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

Speech Emotion Recognition 'in the wild' Using an Autoencoder

Vipula Dissanayake[1], *Haimo Zhang*[1], *Mark Billinghurst*[2], *Suranga Nanayakkara*[1]

[1]Augmented Human Lab, Auckland Bioengineering Institute, The University of Auckland
[2]Empathic Computing Lab, Auckland Bioengineering Institute, The University of Auckland
vipula@ahlab.org, haimo@ahlab.org, mark.billinghurst@auckland.ac.nz, suranga@ahlab.org

Abstract

Speech Emotion Recognition (SER) has been a challenging task on which researchers have been working for decades. Recently, Deep Learning (DL) based approaches have been shown to perform well in SER tasks; however, it has been noticed that their superior performance is limited to the distribution of the data used to train the model. In this paper, we present an analysis of using autoencoders to improve the generalisability of DL based SER solutions. We train a sparse autoencoder using a large speech corpus extracted from social media. Later, the trained encoder part of the autoencoder is reused as the input to a long short-term memory (LSTM) network, and the encoder-LSTM modal is re-trained on an aggregation of five commonly used speech emotion corpora. Our evaluation uses an unseen corpus in the training & validation stages to simulate 'in the wild' condition and analyse the generalisability of our solution. A performance comparison is carried out between the encoder based model and a model trained without an encoder. Our results show that the autoencoder based model improves the unweighted accuracy of the unseen corpus by 8%, indicating autoencoder based pre-training can improve the generalisability of DL based SER solutions.

Index Terms: speech emotion recognition, autoencoders, computational paralinguistics.

1. Introduction

Human voice contains emotions embedded in the forms of linguistic and para-linguistic cues. Consequently, Speech Emotion Recognition (SER) has become a significant research interest. In fact, last two decades have seen a growing trend towards human emotion recognition among human-computer interaction researchers, driven by the popularity of voice-based user interfaces [1, 2, 3, 4].

Among the multiple techniques shown in literature to solve SER, Deep Learning (DL) has so far demonstrated the most promising results [5]. However, unlike in other domains such as image recognition, natural language processing, and automatic speech recognition, DL still has limitations to produce a generalisable model for SER. Among the reasons for that, unavailability of large-scale labelled datasets with wider distributions such as "image-net" [1] (computer vision) and the controlled nature of available SER datasets such as IEMOCAP [6] are prominent. This hinders the development of large-scale models for SER due to overfitting issues and models that work 'in the wild' due to the variable-controlled data collection.

In order to address the generalisability issue and the limitation of labelled data for SER, researchers are proposing various approaches. Although creating a considerable-sized corpus consumes a lot of time and effort, there are crowdsourcing procedures such as iHEARu-Play [7], a gamified program for creating significantly large affective corpora. Apart from collecting more labelled data, there are other techniques to capitalise on available data to create robust SER models such as combining available datasets [8, 9, 10], transfer learning using pre-trained DL models for other tasks [11].

Autoencoder based neural network architectures are known to perform well in learning representations from the data. Autoencoder architectures such as variational autoencoders have been used in SER tasks and proven to work well in learning emotional representations [12, 13]. However, autoencoder based architectures are mostly discussed for within-corpus SER prediction tasks. A little exploration has been conducted with autoencoder based architectures in cross-corpus SER tasks.

In this paper, we investigate a methodology to improve the generalisability of a DL model for SER tasks using an autoencoder-based pre-trained neural network. Building upon a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) network structure proposed in recent literature [10], we train a sparse autoencoder(AE) [14] on a substantial speech corpus extracted from publicly available videos in social media to induce the 'in-the-wild' nature. Later the encoder component of the AE network is used along with the trained weights to build a CNN-LSTM model which we train on an aggregation of five emotional speech corpora and evaluate against a separated portion of the aggregated corpora. In order to examine the generalisability of the proposed methodology, we evaluate our trained model on a separate emotional corpus which has not been used in the training stage. Our study includes a comparison of SER performance with and without AE based training. We demonstrate that a pre-trained autoencoder network can improve the generalisability of a DL based SER model in corss-corpus prediction.

