MULTI-TASK DEEP NEURAL NETWORK WITH SHARED HIDDEN LAYERS: BREAKING DOWN THE WALL BETWEEN EMOTION REPRESENTATIONS

Yue Zhang[1]*, Yifan Liu*[1]*, Felix Weninger*[2]*, Björn Schuller*[1]

[1]Department of Computing, Imperial College London, London, United Kingdom
[2]Nuance Communications, Ulm, Germany
yue.zhang1@imperial.ac.uk

ABSTRACT

Emotion representations are psychological constructs for modelling, analysing, and recognising emotion, being one essential element of affect. Due to its complexity, the boundaries between different emotion concepts are often fuzzy, which is also reflected in the diversification of emotion databases, and their inconsistent target labels. When facing data scarcity as an ever present issue for acoustic emotion recognition, the straightforward method to jointly use the existing data resources is to map various emotion labels onto one common dimensional space; this, however, comes with considerable information loss. To solve the dilemma of data aggregation whilst efficiently exploiting the emotion labels in terms of their original meaning and interrelations, we advocate the usage of multi-task deep neural networks with shared hidden layers (MT-SHL-DNN), in which the feature transformations are shared across different emotion representations, while the output layers are separately associated with each emotion database. On nine frequently used emotional speech corpora and two different acoustic feature sets, we demonstrate that the MT-SHL-DNN method outperforms the single-task DNNs trained with only one emotion representation.

Index Terms— Deep Neural Networks, Multi-task Learning, Affective Computing, Emotion Recognition

1. INTRODUCTION

Emotion has been widely studied in various scientific disciplines like neuroscience, psychology, and cognitive science since long before the emergence of Affective Computing in the second quinquennium of the 1990s [1]. Within the realm of emotion research, the most prevailing emotion modelling concepts are derived from categorical, dimensional, and appraisal-based approaches [2]. The categorical emotion theory postulates that the affect system consists of a limited set of universal basic emotions (e. g., happiness, surprise, fear, sadness, anger, and disgust) [3]. However, many complex and subtle affective states such as sleepiness, stress, and depression cannot be adequately covered by or reduced to these basic emotion categories. Further, a number of studies posit that affective states are not isolated, discrete entities, but they are rather connected in a systematic manner [4, 5, 6]. Hence, dimensional models regard affective experience as a continuum of highly interrelated and often ambiguous states, similar to the spectrum of colour [7, 8]. To date, the most widely used dimensional concept is the circumplex model of affect, which relates all affective states to two independent neurophysiological dimensions: i. e., valence (pleasant–unpleasant) and arousal (activation–deactivation) [4]. In addition, a third dimension of dominance and submissiveness can be integrated into the arousal and valence model, commonly referred to as the PAD (Pleasure, Arousal, Dominance) emotion space [9, 10]. In contrast to the aforementioned emotion theories, the appraisal-driven componential models of emotion were proposed by Scherer et al. [11, 12]. From these different conceptualisations of emotion, heterogeneous emotion databases have been created, each of them having varying descriptive labels or annotation schemes [13, 14].

In Machine Learning, the rich sources of information as conveyed by such diversity of emotion databases are often sacrificed in favour of data aggregation by assigning higher-level emotion descriptors [15] to the dominant two-dimensional arousal and valence model. Thus, some underrepresented or ambiguous emotion classes (e. g., disgust, surprise) are subsumed under the categories 'other' or 'undefined', and excluded from the recognition task to bypass the crux in data aggregation [16]. The general scarcity of labelled data compounds the problem of lacking training data for emotion recognition. This point is particularly crucial when it comes to deep neural networks without special pre-training, which require significant amount of learning data [17]. To break down the wall between different emotion representations across databases, we propose the multi-task deep neural network with shared hidden layers (MT-SHL-DNN), in which the feature transformations computed in the hidden layers are made common for all emotion labelling schemes considered, while the softmax layers functioning as log-linear classifiers are separately assigned to each emotion recognition task. In this way, we achieve large-scale data aggregation without any information loss owing to label mapping or discretisation. On top of that, we show that the MT-SHL-DNN outperforms the single-task DNNs (ST-DNN) trained with single emotion databases, yielding efficient exploitation of label interdependencies.

In the remainder of this contribution, we review state-of-the-art methods in this field before introducing the MT-SHL-DNN approach. Empirical evaluation was carried out on nine frequently used emotion databases and two standard acoustic feature sets as described in the Sections 4 and 5. Finally, we discuss our experimental findings and provide further impulses for future research in the Sections 6 and 7.

