INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Ordinal Triplet Loss: Investigating Sleepiness Detection from Speech

Peter Wu, SaiKrishna Rallabandi, Alan W Black, Eric Nyberg

Language Technologies Institute, Carnegie Mellon University, PA, USA
peterw1@andrew.cmu.edu, {srallaba, awb, ehn } @cs.cmu.edu

Abstract

In this paper we present our submission to the INTERSPEECH 2019 ComParE Sleepiness challenge. By nature, the given speech dataset is an archetype of one with relatively limited samples, a complex underlying data distribution, and subjective ordinal labels. We propose a novel approach termed ordinal triplet loss (OTL) that can be readily added to any deep architecture in order to address the above data constraints. Ordinal triplet loss implicitly maps inputs into a space where similar samples are closer to each other than different ones. We demonstrate the efficacy of our approach on the aforementioned task.[1]

Index Terms: Ordinal Regression, Triplet Loss, Deep Metric Learning

1. Introduction

1.1. Paralinguistics

Paralinguistics refers to the aspects in a speech utterance beyond the linguistic content such as words. Paralinguistic cues such as accentuation are used to convey extra information such as emphasis, focus, expressiveness, and more. Applications of Computational Paralinguistics, the automatic analysis of such information, have grown rapidly over the last decade, spanning both human-human as well as human-machine interactions.

The ComPare Paralinguistics challenges have been playing a significant role in driving progress in the diverse use of paralinguistic information. Besides the traditional tasks such as emotion recognition using suprasegmental verbal and nonverbal aspects of speech, novel tasks such as the detection of speaker traits, deception, conflict, eating, and autism [1, 2, 3, 4] have been introduced. Detecting such information has the potential to not only play a role in assisting technologies with identifying affect but also play a role in detecting abnormalities indicating disorders. Paralinguistic information also has applications in other domains of speech processing such as dialog systems, speech synthesis, voice conversion, and more. In this paper, we present our approach towards one such paralinguistics task - the detection of sleepiness from speech.

The advent of deep learning has brought forth a surge in high-performing speech models [5, 6]. They have shown tremendous improvements in all the aspects of natural language processing (NLP), including speech recognition [5], visual question answering [7], speech synthesis [8], and more. The success of deep architectures in a variety of NLP tasks thus motivates their use in related areas including paralinguistics.

However, these models have been susceptible to learning just surface level associations and biases in the observed data, leading to overfitting and vulnerability to adversarial attacks [9, 10, 11, 12]. Therefore, there has been an interest towards learning algorithms that specifically consider intraclass relationships such as Siamese and triplet loss networks. Siamese networks have shown success in training on limited amounts of complex data [13]. Therefore, we combine a Siamese architecture with ordinal regression techniques in order to effectively train the model based on the given the data constraints.

1.2. Ordinal Data

A significant amount of data generated by our world, from natural forces to human behavior, is effectively continuous. As a result, humans' tendency to bin continuous data [14] has given rise to enormous amounts of ordinal data for applications ranging from healthcare to recommender systems [15, 16]. Thus, while humans tend to assign hard labels, the underlying data generally lies on a continuous spectrum. In order to perform effectively, statistical models must be able to capture the underlying data distribution rather than the humans' potentially subjective, and consequently noisy, discrete values. In a limited data setting where using sheer data size to generalize models is not an option, alternative techniques are required to make full use of the available data.

Leveraging the ordinal nature of a dataset as opposed to treating the classes as categorical is one effective approach for extracting more information from a limited set of samples. Many ordinal regression techniques have been proposed throughout the long-standing history of the field and have been traditionally applied to simpler tasks and non-deep models [17, 18]. For complex data that generally require deeper architectures, the large number of parameters in these ordinal techniques can tend to result in overfitting. Thus, simpler approaches are required in order to effectively integrate ordinal techniques into deep networks.

While treating continuous values as ordinals has good bearings intuitively, it is hard to train deep models that can effectively work with such data since standard classification techniques in deep learning are categorical. Hence, they cannot for example take into account the fact that class 2 is closer to class 3 as opposed to class 8. In order to effectively capture this information in a model, one approach is to construct an output distribution that reflects the relationship between classes. Soft labeling is one such technique that has been empirically shown to be effective with noisy ordinal data [19]. Our proposed approach builds on this idea of leveraging ordinal relations to generalize from limited noisy data, namely via learning the relative distances between the encoded representations of different data samples.

