DEEP NEURAL NETWORKS FOR ACOUSTIC EMOTION RECOGNITION: RAISING THE BENCHMARKS

André Stuhlsatz[1], *Christine Meyer*[2], *Florian Eyben*[3], *Thomas Zielke*[1], *Günter Meier*[2], *Björn Schuller*[3]

[1]Dept. of Mechanical and Process Engineering, Düsseldorf University of Applied Sciences, Germany
[2]Dept. of Electrical Engineering, Düsseldorf University of Applied Sciences, Germany
[3]Institute for Human-Machine Communication, Technische Universität München, Germany
[1,2]{name.surname}@fh-duesseldorf.de, [3]{surname}@tum.de

ABSTRACT

Deep Neural Networks (DNNs) denote multilayer artificial neural networks with more than one hidden layer and millions of free parameters. We propose a Generalized Discriminant Analysis (GerDA) based on DNNs to learn discriminative features of low dimension optimized with respect to a fast classification from a large set of acoustic features for emotion recognition. On nine frequently used emotional speech corpora, we compare the performance of GerDA features and their subsequent linear classification with previously reported benchmarks obtained using the same set of acoustic features classified by Support Vector Machines (SVMs). Our results impressively show that low-dimensional GerDA features capture hidden information from the acoustic features leading to a significantly raised unweighted average recall and considerably raised weighted average recall.

Index Terms— Deep Neural Networks, Generalized Discriminant Analysis, Affective Computing, Emotion Recognition

1. INTRODUCTION

While broadly acknowledged as major contributor to future human-machine and -robot communication and multimedia retrieval systems, it is also well known that computational assessment of human emotion by acoustic properties is a demanding task. We quantified this fact by benchmarks reported on nine frequently used datasets in [7, 8]. To raise these and advance speech-based emotion recognition systems' performance, we introduce a Generalized Discriminant Analysis (GerDA) [13] that is a recently proposed machine learning tool based on Deep Neural Networks (DNNs) for discriminative feature extraction from arbitrary distributed raw data. Even if the dimensionality of the input data is extremely high, as in case of emotion-data considered here, GerDA is able to learn very compact discriminative features. Moreover, GerDA features are optimized for a fast and simple linear classification which is an important requirement for real-time applications [13]. For example, in our experiments, 2D features were extracted from 6 552-dimensional acoustic feature vectors and classified very fast using a simple minimum-distance classifier with a performance superior to the frequently used SVMs.

In the remainder of this paper we introduce GerDA (Sec. 2), the experimental setup (Sec. 3) including databases (Sec. 3.2) and the acoustic feature set (Sec. 3.1), present experimental results (Sec. 4), before concluding (Sec. 5).

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under under grant No. 211486 (SEMAINE). The responsibility lies with the authors.

2. GENERALIZED DISCRIMINANT ANALYSIS

Classical Linear Discriminant Analysis (LDA) seeks a linear transform of arbitrary distributed data $\boldsymbol{x} \in \mathbb{R}^d$ from C classes to Gaussian class-conditionally distributed features $\boldsymbol{h} \in \mathbb{R}^r$. Due to its linear nature, LDA often yields poor classifications on real world data. As generalization of LDA, GerDA maximizes a Fisher discriminant criterion $Q_h(f) := trace\{\mathbf{S}_T^{-1}\mathbf{S}_B\}$ ($\mathbf{S}_T$: total scatter matrix, $\mathbf{S}_B$: between-class scatter matrix) over a nonlinear function space $\mathcal{F}$ including linear transformations as well. While $\mathcal{F}$ is defined by a DNN, i. e., the chosen topology, connections and activation functions, the challenge is to find an optimal mapping $f^* \in \mathcal{F}$, $f^* : \mathbb{R}^d \to \mathbb{R}^r$, $r \leq d$, represented by trained network parameters $\boldsymbol{\Theta}^*$. As is well-known, training DNNs after random initialization almost always results in bad local solutions. We circumvent this drawback by using Restricted Boltzmann Machines (RBM) to pre-optimize the network parameters in a stochastic fashion [6]. After semi-supervised pre-optimization, GerDA DNNs are supervised fine-tuned using a back-propagation algorithm adapted to maximize the criterion Q_h in the feature space explicitly. In the following, for the sake of coherency, we restate the main ideas already proposed in [13] on semi-supervised training of GerDA DNNs:

