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

Representation Learning for Speech Emotion Recognition

Sayan Ghosh[1], *Eugene Laksana*[1], *Louis-Philippe Morency*[2], *Stefan Scherer*[1]

[1]Institute for Creative Technologies, Department of Computer Science
University of Southern California Los Angeles, CA, USA
[2]Language Technologies Institute, Carnegie-Mellon University
Pittsburgh, PA, USA

sghosh@ict.usc.edu, elaksana@ict.usc.edu, morency@cs.cmu.edu, scherer@ict.usc.edu

Abstract

Speech emotion recognition is an important problem with applications as varied as human-computer interfaces and affective computing. Previous approaches to emotion recognition have mostly focused on extraction of carefully engineered features and have trained simple classifiers for the emotion task. There has been limited effort at representation learning for affect recognition, where features are learnt directly from the signal waveform or spectrum. Prior work also does not investigate the effect of transfer learning from affective attributes such as valence and activation to categorical emotions. In this paper, we investigate emotion recognition from spectrogram features extracted from the speech and glottal flow signals; spectrogram encoding is performed by a stacked autoencoder and an RNN (Recurrent Neural Network) is used for classification of four primary emotions. We perform two experiments to improve RNN training : (1) Representation Learning - Model training on the glottal flow signal to investigate the effect of speaker and phonetic invariant features on classification performance (2) Transfer Learning - RNN training on valence and activation, which is adapted to a four emotion classification task. On the USC-IEMOCAP dataset, our proposed approach achieves a performance comparable to the state of the art speech emotion recognition systems.

Index Terms: speech emotion classification, human-computer interaction, computational paralinguistics

1. Introduction

The field of speech emotion recognition has several potential applications, from affective computing to human-computer interfaces, and has witnessed considerable progress recently, partly due to newer datasets recorded with larger number of subjects and improved acquisition technology. Most existing approaches rely on the extraction of standard acoustic features such as pitch, shimmer, jitter and MFCCs (Mel-Frequency Cepstral Coefficients). Temporal characteristics of the data are obtained with statistical functionals which are used as descriptors for segment or utterance-level detection of emotions from speech [1]. Feature extraction is followed by a simple classification stage, with models such as Support Vector Machines (SVM) and Hidden Markov Models (HMM).

There is little effort in obtaining an effective end-to-end representation learning pipeline for speech emotion recognition, where the spectrogram or the time domain waveform are utilized for training a subsequent neural network. One important advantage of representation learning [2] is that features describing the signal are learnt directly from the waveform samples or from the frequency domain representation. This does not require any additional feature extraction, and often generalizes well to unseen data. Moreover, high data collection and annotation costs often necessitate experiments on limited amounts of labeled data with few human subjects. Different training paradigms such as semi-supervised learning [3] and transfer learning [4] have been reported to improve classification accuracy, but have not been extensively explored for emotion classification or affect recognition from speech.

In this paper, we explore spectrogram based representations for speech emotion classification from the USC-IEMOCAP dataset. We experiment with features from the speech spectrogram, but also from the glottal volume velocity spectrogram. Our experiments investigate if classification performance can be improved by filtering out unwanted factors of variation such as speaker identity and verbal content (phonemes) from speech. The frame-based feature representations are obtained by training stacked denoising autoencoders from context windows of spectrograms, and then we learn an utterance-level emotion Bidirectional Long-Short Term Memory (BSLTM)-RNN model. We also study a transfer learning scenario where we leverage additional utterances which have not been labeled with four basic emotions by training an RNN on the valence and activation labels (which comprises all utterances in the dataset), and then adapting the trained network to the four emotion classification task. The primary research questions we wish to investigate in the paper are:
Question 1: Are representations learned from spectrograms successfully discriminative for speech emotion recognition?
Question 2: Do speaker and phonetic invariant representations improve emotion classification accuracy? Do they improve on recognizing emotions which are commonly confused [5] due to similar voice characteristics?
Question 3: Can data insufficiency be addressed by transfer learning from affective attributes such as valence and activation to emotions?

