EMOTION RECOGNITION FROM SPONTANEOUS SPEECH USING HIDDEN MARKOV MODELS WITH DEEP BELIEF NETWORKS

Duc Le and Emily Mower Provost

University of Michigan
Computer Science and Engineering, Ann Arbor, MI 48109
{ducle, emilykmp}@umich.edu

ABSTRACT

Research in emotion recognition seeks to develop insights into the temporal properties of emotion. However, automatic emotion recognition from spontaneous speech is challenging due to non-ideal recording conditions and highly ambiguous ground truth labels. Further, emotion recognition systems typically work with noisy high-dimensional data, rendering it difficult to find representative features and train an effective classifier. We tackle this problem by using Deep Belief Networks, which can model complex and non-linear high-level relationships between low-level features. We propose and evaluate a suite of hybrid classifiers based on Hidden Markov Models and Deep Belief Networks. We achieve state-of-the-art results on FAU Aibo, a benchmark dataset in emotion recognition [1]. Our work provides insights into important similarities and differences between speech and emotion.

Index Terms— emotion classification, deep belief networks, spontaneous speech, FAU Aibo, dynamic modeling

1. INTRODUCTION

Emotion expression is a complex and dynamic process. This complexity has prompted investigations into appropriate modeling strategies to capture the temporal aspects of this behavior. However, the temporal properties of emotion are still not well understood. We address this challenge by utilizing a dynamic frame-level modeling approach with Deep Belief Network (DBN) and assessing the efficacy of our models on FAU Aibo [1], a benchmark dataset in the emotion recognition community.

In the automatic speech recognition (ASR) literature, Mohamed et al. [2] found that acoustic models based on DBN outperformed those based on Gaussian Mixture Model (GMM) on the TIMIT phone recognition task. They argued that DBNs are better at exploiting structural information embedded in high-dimensional data. Motivated by this work, we investigate whether or not emotion recognition, which has to map relatively long speech utterances to high-level emotion classes, can similarly benefit from DBN's modeling power.

We trained a suite of hybrid classifiers which used Hidden Markov Models (HMMs) to capture the temporal property of emotion and DBNs to estimate the emission probabilities. We analyzed the results of these classifiers in terms of the number of HMM states and size of input windows to the DBN. The best result was achieved by combining classifiers using 37-frame windows and different HMM architectures, yielding an unweighted average recall (UAR) of **45.60%** on FAU Aibo's 5-class problem. With speaker-specific z-normalization, we obtained a UAR of **46.36%**. Respectively, these results are significantly better than 44.0% and 44.8% UAR, the current state of the art without [3] and with [4] speaker normalization (one-tailed binomial test, $p \approx 0.002$).

Our experimental results demonstrated that the optimal model parameters for emotion differed from those for speech recognition in some important aspects. Compared to speech, emotion recognition required a smaller learning rate, larger input windows to the DBN, and fewer HMM states. These observations suggest that emotion lies on a different decision space, spans a longer time window, and has less well understood temporal dynamics compared to speech.

2. RELATED WORK

2.1. Aibo Benchmark

We approach the 5-class problem of the FAU Aibo dataset as specified in the 2009 Interspeech Emotion Challenge [5]. FAU Aibo is a spontaneous emotion corpus consisting of speech recordings of 51 children at two different schools, Ohm and Mont, interacting with Sony's pet robot Aibo. The speech recordings were segmented manually into utterances based on syntactic-prosodic criteria, each of which was then assigned one of five class labels: **A**nger, **E**mphatic, **N**eutral, **P**ositive, and **R**est. In the challenge, data from one school (Ohm) was used for training and the other (Mont) was used for testing. In total, the training and test sets have 9,959 and 8,257 utterances, respectively. See [1] for a detailed description of FAU Aibo.

Due to the extreme imbalance of the test set (7.4% **A**, 18.26% **E**, 65.12% **N**, 2.6% **P**, 6.62% **R**), a classifier's per-

formance is measured with unweighted average recall (UAR), defined as the mean recall rate over five emotion classes. The best baseline system in [5] achieved a 38.2% UAR with static modeling using linear-kernel Support Vector Machine (SVM). The best performing system in the 2009 challenge achieved a UAR of 41.65% by employing different GMM training methods [6]. The five trailing participants all came close to the winning classifier (within 0.5% absolute); see [7] for the complete list.

