subsequently, we selected a 3.5-second segment commencing from the new starting frame as the input, ensuring uniform segment lengths. The dataset was partitioned into training, validation, and test sets following an 80:10:10 ratio [13]. We performed data augmentation on the training set to mitigate overfitting, by adding Gaussian noise to the original audio clips to generate two additional samples. Consequently, the augmented training set contained 11,313 audio clips, the validation set comprised 471 clips, and the test set encompassed 478 clips.

For all audio clips, we computed Mel-spectrograms as the original feature maps. Mel-spectrogram is a type of spectral feature containing part of prosodic and segmental information. These spectrograms were of dimensions 148×188, as illustrated in Fig. 1.

3.2. Training with original feature maps

To evaluate the generalization capability of our proposed method across neural network sizes, we employed four CNN-based deep learning models with varying levels of complexity. The most intricate model, referred to as CTRL, boasts the highest number of parameters. CTRL is a parallel network comprising four distinct components: a four-layer CNN (C), four multi-head self-attention Transformer encoders (T), Resnet-18 (R), and a two-layer bidirectional LSTM (L). Remarkably, CTRL encompasses over 10 million trainable parameters, making it one of the most intricate SER networks known to us.

To derive the remaining models, we pruned CTRL. Initially, we removed the Resnet block, resulting in CTL, a parallel network consisting of four-layer CNN, four multi-head self-attention Transformer encoders, and a two-layer bidirectional LSTM. Subsequently, we eliminated the LSTM block to obtain CT, which comprises a parallel network composed of four-layer CNN and four multi-head self-attention Transformer encoders. Lastly, by removing the Transformer block, we obtained a four-layer CNN network.

All four models underwent training using the original feature maps, and we subsequently recorded their test time and accuracy. Hardware consistency is maintained by employing the same single CPU for testing. In practical applications, batch inference time can be used to eliminate the impact of hardware variations.

3.3. Dimensionality reduction

Following the acquisition of the original feature maps and the trained deep learning models, we proceeded to implement steps 1 through 4 as described in Section 2. Notably, in step 4, we made selections for the dimensions of the dimensionality-reduced feature maps, which were set to be 40%, 30%, 20%, and 10% of the original feature maps. Consequently, this led to specific width and height pairs for the reduced feature maps, namely $(104, 108)$, $(90, 93)$, $(74, 75)$, and $(52, 53)$, respectively. These dimensionality-reduced feature maps are visually represented as the yellow rectangles in Fig. 1.

For the purpose of comparison, we also implemented three widely recognized traditional dimensionality reduction methods to reduce the dimensions of the original feature maps to the same dimensions of $(104, 108)$, $(90, 93)$, $(74, 75)$, and $(52, 53)$. These three methods include Principal Component Analysis (PCA) [27], Autoencoders (AE) [28], and Uniform Manifold Approximation and Projection (UMAP) [29].

3.4. Retraining with reduced feature maps

To align with the dimensions of the reduced feature maps, we fine-tuned the parameters of the trained models as minimally as possible, while keeping the hyperparameters unchanged. This fine-tuning involved reducing the kernel sizes, adjusting the pooling layers' stride,

and modifying the dimensions of linear layers. Subsequently, we retrained these adapted models using the reduced feature maps obtained from both our proposed method and the three mainstream methods. Following this retraining, we recorded both the test time and accuracy for evaluation.

4. RESULTS AND ANALYSIS

Table 1 shows the test time and accuracy of the four CNN-based deep learning models trained with original feature maps. The test time refers to the total time for the model to recognize the emotion of the 478 speech clips in the test set.

Table 1: Test time (s) and accuracy $(\%)$ of the four CNN-based models trained with original feature maps

Model	CTRL	CTL	CT	CNN
Time	0.411	0.158	0.11	0.083
Accuracy	80.75	82.50	81.38	78.87

4.1. Comparison with original models

Tables 2 and 3 provide a comparative analysis of test time and test accuracy between models trained with original feature maps (referred to as "original models" below) and models trained with dimensionality-reduced feature maps utilizing our proposed method (referred to as "reduced models" below). To effectively illustrate the impact of our method, we present the results as either time reduction or accuracy loss concerning the original models and the reduced models. This is expressed as:

$$Loss_X = \frac{X_{original} - X_{reduced}}{X_{original}} \times 100\%, \tag{7}$$

where, X represents test time in Table 2 and test accuracy in Table 3.

For all models and dimensions, except for CNN with dimension $(104, 108)$ and CTL with dimension $(52, 53)$, the reductions in test time exceed 50% while incurring a maximum accuracy loss of no more than 13%. This demonstrates the significant enhancement in recognition efficiency achieved by our method. Notably, when we reduce the original feature maps to the same dimensions, the trade-off between test accuracy loss and test time reduction remains consistent across all models, illustrating the generalizability of our method across models with varying complexities.

Table 2: Reduction in test time $(\%)$, between original models and reduced models.

Dimension	CTRL	CTL	CT	CNN
$(104, 108)$	53.53	56.96	50.00	40.96
$(90, 93)$	63.26	70.89	58.18	53.01
$(74, 75)$	70.80	74.68	65.45	71.08
$(52, 53)$	86.62	86.71	81.82	84.34

4.2. Comparison with mainstream methods

Table 4 presents a comparison of test accuracy between models trained with dimensionality-reduced feature maps using our method and those using three mainstream methods.

Table 3: Loss in test accuracy $(\%)$, between original models and reduced models.

