Towards Speech Emotion Recognition "in the wild" using Aggregated Corpora and Deep Multi-Task Learning

Jaebok Kim, Gwenn Englebienne, Khiet P. Truong, Vanessa Evers

Human Media Interaction, University of Twente, Enschede, The Netherlands

{j.kim, g.englebienne, k.p.truong, v.evers}@utwente.nl

Abstract

One of the challenges in Speech Emotion Recognition (SER) "in the wild" is the large mismatch between training and test data (e.g. speakers and tasks). In order to improve the generalisation capabilities of the emotion models, we propose to use Multi-Task Learning (MTL) and use gender and naturalness as auxiliary tasks in deep neural networks. This method was evaluated in within-corpus and various cross-corpus classification experiments that simulate conditions "in the wild". In comparison to Single-Task Learning (STL) based state of the art methods, we found that our MTL method proposed improved performance significantly. Particularly, models using both gender and naturalness achieved more gains than those using either gender or naturalness separately. This benefit was also found in the high-level representations of the feature space, obtained from our method proposed, where discriminative emotional clusters could be observed.

Index Terms: speech emotion recognition, computational paralinguistics, deep learning

1. Introduction

Due to the increasing availability of crowd platforms and smart devices, it becomes realistic to collect an enormous amount of speech data from large and diverse populations. With deep learning technology, it is feasible to build a reliable model that captures more abstract properties of large corpora for Speech Emotion Recognition (SER). However, the performance in the wild is still unreliable partly because of diversity and often unknown contextual factors such as tasks, speakers, and recording conditions that can be encountered in the wild. [1, 2]. The diversity is not always covered in the corpora available to the researcher. Aggregating the corpora in the hope that the model obtained is general and robust enough against these differences has yielded variable results. Currently, speech emotion models work best if they are applied under circumstances that are similar to the ones that the model was trained on [1]. Normalisation [1], selection of samples [3], and transfer learning or adaptation [4, 5, 6] have been studied as methods. It was found that cross-corpus training worked to a certain degree only if corpora have similar contexts. To improve the generalisability of speech emotion models, we propose to use multi-task learning (MTL) that takes the differences between corpora into account as subtasks (i.e. gender and naturalness) and learns from these to make better classifications "in the wild" conditions.

MTL finds a common and essential representation between different tasks and often improves generalisation of a main task [7]. However, the success of MTL heavily depends on the choice of subtasks that may (not) relate to the main task. MTL has been applied in computer vision tasks [8], but for SER tasks, the use of MTL is relatively new. In the field of SER, naturalness [9] and speaker characteristics [10] such as gender and age,

affect the way emotions are expressed, and they are commonly accessible from meta information of emotional speech corpora. However, it is unknown if they are helpful as subtasks in MTL.

In this paper, we investigated whether MTL using gender and naturalness as subtasks improves generalisability of the emotion models trained. Compared to previous studies using deep learning, we validated our method using not only within-corpus but also cross-corpus settings reflecting more realistic challenges. In contrast to other commonly used methods such as transfer learning that requires additional data from test corpora [4, 5, 6], our method aims to generalise the training model without having access to test data which resembles "the wild setting". To the best of our knowledge, this is the first work that investigates gender and naturalness as subtasks of deep MTL in order to model large and variable emotional speech samples.

This paper is structured as follows. We first introduce related previous work in Section 2. Especially, we give a short overview of MTL and high-level representation of emotional speech. Next, we present corpora in Section 3, and describe our proposed learning scheme in Section 4. The results will be reported in Section 5 and concluded in Section 6.

2. Related Work

MTL is a machine learning approach that learns a main task with other related subtasks at the same time by using a shared representation [11, 7, 12]. MTL in deep neural networks (DNN) is similar to single-task learning except for the topology. It allows the learner to use the commonalities among the tasks, which often leads to improved generalisation. Hence, MTL is often regarded as a sort of inductive transfer. It improves generalisation by using the domain information extracted from the training signals of related tasks as inductive biases. However, choosing subtasks is a critical decision for MTL and helpful tasks for essential learning do not necessarily have to be related to the main task [12]. Recently, MTL has been applied to various fields such as computer vision and speech recognition [13, 14, 15]. Particularly, MTL with SVM was applied to SER in a cross corpus setting [16]. Their method did not share hyperplanes but rather information to train separated hyperplanes simultaneously. In [17], Long Short Term Memory (LSTM) based MTL was explored but limited to main tasks such as regression of arousal and valence, and confidence of annotations. In order to see the benefits of MTL, i.e. ability of generalisation, cross-corpus classification experiments simulating "in the wild" conditions should be carried out which have not been done before with MTL.

