DEEP LEARNING FOR ROBUST FEATURE GENERATION IN AUDIOVISUAL EMOTION RECOGNITION

Yelin Kim, Honglak Lee, and Emily Mower Provost *

University of Michigan
Electrical Engineering and Computer Science, Ann Arbor, Michigan, USA
{yelinkim, honglak, emilykmp}@umich.edu

ABSTRACT

Automatic emotion recognition systems predict high-level affective content from low-level human-centered signal cues. These systems have seen great improvements in classification accuracy, due in part to advances in feature selection methods. However, many of these feature selection methods capture only linear relationships between features or alternatively require the use of labeled data. In this paper we focus on deep learning techniques, which can overcome these limitations by explicitly capturing complex non-linear feature interactions in multimodal data. We propose and evaluate a suite of Deep Belief Network models, and demonstrate that these models show improvement in emotion classification performance over baselines that do not employ deep learning. This suggests that the learned high-order non-linear relationships are effective for emotion recognition.

***Index Terms*—** emotion classification, deep learning, multimodal features, unsupervised feature learning, deep belief networks

1. INTRODUCTION

Emotion recognition is the process of predicting the high-level affective content of an utterance from the low-level signal cues produced by a speaker. This process is complicated by the inherent multimodality of human emotion expression (e.g., facial and vocal expression). This multimodality is characterized by complex high-dimensional and non-linear cross-modal interactions [1]. Previous research has demonstrated the benefit of using multimodal data in emotion recognition tasks and has identified various techniques for generating robust multimodal features [2–6]. However, although effective, these techniques do not take advantage of the complex non-linear relationship that exists between the modalities of interest, or alternatively require the use of labeled data. In this work, we apply deep learning techniques to provide robust features for audio-visual emotion recognition.

Emotion recognition accuracy relies heavily on the ability to generate representative features. However, this is a very challenging problem. Emotion states do not have explicit temporal boundaries and emotion expression patterns often vary across individuals [7]. This problem is further complicated by the high dimensionality of the audio-visual feature space. Consequently, accurate modeling generally requires a reduction of the original input feature space. This reduction is commonly accomplished using feature selection, a method that identifies a subset of the initial features that provide enhanced classification accuracy [8]. However, it is not yet clear whether it is more advantageous to select a subset of emotionally relevant features or to capture the complex interactions across all features considered. In this paper, we demonstrate the effectiveness of Deep Belief Networks (DBN) for multimodal emotion feature generation. We learn multi-layered DBNs that capture the non-linear dependencies of audio-visual features while reducing the dimensionality of the feature space.

There has been a substantial body of work on feature representation, extraction, and selection methods in the emotion recognition field in the last decade. Our work is motivated by the discovery of methods for learning multiple layers of adaptive features using DBNs [9]. Research has demonstrated that deep networks can effectively generate discriminative features that approximate the complex non-linear dependencies between features in the original set. These deep generative models have been applied to speech and language processing, as well as emotion recognition tasks [10–12]. In speech processing, Ngiam et al. [13] proposed and evaluated deep networks to learn audio-visual features from spoken letters. In emotion recognition, Brueckner et al. [14] found that the use of a Restricted Boltzmann Machine (RBM) prior to a two-layer neural network with fine-tuning could significantly improve classification accuracy in the Interspeech automatic likability classification challenge [15]. The work by Stuhlsatz et al. [16] took a different approach for learning acoustic features in speech emotion recognition using Generalized Discriminant Analysis (GerDA) based on Deep Neural Networks (DNNs). While the present study is related to recent approaches in multi-modal deep learning and the application of deep learning techniques to emotion data, it focuses on non-linear audio-visual feature learning for emotion, which has not been extensively explored in the emotion recognition domain.

In the current work we present a suite of DBN models to investigate audio-visual feature learning in the emotion domain. We compare two methodologies: (1) unsupervised feature learning (DBN) and (2) secondary supervised feature selection. We first build an unsupervised two-layer DBN, enforcing multi-modal learning as introduced by [13]. We augment this DBN with two types of feature selection (FS): 1) before DBN training to assess the benefit of feature learning exclusively from an emotionally-salient subset of the original features and 2) after DBN training to assess the advantage of reducing the learned feature space in a supervised context. We compare this to the performance of a three-layer DBN model. Our baseline is a Support Vector Machine that uses subsets of the original feature space selected using supervised and unsupervised feature selection. The results provide important insight into feature learning methods for multimodal emotion data.

