EMOTION DETECTION IN SPEECH USING DEEP NETWORKS

Mohamed R. Amer, Behjat Siddiquie, Colleen Richey, and Ajay Divakaran*

SRI International, 201 Washington Rd. Princeton, NJ08540

ABSTRACT

We propose a novel staged hybrid model for emotion detection in speech. Hybrid models exploit the strength of discriminative classifiers along with the representational power of generative models. Discriminative classifiers have been shown to achieve higher performances than the corresponding generative likelihood-based classifiers. On the other hand, generative models learn a rich informative representations. Our proposed hybrid model consists of a generative model, which is used for unsupervised representation learning of short term temporal phenomena and a discriminative model, which is used for event detection and classification of long range temporal dynamics. We evaluate our approach on multiple audio-visual datasets (AVEC, VAM, and SPD) and demonstrate its superiority compared to the state-of-the-art.

Index Terms— Emotion Recognition; Deep Networks; Hybrid Models; CRF; CRBMs

1. INTRODUCTION

Detecting the emotional content in human speech is an important problem and has several interesting applications. Example applications include - use in tutoring systems to detect student state [1]; identifying distressed phone calls automatically [2]. However, recognizing emotions from speech is a very challenging problem, primarily because different people express emotions in different ways. Moreover, a lot of the emotional content in speech is contained in the form of language and when this linguistic content is removed, the "paralinguistic" problem becomes challenging even for humans (e.g. recognizing the emotions from speech content in a language one does not understand) [3].

A standard approach to solving this problem involves extracting framewise low-level descriptors (LLDs) from the audio signal and then using functional features to aggregate these features over the utterance level [1, 4]. A major drawback of these approaches is that they ignore the temporal dynamics of the phenomena both within and across utterances. While there has been some work on modeling the temporal dynamics of affect across utterances [5, 6] and within utterances [7], we show that modeling the multi-scale temporal dynamics leads to significant performance gains. Furthermore, we demonstrate that, starting from low level features such as spectrograms, deep temporal models can learn a rich and representative feature space which often outperforms hand-crafted features.

There are two key *contributions* of our work. First, we propose a hybrid model that consists of a temporal generative model that learns a rich and compact feature representation capable of encoding a variety of low level concepts and a discriminative model for high level reasoning. Secondly, we collect and report recognition results on a new dataset collected in a noisy environment.

Paper organization: Sec. 2 reviews prior work. Sec. 3 defines our model. Sec. 4 specifies our inference and learning algorithms. Sec. 5 compares our approach against the state-of-the art, followed by the conclusion in sec. 6.

2. PRIOR WORK

Most work on signal processing addressed the problem of event detection in speech using shallow models such as Gaussian Mixture Models [8], Dynamic Bayesian Networks, Conditional Random Fields [5, 6], and Support Vector Machines [4]. Even though the aforementioned models achieved good results, they are not expressive enough to capture higher level dynamics and semantics. A common approach for increasing the expressivity of the models is to use deep architectures. Many recent works on vision, speech, natural language processing use hierarchical deep networks to encode the data structure. In our work we focus on non-linear deep networks such as Deep Boltzmann Machines [9] and Neural Networks [10]. These models are known for their strong representational power and have been successfully used for many problems.

Deep Networks have been successfully used for audio analysis, speech recognition, and natural language processing [11]. Non-linear deep networks such as Boltzmann Machines [9] and Neural Networks [10] have the ability to learn a rich feature representation in an unsupervised manner, making them very powerful. Restricted Boltzmann Machines (RBMs) form the building blocks of deep networks models. These models are trained using the Contrastive Divergence (CD) algorithm, which enables deep networks to capture the distributions over the features efficiently and to learn complex representations [13]. RBMs can be stacked together to form deeper networks known as Deep Belief Networks (DBNs), which capture more complex representations. Recently, deep networks were successfully used for the problem of mul-

*The author is a student at Oregon State University and did this work while being an intern at SRI International.

timodal event detection [14], and for Multimodal emotion recognition [15]. However, unlike our proposed approach, these approaches do not explicitly account for the temporal nature of the data.

