INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

On the Correlation and Transferability of Features between Automatic Speech Recognition and Speech Emotion Recognition

Haytham M. Fayek[1], *Margaret Lech*[1], *Lawrence Cavedon*[2]

[1]School of Engineering, RMIT University
[2]School of Science, RMIT University
Melbourne, VIC 3001, Australia
haytham.fayek@ieee.org, {margaret.lech, lawrence.cavedon}@rmit.edu.au

Abstract

The correlation between Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER) is poorly understood. Studying such correlation may pave the way for integrating both tasks into a single system or may provide insights that can aid in advancing both systems such as improving ASR in dealing with emotional speech or embedding linguistic input into SER. In this paper, we quantify the relation between ASR and SER by studying the relevance of features learned between both tasks in deep convolutional neural networks using transfer learning. Experiments are conducted using the TIMIT and IEMOCAP databases. Results reveal an intriguing correlation between both tasks, where features learned in some layers particularly towards initial layers of the network for either task were found to be applicable to the other task with varying degree.

Index Terms: deep learning, emotion recognition, neural networks, speech recognition, transfer learning

1. Introduction

The relationship between Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER) is rather ill-defined. While both tasks use speech as an input modality, the acoustic model in ASR utilizes a few frames to recognize phonemes that are later decoded into a transcription, whereas the acoustic model in SER requires a larger number of frames to recognize emotions. Most of the work carried out on ASR considers the presence of emotions in speech a form of distortion and it has been shown in previous studies that the presence of emotions in speech degrades the accuracy of ASR systems [1]. On the other hand, several studies have reported improvement in SER accuracy when linguistic input is added to the acoustic input [2]. However, since ASR systems perform poorly in the case of emotional speech, this has not yet been fully realized. Therefore, it would be potentially very advantageous to integrate both systems with the aim of improving ASR systems in dealing with emotional speech and at the same time providing linguistic input to SER systems.

With deep learning being the state-of-the-art approach in ASR and SER, a hybrid system using deep learning may be viable. Nevertheless, to construct such a system, the relation between both tasks must first be studied, which could be achieved by studying the relation between the features learned and their relevance to both tasks. Using the notion of transfer learning [3, 4], where a model is trained on a base task and data set and then the trained model is repurposed to another target task or data set, the correlation between the features learned in both tasks could be illuminated.

Deep neural networks tend to learn low-level features in initial layers and transition to high-level features in final layers [5]. Similar low-level features commonly appear across various tasks and data sets, while high-level features are somewhat more tuned to the task or data set at hand, making low-level features more general and easier to transfer from one task or data set to another. In many situations, especially when data in the target task is scarce, the transfer of low-level features from one task or data set to another, followed by learning high-level features, may lead to a boost in performance given that both tasks or data sets share some similarity [6]. Conversely, transferring high-level features and learning low-level ones can be regarded as a form of domain adaptation and can be useful when the tasks are similar or identical but the data distributions are slightly different [4, 7]. In [5], the transferability of features in a computer vision task was experimentally studied, where the generality versus specificity of each layer in a deep neural network was quantified. It was also shown in [5] that there is a correlation between the benefit of feature transfer and the distance between the base task and the target task, such that the improvement due to feature transfer diminishes as the distance between the base and target tasks increase.

Transfer learning has been extensively utilized in ASR across languages and data sets [8]; such as in [9], where multiple languages were used to pre-train deep neural networks and then the trained models were fine-tuned on a target language, demonstrating improved results. Transfer learning has also been used in speaker adaptation, such as in [10, 11]. It has also been employed in SER, as in [12] where transfer learning between databases in SER was studied, indicating a boost in performance relative to independent learning. In [13], the viability of transfer learning between emotion recognition from music and speech was demonstrated. To the best of our knowledge, transfer learning between ASR and SER has not been investigated in prior work.

