Interspeech 2018
2-6 September 2018, Hyderabad

Variational Autoencoders to Learn Latent Representations of Speech Emotion

Siddique Latif[1], *Rajib Rana*[2], *Junaid Qadir*[1], *Julien Epps*[3]

[1]Information Technology University (ITU)-Punjab, Pakistan
[2]University of Southern Queensland, Australia
[3]University of New South Wales, Sydney, Australia

siddique.latif@itu.edu.pk, rajib.rana@usq.edu.au, junaid.qadir@itu.edu.pk, j.epps@unsw.edu.au

Abstract

Learning the latent representation of data in unsupervised fashion is a very interesting process that provides relevant features for enhancing the performance of a classifier. For speech emotion recognition tasks, generating effective features is crucial. Currently, handcrafted features are mostly used for speech emotion recognition, however, features learned automatically using deep learning have shown strong success in many problems, especially in image processing. In particular, deep generative models such as Variational Autoencoders (VAEs) have gained enormous success in generating features for natural images. Inspired by this, we propose VAEs for deriving the latent representation of speech signals and use this representation to classify emotions. To the best of our knowledge, we are the first to propose VAEs for speech emotion classification. Evaluations on the IEMOCAP dataset demonstrate that features learned by VAEs can produce state-of-the-art results for speech emotion classification.

Index Terms: speech emotion classification, Variational Autoencoders, deep learning, feature learning

1. Introduction

Recently speech emotion recognition has received significant attention from both industry and academia. It has various applications in human-computer interaction and analysis of human-human interactions. The speech signal has complex distributions with high variance due to various factors such as speaking style, age, gender, linguistic content, environmental and channel effects, emotional state. Understanding the influence of these factors on the speech signal is a crucial problem for speech emotion recognition. Although considerable attempts have focused on handcrafting features to capture these factors [1], automatic learning of features that are sensitive to emotion needs more exploration.

Deep generative models are recently becoming immensely popular in the deep learning community due to the fact that unlike discriminative approaches, they try to learn the true distribution of the training data and generate new data points (with some variations). In this paper, we are not focused on generating new data but on capitalising the capacity of generative models to learn the true distribution of the data and hence create powerful features, automatically. The most commonly used and efficient generative models are currently Generative Adversarial Nets (GANs) [2] and Variational Autoencoders (VAEs) [3]. While GANs are optimised for generative tasks, VAEs are probabilistic graphical models which are optimised for latent modelling. We therefore focus on VAEs. There have been many attempts to model natural images using generative models [4–6], but only some research has been conducted into learning latent representations of speech generation [7,8], voice conversion [9], and speaker identification [10]. Most importantly, the feasibility of VAEs for speech emotion recognition is largely unexplored.

In this paper, we conduct a preliminary study to understand the feasibility of VAE for learning the latent representation of speech emotion. We also investigate the performance of a variant of VAE known as Conditional Variational Autoencoder (CVAE) [11] for learning the latent representation of speech emotion. To objectively measure the performance of this latent representation, we use Long Short Term Memory (LSTM) to classify speech emotion using the latent representation as features. This simultaneously offers the opportunity to validate the performance of VAE for learning latent representation, and delivers a new VAE-LSTM classification framework. Given that Autoencoders (AE) have been widely used for speech emotion, we implement an AE-LSTM model to compare its classification performance with VAE-LSTM. We also compare the classification performance of VAE-LSTM with the recent results in literature. Our comparisons show that latent representation learned by VAE and its variant CVAE (For brevity we often use the term "VAEs" to represent the pair.) can help achieve state-of-the-art speech emotion classification performance.

