INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Ensembling End-to-End Deep Models for Computational Paralinguistics Tasks: ComParE 2020 Mask and Breathing Sub-Challenges

Maxim Markitantov[1], *Denis Dresvyanskiy*[2,3], *Danila Mamontov*[2,3], *Heysem Kaya*[4], *Wolfgang Minker*[2], *Alexey Karpov*[1]

[1] St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, Russia
[2] Ulm University, Ulm, Germany
[3] ITMO University, St. Petersburg, Russia
[4] Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

m.markitantov@yandex.ru, denis.dresvyanskiy@uni-ulm.de, danila.mamontov@uni-ulm.de, h.kaya@uu.nl, wolfgang.minker@uni-ulm.de, karpov@iias.spb.su

Abstract

This paper describes deep learning approaches for the Mask and Breathing Sub-Challenges (SCs), which are addressed by the INTERSPEECH 2020 Computational Paralinguistics Challenge. Motivated by outstanding performance of state-of-the-art end-to-end (E2E) approaches, we explore and compare effectiveness of different deep Convolutional Neural Network (CNN) architectures on raw data, log Mel-spectrograms, and Mel-Frequency Cepstral Coefficients. We apply a transfer learning approach to improve model's efficiency and convergence speed. In the Mask SC, we conduct experiments with several pretrained CNN architectures on log-Mel spectrograms, as well as Support Vector Machines on baseline features. For the Breathing SC, we propose an ensemble deep learning system that exploits E2E learning and sequence prediction. The E2E model is based on 1D CNN operating on raw speech signals and is coupled with Long Short-Term Memory layers for sequence modeling. The second model works with log-Mel features and is based on a pretrained 2D CNN model stacked to Gated Recurrent Unit layers. To increase performance of our models in both SCs, we use ensembles of the best deep neural models obtained from N-fold cross-validation on combined challenge training and development datasets. Our results markedly outperform the challenge test set baselines in both SCs.

Index Terms: computational paralinguistics, information fusion, neural networks, transfer learning, end-to-end models

1. Introduction

Human speech contains a wide range of non-verbal information about speaker's states and traits. The Computational Paralinguistics Challenge (ComParE), which has been regularly held in the framework of INTERSPEECH, focuses on automatic recognition of various paralinguistic aspects of human speech. Since 2009, every year the organizers present new tasks and provide for participants novel, challenging data related to such problems, as emotion detection [1] and recognition [2], age, gender [3] and dialects [4] recognition, analysis of intoxication and sleepiness [5], as well as speech analysis to detect 'unhealthy' speech in cold conditions [6]. ComParE 2020 consists of three SCs: Mask, Breathing, and Elderly Emotion [7]. In this paper, we propose E2E deep learning approaches for both Mask and Breathing SCs, using the challenge protocol. In [8], we report our results on the Elderly Emotion SC.

This year, the aforementioned two SCs have highly relevant paralinguistic phenomena due to the spread and prevention of coronavirus pandemic (COVID 19). At the time of paper writing, the number of infected people has exceeded 19 million. While people need to wear masks to avoid the infection, the acoustic characteristics of speakers' breathing may be used for automated pre-screening of potentially infected people. In the Mask SC, the task is to tell apart whether a speaker wears a surgical mask or not [7]. Back in 2008, Mendel et al. [9] tried to determine the effect of the mask on speech understanding during surgery. Spectral analysis of speech showed significant differences between speech filtered by a medical mask and unfiltered speech, but the mask had no effect on people's understanding of a spoken content. Ravanelli et al. investigated automatic speech recognition under a mask in various noisy conditions [10]. Saeidi et al. [11, 12] studied effects of wearing a mask on speaker identification. Additionally, they investigated both passive and active effects of wearing the mask on speech by measuring up acoustic properties of the mask material. The research to tell apart speech pronounced by a human with or without a medical/hand-made mask is an open issue today.

Breathing is a vital function of animal species, and historically had been used in medical diagnosis and therapy [13, 14]. While breathing prediction from the acoustic signal has a value on its own, accurate prediction may also help breaking through depression and post-traumatic stress disorder telemonitoring by providing high-level discriminative features such as acoustic and timing patterns [15, 16].

In this work, we investigate various CNNs and pretrained deep neural networks (DNNs) that receive a lot of attention nowadays, as they are powerful machine learning methods and can be used for multiple purposes, and applying the transfer learning approach can significantly increase the performance of these methods [17]. They are effectively used for feature extraction [18, 19, 20], emotion recognition [21, 22], automatic speech recognition [23], and speech synthesis [24].