The pre-optimization of a GerDA DNN, which consists of L layers l_j with N^j units ($1 \leq j \leq L$), is performed by first subdividing the full network into pairs of successive layers $\{l_1, l_2\}$, $\{l_2, l_3\}, \ldots, \{l_{L-1}, l_L\}$. Each pairing is then represented by a single RBM (Figure 1). The RBM's parameters $\boldsymbol{\Theta}^i$ ($1 \leq i \leq L-2$), i. e., the weights and biases, are trained unsupervised via stochastic gradient descent in the Kullback-Leibler divergence,

$$d(P^0||P^\infty; \boldsymbol{\Theta}^i) := \sum_{\boldsymbol{v}^i} P^0(\boldsymbol{v}^i) \log\left(\frac{P^0(\boldsymbol{v}^i)}{P^\infty(\boldsymbol{v}^i; \boldsymbol{\Theta}^i)}\right), \quad (1)$$

where $P^0(\boldsymbol{v}^i)$ denotes an empirical training data distribution and $P^\infty(\boldsymbol{v}^i; \boldsymbol{\Theta}^i)$ is defined as Boltzmann distribution. Binary data $\boldsymbol{v}^i(\boldsymbol{x}_n)$ ($1 \leq n \leq N$) for training each RBM i is generated by the hidden-layer units of the already trained predecessor RBM $i-1$, i. e., $\boldsymbol{v}^i(\boldsymbol{x}_n) = \boldsymbol{h}^{i-1}(\boldsymbol{x}_n)$. Except the first RBM's, inputs are the original training data, i. e., $\boldsymbol{v}^1(\boldsymbol{x}_n) = \boldsymbol{x}_n$. In this way, a stack of RBMs is trained layer-wisely until the top RBM is processed. Because training of RBMs is slow, a so-called Contrastive Divergence (CD) heuristics is used to speed up. Unlike [6], an extended semi-supervised architecture is implemented to effectively pre-optimize GerDA DNNs with respect to a discriminant criterion Q_h: The first RBM is adapted to facilitate real-valued inputs, which is an important requirement for many applications. The last RBM in the stack is extended by

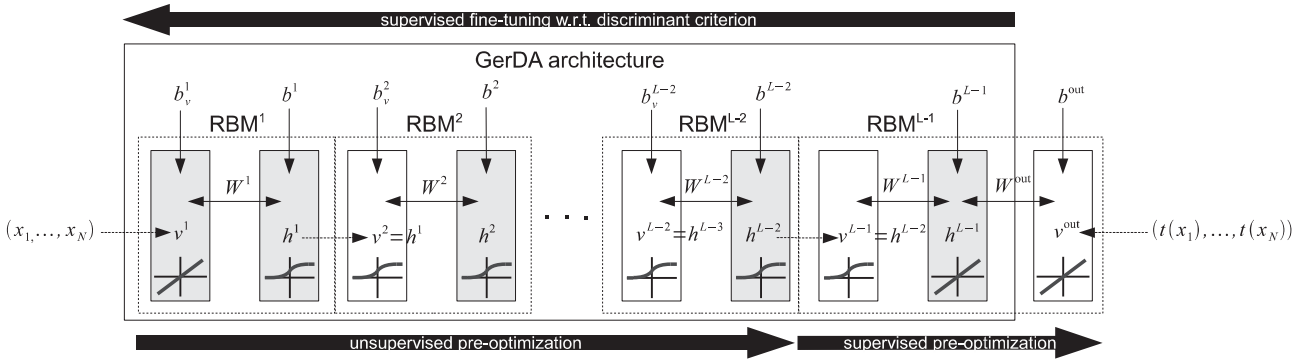

Fig. 1. A GerDA achitecture consisting of multiple layers (filled boxes) connected with weights and biases. Each pairing of layers is represented by a single RBM. The full RBM stack is pre-optimized partly unsupervised and supervised in a layer-wise manner. The resulting network parameters are then used to initialize the GerDA architecture followed by a supervised fine-tuning with respect to a discriminant criterion.