2. Related Work

Autoencoders have been extensively used for emotion recognition (e.g., [12, 13]), however to date, Variational Autoencoders have mainly been used for natural image generation (e.g., [14, 15]). Use of VAEs for speech processing and recognition is very limited. In the speech and audio domain, VAEs have mainly been used for speech generation and transformation [8]. They have also been used to learn phonetic content or speaker identity in speech segments without supervisory data [7, 8]. Moreover, a framework based on VAE was used in [16] to learn both frame-level and utterance-level robust representations. The authors used these salient features along with the other speech features for robust speech recognition. Hsu et al. [9] proposed a VAE based framework for modelling of spectral conversion with unaligned corpora. In this study, the encoder learned the phonetic representation for the speaker, and the decoder reconstructed the designated speaker by removing the demand of parallel corpora for the model training on spectral conversion. Finally, Blaauw et al. [7] used a fully-connected VAE to model the frame-level spectral envelopes of the speech signal. Based on their experiments, the authors found that VAE can achieve similar or comparatively better reconstruction errors than related competitive models like the Restricted Boltzmann machine (RBM).

Many researchers have used LSTMs for speech emotion recognition (e.g., [17, 18]). In many scenarios, LSTMs are

10.21437/Interspeech.2018-1568

more effective than conventionally-employed support vector machines [19]. Researchers have also used LSTM networks on the IEMOCAP speech corpus and have shown that they perform better than powerful methods like Hidden Markov Models [20, 21]. Chernykh et al. [17] used a Connectionist Temporal Classification (CTC) loss function with LSTM networks for emotion classification, and evaluated it on the IEMOCAP dataset. In another notable work [22], Emily et al. also employed the IEMOCAP database for speech emotion recognition. However, the authors have used transfer learning to leverage information from another database to improve the speech emotion accuracy. Transfer learning is out of the scope of this paper, but in future we would investigate if transfer learning can further enhance the accuracy achieved by our approach.

3. Methods

3.1. Generating Speech Features using VAE

Variational Autoencoder (VAE) is a combination of Graphical Models and Neural Networks. It has a similar structure as an Autoencoder (AE) but functions differently. An AE learns a compressed representation of the input and then reconstructs the input from the compressed representation. On the other hand, VAE learns the parameters of a probability distribution representing the input in a latent space. This is done by making the latent distribution as close as possible to a "prior" on the latent variable. The key advantages of the VAE over an AE is that the "prior" allows the injection of domain knowledge, enabling estimation of the uncertainty in the prediction, and making it more suitable for speech emotion recognition.

Formally speaking, given any emotion data X the aim of VAE is to find the probability of X with respect to its latent representation z:

$$P(X) = \int P(X|z)P(z)dz. \tag{1}$$

However, the quantities $P(X|z)$ and $P(z)$ both are unknown. The idea of VAE is to infer $P(z)$ using $P(z|X)$, where $P(z|X)$ is determined using Variational Inference (VI). In VI, $P(z|X)$ is inferred upon minimising the divergence with a known distribution $Q(z|X)$ as follows [3]:

$$\log P(X) = -\{|X - \hat{X}|^2 + KL[Q(z|X)||P(z)]\} \tag{2}$$

As can be seen in (2), the aim of VI is to eventually reduce the reconstruction error and to train the encoder $Q(z|X)$ in such a way that it produces the parameters of the probability distribution for the latent space z based on a known distribution of choice. This will minimise the divergence between $Q(z|X)$ and $P(z)$. For example, if we assume that the latent space will have a normal distribution, we need to train the encoder to generate the mean and covariance. Samples of $P(z|X)$ will be generated using these parameters, which the decoder will use to generate the approximation of X.

Conditional Variational Autoencoder (CVAE): In conventional VAE there is no way to generate specific data, for example a picture of an elephant, if the user inputs an elephant image. This is because the VAE models the latent variable and image directly. To eliminate this problem, the Conditional Variational Autoencoder (CVAE) models both latent variables and the emotion data conditioned on some random variables, c. The encoder is therefore conditioned on two variables X and c: $Q(z|X, c)$ and the decoder is also conditioned to two variables, z and c: $P(X|z, c)$. There are many possibilities for the conditional variable: it could have a categorical distribution expressing the label, or even could have the same distribution as the data.

Despite the capabilities of VAE, we are not particularly interested in generating speech emotion $\hat{X}$. However, when the distance ($|X - \hat{X}|^2$) between the original and the generated emotion becomes smaller than our predefined threshold, we use the parameters of the probability distribution $P(z|X)$ as the features for emotion X. For imposing conditions on the $P(z|X)$ (i.e. to emulate CVAE), we simply concatenate the speech frame representation in LogMel for any particular emotion X with its emotion class label (c) and pass this into the encoder.