2. Related Work

Our experiments in this paper build on prior work in the areas of representation learning and deep neural networks, as well as emotion and affect recognition from speech. Jaitly and Hinton [6] proposed transforming autoencoders to learn acoustic events (onset times and rates) from speech datasets such as Arctic and TIMIT. Graves et al [7] explored recurrent neural networks for speech recognition, obtaining a test set error of 17.7% on TIMIT. Deep learning approaches have also been applied to emotion recognition from speech. Kim et al [8] explored multi-

http://dx.doi.org/10.21437/Interspeech.2016-692

modal deep learning for audiovisual emotion recognition from the IEMOCAP dataset. Han et al [9] performed speech emotion recognition from the IEMOCAP corpus using a combination of DNN (Deep Neural Network) and Extreme Learning Machines. They obtained 20% relative accuracy improvement compared to state-of-the-art approaches.

Lee et al [10] trained recurrent neural networks for speech emotion recognition from IEMOCAP, where the label of each frame is modeled as a sequence of random variables. They obtained a weighted accuracy improvement of 12% compared to a DNN-ELM baseline. Jin et al [11] generated feature representations using standard acoustic and lexical features for emotion recognition from IEMOCAP, obtaining 55.4% accuracy from early fusion of acoustic features (cepstrum and Gaussian supervectors). Trigeorgis et al [12] perform representation learning for end-to-end speech emotion recognition. The present work complements these prior investigations by learning representations for emotion recognition from spectrograms. We have previously explored unsupervised speech representations for affect [13], and in this work further investigate if speaker and phonetic invariant representations of speech are discriminative of emotion, and whether transfer learning from affective attributes to categorical emotions can improve classification performance.

3. Model

3.1. Pretraining with Stacked Denoising Autoencoders

An autoencoder [14] is a neural network typically trained to learn a lower-dimensional distributed representation of the input data. The input dataset of N data points $\{\mathbf{x}_i\}_{i=1}^{i=N}$ is passed into a feedforward neural network of one hidden layer which is a bottleneck with activations $\{y_i\}_{i=1}^{i=N}$ given by $\mathbf{y}_i = \tanh(\mathbf{W}\mathbf{x}_i + \mathbf{b})$. We use $tanh$ activation functions in our work. The output of the autoencoder is $\mathbf{z}_i = \mathbf{W}'\mathbf{y}_i + \mathbf{b}'$ and is generated from the bottleneck layer. Thus the autoencoder is trained using backpropagation, much as in an ordinary feedforward neural network. We use SSE (Sum of Squared Error Loss) $L = \sum_{i=1}^{i=N} \|\mathbf{x}_i - \mathbf{z}_i\|^2$ for training in this paper. Vincent et al [15] introduce denoising autoencoders, where the data point $\mathbf{x}_i$ is corrupted (by randomly setting a fraction of the elements to zero) to produce the input $\tilde{\mathbf{x}}_i$, from which the original clean data point $\mathbf{x}_i$ is set to target $\mathbf{z}_i$ and is reconstructed by the autoencoder. When training denoising autoencoders in a greedy stacked fashion, we have multiple layers with weights $\mathbf{W}^{k-1}$ and $\mathbf{W}^k$ for the k-th hidden layer, where the autoencoder activations at the layer are (where we have $\mathbf{y}_i^0 = \tilde{\mathbf{x}}_i$) $\mathbf{y}_i^k = \tanh(\mathbf{W}^{k-1}\mathbf{y}_i^{k-1} + \mathbf{b}^{k-1})$. Similar to most prior work [16], we used a pyramidal stacking approach for the autoencoders, where the number of neurons is less (generally halved) for the next higher layer. We have kept the same autoencoder output layer size for all feature sets to enable a fair comparison.