2. Background on Methods

2.1. Deep Residual and Recurrent Networks

Since AlexNet [25] and, then, VGG [26] reached state-of-the-art performance in the image recognition domain, the research community realized that, due to the vanishing gradient problem (VGP), it is not efficient just to stack more new convolutional layers in CNN. To tackle VGP, the authors of [27] introduced a simple, but genius idea - "identity shortcut connection", which can skip one or more layers. This idea allows building an arbi-

http://dx.doi.org/10.21437/Interspeech.2020-2666

trarily DNN without a loss of an efficiency in comparison with a less deep network due to its ability to pass forward signals even if all layer's parameters equal zero. Thus, a DNN can by itself "decide" how deep it should be.

Recurrent neural networks (RNNs) are effectively used in cases when time dependent sequences need to be modeled and processed, such as speech, handwriting, text, heartbeats, EEG, or any other numerical sequences. The most commonly used types of RNNs are Long Short-Term Memory (LSTM) models [28] and Gated Recurrent Units (GRU) [29], which solve the vanishing gradient problem of standard RNNs and model long-term dependencies much better than RNNs. In turn, a GRU has less parameters than a LSTM because there is no output gate.

2.2. N-Fold Cross-Validation for Model Optimization

N-Fold cross-validation (CV) is a statistical method for evaluating a generalization performance [30], that allows increasing size of the training data, and as a result the recognition accuracy. In the Mask SC, we have slightly modified the classical CV approach. Firstly, we shuffle and split both training and the development sets into $N/2$ parts using stratification. In the Breathing SC, we split training and development parts without any shuffling. Then we use these N parts for CV. Separate partitioning of the training and development sets maximizes the speaker disjunction.

3. Proposed Approaches

3.1. The Mask Sub-Challenge

Figure 1 shows pipeline of our approach for the Mask SC. Note that in one sub-system we additionally used Support Vector Machines (SVM) exploiting baseline features BoAW [31].

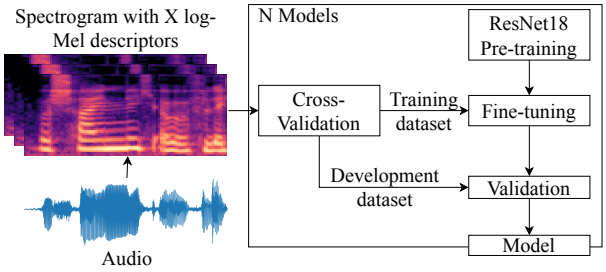

Figure 1: *Pipeline for the Mask Sub-Challenge*

3.1.1. Acoustic Features

In the Mask SC, we used all the baseline features (openSMILE [32, 33], BoAW [31], DeepSpectrum [34], AuDeep [35, 36]) with SVM using linear and non-linear kernels. Additionally, 30 Mel-Frequency Cepstral Coefficients (MFCCs) with their Δ and $\Delta\Delta$ features, as well as 64 Mel-frequency bands (log-Mel) with the window width of 32 ms and an overlap of 10 ms were extracted from each audio file. MFCCs and log-Mels were converted to images and used as inputs for pretrained DNNs. All the features have been normalized per batch with Z-score normalization.

3.1.2. Models

In the Mask SC, we compared results of pretrained AlexNet, ResNet18, ResNet34, ResNet50, ResNet101 and VGG-16 architectures on log-Mel spectrograms. However, in the following experiments, we used ResNet18 architecture, since it gave a better performance on the Mask SC development set, as well as log-Mels.

The first system included four ResNet18 models with 3 extra fully-connected (FC) layers. We removed the last layer and appended 3 fully-connected layers with dropout layers. After that, we fine-tuned all the layers. For simplicity, we call this architecture of neural networks ResNet18v1 in this paper. We applied the Adam optimizer for all the neural networks. We calculated mean Unweighted Average Recall (UAR) over all CV folds in order to choose a model with the best performance. The second system had four ResNet18 models with 1 fully-connected layer with number of neurons equal to the number of classes. We call this neural network ResNet18v2. Stochastic gradient descent (SGD) and Adam were used as optimizers for neural networks; in addition, we used the models with highest UAR scores on each CV fold. The third system is composed of ResNet18v2 and SVM with non-linear kernel which are trained on the 64 log-Mels and on the baseline features BoAW-2000 respectively. We used the first 2 folds to train ResNet18v2 (64 log-Mels), and the remaining folds to train SVM (BoAW-2000). This approach allowed us to increase performance during validation. In the following systems, we increased the number of folds to 10. The fourth system consisted of ten ResNet18v2 models with Adam and SGD optimizers on different folds; we used neural networks with the best performances on each CV fold.