3.2. Speech Emotion Classification using LSTM

LSTM can model a long range of contexts due to the presence of a special structure called the memory cell. Emotions in speech are context-dependent, therefore the ability to model contextual information makes LSTM suitable for speech emotion recognition [23].

The LSTM memory cell is built into a memory block, which constitutes the hidden layers of LSTM. There are three gate units in the memory cell - the input, output, and forget gate, which are used to perform reading, writing, and resetting of information, respectively. When the feature representations from the VAE are input to the LSTM, the input gate enables a memory block to selectively control the incoming information and store in the internal memory. The output gate decides what part of the information will be output, and a forget gate selectively clears the speech emotional contents from the memory cell.

To use LSTM for emotion classification, its output vector (end layer) is projected onto a vector with a length of the number of emotion classes. Projection is done using simple functions $Q = Wx$, where $x \in R^n$ is the LSTM output vector, $W \in R^{m \times n}$ is a weight vector and $Q \in R^m$ is the vector having the same length as the number of classes m. The vector Q is then mapped onto a probability vector with values in $[0, 1]$ having sum of the probabilities equates to 1. The highest probability indicates the identified class.

The overall classification framework has been shown in Figure 1. Previous studies have concluded that the performance of the LSTM model can be enhanced by using more predictive and knowledge-inspired features despite the limited training examples [19, 23, 24]. Therefore, LSTM is a natural choice for us to use with features generated by VAEs.

4. Experimental Setup

4.1. Speech Corpus

For experimentation, we selected the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [26] dataset, which is widely used for speech emotion recognition. IEMOCAP is a multimodal corpus containing recordings of ten actors over five sessions. Each session contains one female and one male speaker. The data includes two types of dialogues: scripted and non-scripted. In the non-scripted dialogue, the speakers were instructed to act without pre-written scripts. For the scripted dialogue data, the actors followed a pre-written script. Annotation was performed by 3-4 assessors based on both video and audio streams. Each utterance was annotated using 10 categories: neutral, happiness, sadness, anger, surprise, fear, disgust frustration, excited, and other. To better compare the results with

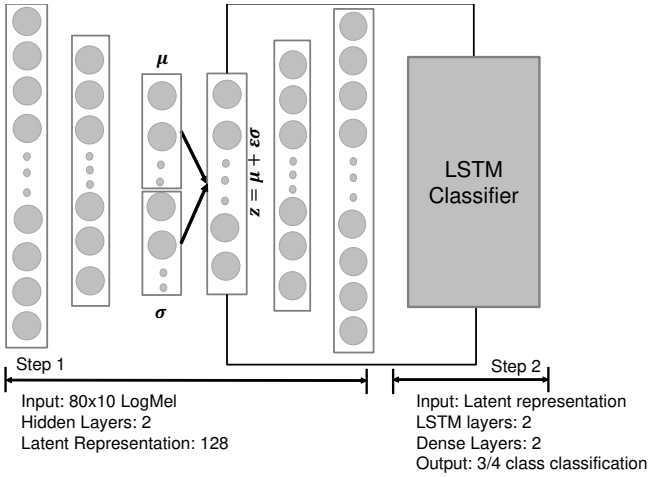


Figure 1: *Overall Classification Framework.*

related work, we computed our results for improvised, scripted and complete data (including both improvised and scripted). We considered four emotions: neutral, happiness, sadness, and anger, by combining happiness and excited as one emotion, following the state-of-the-art studies on this corpus [25, 27].

IEMOCAP data were also annotated on three continuous dimensions: Arousal (A), Power (P), and Valence (V). For comparison of our classification results with the state-of-the-art approaches in [19, 23], we also consider the above emotion dimensions. However, to maintain it as a classification problem, like [19,23], within each dimension we created three categories: low (values less than 3), mid (values equal to 3) and high (values greater than 3).