Since the challenge, several improvements have been made on this 5-class problem. Hassan et al. [8] achieved a 42.7% UAR by applying importance weights within a SVM to compensate for differences between training and testing conditions. Attabi et al. [3] used GMM to model multiple windowed spectrum estimates of Perceptual Linear Prediction (PLP) coefficients, resulting in a 44.0% UAR. The best known result, 44.8% UAR, was achieved with a two-pass system in which a high-level SVM classified each test utterance using ranking scores obtained from five low-level SVMs, one for each emotion [4]. However, this work made use of speaker identity information in the test set, which was not available in the original challenge.

Our work differs from the above papers in that we used dynamic frame-level modeling with DBN-based acoustic models, whereas most previous work used GMM-based acoustic models and/or static modeling with SVM.

2.2. Deep Belief Networks

A DBN consists of a stack of Restricted Boltzmann Machines (RBMs) trained greedily layer by layer. RBMs are undirected graphical models with two sets of visible and hidden units connected as a complete bipartite graph. Two types of RBMs, Bernoulli and Gaussian, are commonly used. In Bernoulli RBMs, both visible and hidden units are binary: $\mathbf{v} \in \{0,1\}^D$ and $\mathbf{h} \in \{0,1\}^K$, where D and K denote the number of visible and hidden units, respectively. In Gaussian RBMs, visible units can take on real numbers: $\mathbf{v} \in \mathbb{R}^D$. The joint probability distribution between $\mathbf{v}$ and $\mathbf{h}$ can be written as:

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

where Z is a normalization constant and $E(\mathbf{v},\mathbf{h})$ is an energy function. For Bernoulli RBMs, the energy function is:

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{D}\sum_{j=1}^{K} W_{ij}v_i h_j - \sum_{i=1}^{D} b_i v_i - \sum_{j=1}^{K} a_j h_j \quad (2)$$

where W_{ij} denotes the weight of the undirected edge connecting visible node v_i and hidden node h_j, and $\mathbf{a}$ and $\mathbf{b}$ are the bias terms for the hidden and visible units, respectively. For Gaussian RBMs, assuming the visible units have zero mean and unit variance, the energy function is:

$$E(\mathbf{v},\mathbf{h}) = \sum_{i=1}^{D}\frac{(v_i - b_i)^2}{2} - \sum_{i=1}^{D}\sum_{j=1}^{K} W_{ij}v_i h_j - \sum_{j=1}^{K} a_j h_j \quad (3)$$

An RBM is pre-trained generatively to maximize the data log-likelihood $\log P(\mathbf{v})$. The hidden layer's output of one RBM can be treated as input to another RBM which is then trained separately from the previous model. This stack of generatively pre-trained RBMs constitutes a DBN which can then be discriminatively fine-tuned as an Artificial Neural Network (ANN). The weights initialized by pre-training help the model avoid bad local minima, which can be a serious problem for deep networks. In this paper, we also refer to a pre-trained ANN as a DBN. See [9–11] for a detailed description of DBNs and training methodologies.

2.3. Deep Learning and Emotion Recognition

Deep learning techniques have found recent successes in various communities including computer vision [12–15], speech and language processing [2, 16–18], and emotion recognition [19–22]. Stuhlsatz et al. [19] used generatively pre-trained ANNs to learn discriminative features of low dimension and found improvement in both weighted and unweighted recall on multiple emotion corpora. Schmidt and Kim [20] used DBNs to learn high-level features directly from magnitude spectra and achieved good performance on music emotion recognition compared to other feature extraction schemes. Brueckner and Schuller [21] applied static modeling with DBN on the 2012 Interspeech likability classification task [23] and found that using RBM as the first network layer significantly improved the baseline result. More recently, Kim et al. [22] used DBNs to capture non-linear feature interactions in audiovisual data and found improvement over baselines that did not use deep learning.

While most previous work focused on static modeling using DBN either directly or indirectly as a feature extraction tool, our work investigated dynamic frame-level modeling using DBN-based acoustic models in conjunction with HMMs.

3. DATA

Our experiments were performed on the FAU Aibo dataset and followed the 2009 emotion challenge guidelines [5]. We used utterances from one school (Ohm) for training and the other (Mont) for testing. We constructed a held-out validation set by randomly selecting 6 out of 26 speakers from the original training set. As a result, the validation set had 1,690 utterances and the training set had 8,269.