In this paper, we investigate the correlation and transferability of features learned in deep Convolutional Neural Networks (ConvNets) [14] between ASR and SER using transfer learning. We train two base models and iteratively quantify how each layer in each model is relevant to the other task. This could be achieved by transferring the learned layers from the base model to the target model on the other task, holding l layers constant and fine-tuning the remaining layers, as summarized in Figure 1. If the constant transferred layers are relevant to the target task, one can expect an insignificant or no drop in performance and vice versa. By iteratively varying the number of constant layers l, we can infer a correlation between both tasks.

http://dx.doi.org/10.21437/Interspeech.2016-868

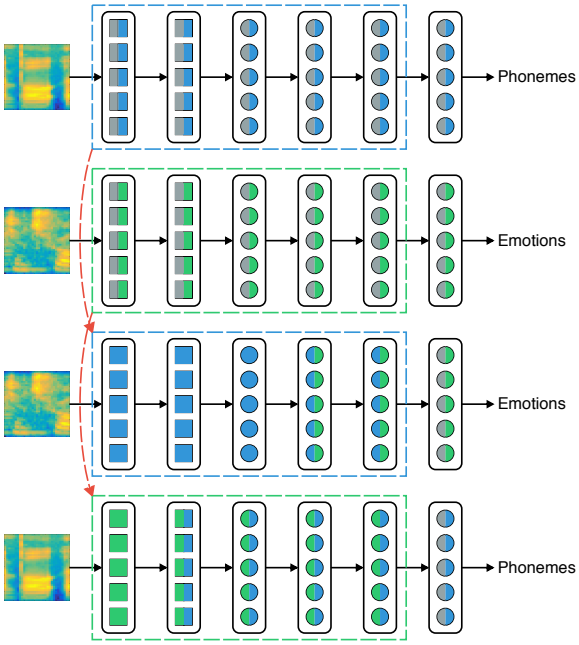

Figure 1: *Transfer Learning between ASR and SER. The parameters of the first and second models are initialized randomly and trained for ASR and SER respectively. The third and fourth models are examples of transfer learning. Note that the final layer in both models can not be transferred since the number of output classes are different. The parameters of the third model are initialized from the trained ASR model and the final three layers are fine-tuned for SER. The parameters of the fourth model are initialized from the trained SER model and all layers except the first layer are fine-tuned for ASR.*

The contributions of this paper are:

1. We comprehensively study and demonstrate the feasibility of transfer learning between ASR and SER, which has not been previously reported.
2. We propose a single acoustic model architecture that compares favorably with current state-of-the-art results on both tasks.
3. We show to what extent features can be used between both tasks, which paves the way for hybrid systems.

2. Experimental Setup

In this section, the data, pre-processing and systems used for ASR and SER are described in Sections 2.1 and 2.2 respectively. The ConvNet acoustic model architecture and training procedure, which are common to both tasks, are detailed in Section 2.3. Section 2.4 presents the transfer learning methodology.

2.1. Automatic Speech Recognition

The TIMIT database [15] was selected for the ASR task. The database contains recordings of 630 speakers from eight major American English dialects, each reading ten phonetically rich sentences. The complete 462-speaker training set without SA utterances was used for training. Hyperparameters were cross-validated on the 50-speaker development set, which was also used for early stopping during training. Results are reported on the 24-speaker core test set. As usual, the 61 phonemes were mapped to 48 phonemes for training and were then mapped to 39 phonemes for scoring, as detailed in [16].

Speech was analyzed using a 25 ms Hamming window with a stride of 10 ms. Log Fourier transform based filter banks with 40 coefficients distributed on a Mel scale were extracted from each frame. The mean and variance of each coefficient were normalized using the mean and variance computed on the training set only. No speaker dependent operations were performed. Kaldi [17] was used to pre-process the data and produce force-aligned frame labels by training a monophone Gaussian Mixture Model - Hidden Markov Model (GMM-HMM) system with Mel-Frequency Cepstral Coefficients (MFCCs).

The ASR system had a hybrid architecture [18] in that a ConvNet was trained as the acoustic model to produce pseudo-likelihoods for the states of a three-state HMM with a bigram language model trained on the training set.