Temporal Deep Networks are capable of capturing the representation of a more temporally rich set of problems. Temporal Deep Networks include Conditional RBMs (CRBMs) [16], and Temporal RBMs (TRBMs) [17]. CRBMs and TRBMs have been successfully used in the audio domain, for example, phone recognition [18], and polyphonic music generation [19]. Recently, Deep Stacking Networks [20], a special type of deep model equipped with parallel and scalable learning, have been successfully used for frame-level phone classification [21], phone recognition, and information retrieval [10].

Deep Architectures can be divided into three groups, generative, discriminative, or hybrid models [11]. In hybrid architectures, the goal is discrimination, which is assisted with the outcomes of generative architectures via better optimization or/and regularization, combining the advantages of generative and discriminative models. Hybrid architectures have been successfully used in speech recognition and natural language processing [10, 22, 23]. The learning of the hybrid model parameters could be done in three ways: staged [24]; iterative [25]; joint [26]. The staged learning allows scalable learning of the model parameters. In our case, we use a staged hybrid architecture, where the generative part (CRBM) captures short term (intra-utterance) dynamics, while the discriminative part (CRF) captures the long term (inter-utterance) correlations. We have recently proposed a similar hybrid model for the problem of multi-modal event detection [27].

In this paper we evaluate the model for speech analysis on challenging audio datasets, especially in the context of inter-utterance and intra-utterance temporal dynamics. In the next section, we formulate our Hybrid model using a combination of a CRBM and a CRF.

3. THE HYBRID MODEL

The hybrid model allows us to combine the advantages of generative as well as discriminative models leading to a stronger classifier compared to purely generative models. Let $\mathbf{y}_t$ be the multi-class label vector at time t, $\mathbf{v}_t$ is the vector of raw features at time t, and $\mathbf{h}_t$ is a vector of the latent hidden variables. $\mathbf{v}_{<t}$ is the concatenated history vector of the visible. We define our hybrid model as:

$$\underbrace{p(\mathbf{y}_t, \mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t})}_{\text{Hybrid}} = \underbrace{p_{\text{D}}(\mathbf{y}_t|\mathbf{h}_t)}_{\text{Discriminative}} \cdot \underbrace{p_{\text{G}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t})}_{\text{Generative}}. \quad (1)$$

Our hybrid model $p(\mathbf{y}_t, \mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t})$ shown in (1) consists of two terms, a generative term $p_{\text{G}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t})$, and a discriminative term $p_{\text{D}}(\mathbf{y}_t|\mathbf{h}_t, \mathbf{v}_t)$. Fig. 1(b) shows an illustration of our hybrid model. In the following subsections, we first specify our generative CRBM model, followed by the discriminative CRF model.

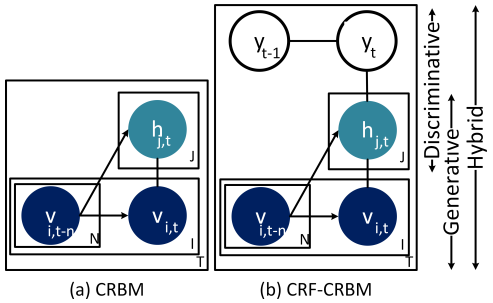

Fig. 1. (a) shows the CRBM model, where v are the visible nodes, h are the hidden nodes. (b) shows our hybrid model (CRF-CRBM), with a per frame label y.