3.2. Classification with BLSTM-RNN

Recurrent Neural Networks (RNN) are suitable for learning time series data. While RNN models are effective at learning temporal correlations, they suffer from the vanishing gradient problem which increases with the length of the training sequences. To resolve this problem, LSTM (Long Short Term Memory) RNNs were proposed by Hochreiter et al [17] to model long term temporal dependencies. In our paper, we use BLSTM (Bidirectional LSTM)-RNNs for sequence classification with a target replication scheme, where the target class for each time step is assigned to the emotion category of the entire utterance. For prediction of an input sequence $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_T\}$ the predicted emotion category is obtained by a majority voting scheme on the predictions of individual time steps in the sequence, which correspond to context frames in the utterance. The BLSTM-RNN has two layers, each of cell size 30, with a four-dimensional softmax layer on top for emotion classification. We perform validation experiments on the RNN to find the best model, where each hyper-parameter setting is obtained from random sampling on the grid below:

- *Learning Rate* : [6e-6,8e-6,1e-5,2e-5,4e-5]
- *Momentum* : [0.7,0.8,0.9]
- *Input noise variance* : [0.0,0.1,0.2,0.3]
- *Weight noise variance* : [0.0,0.05,0.1,0.15,0.2]
- *Batch size*: 1300 utterances
- *Maximum Epochs* : 100

To improve generalizability of the BLSTM, random noise is added to the input sequences and the model weights in every epoch, and can be controlled by the noise variance hyperparameters. We have used the BLSTM-RNN implementation from the CURRENNT toolbox [18].

4. Dataset and Feature Extraction

For our experiments, we used the USC-IEMOCAP dataset [19], which is a well-known dataset for speech emotion recognition comprising of acted and spontaneous multimodal interactions of dyadic sessions between actors, where conversations are scripted as well as improvised. The dataset consists of around 12 hours of speech from 10 human subjects, and is labeled by three annotators for emotions such as *Happy*, *Sad*, *Angry*, *Excitement*, *Neutral* and *Surprise*, along with dimensional labels such as *Valence* and *Activation*.

We perform classification experiments only on four basic emotions - *Neutral*, *Angry*, *Sad* and *Happy*, with a total of 5531 utterances belonging to these categories (Happy: 1636 Angry: 1103 Sad: 1084 Neutral: 1708). These standard emotion categories were chosen since they are most commonly used for emotion recognition from the IEMOCAP dataset [8], [11]. Each utterance is approximately 2-5 seconds in duration, with short periods of silence before and after speech in all utterances. The affective dimensions (valence, activation and dominance) are annotated on a Likert scale of 1-5, where we have averaged out the dimension ratings across all three annotators. We are also interested in analyzing the performance of our proposed approach over *improvisational* utterances from the dataset. Improvisational utterances correspond to conversations not generated from a predefined script and are hence more spontaneous and similar to natural speech.

4.1. Glottal Source Waveform

Paralinguistic and affective attributes such as emotion, valence and activation should be speaker and phonetic invariant, and not sensitive to changes in the speaker's identity or verbal content (phoneme or words being uttered). We wish to investigate whether filtering out the factors of variation (speaker identity and phonetic information) from the speech signal prior to training of the denoising autoencoder and the BLSTM-RNN could improve classification performance. The glottal source waveform has this property and is obtained by glottal inverse filtering of the speech signal using the Iterative Adaptive Inverse Filtering (IAIF) algorithm [20]. While the signal obtained through inverse filtering may be an approximate of the actual glottal waveform and potentially result in experimental bias, we have chosen the IAIF algorithm because it is widely used in the literature. Besides, we are also motivated by the recent success of glottal flow based features such as Normalized Open Quotient (NAQ), and Quasi-open Quotient (QOQ), which have been shown to be discriminative at tasks such as depression assessment [21], and voice quality classification [22].

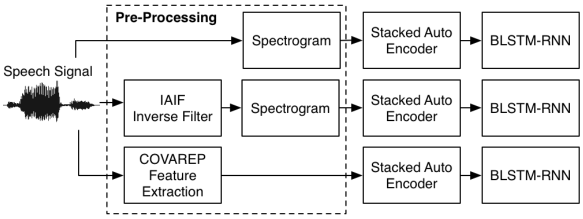

Figure 1: Experimental setup for emotion recognition

4.2. Spectrogram Representations

Spectrograms were extracted from both (1) the speech waveforms and (2) glottal flow waveforms. We used a frame width of 20 ms and a frame overlap of 10 ms for extraction. The spectrograms consist of 128 FFT (Fast Fourier Transform) bins for each frame and we subsequently stack a context window of five adjacent frames together (resulting in a 640-dimensional vector) to better capture contextual information. The 640-dimensional feature vectors are input to the stacked autoencoder which has an architecture of 640-320-160-64. This is motivated by prior work [23],[24] which report that longer temporal context improves emotion classification performance. We used a log-scale in the frequency domain, since a higher emphasis in lower frequencies has been shown to be more significant for auditory perception.