2. Related Work

Speech emotion recognition is considered a challenging task in Human-Computer Interaction (HCI) and speech signal processing domain. Researchers have been proposing different methodologies, corpora and theories on SER [15, 1, 16, 17]. The early stage of SER research have used handcrafted speech features such as pitch, energy and jitter of speech and low-level spectral features such as Mel-frequency Cepstral Coefficients (MFCC) to train classic machine learning models namely support vector machines and Hidden Markov Model (HMM), to classify emotions of the selected corpus [2, 18, 3]. There is an increasing trend in most recent literature to use deep neural networks over classic machine learning approaches for the classification, and those researches show the state-of-the-art results in SER. However most DL approaches also use handcrafted speech features, low-level spectral features and raw audio signal as the input to the algorithm.

[1] *http://www.image-net.org/*

http://dx.doi.org/10.21437/Interspeech.2020-1356

Table 1: *Emotion class mapping*

Mapped Class	Original Class					
	IEMOCAP	EMOVO	EmoDB	RAVDESS	SAVEE	TESS
Positive	*happy, joy, surprise*	*joy, surprise*	*happy*	*happy, surprise*	*happy, surprise*	*happy, pleasant surprise*
Negative	*angry, disgust, fear, frustration, sad*	*angry, disgust, fear, sad*	*angry, disgust, fear, sad*	*angry, disgust, fear, sad*	*angry, disgust, fear, sad*	*angry, disgust, fear, sad*
Neutral	*neutral*	*neutral*	*boredom, neutral*	*calm, neutral*	*neutral*	*neutral*

One major issue observed in the DL approaches is the risk of overfitting the model to the training corpus when there is limited amount of labeled data [19]. Hence, it shows better classification results for the corpus that has been used in the training process; still, the classifier fails to predict accurately for other corpora. In other words, the classifier is not generalisable.

To address the scarcity of generalisability of SER algorithms, researchers have explored multiple methodologies. 1) Collecting and annotating new data is a trivial solution for the lack of data. Researchers have been exploring possibilities using crowdsourcing technologies and scraping available data from the internet [7, 20]. However, it is expensive and consumes much time to create a big enough dataset with proper data distribution [21]. 2) Corpora aggregation, where researchers have explored the possibility of combining multiple corpora to expand data distribution and increase the number of labelled data in the training process [9]. 3) Data augmentation, manipulating existing labelled data to generate multiple slightly different copies of available data has been another general solution for fewer annotated data situations which has been tried out in the speech domain [17, 16, 22]. However, Bao et al. [16] have observed a bias in augmented data, which leads the SER model to overfit to that particular bias. 4) Transfer learning can be defined as repurposing already a well-trained model for another task, which has been successfully explored in other domains, such as computer vision and natural language processing. Recently some experiments have been carried out using transfer learning for SER tasks to improve the performance of models [23, 24, 25, 11]. However, Rosenstein et al. [26] argue that dissimilarity between the source and destination task can hurt the performance of the new model. 5) Multi-task learning, instead of training the DL model to learn a single specific task of SER, researchers have tried training a single model to perform multiple tasks. Multi-task learning has also been shown promising results as a solution for overfitting issue for SER tasks [9, 27, 28]. Deng et al. [29] have proposed autoencoder based semi-supervised multi-task learning approach, however method proposed in this paper has taken the temporal and spatial nature of speech signals into consideration by using LSTM & CNN where previous has only focused on spatial nature.