We used the Hidden Markov Toolkit (HTK) to extract Mel Frequency Cepstral Coefficients (MFCC) from each utterance using a 25-ms Hamming window and 10-ms frame rate. Each audio frame was represented by a 39-dimensional real vector consisting of 12 MFCCs and energy, along with their first and second temporal derivatives. We performed z-normalization over each speaker in the training data. Since speaker identity information for test utterances was not available in the challenge, we did not use per-speaker normalization for the test

set to keep our results comparable with previous work. We instead performed z-normalization over the entire test data.

Finally, we augmented each audio frame with a varying number of its nearest neighbors, resulting in context windows of lengths **7**, **11**, **17**, **27**, **37**, **47**, **57**, **67**, and **77**. The first five frame sizes were used in [2]; the last four are our extensions to explore the advantage of including additional context for emotion recognition.

4. PROPOSED METHOD

4.1. Hybrid DBN-HMMs

We model each emotion as a left-to-right HMM with **1**, **3**, or **5** states and 16 Gaussian mixture components. We use an auxiliary HMM to capture background noise at the beginning and end of each utterance. After training the HMMs with standard Baum-Welch re-estimation, we force-aligned the training set to produce a mapping between audio frames and HMM states, which are used to fine-tune the DBNs. After this stage, the HMM transition probabilities stay fixed and the GMM is replaced by DBN as the acoustic model.

We fix the DBN architecture for all experiments, using 5 hidden layers (first layer is a Gaussian RBM; all other layers are Bernoulli RBMs) and 1024 units per layer; the only variable is the input vector size, which depends on the context window length. We use the same hyperparameters for generative pre-training as in [2]. The Gaussian RBM ran for 225 epochs with 0.0001 learning rate. All other layers ran for 75 epochs with 0.001 learning rate. A 0.9 momentum, 0.0002 L2 weight cost, and a minibatch size of 128 were used.

After pre-training, we add a logistic regression layer and train the model discriminatively as a feed-forward ANN using stochastic gradient descent with the same minibatch size. We follow the setup in [2] with a few modifications, which will be explained in the next paragraph. The learning rate starts at 0.01. At the end of each epoch, we perform recognition on the validation set using the HMM architecture discussed earlier, replacing the GMM with the current DBN. If the UAR (unweighted average recall) goes down, the model parameters are returned to their values at the beginning of the epoch and the learning rate is halved. This continues until the learning rate falls below 0.0001. We continue to use a 0.9 momentum and 0.0002 L2 weight cost. In the end, we obtain a feed-forward ANN that outputs the emission probability of each HMM state given a context window of audio frames.

Compared to [2], our setup has two main differences. First, we used a smaller learning rate during fine-tuning, which we found was important to avoid getting stuck in bad local minima. Second, because FAU Aibo is very unbalanced, it was necessary to normalize the DBN output with priors over HMM states computed from the training set.

4.2. Combining Different Classifiers

In this work we assess three HMM models (1, 3, 5 states) and nine context window sizes, resulting in 27 different models. We hypothesize that there is no one best architecture for all emotions or utterances. This implies that we can benefit from combining results of different classifiers. We test this hypothesis and address two related questions. First, how can we devise a confidence measure for each utterance that correlates well with prediction accuracy? Second, when taking the decisions of multiple classifiers into consideration, how can we weigh them such that the better models are trusted more?

We first enable the classifier to output a probability distribution over emotion classes given a speech utterance. We fit the temporal behavior of an utterance to each emotion class to obtain an average per-frame log-likelihood, referred to as activation, and normalize these activations into a probability distribution. Let $\mathbf{a^C}(\mathbf{x}) \in \mathbb{R}^{|L|}$ be the activation vector of classifier C for utterance x, where L is the set of emotion labels and $a_i^C(x)$ denotes the activation of C for emotion i given utterance x. Because activation is $\log$ of a very small probability, its value is always negative and requires an unconventional normalization method to keep the ratios consistent. The probability that classifier C would assign class label i to utterance x is defined as:

$$P_C(l = i|x) = \left(\sum_{j \in L} \frac{a_i^C(x)}{a_j^C(x)} \right)^{-1} \tag{4}$$

The maximum probability in this distribution can be interpreted as how confident the classifier is for a given utterance: higher probability means higher confidence.