4.3. Baseline Feature Extraction

For the baseline approach, we extract commonly used speech features for affect recognition, such as Mel-Frequency Cepstral Coefficients (MFCCs) and prosodic/voice quality features [22] from the COVAREP toolbox [25] version 1.4.1. The reader is requested to refer to [26] for a detailed description of the features, which include F0 (Fundamental Frequency), NAQ (Normalized Amplitude Quotient), QOQ (Quasi-open Quotient) and MFCCs with delta coefficients. A context window vector of size 175 is created by stacking 35 COVAREP feature vectors from five adjacent frames, and are input to a stacked autoencoder with an architecture of 175-64.

5. Experimental Setup

The utterances in the IEMOCAP dataset are split into five sessions, where each session consists of a dyadic conversation between a male and a female speaker. The experiments in our paper are performed in a leave one session out strategy, similar to [10]. Since there are 10 speakers in the dataset, each session consists of 2 speakers. For each fold, utterances from eight speakers (four sessions) correspond to the training set, and from the remaining session, hyper-parameter validation is performed on one speaker, testing is performed on the other speaker, and vice-versa. We have reported weighted and unweighted classification accuracy for the entire testing set (*scripted+improvised*) and a subset consisting only of *improvised* utterances. Weighted accuracy is the accuracy over all testing utterances in the dataset, and unweighted accuracy is the average accuracy over each emotion category (Happy, Angry, Sad and Neutral).

For each session, stacked autoencoder pretraining is performed over all training set utterances. The BLSTM-RNN training for the emotion classification task is done only for utterances which have been labeled with the primary emotion categories. Refer to Figure 1 for a pictorial overview of our approach. We compare our proposed approach to the following baselines :
(1) The **DNN-ELM** approach in [9] where the authors train a Deep Neural Network (DNN) with an ELM (Extreme Learning Machine).
(2) The **RNN-ELM** approach in [10] where the authors train a Recurrent Neural Network (RNN) with an ELM (Extreme Learning Machine).
(3) Jin et al. [11] where acoustic and lexical features are fused to create higher level representations. They also use standard features, along with techniques such as BoW (Bag of Words Modeling). For fairness of comparison we report their results obtained with acoustic features.
(4) **Our proposed model**, where we extract features from the COVAREP toolbox and stack them to create context window descriptors of five frames, instead of obtaining spectrogram representations.

Transfer Learning Valence and Activation to Emotions: In our paper, we perform classification experiments on four categories of emotion - *Happy*, *Angry*, *Sad* and *Neutral*. However, 4328 utterances in the IEMOCAP dataset are annotated with valence and activation intensities, even though they do not belong to the four primary emotion categories. The correlation of affective dimensions and emotion categories has been extensively studied in the literature [27], and we wish to investigate if features which are representation learned for the task of affective dimension regression are also discriminative of emotion classification.

For each utterance in the dataset, we obtain an aggregate score for the valence and activation dimensions by averaging all the annotator ratings, and train a BLSTM-RNN network as a regression model. Similar to emotion classification, we replicate the targets at every timestep, where each target is a two dimensional vector consisting of the valence and activation score for the utterance. The network architecture is identical to the BLSTM classifier, with the exception of a two-dimensional feedforward linear unit in the topmost layer. The SSE (Sum of Squared Errors) is used for training the network, and subsequently the top layer is replaced with a softmax layer. Training is resumed in the *adaptation* stage, where instead of the valence and activation, the target is an emotion label belonging to one of the four basic emotion categories. Transfer learning not only enables the RNN to make use of additional unlabeled data, but also provides us an insight into the correlation between emotions and affective dimensions.

