ETS System for AV+EC 2015 Challenge

Patrick Cardinal
École de Technologie Supérieure (ÉTS)
1100 rue Notre-Dame Ouest
Montréal, Québec, Canada
Patrick.Cardinal@etsmtl.ca

Najim Dehak
Massachusetts Institute of Technology (MIT)
77 Massachusetts avenue
Cambridge, Massachusetts, US
najim@csail.mit.edu

Alessandro Lameiras Koerich
École de Technologie Supérieure (ÉTS)
1100 rue Notre-Dame Ouest
Montréal, Québec, Canada
alessandro.koerich@etsmtl.ca

Jahangir Alam
Centre de recherche informatique de Montréal (CRIM)
405, avenue Ogilvy
Montréal, Québec, Canada
jahangir.alam@crim.ca

Patrice Boucher
École de Technologie Supérieure (ÉTS)
1100 rue Notre-Dame Ouest
Montréal, Québec, Canada
patrice.boucher.1@etsmtl.net

ABSTRACT

This paper presents the system that we have developed for the AV$^+$EC 2015 challenge which is mainly based on deep neural networks (DNNs). We have investigated different options using the audio feature set as a base system. The improvements that were achieved on this specific modality have been applied to other modalities. One of our main findings is that the frame stacking technique improves the quality of the predictions made by our model, and the improvements were also observed in all other modalities. Besides that, we also present a new feature set derived from the cardiac rhythm that were extracted from electrocardiogram readings. Such a new feature set helped us to improve the concordance correlation coefficient from 0.088 to 0.124 (on the development set) for the valence, an improvement of 25%. Finally, the fusion of all modalities has been studied using fusion at feature level using a DNN and at prediction level by training linear and random forest regressors. Both fusion schemes provided promising results.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Theory

AVEC'15 October 26, 2015, Brisbane, Australia

DOI: http://dx.doi.org/10.1145/2808196.2811639.

Keywords

Affective computing; Emotion Recognition; Speech, Facial Expressions and Physiological Signals

1. INTRODUCTION

Emotion detection is viewed as a problem which consists in recognizing the emotional state of a person through corresponding emitted signals. People tend to interpret the emotional state of other persons through the use of non-verbal cues, which are multi-modal. Indeed, emotion can be perceived from the voice of a speaker via prosodic cues or visually by interpreting facial expressions or gestures of people around us. All of these modalities contain affective information that can be used to automatically infer the emotional state of someone. Several applications could benefit from automated emotion detection software. In order to have in the future an emotion detection system deployed in real life, the reliability of the detection results is crucial. This work presents our contribution to the 2015 edition of the AV$^+$EC challenge. For this challenge, the focus is on multimodal emotion recognition in terms of continuous values in the two dimensional space of arousal and valence [9]. The database used for this challenge is RECOLA, a multimodal corpus of spontaneous collaborative and affective interactions [10].

In this work, we explore the use of Deep Neural Networks (DNNs) which have been shown to be very efficient in many speech processing tasks [7, 6] where a large amount of data is available but also in systems where the amount of training data is limited [11]. In the present work, we have mainly carried out experiments on the audio features of subjects and have extended our procedure to other modalities as well. An additional feature set, derived from electrocardiogram (ECG) readings, is also presented. The paper is organized as follows. Section 2 describes the technology used to create the DNN and the methodology for training them. Sections 3, 4 and 5 outline the experiments and presents results for each modality: audio, physiological and video features. Section 6 describes the approaches used for combining the results. Finally, the conclusion is presented in the last section.

2. METHODOLOGY

The learning architecture for all experiments is the neural network. The DNN software used in this work is based on the Theano library [2, 3]. As described in [3], Theano is a freely available compiler for mathematical symbolic expressions in Python. Symbolic expressions are compiled, taking advantage of the speed of optimized native languages, and then executed transparently on CPUs or GPUs.

The development set has been split into two subsets to ensure generalization capabilities of trained DNNs. The first subset contains three randomly selected sessions used during the training phase to determine when to stop the training process. The best set of parameters is determined from the predicted accuracies on the second subset of the development set (6 remaining files). However, the final DNN used the whole development set as stopping criteria. The use of this model on the test set afforded good results. Owing to time constraints, a full cross-validation scheme has not been performed. Note that the Concordance Correlation Coefficient (CCC) for the development set reported in this work is computed on the whole development set.