2. Related Work

2.1. Speech Techniques

Typical approaches for classification and prediction of paralinguistic features include extraction of low level descriptive features such as Mel-Frequency Cepstral Coefficients (MFCCs), log Mel-scale filter banks energies (FBANK) and several suprasegmental acoustic features that can be extracted using the

[1] Code is available at https://github.com/peter-yh-wu/ordinal

http://dx.doi.org/10.21437/Interspeech.2019-2278

openSMILE tool [20] followed by a classification model such as an SVM, decision tree, or neural network. While low level features act as general purpose feature sets, automatically derived neural representations using unsupervised learning [21] have the potential to further increase model performances. Recently there has been a surge in the use of pretrained generative models such as ELMo [22], BERT [23], and more. These features usually embed the task relevant information from the entire utterance in a compact form. In accordance with this trend, end-to-end learning models have been employed in paralinguistics tasks [24, 25].

2.2. Ordinal Regression

Each audio sample in the dataset is labeled with a number based on the KSS scale [26]. Since numbers on this scale follow a clear ranking, approaches in ordinal regression can be applied to this task. Namely, instead of penalizing all incorrect labels equally as in traditional multi-class classification, we can leverage the intuition that an incorrectly predicted class $\hat{y}$ that is numerically closer to the actual class y should be penalized less than a farther $\hat{y}$. Two primary ordinal regression techniques that have been applied to statistical models include ordistic loss, which represents the output distribution as a mixture of Gaussians, and a thresholding-based approach which learns the decision boundary between adjacent classes [18]. Since both approaches involve many parameters, utlizing them in a deep architecture can lead to overfitting.

Soft labels have been shown to not only work effectively with neural models, but also help with convergence and training on noisy data [27, 19]. While not originally created for ordinal tasks, empirical results suggest that soft labelling can be effectively applied to ordinal regression problems [19]. In this paper, we show why soft labelling is particularly effective for ordinal tasks and propose a general deep approach that learns ordinal relationships through soft labels and relative distance constraints.

2.3. Deep Metric Learning

Deep metric learning (DML) encompasses approaches that capture the similarity between datapoints via deep architectures. One such technique is the triplet loss function [28], which constrains models to map input data from the same class to similar locations in an embedding space and data from different classes to separate locations. Specifically, the loss function for a triple (x_a, x_p, x_n) with respective classes $y_a = y_p \neq y_n$ is given by

$$\|f(x_a) - f(x_p)\|_2^2 - \|f(x_a) - f(x_n)\|_2^2 + \alpha,$$

where $\|\cdot\|$ is the Euclidean norm, $f(x)$ is the encoded representation of x, and α is a hyperparameter representing the margin between same-class and different-class pairs.

Previous works have shown the effectiveness of triplet loss and Siamese architectures in limited data settings [13]. Siamese networks perform well in such cases since they keep the number of model parameters low through weight sharing and effectively increase the dataset size through accepting multiple inputs at a time. Additionally, by encouraging input representations to cluster spatially by their class labels, these approaches can implicitly accentuate features useful for downstream classification tasks.

Like many other DML techniques, triplet loss is designed for categorical data, and consequently does not leverage any properties of ordinal data. We propose an augmented loss function, which we refer to as ordinal triplet loss in later sections, that captures the ordered nature of a collection of data through accounting for the absolute difference between class labels in its relative distance constraints.

3. Proposed Approach

Our proposed OTL approach is mainly comprised of two parts: soft labelling and an ordinal triplet loss function. Previous works have demonstrated the superiority of soft labels over hard labels for tasks with noisy data [19]. In Section 3.1., we show that soft labels are especially suited for ordinal tasks via a statistical interpretation. The ordinal triplet loss function serves to encourage the model to learn representations specific to the ordinal task at hand by adding a loss constraint to a hidden layer. We discuss the formulation of the loss function in Section 3.2 and how to integrate it into a deep architecture in Section 3.3.

3.1. Soft Labels

Results from Zhang et al [19] suggest that soft labels are well-suited for tasks with noisy, complex data. We reformulate their approach through a statistical lens in order to evince its particular effectiveness for ordinal tasks.