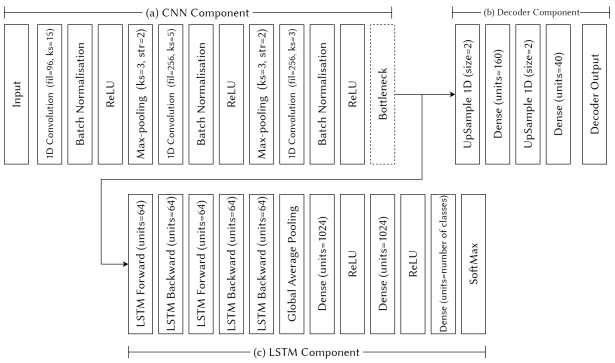


Figure 1: *Network components and architectures. 'fil', 'ks', and 'str' stands for hyperparameters respectively 'filters', 'kernel size', and 'strides'. CNN component and decoder components are compiled together as an autoencoder model, while CNN component and LSTM component are compiled as CNN-LSTM model.*

Researchers have used cross corpora testing to simulate the generalisability of their models and simulate 'in the wild' condition [10, 9]. An exploration of Parry et al. [10] on how the generalisability of SER model varies with the DL model architecture has resulted that a combined architecture of CNN and LSTM provides a more generalisable SER model. As an extension to the combined DL architecture introduced by Parry et al. [10], we explore the impact of pre-training with autoencoder to the generalisability. We are trying to address the effects of limited data using autoencoder and transfer learning; also our approach tries to minimise the mismatch between source and destination of transfer learning stage by using similar kind of data in both modalities.

3. Methodology

3.1. Data

Emo-DB [30], EMOVO [31], IEMOCAP [6], RAVDESS [32], SAVEE [4], TESS [33], and VoxCeleb [34] corpora were used in this study. Following prior work [10], first five corpora were split into training, validation and testing partitions while the TESS corpus was kept aside during training and validation phases. To ensure, training, testing and validation data have no overlapping datapoints, Emo-DB, audio from actor #15 and #16 used as testing and validation partition respectively while audio of rest of the actors used as training partition. In EMOVO, audio of the first two male and female speakers were used as training partition, while data of third male speaker and third female speaker used as validation and testing partitions. Data of the

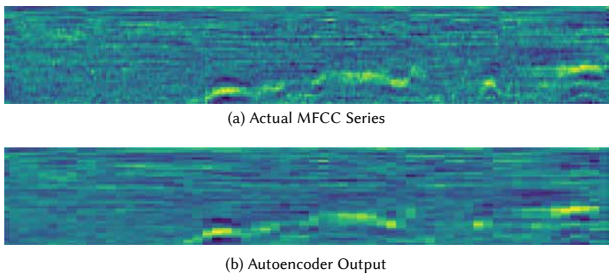


Figure 2: *Input and output to the trained autoencoder network.*

Table 2: *Cross-corpus speech emotion prediction performance. CNN-LSTM ([10]) shows the results published in recent literature [10]. *left-out corpus.*

Model	Unweighted Accuracy (%)							
	IEMOCAP	EMOVO	EmoDB	EPST	RAVDESS	SAVEE	Average	TESS*
CNN-LSTM (Literature)	**50.31**	53.24	**69.72**	51.81	53.08	**72.66**	53.35	49.48
CNN-LSTM (without encoder)	42.72	51.94	69.71	-	50.42	50.42	53.63	50.34
CNN-LSTM (with encoder)	46.79	**56.74**	67.58	-	**56.71**	56.67	**56.90**	**58.06**

first five sessions of IEMOCAP were assigned to the training; validation and testing partitions contained half of the sixth session each. RAVDESS training partition consisted of recordings from the first 22 actors while data of actor #23 and actor #24 were used in testing and validation partitions. Speaker 'DC', 'JE', 'JK' of SAVEE corpus assigned to the training portion while data of actor 'KL' split equally for testing and validation. Training partitions of the first five corpora were aggregated together and randomised to use as the training set while a similar process was applied to the validation portions to create the validation set.

The TESS corpus was not split as it was not used in the training and validation phases and is kept as a particular testing corpus to simulate 'in the wild' condition. Audio clips of the VoxCeleb corpus, which consists of a large set of speech audio clips of celebrities extracted from social media, was used in the training stage of the autoencoder model. No split or data clearance was applied to this corpus.

