Modeling gender information for emotion recognition using Denoising autoencoder

Rui Xia, Jun Deng, Björn Schuller, Yang Liu

Angaben zur Veröffentlichung / Publication details:

Xia, Rui, Jun Deng, Björn Schuller, and Yang Liu. 2014. "Modeling gender information for emotion recognition using Denoising autoencoder." In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4-9 May 2014, Florence, Italy*, 990–94. Piscataway, NJ: IEEE. https://doi.org/10.1109/icassp.2014.6853745.

MODELING GENDER INFORMATION FOR EMOTION RECOGNITION USING DENOISING AUTOENCODER

Rui Xia[1,2], *Jun Deng*[2], *Björn Schuller*[2,3] , *Yang Liu*[1]

[1] Computer Science Department,The University of Texas at Dallas, USA
[2] Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany
[3] Department of Computing, Imperial College London, U.K.

ABSTRACT

The Denoising autoencoder (DAE) has been successfully applied to acoustic emotion recognition lately. In this paper, we adopt the framework of the modified DAE introduced in [1] that projects the input signal to two different hidden representations, for neutral and emotional speech respectively, and uses the emotional representation for the classification task. We propose to model gender information for more robust emotional representation in this work. For neutral representation, male and female dependent DAEs are built using non-emotional speech with the aim of capturing distinct information between the two genders. The emotional hidden representation is shared for the two genders in order to model more emotion specific characteristics, and is used as features in a back-end classifier for emotion recognition. We propose different optimization objectives in training the DAEs. Our experimental results show improvement on unweighted accuracy compared with previous work using the modified DAE method and the classifiers using the standard static features. Further performance gain can be achieved by structural level system combination.

Index Terms— Emotion recognition, Denoising autoencoder, Gender

1. INTRODUCTION

There has been a lot of research efforts lately on identifying paralinguistic information in human speech (information beyond words). Many challenges [2, 3, 4] related to paralinguistic tasks have been organized and attracted many researchers. Automatic emotion recognition is one of such paralinguistic tasks (others include various speaker states, age, etc.). To accurately detect emotion, front-end feature extraction and back-end classification are two major parts. In the front-end, it is important to extract a robust feature representation which captures emotional cues. Previous work (e.g., [5]) has shown that static features extracted by applying various functionals to large amounts of low level descriptors (LLD) can yield competitive performance on emotion recognition tasks. Many studies have been conducted to investigate complementary features in addition to these static features, such as gaussian mixture model (GMM) related features [6, 7], bag-of-word sentiment categories as lexicon features [8], and facial related features [9]. In the classification stage, standard classifiers such as support vector machines (SVM) have been very popular. In addition, ensemble methods have been used to take advantages of strength of multiple classifiers and shown good results. For example, in [10], a particle filtering based method is used for fusion of audio, visual and lexicon features. Besides these, prior studies also investigated transfer learning [11] and active learning [12, 13] approaches for the emotion recognition task. With growing interest in deep neutral network (DNN) recently, deeper structure by stacking autoencoders or Restricted Boltzmann Machines (RBM) has also been successfully used in many fields including emotion recognition task [14, 15, 16, 17, 18, 19]. In our previous work [1], we proposed to use the denoising autoencoder (DAE) and its modified version for emotion recognition. By introducing two hidden representations, one used to capture neutral information and the other for emotional cues, we demonstrated that a more robust feature representation can be extracted, yielding a performance gain on emotion recognition.

In this study, we adopt the modified DAE framework as in [1], but propose to better model the neutral projection in order to consider gender information. In speech recognition, gender dependent acoustic models are sometimes used in order to model the huge difference between male and female speech (vocal tract characteristics, pitch, etc.). For emotion recognition, there has been little prior work on modeling gender information. In [20], gender-dependent emotion recognizers are trained. In our method, we train female and male dependent DAEs separately by using their corresponding non-emotional data. These gender dependent parameters are used in the modified DAE framework to estimate the emotional projection. In addition, we propose different cost functions when pre-training the gender dependent DAEs, which are meant to either capture the shared information between the two genders or the distinct features between them. Our experimental results show that our proposed method has better performance compared with results in previous work.