In the fifth system, we used a weighted fusion (WF) to all applied neural networks. In each fold, we weighted predictions of two networks (ResNet18v2 with Adam, and Resnet18v2 with SGD), where the fusion weight is optimized on the respective validation set. Then, we calculated mean prediction of the fold-wise decisions.

All the neural networks were trained to minimize the binary cross-entropy objective. The mini-batch size was 16. We set an initial learning rate of 0.0001 for the Adam and 0.001 with momentum 0.9 for SGD optimizer. We also decreased the learning rate, when the validation loss does not improve for two successive epochs. The training process was stopped after 120 epochs. The models with the smallest validation loss were chosen.

3.2. The Breathing Sub-Challenge

The pipeline of the proposed approach for the Breathing SC is illustrated in Figure 2. In this SC, we used two E2E approaches. The first approach is based on 1D CNN + LSTM RNN model. It has 1D CNN that uses directly the raw data. The second approach is based on the pretrained ResNet18 model with two stacked RNN-GRU layers above (ResNet18 + GRU) and uses log-Mel spectrograms. As a final submission, we present feature and decision level fusion applied on best models from both first and second approaches. All the models used a loss function defined as 1-*r*, where *r* is the Pearson's correlation coefficient (PCC).

3.2.1. 1D CNN + LSTM model

Since there are no available pretrained 1D CNN models on raw acoustic data, we created our own. In order to capture temporal dependencies from 1D CNN model's embeddings more efficiently, we have also stacked two LSTM layers above it. Then, to flatten the output, a dense layer of one neuron with $tanh$ activation function was added. Thus, we propose E2E sequence-to-sequence model, which directly maps input data in

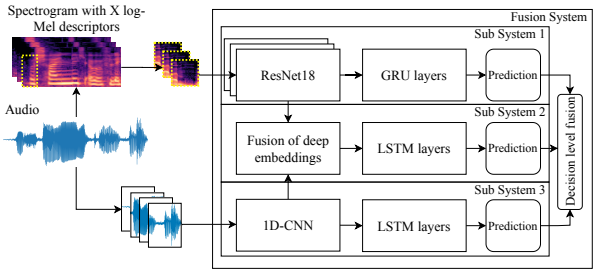

Figure 2: *Pipeline for the Breathing Sub-Challenge*

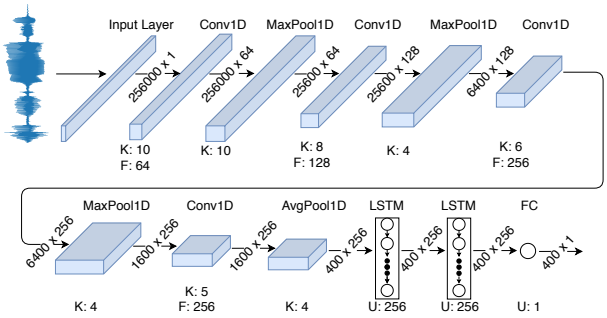

Figure 3: *Architecture of 1D CNN + LSTM medium-size model in the Breathing Sub-Challenge. K - kernel size, F - number of filters (kernels), U - number of units.*

the waveform format into a sequence of breath belt signals.

Additionally, to get more data, we cut each audio file into several parts overlapping in 2/5 of the part length. Overlaps of outputs were averaged after prediction. Hereinafter, we call the length of part as window size and the shift (how much one part overlaps with another one) as step of window. To choose the best window size in terms of PCC, we conducted tests with each proposed 1D CNN + LSTM model. The test results showed best models performance on window sizes of 16 and 24 seconds.

Overall, for case of raw data we have built two different types of 1D CNNs in terms of the complexity: medium and complex ones with around 1.4 M and 3.5 M parameters correspondingly. The first two submitted systems had the medium complexity. As an example, Figure 3 shows the medium model with data input shape of 16 seconds (one second includes 16K samples in a raw waveform and 25 breath belt signals). After each convolutional layer, dropout with a probability of 0.3 was applied. It should be noted that the first submitted model had one convolutional layer less, as well as less amount of convolutional filters on each layer.