3.1. The Conditional Restricted Boltzmann Machines

CRBM defines a probability distribution $p_{\text{G}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t})$ as a Gibbs distribution. Let $\mathbf{v}$ be a vector of visible nodes and $\mathbf{h}$ a vector of hidden nodes. The CRBM architecture is defined as fully connected between layers, with no lateral connections. This architecture implies that $\mathbf{v}$ and $\mathbf{h}$ are factorial given one of the two vectors. This allows for the exact computation of $p(\mathbf{v}|\mathbf{h})$ and $p(\mathbf{h}|\mathbf{v})$. CRBMs takes into account history from the previous time instances $[(t-N), \ldots, (t-1)]$ at time (t). This is done by treating the previous time instances as additional inputs. The additional inputs from previous time instances are modeled as directed autoregressive edges from the past N visible nodes and the past M hidden nodes, where, N does not have to be equal to M. The concatenated history vector is defined as $\mathbf{v}_{<t}$. Fig. 1(a) shows an illustration of our CRBM model. Doing so does not complicate inference[1].

We define our generative CRBM as:

$$p_{\text{G}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t}; \boldsymbol{\theta}_{\text{G}}) = \exp[-E_{\text{G}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t}; \boldsymbol{\theta}_{\text{G}})]/Z(\boldsymbol{\theta}_{\text{G}}),$$

$$\begin{aligned} E_{\text{G}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t}; \boldsymbol{\theta}_{\text{G}}) &= -\textstyle\sum_i (c_{i,t} - v_{i,t})^2/2 \\ &\quad - \textstyle\sum_j d_{j,t} h_{j,t} - \sum_{i,j} v_{i,t} w_{i,j} h_{j,t}, \end{aligned} \quad (2)$$

$$\begin{aligned} Z(\boldsymbol{\theta}_{\text{G}}) &= \textstyle\sum_{\mathbf{v},\mathbf{h}} \exp[-E_{\text{G}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t}; \boldsymbol{\theta}_{\text{G}})] \\ \boldsymbol{\theta}_{\text{G}} &= \{\mathbf{a}, \mathbf{b}, A, B, W\}, \end{aligned}$$

where $c_{i,t} = a_{i,t} + \sum_n A_{n,i} v_{n,<t}$ and $d_{j,t} = b_{j,t} + \sum_m B_{m,j} v_{m,<t}$ and A and B are matrices of concatenated vectors of previous time instances of $\mathbf{a}$ and $\mathbf{b}$.

Since our data is continuous, we define $p(v_i|\mathbf{h})$ as a multivariate Gaussian distribution with zero mean and unit covariance $p(v_i|\mathbf{h}) = \mathcal{N}(a_i + \sum_j h_j w_{ij}, 1)$. The conditional $p(h_j = 1|\mathbf{v})$ is defined as a logistic function $p(h_j = 1|\mathbf{v}) = \sigma(b_j + \sum_i v_i w_{ij})$[2], since we want the hidden layer to be binary (empirically proven to be better [16]).

The parameters of our generative model $\boldsymbol{\theta}_{\text{G}}$ are $\{\mathbf{a}, A\}$ and $\{\mathbf{b}, B\}$ the the biases for $\mathbf{v}$ and $\mathbf{h}$ respectively and the network weights' $\{W\}$.

[1] Some approximations have been made to facilitate efficient training and inference, more details are available in [16].

[2] The logistic function $\sigma(\cdot)$ for a variable x is defined as $\sigma(x) = (1 + exp(-x))^{-1}$.

3.2. Conditional Random Fields

The CRBMs are very effective for learning and representing short term temporal phenomena. In our problem we also need to model long range temporal dynamics. With this requirement in mind, we choose Conditional Random Fields (CRFs) [28] as our discriminative model. Let $\mathbf{y}_t$ be the label of the sequence at time t and $\mathbf{h}_t$ to be the output of the CRBM (2), which serves as an input to the CRF as shown in Fig. 1(b), f_j^1 is a transition feature function, f_k^2 is a state feature function, with ω_j^1 the transition component of the parameters and ω_k^2 the state component of the parameters, and Z is the partition function to ensure the proper normalization of the model. Our discriminative term, modeled by $p_{\mathrm{D}}(\mathbf{y}_t|\mathbf{h}_t;\boldsymbol{\theta}_{\mathrm{D}})$ is defined as:

