STUDY OF DENSE NETWORK APPROACHES FOR SPEECH EMOTION RECOGNITION

Mohammed Abdelwahab and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Department of Electrical Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

mxa129730@utdallas.edu, busso@utdallas.edu

ABSTRACT

Deep neural networks have been proven to be very effective in various classification problems and show great promise for emotion recognition from speech. Studies have proposed various architectures that further improve the performance of emotion recognition systems. However, there are still various open questions regarding the best approach to building a speech emotion recognition system. Would the system's performance improve if we have more labeled data? How much do we benefit from data augmentation? What activation and regularization schemes are more beneficial? How does the depth of the network affect the performance? We are collecting the MSP-Podcast corpus, a large dataset with over 30 hours of data, which provides an ideal resource to address these questions. This study explores various dense architectures to predict arousal, valence and dominance scores. We investigate varying the training set size, width, and depth of the network, as well as the activation functions used during training. We also study the effect of data augmentation on the network's performance. We find that bigger training set improves the performance. Batch normalization is crucial to achieving a good performance for deeper networks. We do not observe significant differences in the performance in residual networks compared to dense networks.

Index Terms— Speech emotion recognition, Deep Neural Networks.

1. INTRODUCTION

Speech emotion recognition has greatly benefited from advancements made in *deep neural networks* (DNNs) [1–4]. However, unlike successful classification problems that have used DNNs, such as image recognition, the lack of large naturalistic emotional databases has prevented the effective use of novel algorithms to train DNNs. Current DNN solutions for emotion speech recognition rely on standard activation functions for networks which are generally trained with less than four layers [3, 5].

A key limitation of current databases is the size of the corpus [6, 7]. Would the performance of the system improve if we have more labeled data? The intuition suggests that increasing the training set should increase the performance. However, it is not clear the value of additional data, as the classification performance may saturate. How much do we benefit from data augmentation? While data augmentation has played an important role in other fields, few studies have used it for speech emotion recognition [8, 9]. How does the depth of the network affect the performance? Current DNN solutions considers few layers. Many of the advances in DNNs in terms of regularization (e.g., batch normalization), activations (e.g., *exponential linear unit* (ELU)) and structures (e.g., *deep residual network* (RestNet)) are specially effective on deeper networks. What activation and regularization schemes are more beneficial for speech emotion recognition? These questions cannot be addressed with current emotional corpora, which are limited in size, naturalness and diversity.

We are collecting the MSP-Podcast corpus [7], a large spontaneous speech emotional database, which currently has over 30 hours of recordings. The size, diversity and naturalness of the corpus provides an ideal resource to study the aforementioned questions. This study explores practices in training DNN that work well in speech emotion recognition (prediction of arousal, valence and dominance). We evaluate the benefits gained when training with more data, changing the number of layers, and nodes of a fully-connected network. We evaluate whether deeper networks with more capacity are able to exploit the benefits of training with more data. We consider the value of using speed perturbation to augment available labeled data. Finally, we evaluate the use of batch normalization, alternative activation functions such as ELU, and novel architectures such as RestNet.

The experimental evaluation shows that increasing the training data is the most effective way to increase performance. As we increase the training set, we observe a positive trend that does not saturate, which validates our effort to increase the size of the MSP-Podcast corpus. The evaluation also shows that normalization between layers is crucial for deep networks to prevent degradation of performance when more layers are added. Contrary to our expectations, we do not observe better performance for ELU and RestNet compared to conventional fully connected networks trained with *rectified linear units* (ReLUs). The differences in prediction performance are not statistically difference, even when we train deeper structures. Also, we do not observe performance improvements when using data augmentation. As more resources for speech emotion recognition become available, the analysis in this study provides useful guidelines to build robust speech emotional recognizers.

2. RELATED WORK

Emotion recognition from speech is a growing research field. While new databases are being recorded, the amount of labeled data is relatively small compared to other fields that rely on thousands or millions of hours of labeled data. This lack of data caused the models trained for this task to be simpler and heavily regularized to avoid overfitting.