6. Results and Discussions

As described in the previous sections, we conduct classification experiments to compare the performance of representations learned from the speech and glottal flow spectrograms with (1) state-of-the-art baselines [9], [10] and [11] (2) acoustic and voice quality features extracted using the COVAREP toolbox [25]. In Table 1, we present test accuracies in the leave-one-session-out setup for different emotions and feature sets. From an examination of the results, we observe that *Happy* and *Angry* classes frequently are confused in the classification, while highest performance is achieved for the *Sad* category.

Table 1: Test Accuracies reported for different feature sets on overall (improvised+scripted) and improvised utterances. Results for our proposed approach are shown in bold

Utterance Category	Feature Set	Weighted	Unweighted	Happy	Angry	Sad	Neutral
Overall	COVAREP	48.19	50.26	36.13	56.98	65.57	42.38
	Spectrogram (Speech)	**48.70**	**49.75**	**35.15**	**43.16**	**67.13**	**53.57**
	Spectrogram (Glottal)	**50.47**	**51.86**	**36.98**	**53.58**	**64.27**	**52.63**
Improvised	COVAREP	49.64	51.84	38.47	58.50	66.05	44.35
	Spectrogram (Speech)	**51.94**	**51.85**	**39.60**	**42.91**	**69.85**	**55.05**
	Spectrogram (Glottal)	**52.82**	**54.56**	**44.24**	**56.06**	**65.73**	**52.23**

	Hap	Ang	Sad	Neu
Hap	35.1	24.9	14.4	25.4
Ang	35.7	43.1	3.9	17.2
Sad	7.0	2.8	67.1	23.0
Neu	20.0	6.5	19.8	53.5

(a) Speech confusion matrix

	Hap	Ang	Sad	Neu
Hap	36.9	18.5	15.1	29.2
Ang	24.2	53.5	4.2	17.9
Sad	8.3	2.5	64.27	24.9
Neu	20.8	4.5	22.0	52.6

(b) Glottal confusion matrix

Figure 2: Confusion matrices obtained for speech and glottal representations

Representation learning from spectrograms: In [10], the authors report the best unweighted and weighted accuracies obtained by the DNN-ELM model to be 52.13% and 57.91% respectively. The corresponding performances obtained by their proposed RNN model (without ELM) are around 56% and 58% respectively, which are comparable to our accuracies. In [10] it is not clearly specified which speaker in each session was used for validation and testing respectively. In our experiments, we select one speaker for validation and one for testing. We then repeat the experiment with the speakers switched and include the average performance of both test sets in our evaluation. The RNN-ELM approach in [10] also trains and tests only on *improvised* utterances, whereas our approach trains on both *improvised* as well as *scripted* utterances. The acoustic fusion approach in [11] report 10 fold leave-one-speaker-out validation accuracies on all utterances (scripted as well as improvised), but do not explicitly evaluate on a testing set. The performances they report (weighted accuracy of 49% for cepstral BoW to 55.4% for feature fusion) are comparable to the validation accuracy of 57% which we obtain in the leave-one-session-out scenario. These findings show that emotion categories can be directly learned from low-level spectrogram representations.

Speaker and Phonetic Invariance: The representations obtained from the glottal flow signal outperform those obtained from the speech signal by 1.77% in terms of weighted accuracy, which shows that it is beneficial to filter out factors of variation such as speaker identity and phonetic information from the speech prior to emotion classification. To examine the difference in greater detail between the performances of the glottal and speech representations, we present both confusion matrices obtained by testing on all utterances in Figure 2. Each row of the tables consists of the fraction of the ground truth emotion labeled utterances which have been confused with other emotions during prediction. We observe that glottal flow representations reduce the confusion between *Happy* and *Angry* categories to a major extent. This improves their respective classification accuracies by 1.8% (for *Happy*) and 10.4% (for *Angry*). We believe that this improvement stems from the fact that as reported in [5], *Happy* and *Angry* have similar acoustic characteristics. Further, they share a similar level of activation when considering their location in the valence-activation space. We believe that glottal flow representations are less prone to confuse *Happy* and *Angry*, as they capture differences on the valence dimension better. Based on these findings we can confirm research question 2 and conclude that representations learned from speech can be improved by filtering speaker and phonetic factors of variation prior to classification.