real-valued output units to supervised learn specific targets codes $\boldsymbol{t}(\boldsymbol{x}_n) \in \mathbb{R}^C$ that can be shown to be equivalent (asymptotically) to an implicit maximization of Q_h in the space spanned by $\boldsymbol{h}^L \in \mathbb{R}^r$. As the objective is to extract real-valued features, we modeled the hidden units $\boldsymbol{h}^L$ of the extraction layer real-valued, too. In the end, the topmost RBM's output units are discarded again, and the remaining parameters $(\boldsymbol{b}^1, \ldots, \boldsymbol{b}^{L-1}, \mathbf{W}^1, \ldots, \mathbf{W}^{L-1})$ serve as starting point for the fine-tuning via back-propagation adapted to maximize the discriminant criterion Q_h in the feature space directly.

3. EXPERIMENTAL SETUP

For the following investigations, the same setup as introduced in [8] is used facing GerDA with different emotion groups as well as two binary meta-groups. Static acoustic features from time-varying speech signals are obtained from a broad range of emotional speech databases by a specific pre-processing. The acoustic features are then used to compute GerDA features subsequently classified by a Mahalanobis minimum-distance classifier. The resulting performances are compared with a pair-wise multiclass SVM using a polynomial kernel on the acoustic features.

3.1. Acoustic Features

For each considered emotion recognition task, acoustic feature vectors of 6 552 dimensions were extracted using the openEAR toolkit [3] as 39 functionals of 56 acoustic Low-Level Descriptors (LLDs) including first and second order delta regression coefficients. Table 2 summarizes the statistical functionals which were applied to the LLDs shown in Table 1 to map a time series of variable length onto a static feature vector. Additionally, speaker (group) standardization was carried out.

3.2. Emotional Speech Databases

As benchmark databases, we chose nine among the most frequently used that range from acted over induced to spontaneous affect portrayal. For better comparability of obtained performances among corpora, we additionally map the diverse emotion groups onto the two most popular axes in the dimensional emotion model as in [7, 8]: arousal (i. e., passive ("-") vs. active ("+")) and valence (i. e., negative ("-") vs. positive ("+")). These mappings are not straight forward—we favor better balance among target classes. We further discretized into the four quadrants (q) 1–4 of the arousal-valence plane for continuous labeled corpora. In the following, each set is shortly introduced including the mapping to binary arousal/valence by "+" and "-" per emotion and its number of instances.

The *Danish Emotional Speech* (DES) database [2] contains pro-

Table 1. 33 Low-Level Descriptors (LLD) used.

Feature Group	Features in Group
Raw Signal	Zero-crossing-rate
Signal energy	Logarithmic
Pitch	Fundamental frequency F_0 in Hz via Cepstrum and Autocorrelation (ACF). Exponentially smoothed F_0 envelope.
Voice Quality	Probability of voicing ($\frac{ACF(T_0)}{ACF(0)}$)
Spectral	Energy in bands 0–250 Hz, 0–650 Hz, 250–650 Hz, 1–4 kHz 25 %, 50 %, 75 %, 90 % roll-off point, centroid, flux, and rel. pos. max. / min.
Mel-spectrum	Band 1–26
Cepstral	MFCC 0–12

Table 2. 39 functionals applied to LLD contours.

Functionals	#
Respective rel. position of max./min. value	2
Range (max.-min.)	1
Max. and min. value - arithmetic mean	2
Arithmetic mean, Quadratic mean, Centroid	3
Number of non-zero values	1
Geometric, and quadratic mean of non-zero values	2
Mean of absolute values, Mean of non-zero abs. values	2
Quartiles and inter-quartile ranges	6
95 % and 98 % percentile	2
Std. deviation, variance, kurtosis, skewness	4
Zero-crossing rate	1
# of peaks, mean dist. btwn. peaks, arth. mean of peaks, arth. mean of peaks - overall arth. mean	4
Linear regression coefficients and error	4
Quadratic regression coefficients and error	5

Table 3. Overview of the selected emotion corpora (Lab: labelers, Rec: recording environment, f/m: (fe-)male subjects).