For the third submission we built a complex model with the architecture similar to ResNet, but in 1D manner. Thus, the 1D CNN model was replaced with seven ResNet-like residual blocks. LSTM part of the model remained the same except the first LSTM layer; it had 512 neurons vs. 256 in the second system. We chose Adam as an optimizer with the learning rate of 0.001. Window sizes were chosen as 16 and 24 seconds, and mini-batch size - 40 and 26 correspondingly. We stopped the training process after 150 epochs. The model with the best value of the PCC on the validation set was used for the first test submission. Our first system showed better values of the PCC on the development set in comparison with the baseline result. However, its effectiveness on the test set was worse than the baseline. To reduce prediction variance, we applied CV with 4 folds on combined training and development sets for all further submissions. The proposed ensemble of 4 models was applied to make final predictions on the test set.

3.2.2. ResNet18 + GRU model

In this approach, we used pretrained 2D CNN ResNet18 with stacked 2 GRU layers above (512 and 64 units respectively). The last fully connected layer in ResNet18 has been replaced by GRU layers. We used GRUs instead of LSTM units because they have fewer parameters to train. During the training process we used dropout with a probability of 0.3 between ResNet18's output and the first GRU layer as well as between GRU layers. The model contains approximately 12 M trainable weights. We extracted Mel-spectrograms from the raw speech signal. It was computed by the window of 40 ms with 20 ms overlap and 128 Mel-frequency bands. Then these log-Mels were cut into pieces of 1 second, that the middle of each such piece corresponds to one of the target values. Note that in order to be able to get log-Mel for target values at the edges of audio recordings, where the target value would correspond to the middle, we extended the raw speech signal of each participant by 0.5 sec at the beginning and at the end by mirroring. Since we had to build sequence-to-sequence model, we passed through ResNet18 the sequence of log-Mels with a certain window size. We have tested several window lengths of {14, 16, 20, 24, 32, 40, 80} sec. The best results on the development set were achieved with the window length of 80 seconds. The step of window was taken as 1/2 of a window length. Due to computing power limitations, wider windows could not be used. Also we did not use mini-batches and trained on each sample. The weights of the first and second ResNet18 blocks were frozen, only the weights of 3 and 4 blocks were trained. As optimizer we have used Adam with a learning rate of 0.002, and 20 epochs of training. The ResNet18 + GRU model was used for the fourth submission.

3.2.3. Fusion Techniques

To combine the best systems, we used both fusion of deep embeddings (DE) and decision level fusion schemes. For DE fusion, we extracted deep embeddings from last layers of the best 1D CNN and Resnet18, concatenated them and trained a 2-layer LSTM network with 512 and 256 units, respectively. Before the training, obtained DEs were normalized.

Furthermore, the decision level fusion was applied as follows: we evaluated weighted mean prediction from two best systems and DE fusion system per each fold independently. To choose weights for each fold, we generated triplets from a Dirichlet distribution (using python's NumPy library function *numpy.random.dirichlet*) 1000 times and took the best one in terms of PCC. Thus, per each fold we have 3 models with corresponding weights. Final predictions on the test set were made independently by fold-wise fusion systems, which were averaged as in the other submissions.

4. Experimental Results

In the paper, we do not present results obtained on the development sets, since we used all data in N-Fold CV, which significantly increased the performance. Therefore, we report mean value of N-fold CV. Table 1 summarizes results of the submitted systems in both SCs. The official performance measure for the Breathing SC is PCC, while it is UAR for the Mask SC.

Table 1: *Cross-validation and test set results for the Mask and Breathing Sub-Challenges. X log-Mel: Spectrogram with X log-Mel descriptors. WF: weighted fusion.* ***UAR****: Unweighted Average Recall. PCC: Pearson's correlation coefficient. Perf(ormance).*