2. RELATED WORK

The first benchmark for acoustic emotion recognition with deep neural networks was set by Stuhlsatz et al. [18], showing significant improvement over the performance of Support Vector Machines (SVMs). However, without the supporting function by data aggregation and

The research leading to these results has received funding from the European Unions Framework Programme HORIZON 2020 under the Grant No. 645378 (ARIA-VALUSPA).

multi-task learning, their proposed data-driven Generalised Discriminant Analysis (GerDA) based on DNNs is subject to serious limitation due to the high number of emotion classes and the relatively small number of available examples. Deng et al. [19] introduced for the first time the shared-hidden-layer autoencoders (SHLA) approach for learning common feature spaces shared across the training and test set to diminish their discrepancy resulting from different corpora. However, unlike our study, this work does not exploit the multi-task learning capabilities of DNNs. In the related field of Automatic Speech Recognition (ASR), Huang et al. [20] demonstrated that multilingual DNNs with shared hidden layers can significantly reduce word error rate over monolingual DNNs trained using only the language specific data. Further, it was shown that the shared hidden layers can be effectively transferred to improve recognition accuracy of new languages. Most recently, the shared-hidden-layer DNN approach has been effectively applied to language modelling [21]. Furthermore, recurrent DNN based transfer learning from dimensional to categorical emotion attributes of a single emotion database has been investigated in the work [22], with moderate success; yet this study does not exploit multi-task learning. To the authors' best knowledge, supervised multi-task learning with DNNshas never been applied to acoustic emotion recognition before.

3. MULTI-TASK SHARED-HIDDEN-LAYER DNN

Figure 1 depicts the structure of the MT-SHL-DNN, in which acoustic input features are transformed through the cross-task hidden layers. Unlike standard DNNs, multiple output layers are used for emotion classification according to various conceptualisations, such as the valence/arousal dimensions, or a set of affective states. These emotion target schemes are viewed as tasks in the multi-task learning framework. Each output layer is assigned to a specific task, with the number of nodes corresponding to the task's number of emotion classes. The input and hidden layers are shared across all the tasks.

The motivation for using this network structure is two-fold: First, multi-task learning acts as a regularisation for the network training, since the hidden layer representation is coerced to be predictive for a broad range of emotion representations. Second, it allows an utterance to be interpreted in manifold ways according to various emotion representations. This is unlike the application of shared-hidden-layer DNNs to multi-lingual automatic speech recognition [20] or language modeling [21], where usually utterances are transcribed only in a single language. In case that such multi-dimensional interpretation of an emotional speech utterance is desired, the proposed multi-task network structure is far more efficient than using a set of single-task networks, since the input-to-hidden and hidden-to-hidden connections have to be computed only once for each input vector, and the number of parameters in each output layer is small.

Mathematically, the output $\mathbf{y} = \mathcal{N}(\mathbf{x})$ of the MT-SHL-DNN $\mathcal{N}$ on an acoustic feature vector $\mathbf{x}$ is composed of sub-vectors for each task $1, \ldots, C$:

$$\mathbf{y} = \left[\mathbf{y}^{(1)}; \mathbf{y}^{(2)}; \ldots; \mathbf{y}^{(C)}\right], \quad (1)$$

where each $\mathbf{y}^{(c)}$, $c = 1, \ldots, C$ corresponds to a transformation of the last hidden layer activation with a task-specific weight matrix $\mathbf{W}_H^{(c)}$:

$$\mathbf{y}^{(c)} = \mathcal{H}(\mathbf{W}_H^{(c)}\mathbf{h}) = \mathcal{H}(\mathbf{W}_H^{(c)}\mathcal{G}(\mathbf{W}_{H-1}(\cdots\mathcal{G}(\mathbf{W}_1\mathbf{x})))), \quad (2)$$

with an output layer activation function $\mathcal{H}$ and a hidden layer activation function $\mathcal{G}$. Biases are omitted in the above equation for readability, but were used in our experiments. Optimisation of the parameters $\mathcal{W} = \{\mathbf{W}_1, \ldots, \mathbf{W}_{H-1}, \mathbf{W}_H^{(1)}, \ldots, \mathbf{W}_H^{(C)}\}$ of the MT-SHL-DNN is done via error back-propagation and stochastic gradient descent (SGD). Assuming that each acoustic feature vector $\mathbf{x}_i$ in the training set belongs to exactly one task c_i with label $\mathbf{z}_i$, the loss function to be minimised is