The results show that the DBN models outperform the baseline models. Further, our results demonstrate that the three-layer DBN outperforms the two-layer DBN models for emotionally subtle data.

*This work is supported by the National Science Foundation (NSF RI 1217183)

This suggests that unsupervised feature learning can be used in lieu of supervised feature selection for this data type. In addition, the relative performance improvement of the three-layer model for subtle emotions suggests that these complex feature relationships are particularly important for identifying subtle emotional cues. This is an important finding given the challenges inherent in and need for recognizing emotions elicited in realistic scenarios [17].

2. RELATED WORK

2.1. Feature Selection in Emotion Recognition

In this section, we discuss the feature selection techniques that are used extensively in emotion research including: Forward Selection, Information Gain (IG), and Principal Component Analysis (PCA). These techniques are either supervised (forward selection and IG) or use representations based on the linear dependencies between the original features (PCA).

Forward feature selection is a greedy algorithm that sequentially selects features that increase the overall classification accuracy. This method has been widely used in many machine learning applications, including emotion recognition tasks [18]. Although this method can identify a subset of good features for classification, it may not be suitable if there are groups of features with complex relationships due to the greedy nature of the approach. IG based feature selection methods are also commonly used in emotion recognition [19,20]. This method ranks features by calculating the reduction in the entropy of class labels given knowledge of each feature. In general, however, it does not search for feature interactions. Furthermore, both forward selection and IG methods require labeled data during the feature selection process.

PCA and its variants (e.g., Principal Feature Analysis, or PFA [21]) are broadly used in the emotion recognition literature [22–24]. PCA finds a linear projection of the base feature set to a new feature space where the new features are uncorrelated. The feature set can be reduced to retain a majority of the variance in the original feature space. Although this unsupervised method has been widely used in many emotion applications, the limitation is in its linear projection of the base features, which tends to obscure the emotion content [25]. PFA is an extension of PCA. It clusters the data in the PCA space and returns final features closest to the center of each cluster. This results in a feature set that maintains an approximation of the variance of the original set, while minimizing correlations between features. We leverage IG for our proposed deep learning feature selection methods, and IG and PFA for the baseline models.

2.2. Unsupervised Feature Learning and Deep Learning

Deep learning techniques (See [9] for a survey) have become increasingly popular in various communities including speech and language processing [10–12] and vision processing [26–30]. This progress has been facilitated by the recent discovery of more effective learning algorithms for constructing DBNs in an unsupervised context, for example exploiting single-layer building blocks such as Restricted Boltzmann Machines (RBMs) [31]. DBNs [32] learn hierarchical representation from data and can be effectively constructed by greedily training and stacking multiple RBMs.

RBMs are undirected graphical models that represent the density of input data $\mathbf{v} \in \mathbb{R}^D$ (referred to as "visible units") using binary latent variables $\mathbf{h} \in \{0,1\}^K$ (referred to as "hidden units"). In the RBM, there are no connections between units in the same layer, which makes it easy to compute the conditional probabilities.

In this work, we use Gaussian RBMs that employ real-valued visible units for training the first layer of the DBNs. We use Bernoulli-Bernoulli RBMs that employ binary visible and hidden units for training the deeper layers. In a Gaussian RBM, the joint probability distribution and energy function of $\mathbf{v}$ and $\mathbf{h}$ is as follows:

$$P(\mathbf{v},\mathbf{h}) = \frac{1}{Z}\exp(-E(\mathbf{v},\mathbf{h})) \tag{1}$$

$$E(\mathbf{v},\mathbf{h}) = \frac{1}{2\sigma^2}\sum_i v_i^2 - \frac{1}{\sigma^2}\left(\sum_i c_i v_i + \sum_j b_j h_j + \sum_{i,j} v_i W_{ij} h_j\right) \tag{2}$$

where $\mathbf{c} \in \mathbb{R}^D$ and $\mathbf{b} \in \mathbb{R}^K$ are the biases for visible and hidden units, respectively, $\mathbf{W} \in \mathbb{R}^{D\times K}$ are weights between visible units and hidden units, σ is a hyper-parameter, and Z is a normalization constant. The conditional probability distributions of the Gaussian RBM are as follows:

$$P(h_j = 1|\mathbf{v}) = sigmoid\left(\frac{1}{\sigma^2}(\sum_i W_{ij} v_i + b_j)\right) \tag{3}$$

$$P(v_i|\mathbf{h}) = \mathcal{N}\left(v_i; \sum_j W_{ij} h_j + c_i, \sigma^2\right) \tag{4}$$