2.2. Speech Emotion Recognition

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [19] was selected for the SER task. The database comprises 12 hours of audio-visual recordings divided into five sessions. Each session is composed of two actors, a male and a female, performing emotional scripts as well as improvised scenarios. In total, the database comprises 10039 utterances with an average duration of 4.5 s. Utterances were labeled by three annotators using categorical labels. Ground truths labels were obtained by majority voting, where 74.6% of the utterances were agreed upon by at least two annotators. Utterances that were labelled differently by all three annotators were discarded. To be consistent with other studies on this database [20, 21, 22], we included utterances that bore only the following four emotions: *anger*, *happiness*, *sadness* and *neutral*, with *excitement* considered as *happiness*; amounting to 5531 utterances.

An eight-fold Leave-One-Speaker-Out (LOSO) cross-validation scheme [23] was employed in all our experiments using the eight speakers in the first four sessions. Both speakers in the fifth session were used to cross-validate the hyperparameters of our model and to apply early stopping during training and therefore were not included in the cross-validation folds so as to not bias the results [24].

Audio was analyzed using a 25 ms Hamming window with a stride of 10 ms. Log Fourier transform based filter banks with 40 coefficients distributed on a Mel scale were extracted from each frame. The mean and variance were normalized per coefficient for each fold using the mean and variance computed using the training subset only. No speaker dependent operations were performed.

Since the data was labeled at an utterance level, all frames in an utterance inherited the utterance label. A voice activity detector was then used to label silent frames and *silence* was added as an additional class to the four previously mentioned emotion classes; i.e. a frame has either the same label as its parent utterance or the *silence* label. This scheme was adopted to be consistent with the ASR task since silence is considered a class in TIMIT and is scored. Moreover, the presence of silence (and other disfluencies) has proved to be an effective cue in emotion recognition [25].

The system comprised only an acoustic model as commonly employed in SER [26]. The acoustic model is a ConvNet identical to the model used in ASR.

2.3. Convolutional Neural Network Acoustic Model

Acoustic models in ASR usually utilize 9–21 frames of speech, which corresponds to approximately 100–220 ms. However, acoustic models in SER typically require a longer duration, with most studies using a complete utterance and some suggesting a duration of 250 ms to be sufficient [27]. We experimented with various numbers of frames using the validation sets in both tasks and empirically found that 31 frames was a good trade-off between both tasks. Temporal derivatives were not appended to the input as is usually done in ASR, since we observed that appending temporal derivatives degraded the accuracy of SER.

A ConvNet was chosen for the acoustic model, since features learned in ConvNets are more invariant to small changes facilitating transfer learning. The architecture of the ConvNet is detailed in Table 1. We used two convolutional and max pooling layers, followed by four fully-connected layers with Batch Normalization (BatchNorm) [28] layers and Rectified Linear Units (ReLUs) [29] interspersed between them and a softmax output layer. The output layer had 144 classes (i.e. 48 phonemes $\times$ 3 states) in the case of ASR, and 5 classes in the case of SER.

Table 1: *Convolutional Neural Network Architecture.*

No.	Type	Size	Other
1	Convolution	$64, 5 \times 4$	$l_2 = 1 \times 10^{-3}$
	BatchNorm	-	-
	ReLU	-	-
	Max Pooling	2×2	Stride = 2
2	Convolution	$128, 3 \times 3$	$l_2 = 1 \times 10^{-3}$
	BatchNorm	-	-
	ReLU	-	-
	Max Pooling	2×2	Stride = 2
3	Fully-Connected	1024	-
	BatchNorm	-	-
	ReLU	-	-
	Dropout	-	$P = 0.6$
4	Fully-Connected	1024	-
	Batch Norm	-	-
	ReLU	-	-
	Dropout	-	$P = 0.6$
5	Fully-Connected	1024	-
	BatchNorm	-	-
	ReLU	-	-
	Dropout	-	$P = 0.6$
6	Fully-Connected	144/5	-
	Softmax	-	-