In a K-class ordinal task, we can uniformly scale a class label $k \in \{0, 1, \ldots, K-1\}$ to the interval $[0, 1]$, i.e. mapping class k to $k/(K-1)$, without losing generality. Additionally, through associating a datapoint in original class k with a pair $(k/(K-1), 1 - k/(K-1))$ that sums to 1, we can reinterpret the class as a combination of binary labels. In other words, we can interpret the datapoint as being a combination of $k/(K-1)^{th}$ of a class-0 datapoint and $1 - k/(K-1)^{th}$ of a class-1 datapoint. Assuming that the binary classes are generated from a Bernoulli distribution, we can express the likelihood of a set of data $\{x_1, x_2, \ldots, x_B\}$ with respective classes $\{y_1, y_2, \ldots, y_B\}$ as

$$\prod_{i=1}^{B} f(x_i)^{\frac{y_i}{K-1}} (1 - f(x_i))^{1 - \frac{y_i}{K-1}},$$

where $f(x_i)$ is the model output for datapoint x_i. We can thus maximize this likelihood by training the model using the class pairs via cross-entropy loss. During test time, we invert the class-to-soft-label function to retrieve class predictions, namely mapping a pair $(\hat{p}, 1-\hat{p})$ to $\lceil \hat{p}(K-1) \rfloor$, where $\lceil \cdot \rfloor$ is the nearest integer function.

Training the model in this matter naturally penalizes class predictions more the farther they are from the true class, thus capturing the ordinal nature of the data. In fact, due to the curvature of the log likelihood function, loss penalties approximately increase exponentially with respect to distance to the middle class, capturing the central tendency bias inherent in datasets using the Likert scale. It is worth noting that this soft label formulation works with ordered data in general, including continuous data.

3.2. Ordinal Triplet Loss

Ordinal triplet loss augments the traditional triplet loss function [28] by capturing ordinal relations, thus further utilizing properties in a limited corpus. Namely, the function adds a constraint ensuring that datapoints with farther class labels have larger distances between them in their embedded space. Each input triplet is comprised of an anchor sample x_a, another sample x_s, and a sample x_d constrained to have a class farther from x_a than x_s. In other words, their respective class labels satisfy

$$|y_a - y_d| > |y_a - y_s| + \alpha,$$

where $\alpha \in \mathbb{N}$ is a hyperparameter. Since x_s does not need to have the same class as x_a, the resulting set of possible triplets is noticeably larger than that of the traditional triplet loss formulation. When appropriate techniques described in Section 3.4 are applied to select which triplets to train, this expanded set of triplets can help the model generalize better. The ordinal triplet loss for a triplet (x_a, x_s, x_d) is given by

$$\sigma(\|f(x_a) - f(x_d)\| - \|f(x_a) - f(x_s)\|),$$

where $f(x)$ is the encoded representation of x, $\|\cdot\|$ is the Euclidean norm, and σ is the logistic function, given by $\sigma(x) = \log(1 + e^{-x})$. Conceptually, the loss function penalizes cases where the model maps the x's to representations where x_a is closer to x_d than x_s. The logistic function serves to make the loss function differentiable. Like the soft label approach, ordinal triplet loss can be applied to continuous data as well.

3.3. Network Architecture

We use an architecture similar to that of Zhang et al [19] to train our model, replacing their loss functions with ordinal triplet loss. Namely, the model receives triplet inputs and jointly optimizes the ordinal triplet loss function, which uses all three inputs, and the soft label cross-entropy loss, which uses only the anchor samples. Each iteration, the model embeds all inputs using an encoder f before applying ordinal triplet loss, and passes the anchor sample embeddings through an MLP g before applying the soft label cross-entropy loss. We add a batch norm layer between f and g to help with convergence. The loss function for a batch $\{(x_1, y_1), (x_2, y_2), \ldots, (x_B, y_B)\}$ is given by

$$\frac{1}{B}\left(\sum_{i=1}^{B} l_s(x_a^{(i)}, y_a^{(i)}) + \beta \sum_{i=1}^{B} l_t(x_a^{(i)}, x_s^{(i)}, x_d^{(i)})\right),$$

where l_t is the ordinal triplet loss function, l_s is the soft label cross-entropy loss function, and β is a hyperparameter describing how much to weigh the ordinal triplet loss.

Conceptually, f serves to separate embeddings in a manner that captures ordinal relations in order to help g in the downstream classification task. As with other Siamese architectures [13], the weight sharing between elements in each triplet and the increased number of possible inputs via grouping samples into tuples aims to help with training effectively on limited amounts of complex data.