The data was normalized using Z-score. The mean $\bar{X}_{train}$ and the standard deviation $std(X_{train})$ are determined on the entire training set. Features of all sets are then normalized using the following formula:

$$x_i = \frac{x_i - \bar{X}_{train}}{std(X_{train})}$$

The data normalization scheme described in [9], that consists in applying normalization on a session basis, does not perform well in our case.

The evaluation measure used in the challenge is the CCC which is defined as:

$$\rho_{xy} = \frac{2 * s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2}$$

where x and y are the sets for which the correlation is calculated, s_x^2 and s_y^2 are the variances calculated on sets x and y respectively and $\bar{x}$ and $\bar{y}$ are the means of sets x and y, respectively.

3. AUDIO

In this section, experiments using audio features are described. The first task was to determine whether it is best to train a single neural network for both dimensions (arousal and valence) compared to one for each dimension. The results of Table 1 show that training a neural network for each dimension leads to better predictions.

Table 1 shows the results of experiments conducted on the audio feature set. Results are given using output of the DNN (RAW) and outputs filtered with median filter. The window size of the median filter has been determined using the development set in the same way described in [9]. For convenience, the results obtained on the neural network of [9] is reported in the table.

The base system is a neural network trained as a baseline for each dimension. The neural network for arousal has three hidden layers of 64 neurons each while the DNN for valence has 2 hidden layers of 128 and 64 neurons. For arousal, the base system improved the CCC by 16% on the development set and by 11% on the test set. In the case of valence, the improvement is 62% on the development set and 58% on the test set with respect to the baseline. Note that the test data was not median-filtered. The results on the test set was convincing enough to pursue experiments with the same procedure.

Table 1: Results on the development and test set (when available) for audio feature set using a DNN.

System	Dev		Test
	RAW	Filtered	
Arousal			
Baseline (CCC_{NN})	–	*0.214*	*0.139*
No Context	0.255	–	0.156
3 frame context	0.280	0.354	–
7 frames context	0.223	0.288	–
No sil., No context	0.123	0.151	–
No sil., 3 frame context	0.121	0.148	–
Valence			
Baseline (CCC_{NN})	–	*0.058*	*0.035*
No Context	0.153	–	0.084
3 frame context	0.170	0.192	–
7 frame context	0.145	0.170	–
No silence	0.205	0.246	–
No silence, 3 frame context	0.129	0.156	–

3.1 Frame Stacking

Frame stacking consists of concatenating frames in a given window as shown in Figure 1. For example, a context of 3 frames means that frames at times $t-1$, t and $t+1$ are concatenated to create one feature vector at time t. Frame stacking allows the DNN to use contextual information when learning the prediction function. In the case of supplied audio features, using a context of 3 frames means that information over a span of 3080 ms are given to the DNN.

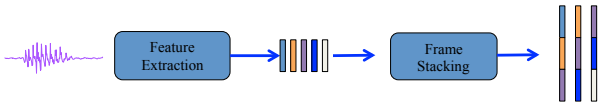

Figure 1: *Frame stacking process.*

The third and fourth lines of Table 1 show the results when using a context window of 3 and 7 frames for the arousal. Results for the valence are given on lines 9 and 10. Using a context window of 3 frames leads to an improvement of the CCC over the baseline for both dimensions. The improvement is 9% for the arousal and 10% for the valence. However, using a context window of 7 frames impairs the results. A possible explanation could be that the amount of training data is insufficient to handle such high dimensional feature vectors. However, experiments conducted on other modalities seem to show that the dimension vector might not be the main reason for this behaviour. Indeed, another possible explanation is that in this situation, a very limited amount (in this case 3 frames) is required to yield satisfactory results.

3.2 Silence Removing

Audio recordings contain silence and the voice of an "external" person. That means that for these frames, the learning system tries to match arousal and valence values to silence or the voice of the wrong person. The following experiments explored how removing non-speaker frames affects the quality of the arousal and valence predictions. Fortunately, the voice of the other person is quite far from the microphone. This characteristic allows the use of a voice activity detector (VAD) for removing non-speech segments (voice of the other speaker is low enough to be considered as background noise). Here is a description of the VAD used for this work.