4.2. Speech Data Processing

We consider the LogMel speech frame representation, as used in [25,28]. Again following the above studies, a Hamming window of length 25ms with 10ms frame-shift was applied to the speech signal, and the discrete Fourier transform coefficients were computed. We then computed 80 mel-frequency filterbanks. The feature set was formulated by taking the logarithmic power of each mel-frequency band energy.

4.3. Configuration of VAE and LSTM

We input speech segments of length 100ms into the VAE for latent representation of data. This speech segment of 800 features is represented in a latent space of 128. We used two encoding layers with 512 and 256 hidden units respectively. The number of hidden units were chosen based on intuition from prior work on autoencoders [3] and on speech recognition using VAEs [8].

We used the Adam (adaptive moment estimation) optimiser, which is a Stochastic Optimisation Algorithm widely used to update network weights iteratively based on the training data [29]. The values of the various parameters used in the Adam optimiser were as follows: $\beta1$=0.999 and $\beta2$=0.99, ϵ=10^{-8} and learning rate =10^{-3}. These values were chosen in an iterative manner to obtain the minimum reconstruction loss of the autoencoder networks. We used the reparameterization trick [3] to approximate the latent space z with normally distributed δ by setting $z = \mu + \delta \odot \sigma$, where $\odot$ denotes element-wise multiplication, $\delta \sim N(0, 1)$, and $z \sim N(\mu, \sigma)$.

In CVAE we conditioned the VAE on the categorical emotion labels. To benchmark the performance of VAE, we also used a conventional autoencoder (AE) having the same architecture (i.e., hidden units, layers and model parameters), except for the Gaussian layer, which was replaced with a fully connected layer.

Our LSTM model consisted of two consecutive LSTM layers with the activation of the hyperbolic tangent. The hidden states of the second LSTM layer were connected to the dense layer and the outputs of the dense layer were fed into the softmax layer for classification of both categorical and dimensional class labels. The network parameters were chosen through cross-validation experiments. As a common setup, we used the Adam optimiser [29] with default learning rate of 10^{-3} by following [30]. To avoid overfitting, we used early stopping criteria with the maximum number of epochs equal to 20. All the experiments were performed using an Nvidia Quadro M5000 with 8 GB memory.

5. Results

The latent representations generated by both VAEs and AE were input to an LSTM network for classification. The segment-level latent representations obtained by autoencoder networks were merged into the whole utterance-level features for classification of emotions as in [31, 32]. Because the IEMOCAP corpus did not have a split of training and testing data, we investigated the performance of our model by training it in the speaker-independent manner. This also allowed us to compare our results with previous studies. We adopted a leave-one-session-out cross-validation approach and evaluated the models for both weighted accuracy (WA) and unweighted accuracy (UA) for categorical dimensions. For dimensional annotations, we followed evaluations strategies in [19,23] to be able to compare with these studies. We report the F-measures scores over the test dataset. The models were trained using 90% of data and testing was performed on the remaining 10% of unseen data.

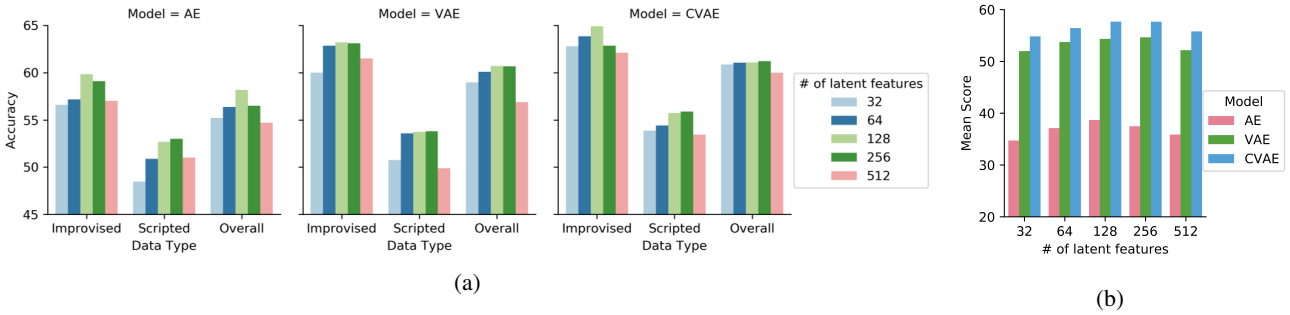
5.1. Classification Performance for Categorical Emotions

Table 1 shows the five-fold classification results on different subsets of the IEMOCAP data. It can be noted that the features learned by VAE produces better classification performance when compared with the conventional autoencoder. The representations learned by CVAE are highly predictive, which further outperform that learned by VAE.