Corpus	Language	Speech	Emotion	# Arousal		# Valence		# All	h:mm	# m	# f	# Lab	Rec	kHz
				-	+	-	+							
ABC	German	fixed	acted	104	326	213	217	430	1:15	4	4	3	studio	16
AVIC	English	free	natural	553	2449	553	2449	3002	1:47	11	10	4	studio	44
DES	Danish	fixed	acted	169	250	169	250	419	0:28	2	2	–	studio	20
EMOD	German	fixed	acted	248	246	352	142	494	0:22	5	5	–	studio	16
eNTER	English	fixed	induced	425	852	855	422	1277	1:00	34	8	2	studio	16
SAL	English	free	natural	884	808	917	779	1692	1:41	2	2	4	studio	16
Smart	German	free	natural	3088	735	381	3442	3823	7:08	32	47	3	noisy	16
SUSAS	English	fixed	natural	701	2892	1616	1977	3593	1:01	4	3	–	noisy	8
VAM	German	free	natural	501	445	875	71	946	0:47	15	32	6/17	noisy	16

fessionally acted nine Danish sentences, two words, and chunks that are located between two silent segments of two passages of fluent text. Emotions contain angry (+/-, 85), happy (+/+, 86), neutral (-/+, 85), sadness (-/-, 84), and surprise (+/+, 79). The *Berlin Emotional Speech Database* (EMOD) [1] features professional actors speaking ten emotionally undefined sentences. 494 phrases are commonly used: angry (+/-, 127), boredom (-/-, 79), disgust (-/-, 38), fear (+/-, 55), happy (+/+, 64), neutral (-/+, 78), and sadness (-/-, 53). The eNTERFACE (eNTER) [14] corpus consists of recordings of naive subjects from 14 nations speaking pre-defined spoken content in English. The subjects listened to six successive short stories eliciting a particular emotion out of angry (+/-, 215), disgust (-/-, 215), fear (+/-, 215), happy (+/+, 207), sadness (-/-, 210), and surprise (+/+, 215). The *Airplane Behaviour Corpus* (ABC) [10] is based on induced mood by pre-recorded announcements of a vacation (return) flight, consisting of 13 and 10 scenes. It contains aggressive (+/-, 95), cheerful (+/+, 105), intoxicated (+/-, 33), nervous (+/-, 93), neutral (-/+, 79), and tired (-/-, 25) speech. The *Speech Under Simulated and Actual Stress* (SUSAS) database [5] serves as a first reference for spontaneous recordings. Speech is additionally partly masked by field noise in the chosen actual stress speech samples recorded in subject motion fear and stress tasks. SUSAS content is restricted to 35 English air-commands in the speaker states high stress (+/-, 1 202), medium stress (+/-, 1 276), neutral (-/+, 701), and scream (+/-, 414). The *Audiovisual Interest Corpus* (AVIC) [9] consists of spontaneous speech and natural emotion. In its scenario setup, a product presenter leads subjects through a commercial presentation. AVIC is labelled in "level of interest" (loi) 1–3 having loi1 (-/-, 553), loi2 (+/+, 2279), and loi3 (+/+, 170). The *Belfast Sensitive Artificial Listener* (SAL) data is part of the final HUMAINE database. We consider the subset used, e. g., in [15] with an average length of 20 minutes per speaker of natural human-SAL conversations. The data has been labeled continuously in real time with respect to valence and activation using a system based on FEELtrace. The annotations were normalized to zero mean globally and scaled so that 98 % of all values are in the range from -1 to +1. The 25 recordings have been split into turns using an energy based Voice Activity Detection. Labels for each obtained turn are computed by averaging over the complete turn. Per quadrant the samples are: q1 (+/+, 459), q2 (-/+, 320), q3 (-/-, 564), and q4 (+/-, 349). The *SmartKom* (Smart) [12] corpus consists of Wizard-Of-Oz dialogs. For our evaluations we use dialogs recorded during a public environment technical scenario. It is structured into sessions which contain one recording of approximately 4.5 min length with one person and labelled as anger/irritation (+/-, 220), helplessness (+/-, 161), joy/gratification (+/+, 284), neutral (-/+, 2179), pondering/reflection (-/+, 643), surprise (+/+, 70), and unidentifiable episodes (-/+, 266). Finally, the *Vera-Am-Mittag* (VAM) corpus [4] consists of recordings taken from a German TV talk show. The audio recordings were manually segmented to the utterance level, whereas each utterance contained at least one phrase. The labeling bases on a discrete five point scale for valence, activation, and dominance. Samples among quadrants are q1 (+/+, 21), q2 (-/+, 50), q3 (-/-, 451), and q4 (+/-, 424). Further details on the corpora are summarized in Table 3 and found in [8]. Note that in the ongoing, balancing of the training partition is used.