3.2.1 Voice Activity Detector

In order to remove non-speech frames we use a Gaussian mixture model (GMM)-based unsupervised VAD (voice activity detector), shown in Figure 2, described in [1]. This VAD is conceptually similar to the VQ-based self adaptive VAD proposed in [8]. In VQ-based VAD speech and non-speech models are estimated using k-means (with k = 16) clustering whereas in this case they are trained using 16-component GMMs with diagonal covariance matrices. In unsupervised GMM-based VAD k-means clustering is used just for initialization.

In this VAD, producing speech/non-speech VAD segmentations for an audio recording at hand involves the following steps [1]:

- compute the log energy E^{log} frame by frame, sort the energies and take the lowest and highest (e.g., 10% of all frames in each case) energy frame indices.
- determine the energy threshold from the sorted energies using following formula:

$$\Theta = \frac{\Theta_1 + \Theta_2}{2} \quad (1)$$

  where Θ_1 and Θ_2 represent those values of sorted energy that correspond to the 20% and 80% length (or indices) of the sorted energy vector.
- compute the MFCC (12-dimensional including the 0th cepstral coefficients, no feature normalization is applied) features from the observed signal.
- train a 16-components GMM for speech $\lambda^s = (w_c^s, \mu_c^s, \Sigma_c^s)$ by taking the MFCCs corresponding to the highest energy frame indices. Similarly, by taking MFCCs that corresponds to the lowest energy frame indices train a 16-components GMM for non-speech $\lambda^{ns} = (w_c^{ns}, \mu_c^{ns}, \Sigma_c^{ns})$ where $c = 1, 2, ..., C$ is mixture component index and represents number of mixture components.
- compute speech log likelihood LL_s of each feature with respect to the trained speech model λ^s . Similarly, given trained non-speech model λ^{ns} compute non-speech log likelihood LL_{ns}.
- Compute the log likelihood ratio LLR by simply subtracting non-speech log likelihood from the speech log likelihood. Smooth the LLR using a moving averaging filter with a window of 25-frames. Determine a threshold Θ_{llr} from the sorted likelihood ratio using equation (1).
- Choose speech if $LLR > \theta_{llr}$ and $E^{log} \geq \theta$ otherwise non-speech.
- Then hangover scheme is used to prevent speech leakage. The hangover scheme does this by reducing the risk of a low-energy portion of speech being falsely classified as non-speech. The final VAD labels (contains only speech frames) are then obtained using an endpoint detection algorithm.

3.2.2 Results

The VAD considered that 60% of frames were either silence or the other person's voice. That means that the training set has been reduced to approximately 18 minutes, compared to 45 minutes of the original training set. The results show a substantial reduction of the CCC for the arousal but a large improvement for the valence. The use of a context window of 3 frames reduces the CCC. This could be caused by too small an amount of data.

These results are interesting and have been used as another prediction set when fusing modalities together; see Section 6.2.2. However, considering the small amount of data remaining after the silence removal process, more experiments need to be performed for investigating the real effect of noise frames on the predictions.

4. PHYSIOLOGICAL SIGNALS

In this section, we describe the experiments conducted on physiological signals.

Neural networks have been trained for both arousal and valence dimensions. The training process was the same as for audio experiments. Table 2 and 3 shows the results on electrocardiogram (ECG) and electrodermal activity (EDA) feature sets.

Table 2: Results on the development and test set (when available) for a DNN trained with the ECG feature set.

System	Dev		Test
	RAW	Median Filtered	
Arousal			
Baseline (CCC_{NN})	–	0.218	0.161
no context	0.194	0.203	–
3 frame context	0.236	0.246	–
7 frames context	0.252	0.262	–
Valence			
Baseline (CCC_{NN})	–	0.153	0.121
no context	0.085	0.088	–
3 frame context	0.092	0.099	–
7 frame context	0.074	0.077	–

As it was the case with the audio features, the use of a context window improved the CCC for both arousal and valence dimensions. For the arousal dimension, the best configuration corresponds to a context window of 7 frames. An improvement of 23% and 15% has been achieved on the ECG and EDA features respectively. In the case of the valence dimension, a context window of 3 frames yields the better results. An improvement of 8% and 9% has been obtained on both feature sets.