We then need a confidence measure for an entire classifier architecture. A reasonable approach would be to look at how well the classifier performs for each emotion. However, because the validation set contained very few instances of the minority classes, we found this approach to be unreliable. Instead, we define w_C, the weight of classifier architecture C, as its UAR on the validation set.

Given a set of classifiers $\mathbb{C}$ and an utterance x, we can compute the probability that the assigned emotion label is i:

$$P(l = i|x, \mathbb{C}) = \frac{\sum_{C \in \mathbb{C}} w_C P_C(l = i|x)}{Z(x, \mathbb{C})} \tag{5}$$

$$Z(x, \mathbb{C}) = \sum_{j \in L} \sum_{C \in \mathbb{C}} w_C P_C(l = j|x) \tag{6}$$

We have 27 models with 2^{27} possible combinations, making it impractical to perform an exhaustive test. We combined the classifiers according to two different schemes, under the hypothesis that models with different parameters are better at classifying different types of utterances, thus combining them

can be advantageous. In Combination Scheme A, we combined models with the same number of HMM states but differing context window sizes. In Combination Scheme B, we combined models with the same number of context window size but differing numbers of HMM states.

5. RESULTS AND DISCUSSION

5.1. Computation Time

Training DBNs of the sizes used in this work was computationally expensive. We achieved significant speed-up by enabling GPU acceleration with Theano [24]. For 37-frame input windows, pre-training the deepest layer took 11 mins / epoch and fine-tuning took 10 mins / epoch. In our system, a single GeForce GT 630 GPU learned at roughly 10 times faster than a 3.2 GHz Intel i7 core.

5.2. Single Classifier Performance

In this section we evaluate the performance of our 27 stand-alone classifiers on the FAU Aibo test set.

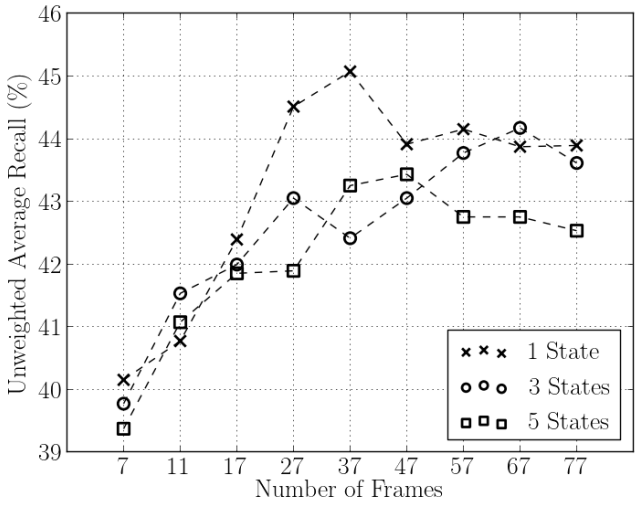

Fig. 1: Unweighted average recall on FAU Aibo test set of 27 stand-alone DBN-HMM classifiers.

Figure 1 summarizes the effect on UAR as we varied the window size and number of HMM states. The best result, **45.08%** UAR, was achieved with 1-state HMMs and 37-frame input windows. The best performance occurred with windows longer than or equal to 27 frames. In ASR, Mohamed et al. [2] reported that the best frame lengths for TIMIT phone recognition were 11, 17, and 27, which cover the range of 110-270ms, the average size of phones or syllables. In contrast, our results showed that emotion recognition works well with larger context windows covering up to 770ms, suggesting that emotion spans a longer time frame. This result echoes the findings in [25–27], which showed that window lengths of 1 second perform well on and are sufficient for emotion recognition.

It is interesting to note that the best results for each HMM architecture (37 frames for 1-state, 67 frames for 3-state, and 47 frames for 5-state) demonstrated that architectures with fewer HMM states performed better. For a long time, the 3-state left-to-right HMM has been used successfully in speech recognition to model basic acoustic units such as phonemes. This makes sense intuitively because speech is usually produced in a continuous manner, thus modeling speech production as a linear generative process should prove beneficial. However, the fact that the same technique did not work as well on emotion recognition suggests that emotion is produced with different temporal dynamics and might be better modeled with non-unidirectional HMMs. We will further investigate this hypothesis in future work.