The posteriors of the hidden units given visible units (Equation 3) form the generated features used in the classification framework. The parameters of the RBM ($\mathbf{W}$, $\mathbf{b}$, $\mathbf{c}$) are learned using contrastive divergence as in [33]. We use sparsity regularization [26] to penalize a deviation of expected activation of the hidden units from a low fixed level p. Given a training set $\{\mathbf{v}^{(1)},...,\mathbf{v}^{(m)}\}$, we include a regularization penalty of the form:

$$\lambda \sum_{j=1}^{K} \left| p - \frac{1}{m}\sum_{l=1}^{m} \mathbb{E}\left[h_j^{(l)}|\mathbf{v}^{(l)}\right]\right|^2 \tag{5}$$

where $\mathbb{E}[\cdot]$ is the conditional expectation given the data, λ is a regularization parameter, and p is a constant that specifies the target activation of the hidden unit h_j [26].

3. DATA

3.1. IEMOCAP Data

In this work, we use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database [34]. This database contains both motion capture markers and audio data from five pairs of actors (male-female). The subjects' facial movements were recorded using 53 infrared facial markers. The actors performed from scripted scenes and improvised scenarios. This method of collection allowed for both control over the affective content and naturalistic speaking styles.

The data were evaluated using categorical and dimensional labels (only categorical labels are used in this study). The categorical ground truth of the data was labeled by at least three evaluators. In this work, we only consider utterances with labels from the following set: *Angry, Happy, Neutral, Sad*. We use three types of utterances in this paper: (1) prototypical data (complete agreement on the affective state from evaluators), (2) non-prototypical data (majority agreement), and (3) a combined set of these two data types. There are 1430 utterances in prototypical data (Angry: 284, Happy: 707, Neutral: 123, Sad: 316 utterances) and 1588 utterances in non-prototypical data (Angry: 316, Happy: 498, Neutral: 455, Sad: 319 utterances), resulting in 3018 utterances in the combined set.

3.2. Audio-Visual Feature Extraction

The original audio features include both prosodic and spectral features, such as pitch, energy and mel-frequency filter banks (MFBs). MFBs have been shown to be better discriminative features than mel-frequency cepstral coefficients (MFCCs) in emotion recognition [35]. The original video features are based on Facial Animation Parameters (FAP), part of the MPEG-4 standard. FAPs describe the movement of the face using distances between particular points on the face. They have been widely used to capture facial expressions in the emotion recognition literature. The subset is chosen to include emotionally meaningful movements (e.g., eye squint, smile, etc.).

The final features are statistical functionals of the raw audio-visual features. These include mean, variance, lower and upper quantiles, and quantile range, giving a total of 685 features. Of these 685 features, 145 are auditory features and 540 are video features. The features are normalized on a per-speaker basis to mitigate speaker variation [20].

4. PROPOSED METHOD

4.1. Cross-Validation and Performance Evaluation

We use leave-one-speaker-out cross validation to ensure that the models are not overtraining to the affective styles of a particular speaker. We pre-train the DBN models (unsupervised) and search for the best hyper-parameters including: sparsity parameters and the number of final output nodes. We select our hyper-parameters using cross validation over the training data. We fix the number of hidden nodes of the two-layer DBNs, the sigma parameter for the first-layer Gaussian RBMs, and the L2 regularization parameter (Section 4.3). We select the best hyper-parameters for each data type: prototypical, non-prototypical, and combined.

We evaluate the performance of the baseline and DBN systems using Unweighted Accuracy (UA). UA is an average of the recall for each emotion class [17]. The unweighted accuracy better reflects overall accuracy in the presence of class imbalance.

4.2. Baseline Models

Our baseline models are two SVMs with radial basis function (RBF) kernels. The SVMs do not use features generated via deep learning techniques. The SVMs have radial basis function (RBF) kernels and are implemented using the Matlab Bioinformatics Toolkit. We train four emotion-specific binary SVMs in a self-vs.-other approach. The final emotion class label is assigned by identifying the model in which the test point is maximally far from the hyperplane as in as in [20].