Data augmentation has been used to reduce overfitting on training data and improve the robustness of the trained models. Ko et al. [10] considered *Vocal Tract Length Perturbation* (VTLP), tempo perturbation and speed perturbation for audio augmentation in speech recognition. They showed that speed perturbation provided better performance improvement. Aldeneh and Provost [9] showed that speed perturbation augmentation gives significant improvement over no augmentation for speech emotion recognition using *convolutional neural networks* (CNN). Keren et al. [8] showed that replacing original examples with shorter overlapped examples also resulted in a performance boost in CNN.

This work was funded by NSF CAREER award IIS-1453781.

Another important aspect of DNN is the activation function. A common approach is ReLU, where the activation function is the identity for positive values and zero for negative values. It has become a popular choice due to its sparseness and having a derivate of 1 for positive values. These properties lead to faster training and convergence compared to sigmoid and tanh activation functions. However, ReLU units are non negative, so they have a positive mean activation. This issue introduces a bias shift that has to be compensated for in the following layers. To address this problem, new activation functions have been introduced [11–14]. In this study, we consider ELU [14] that introduces saturated negative values. the function is defined as follows

$$f(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha(exp(x) - 1) & \text{if } x < 0 \end{cases} \quad (1)$$

The authors mentioned that ELU units are most effective for networks of five or more layers. ReLU has been used extensively in speech emotion recognition [15–18]. We are not aware of any work using ELU for speech emotion recognition, since most of the networks in speech emotion recognition have between two and four layers. This study evaluates the benefits of ELU as we add more layers.

DNNs are difficult to train. As the network gets deeper, we notice performance degradation. To address this problem, several approaches have been proposed. Ioffe and Szegedy [19] introduced *batch normalization* (BN) that normalizes the output of each layer, leading to better gradient flow, faster training and lower sensitivity to the initialization of the parameters. In practice, batch normalization does not always lead to improvements. It seems to depend on the classification problem. Recently, more speech emotion recognition studies have used batch normalization [8, 20]. This study evaluates whether batch normalization is beneficial for speech emotion recognition, especially for deeper networks.

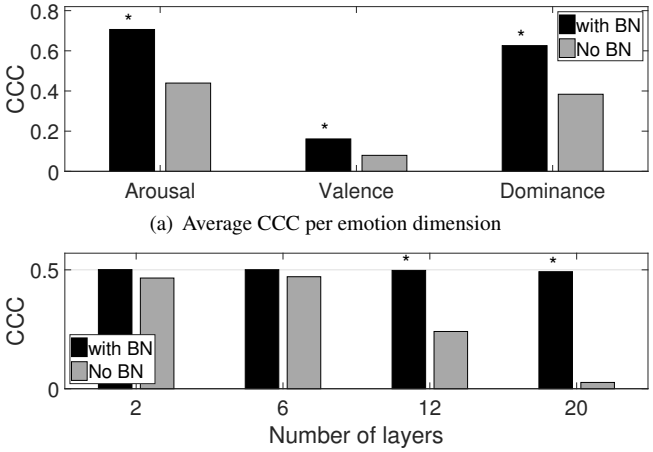
An appealing architecture for DNN with multiple layers is to use residual network. He et al. [21] proposed residual block which is made of a few dense layers with a skip connection. The skip connection passes the input of the block directly to the output of the last dense layer in the block. The skip connection allows the optimal representation learned at one layer to be maintained by the deeper layers. The authors showed that residual networks can achieve state of the art performance in image recognition and train even deeper networks. We are not aware of speech emotion recognition using residual networks. We consider residual networks to determine if our problem benefits from the residual network structure as we add more layers.

The closest study to our work is the work of Fayek et al. [22], which evaluated regularization and data augmentation for deep recurrent networks. Our study also explore the depth of the networks, activation functions, normalization and data augmentation. A main difference is that our study focuses on segment-based prediction instead of frame-based prediction, since frame-based formulation normally requires frame-level annotations that are usually absent from the majority of the speech emotion recognition corpora.