3.4. Implementation Details

Since the number of possible triplets is cubic with respect to the number of data samples, training using the traditional epoch formulation is impractical. Thus, we choose datapoints using an ordinal version of the triplet loss semi-hard sampling approach [28]. Namely, given an (x_a, x_s) pair, we select the x_d with the minimum $\|f(x_a) - f(x_d)\|$ that satisfies

$$\|f(x_a) - f(x_d)\| > \|f(x_a) - f(x_s)\|,$$

as well as the class label constraint $|y_a - y_d| > |y_a - y_s| + \alpha$.

4. Experiments

We describe in the following sections the experiments we conducted to achieve our best model. Our experiments generally proceeded in four parts: selecting features to train our models, modifying them to improve convergence, experimenting with soft labelling, and finally testing our proposed ordinal triplet loss formulation. All experiments used the Adam optimizer and a learning rate scheduler which decreased the rate by a factor of 0.1 after 10 epochs of no improvement.

4.1. Feature Selection

Table 1 describes the experiments we conducted to select the best features to use for our model. Features tested include the ComParE baseline features, SoundNet features, MFCCs, and raw waveforms. Of the ComParE baseline features, we observed that ComParE, BoAW-2000, and auDeep-fused yielded the best performances for both neural and statistical models. SoundNet features are extracted from the pretrained network with the same name [21]. We used the MFCCs to train a multilayer LSTM augmented with an attention mechanism. The raw waveforms were used to train a deep network comprised of two convolutional layers followed by a multi-layer LSTM. For the SoundNet and baseline features, we used MLPs structured such that each subsequent layer in the network has approximately half the number of units as the previous one. SVM results for ComParE, BoAW-2000, and auDeep-fused are based on those reported in the challenge paper [29]. We observed that of the tested features, the three listed baseline features yielded the best results, as bolded in the table.

Table 1: *Performance on Different Features*

	Model	Spearman (Devel)
SoundNet	MLP	0.030
ComParE	SVM	0.251
	MLP	**0.300**
BoAW-2000	SVM	0.269
	MLP	**0.313**
auDeep-fused	SVM	0.261
	MLP	**0.329**
MFCC	Attention LSTM	0.018
Raw Waveform	CNN LSTM	0.031

4.2. Data Modification

Table 2 on the next page describes the experiments we conducted to modify the input data. Namely, we tested upsampling and weighting the classification loss by class label frequencies as potential approaches to reconcile the skewed data distribution. We also tested applying PCA on the input features before feeding them into the model as a potential approach to reduce the high dimensionality of the features. For all the experiments in this section, we used MLPs with the halving property described in the previous section. We observed that these data modification approaches did not consistently improve the model, and thus did not use them in subsequent experiments.

4.3. Impact of Soft Labels

Table 3 describes the results from using the soft labelling formulation. All experiments in this section also used MLPs with the halving property described earlier. We observe that models trained on soft labels perform noticeably better than models trained on hard labels for two of the three feature types.

Table 2: *Data Modifications*

	Features	Spearman (Devel)
Upsampling	ComParE	0.271
	BoAW-2000	0.308
	auDeep-fused	0.303
PCA	ComParE	0.279
	BoAW-2000	0.325
	auDeep-fused	0.254
Weighted Loss	ComParE	0.279
	BoAW-2000	0.301
	auDeep-fused	0.243

Table 3: *Soft Labels*

Features	Spearman (Devel)
ComParE	**0.311**
BoAW-2000	**0.333**
auDeep-fused	0.322

4.4. Impact of Ordinal Triplet Loss

Table 4 below summarizes our results using ordinal triplet loss. We train all models in this formulation using the Adam optimizer with learning rate 10^{-7}, the joint loss described in Section 3.3, batch sizes of 64, and early stopping with a patience of 10. For our models trained via ordinal triplet loss, f is an MLP with input dimensions halved for each subsequent layer, and g is comprised of two fully connected layers. We observe that utilizing ordinal triplet loss yields noticeable improvement in model performance with respect to the BoAW-2000 feature set.

Table 4: *Ordinal Triplet Loss*

Features	Spearman (Devel)
ComParE	0.308
BoAW-2000	**0.343**
auDeep-fused	0.323

4.5. Analysis of Results

Figure 1 plots the t-SNE visualization of the training data in our model's embedding space. Lighter points represent data samples with higher class labels. The model is able to successfully learn a space that captures desirable ordinal relations, generally mapping data with closer class labels to closer locations in the embedding space.