System ID	Features/Input	Model Description	CV Perf.	Test Set Perf.
The Mask Sub-Challenge, UAR [%] (baseline UAR=71.8%)				
1	64 log-Mel	ResNet18v1	77.23	74.00
2	64 log-Mel	ResNet18v2 (Adam + SGD)	79.86	75.10
3	64 log-Mel + BoAW-2000	ResNet18v2 (SGD) + SVM	78.55	74.90
4	64 log-Mel	ResNet18v2 (Adam + SGD)	82.17	75.30
5	64 log-Mel	WF(ResNet18v2 (Adam), Resnet18v2 (SGD))	**84.32**	**75.90**
The Breathing Sub-Challenge, PCC (baseline PCC=0.731)				
1	Raw signal	1D CNN + LSTM (single model, without N-Fold CV)	0.545	0.660
2	Raw signal	1D CNN + LSTM (medium-scale structure)	0.607	0.744
3	Raw signal	1D CNN + LSTM (complex structure)	0.621	0.718
4	128 log-Mel	ResNet18 + GRU	0.580	0.734
5	Raw signal + 128 log-Mel	Fusion System (see Figure 2)	**0.640**	**0.763**

4.1. The Mask Sub-Challenge

In the Mask SC, the first submission was made with the ResNet18v1 trained on 64 log-Mel, which gave the mean CV UAR=77.23% and test set UAR=74.00%. The second submission was made via a combination of four ResNet18v2 with different optimizers that resulted in UAR of 79.13% and 75.10% for CV and test set, respectively. The third submission fused the probability scores from two ResNet18v2 neural networks and SVMs with non-linear kernels trained on 64 log-Mel and BoAW-2000 features, respectively. This system gave the mean CV UAR=78.55% and the test set UAR=74.90% that was slightly worse than the second submission. The fourth submission was made via a combination of ten ResNet18v2 models trained with alternative (Adam and SGD) optimizers, which lead to UAR scores of 82.17% and 75.30% for CV and test set, respectively. This system was also trained using 64 log-Mel features. As a final, fifth submission, we used the weighted fusion of CV folds of ResNet18v2 models with various optimizers. As expected, the WF fusion outperformed the fourth submission results. Our final system yielded a mean CV UAR=84.32% and the test set UAR=75.90%, and has significantly outperformed the challenge baseline result by an absolute difference of 4.1%.

4.2. The Breathing Sub-Challenge

As mentioned above, the model for the first submission used 1D CNN + LSTM architecture, processed raw signals and had less parameters in comparison with other 1D CNN + LSTM architectures applied. This yielded a test set PCC of 0.660. For the second submission we restructured 1D CNN + LSTM model by increasing the number of filters, as well as the number of convolutional and pooling layers. We also used the CV approach to get an ensemble of four 1D CNN + LSTM models obtained from 4-folds, where each model is trained with more data compared to the initial system. The proposed system has significantly outperformed the baseline system and gave PCC=0.744 on the test set. The mean PCC value in CV was 0.607.

Assuming that a more complex model may be more effective, the third submission was made using a ResNet-like model with 3.5 M parameters. For this trial, we also used the CV approach to get an ensemble. The mean CV PCC=0.621; however, on the test set it reached a lower PCC (0.718) compared to the simpler model. The possible reason of this failure may be caused insufficient data amount for training such huge model.

One way to tackle with small data amount is using the transfer learning approach. Our fourth submission was made using pretrained ResNet18 + GRU approach that gave the mean CV PCC=0.58 and outperformed the SC baseline result with PCC=0.734. As the final fifth submission, we used both feature and decision level fusion of models from the second and fourth trials. The mid-level fusion was conducted via DE extracted from both 1D CNN and Resnet18 models by LSTM model, which gave a mean PCC=0.598. Then, we applied a decision level fusion by weighted means of each fold separately. In the end, we averaged four test predictions obtained from each fold. Our final system showed the mean CV PCC=0.640 and the test set PCC=0.763, and has significantly outperformed the baseline test set PCC score of 0.731.

5. Conclusions

In this paper, we investigated advantages of different deep neural network (DNN) approaches on several types of features, with a special focus on ensembling various DNNs, including E2E models, for the Mask and Breathing SCs addressed by INTERSPEECH ComParE 2020. The ensemble models increase the generalization performance, when the individual models are diverse and sufficiently accurate. DNNs provide an outstanding performance, when their hunger for the data is satisfied. We have shown that by effectively increasing the training data via N-Fold CV, careful selection of pretrained models and neat design of E2E neural networks, state-of-the-art DNN ensembles can be obtained. Moreover, by fusing them, we have reached UAR=75.90%, and PCC=0.763 in the Mask and the Breathing SCs, respectively, both of which markedly outperformed the competitive challenge baselines. Scripts of this work can be found at GitHub[1].