2. METHOD

In this section, we first briefly introduce the method in previous work [1] that uses the denoising autoencoder (DAE) for emotion recognition. Then, based on this previous approach we propose a method considering gender variability.

2.1. Previous Work: Modified DAE for Emotion Recognition

The traditional DAE introduced by Vincent et al. [21] aims to learn a mapping function from an input to a hidden representation, which can capture the main variation of the data. There are two stages for training DAE, pre-training and fine-tuning. The unsupervised pre-training stage iteratively minimizes the loss function between the original input and the reconstructed input. To make a prediction, a softmax layer can be added on top of the hidden layer. Given the

Correspondence should be addressed to {rx,yangl}@hlt.utdallas.edu, {jun.deng,schuller}@tum.de

predicted labels and the ground truth, a supervised fine-tuning stage is applied to further update parameters. The details of the learning algorithm can be found in [21].

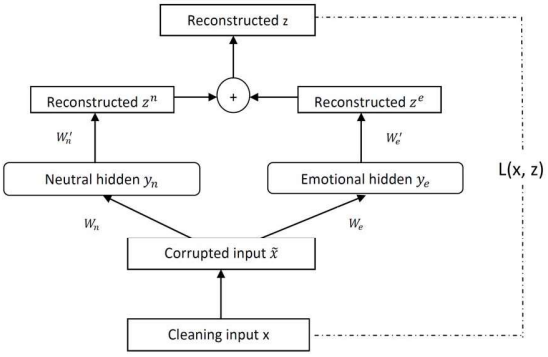

Fig. 1. Modified DAE structure with two hidden representations for emotion recognition.

In [1], we proposed a modified DAE framework for emotion recognition, as shown in Figure 1. The major difference between this and the traditional DAE is that in the hidden layer we proposed to project the input into two hidden representations, y_n and y_e. y_n is called neutral hidden and designed to capture neutral information that may be contained in all emotional speech. The other one y_e is called emotional hidden, which encodes emotional information.

During training, two parameter sets need to be estimated. The parameter set, $\theta_n(W_n, b_n, b_n^{'})$, associated with the projection to neutral hidden is pre-learned by a traditional DAE using a large neutral based corpus. The parameter set of emotional hidden representation, $\theta_e(W_e, b_e, b_e^{'})$, is estimated via pre-training and fine turning. It is initialized with random values and pre-trained based on the following steps:

- Encoding: project corrupted input to hidden representations

$$y_n = s(W_n\widetilde{x} + b_n), \quad (1)$$
$$y_e = s(W_e\widetilde{x} + b_e). \quad (2)$$

- Decoding: reconstruct inputs from hidden representations

$$z_n = s(W_n^T y_n + b_n^{'}), \quad (3)$$
$$z_e = s(W_e^T y_e + b_e^{'}). \quad (4)$$

- Combine: make new reconstructed input with linearly weighted combination

$$z = \alpha * z_e + (1 - \alpha) * z_n \quad (5)$$

- Learning: minimize the loss function and update θ_e

$$L(x, z) = |z - x|^2. \quad (6)$$

where s is sigmoid function ($s(x) = (1 + exp(-x))^{-1}$). The loss function $L(x, z)$ is defined as the squared error between the new reconstructed input and the original input. Stochastic gradient descent algorithm is applied to minimize the cost. Note that here we only update the parameter set θ_e and fix the other parameter set θ_n.

After pre-training, a softmax layer is added on top of the emotional hidden layer for classification. Parameters in θ_e are fine-tuned based on the predictions and the corresponding ground truth. For emotion recognition, the emotional hidden representations are used as features with support vector machine (SVM) as the back-end classifier. Experimental results in [1] showed improved emotion recognition performance using this modified DAE method, suggesting the emotional hidden projection in this framework is a better feature representation.

2.2. Proposed Method

Based on research in speech recognition, it is expected that there may be significant differences between male and female speech, including emotional speech. Therefore we investigate if modeling gender information benefits speech emotion recognition. One way to do this would be using the same above framework for male and female speech separately, i.e., build gender dependent neutral and emotion models. However, since the typical emotional speech data sets are rather small, we expect that splitting the data this way will result in too little data for training the emotion models. Therefore in this study, we propose to model gender information by using gender dependent neutral models, but shared emotion models, based on the modified DAE framework described above. Additionally, when training the gender specific neutral models, we propose to consider the relationship between the two genders. The following describes our method in details.