As deep learning has gained a lot of interest and showed promising results in various fields of automatic speech analysis [18], it is being actively investigated in the field of SER too. Especially, by using representation learning [19], there has been effort to extract unsupervised features which generalise

Table 1: *Overview of the selected corpora (the number of utterances)*

Corpus (ID)	Speakers	Emotion				Gender		Naturalness		Languages
		neutral	happy	sad	angry	female	male	natural	acted	
AIBO (A)	51	10967	889	0	1492	7579	5769	13348	0	German
EMODB (E)	10	77	61	58	97	160	133	0	293	German
ENTERFACE (F)	43	0	208	422	211	200	641	0	841	English
LDC (L)	7	80	180	161	139	320	240	0	560	English
IEMOCAP (I)	10	1708	595	2168	2206	336	6341	3177	3500	English
total	121	12832	1933	2809	4145	8595	13124	16525	5194	

emotional speech rather than engineered features (e.g. pitch) [20, 21]. [22, 23] proposed an intrinsic way to build a high-level representation of emotion using the engineered features. They extracted segment-level engineered features (e.g. Mel-Frequency Cepstral Coefficients (MFCC) and pitch) and modelled probabilities of emotional categories using Deep Neural Network (DNN) [22] and Bi-directional LSTM (BLSTM) [23]. Then, functionals of the probabilities were used to extract utterance-level features, denoted as high-level feature representation. Extreme Learning Machine (ELM) using the utterance-level features outperformed conventional approaches such as HMM, SVM, and BLSTM. More recently, [20] showed that unsupervised representation learning has limitation in complex subsequent structures for affect compared to [23]'s approach. Therefore, we chose [22, 23]'s architectures as baselines (single task learning) and compared it to our proposal and will investigate the effectiveness of MTL in various settings.

3. Data

In order to test our method with varying contexts, we used multiple corpora. We decided to focus on four representative emotional categories: neutral, happy, sad, and angry. We also aimed for variation in gender (female or male) and naturalness (natural or acted). We selected six corpora meeting our requirements: LDC Emotional Prosody (L) [24], eNTER-FACE (F) [25], EMODB (E) [26] FAU-aibo emotion corpus (A) [27], and IEMOCAP (I) [28]. Table 1 summarizes the corpora selected. IEMOCAP has both acted and natural (improvised) emotional speech [28], while EMODB, ENTERFACE, and LDC only have acted speech. Since AIBO has only emotional speech of children, we further categorised the gender of FAU-aibo corpus to female-child and male-child. While languages pose cultural differences in emotional expressions [1], we discard languages as a subtask since the selected corpora have only Germanic languages.

4. Method

4.1. Features

Based on previous work [22, 23], we chose a feature set: F0, voice probability, zero-crossing-rate, 12-dimensional MFCC with energy and their first time derivatives, totalling 32 features. As shown [22, 23], lower level features (e.g. mel-spectrogram) did not give good performance. Hence, we excluded them in the subsequent experiments. First, we normalised the gain of utterances. Then, we extracted the features for every frame using a 25-ms window sliding at 10-ms.

4.2. Generalisation using deep multi-task learning

In this study, we built high-level representation or features of emotional states for each utterance [22, 23]. First the frame-level acoustic features were fed into a shared network for multiple tasks: emotion, gender, and naturalness. To examine the effect of multi-task learning in various architecture, we used DNN and LSTM. Let us denote them, DNN-MTL and LSTM-MTL, respectively. The shared network was composed of 2 hidden layers with 256 cells for LSTM-MTL and 3 hidden layers of 256 nodes for DNN-MTL. While DNN-MTL does not model temporal dynamics of emotional speech, it uses a context window with a size of 250ms [22]. We did not find statistical differences in the performances between LSTM and Bi-directional LSTM (BLSTM). Moreover, later experiments indicated no more gains with higher number of layers and nodes. Using these shared networks, we optimised the cost functions of the tasks, defined as:

$$\epsilon_{total} = \epsilon_{main} + \sum_{i=1}^{N} \lambda_i * \epsilon_{sub_i} \qquad (1)$$

where ϵ is a cost function, λ_i is a non-negative weight for subtask, and N is the total number of subtasks. Since we do not have pre-knowledge about which task contributes more to the performance of the main task, we empirically set the same weight (.1) for the cost of gender and naturalness.

