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

Learning Higher Representations from pre-trained Deep Models with Data Augmentation for the COMPARE 2020 Challenge Mask Task

Tomoya Koike[1], *Kun Qian*[1*], *Björn W. Schuller*[2,3], *and Yoshiharu Yamamoto*[1]

[1]Educational Physiology Laboratory, The University of Tokyo, Japan
[2]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[3]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
{tommy, qian, yamamoto}@p.u-tokyo.ac.jp, schuller@ieee.org

Abstract

Human hand-crafted features are always regarded as expensive, time-consuming, and difficult in almost all of the machine-learning-related tasks. First, those well-designed features extremely rely on human expert domain knowledge, which may restrain the collaboration work across fields. Second, the features extracted in such a brute-force scenario may not be easy to be transferred to another task, which means a series of new features should be designed. To this end, we introduce a method based on a transfer learning strategy combined with data augmentation techniques for the COMPARE 2020 Challenge *Mask* Sub-Challenge. Unlike the previous studies mainly based on pre-trained models by image data, we use a pre-trained model based on large scale audio data, i. e., AudioSet. In addition, the *SpecAugment* and *mixup* methods are used to improve the generalisation of the deep models. Experimental results demonstrate that the best-proposed model can significantly ($p < .001$, by one-tailed z-test) improve the unweighted average recall (UAR) from 71.8 % (baseline) to 76.2 % on the test set. Finally, the best result, i. e., 77.5 % of the UAR on the test set, is achieved by a late fusion of the two best proposed models and the best single model in the baseline.

Index Terms: Computational Paralinguistics, Speech under Mask, Deep Learning, Data Augmentation, Transfer Learning

1. Introduction

For computational paralinguistic tasks, extracting efficient and robust representations from the audio data is a prerequisite [1]. In the past decade, *deep learning* (DL) [2] has dramatically changed machine learning (ML) from classic paradigms of human hand-crafted features plus shallow models to training deep models which can extract higher representations directly from the data itself via a series of non-linear transformations. In particular, using a pre-trained convolutional neural network (CNN) [3] to learn higher representations under a *transfer learning* [4] paradigm has become increasingly popular. Recently, pre-trained CNNs have been demonstrated to be efficient to fulfil the tasks of snore sound classification [5,6], heart sound classification [7], and acoustic scene classification [8, 9]. Even though the results in the aforementioned studies were encouraging, one factor was ignored: that the pre-trained CNNs used previously were based on image recognition tasks rather than audio classification tasks. Therefore, some high-level features inherited in the audio data may not be extracted from those CNNs pre-trained on image data. Motivated by our most recent work in [10], we introduce an audio-based pre-trained model to the field of computational paralinguistic tasks. We use the data provided by the COMPARE 2020 Challenge **Mask** Sub-Challenge (**MSC**) [11]. In the **MSC**, models should be trained to fulfil the task of telling if the speaker is wearing a surgical mask or not. This application may not only facilitate forensics and communication between surgeons [12], but also trigger potential *computer audition* (CA) based intelligent systems to combat the COVID-19 pandemic [13–15] (e. g., automatic checking if a person is obeying the rule of wearing a mask in a public scenario).

The main contributions of this paper can be summarised as: First, we introduce the large-scale pre-trained audio neural networks (PANNs) [16] to the **MSC**. To the best of our knowledge, it is the first work on using audio-based pre-trained deep models for the COMPARE challenges. Second, to improve the generalisation of the deep models, we use and compare two data augmentation techniques, i. e., SpecAugment [17] and mixup [18]. Third, we use a late fusion strategy, including the *snapshot ensembles* [19] which contributes as important part of late fusion without additional training time.

The following section will introduce the related work and the background of this study. Then, the database, methods, and toolkits we use will be described in Section 3. The experimental results are shown in Section 4 followed by a discussion in Section 5. Finally, we give conclusions in Section 6.

2. Related Work

In the last year's COMPARE Challenge [20], two winners had used a CNN as the feature extractor [21, 22]. Nevertheless, building a well-designed CNN architecture is not an easy task the same as extracting human hand-crafted features. One of the solutions is to use pre-trained CNN models in a transfer learning paradigm.

As introduced in Section 1, numerous pre-trained CNN models were based on image data, which may restrain the capacity to learn higher representations from audio data. Kong *et al.* proposed the PANNs which are pre-trained on large-scale audio data, i. e., the AudioSet [23]. In one of our most recent studies [10], it was demonstrated that using pre-trained models based on audio data can be superior to models pre-trained on image data. Motivated by this success, we introduce PANNs in this year's COMPARE Challenge. Moreover, to improve the generalisation of the deep models, we involve a series of data augmentation techniques in this study.