$$\begin{gathered} p_{\mathrm{D}}(\mathbf{y}_t|\mathbf{h}_t;\boldsymbol{\theta}_{\mathrm{D}}) = \exp[E_{\mathrm{D}}(\mathbf{y}_t|\mathbf{h}_t;\boldsymbol{\theta}_{\mathrm{D}})]/Z(\boldsymbol{\theta}_{\mathrm{D}}), \\ E_{\mathrm{D}}(\mathbf{y}_t|\mathbf{h}_t;\boldsymbol{\theta}_{\mathrm{D}}) = \textstyle\sum_j \omega_j^1 f_j^1(\mathbf{y}_{t-1},\mathbf{y}_t,\mathbf{h}_t) \\ + \textstyle\sum_k \omega_k^2 f_k^2(\mathbf{y}_t,\mathbf{h}_t), \\ Z(\boldsymbol{\theta}_{\mathrm{D}}) = \textstyle\sum_{\mathbf{y}} E_{\mathrm{D}}(\mathbf{y}|\mathbf{h}_t;\boldsymbol{\theta}_{\mathrm{D}}), \quad \boldsymbol{\theta}_{\mathrm{D}} = \{\boldsymbol{\omega}^1,\boldsymbol{\omega}^2\}. \end{gathered} \tag{3}$$

In the following section we specify our inference and learning algorithms.

4. INFERENCE AND LEARNING

Inference on a CRBM is not much different from that on a RBM. The hidden nodes at time (t) are conditionally independent given data at previous time instances $[(t-N),\ldots,(t-1)]$. Inference is done in a layer-wise manner by activating a hidden layer given the visible layer using the conditional independence advantage of the CRBM model $p(h_j=1|\mathbf{v})$. Given a new observation sequence $\mathbf{h}_t$ and model parameters θ obtained during training, our goal is to predict the label $\hat{\mathbf{y}}_t$. This can be computed by maximizing over all labels as follows:

$$\hat{\mathbf{y}}_t = \arg\max_{\mathbf{y}} p_{\mathrm{D}}(\mathbf{y}_t|\mathbf{h}_t;\boldsymbol{\theta}_{\mathrm{D}}). \tag{4}$$

Note that the CRF model assigns a label to each node of the sequence.

Learning the RBM parameters using maximum likelihood learning is quite slow. However, learning can be significantly speeded up if we approximately follow the gradient of another function, in this case Contrastive Divergence (CD) [29]. The learning rules are derived using CD, where $\langle\cdot\rangle_{data}$ is the expectation with respect to the data distribution and $\langle\cdot\rangle_{recon}$ is the expectation with respect to the reconstructed data. The reconstruction is generated by first sampling $p(h_j=1|\mathbf{v})$ for all the hidden nodes in parallel. The visible nodes are then generated by sampling $p(v_i|\mathbf{h})$ for all the visible nodes in parallel.

$$\begin{gathered} \Delta w_{i,j} \propto \langle v_i h_j\rangle_{data} - \langle v_i h_j\rangle_{recon}, \\ \Delta a_i \propto \langle v_i\rangle_{data} - \langle v_i\rangle_{recon}, \\ \Delta b_j \propto \langle h_j\rangle_{data} - \langle h_j\rangle_{recon}, \\ \Delta A_{k,i,t-n} \propto v_{k,t-n}(\langle v_{i,t}\rangle_{data} - \langle v_{i,t}\rangle_{recon}), \\ \Delta B_{i,j,t-m} \propto v_{i,t-m}(\langle h_{j,t}\rangle_{data} - \langle h_{j,t}\rangle_{recon}). \end{gathered} \tag{5}$$

The update equations of the dynamically changing bases $\Delta\mathbf{c}$ and $\Delta\mathbf{d}$ are obtained by first updating $\Delta A_{k,i,t-q}$ and $\Delta B_{i,j,t-q}$ and then combining them with $\Delta a_{i,t}$ and $\Delta b_{j,t}$. The discriminative learning is done by maximum likelihood estimation of $\boldsymbol{\theta}_{\mathrm{D}}$ in (3) using [30].