Next, to build the high-level features, statistical functionals were applied to the sequential outputs of softmax layers. We used the same 4 functionals (e.g. min, max, mean, and etc.) proposed in [22]. We aim to utilise the diversity of the contexts for improved generalisation of the emotion models trained and mainly examine the effect of the generalisation, not an extended representation including gender and naturalness. Hence, we discarded output layers of the contexts after training the shared network. Finally, our representation included emotional categories (4 classes), totalling 16 (4 classes x 4 functionals) high-level
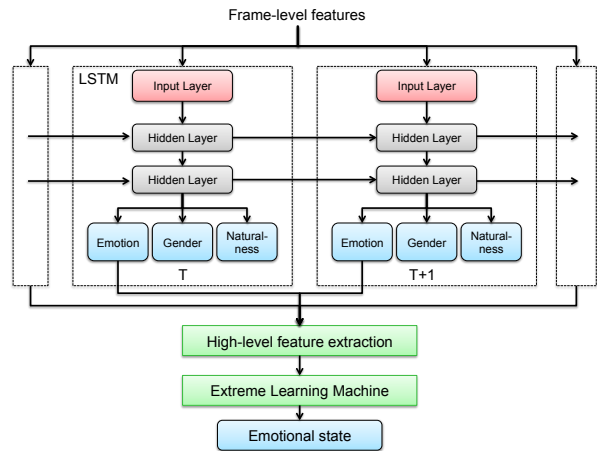

Figure 1: *Block diagram of LSTM-MTL, T = time*

Table 2: *Unweighted accuracy of within-corpus experiments; Mean value of UA over corpora (M) (**bold fonts for significant gains compared to baseline**)*

ID	Baseline		Proposed	
	DNN-STL	LSTM-STL	DNN-MTL	LSTM-MTL
A	$.405 \pm .03$	$.489 \pm .06$	$.409 \pm .11$	$\mathbf{.520 \pm .07}$
I	$.432 \pm .12$	$.529 \pm .14$	$\mathbf{.461 \pm .04}$	$\mathbf{.569 \pm .13}$
E	$.888 \pm .07$	$.921 \pm .10$	$.842 \pm .11$	$.925 \pm .08$
F	$.867 \pm .10$	$.953 \pm .15$	$.860 \pm .12$	$.953 \pm .15$
L	$.554 \pm .12$	$.552 \pm .15$	$.537 \pm .13$	$.564 \pm .15$
M	$.629$	$.689$	$.622$	$\mathbf{.706}$

features for the proceeding ELM (the same setting of ELM in [23]). Figure 1 describes the topology of LSTM-MTL.

5. Experiments and Result

5.1. Experiment setup

We compared DNN-MTL and LSTM-MTL to single task learning (STL) based methods: DNN-STL [22] and LSTM-STL [23]. Since we aim to examine the ability to generalise emotional speech in the diverse contexts, we composed various validation settings: WITHIN-CORPUS and CROSS-CORPUS. First, the WITHIN-CORPUS setting is leave-one-speaker-out cross-validation (LOSOCV). Although this might not be an optimal setting for MTL because of the smaller sizes of the corpora, we included the results as baselines and for completeness.

In the CROSS-CORPUS (leave-one-corpus-out-cross-validation) condition, we test on a corpus that was not included as training data. Optimising models without any access to a testing corpus is such a challenging task. Particularly, AIBO and ENTERFACE do not cover the complete set of 4 emotional categories and show a severe unbalance in number of samples as can be seen in Table 1. Moreover, we examine the contribution of each subtask to the emotion detection task. We compared the baseline (STL) and the proposed methods using either gender (GENDER-MTL) or naturalness (NATURALNESS-MTL), and both of them (ALL-MTL) as subtasks. In all conditions, we used 10% of training data for optimising parameters and excluded them from training data.