Many previous works, especially those focusing on static modeling, successfully used SMOTE [28] to balance out the training set by simultaneously up-sampling minority classes and down-sampling majority classes. However, we found that SMOTE did not help our DBN fine-tuning process as it made the models more likely to overfit.

5.3. Classifier Ensemble Performance

In this section we investigate the effect of combining different models to do recognition on the FAU Aibo test set.

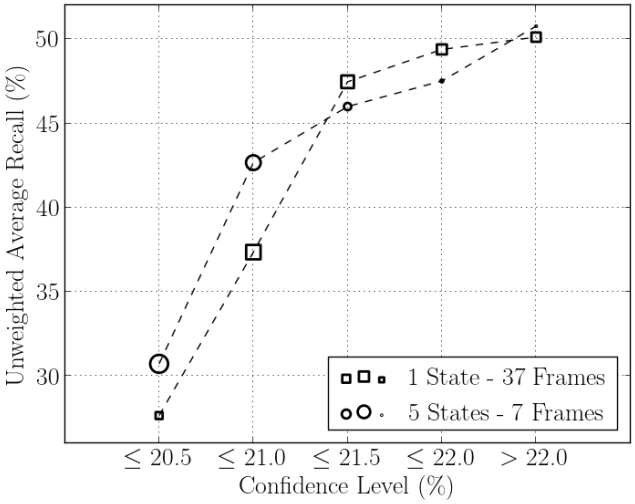

Fig. 2: Unweighted average recall on FAU Aibo test set as a function of confidence level. The best (1-state, 37-frame) and worst (5-state, 7-frame) models are being compared. Marker size denotes the number of utterances whose confidence levels fall within the specified range.

Our method of combining different classifiers relies on the confidence measure introduced in Section 4.2, hence it is worthwhile to analyze its behavior in some detail. Figure 2 plots the UAR of utterances falling in certain ranges

Table 1: Unweighted average recall on FAU Aibo test set of combined classifiers.

Combination Scheme A: Same state number, different window sizes

States	1	3	5
UAR (%)	45.28	44.90	43.52
Gain (%)	+0.20	+0.72	+0.08

Combination Scheme B: Same window size, different state numbers

Frames	7	11	17	27	37	47	**57**	67	77
UAR (%)	42.40	43.42	44.34	45.00	44.38	44.34	**45.60**	44.68	44.74
Gain (%)	+2.24	+1.88	+1.94	+0.48	−0.70	+0.42	+**1.44**	+0.50	+0.84

Numbers in the *Gain* row denote the absolute changes in UAR with respect to the best classifier in the combination.

of confidence levels for our best (1-state, 37-frame) and worst (5-state, 7-frame) models. As can be seen, our confidence measure exhibited two desirable characteristics. One, it correlated well with the UAR of both models, regardless of how good they were. Two, the worse model was less certain in general as most of its decisions had lower confidence values. It should be noted that although our confidence measure spans a narrow range, its absolute value is inconsequential; it is the relative difference between the classifiers' confidence scores that matters.

Table 1 shows the performance of our combined classifiers. The best result, **45.60%** UAR, was achieved by combining models using 57-frame input windows and different HMMs. Combining classifiers with differing numbers of HMM states (Scheme B) yielded significantly higher gain than combining those with different context window lengths (Scheme A). This observation is particularly interesting because it implies that models using different HMMs can better capture emotion variation than those using different context windows. We will explore this phenomenon in more detail in future work.

A weakness of our combination scheme lies in how we weigh a classifier as a whole. We currently assign weights to classifiers simply based on their UAR on the validation set. However, this approach is unreliable because the validation set contains very few instances of the minority classes, and good performance on the validation set does not guarantee good performance on the test set. In the future we will explore additional methods to combine different classifiers, such as training a secondary model that works directly with a utterance's confidence measures [29].

5.4. Effect of Speaker Normalization

To keep our results comparable with previous work on FAU Aibo, we did not make use of speaker identity information of the test utterances. However, it is reasonable to assume that a personalized emotion recognition system in the real world would have access to this information, rendering speaker normalization for test data possible. Because the first layer of our DBN assumes the input data to have zero mean and unit variance, speaker normalization should have positive impact on the models' performance.