3.2. Model Architectures

The structure of neural network models used in this experiment was inspired by the CNN-LSTM model proposed by Parry et al. [10], which consumes Mel-Frequency Cepstral Coefficients (MFCC) as the input and the output emotion class. All the network architectures and hyperparameters are presented in Figure 1. The CNN-LSTM model [10] is split into two separate networks, CNN and LSTM; the output of the CNN component (a) is named as the bottleneck layer, while it is fed to a decoder neural network (b) as well as to the LSTM network (c). Newly proposed decoder network consumes the bottleneck layer, and it is designed to output a vector similar to input MFCC of the CNN component. Two combinations of the neural network components compile into two neural network models with a shared network component, as shown in Figure 1. We recognise the first model compiled with CNN and decoder component as autoencoder network (AE), while the latter model compiled with CNN and LSTM components as CNN-LSTM network.

3.3. Experiment Setup

We used Mel-Frequency Cepstral Coefficients (MFCC) as low-level features for both of our models. Our models consumed only 40 MFCCs which were calculated for each input utterance on 25ms window and 10ms of shift. Each feature vector was normalised to zero mean and unit variance in utterance level and padded to a minimum of 100 frames.

All our deep learning models were implemented using Python 3.7 and Tensorflow 2.1[2] framework and optimised with adam optimiser [35] with similar hyperparameters (learning rate=0.001). The autoencoder network was compiled to optimise with mean squared error (MSE) loss function while the CNN-LSTM model was compiled to optimise with categorical cross-entropy loss. The rationale for choosing different loss was to have a suitable optimisation in the particular purpose of training.

To unify the emotion classes defined differently in different corpora, we used, we remapped all emotion class into three high-level categories (positive, negative and neutral). Due to the imbalance of examples for each high-level class, we used unweighted accuracy (mean accuracy of each class) as the evaluation metric for the emotion prediction task.

3.4. Autoencoder Training

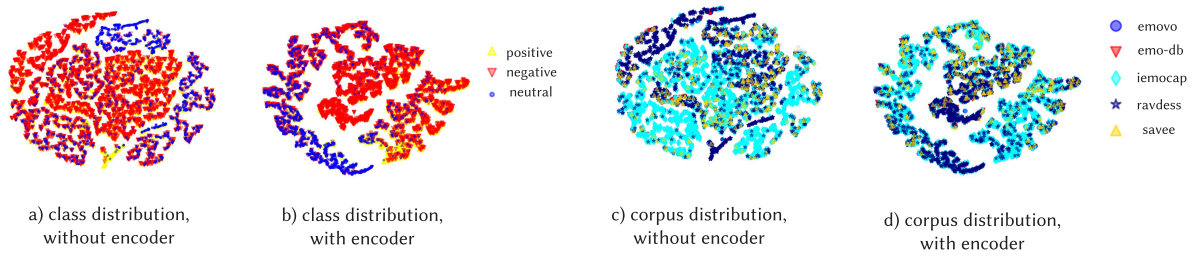
The autoencoder model was trained with a substantial sized speech corpus from VoxCeleb, which had nearly 150,000 uttarances. The rationale of using an autoencoder architecture was to teach the network to learn a representation of generic speech data. The CNN component of the network, which we recognised as the encoder, was optimised by learning a complex representation of the given MFCC series, while the decoder component was optimised to decode the encoder's output back to the corresponding MFCC series. Figure 2 shows output (b) of the trained autoencoder for an input sequence (a) of MFCCs. After the training process, we discard the decoder component and reuse the already learnt encoding representation for SER task.

3.5. Speech Emotion Recognition Training

The CNN-LSTM model was trained to predict emotions. For this training, we used the aggregated training set and the validation set reported in the data section. Since different corpora had different emotion class labels, instead of dropping data, we mapped all those classes to three high-level classes: positive, negative, and neutral, using a mapping inspired by recent literature (see Table 1) [10].