5. EXPERIMENTS

For evaluating the performance of our approach, we compare the performance of different combinations of hybrid models SVM-RBM, CRF-RBM, SVM-CRBM, and CRF-CRBM (Fig. 1) against models trained on the raw features, which demonstrates the importance of the generative model with and without temporal modeling. Furthermore, we compare our approach against the state-of-the-art on multiple datasets.

Datasets and Implementation Details: We evaluate our model on three different datasets. Each dataset contains audio data. For one of the datasets we use hand-crafted (predefined) features, while for the remaining two datasets we rely on raw features (audio spectrograms). Our experiments shows that we can learn good feature representations from CRBMs using both hand-crafted as well as raw features.

AVEC [4] is an audio-visual dataset for single person affect analysis. The dataset involves users interacting with emotionally stereotyped virtual characters operated by a human. The dataset has been annotated with binary labels for four different affective dimensions - Activation, Expectation, Power and Valence. The dataset is divided into two sets, 31 sequences for training and 32 sequences for testing. The dataset comes with pre-computed audio and video features; refer to [4] for details. For the purposes of our experiments, we focus on the *Audio Subchallenge* which uses the audio features exclusively. We apply PCA on the extracted features and reduce each of the audio features to 100 dimensions. We choose a CRBM with a temporal order N = 8, with the first hidden layer being over-complete, consisting of 150 nodes. We use the AVEC dataset to compare against [5, 6, 8].

VAM: The Vera-Am-Mittag (VAM) corpus [31] consists of recordings taken from a talk show. The corpus contains audio-visual data from spontaneous and unstaged discussions between the guests of the talk show. The audio recordings were manually segmented to the utterance level and each utterance has been labeled with three emotion primitives: Valence, Activation and Dominance, on a continuous valued scale by multiple human annotators. Here again we use only the audio data for recognizing emotions. From the audio clips, we extract the spectrogram using OpenSMILE [32] and simply use the spectrogram as our feature. We choose a CRBM with a temporal order N = 8, with the first hidden layer being over-complete, consisting of 100 nodes.

SPD: We collected the Seattle Police Database (SPD), which contains audio recordings of police-civilian interactions. Unlike *AVEC* and *VAM*, the *SPD* has been recorded in "in the wild" and hence contains a considerable amount of noise. The recordings were segmented to an utterance level and were annotated by multiple annotators as belonging to four different categories: *calm*, *slightly-agitated*, *agitated* and

extremely-agitated. Here again we use a spectrogram as our feature vector and use a CRBM with a temporal order N = 8, with the first hidden layer being over-complete, consisting of 100 nodes.

Results: *Table 1* shows results on AVEC dataset. We report the weighted average classification accuracy of each of the four affective labels as well as the mean classification accuracy on the *Audio Subchallenge*. We use the training set for learning the classifier and report the results on the development set. We compare our approach against the baseline [4], as well as previously published results on this dataset [5, 6, 8] and we can see that our approach performs favorably compared to the state-of-the-art.

Table 2 shows the classification results on the VAM dataset. We use Leave-One-Speakers-Group-Out (LOSGO) cross-validation as in [3, 7] and compare the unweighted Accuracy. Our results are comparable to [3, 7] despite the fact that we use a spectrogram as our feature vector, while both [3, 7] use a hand-crafted 6552 dimensional feature vector. Moreover, we can also see that deep learning {SVM-RBM, SVM-CRBM, CRF-RBM, CRF-CRBM} significantly improves performance over the raw features {(SVM-RAW, CRF-RAW)}.