We used our trained models to classify test utterances normalized at the speaker level. We found that the UAR of most classifiers went up by roughly 0.5 to 1%. The best UAR was **46.36%**, a result of combining models with 37-frame input windows and different HMM architectures. It should be noted that even if speaker information is not available, many techniques exist to assign approximate identities to speech utterances. See [30–32] for a survey on automatic speaker identification and recognition.

6. CONCLUSION AND FUTURE WORK

In this paper we investigated dynamic frame-level modeling with hybrid DBN-HMM classifiers on the FAU Aibo spontaneous emotion corpus and achieved state-of-the-art performance on the 5-class problem. We showed that although closely related, emotion and speech recognition have fundamental differences made evident by their different behaviors with respect to context window sizes and number of HMM states. These observations provide important insights to better understand emotion and improve recognition technologies.

For future work we plan to investigate additional features such as raw Mel Filter Bank coefficients [2,33], tune the DBN architecture, and try other approaches of combining different classifiers. We also plan to look more closely at the unique specializations of individual classifiers, which will provide valuable information about the inner workings of the systems and shed some light on the nature of emotion expression.

7. REFERENCES

[1] S. Steidl, *Automatic classification of emotion related user states in spontaneous children's speech*, Ph.D. thesis, 2009.

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

[3] Y. Attabi, J. Alam, P. Dumouchel, P. Kenny, and D. O'Shaughnessy, "Multiple windowed spectral features for emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013.

[4] H. Cao, R. Verma, and A. Nenkova, "Combining ranking and classification to improve emotion recognition in spontaneous speech," in *Proc. of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012.

[5] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brighton, United Kingdom, 2009, pp. 312–315.

[6] M. Kockmann, L. Burget, and J. Černocký, "Brno university of technology system for interspeech 2009 emotion challenge," in *Proc. of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brighton, United Kingdom, 2009, number 9, pp. 348–351.

[7] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 910, pp. 1062 – 1087, 2011, Sensing Emotion and Affect - Facing Realism in Speech Processing.

[8] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: Compensating for covariate shift," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 7, pp. 1458–1468, 2013.

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

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

[11] G. E. Hinton, "A practical guide to training restricted boltzmann machines," Tech. Rep., 2010.

[12] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area v2," in *Proc. of the 22nd Annual Conference on Neural Information Processing Systems (NIPS)*. 2008, MIT Press.

[13] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. of the 26th Annual International Conference on Machine Learning (ICML)*, Montreal, QC, Canada, 2009, pp. 609–616, ACM.

[14] Y. Tang and C. Eliasmith, "Deep networks for robust visual recognition," in *Proc. of the 27th Annual International Conference on Machine Learning (ICML)*, Haifa, Israel, 2010, pp. 1055–1062, Omnipress.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. of the 26th Annual Conference on Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, USA, 2012, pp. 1106–1114.

[16] G. E. Hinton, D. Li, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

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

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

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

[20] E. M. Schmidt and Y. E. Kim, "Learning emotion-based acoustic features with deep belief networks," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2011, pp. 65–68.

[21] R. Brueckner and B. Schuller, "Likability classification - a not so deep neural network approach," in *Proc. of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012.

[22] Y. Kim, H. Lee, and E. Mower Provost, "Deep learning for robust feature generation in audio-visual emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada.

[23] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The interspeech 2012 speaker trait challenge," in *13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012.

[24] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proc. of the Python for Scientific Computing Conference (SciPy)*, Austin, TX, USA, June 2010.

[25] B. Schuller and L. Devillers, "Incremental acoustic valence recognition: an inter-corpus perspective on features, matching, and performance in a gating paradigm," in *Proc. of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Makuhari, Japan, 2010, pp. 801–804.

[26] J. H. Jeon, R. Xia, and Y. Liu, "Sentence level emotion recognition based on decisions from subsentence segments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 4940–4943.

[27] E. Mower and S. Narayanan, "A hierarchical static-dynamic framework for emotion classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 2372–2375.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research (JAIR)*, vol. 16, pp. 321–357, 2002.

[29] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *Journal of Artificial Intelligence Research (JAIR)*, vol. 11, pp. 169–198, 1999.

[30] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, FL, USA, 2002, vol. 4, pp. IV–4072–IV–4075.

[31] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, Jan. 2010.

[32] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *Circuits and Systems Magazine*, vol. 11, no. 2, pp. 23–61, 2011.

[33] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Proc. of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Antwerp, Belgium, 2007, p. 22252228.