$$\mathcal{L}^{\text{MT}} = \sum_i \mathcal{L}(\mathbf{y}^{(c_i)}, \mathbf{z}_i), \quad (3)$$

i. e., the outputs for the tasks $c \neq c_i$ are irrelevant. Hence, the forward and backward propagation mechanism needs to evaluate only one output layer per input vector. In order to increase efficiency, we can limit each mini-batch to be used in SGD to input vectors belonging to a single task, which we will refer to as *homogeneous* mini-batches. Alternatively, we conjecture that we might be able to improve performance by using *stratified* mini-batches, where every mini-batch is representative of the full training set in that it contains input vectors from each task.

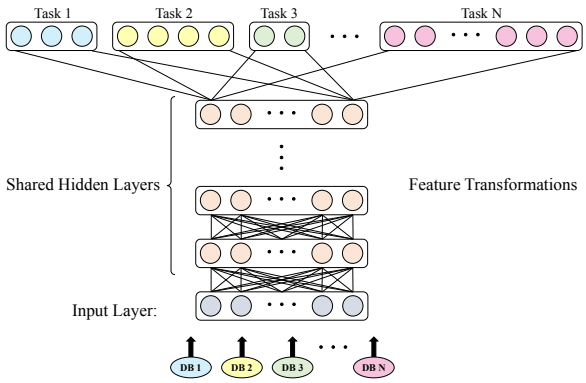

Fig. 1. Structure of the multi-task deep neural network with shared hidden layers (MT-SHL-DNN) applied on N emotion databases (DB), each with its own emotion representation (EMO REP).

4. EMOTIONAL SPEECH DATABASES

We retain the choice of the nine most frequently used emotional speech databases from the previous benchmark work [18], ranging from enacted over induced to spontaneous emotion portrayal. The *Airplane Behaviour Corpus (ABC)* [23] is an audio-visual emotion database created for the special target application of public transport surveillance. The recordings are based on mood induction by pre-recorded announcements during a simulated vacation flight. The *Audiovisual Interest Corpus (AVIC)* [24] features the level of interest of subjects in the scenario in which a product presenter led them through a commercial advertisement. The *Danish Emotional Speech (DES)* database [25] contains recordings of four professional actors, each of them articulated two isolated words, nine sentences, and two passages of fluent text in different emotions. The *Berlin Emotional Speech Database (EMOD)* [26] is targeted at a number of basic emotionsexpressed by German actors. The *eNTERFACE (eNTER)* database [27] involves subjects speaking pre-defined content in English, after having listened to six short stories, each of them eliciting a particular emotion. The *Belfast Sensitive Artificial Listener (SAL)* is a subset of the HUMAINE database [28, 29] containing audiovisual recordings from natural human-computer conversations. The *SmartKom (Smart)* corpus [30] features spontaneous speech produced

Corpus	Content	# / Emotion							# Inst	hh:mm	#Subj	Cond	kHz
ABC	German fixed	aggr 95	chee 105	into 33	nerv 93	neut 79	tire 25	-	430	1:15	8 4f	acted studio	16
AVIC	English free	loi1 553	loi2 2278	loi3 170	-	-	-	-	3002	1:47	21 10f	spont normal	44.1
DES	Danish fixed	angr 127	happ 86	neut 85	sad 84	surp 84	-	-	419	0:28	4 2f	acted normal	20
EMOD	German fixed	angr 127	bore 79	disg 38	fear 55	happ 64	neut 78	sadn 53	494	0:22	10 5f	acted studio	16
eNTER	English fixed	angr 215	disg 215	fear 215	happ 207	sadn 210	surp 215	-	1277	1:00	42 8f	acted normal	16
SAL	English free	q1 459	q2 320	q3 564	q4 349	-	-	-	1692	1:41	4 2f	spont normal	16
Smart	German free	angr 220	help 161	joy 284	neut 2179	pond 643	surp 70	unid 266	3823	7:08	79 47f	spont noisy	16
SUSAS	English fixed	hist 1202	meds 1276	neut 701	scre 414	-	-	-	3593	1:01	7 3f	mixed noisy	8
VAM	German free	q1 21	q2 50	q3 451	q4 424	-	-	-	946	0:47	47 32f	spont norm	16