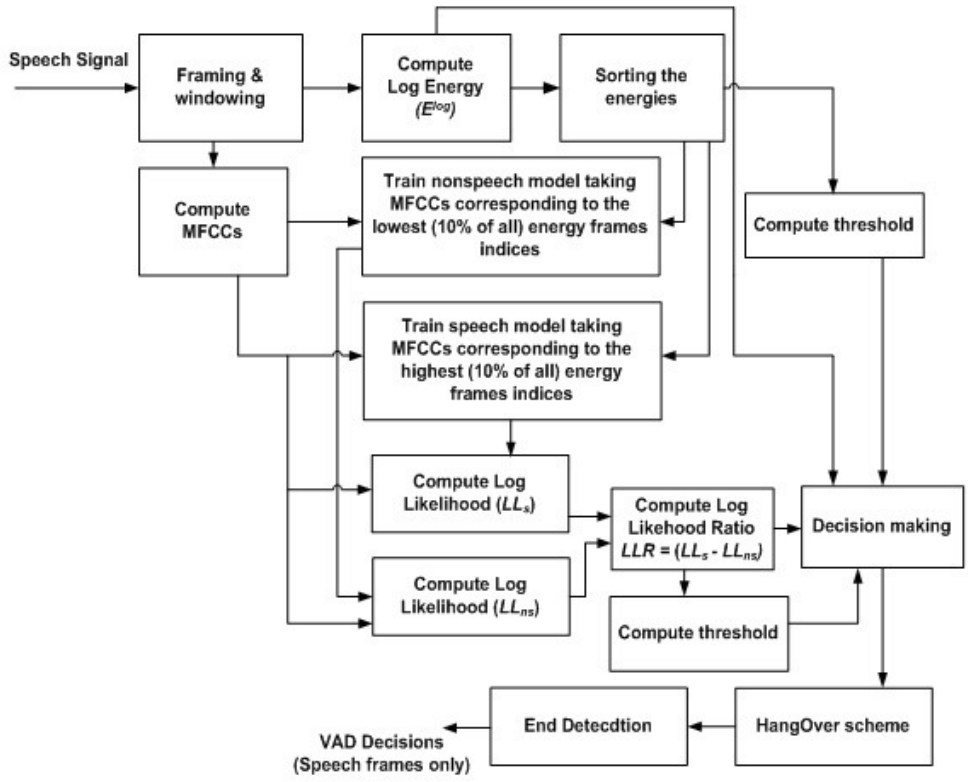

Figure 2: ***Block diagram of Gaussian mixture model (GMM)-based unsupervised voice activity detector.***

Table 3: Results on the development and test set (when available) for a DNN trained on the EDA feature set.

System	Dev		Test
	RAW	Median Filtered	
Arousal			
Baseline (CCC_{NN})	–	0.078	0.079
No frame context	0.099	0.101	–
context of 3 frames	0.099	0.104	–
Context of 7 frames	0.116	0.128	–
Valence			
Baseline (CCC_{NN})	–	0.166	0.156
No frame context	0.147	0.151	–
context of 3 frames	0.162	0.169	–
Context of 7 frames	0.144	0.152	–

4.1 Unstable Fraction of Cardiac Intervals

A preliminary experiment, using a new feature called Unstable Fraction of Cardiac Intervals (UFCI), has been conducted. This parameter has been studied in our research work on automatic stress detection.

In preliminary experiments, we observed that the mean squared successive difference (MSSD) of cardiac intervals is a good indicator of stress and is fundamentally sensitive to the heart rate value (the MSSD tends to increase as the heart rate decreases) notwithstanding the level of stress of the subject. In order to reduce this bias, we proposed to divide the MSSD by the mean length of cardiac intervals:

$$UFCI = \frac{rMSSD}{\sum_n^N I(n)/N} = \frac{N}{\sum_n^N I(n)} \sqrt{\frac{1}{N-1} \sum_{n=2}^{N} (I(n) - I(n-1))^2}$$

where I is the length of interval n among N intervals which are taken on a fixed time window frame (approximately 15 seconds in experiments for this work).

Table 4 shows that the UFCI feature benefits valence prediction with an improvement of 25% of the CCC. However, on the arousal, this feature reduces the CCC by 19%.

Table 4: Results on the development set and test set (when available) for a DNN trained with the UFCI feature set.