Transfer Learning: To address research question 3, we pretrain the BLSTM-RNN as a regression model for valence and activation on the entire training set (for four sessions in each fold), and subsequently finetune it for the four category emotion recognition task. We found that it was necessary to scale down the BLSTM weights inside the pretrained network to improve performance on the adaptation task. Further for finetuning, the hyper-parameters also have to be validated over a grid similar to the setting described in Section 3.2. Since the glottal flow spectrogram performs the best among the competing feature sets, we conduct the transfer learning experiment only on the glottal flow representations. We obtained a weighted accuracy of 51.64% and an unweighted accuracy of 52.89% on the test set, with emotion wise accuracies of 38.03% (Happy), 55.40% (Angry), 67.35% (Sad) and 50.81% (Neutral). While the transfer learned representations achieve an improvement of 1.17% and 1.03% in weighted and unweighted accuracies respectively, it is not a significant boost compared to directly learning categorical emotions from the glottal flow representations.

7. Conclusions

In this paper, we have investigated representation learning for categorical emotion recognition from spectrograms of the speech and glottal flow signals. Our experiments indicate that representation learned features are highly discriminative of emotion classification and are comparable to state-of-the-art approaches. We also find that filtering out speaker and phonetic information by inverse filtering reduces confusion between *Happy* and *Angry* categories, and that transfer learning from valence and activation to emotion categories provides a marginal improvement in performance. Overall, we believe our findings are encouraging, in particular with respect to potential performance improvement in a multi-classifier system due to the diversity in observed errors. In addition, we plan to explore more extensive transfer learning experiments with much larger datasets.

8. Acknowledgements

The work depicted here is sponsored by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005. Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred. The Tesla K40 used for this research was donated by the NVIDIA Corporation.

9. References

[1] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge." Citeseer, 2009.

[2] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1798–1828, 2013.

[3] M. F. A. Hady and F. Schwenker, "Semi-supervised learning," in *Handbook on Neural Information Processing*. Springer, 2013, pp. 215–239.

[4] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 759–766.

[5] R. J. Davidson, K. R. Scherer, and H. Goldsmith, *Handbook of affective sciences.* Oxford University Press, 2003.

[6] N. Jaitly and G. E. Hinton, "A new way to learn acoustic events," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[7] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6645–6649.

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

[9] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," *Proceedings of INTERSPEECH, ISCA, Singapore*, pp. 223–227, 2014.

[10] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4749–4753.

[12] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, S. Zafeiriou *et al.*, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[13] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Learning representations of affect from speech," *arXiv preprint arXiv:1511.04747, International Conference on Learning Representations (ICLR) 2016 Workshop*, 2015.

[14] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," *Unsupervised and Transfer Learning Challenges in Machine Learning, Volume 7*, p. 43, 2012.

[15] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.

[16] J. Masci, U. Meier, G. Fricout, and J. Schmidhuber, "Multi-scale pyramidal pooling network for generic steel defect classification," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE, 2013, pp. 1–8.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] F. Weninger, J. Bergmann, and B. Schuller, "Introducing currennt: The munich open-source cuda recurrent neural network toolkit," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 547–551, 2015.

[19] 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, pp. 335–359, 2008.

[20] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[21] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[22] S. Scherer, J. Kane, C. Gobl, and F. Schwenker, "Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification," *Computer Speech & Language*, vol. 27, no. 1, pp. 263–287, 2013.

[23] Y. Kim and E. M. 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*. IEEE, 2013, pp. 3677–3681.

[24] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 216–221.

[25] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarepa collaborative voice analysis repository for speech technologies," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 960–964.

[26] M. Brilman and S. Scherer, "A multimodal predictive model of successful debaters or how i learned to sway votes," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*. ACM, 2015, pp. 149–158.

[27] R. Sun and E. I. Moore, "Empirical study of dimensional and categorical emotion descriptors in emotional speech perception," in *Twenty-Fifth International FLAIRS Conference*, 2012.