3. Materials and Methods

An overview of our proposed methods is shown in Figure 1. We use the PANNs to extract higher representations from the log-

* corresponding author

http://dx.doi.org/10.21437/Interspeech.2020-1552

Mel spectrograms of the audio clips. Then, data augmentation methods are used to improve the generalisation of the models.

3.1. PANNs: AudioSet Pre-training

Transfer learning has become the dominant method to get robustness in many fields such as computer vision [24], natural language processing [25], and speech recognition [26]. PANNs [16] are one of the pre-trained CNN models for audio-related tasks, which is characterized in terms of being trained with the AudioSet [23] dataset. The CNN model we chose from PANNs is composed of 6 convolutional blocks and two fully connected layers, which are 14 layers in total. One convolutional block is made up of two 3×3 convolutional filters, two batch normalisation layers, and an average pooling layer at the end of a block, shown as a figure in the middle of Figure 1. PANNs score high accuracy in many audio-related tasks by pre-training with AudioSet. We compare PANNs with ResNet [27], in both cases with or without pre-training. We use the weights pre-trained on AudioSet as a pre-training, which are distributed by the authors of PANNs.

3.2. SpecAugment

Data augmentation has been proposed as a method to generate additional training data for Automatic Speech Recognition [28, 29]. In SpecAugment [17] (see Figure 2), a log-Mel spectrogram as an input feature is masked with a block of consecutive time steps or Mel frequency channels. This augmentation simulates the disfunction of a microphone at a particular time as a mask of time steps or the disappearance in some frequency bands due to an echo as a mask of frequency steps. Time warping, which is a deformation of the time-series in the time steps, is also proposed in SpecAugment, but not used in this paper. We explore the rate of the drop both in the time and frequency domain.

3.3. Mixup

Mixup [18] (see Figure 3) is the method in which an augmented input and label are generated from the random mix of two inputs and corresponding labels. We can denote mixup method as:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j,$$

where an augmented input is $\tilde{x}$, the corresponding label is $\tilde{y}$, x_i, and x_j are raw inputs, and y_i and y_j are one-hot encodings of corresponding labels. The random mixing proportion λ is chosen from a beta distribution with the parameter α, which is denoted as $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$, for $\alpha \in (0, \infty)$. By linearly interpolating a new instance between the inputs, the model is expected to predict the middle of two labels without sharp transitions. We explore several values of α in the experiments.

3.4. Late Fusion

When predicting on the test dataset, the standard procedure is to train the model with the combination of the train dataset and the development dataset. However, this procedure makes it impossible to stop the training before over-fitting. To attenuate the over-fitting, we can ensemble the predictions from several snapshots of the model weights at some points in epochs with reference to Snapshot Ensembles [19]. It should be noted that Snapshot Ensembles can be applied with no additional computational cost to train the model, because the difference from the normal training is only saving snapshots in some epochs while the normal majority vote by several different models cannot. If the number of saved snapshots is denoted as n, it takes time by n times on test predictions as the same with a common majority vote by several different models. We set the number of snapshots, which is the same as the number of ensembles, as 5. Those five snapshots are the weights from 10, 15, 20, 25, and 30 epochs. The late fusion of the proposed models and the (best) baseline model is also evaluated.

4. Experimental Results

For the Mask Sub-Challenge – as outlined above – we aim to build a robust model for binary classification to predict whether the person who speaks wears a mask or not. The evaluation metric for this sub-challenge is unweighted average recall (UAR) [11].

The structure of this section is as follows. First, the experimental setup, including pre-processing and model training, are presented in 4.1. Second, we compare and analyse the results on two CNN models with or without pre-training, different window size and stride of STFT, varying drop-rate of the augmentation and alpha of the beta distribution for mixup. Lastly, fused scores on the development and evaluation set are reported in 4.2.

4.1. Setup

As a pre-processing, we calculate a log-Mel spectrogram from an audio input with 200 Mel filters. We explored the window size and window stride of the short-time Fourier transform (STFT) as hyper-parameters. Binary cross-entropy is taken as a loss function. The optimiser of the model is Adam [30], and the learning rate is set to decrease in each epoch by 0.99 times, starting from 0.0001. After deciding to use PANNs as a classifier, we train the model for 30 epochs from the weights which are pre-trained on AudioSet [23] except for the last fully connected layer. This setting is kept in all of the experiments we note after here. Python 3.7.6 and Pytorch 1.4.0 are used to implement those methods and models above. For the reproducibility of the experiments, we use our open source toolkit, i. e., DEEPSELF [31].

4.2. Results

We explore four kinds of experiments: Pre-training and CNN model selection, log-Mel window size and stride, SpecAugment [17], and mixup [18]. The comparison between ResNet [27] and PANNs, with or without pre-training, is conducted with the results shown in Figure 4(a). ResNet is pre-trained on ImageNet [32], and the PANNs are – as outlined – pre-trained on AudioSet [23]. As shown in Figure 4(b), the best results in UAR on the development dataset are 68.4 % UAR with 50 ms window size and 2 ms window stride.