Table 3 shows the classification results on the SPD dataset. The dataset consists of 4 different classes: Neutral(N), Slightly Agitated (SA), Agitated (A) and Extremely Agitated (EA) and contains 1072 samples - 421(N), 391(SA), 226(A), 34(EA). We divide the dataset into training and test sets consisting of about 50% of the total data and compare the mean classification accuracy on the test set in 6 different scenarios. The first five scenarios are binary classification problems - T1: Neutral as one class vs Slightly Agitated, Agitated and Extremely Agitated as the other class, T2: Neutral as one class vs Agitated and Extremely Agitated as the other class, T3: N vs SA, T4: N vs A, T5: N vs EA. The sixth scenario T6, is a multi-class classification problem. From the results we can see that it is easier to distinguish between the more extreme classes and the performance goes down when Slightly Agitated is added. Here again we can see that Deep Learning results in an increased performance compared to raw features.

Discussion: In AVEC, the *CRF-CRBM* model gives the best performance. This is due to the presence of long term dynamics justifying the use of CRF model. In VAM and SPD, using deep learning improves the overall feature representation (short term dynamics), but due to the lack of long term dynamics, an SVM classifier is sufficient to capture the emotions.

6. CONCLUSION

We have proposed a hybrid model comprising of temporal generative and discriminative models for detection and recognition of emotional content in speech. We employ a deep networks based temporal generative model which enables us to learn a rich feature representation to model the short term (intra-utterance) temporal characteristics. The discriminative component of our model consists of a CRF, which enables modeling long range (inter-utterance) temporal dependencies leading to a superior classification performance. An extensive experimental evaluation on three different datasets demonstrates the superiority of our approach over the state-of-the-art.

Model	A	E	P	V	Mean
baseline [4]	63.7	63.2	65.6	58.1	62.6
HMM [8]	66.9	62.9	63.2	65.7	64.6
LDCRF [5]	74.9	68.4	67.0	63.7	68.5
HCRF [6]	73.4	65.5	68.7	**70.0**	**69.4**
SVM-RAW (ours)	63.7	63.2	65.6	58.1	64.8
CRF-RAW (ours)	**76.9**	65.5	68.7	61.7	68.1
SVM-RBM (ours)	62.1	63.0	64.0	58.3	61.8
CRF-RBM (ours)	73.2	67.7	68.3	61.1	67.6
SVM-CRBM (ours)	67.0	67.6	65.2	63.3	65.8
CRF-CRBM (ours)	72.3	**69.2**	**70.5**	65.3	**69.2**

Table 1. Classification accuracy (in %) on the AVEC dataset.

Model	A	P	V	Mean
HMM/GMM-RAW [7]	76.5	N.A.	49.2	N.A.
GerDA-RAW [3]	**78.4**	N.A.	**52.4**	N.A.
SVM-RAW [3]	72.1	N.A.	48.1	N.A.
CRF-RAW (ours)	71.0	69.5	50.0	63.5
SVM-RBM (ours)	76.2	**72.7**	50.0	**66.3**
CRF-RBM (ours)	74.7	70.5	50.0	65.0
SVM-CRBM (ours)	75.1	72.3	50.0	65.8
CRF-CRBM (ours)	73.8	70.0	50.0	64.6

Table 2. Classification accuracy (in %) on the VAM dataset.

Model/ Scenario	SVM-RAW	CRF-RAW	SVM-RBM	CRF-RBM	SVM-CRBM	CRF-CRBM
T1	63.5	72.8	72.8	**75.3**	71.5	73.8
T2	81.5	83.5	81.2	**85.6**	83.3	83.7
T3	59.5	63.4	63.9	64.9	**65.9**	62.8
T4	82.1	81.6	82.1	79.7	**82.4**	78.1
T5	95.2	97.1	**97.4**	95.9	96.5	85.1
T6	51.6	52.2	52.3	52.6	**53.8**	51.9
Mean	72.2	75.1	75.0	**75.6**	**75.6**	72.6

Table 3. Classification results (in %) on the SPD dataset.