We followed two training strategies: 1) training the CNN-LSTM model from scratch and 2) reusing the already trained CNN component in the autoencoder and training only the LSTM component. Both modalities are trained and validated

[2] *https://www.tensorflow.org/*



a) class distribution, without encoder

b) class distribution, with encoder

c) corpus distribution, without encoder

d) corpus distribution, with encoder

Figure 3: *Distributions in last hidden layer*

with same data partitions and tested against separate partitions for testing, further models were tested against the TESS corpus which had not been used in the training and validation phases.

4. Results & Discussion

4.1. Speech Emotion Recognition Training

We tested out the performance of SER model trained from scratch, which is performed in two stages, 1) within corpora and 2) cross-corpus. Results with and without pre-trained encoder based on VoxCelab are tabulated in Table 2 with reference to a similar study from the literature ('CNN-LSTM (Literature)'). Although EPST [36] corpus was not available for the current study, we maintained a similar condition as much as possible with the published method by Parray et al. As expected, results of within corpora and cross-corpus tests indicate similar results to the results shown in the literature. There exist some mismatches in some test set of some corpora which will be discussed further under discussion.

In addition, evaluated the SER model with pre-trained encoder (third row in Table 2). Results of within corpora test show improved unweighted accuracy over the CNN-LSTM trained from scratch (second row in Table 2). Finally, the cross-corpus testing results have 8% improvement over the without encoder methodology and 8.5% improvement over the results published in the literature.

4.2. Visualisation Study

To understand the reasons for the improved generalisability of the encoder based training method, we present an investigation of visualisation of the vector representation in the last hidden layer of the CNN-LSTM network. t-Distributed Stochastic Neighbour Embedding (t-SNE) [37], a machine learning algorithm for visualisation is used to reduce the dimensions of the final hidden layer to two-dimensions and visualise the vector distribution based on the emotion label and the corpus of origin. Visualisations are shown in Figure 3. Sub-figures (a) and (b) demonstrate how emotional classes are distributed in the from-scratch training and encoder based training, respectively. These two sub-figures reveal some clusters based on classes, which are expected at this stage of the trained neural network. However, the visualisation of encoder-based methodology shows much clearer class cluster boundaries than from-scratch training methodology, which can be related to the improvement of accuracy observed in the results. Corpus-based clustering in Figure 3 (c) and (d) visualise from-scratch training and encoder based training respectively. These two sub-figures indicate some form of grouping, which suggest that the encoder-based CNN-LSTM network has learned a little information about corpora variation. However, distribution does not show significant overfitting to corpora variation, as representations of each corpus distributed over the plane.

4.3. Discussion

Findings of the studies presented in the paper suggest that autoencoder based training methodology has improved the generalisability of SER model compared to same network architecture trained from scratch. A possible explanation for this might be that autoencoder based training expanded the amount of possible data to be used in the training process. As mentioned in the literature, the limited availability of labelled emotional data leads a neural network to overfit and perform poorly to unseen data. In this study, the autoencoder training stage had a substantial amount of spoken utterances compared to the number of utterances available for the emotion recognition task. Although the amount of labelled emotional data was similar in both training conditions, the influence of the pre-trained encoder would have a broader effect on the LSTM component when training for SER.

It was observed that the within-corpus evaluation of SAVEE corpus had poor prediction performance compared to the results published in the literature. This inconsistency may be due to the methodology of training, validation, and testing set splitting. In the literature, splitting was random; however, in this study, we strictly maintained speaker separation across training, validation and testing datasets under an assumption that would grant more space to generalise the model. Effect of an individual variant in training, validation, and testing set could have possibly lead to this inconsistency.

5. Conclusion & Future Work

In this paper, we presented an evaluation of an autoencoder based training methodology for cross-corpus emotion recognition. Our studies demonstrated that pre-trained encoder had improved the generalisability of a speech emotion recognition model and performed comparatively well compared to a model trained from scratch for unseen data distribution. A visualisation study has confirmed improved class boundaries at the final hidden layer of the neural network.