The range of the drop-rate of SpecAugment is 0.0, 0.02, and 0.05 in time steps and frequency channels, listed with the corresponding results in Figure 4(c). We set 50 ms and 2 ms as window size and window stride of the spectrogram and started this experiment. Thus, the score with no drop (where the values of the drop-rate in both the time steps and the frequency channels are zero in Figure 4(c)) is the same setting as for the best result in Figure 4(b). The results show that the SpecAugment does not contribute to the score in the Mask Sub-Challenge.

We further explore the parameter of mixup, i. e., α of the beta distribution, from 0.0 to 0.3 by 0.1. The model achieves

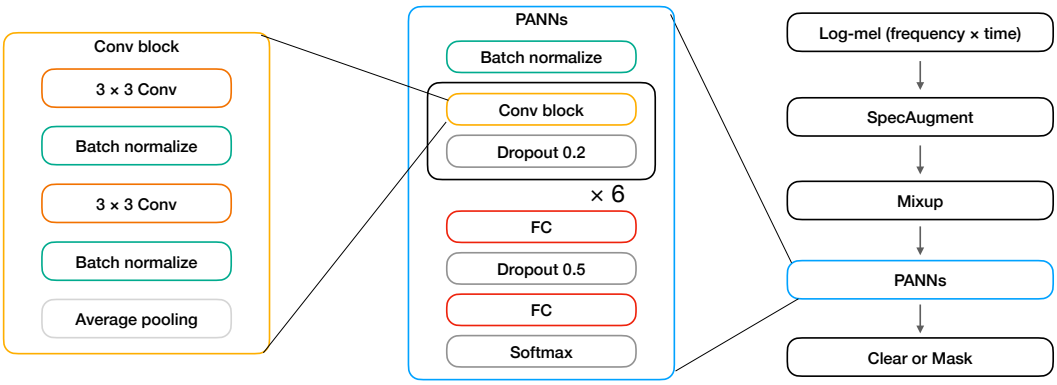


Figure 1: *Overview of our proposed system at the right, and PANN's structure in the middle and left. An input audio clip is preprocessed into log-Mel and dropped at some rate by SpecAugment. Mixup follows that procedure, and PANNs extract features from them and classify whether the speaker wears a mask or not.*

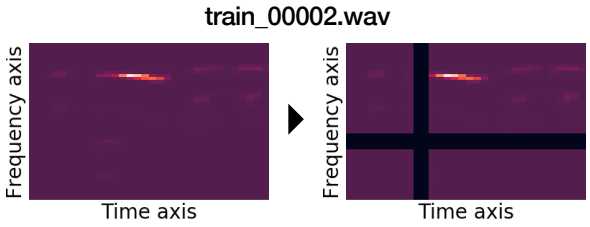


Figure 2: *One example of SpecAugment with two training data. Left: Log-mel spectrogram of train_00002.wav. Right: Augmented spectrogram generated from the log-Mel spectrogram of the instance train_00002.wav. Black bands in frequency axis and time axis are the dropped steps assuming the real environment of recording.*

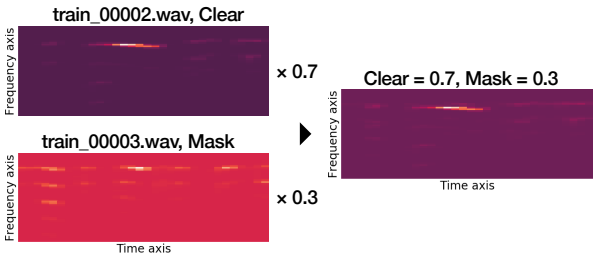


Figure 3: *One example of mixup with two training data. Left upper: Log-mel spectrogram of train_00002.wav which label is clear speech. Left Bottom: Log-mel spectrogram of train_00003.wav which label is mask. Right: A mixed spectrogram of the two spectrograms.*

Table 1: *The other evaluation metrics (in [%]) of the best model on the development set. The notations of the metrics are: F1-score (F1), accuracy (Acc), precision (Prec), specificity (Spec), and sensitivity (Sens).*

	F1	Acc	Prec	Spec	Sens
Pre-trained ResNet	70.4	67.0	68.8	60.7	72.2
Pre-trained PANNs	70.8	68.5	71.6	66.8	70.0
+ mixup	70.4	68.8	72.9	69.8	68.0

Table 2: *Confusion matrix (normalised: in [%]) of the best single model on the dev set.*