System	Dev		Test
	RAW	Median Filtered	
Arousal			
Baseline (CCC_{NN})	–	0.218	0.161
with UFCI feature	0.204	0.213	–
UFCI + 3 frame context	0.195	0.202	–
Valence			
Baseline (CCC_{NN})	–	0.153	0.121
with UFCI feature	0.099	0.101	–
UFCI + 3 frame context	0.118	0.124	–

5. FACIAL EXPRESSIONS

In this section, experiments conducted on visual information are described. Neural networks have been trained for both arousal and valence dimensions. The training process was the same as for audio experiments. Tables 5 and 6 show the CCC of predictions produced by DNNs trained on geometric and appearance features.

Table 5: Results on the development and test set (when available) for a DNN trained with geometric feature set.

System	Dev		Test
	RAW	Median Filtered	
Arousal			
Baseline (CCC_{NN})	–	0.178	0.149
No Context	0.085	0.090	–
Context of 3 frames	0.094	0.103	–
Valence			
Baseline (CCC_{NN})	–	0.325	0.292
No Context	0.356	0.389	–
Context of 3 frames	0.348	0.376	–

Table 6: Results on the development and test set (when available) for a DNN trained with appearance feature set.

System	Dev		Test
	RAW	Median Filtered	
Arousal			
Baseline (CCC_{NN})	–	0.079	0.017
No frame context	0.131	0.144	–
context of 3 frames	0.148	0.173	–
Valence			
Baseline (CCC_{NN})	–	0.273	0.234
No frame context	0.226	0.243	–
context of 3 frames	0.240	0.263	–

In light of the results obtained on other modalities, experiments conducted with a frame context have been limited to a context of 3 frames. For both sets of features, the use of a context window improved the CCC on the development set, excepting the valence dimension with geometric features for which a reduction of approximately 3.3% has been observed.

6. FUSION

In this section, experiments conducted on fusing modalities are described.

6.1 Features Fusion

A first experiment explores fusion of features and training a single model for all modalities. For this approach, all feature sets have been concatenated to create large feature vectors of 682 dimensions. For these experiments, no dimension reduction has been performed and all frames have been used. A DNN has been used for this experiment. Table 7 shows the result for the valence and arousal.

In the case of the arousal, an insignificant improvement over the audio modality (the best one) is obtained by using features of all modalities. The improvement obtained on the valence is slightly better than the one obtained using geometric features but it is not significantly better. We can conclude from these experiments that fusing features does not improve the prediction accuracy, at least when only a small amount of data is available.

Table 7: Results on the development set from feature set of all modalities.

System	CCC on Dev set	
	RAW	Median Filtered
Arousal		
No context	0.296	0.327
3 frame context	0.320	0.357
Valence		
No context	0.288	0.292
3 frame context	0.357	0.398

6.2 Score fusion

6.2.1 *Fusion using Random Forests*

Additional experiments have been conducted to explore fusion of system outputs using random forest. For these experiments, a random forest (Weka implementation has been used) model has been trained with predictions made on the predictions made from development sets of each modality, on which predictions made from the concatenation of all modality features has been added.

Random forests is an ensemble regression approach proposed by Leo Breiman [4] that employs tree-structured predictors constructed using randomness. The randomness is obtained by randomly selecting features at each tree node to grow each tree, which causes perturbations in the induced models. Final predictions are obtained by aggregating (voting) over the ensemble.

The model has been tested by dividing the development set into 2 sections. The first section, made up by the aggregation of subsets 1 to 4, has been used to generate a random forest of size 100 and the second section, made up by aggregating subsets 5 to 9, has been used to test the correctness of predictions made. A sanity check has been done by reversing the role of both sections. Table 8 shows the results achieved with Random Forests for fusing the outputs from the 6 models (one for each modality and another one for the fusion of all modalities at feature level).

Table 8: Results on the development and test sets for the fusion of all modalities using random forests.

Dimension	CCC (Dev)	CCC (Test)
	Median Filtered	
Arousal	0.967	0.246
Valence	0.979	0.303

Results show very good performance on the development set. Unfortunately, the results on the test set are lower than those of the baseline. Investigations will be made to determine why the model does not perform well on the test set.