3. DATABASE AND ACOUSTIC FEATURES

3.1. The MSP-Podcast Corpus

The MSP-Podcast corpus is a collection of publicly available podcasts with creative commons license [7]. The database includes spontaneous interactions about various topics such as debates, movie discussion and sport shows, recorded by hundreds of speakers under different conditions. Each audio recording is segmented into several utterances of duration ranging from 2.75 to 11 seconds. For quality control, segments containing music, multiple overlapping speakers or having low signal to noise ratio are discarded. The utterances are annotated on *Amazon mechanical turk* (AMT) using a crowdsourcing protocol inspired by the study of Burmania et al. [23]. The utterances are annotated for both categorical emotions and continuous emotional dimensions. Emotional dimensions are annotated using seven likert scales for arousal, valence and dominance. Each utterance is annotated by at least five evaluators. The emotional dimension labels are assigned as the average of the available annotations. The collection of this corpus is an ongoing effort. This study uses the version 1.0 of the corpus, which consists of 20,045 labeled utterances, focusing on the continuous emotional dimensions (34 hrs, 15 min). The test set has 6,069 segments from 50 speakers (25 males, 25 females), the development set has 2,226 segments from 15 speakers (10 males, 5 females) and the training set has the remaining 11,750 segments. This partition attempts to create speaker independent sets for the evaluations.



(a) Average CCC per emotion dimension

(b) Average CCC per number of layers

Fig. 1. Average CCC for networks trained with batch normalization (BN) and networks trained without batch normalization.

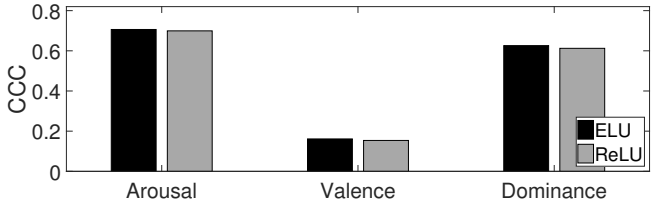
3.2. Acoustic Features

We use the feature set proposed for the INTERSPEECH 2013 Computational Paralinguistics Challenge (ComParE) [24]. This set consists of 65 *low-level descriptors* (LLD) extracted from speech, including prosodic, spectral, energy and voice quality features. High level statistical functions are then applied to generate 6,373 segment level *high-level descriptors* (HLD) features. The acoustic features are extracted using OpenSMILE [25].

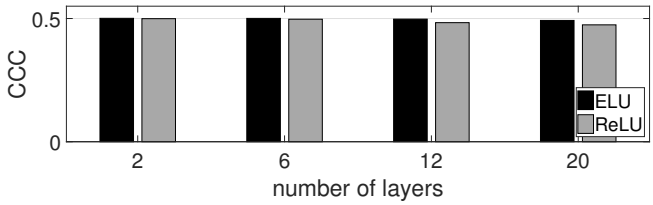
We normalize the features to have zero mean and a standard deviation of one. The mean and the variance of the data is calculated using 95% of the data to avoid outliers skewing the values. After normalization, we set values of any sample greater than 10 times the standard deviation to zero to reduce the effect of noisy outliers.

4. EXPERIMENTAL FRAMEWORK AND RESULTS

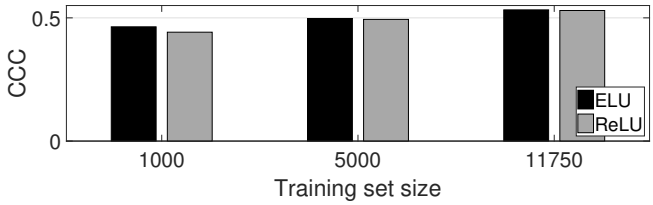
All experiments use Keras with TensorFlow backend, using GPU with Adam optimizer [26]. We train a separate model for each emotional dimension (arousal, valence, dominance). We train the models to maximize the *concordance correlation coefficient* (CCC) between the continuous emotional label of the segment (x) and its estimate (y). The CCC is defined as:



(a) Average CCC per emotion dimension



(b) Average CCC per number of layers



(c) Average CCC per training set size

Fig. 2. Average CCC for networks trained with ELU and ReLU.

$$\rho_c(x, y) = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (2)$$

where μ_x and μ_y are the means of x and y, σ_x^2 and σ_y^2 are the variances of x and y, and ρ is the Pearson correlation between x and y.