In Table 1, we also compare different approaches on IEMOCAP used in the literature with our proposed approach. Lee et al. [18] proposed an extreme learning machine (ELM) based RNN model using bidirectional-LSTM (BLSTM) model and

Table 1: *Accuracy (%) comparison amongst different models for categorical classification.*

Data	AE-LSTM		VAE-LSTM		**CVAE-LSTM**		Attentive CNN [25] (WA)			BLSTM [17] (WA)	BLSTM [18] (WA)
	WA	UA	WA	UA	WA	UA	LogMel	MFCC	eGeMAPS		
Improvised	59.84	58.32	63.21	60.91	**64.93**	62.81	61.716	61.35	61.27	54	62.85
Scripted	52.68	48.52	53.74	52.23	55.71	53.50	52.64	53.19	53.19	NA	NA
Complete Data	58.16	55.42	60.71	56.08	61.08	58.10	54.86	55.12	54.78	NA	NA



(a) (b)

Figure 2: *Results using different number of latent features both on categorical and dimensional annotations. Figure 2a shows the effect of different number of features on categorical classification accuracy and 2b presents the corresponding trend of mean score for dimensional annotation.*

achieved 62.85% accuracy. The authors used low-level acoustic features and MFCC along with their derivatives, as a feature set to the model. In [25], authors used different types of features and evaluated single view (SV) as well as multi-view (MV) attentive CNN on IEMOCAP data using four emotions (as we used). We mention their best results (SV or MV) in the table. Chernykh et al. [17] used three different type of features (MFCC, chromagram, and spectrum properties) and report 54% accuracy using BLSTM. Using CVAE derived features, we achieve 64.93% accuracy, which is very competitive with respect to the literature.

5.2. Classification Performance for Dimensional Emotions

Table 2 presents the 10-fold cross-validation results on dimensional annotation using IEMOCAP data, where "Mean" represents the arithmetic mean of all three emotional dimensions: Arousal ("A"), Power ("P"), and Valence ("V"). The results are calculated on the basis of classifying the three subcategories: low, mid and high within each emotion dimension. We compare the performance of our proposed methods with an autoencoder model and also with some recent studies in the literature. Both VAE-LSTM and CVAE-LSTM significantly outperform the AE-LSTM model, while CVAE-LSTM producing the best performance.

Table 2: *Results on IEMOCAP data for dimensional annotations.*

Method	A (%)	P (%)	V (%)	**Mean (%)**
AE-LSTM	42.21	38.25	35.58	38.68
VAE-LSTM	61.35	53.18	48.46	54.33
CVAE-LSTM	62.73	53.84	52.69	**56.42**
DN features [19]	41.6	37.8	34.0	37.8
DN+LLD features [19]	53.9	51.6	39.5	48.3
eGeMAPS [23]	60.1	52.2	46.6	53
Hierarchical Feature Fusion [23]	61.7	52.8	51.2	55.3

Studies [19,23] that we have compared with in Table 2 used different types of features, such as knowledge-inspired disfluency and nonverbal vocalization (DN) features, and statistical Low-Level Descriptor (LLD) features, as an input to the LSTM model. The highest score they achieve is 55.3% (Mean score), which we closely outperform using our proposed CVAE-LSTM model (Mean score 56.42%).

5.3. Number of Latent Features Versus Accuracy

In all the results reported above, we have used a latent space size 128, which essentially means we have used 128 set of mean and variances (since, $z = \mu + \delta \odot \sigma$) of a normal distribution as latent features. However, we also investigate the impact of a higher and lower number of latent features.