6.2.2 Fusion Using Linear Regression

Considering results obtained with random forests, additional experiments have been conducted to explore fusion with a linear regression model. In a first instance, fusion using predictions made from each modality model has been used to train the fusion model. In the second experiment, the predictions obtained from the model built with features from each modality has been added. Finally, predictions made with audio features after removing the silence have also been added in the fusion model.

The model used for the fusion is *fitlm*, a linear model implemented in Matlab. The best model has been chosen using the one-leave-out cross-validation scheme. The results are shown in Table 9.

Table 9: Results on the development and test sets for the fusion of all modalities using linear regression.

System	CCC (Dev)		CCC (Test)
	RAW	Filtered	
Arousal			
Baseline	–	0.476	0.444
5 modalities	0.402	0.403	–
+ feature fusion	0.433	0.439	–
+ audio no silence	0.441	0.446	0.293
Valence			
Baseline	–	0.461	0.382
5 modalities	0.566	0.573	–
+ feature fusion	0.555	0.563	–
+ audio no silence	0.575	0.584	0.290

The best model for the arousal was a second degree polynomial for each predicator. The best one for the valence was a fourth degree polynomial for each predicator.

The system improves on the baseline for the valence but is quite low for the arousal. A possible explanation for the low CCC in the latter case could be the low CCC obtained using geometric features. Indeed, according to the results reported in [9], the geometric feature set is the most important as it accounts for approximately 35% of the total CCC. On the other hand, we obtained quite good results on important modalities for the valence. The results on ECG was low but the CCC on geometric features was higher in our system and both ECG and geometric features account for the same amount in the fusion.

The results also show that using predictions made by the fusion of features helps for the arousal but not for the valence. A possible explanation is that the DNNs trained on all features have been capable of extracting more information from the geometric features. In this case, the advantage of using these predictions would completely disappear with a better predictor trained on geometric features.

In the last experiment, predictions made with audio features from which silence frames have been removed have been used in addition to other modalities. Missing frames have been filled with zeros to ensure that all prediction sets are the same size. The results show that using these predictions improved the results for both the valence and the arousal. That could mean that although the amount of "real" audio data was small, information extracted from it is useful in a global system. This last system was the last one submitted for the challenge. As was the case for the fusion using random forest, the results on test set was not as good as expected. By analyzing the weights of a linear regression model trained on these data, it has been found that the weight of the ECG modality was very small but with a large variance while all others were stable. This could be an indication of ill behaviour of this particular model, a situation that requires further investigation.

7. CONCLUSION AND FUTURE WORK

Our results using DNNs to predict the arousal and the valence from different modalities have been presented. It has been found that using frame stacking improves the performance of the systems. In the case of ECG, we presented a new feature which improves the precision of the valence prediction by 25%.

Considerable work remains to be done to obtain a reliable system. Using the DNN concept, the use of bottleneck features will be explored. In speech recognition, this technique allowed to obtain good results on small data sets [11] and seems to show the ability to perform speaker adaptation [5], a feature that could be beneficial to the present problem. Other features and models will also be explored.

8. ACKNOWLEDGEMENTS

We would like to thank Leo Liu for his work on the development of the DNN library, especially for his contribution in adding regression capability specifically for this work.

9. REFERENCES

[1] J. Alam, P. Kenny, P. Ouellet, T. Stafylakis, and P. Dumouchel. Supervised/unsupervised voice activity detectors for text-dependent speaker recognition on the rsr2015 corpus. In *Proceedings of Odyssey Speaker and Langauge Recognition Workshop*, June 2014.

[2] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[3] 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 *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010. Oral Presentation.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] P. Cardinal, N. Dehak, Y. Zhang, and J. Glass. Speaker adaptation using the i-vector technique for bottleneck features. In *Proceedings of Interspeech 2015*, 2015.

[6] G. E.Dahl, Y. Dong, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30 –42, january 2012.

[7] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magagazine*, 29(6):82–97, 2012.

[8] T. Kinnunen and P. Rajan. A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data. In *Proceedings of Internation Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7229–7233, May 2013.

[9] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio,Video, and Physiological Data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC), ACM MM*, Brisbane, Australia, October 2015.

[10] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. In *Proceedings of Face & Gestures 2013, 2nd IEEE International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*, Shanghai, China, April 2013.

[11] Y. Zhang, E. Chuangsuwanich, and J. Glass. Extracting deep neural network bottleneck features using low-rank matrix factorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.