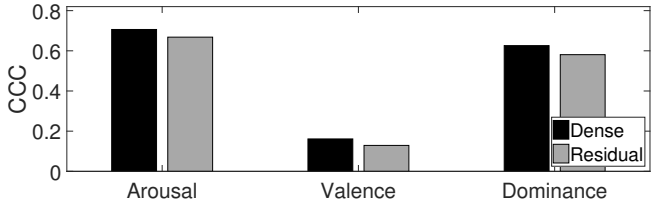
All the networks take an input vector of size 6,373 and output a scalar value. For all the networks, we set the batch size to 256. The learning rate is set to 1e-3 for the first 100 epochs, and then linearly annealed till it reaches zero. Dropout layers are introduced at the input with rate of 20%, between layers with rate of 50%, and maxnorm of four as a weight constraint. We train each network four times with a different seed to account for different initializations.

We train the networks with different number of layers (2, 6, 12, 20), number of nodes per layer (256, 1024), and number of training samples (1,000, 5,500, 11,750) to have a better understanding of the effects introduced by the implementation options.

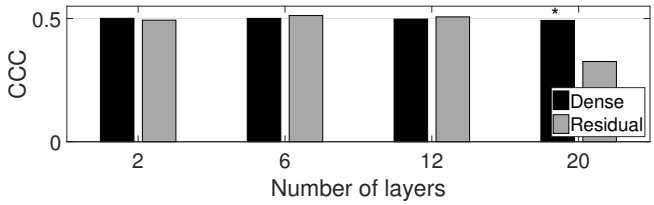
4.1. Batch Normalization

We consider the difference in the performance of the network with and without batch normalization. Batch normalization layer is added before the input of each layer.

Figure 1(a) shows the average CCC for networks trained with and without batch normalization for each emotion dimension. Each bar is the average CCC of 96 trials (4 initializations x 4 layers x 3 train size x 2 nodes= 96), the asterisk indicates statistical significance, (p-value < 0.01 under matched pair t-test). We can see that for all emotion dimensions batch normalization is essential to achieve good performance. Figure 1(b) shows that without batch normalization, the networks performance degraded as more layers are added, failing to train when we use 20 layers. However, with batch normalization, the network performance is more consistent across different number of layers. We have included batch normalization in all network variations that are considered in the remaining experiments.



(a) Average CCC per emotion dimensions



(b) Average CCC per number of layers

Fig. 3. Average CCC for dense networks compared to residual networks.

4.2. Activation functions (ReLU vs ELU)

Figure 2(a) shows the average CCC achieved when we use either ReLU or ELU. On average, we do not observe statistical differences when using ELU or ReLU. Figure 2(b) shows that for deeper layers, on average, ELU provides slightly better performance. The difference is more prominent when the training set size is small (Figure 2(c)). The rest of the evaluations are implemented with ELU.

4.3. Architecture Comparison (Dense vs Residual Networks)

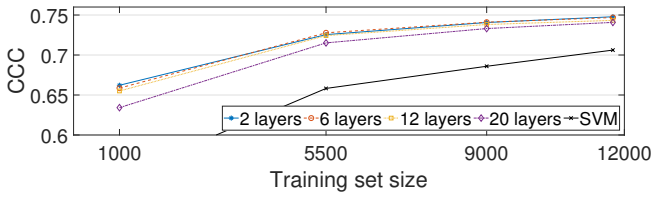
We compare fully connected networks with residual networks. We use a residual block comprised of 2 layers. We use projection shortcut for the first residual block due to the difference in feature map size. We use full pre-activation component ordering as proposed by Kaiming et al. [27].

Figure 3(a) shows the average CCC for dense networks compared to residual networks. We do not observe any statistical difference between both network structures. While both network structures have the same pearson correlation coefficient, residual networks tend to have a higher mean square error. This explains why dense networks have a slightly higher concordance correlation. Figure 3(b) shows that for a 20 layer network, on average, residual networks performs significantly worse. This drop in performance is noticed only when the training set size is small. The average CCC performance across emotional dimensions for a 20 layer network trained with 1,000 samples is 0.46 for dense networks, and 0.13 for residual networks. For dense networks, we observe an average concordance correlation coefficient that is consistent across different number of layers.

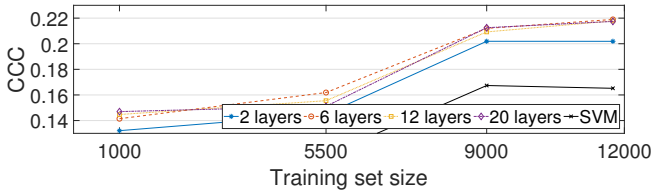
4.4. Training Set Size

To investigate how the training size affects the network's performance, we train the networks with 1,000, 5,500, 9,000, and 11,750 samples. As a reference, we also train a *support vector regression* (SVR) with RBF kernel. We set the parameters optimizing performance on the validation set ($C = 10$, $\epsilon = 0.01$).