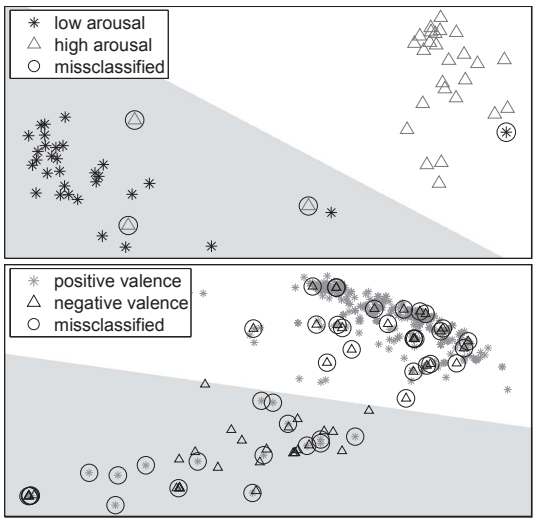

Fig. 2. Classification of 2D GerDA features of EMOD arousal, speaker 4 (top) and AVIC valence, speaker group 2 (bottom).

4. EXPERIMENTAL RESULTS

For all databases, test-runs are carried out in Leave-One-Speaker-Out (LOSO) or Leave-One-Speakers-Group-Out (LOSGO) manner to face speaker independence, as required by most applications. In the case of 10 or fewer speakers in one corpus we apply the LOSO strategy; otherwise, namely for AVIC, eNTERFACE, SmartKom, and VAM, we select 5 speaker groups with utmost equal number of male and female speakers and samples per group for LOSGO evaluation. As evaluation measures, we employ the Weighted (WA)

Table 4. Unweighted (UA) and Weighted (WA) Accuracy of the SVM (upper line, each) and GerDA (lower line, each) based acoustic emotion recognition. Raised benchmarks using GerDA with a simple minimum-distance classification are bold typed.

Corpus		All		Arousal		Valence	
	[%]	UA	WA	UA	WA	UA	WA
ABC	SVM	55.5	61.4	61.1	70.2	70.0	70.0
	GerDA	**56.1**	**61.5**	**69.3**	**80.6**	**79.6**	**79.0**
AVIC	SVM	56.5	68.6	66.4	76.2	66.4	76.2
	GerDA	**59.9**	**79.1**	**75.6**	**85.3**	**75.2**	**85.5**
DES	SVM	59.9	60.1	87.0	87.4	70.6	72.6
	GerDA	56.7	56.6	**90.0**	**90.3**	**71.7**	**73.7**
EMOD	SVM	84.6	85.6	96.8	96.8	87.0	88.1
	GerDA	79.1	81.9	**97.6**	**97.4**	82.2	87.5
eNTER	SVM	72.5	72.4	78.1	79.3	78.6	80.2
	GerDA	61.1	61.1	77.0	**80.8**	74.4	79.7
SAL	SVM	29.9	30.6	55.0	55.0	50.0	49.9
	GerDA	**35.9**	**34.3**	**65.1**	**66.4**	**57.7**	**53.0**
Smart	SVM	23.5	39.0	59.1	64.1	53.1	75.6
	GerDA	**25.0**	**59.5**	55.2	**79.2**	52.2	**89.4**
SUSAS	SVM	61.4	56.5	63.7	77.3	67.7	68.3
	GerDA	58.7	53.6	**68.2**	**83.3**	**74.4**	**75.0**
VAM	SVM	37.6	65.0	72.4	72.4	48.1	85.4
	GerDA	**39.3**	**68.0**	**78.4**	**77.1**	**52.4**	**92.3**
Mean	SVM	53.5	59.9	71.1	75.4	64.5	68.3
	GerDA	52.5	**61.7**	**75.2**	**82.3**	**68.9**	**79.5**