Table 1. Overview of the selected emotion databases

Abbreviations: Inst: instances; Subj: subjects (m: male, f: female); Cond: condition; spont.: spontaneous speech; aggr: aggressive, angr: anger, bore: boredom, chee: cheerful, disg: disgust, happ: happy, help: helplessness, hist: high stress, into: intoxicated, loi13: level of interest 13, meds: medium stress, nerv: nervousness, neut: neutral, pond: pondering, q1 q4: quadrants in the arousal-valence plane, sadn: sadness, surp: surprise, tire: tired, unid: unidentifiable

through Wizard-of-Oz dialogues. The *Speech Under Simulated and Actual Stress (SUSAS)* database [31] was designed for stress-inducing scenarios from four domains: Simulated Stress, Calibrated Work Load Tracking Task, Acquisition and Compensatory Tracking Task, and Amusement Park Roller-Coaster. The *Vera-An-Mittag (VAM)* corpus [32] was extracted from spontaneous and emotionally coloured speech of guests during a German TV talk show. A detailed overview of the databases is given in Table 4. From this, we can make two observations: It can clearly be seen that the overlap in classes is low, and any attempt to map the representations to each other would inevitably result in a loss of information in the label space. By data aggregation, we can leverage 15 hours of speech data while individual databases would have around one hour of data with the exception of the larger SmartKom database. This clearly illustrates the potential of the proposed multi-task learning approach.

5. ACOUSTIC FEATURES

We evaluated the effectiveness of the MT-SHL-DNN method on two acoustic feature sets, which serve as standard references in speech emotion recognition. The ComParE set of supra-segmental (utterance-level) acoustic features contains 6 373 static features, which are obtained by computing statistical functionals over low-level descriptor (LLD) contours. For feature extraction, we used openSMILE in its 2.1 release [33]. To alleviate overfitting, we further applied the "extended Geneva Minimalistic Acoustic Parameter Set" (eGeMAPS) [34], which was selected for its potential to index affective physiological changes in voice production. In total, the eGeMAPS set contains 88 features, including spectral (MFCC 1–4, spectral flux) and frequency related parameters (formant 2–3 bandwidth).

6. EXPERIMENTS AND RESULTS

Our experiments were devoted to verifying the effectiveness of multi-task vs single-task training, and experimenting with different batching schemes as well as acoustic feature sets. As evaluation measure, we used Unweighted Average Recall (UAR) because it is also meaningful for highly unbalanced distributions of instances among classes, compared to weighted average recall ('conventional' accuracy). We used the Scikit-learn machine learning module [35] and the deep learning library Keras which is capable of running on top of either Theano or TensorFlow.

In the pre-processing stage, the data sets were normalised to zero mean and unit standard deviation. When using homogeneous mini-batches, each database is divided into 20 mini-batches. When using stratified mini-batches, all training instances are partitioned into stratified mini-batches of 100 samples each. The creation of mini-batches is done randomly at the start of each epoch.

For better comparison with the previous benchmark results [18], we performed Leave-one-Speaker-Out (LOSO) validation for databases with 10 or less subjects, otherwise Leave-One-Speakers-Group-Out (LOSGO) to ensure speaker independence. Here, it is noted that Stuhlsatz et al. [18] selected 5 speaker groups with utmost equal number of male and female speakers and samples per group for LOSGO evaluation, whereas we partitioned all training data into 5 gender-stratified speaker groups.

Moreover, we investigated the performance of the MT-SHL-DNN method on different feature sets (cf. Section 5). Based on previous experience, the network topology included five hidden layers ($H = 5$) with rectified linear activation function for the hidden layers and softmax activation function for the output layers. The loss function $\mathcal{L}$ in the multi-task objective (3) was the cross-entropy. With the large ComParE feature set, we used 2 048 neurons per hidden layer, and the training process was conducted for 80 epochs. The learning rate was set to 0.01 and Nesterov momentum with coefficient 0.9 was used. The DNNs for the eGeMAPS feature set had 256 neurons at each layer and were trained for 500 epochs. In a preliminary experiment, we had found that with this reduced feature set, larger networks did not perform better than this small topology. For both feature sets, dropout with probability of 0.5 was used on the input layer to improve generalisation.