As a common setting, we utilised stochastic optimisation with a mini-batch of 128 samples, Adam method [29], and a fixed learning rate of $3 \cdot 10^{-3}$. We used categorical cross-entropy for the cost function. To prevent over-fitting, we used dropout [30] with $p = .5$ and early-stopping [31]. As an evaluation matrix, we used Unweighted Accuracy (UA) to consider the unbalanced number of samples between classes. Lastly, we used Wilcoxon signed-rank paired test (0.95 of confidence level) [32] to see statistical significance of gains of DNN-MTL and LSTM-MTL over baselines.

5.2. Result

Table 2 summarised the results of WITHIN-CORPUS experiments. Overall, LSTM-MTL (.706) outperformed DNN-STL (.629) and LSTM-STL (.689). There was no significant gains for EMODB, ENTERFACE, and LDC that have smaller number of samples compared to those of AIBO and IEMOCAP. For AIBO and IEMOCAP, LSTM-MTL showed significant gains ($p < .05$) over LSTM-STL. In short, MTL could be effective for training a single corpus if the size is sufficiently large.

Table 3 summarised the results of CROSS-CORPUS experi-

ments. Since IEMOCAP has both acted and natural emotional speech, we divided it into acted (IA) and natural (IN). First, we compared Baseline (STL) and ALL-MTL that uses both gender and naturalness as auxiliary tasks. Both DNN-MTL and LSTM-MTL showed gains for all corpora (except for AIBO by LSTM-MTL) while the gains varied on the corpora and architectures. DNN-MTL outperformed DNN-STL by **11%**, **12%**, **12%** for LDC (L), improvised (IN) and scripted (IA) IEMOCAP corpora, respectively. LSTM-MTL improved LSTM-STL by **13%**, **12%** for EMODB (E) and scripted (IA) IEMOCAP corpora. The overall mean of gains of DNN-MTL (LSTM-MTL) over DNN-STL (LSTM-STL) was **7**.4 (**5**.4)%, which is statistically significant and more superior to the gains reported in the within-corpus setting.

Next, we examined the performance of MTL using only either gender or naturalness as an auxiliary task in order to see the contribution of each task to the main task. As shown, Gender-MTL and Naturalness-MTL showed smaller overall gains compared to ALL-MTL. For some corpora (e.g. AIBO), Naturalness-MTL showed even better performance than ALL-MTL; however, hurt of generalisation was also reported depending on the corpora and architecture. Moreover, there was no significant difference between the performance of Gender-MTL and Naturalness-MTL.

Lastly, we investigated dependency of our methods on the type of testing corpora: acted and natural. DNN-MTL and LSTM-MTL obtained more overall gains from testing acted corpora. However, AIBO has a missing category and the severe unbalanced number of samples for classes (82% of the data was neutral in Table 1). In addition, acted corpora often do not have prototypical but diverse emotional expressions [3]. Hence, we do not conclude that effect of our methods is limited to only typical expressions.

5.3. Visualisation of representations

We visualised high-level representations of the selected corpora to see the generalisation ability of our method proposed in a feature space. To this end, we employed t-distributed stochastic neighbour embedding (T-SNE) [33] that is a non-linear dimensionality reduction technique embedding high-dimensional data into a space of two or three dimensions. First, we aggregated the selected corpora and shuffled the utterances in a random manner. Next, we split the shuffled data into training (80%), validation (10%), and testing (10% of the whole data) sets. We fed the training data into each model and optimised the parameters using the validation data. Then, the aggregated data was fed into the trained models to obtain the high-level representations as explained in Section 4. In Figure 2, we compared high level representations learnt from of DNN-STL (a), DNN-MTL (b), LSTM-STL (c), and LSTM-MTL (d). To look at benefits of these representations in a more quantitative way, we summarised their confusion matrix of the testing set in Table 4.

As shown in Figure 2, the proposed methods, (b) and (d), showed relatively more discriminative clusters compared to those of (a) and (c). While (a) showed mixed data points of neutral, sad, and angry, (b) showed more separated clusters of these categories. In Table 4 (b) showed large gains of sad and angry (**52%** and **21%**, respectively). When we compared (c) and (d), we could find a more separated cluster of sad that achieved a gain of **19%**. In overall, DNN(LSTM)-MTL outperformed DNN(LSTM)-STL by large gains **17**(**6**)% and showed more generalised representations of the large aggregated corpus compared to DNN(LSTM)-STL.