2.2.1. *Gender Dependent Neutral DAEs*

Assume we have non-emotional speech training utterances for female, x_f, and male, x_m. From these, the corrupted input sets, $\widetilde{x_f}$ and $\widetilde{x_m}$, are generated by adding Gaussian noise. To build gender dependent DAEs, instead of simply estimating model parameters using the two sets separately, we propose to consider the relationship between the male and female data. Figure 2 shows our proposed method – we use the female model as an example here to explain the method. Note that this DAE corresponds to the left part of Figure 1, i.e., the neutral model.

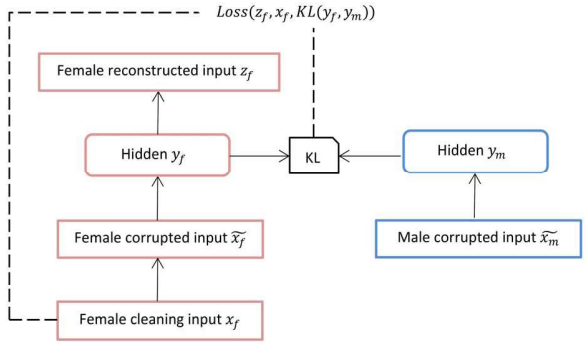

Fig. 2. Proposed DAE for the female neutral model.

To train the female dependent DAE, we first encode female corrupted instances $\widetilde{x_f}$ to the hidden representation and reconstruct it as follows:

$$y_f = s(W_n\widetilde{x_f} + b_n), \quad (7)$$
$$z_f = s(W_n^T\widetilde{y_f} + b_n^{'}). \quad (8)$$

Here, we use the pre-learned parameter set θ_n (the same for female and male models) as the initial value instead of training from scratch.

To train the female model parameters, we consider male speech information as well, and define two loss functions $Loss_1$ and $Loss_2$ as follows:

$$Loss_1(x_f, z_f) = |z_f - x_f|^2 - \beta * KL(p||q), \quad (9)$$

$$Loss_2(x_f, z_f) = |z_f - x_f|^2 + \gamma * KL(p||q) \quad (10)$$

where β and γ are hyper parameters, p and q are calculated as follows:

$$p = \frac{1}{M}\sum_{l=1}^{M} s(W_n x_{f_l} + b_n), \quad (11)$$

$$q = \frac{1}{N}\sum_{k=1}^{N} s(W_n x_{m_k} + b_n). \quad (12)$$

where M and N represent the number of instances for female and male sets. With p and q, $KL(p||q)$ is defined as:

$$KL(p||q) = \frac{1}{S}\sum_{i=1}^{S}(p_i log(\frac{p_i}{q_i}) + (1-p_i) * log(\frac{1-p_i}{1-q_i})), \quad (13)$$

where S means the number of the hidden nodes.

There are two components in these loss functions(Equation 9 and 10). The first part is the standard squared loss function – it calculates the reconstruction cost between the original input x_f and the reconstructed z_f. This is the same as the standard or the modified DAE. The second term in the above loss functions considers information from the other gender. p and q, as calculated in Equation 11 and 12, represent the average of the hidden representations over the training set for female and male respectively. We introduce a distance function (Equation 13) for these two average representations, based on KL-divergence: it is the average KL-divergence between each of the corresponding hidden nodes for the two genders.

The two loss functions are based on two different considerations: $Loss_1$ (Equation 9) aims to increase $KL(p||q)$, which means increasing the difference between the two hidden representations, i.e., forcing the models to learn differences between genders. $Loss_2$ (Equation 10) tries to minimize $KL(p||q)$ in order to make the DAEs to encode some shared information in both genders. The motivation behind these two is based on the assumption that female and male speech may have shared information, as well as distinct gender specific information. Note that the added KL term in the loss function compared to the standard squared loss can also be treated similarly as a regularization term in many optimization problems.