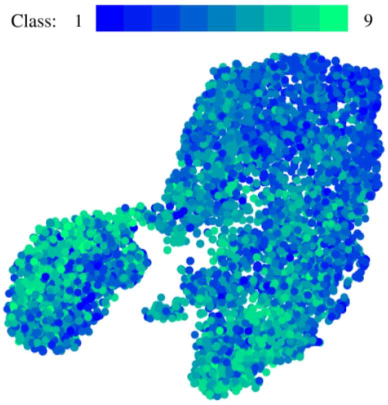

Figure 1: *t-SNE Visualization of Embedding Space*

5. Conclusion

In this work, we present ordinal triplet loss as an effective way to train deep architectures on noisy, complex, ordered data. We show mathematically and empirically that soft labels work particularly effectively in ordinal regression tasks. We propose an ordinal triplet loss function that captures ordinal relations in its embedding space, which we validate empirically on the Sleepiness dataset. Finally, we show that our proposed approach performs well on the Sleepiness dataset. In the future, we are interested in exploring how well our approach performs on continuous data in order to show an effective deep technique on complex regression tasks.

6. References

[1] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[2] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2794–2797.

[3] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[4] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, 2017, pp. 3442–3446.

[5] D. Amodei *et al.*, "Deep speech 2 : End-to-end speech recognition in english and mandarin," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 173–182. [Online]. Available: http://proceedings.mlr.press/v48/amodei16.html

[6] A. van den Oord *et al.*, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmssan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 3918–3926. [Online]. Available: http://proceedings.mlr.press/v80/oord18a.html

[7] K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges," *Computer Vision and Image Understanding*, vol. 163, pp. 3 – 20, 2017, language in Vision. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314217301170

[8] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *CoRR*, vol. abs/1703.10135, 2017.

[9] H. Chen, H. Zhang, P.-Y. Chen, J. Yi, and C.-J. Hsieh, "Attacking visual language grounding with adversarial examples: A case study on neural image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2587–2597. [Online]. Available: https://www.aclweb.org/anthology/P18-1241

[10] Y. Goyal, T. Khot, A. Agrawal, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," *International Journal of Computer Vision*, vol. 127, no. 4, pp. 398–414, Apr 2019. [Online]. Available: https://doi.org/10.1007/s11263-018-1116-0

[11] A. Kuhnle, H. Xie, and A. A. Copestake, "How clever is the film model, and how clever can it be?" in *ECCV Workshops*, 2018.

[12] A. Agrawal, A. Kembhavi, D. Batra, and D. Parikh, "C-vqa: A compositional split of the visual question answering (vqa) v1.0 dataset," *CoRR*, vol. abs/1704.08243, 2017.

[13] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," 2015.

[14] J. Tee and D. P. Taylor, "Is Information in the Brain Represented in Continuous or Discrete Form?" *arXiv e-prints*, May 2018.

[15] H. R. Marateb, M. Mansourian, P. Adibi, and D. Farina, "Manipulating measurement scales in medical statistical analysis and data mining: A review of methodologies," in *Journal of research in medical sciences : the official journal of Isfahan University of Medical Sciences*, 2014.

[16] P. Melville and V. Sindhwani, *Recommender Systems*. Boston, MA: Springer US, 2017, pp. 1056–1066. [Online]. Available: https://doi.org/10.1007/978-1-4899-7687-1_964

[17] W. Chu and Z. Ghahramani, "Gaussian processes for ordinal regression," *J. Mach. Learn. Res.*, vol. 6, pp. 1019–1041, Dec. 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=1046920.1088707

[18] J. Rennie and N. Srebro, "Loss functions for preference levels: Regression with discrete ordered labels," *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, 01 2005.

[19] B. Zhang, Y. Kong, G. Essl, and E. M. Provost, "f-similarity preservation loss for soft labels: A demonstration on cross-corpus speech emotion recognition."

[20] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[21] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016.

[22] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. of NAACL*, 2018.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pretraining of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

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

[25] B. Schuller, S. Steidl, A. Batliner, P. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," in *INTERSPEECH 2018 – 18th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, Proceedings*, 2018.

[26] A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro, *Karolinska Sleepiness Scale (KSS)*. New York, NY: Springer New York, 2012, pp. 209–210. [Online]. Available: https://doi.org/10.1007/978-1-4419-9893-4_47

[27] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.

[28] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," 06 2015, pp. 815–823.

[29] B. W. Schuller *et al.*, "The interspeech 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds orca activity," in *Proceedings INTERSPEECH 2019, Graz, Austria*, 2019.