UAR [%]	stratified		homogeneous	
Corpus	SHL	SHL+Retrain	SHL	SHL+Retrain
ABC	**56.11**	54.94	51.26	52.36
AVIC	56.62	**57.23**	56.28	56.08
DES	52.16	54.67	53.96	**55.39**
EMOD	**82.34**	78.63	78.47	81.09
eNTER	66.32	69.11	68.32	**69.19**
SAL	28.78	30.13	**31.05**	30.75
SMART	24.06	24.88	25.70	**25.93**
SUSAS	62.51	63.10	62.41	**64.31**
VAM	38.27	39.39	**40.61**	40.40
Mean	51.91	52.45	52.01	**52.84**

Table 2. Emotion recognition performance in terms of UAR on the ComParE feature set; Mini-batches either stratified including instances from all corpora or homogeneous containing samples from only one database; SHL: performance of the MT-SHL-DNN; SHL+Retrain: single-task DNNs trained from MT-SHL-DNN

UAR [%]	eGeMAPS		ComParE	
Corpus	ST	SHL+Retrain	ST	SHL+Retrain
ABC	46.51	**46.74**	50.80	**52.36**
AVIC	52.39	**54.42**	55.85	**56.08**
DES	53.23	**53.31**	**55.94**	55.39
EMOD	**73.21**	71.30	78.75	**81.09**
eNTER	61.05	**62.94**	68.95	**69.19**
SAL	**32.53**	31.05	28.27	**30.75**
SMART	20.64	**21.95**	24.54	**25.93**
SUSAS	**61.26**	60.58	62.94	**64.31**
VAM	37.78	**38.42**	38.10	**40.40**
Mean	48.73	**48.97**	51.57	**52.84**

Table 3. Single-task (ST) DNN performance in terms of UAR compared to retrained MT-SHL-DNN with homogeneous mini-batches, using the eGeMAPS or ComParE feature set

Finally, since the proposed multi-task learning approach allows us to leverage multiple training databases as well as regularise the network training, we hypothesise that the trained MT-SHL-DNN network can be used as a starting point to train improved single-task networks. To this end, we created single-task networks with parameters $\mathcal{W}^{(c)} = \{\mathbf{W}_1, \ldots, \mathbf{W}_{H-1}, \mathbf{W}_H^{(c)}\}$ from the trained shared hidden layers and a single trained output layer for task c. Then, we re-trained the entire parameter set $\mathcal{W}^{(c)}$ on a single emotion database.

In Table 2, two observations can be made. First, training with homogeneous mini-batches seems to be slightly better than stratified mini-batches. This is encouraging as homogeneous mini-batches can be trained more efficiently. Second, for both batching methods, training a task-specific network based on the MT-SHL-DNN improves the accuracy compared to the MT-SHL-DNN itself. This is not surprising, as fine-tuning the hidden layer weights for each database leads to a blow-up of the effective number of parameters that is proportional to the number of databases. Conversely, the comparison with the single-task DNNs in Table 3 shows that the results obtained by the MT-SHL-DNN are superior to those of the set of single-task DNNs, on average across the nine databases (52.01 vs 51.57 % UAR). Although the improvement is not statistically significant according to a one-sided Wilcoxon signed-rank test, it is remarkable given the compactness of the MT-SHL-DNN compared to the set of nine single-task DNNs. Moreover, Table 3 confirms the usefulness of the the MT-SHL-DNN as an initialisation for training the set of single-task DNNs, leading to improved accuracy for both evaluated feature sets. For the ComParE feature set, the improvement across databases is significant at the 0.01 level (from 51.57 to 52.84 % UAR).

7. CONCLUDING REMARKS

In this paper, we applied for the first time multi-task deep neural networks with shared hidden layers (MT-SHL-DNN) to acoustic emotion recognition. On nine frequently used emotional speech databases, we showed that the MT-SHL-DNN yields remarkable performance compared to single-task DNNs trained with only one emotion classification scheme, and that an effective data aggregation scheme is obtained by training DNNs staring from a MT-SHL-DNN. The robustness of the approach was validated on the ComParE and eGeMAPS feature set.

In future work, we aim to extend the MT-SHL-DNN approach for acoustic emotion recognition by using multiple softmax and linear layers for combined classification and regression tasks. For this purpose, the cross entropy loss and sum of squared errors need to be appropriately weighted. Moreover, we will investigate cross-emotion model transfer to improve recognition accuracy of new emotions. Finally, we will explore the MT-SHL-DNN approach for a broader range of speaker state and trait recognition tasks, such as the ones outlined in the work [36].