During training, the stochastic gradient descent algorithm is applied to minimize the loss function and update parameters. Since we use batch mode to train the DAE, when training the female models, M is equal to the number of instances in each minibatch. N is the total number of instances in the male set x_m. After training, two groups of the estimated parameter sets based on different loss functions can be obtained. $\theta_{(f|m)L_1}(W_{(f|m)L_1}, b_{(f|m)L_1}, b'_{(f|m)L_1})$ denote female or male parameter sets trained by using $Loss_1$. $\theta_{(f|m)L_2}(W_{(f|m)L_2},\ b_{(f|m)L_2}, b'_{(f|m)L_2})$ are for DAEs learned based on $Loss_2$.

2.2.2. *Emotional Hidden Representation and Emotion Recognition*

For building the emotion model, we use the same modified DAE framework. Different from the previous work, for projection to the neutral hidden representation, we apply gender specific parameter sets (different depending on the loss functions used) on inputs with the known gender label in the training set. Projection parameter sets from input to the emotional hidden are initialized with randomized value and iteratively estimated. Pre-training minimizes the squared loss between the reconstructed z and the input, and fine-tuning updates parameters for the emotional hidden projection to minimize the emotion classification error.

After pre-training and fine-tuning, emotional hidden representations are used as new features for emotion recognition with standard classifiers. Again, we use the same emotional hidden representation for male and female data. This is meant to capture gender independent but emotion specific information. During testing, we do not need to know the gender label for the test instances because features are generated by only passing instances into the emotional projection.

2.2.3. *Combination*

In the above, we mentioned using different loss functions will result in different models. Here we propose a combination method on the structure level to combine the different losses. Rather than projecting the input to one neutral hidden representation in the modified DAE framework, we project the input to two neutral hidden projections. The parameter set for each projection is associated with the corresponding loss function, $\theta_{(m|f)L_1}$ and $\theta_{(m|f)L_2}$. Given $\theta_{(m|f)L_1}$ and $\theta_{(m|f)L_2}$, two neutral reconstructed inputs z_{L_1} and z_{L_2} can be calculated as follows:

$$z_{L_1} = s(s(W_{(f|m)L_1}\widetilde{x} + b_{(f|m)L_1})W^T_{(f|m)L_1} + b'_{(f|m)L_1}), \quad (14)$$

$$z_{L_2} = s(s(W_{(f|m)L_2}\widetilde{x} + b_{(f|m)L_2})W^T_{(f|m)L_2} + b'_{(f|m)L_2}). \quad (15)$$

Then, we combine z_{L_1} and z_{L_2} linearly with equal weights to obtain the neutral reconstructed input z_n as:

$$z_n = 0.5 * z_{L1} + 0.5 * z_{L2}. \quad (16)$$

Parameter training is similar to above, except now in pre-training we use the combined z_n. Estimation of z_e is the same as before, via pre-training and fine tuning.

3. EXPERIMENTS

3.1. Features

We use the static features extracted with openSMILE [22] as the input signal in the DAE framework. There are 1,584 features in total, as used in the INTERSPEECH 2010 Paralinguistic Challenge [23]. Since the feature values have very different ranges, we normalized all the features to the range of 0 to 1 before using them as input to the DAE. Details of the features can be found in [23]. Table 1 summarizies these features.

3.2. Data

The interactive Emotional Dyadic Motion Capture (USC IEMOCAP) database [24] is used in this study. This corpus has approximately 12 hours of audiovisual data, including video, speech, motion capture of face, and text transcriptions. It has 10 professional actors (5 male and 5 female) acting in two different scenarios: scripted play and spontaneous dialog, in their dyadic interaction. Each interaction is around 5 minutes long, and is segmented into sentences. These sentences are labeled by at least 3 annotators. We use four emotion categories in this study: angry, happy, sad and neutral. Note that we merged Happy and Excited in the original annotation into one class: happy. Only the utterances with the majority agreement are used in the experiments. There are 5,531 utterances in this database.