Figure 2a and 2b show the trend of results using different number of latent features for categorical and dimensional emotions, respectively. Across all of AE, and VAEs, a very small number of features (32) perform poorly. However, a very large number of features (512) does not produce the best performance as well. Within this lower and higher bound, only an insignificant improvement can be observed with the increase of number of features. Based on these results we conclude that a suitable number of latent features needs to be determined empirically to avoid selecting a very small or a very large number of features.

6. Conclusion

In this paper we demonstrate that VAEs can effectively learn latent representation of speech emotion, which offers great potential for learning powerful features, automatically. We show that this helps achieve high classification accuracy when combined with a classifier of natural choice, LSTM, as LSTM has the intrinsic capacity to model contextual information like speech emotion, also an LSTM model can be enhanced by using more predictive and knowledge-inspired features. We analyse both categorical and dimensional emotions and comparing the emotion classification results with that of a widely used AE-LSTM model, we show that VAEs offer great promise by producing state-of-the-art results. We also analyse the impact of the number of latent features on classification accuracy with a view to determining the optimal number of features. However, we conclude that the suitable number of features needs to be determined empirically. Overall, the preliminary results presented in this paper demonstrate that it is highly feasible to automatically learn features for speech emotion classification using deep learning techniques, which will potentially motivate researchers to further innovate in this space.

7. Acknowledgements

This research is partly supported by Advance Queensland Research Fellowship, reference AQRF05616-17RD2.

8. References

[1] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[2] 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*, 2014, pp. 2672–2680.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[4] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[5] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," *arXiv preprint arXiv:1512.09300*, 2015.

[6] E. L. Denton, S. Chintala, R. Fergus *et al.*, "Deep generative image models using a laplacian pyramid of adversarial networks," in *Advances in neural information processing systems*, 2015, pp. 1486–1494.

[7] M. Blaauw and J. Bonada, "Modeling and transforming speech using variational autoencoders." in *INTERSPEECH*, 2016, pp. 1770–1774.

[8] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," *arXiv preprint arXiv:1704.04222*, 2017.

[9] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*. IEEE, 2016, pp. 1–6.

[10] J. Villalba, N. Brümmer, and N. Dehak, "Tied variational autoencoder backends for i-vector speaker recognition," *Proc. Interspeech 2017*, pp. 1004–1008, 2017.

[11] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, 2015, pp. 3483–3491.

[12] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 511–516.

[13] 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, 2018.

[14] X. Hou, L. Shen, K. Sun, and G. Qiu, "Deep feature consistent variational autoencoder," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE, 2017, pp. 1133–1141.

[15] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," in *Advances in Neural Information Processing Systems*, 2016, pp. 3738–3746.

[16] S. Tan and K. C. Sim, "Learning utterance-level normalisation using variational autoencoders for robust automatic speech recognition," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 43–49.

[17] V. Chernykh, G. Sterling, and P. Prihodko, "Emotion recognition from speech with recurrent neural networks," *arXiv preprint arXiv:1701.08071*, 2017.

[18] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition." in *INTERSPEECH*, 2015, pp. 1537–1540.

[19] L. Tian, J. D. Moore, and C. Lai, "Emotion recognition in spontaneous and acted dialogues," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 698–704.

[20] M. Wöllmer, A. Metallinou, N. Katsamanis, B. Schuller, and S. Narayanan, "Analyzing the memory of blstm neural networks for enhanced emotion classification in dyadic spoken interactions," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4157–4160.

[21] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[22] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," *arXiv preprint arXiv:1706.03256*, 2017.

[23] L. Tian, J. Moore, and C. Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 565–572.

[24] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 65–72.

[25] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *arXiv preprint arXiv:1706.00612*, 2017.

[26] 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.

[27] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, 2017.

[28] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] J. Kim, K. P. Truong, G. Englebienne, and V. Evers, "Learning spectro-temporal features with 3d cnns for speech emotion recognition," *arXiv preprint arXiv:1708.05071*, 2017.

[31] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[32] Y. Zhao, X. Jin, and X. Hu, "Recurrent convolutional neural network for speech processing," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5300–5304.