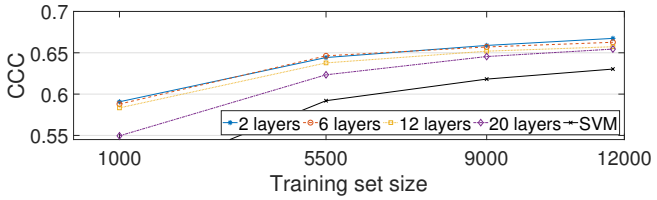
Figure 4 shows the average CCC achieved by dense networks with varying depth as the training set size increases for each emotion dimension. We consistently observe lower performance for SVR. For arousal and dominance, we notice that when the training set is



(a) Average CCC for arousal recognition by networks of different depth

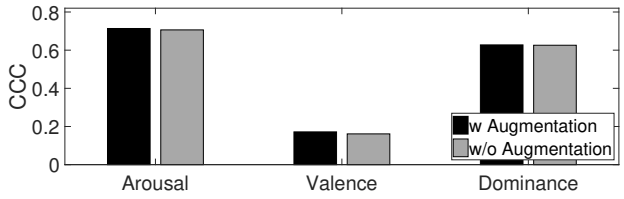


(b) Average CCC for valence recognition by networks of different depth

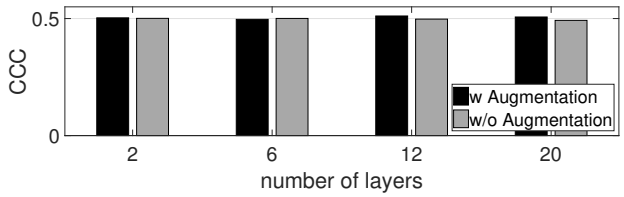


(c) Average CCC for dominance recognition by networks of different depth

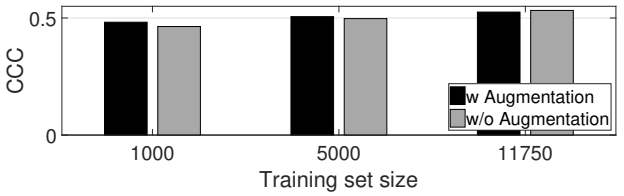
Fig. 4. Average CCC achieved by dense networks with varying depth as training set size increases.



(a) Average CCC per emotion dimensions



(b) Average CCC per number of layers



(c) Average CCC per training set size

Fig. 5. Average CCC for dense networks trained with and without data augmentation.

small the 20-layer network performs worse than the conventional 2-layer network. However, as the training set size increases, the performance gap between the networks diminishes. We expect the deep network to outperform the shallow network when more training data becomes available. Since valence estimation is a harder problem [28], we notice that deeper networks outperform the conventional 2-layer network. It is also important to note that for valence, the 2-layer network performance has started to saturate around 9,000 training examples. This trend is not observed in deeper networks which have more capacity.

It is clear that the networks benefit when we add more training data. This is reflected in the increase of the average concordance correlation coefficient achieved by the networks across all emotion dimensions. The network's performance gain diminishes as the training set size increases. However, the positive trend is consistent. We do not see evidence of saturation, which suggest that increasing the MSP-Podcast will lead to better models for speech emotion recognition. At some point, we expect the performance to saturate, as shown in other tasks such as audio event detection [29].

4.5. Data Augmentation

Figure 5 shows the average CCC for dense networks trained with and without speed perturbed data augmentation for each of the emotion dimensions. While, on average, we do not observe improvements when training with data augmentation for all network depths, we notice that data augmentation provides a small benefit for very deep layers when the training set size is small. The average CCC performance across emotional dimensions for a 20 layer network trained with 1,000 samples is 0.48 with augmentation, and 0.46 without augmentation.