Table 1. *Acoustic feature sets: 38 low-level descriptors (LLD) and 21 functionals.*

Descriptors	Functionals
PCM loudness	Position max./min.
MFCC [0-14]	arith. mean, std. deviation
log Mel Freq. Band [0-7]	skewness, kurtosis
LSP Frequency [0-7]	lin. regression coeff. 1/2
F0	lin. regression error Q/A
F0 Envelope	quartile 1/2/3
Voicing Prob.	quartile range 2-1/3-2/3-1
Jitter local	percentile 1/99
Jitter consec. frame pairs	percentile range 99-1
Shimmer local	up-level time 75/90

3.3. Experimental Setup

We conduct leave-one-speaker-out cross validation for the emotion recognition experiments. Normalization of features is based on all the training set, instead of a speaker-wised manner. To pre-train gender independent θ_n in the DAE, we use the Wall Street Journal (WSJ) corpus (about 78K instances) to train with the traditional DAE. The learning rate is set to 0.01 and the number of training epochs is 30. Each minibatch contains 1000 instances. The reason of using WSJ corpus here, rather than the IEMOCAP data, is because we want to train a general DAE to represent neutral speech. Then, to train the gender dependent DAE parameter sets $\theta_{(m|f)L_1}$ and $\theta_{(m|f)L_2}$, we use 10 iterations with 0.01 as the learning rate. This is done using the training set from the IEMOCAP data. The hyper parameters β and γ are set as 0.01 and 0.1 respectively. After that, we use $\theta_{(m|f)L_1}$ and $\theta_{(m|f)L_2}$ as parameters of the neutral projections in the modified DAE method. In the pre-training stage, the number of training iterations is set to 20 and the learning rate is 0.01. In the fine-tuning stage, we use 12 iterations with 0.05 as the learning rate. The weight combination parameter α is 0.7, which means the emotional reconstruction has more weight than the neutral one. For all the DAE models, a corruption level of 0.1 is used to obtain the corrupted signal from the original features. The number of hidden nodes is 800, the same as that used in previous work [1]. SVMs with radial basis function (RBF) kernels are used as the classifier for the new features from the hidden representation in the proposed DAE method.

3.4. Experiment Results

Table 2 shows the emotion classification results based on the unweighted average recall (UAR), a metric that has been used as the standard measurement in the INTERSPEECH Emotion Challenges. This is the average of the results for each emotion class. For a comparison, the first two results in 2 are from systems using the original static features and emotional hidden projection as features in the previous modified DAE method [1]. The following rows show the results with features extracted by using the gender dependent neutral projection learned based on different loss functions, and system combination. The last row shows the result using the new gender dependent neutral projection as in our proposed method; however, for the loss function, we do not include the KL cost, and thus it is just a standard squared loss for each gender separately, without considering their relationship during DAE training.

From Table 2 we can see that features extracted based on our proposed method can yield better results on UAR, 1.5% and 1.2% improved, using the two loss functions respectively, compared to the previous modified DAE method. System combination on the structure level yields further gain. Our method has a significant improvement compared with our previous work (p-value < 0.05 with one tailed z-test) and passes the significance level of 0.01 compared with the baseline static features. When KL cost is not used in the loss function, there is a performance degradation compared to when it is used in our proposed method, indicating the effectiveness of modeling the relationship between the two genders. Table 3 shows the accuracy for each emotion class. We notice that there is more improvement for the 'neutral' and 'sad' classes, and less for 'angry' and 'happy'. This suggests we may need to build models taking into account the valence or arousal dimension. Finally, we extended the above gender dependent framework to speaker dependent ones. Using the same data set, our experimental results show similar performance as when using gender dependent neutral models. This might be because each speaker has very little data, limiting the potential advantage of building speaker dependent models. We will continue to investigate this in the future work.

Table 2. *Emotion classification results (in %).*

System		UAR
Static features		59.7
Previous modified DAE [1]		61.4
New DAE	$Loss_1$	62.9
	$Loss_2$	62.6
	System combination	**63.1**
	No KL in Loss	62.2

Table 3. *Accuracy in % for each emotion category.*

System		angry	happy	neutral	sad
Static features		66.0	52.3	53.0	67.5
Previous modified DAE [1]		68.3	58.2	54.0	65.2
New DAE	$Loss_1$	69.4	58.4	56.3	67.3
	$Loss_2$	68.8	57.8	56.5	67.0
	System combination	69.5	58.8	56.8	67.3

4. CONCLUSION

In this paper, we proposed to consider gender information to model neutral projection and better capture emotion specific features in the DAE framework. When training gender dependent models, we use KL-divergence between the hidden representations of the male and female speakers as the additional cost in the objective function to measure correlation between the two genders. Emotional projection is trained under the modified DAE framework with gender dependent DAEs used for neutral projection. Our experiments show that using the new emotional projection as features yielded better system performance, suggesting the benefit of modeling gender variability for emotion recognition.