Both models employ feature selection. The first uses IG [36] and the second uses PFA [21] feature selection (a supervised and unsupervised feature selection technique, respectively). IG is applied to each emotion class, resulting in four sets of emotion-specific features. Each emotion-specific SVM uses the associated emotion-specific feature subset. The number of features is chosen over {60, 120, 180} for each data type.

We optimized the baselines using leave-one-subject-out cross-validation for each data type (prototypical, non-prototypical, and combined data). The parameters include the number of selected features using IG and PFA, the value of the box constraint (C=1) for the soft margin in the SVM, and the scaling factor (sigma=8) in the RBF kernel.

We also compare our results with the maximal accuracy achieved from a previous work of Metallinou et al. [37], which utilizes the same IEMOCAP database as our work and introduces a decision-level Bayesian fusion over models using face, voice, and head movement cues. Although Metallinou's work used a different subset of the IEMOCAP database, this comparison supports the strong performance of our proposed method.

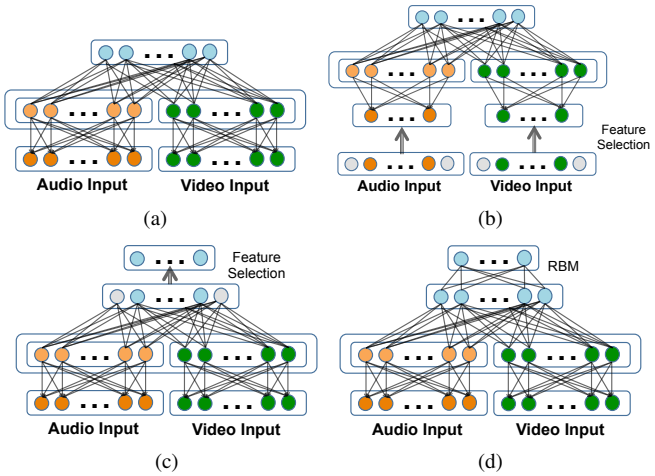

Fig. 1. Illustration of proposed models: (a) DBN2, (b) FS-DBN2, (c) DBN2-FS, and (d) DBN3.

4.3. Deep Belief Network Models

We experiment with four different DBN models in order to explore different non-linear dependencies between audio and motion-capture features. We also assess the utility of feature selection methods in these deep architectures (Figure 1).

Our basic DBN is a two-layer model and is a building block for the other variants. It learns the audio and video features separately in the first hidden layer. The learned features from the first layer are concatenated and used as the input to the second hidden layer as introduced in [13]. We call this the *DBN2* model (Figure 1(a)).

The other three DBN models are based on DBN2. Two involve feature selection and one is a three-layer DBN model. The two-layer models use supervised feature selection (IG) either prior to or post the unsupervised pre-training. The three-layer model reduces the feature dimensionality using a third RBM layer, invoking unsupervised feature learning. Thus, the three-layer model captures additional high-order non-linear dependencies of all features, whereas the models employing supervised feature selection use only emotionally salient features. The variants are defined as follows:

- FS-DBN2 is a two-layer DBN with feature selection prior to the training of the DBN2 model (Figure 1(b)).
- DBN2-FS is a two-layer DBN with feature selection on the final RBM output nodes (Figure 1(c)).
- DBN3 is a three-layer DBN that stacks an additional RBM on the second-layer RBM output nodes of the DBN2 model (Figure 1(d)).

The number of hidden units in the first layer is approximately 1.5x overcomplete for each audio feature (300 units from 145 audio features) and video feature (700 units from 540 video features), resulting in 1000 concatenated first layer hidden units. The number of second hidden units is fixed at 200 for DBN2, DBN2-FS, and DBN3.

Table 1. *Unweighted classification accuracy (%) for combined, non-prototypical, and prototypical data*

	Baseline		Proposed DBNs			
	IG	PFA	DBN2	DBN2-FS	DBN3	FS-DBN2
Combined	64.42	64.45	65.25	66.12	65.71	65.89
Non-Prot	55.81	55.99	56.89	56.97	57.70	56.07
Prot	73.38	70.02	70.46	72.96	73.78	72.77

For FS-DBN2, the number of second hidden units is fixed to 150 because the number of visible units is smaller compared to the other three DBN models.