Corpus ID		Baseline (STL)		Proposed (MTL)					
				ALL-MTL		Gender-MTL		Naturalness-MTL	
Train	Test	DNN-STL	LSTM-STL	DNN-MTL	LSTM-MTL	DNN-MTL	LSTM-MTL	DNN-MTL	LSTM-MTL
{E,F,L,IN,IA}	A	.310	.355	**.351**	.352	.308	.338	**.401**	.326
{A,F,L,IN,IA}	E	.359	.288	**.388**	**.425**	**.392**	**.384**	.286	.267
{A,E,L,IN,IA}	F	.306	.332	**.327**	**.351**	.300	.318	**.337**	.326
{A,E,F,IN,IA}	L	.239	.251	**.347**	**.267**	**.251**	.251	**.249**	.244
{A,E,F,L,IA}	IN	.361	.455	**.481**	**.484**	.331	.397	**.488**	**.464**
{A,E,F,L,IN}	IA	.378	.341	**.498**	**.464**	**.405**	**.445**	.378	**.353**
Mean of natural: A,IN		.335	.405	**.416**	**.418**	.319	.367	**.445**	.395
Mean of acted: E,F,L,IA		.320	.303	**.390**	**.377**	**.337**	**.349**	.312	.298
Overall mean		.325	.337	**.399**	**.391**	**.331**	**.355**	**.356**	.330

(a) DNN-STL   (b) DNN-MTL
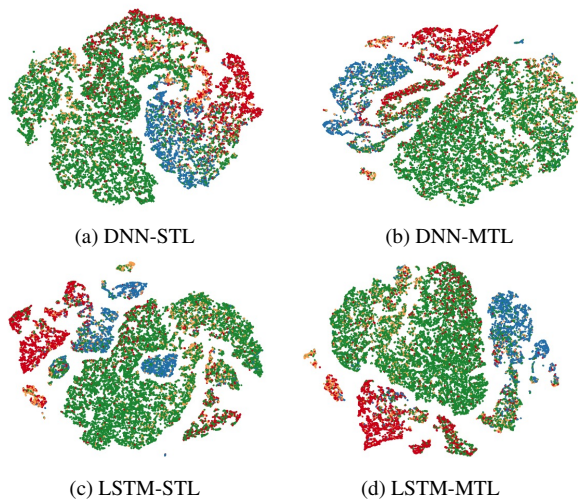
(c) LSTM-STL   (d) LSTM-MTL

Figure 2: *The result of T-SNE for high-level features of the aggregated corpora; coloured by emotional categories (green: neutral, orange: happy, blue: sad, and red: angry)*

Table 4: *Confusion matrix of the testing data and unweighted accuracy (UA)*

	(a) DNN-STL				(b) DNN-MTL			
	N	H	S	A	N	H	S	A
N	.99	.00	.01	.01	.94	.00	.04	.02
H	.94	.01	.01	.04	.80	.02	.06	.12
S	.78	.00	.18	.04	.16	.01	**.70**	.14
A	.73	.00	.00	.26	.44	.02	.06	**.47**
UA	.359				**.534**			
	(c) LSTM-STL				(d) LSTM-MTL			
	N	H	S	A	N	H	S	A
N	.90	.03	.03	.05	.91	.01	.04	.04
H	.64	.12	.08	.16	.64	**.15**	.10	.10
S	.26	.03	.65	.07	.09	.03	**.84**	.04
A	.35	.03	.03	.60	.29	.06	.05	.61
UA	.565				**.628**			

5.4. Summary and discussion

In summary, the overall gain in the within-corpus setting was not significant. For the DNN topology, the performance slightly dropped (Table 2). However, the gains for relatively larger corpora such as AIBO and IEMOCAP were still significant. Moreover, the overall gains were much larger for the cross-corpora setting (Table 3). While we could not find a significant difference between the performance of GENDER-MTL and NATURALNESS-MTL, the combination (ALL-MTL) outperformed single-task learning based methods regardless of topology. There was no hurt of generalisation by ALL-MTL. Hence, we concluded that MTL is more effective for larger corpora in a similar way of other applications [12]. Also, when all tasks reach best performance at approximately the same time of training, the performance of the main task could be maximised [12]. Hence, a control of separated learning rates should be addressed instead of the same fixed learning rate for all tasks. Moreover, complicated networks including private layers for subtasks potentially increase the performance [12] as shown in other applications [8]. Lastly, we should investigate how MTL affects discrimination of specific emotional categories as future work.