The parameters of the convolutional and fully-connected layers were initialized from a Gaussian distribution with zero mean and $\sqrt{2/n}$ standard deviation, where n is the number of inputs to the layer, as recommended in [30]. Mini-batch stochastic gradient descent with a batch size of 256 and RMSProp [31] per-parameter adaptive learning rate were used to optimize the ConvNet parameters with respect to a cross entropy cost function. The base learning rate was set to 1×10^{-3} and 1×10^{-4} for ASR and SER respectively and the decay rate was set to 0.99. Convolutional layers were regularized using l_2 weight decay with penalty $\lambda = 1 \times 10^{-3}$; fully-connected layers were regularized using dropout [32] with a retention probability $P = 0.6$. The validation set was used to perform early stopping during training, such that training halts if the accuracy ceases to improve in 3 epochs; the model with the best accuracy on the validation set during training was selected. Training was carried out on a cluster of NVIDIA Tesla K40 Graphics Processing Units (GPUs).

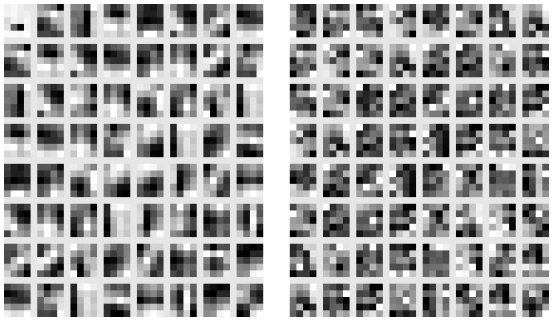

Figure 2: *Learned Features from ASR (left) and SER (right).*

2.4. Transfer Learning

After training the ConvNet acoustic models individually on each task, the trained model parameters in all layers except the output layer were copied to the other model for the other task, as depicted in Figure 1. The output layer was randomly initialized, since the number of output classes in both tasks were different. Subsequently, the first l layers were held constant, where $l \in \{0, 1, \ldots, 5\}$ and the remaining layers were fine-tuned on the target task. By iteratively varying the number of constant layers l and measuring the difference in performance between the baseline model trained from random initialization and the transferred model with l constant layers, we can quantify the relevance of features learned in the constant layers for the base task to the target task. If the constant transferred layers are relevant to the target task, one can expect an insignificant or no drop in performance and vice versa.

In the case of $l = 5$, the transferred model can be regarded as a feature extractor in that only the output layer, which was randomly initialized, was fine-tuned. In the case of $1 \leq l \leq 4$, the output layer was first fine-tuned for one epoch to avoid back-propagating gradients from randomly initialized weights to previous layers and subsequently the final $(L - l)$ layers (including the output layer) were fine-tuned simultaneously until the previously mentioned early stopping criterion was met, where $L = 6$ is the total number of layers in the ConvNet. Similarly, when $l = 0$, the output layer was first fine-tuned for one epoch and then all layers of the network were fine-tuned simultaneously with the output layer; in this case, the transferred model can be regarded as an initialization to the target model.

3. Results and Discussion

For the ASR task, we report the Frame Error Rate (FER), which is the output of the acoustic model and the Phone Error Rate (PER) which is the output of the system after decoding. For the SER task, we report the Error (E) and Unweighed Error (UE) ($1-$ Unweighted Average Recall (UAR) [23]) to reflect imbalanced classes. These metrics were the average of the eight-fold LOSO cross-validation scheme, except for the ASR baseline since the data split was predefined.

Our baseline ASR achieves an FER of 30.53% and 31.61% and a PER of 18.71% and 20.18% on the development and test sets respectively. Our baseline SER achieves an error of 44.63% and 46.44% and an unweighted error of 46.34% and 48.96% on the validation and test sets respectively. Both baseline results

compare favorably with current state-of-the-art. Slightly better results could be achieved if the number of frames was tuned for each task separately but this would complicate transfer learning between both tasks, which is the main objective of this paper. The learned features in the first layer of the ConvNet for both tasks are visualized in Figure 2. It is evident that features in ASR are smoother and more attuned to detect rapid variations, which is why temporal derivatives are popular in ASR. However, features in SER did not indicate a similar pattern and we hypothesize that this can be attributed to the longer period in which emotions manifest in speech.