6. Acknowledgements

The research was supported by RFBR project No. 20-04-60529 (the Mask SC); by RSF project No. 18-11-00145 (acoustic modeling in the Breathing SC); by the German Federal Ministry of Education and Research project "RobotKoop: Cooperative Interaction Strategies and Goal Negotiations with Learning Autonomous Robots"; as well as by the Government of Russia, grant No. 08-08.

[1] https://github.com/DresvyanskiyDenis/compare20-MB

7. References

[1] B. W. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *INTERSPEECH*, 2009, pp. 312–315.

[2] B. W. 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 *INTERSPEECH*, 2013, pp. 148–152.

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

[4] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson *et al.*, "The INTERSPEECH 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity," in *INTERSPEECH*, 2019, pp. 2378–2382.

[5] B. W. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *INTERSPEECH*, 2011, pp. 3201–3204.

[6] B. W. 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 *INTERSPEECH*, 2017, pp. 3442–3446.

[7] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly emotion, Breathing & Masks," in *INTERSPEECH*, Shanghai, China, October 2020, to appear.

[8] G. Sogancioglu, O. Verkholyak, H. Kaya, D. Fedotov, T. Cadee, A. A. Salah, and A. Karpov, "Is Everything Fine, Grandma? Acoustic and Linguistic Modeling for Robust Elderly Speech Emotion Recognition," in *INTERSPEECH*, Shanghai, China, October 2020, to appear.

[9] L. L. Mendel, J. A. Gardino, and S. R. Atcherson, "Speech understanding using surgical masks: a problem in health care?" *Journal of the American Academy of Audiology*, vol. 19, no. 9, pp. 686–695, 2008.

[10] M. Ravanelli, A. Sosi, M. Matassoni, M. Omologo, M. Benetti, and G. Pedrotti, "Distant talking speech recognition in surgery room: The domhos project," in *Proc. AISV*, 2013, 13 pages.

[11] R. Saeidi, T. Niemi, H. Karppelin, J. Pohjalainen, T. Kinnunen, and P. Alku, "Speaker recognition for speech under face cover," in *INTERSPEECH*, 2015, pp. 1012–1016.

[12] R. Saeidi, I. Huhtakallio, and P. Alku, "Analysis of face mask effect on speaker recognition," in *INTERSPEECH*, 2016, pp. 1800–1804.

[13] F. Goldman-Eisler, "Speech-breathing activity-a measure of tension and affect during interviews," *British Journal of Psychology*, vol. 46, no. 1, pp. 53–63, 1955.

[14] A. Lewandowski and A. I. Gillespie, "The relationship between voice and breathing in the assessment and treatment of voice disorders," *Perspectives of the ASHA Special Interest Groups*, vol. 1, no. 3, pp. 94–104, 2016.

[15] A. A. Akdag Salah, M. Ocak, H. Kaya, E. Kavcar, and A. A. Salah, "Hidden in a breath: Tracing the breathing patterns of survivors of traumatic events," in *Digital Humanities*, 2019, 4 pages.

[16] H. Kaya, D. Fedotov, D. Dresvyanskiy, M. Doyran, D. Mamontov, M. Markitantov, A. A. Akdag Salah, E. Kavcar, A. Karpov, and A. A. Salah, "Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC '19, 2019, pp. 27–35. [Online]. Available: https://doi.org/10.1145/3347320.3357691

[17] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[18] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *INTERSPEECH*, 2011, pp. 237–240.

[19] S. Dalmia, X. Li, F. Metze, and A. W. Black, "Domain robust feature extraction for rapid low resource asr development," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 258–265.

[20] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*. IEEE, 2013, pp. 8614–8618.

[21] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition," in *INTERSPEECH*, 2013, pp. 3366–3370.

[22] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, "Bimodal speech emotion recognition using pre-trained language models," *arXiv preprint arXiv:1912.02610*, 2019, 10 pages.

[23] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2011.

[24] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017, 10 pages.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014, 14 pages.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014, 5 pages.

[30] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer New York, 2009, ch. Ensemble Learning, pp. 605–624.

[31] M. Schmitt and B. Schuller, "Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 1–5, 2017.

[32] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[33] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in Emotion Science*, vol. 4, pp. 1–12, 2013.

[34] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. W. Schuller, "Snore sound classification using image-based deep spectrum features," in *INTERSPEECH*, 2017, pp. 3512–3516.

[35] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *Proc. of the DCASE 2017 Workshop*, 2017, 5 pages.

[36] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "audeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6340–6344, 2017.