6. Conclusions

In this paper, we proposed generalisation of emotional models using large aggregated speech corpora and deep multi-task learning of commonly accessible contexts: gender and naturalness. We tackled a practical issue in the wild, that is, aggregating small but diverse corpora. To this end, we obtained high level representation of emotional speech using DNN and LSTM that utilise gender and naturalness as subtasks. We examined our method in various settings, within-corpus and cross-corpus. In the within corpus setting, the proposed method achieved significant gains for relatively larger corpora. However, in the cross-corpus setting, larger gains were reported in most of the corpora. Particularly, the combination of gender and naturalness as subtasks resulted the best gain and no hurt of generalisation regardless of its topology. Moreover, we visualised the high-level representation obtained from the proposed method in a feature space by using t-distributed stochastic neighbour embedding and found clear clusters of emotional utterances, resulting in significant gains. We concluded that our method is applicable to various topologies and corpora but potentially more effective for larger corpora. Potential improvement can be achieved using sophisticated architectures (e.g. a private network for each task) and independent learning rates for subtasks.

7. Acknowledgements

8. References

[1] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: variances and strategies," *Affective Computing, IEEE Transactions on*, vol. 1, no. 2, pp. 119–131, 2010.

[2] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[3] B. Schuller, Z. Zhang, F. Weninger, G. Rigoll *et al.*, "Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization," in *Proc. 2011 Afeka-AVIOS Speech Processing Conference, Tel Aviv, Israel*, 2011.

[4] Z. Zhang, F. Weninger, M. Wollmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 523–528.

[5] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *Signal Processing Letters, IEEE*, vol. 21, no. 9, pp. 1068–1072, 2014.

[6] J. Deng, Z. Zhang, and B. Schuller, "Linked source and target domain subspace feature transfer learning–exemplified by speech emotion recognition," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 761–766.

[7] R. Caruna, "Multitask learning: A knowledge-based source of inductive bias," in *Machine Learning: Proceedings of the Tenth International Conference*, 1993, pp. 41–48.

[8] B. Jou and S.-F. Chang, "Deep cross residual learning for multitask visual recognition," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 998–1007.

[9] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 474–477.

[10] L. R. Brody, "Gender differences in emotional development: A review of theories and research," *Journal of Personality*, vol. 53, no. 2, pp. 102–149, 1985.

[11] J. Baxter, "A model of inductive bias learning," *J. Artif. Intell. Res.(JAIR)*, vol. 12, no. 149-198, p. 3, 2000.

[12] R. Caruana, "Multitask learning," in *Learning to learn*. Springer, 1998, pp. 95–133.

[13] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 6965–6969.

[14] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *European Conference on Computer Vision*. Springer, 2014, pp. 94–108.

[15] A. Evgeniou and M. Pontil, "Multi-task feature learning," *Advances in neural information processing systems*, vol. 19, p. 41, 2007.

[16] B. Zhang, E. M. Provost, and G. Essi, "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5805–5809.

[17] F. Eyben, M. Wöllmer, and B. Schuller, "A multitask approach to continuous five-dimensional affect sensing in natural speech," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 1, p. 6, 2012.

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

[19] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1798–1828, 2013.

[20] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Learning representations of affect from speech," *arXiv preprint arXiv:1511.04747*, 2015.

[21] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, S. Zafeiriou *et al.*, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[22] I. T. Kun Han, Dong Yu, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proceedings of INTERSPEECH*, 2014.

[23] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proceedings of INTERSPEECH*, 2015.

[24] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell, "Emotional prosody speech and transcripts," *Linguistic Data Consortium, Philadelphia*, 2002.

[25] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.

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

[27] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. J. Russell, and M. Wong, "You stupid tin box-children interacting with the aibo robot: A cross-linguistic emotional speech corpus." in *Proceedings of LREC*, 2004.

[28] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[29] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[31] L. Prechelt, "Automatic early stopping using cross validation: quantifying the criteria," *Neural Networks*, vol. 11, no. 4, pp. 761–767, 1998.

[32] J. D. Gibbons and S. Chakraborti, *Nonparametric statistical inference*. Springer, 2011.

[33] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.