and Unweighted (UA) Accuracy as demanded in [11]. The latter measure better reflects unbalance among classes. The results using GerDA features in a Mahalanobis minimum-distance classifier and acoustic features in a polynomial SVM are given in Table 4 for all emotion classes contained per database and for the clustered two-class arousal/valence binary cover-classes tasks. As an example, in Figure 2, 2D GerDA features of speaker 4 of the EMOD binary arousal task (top) and of speaker group 2 of the binary valence AVIC task as well as the resulting classification boundaries are plotted. The increased difficulty in the spontaneous valence task is clearly visible in comparison to the commonly known to be easier acted and arousal task.

5. CONCLUSIONS

In this paper we introduced Generalized Discriminant Analysis (GerDA) based on Deep Neural Networks (DNNs) for the task of acoustic emotion recognition. Overall, the results averaged over nine databases and a total of 15 680 test instances show a highly significant improvement over the previously reported baselines by SVMs: In a one-tailed test considering weighted accuracy, GerDA outperforms the SVM for all classes at a level of 0.05, for the two-class arousal and valence tasks using 2D GerDA features, the level is $\ll 10^{-3}$. The breakdown in the All-tasks of EMOD and eNTERFACE may be due to the high number of classes and the relatively small number of available examples. Because GerDA is a data-driven tool, a sufficient amount of information must be provided to obtain highly compact and discriminative features.

In future work we aim at comparison with further neural network approaches, such as long short term memory architectures and hierarchical architectures to better cope with the decreased gain of GerDA in multi-class settings with little available training data.

6. REFERENCES

[1] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Proc. Interspeech*, Lisbon, 2005, pp. 1517–1520.

[2] I. S. Engbert and A. V. Hansen, "Documentation of the danish emotional speech database des," Tech. Rep., Center for PersonKommunikation, Aalborg University, Denmark, 2007.

[3] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," in *Proc. Affective Computing and Intelligent Interaction (ACII)*, Amsterdam, The Netherlands, 2009, IEEE.

[4] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German Audio-Visual Emotional Speech Database," in *Proc. IEEE ICME*, Hannover, Germany, 2008, pp. 865–868.

[5] J.H.L. Hansen and S. Bou-Ghazale, "Getting started with susas: A speech under simulated and actual stress database," in *Proc. EUROSPEECH-97*, Rhodes, Greece, 1997, vol. 4, pp. 1743–1746.

[6] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," in *SCIENCE*, vol. 313, pp. 504507, 2006.

[7] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies," *IEEE Transactions on Affective Computing*, vol. 1, 2010.

[8] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic Emotion Recognition: A Benchmark Comparison of Performances," in *Proc. IEEE ASRU*, Merano, Italy, pp. 552–557, 2009.

[9] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application," *Image and Vision Computing Journal*, vol. 27, pp. 1760–1774, 2009.

[10] B. Schuller, M. Wimmer, D. Arsic, G. Rigoll, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proc. ICASSP 2007*, Honolulu, Hawaii, USA, 2007, vol. II, pp. 733–736, IEEE.

[11] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. Interspeech*, Brighton, UK, 2009, ISCA.

[12] S. Steininger, F. Schiel, O. Dioubina, and S. Raubold, "Development of user-state conventions for the multimodal corpus in smartkom," in *Proc. Workshop on Multimodal Resources and Multimodal Systems Evaluation*, Las Palmas, 2002, pp. 33–37.

[13] A. Stuhlsatz, J. Lippel, and T. Zielke "Discriminative Feature Extraction with Deep Neural Networks," *In Proc. IJCNN 2010*, Barcelona, Spain, 2010.

[14] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *Proc. IEEE Workshop on Multimedia Database Management*, Atlanta, 2006.

[15] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. 9th Interspeech 2008*, Brisbane, Australia, 2008, pp. 597–600, ISCA.