The sparseness parameters are selected using leave-one-speaker-out cross-validation, while all other parameters (including hidden layer size and weight regularization) are kept fixed (See Section 4.1 for details). Since the number of video features is larger than the number of audio features, we select the sparsity parameters of bias for audio data and video data over $\{0.1, 0.2\}$ and $\{0.02, 0.1\}$, respectively. Also, the sparsity parameters of λ are selected over $\{2, 6, 10\}$ for audio features, while λ sparsity parameters are fixed at 5 for video features. Our preliminary results demonstrated that the λ value for the video features did not noticeably affect the results. The number of features selected at the final level (DBN2-FS) and the number of hidden units at the final level (DBN3) are selected over $\{50, 100, 150\}$.

For FS-DBN2, a total of 100 audio features and 200 video features are chosen using IG. We first pre-train a sparse RBM with 100-200 nodes for the audio features and 200-600 nodes for the video features. We select the sparsity parameters of bias over $\{0.1, 0.5\}$ for each RBM. λ is fixed as 5. Next, we concatenate the learned features and pre-train a first layer of DBN with 800 output nodes and the second layer with 150 nodes (Bernoulli-Bernoulli).

The output of each DBN is classified using the same SVM structure used in the baseline (Section 4.2).

5. RESULTS AND DISCUSSION

A summary of the emotion classification results can be seen in Table 1. The DBN models for the combined data achieve UAs ranging from 65.25% (DBN2) to 66.12% (DBN2-FS). All DBN models outperform the baseline models (the two baseline models perform comparably). The performance gap between the maximal UAs of proposed models and the PFA baseline is 1.67%.

The DBN models for the non-prototypical data achieve accuracies ranging from 56.70% (FS-DBN2) to 57.70% (DBN3). All DBNs outperform the baseline models (which again perform comparably). The performance gaps between the UAs of proposed models and baseline models range from 1.71% to 1.89%. We obtain a slight performance gain when using DBN3 compared to both DBN2-FS and FS-DBN2 for subtle or non-prototypical utterances (0.73% and 1.63% increase, respectively). This result is important given that the DBN3 model does not use any labeled data (unsupervised feature learning), whereas the FS-DBN2 model learns a new set of features from a previously identified subset of emotionally salient features and the DBN2-FS invokes feature selection at the output. This demonstrates that we can effectively use unsupervised feature learning, rather than supervised feature selection, for emotion recognition, even for emotionally subtle utterances (non-prototypical).

The DBN models for the prototypical data achieve accuracies ranging from 70.46% (DBN2) to the maximum of 73.78% (DBN3). The performance gap between the maximal UAs of the proposed models and maximal UAs of the baseline models (73.38% with IG) is 0.40%. The baseline models themselves achieve differing levels of accuracy; the IG baseline outperforms the PFA baseline by 3.36%. This may suggest that in emotionally clear utterances, supervised feature selection (emotion-specific IG) is preferable to unsupervised feature selection (PFA). The accuracy of the DBN3 model indicates that unsupervised feature learning can achieve comparable performance to supervised feature selection for emotionally clear utterances. Further, the DBN3 outperforms unsupervised feature selection (PFA baseline) by 3.76%, highlighting the potential importance of feature learning rather than unsupervised feature reduction for emotionally clear data.

The deep learning method performs comparably to the previous work of Metallinou et al. [37], 62.42%. Direct comparisons are not possible due to differences in the data subsets considered.

6. CONCLUSIONS

In this work, we investigate the utility of deep learning techniques for unsupervised feature learning in audio-visual emotion recognition. Our results demonstrate that DBNs can be used to generate audio-visual features for emotion classification, even in an unsupervised context. The comparison of the classification performances between the baseline and the proposed DBN models demonstrate that it is important to retain complex non-linear feature relationships (using deep learning techniques) in emotion classification tasks. The strongest performance gain is observed in the non-prototypical data. This is important in applications of automatic emotion recognition systems where emotional subtlety is common.

In our future work, we will investigate the comparative advantage of deep learning techniques with additional emotion corpora. We will also investigate deep modeling in the context of dynamic feature generation. Finally, the visualization of complex dependencies between either features or weights between hidden nodes of the DBNs may open a new gateway for the interpretation of audio-visual emotion data.