3.1. Speech Emotion Recognition to Automatic Speech Recognition

The results of transferring the trained SER ConvNet to ASR and iteratively varying the number of constant layers l are listed in Table 2 and plotted in Figure 3.

Table 2: *Transfer Learning Performance SER to ASR.*

No. Constant Layers (l)	FER		PER	
	Dev	**Test**	**Dev**	**Test**
Baseline	30.53%	31.61%	18.71%	20.18%
5	71.09%	71.64%	61.15%	61.82%
4	53.26%	53.92%	42.96%	44.13%
3	40.29%	40.97%	28.81%	30.48%
2	31.75%	32.87%	20.08%	21.85%
1	30.83%	32.01%	18.99%	20.94%
0	30.62%	31.65%	18.73%	20.57%

Several observations can be made from the results. Using the trained SER ConvNet as a feature extractor ($l = 5$) yielded better performance than expected, suggesting that some features learned for SER were linguistic rather than purely acoustic. As for $1 \leq l \leq 4$, the results demonstrate two obvious trends: features learned in the first two layers could be transferred without fine-tuning and perform comparably well with minor degradation compared to the baseline model; while the subsequent layers indicate an almost linear correlation in that the performance drops significantly for each layer transferred without fine-tuning. The success of the first two layers may be attributed to two factors: features learned in the first two layers were general enough to be relevant to ASR; and convolutional layers are easier to transfer than fully-connected layers as they are less sensitive to small variations. Fine-tuning the entire model ($l = 0$) did not yield an improvement.

3.2. Automatic Speech Recognition to Speech Emotion Recognition

The results of transferring the trained ASR ConvNet to SER and iteratively varying the number of constant layers l are listed in Table 3 and plotted in Figure 3.

The results demonstrate several interesting trends. As expected, using the trained ASR ConvNet as a feature extractor ($l = 5$) yielded high error rates. This is likely due to the filtering of the paralinguistic aspects of speech in deep layers of the ASR acoustic model as paralinguistics hinder classification of the lexical components of speech, which coincides with observations made in [33]. As for $1 \leq l \leq 4$, features learned in the first two layers could be transferred without fine-tuning and perform similarly to the baseline model, which is particularly interesting since these layers were trained for ASR and were not fine-tuned for SER. Transferring subsequent layers without fine-tuning resulted in moderate degradation compared to the baseline model. Fine-tuning the entire model ($l = 0$) did yield an insignificant improvement.

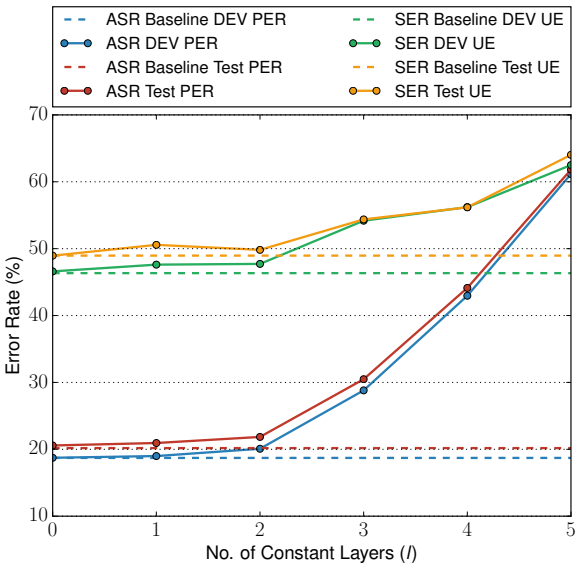

Figure 3: *Transfer Learning Performance.*

Table 3: *Transfer Learning Performance ASR to SER.*

No. Constant Layers (l)	E		UE	
	Dev	**Test**	**Dev**	**Test**
Baseline	44.63%	46.44%	46.34%	48.96%
5	52.55%	59.20%	62.50%	64.03%
4	51.94%	53.34%	56.21%	56.18%
3	50.22%	52.01%	54.18%	54.37%
2	47.39%	48.50%	47.72%	49.82%
1	46.37%	48.36%	47.61%	50.57%
0	45.26%	46.97%	46.60%	48.95%