Dimension	CTRL	CTL	CT	CNN
$(104, 108)$	2.41	3.81	4.29	2.74
$(90, 93)$	6.82	9.47	7.54	3.00
$(74, 75)$	9.15	12.01	8.74	5.39
$(52, 53)$	12.17	19.95	12.51	9.81

Across all models and dimensions, except for CTRL with dimension $(52, 53)$, our method consistently achieves the highest test accuracy after dimensionality reduction. This outcome underscores the superior performance of our method compared to the three mainstream dimensionality reduction methods in the context of CNN-based SER networks.

Table 4: Comparison of test accuracy $(\%)$, between models reduced by mainstream methods and models reduced by our method.

Model	Dimension	PCA	AE	UMAP	Ours
CTRL	$(104, 108)$	55.16	73.57	56.69	**78.80**
	$(90, 93)$	56.69	70.43	57.46	**75.24**
	$(74, 75)$	56.35	71.34	57.25	**73.36**
	$(52, 53)$	61.37	**71.27**	62.27	70.92
CTL	$(104, 108)$	44.21	59.48	48.26	**79.36**
	$(90, 93)$	52.93	59.14	49.09	**74.69**
	$(74, 75)$	53.35	60.81	52.23	**72.59**
	$(52, 53)$	54.60	64.57	60.46	**66.04**
CT	$(104, 108)$	57.39	61.79	49.02	**77.89**
	$(90, 93)$	54.67	60.39	54.25	**75.24**
	$(74, 75)$	53.56	60.95	54.88	**74.27**
	$(52, 53)$	55.58	64.64	59.14	**71.20**
CNN	$(104, 108)$	60.95	64.44	55.72	**76.71**
	$(90, 93)$	57.95	65.83	55.09	**76.50**
	$(74, 75)$	60.18	65.34	54.11	**74.62**
	$(52, 53)$	61.30	66.39	58.79	**71.13**

5. CONCLUSIONS

In this paper, we introduce an innovative gradient-based approach for dimensionality reduction of acoustic features. Our method effectively reduces the dimensionality of original feature maps by considering feature activation strength. The approach significantly enhances the efficiency of the recognition process and exhibits adaptability to CNN-based models of varying complexities. Furthermore, it outperforms three mainstream dimensionality reduction methods across various dimensions, making it suitable for real-time SER systems on low-performance embedded systems.

Looking ahead, our future research endeavors will concentrate on two primary avenues. Firstly, we intend to evaluate the method on diverse datasets to confirm its generalization capability. Secondly, we plan to extend the method to encompass non-CNN-based networks, broadening its applicability and utility in the realm of SER.

6. ACKNOWLEDGEMENTS

Thanks to Shandong Hanju Machinery Manufacturing Co., Ltd. for supporting the research and to Binghui Wu for useful discussions.

7. REFERENCES

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM (CACM)*, vol. 61, no. 5, pp. 90–99, 2018.

[2] J. Kim and R. A. Saurous, "Emotion recognition from human speech using temporal information and deep learning," in *INTERSPEECH*, 2018, pp. 937–940.

[3] Q. Luo and H. Tan, "Facial and speech recognition emotion in distance education system," in *The 2007 International Conference on Intelligent Pervasive Computing (IPC 2007)*, Jeju, Korea (South), 2007, pp. 483–486.

[4] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, "Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 342–356, 2021.

[5] S. N. Zisad, M. S. Hossain, and K. Andersson, "Speech emotion recognition in neurological disorders using convolutional neural network," in *International Conference on Brain Informatics*, 2020, pp. 287–296.

[6] M. Ayadi, M. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[7] F. Eyben 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, 1 April-June 2016.

[8] 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, Dec. 2014.

[9] G. Trigeorgis et al., "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5200–5204.

[10] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using fourier parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69–75, 1 Jan.-March 2015.

[11] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[12] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 International Conference on Platform Technology and Service (PlatCon)*, Busan, Korea (South), 2017, pp. 1–5.

[13] S. Han, F. Leng, and Z. Jin, "Speech emotion recognition with a resnet-cnn-transformer parallel neural network," in *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*, Beijing, China, 2021, pp. 803–807.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[15] 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," `https://doi.org/10.1371/journal.pone.0196391`, 2018.

[16] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "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, Dec. 2017.

[17] P. Singh, R. Srivastava, K. P. S. Rana, and V. Kumar, "A multimodal hierarchical approach to speech emotion recognition from audio and text," *Knowledge-Based Systems*, vol. 229, 2021.

[18] S. Ntalampiras, "Toward language-agnostic speech emotion recognition," *Journal of the Audio Engineering Society*, vol. 68, no. 1/2, pp. 7–13, 2020.

[19] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, Nov. 2017.

[20] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 818–833.

[21] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 5115–5119.

[22] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[23] P. T. Jiang, C. B. Zhang, Q. Hou, M. M. Cheng, and Y. Wei, "Layercam: Exploring hierarchical class activation maps for localization," *IEEE Transactions on Image Processing*, vol. 30, pp. 5875–5888, 2021.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[25] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," `http://kahlan.eps.surrey.ac.uk/savee/`, April 2, 2015.

[26] K. Dupuis and M. K. Pichora-Fuller, "Toronto emotional speech set (tess)," `https://doi.org/10.5683/SP2/E8H2MF`, 2010.

[27] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.

[28] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, 5 April 2016.

[29] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, pp. 861, 2018.