7. ACKNOWLEDGEMENTS

Thanks to Kihyuk Sohn, David Escott, Krithika Swaminatham, and Nattavut Yampikulsakul for many helpful discussions.

8. REFERENCES

[1] G.W. Taylor, G.E. Hinton, and S.T. Roweis, "Modeling human motion using binary latent variables," *Advances in neural information processing systems*, vol. 19, pp. 1345, 2007.

[2] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.-M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 205–211.

[3] D. Ververidis and C. Kotropoulos, "Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition," *Signal Processing*, vol. 88, no. 12, pp. 2956–2970, 2008.

[4] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2005, pp. 474–477.

[5] M. Wimmer, B. Schuller, D. Arsic, G. Rigoll, and B. Radig, "Low-level fusion of audio and video feature for multi-modal emotion recognition," in *3rd International Conference on Computer Vision Theory and Applications. VISAPP*, 2008, vol. 2, pp. 145–151.

[6] M. Pantic, G. Caridakis, E. André, J. Kim, K. Karpouzis, and S. Kollias, "Multimodal emotion recognition from low-level cues," *Emotion-Oriented Systems*, pp. 115–132, 2011.

[7] C.N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, pp. 1–23, 2012.

[8] M. El Ayadi, M.S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[9] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[10] N. Morgan, "Deep and wide: Multiple layers in automatic speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 7–13, 2012.

[11] A. Mohamed, G.E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 14–22, 2012.

[12] G. Sivaram and H. Hermansky, "Sparse multilayer perceptron for phoneme recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 23–29, 2012.

[13] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A.Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011, pp. 689–696.

[14] R. Brueckner and B Schuller, "Likability classification - a not so deep neural network approach," in *Proceedings of INTERSPEECH*, 2012.

[15] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, et al., "The interspeech 2012 speaker trait challenge," *Interspeech, Portland, Oregon*, 2012.

[16] A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5688–5691.

[17] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. Interspeech*, 2009, pp. 312–315.

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

[19] T. Polzehl, S. Sundaram, H. Ketabdar, M. Wagner, and F. Metze, "Emotion classification in childrens speech using fusion of acoustic and linguistic features," *Proceedings of INTERSPEECH-2009, Brighton, UK*, pp. 340–343, 2009.

[20] E. Mower, MJ. Mataric, and SS. 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.

[21] Y. Lu, I. Cohen, XS. Zhou, and Q. Tian, "Feature selection using principal feature analysis," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 301–304.

[22] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, "Of all things the measure is man: Automatic classification of emotions and inter-labeler consistency," in *Proc. ICASSP*, 2005, vol. 1, pp. 317–320.

[23] A. Metallinou, C. Busso, S. Lee, and S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2474–2477.

[24] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling," in *Proc. of INTERSPEECH Conference*, 2010, pp. 2362–2365.

[25] C. Busso and S.S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: a single subject study," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2331–2347, 2007.

[26] H. Lee, C. Ekanadham, and A. Ng, "Sparse deep belief net model for visual area v2," *Advances in neural information processing systems*, vol. 20, pp. 873–880, 2008.

[27] Y. Tang and C. Eliasmith, "Deep networks for robust visual recognition," in *International Conference on Machine Learning*. Citeseer, 2010, vol. 28.

[28] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Unsupervised learning of hierarchical representations with convolutional deep belief networks," *Communications of the ACM*, vol. 54, no. 10, pp. 95–103, 2011.

[29] K. Sohn, D.Y. Jung, H. Lee, and A.O. Hero, "Efficient learning of sparse, distributed, convolutional feature representations for object recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2643–2650.

[30] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1106–1114.

[31] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1, pp. 194–281, 1986.

[32] G.E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[33] G.E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[34] 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.

[35] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," *Proceedings of Interspeech 2007*, pp. 2225–2228, 2007.

[36] W. Duch, J. Biesiada, T. Winiarski, K. Grudzinski, and K. Grabczewski, "Feature selection based on information theory filters," in *Neural Networks and Soft Computing: Proceedings of the Sixth International Conference on Neural Networks and Soft Computing, Zakopane, Poland, June 11-15, 2002*. Physica Verlag, 2003, vol. 1, p. 173.

[37] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2462–2465.