5. CONCLUSIONS

The collection of large emotional speech databases such as the MSP-Podcast corpus are opening new research opportunities to explore better strategies to train DNNs for emotion recognition. This study explored the performance of regression models for arousal, valence and dominance as a function of the number of layers, and the size of the training set. The study also evaluated the use of alternative activation functions, batch normalization and residual networks. We observed that the most effective approach to improve performance is to increase the size of the training set. The study showed that batch normalization between layers is needed, especially as we increase the number of layers. With the current size of the MSP-Podcast (34h,15m), deeper networks were not able to capture the intricacy of the data to outperform simpler networks. Data augmentation is a viable option when the training size is limited. However, its benefit was not observed as we increase the size of the training set. We did not observe improvements using residual networks over conventional fully connected networks, or ELU over ReLU.

The data collection of the MSP-Podcast corpus is an ongoing effort. As the training set increases, it will be interesting to evaluate whether the patterns observed in this study are still consistent. We expect to observe better performance with more data. We are particularly interested in observing saturation of performance, and whether DNN with more capacity can raise the performance level. We will consider more robust normalization techniques such as layer or weight normalization. We will also study other data augmentation techniques, such as using synthetic data generated with *generative adversarial networks* (DNNs). Another open question is to evaluate the benefits of end-to-end networks, where the acoustic features are learned by the network.

6. REFERENCES

[1] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. Mower Provost, "Progressive neural networks for transfer learning in emotion recognition," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1098–1102.

[2] S. Parthasarathy, R. Lotfian, and C. Busso, "Ranking emotional attributes with deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 4995–4999.

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

[4] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech 2014*, Singapore, September 2014, pp. 223–227.

[5] N. Lane and P. Georgiev, "Can deep learning revolutionize mobile sensing?," in *International Workshop on Mobile Computing Systems and Applications (HotMobile 2015)*, Santa Fe, NM, USA, February 2015, pp. 117–122.

[6] H. Cao, D.G. Cooper, M.K. Keutmann, R.C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, October-December 2014.

[7] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. To appear, 2018.

[8] G. Keren, J. Deng, J. Pohjalainen, and B. Schuller, "Convolutional neural networks with data augmentation for classifying speakers' native language," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 2393–2397.

[9] Z. Aldeneh and E. Mower Provost, "Using regional saliency for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 2741–2745.

[10] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 3586–3589.

[11] A. Maas, A. Hannun, and A. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *International Conference on Machine Learning (ICML 2013)*, Atlanta, GA, USA, June 2013, vol. 30, pp. 1–9.

[12] K. Konda, R. Memisevic, and D. Krueger, "Zero-bias autoencoders and the benefits of co-adapting features," in *International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, USA, May 2015, pp. 1–11.

[13] L. He, D. Jiang, L. Yang, E. Pei, P Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *International Workshop on Audio/Visual Emotion Challenge (AVEC 2015)*, Brisbane, Australia, October 2015, pp. 73–80.

[14] D. Clevert, T. Unterthiner, and H. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *International Conference on Machine Learning (ICML 2016)*, San Juan, Puerto Rico, May 2016, pp. 1–14.

[15] S. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, R.C. Ferrari, M. Mirza, D. Warde-Farley, A. Courville, P. Vincent, R. Memisevic, C. Pal, and Y. Bengio, "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, June 2016.

[16] 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, January-March 2017.

[17] N. Lane, P. Georgiev, and L. Qendro, "Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning," in *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2015)*, Osaka, Japan, September 2015, pp. 283–294.

[18] H. Fayek, M. Lech, and L. Cavedon, "Towards real-time speech emotion recognition using deep neural networks," in *International Conference on Signal Processing and Communication Systems (ICSPCS 2015)*, Cairns, Australia, December 2015, pp. 1–5.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (PMLR 2015)*, Lille, France, July 2015, vol. 37, pp. 448–456.

[20] H. Fayek, M. Lech, and L. Cavedon, "On the correlation and transferability of features between automatic speech recognition and speech emotion recognition," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 3618–3622.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, Las Vegas, NV, USA, June-July 2016, pp. 770–778.

[22] H. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, August 2017.

[23] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[24] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[26] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, San Diego, CA, USA, May 2015, pp. 1–13.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision (ECCV 2016)*, Amsterdam, The Netherlands, October 2016, pp. 630–645.

[28] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 1179–1182.

[29] S. Hershey, S. Chaudhuri, D.P.W. Ellis, J.F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R.A. Saurous, B. Seybold, M. Slaney, R.J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 131–135.