4. Conclusion and Future Work

The correlation between Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER) was quantified by studying the relevance of features learned between both tasks in deep convolutional neural networks using transfer learning. Results attested to the feasibility of transfer learning between both tasks and demonstrated to what extent features can be used between both tasks. It was demonstrated that initial layers in the network are transferable between both tasks and the relevance of features decays gradually through deep layers.

Future work comprises the formulation of an ASR/SER hybrid system using a multi-task learning framework as well as incorporating other speech processing tasks into this study.

5. Acknowledgments

This research was funded by the Vice-Chancellor's PhD Scholarship (VCPS) from RMIT University. We gratefully acknowledge the support of NVIDIA Corporation with the donation of one of the Tesla K40 GPUs used for this research. We would like to thank Navdeep Jaitly for the helpful discussions.

6. References

[1] R. Fernandez, "A computational model for the automatic recognition of affect in speech," Ph.D. dissertation, School of Architecture and Planning, Massachusetts Institute of Technology, 2004.

[2] C. M. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 293–303, March 2005.

[3] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 7 1997.

[4] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," *JMLR: Proc. Unsupervised and Transfer Learning challenge and workshop*, pp. 17–36, 2012.

[5] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3320–3328.

[6] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," *CoRR*, vol. abs/1403.6382, 2014.

[7] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 513–520.

[8] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," *CoRR*, vol. abs/1511.06066, 2015.

[9] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in dnn-based lvcsr," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, Dec 2012, pp. 246–251.

[10] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid nn/hmm systems," in *INTERSPEECH*, 2010, p. 526529.

[11] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, Dec 2012, pp. 366–369.

[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*, Sept 2013, pp. 511–516.

[13] E. Coutinho, J. Deng, and B. Schuller, "Transfer learning emotion manifestation across music and speech," in *Neural Networks (IJCNN), 2014 International Joint Conference on*, July 2014, pp. 3592–3598.

[14] Y. LeCun and Y. Bengio, "The handbook of brain theory and neural networks," M. A. Arbib, Ed. Cambridge, MA, USA: MIT Press, 1998, ch. Convolutional Networks for Images, Speech, and Time Series, pp. 255–258.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, vol. 93, 1993.

[16] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 11, pp. 1641–1648, Nov 1989.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 12 2011, iEEE Catalog No.: CFP11SRW-USB.

[18] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 10, pp. 1533–1545, Oct 2014.

[19] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[20] M. Shah, C. Chakrabarti, and A. Spanias, "A multi-modal approach to emotion recognition using undirected topic models," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, June 2014, pp. 754–757.

[21] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 183–196, April 2013.

[22] Y. K., H. L., and E. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 3687–3691.

[23] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge." in *INTERSPEECH*, vol. 2009, 2009, pp. 312–315.

[24] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of database systems*. Springer, 2009, pp. 532–538.

[25] L. Tian, C. Lai, and J. Moore, "Recognizing emotions in dialogues with disfluencies and non-verbal vocalisations," in *Proceedings of the 4th Interdisciplinary Workshop on Laughter and Other Non-verbal Vocalisations in Speech*, vol. 14, 2015, p. 15.

[26] H. M. Fayek, M. Lech, and L. Cavedon, "Towards real-time speech emotion recognition using deep neural networks," in *Signal Processing and Communication Systems (ICSPCS), 2015 9th International Conference on*, December 2015.

[27] Y. Kim and E. Provost, "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 3677–3681.

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

[29] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *JMLR W&CP: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*, 4 2011.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *arXiv preprint arXiv:1502.01852*, 2015.

[31] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, vol. 4, 2012.

[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[33] A.-R. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 4273–4276.