8. REFERENCES

[1] R.W. Picard and R. Picard, *Affective computing*, vol. 252, MIT press Cambridge, 1997.

[2] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. of EmoSPACE*, Santa Barbara, CA, 2011, pp. 827–834, IEEE.

[3] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[4] J.A. Russell, "A circumplex model of affect," *Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[5] R. Plutchik, *Emotion: A psychoevolutionary synthesis*, Harpercollins College Division, 1980.

[6] R.J. Larsen and E. Diener, "Promises and problems with the circumplex model of emotion.," *Review of Personality and Social Psychology*, vol. 13, pp. 25–59, 1992.

[7] J. Posner, J.A. Russell, and B.S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 03, pp. 715–734, 2005.

[8] J.A. Russell and B. Fehr, "Fuzzy concepts in a fuzzy hierarchy: varieties of anger," *Personality and Social Psychology*, vol. 67, no. 2, pp. 186, 1994.

[9] J. Jia, S. Zhang, F. Meng, Y. Wang, and L. Cai, "Emotional audio-visual speech synthesis based on pad," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 3, pp. 570–582, 2011.

[10] J.R. Fontaine, K.R. Scherer, E.B. Roesch, and P.C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[11] K.R. Scherer, A. Schorr, and T. Johnstone, *Appraisal processes in emotion: Theory, methods, research*, Oxford University Press, Oxford/New York, 2001.

[12] K.R. Scherer, "What are emotions? and how can they be measured?," *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[13] D. Ververidis and C. Kotropoulos, "A state of the art review on emotional speech databases," in *Proc. of Richmedia Conference*. Citeseer, 2003, pp. 109–119.

[14] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, no. 1, pp. 33–60, 2003.

[15] R. Buck, "The biological affects: a typology," *Psychological Review*, vol. 106, no. 2, pp. 301, 1999.

[16] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. of INTERSPEECH*, Lyon, France, 2013, pp. 148–152, ISCA.

[17] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[18] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *Proc. of ICASSP*, Prague, Czech Republic, 2011, IEEE, pp. 5688–5691.

[19] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *Proc. of ICASSP*, Florence, Italy, 2014, pp. 4851–4855, IEEE.

[20] J.T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. of ICASSP*, Vancouver, Canada, 2013, IEEE, pp. 7304–7308.

[21] A. Ragni, E. Dakin, X. Chen, M.J.F. Gales, and K.M. Knill, "Multi-language neural network language models," in *Proc. of INTERSPEECH*, San Francsico, CA, 2016, pp. 3042–3046, ISCA.

[22] S. Ghosh, E. Laksana, L.P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," in *Proc. of INTERSPEECH*, San Francisco, CA, 2016, pp. 3603–3607, ISCA.

[23] B. Schuller, M. Wimmer, D. Arsić, G. Rigoll, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proc. of ICASSP*, Honolulu, HI, 2007, vol. 2, pp. 733–736, IEEE.

[24] 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, Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, vol. 27, no. 12, pp. 1760–1774, 2009.

[25] I.S. Engberg and A.V. Hansen, "Documentation of the danish emotional speech database DES," *Tech. report, Center for Person Kommunikation, Denmark*, p. 22, 1996.

[26] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. of INTERSPEECH*, Lisbon, Portugal, 2005, vol. 5, pp. 1517–1520, ISCA.

[27] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *Proc. of IEEE Workshop on Multimedia Database Management*, Atlante, 2006, IEEE, pp. 8–8.

[28] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.C. Martin, L. Devillers, S. Abrilian, A. Batliner, et al., "The humaine database: addressing the collection and annotation of naturalistic and induced emotional data," in *Proc. of International Conference on Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 488–500.

[29] 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. of INTERSPEECH*, Brisbane, Australia, 2008, pp. 597–600, ISCA.

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

[31] J.H. Hansen, S.E. Bou-Ghazale, R. Sarikaya, and B. Pellom, "Getting started with SUSAS: a speech under simulated and actual stress database.," in *Proc. of Eurospeech*, Rhodes, Greece, 1997, vol. 4, pp. 1743–1746.

[32] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *Proc. of IEEE International Conference on Multimedia and Expo*, Hannover, Germany, 2008, pp. 865–868, IEEE.

[33] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of ACM MM*, Barcelona, Spain, 2013, pp. 835–838.

[34] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, 2015, 14 pages.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[36] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, Wiley, 2013.