In future, it would be worth to explore how the generalisability of proposed method behave based on 1) the structure of the encoder, 2) targeted learning task of the encoder, and 3) training data augmentation with natural speech data.

6. Acknowledgements

This work is supported by the Assistive Augmentation research grant under the Entrepreneurial Universities (EU) initiative of New Zealand.

7. References

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

[2] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.*, vol. 2. IEEE, 2003, pp. II–1.

[3] Y.-L. Lin and G. Wei, "Speech emotion recognition based on hmm and svm," in *2005 international conference on machine learning and cybernetics*, vol. 8. IEEE, 2005, pp. 4898–4901.

[4] S. Haq and P. J. Jackson, "Multimodal emotion recognition," in *Machine audition: principles, algorithms and systems*. IGI Global, 2011, pp. 398–423.

[5] 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. 117 327–117 345, 2019.

[6] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[7] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "ihearu-play: Introducing a game for crowdsourced data collection for affective computing," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 891–897.

[8] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote," in *in Proc. Interspeech 2011*, 2011.

[9] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Towards speech emotion recognition "in the wild" using aggregated corpora and deep multi-task learning," in *Proc. Interspeech 2017*, 2017, pp. 1113–1117.

[10] 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 *Proc. Interspeech 2019*, 2019, pp. 1656–1660.

[11] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," in *Proc. Interspeech 2018*, 2018, pp. 257–261.

[12] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," in *Proc. Interspeech 2018*, 2018, pp. 3107–3111.

[13] S. Parthasarathy, V. Rozgic, M. Sun, and C. Wang, "Improving emotion classification through variational inference of latent variables," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7410–7414.

[14] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proceedings of ICML workshop on unsupervised and transfer learning*, 2012, pp. 37–49.

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

[16] F. Bao, M. Neumann, and N. T. Vu, "CycleGAN-Based Emotion Style Transfer as Data Augmentation for Speech Emotion Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2828–2832.

[17] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. W. Schuller, "Towards Robust Speech Emotion Recognition Using Deep Residual Networks for Speech Enhancement," in *Proc. Interspeech 2019*, 2019, pp. 1691–1695.

[18] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Speech emotion recognition using hidden markov models," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[19] Z. Zhang, B. Wu, and B. Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6705–6709.

[20] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[21] A. Tarasov, S. J. Delany, and C. Cullen, "Using crowdsourcing for labelling emotional speech assets," in *W3C workshop on Emotion ML*, 2010.

[22] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "Cnn+lstm architecture for speech emotion recognition with data augmentation," in *Proc. Workshop on Speech, Music and Mind 2018*, 2018, pp. 21–25.

[23] M. E. A. Elshaer, S. Wisdom, and T. Mishra, "Transfer learning from sound representations for anger detection in speech," *CoRR*, vol. abs/1902.02120, 2019.

[24] P. Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 265–275, 2019.

[25] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Cross corpus speech emotion classification- an effective transfer learning technique," *CoRR*, vol. abs/1801.06353, 2018.

[26] M. T. Rosenstein, Z. Marx, L. P. Kaelbling, and T. G. Dietterich, "To transfer or not to transfer," in *NIPS 2005 workshop on transfer learning*, vol. 898, 2005, pp. 1–4.

[27] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multi-task learning," in *Proc. Interspeech 2018*, 2018, pp. 951–955.

[28] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proc. Interspeech 2017*, 2017, pp. 1103–1107.

[29] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 31–43, 2017.

[30] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[31] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "Emovo corpus: an italian emotional speech database," in *International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association (ELRA), 2014, pp. 3501–3504.

[32] 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, 2018.

[33] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2010.

[34] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations*, 2015.

[36] M. Liberman and et al., "Emotional prosody speech and transcripts ldc2002s28," *Web Download Philadelphia: Linguistic Data Consortium*, 2002.

[37] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.