Acknowledgement:

This work was supported by The Defense Advanced Research Projects Agency under Army Research Office Contract Number W911NF-12-C-0001. The views, opinions, and/or findings contained in this paper are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the DARPA or the DoD.

7. REFERENCES

[1] D. Litman and K. Forbes, "Recognizing emotions from student speech in tutoring dialogues," in *ASRU*, 2003.

[2] P. Belin, S. Fillion-Bilodeau, and F. Gosselin, "The montreal affective voices: a validated set of nonverbal affect bursts for research on auditory affective processing," in *Behavior Research Methods*, 2008.

[3] A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *ICASSP*, 2011.

[4] B. Schuller and et al., "Avec 2011 -the first international audio visual emotion challenge," in *ACII*, 2011.

[5] G. Ramirez, T. Baltrusaitis, and L. P. Morency, "Modeling latent discriminative dynamic of multi-dimensional affective signals," in *ACII*, 2011.

[6] B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney, "Affect analysis in natural human interactions using joint hidden conditional random fields," in *ICME*, 2013.

[7] Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Gerhard Rigoll, and Andreas Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *ASRU*, 2009.

[8] M. Glodek and et al., "Multiple classifier systems for the classification of audio-visual emotional states," in *ACII*, 2011.

[9] Y. Bengio, "Learning deep architectures for ai," in *FTML*, 2009.

[10] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network leaning for speech recognition and related applications: An overview," in *ICASSP*, 2013.

[11] L. Deng and D. Yu, "Deep learning for signal and information processing," in *FTML*, 2013.

[12] L. Zhu, Y. H. Chen, and A. Yuille, "Recursive compositional models for computer vision," *Journal of Mathematical Imaging and Vision*, 2011.

[13] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," in *NC*, 2006.

[14] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A.Y. Ng, "Multimodal deep learning," in *ICML*, 2011.

[15] Yelin Kim, Honglak Lee, and Emily Mower Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *ICASSP*, 2013.

[16] G. W. Taylor and et. al., "Modeling human motion using binary latent variables," in *NIPS*, 2007.

[17] I. Sutskever and G. E. Hinton, "Learning multilevel distributed representations for high-dimensional sequences," in *AISTATS*, 2007.

[18] A. R. Mohamed and G. E. Hinton, "Phone recognition using restricted boltzmann machines," in *ICASSP*, 2009.

[19] N. B. Lewandowski, Y. Bengio, and P. Vincent, "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription," in *ICML*, 2012.

[20] L. Deng, D. Yu, and J. Platt, "Scalable stacking and learning for building deep architectures," in *Interspeech*, 2012.

[21] B. Hutchinson, L. Deng, and D. Yu, "Tensor deep stacking networks," in *TPAMI*, 2013.

[22] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," in *ICASSP*, 2012.

[23] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Interspeech*, 2011.

[24] N. Smith and M. Gales, "Speech recognition using svms," in *NIPS*, 2002.

[25] A. Fujino, N. Ueda, and K. Saito, "Semi-supervised learning for a hybrid generative/discriminative classifier based on the maximum entropy principle," in *TPAMI*, 2008.

[26] H. Larochelle and Y. Bengio, "Classification using discriminative restricted boltzmann machines," in *ICML*, 2008.

[27] M. R. Amer, B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney, "Multimodal fusion using dynamic hybrid models," in *WACV*, 2014.

[28] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001.

[29] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," in *NC*, 2002.

[30] Mark Schmidt, "Ugm: Matlab code for undirected graphical models," 2012.

[31] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *ICME*, 2008.

[32] Florian Eyben, Martin Wollmer, and Bjorn Schuller, "opensmile: The munich versatile and fast open-source audio feature extractor," in *ACM MM*, 2010.