5. ACKNOWLEDGMENT

This work is supported by U.S. Air Force Award FA9550-10-1-0388, NSF award 1225629, and DARPA contract FA8750-13-2-0041. Any opinions expressed in this work are those of the authors and do not necessarily reflect the views of the funding agencies.

6. REFERENCES

[1] Rui Xia and Yang Liu, "Using denoising autoencoder for emotion recognition.," in *Proceedings of INTERSPEECH*, 2013.

[2] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon, "Emotion recognition in the wild challenge 2013," in *Proceedings of ICMI*, 2013.

[3] Björn Schuller, Stefan Steidl, and Anton Batliner, "The interspeech 2009 emotion challenge.," in *Proceedings of INTERSPEECH*, 2009, pp. 312–315.

[4] Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic, "Avec 2012: the continuous audio/visual emotion challenge," in *Proceedings of ICMI*, 2012, pp. 449–456.

[5] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[6] Tauhidur Rahman, Soroosh Mariooryad, Shalini Keshavamurthy, Gang Liu, John HL Hansen, and Carlos Busso, "Detecting sleepiness by fusing classifiers trained with novel acoustic features.," in *Proceedings of INTERSPEECH*, 2011, pp. 3285–3288.

[7] Rui Xia and Yang Liu, "Using i-vector space model for emotion recognition.," in *Proceedings of INTERSPEECH*, 2012.

[8] Viktor Rozgic, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, Aravind Namandi Vembu, and Rohit Prasad, "Emotion recognition using acoustic and lexical features.," in *Proceedings of INTERSPEECH*, 2012.

[9] Emily Mower, Maja J Mataric, and Shrikanth Narayanan, "A framework for automatic human emotion classification using emotion profiles," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1057–1070, 2011.

[10] Arman Savran, Houwei Cao, Miraj Shah, Ani Nenkova, and Ragini Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," in *Proceedings of ICMI*. ACM, 2012, pp. 485–492.

[11] Jun Deng, Zixing Zhang, Erik Marchi, and Björn Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proceedings of ACII*, Geneva, Switzerland, September 2013, HUMAINE Association, pp. 511–516, IEEE.

[12] Zixing Zhang, Jun Deng, Erik Marchi, and Björn Schuller, "Active learning by label uncertainty for acoustic emotion recognition.," in *Proceedings of INTERSPEECH*, 2012.

[13] Wenjing Han, Haifeng Li, Huarbin Ruan, Lin Ma, Jiayin Sun, and Björn Schuller, "Active learning for dimensional speech emotion recognition," in *Proceedings of INTERSPEECH*, 2013.

[14] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[15] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of ICASSP*, 2013.

[16] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng, "Semantic compositionality through recursive matrix-vector spaces," in *Proceedings of EMNLP*. Association for Computational Linguistics, 2012, pp. 1201–1211.

[17] Yelin Kim and Emily Mower Provost, "Deep learning for robust feature generation in audio-visual emotion recognition," in *Proceedings of ICASSP*, 2013.

[18] Duc Le and Emily Mower Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks.," in *Proceedings of Automatic Speech Recognition and Understanding (ASRU)*, 2013.

[19] Angeliki Metallinou, Martin Wollmer, Athanasios Katsamanis, Florian Eyben, Björn Schuller, and Shrikanth Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *Affective Computing, IEEE Transactions on*, vol. 3, no. 2, pp. 184–198, 2012.

[20] Thurid Vogt and Elisabeth André, "Improving automatic emotion recognition from speech via gender differentiation," in *Proceedings of Language Resources and Evaluation Conference (LREC 2006), Genoa*. Citeseer, 2006.

[21] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, pp. 3371–3408, 2010.

[22] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[23] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller, and Shrikanth S Narayanan, "The interspeech 2010 paralinguistic challenge.," in *Proceedings of INTERSPEECH*, 2010, pp. 2794–2797.

[24] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.