Pred ->	**Clear**	**Mask**
Clear	69.8	32.2
Mask	32.0	68.0

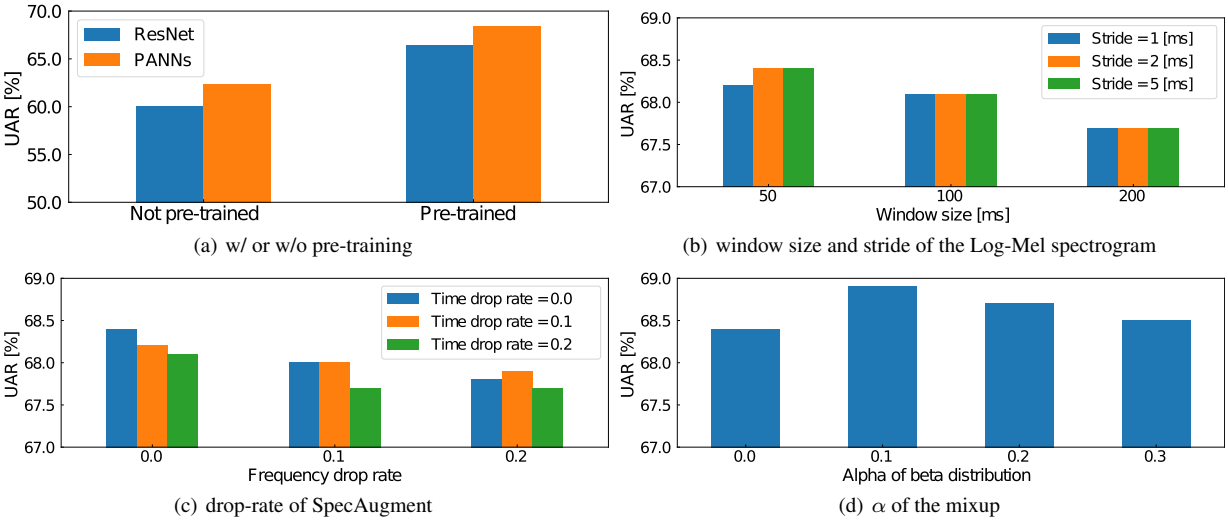
68.9 % in UAR on the development set with α equalling 0.1, as shown in Figure 4(d).

In Table 3, "Baseline single best" is the result by DEEPSPECTRUM and SVM [11], which reaches the highest score in Test UAR with a single model. "Baseline fusion best" stems from the majority vote of all baseline models [11]. Our proposed method is denoted as PANNs plus mixup, which scored 68.9 % and 76.2 % in UAR on the development dataset and the test dataset, respectively. Majority vote of predictions from Snapshot Ensembles scored 75.6 % as a UAR on the test set, which exceeds the score from "Baseline fusion best" by 5.4 %. The late fusion listed in the last row of Table 3 is composed of three models, which are DEEPSPECTRUM, PANNs plus mixup, and Snapshot Ensembles of PANNs plus mixup.

5. Discussion

Our proposed model can beat the best baseline model in this study (76.2 % of UAR vs 71.8 % of UAR). It is worth mentioning that Snapshot Ensembles from a single model outperformed the fusion of 15 baseline models. This is, because the training time does not increase with the former approach while the latter increases, as the number of the models to vote increases. From the result that Snapshot Ensembles did not improve the robustness on the test set compared with the single best model, we could assume that the model was not trained enough to over-fit to the training set. We should still investigate further about this point with other databases.

From Section 4.2, we can say that mixup contributed to improving robustness of the model while SpecAugment did not. This could be because of the dropout layers in PANNs which



(a) w/ or w/o pre-training

(b) window size and stride of the Log-Mel spectrogram

(c) drop-rate of SpecAugment

(d) α of the mixup

Figure 4: *The initial experimental results on comparing and selecting pre-trained models on the dev. set.*

Table 3: *The results (UARs in [%]) achieved by the models via the late fusion strategy. The "Late fusion" result was obtained by the majority vote of three predictions which are "Baseline single best", "PANNs + mixup" and "+ Snapshot Ensembles"*

	Dev	Test
Baseline single best [11]	63.4	70.8
Baseline fusion best [11]	—	71.8
PANNs + mixup	68.9	76.2
+ Snapshot Ensembles	—	75.6
Late fusion	—	**77.5**

make the model robust to the perturbation of the input at the feature level by dropping some inputs of the dropout layers. Candidates of the length of window stride in Figure 4(b) are relatively shorter than [21]. However, the result shows that the relatively higher resolution of the time-domain was attributed to the score regardless of some frequency leakage.

In this study, it is demonstrated that, using a pre-trained model by audio data can benefit more for the **MSC** than a pre-trained model by image data. This finding is consistent with our previous work in health-related audio classification tasks [10]. In addition, *data augmentation* is essential to improve the generalisation of the deep models. Further studies are necessary to reveal situations which require data augmentation like this MSC dataset does. Finally, late fusion is helpful to improve the performance.

PANNs plus mixup is the weight after 30 epochs training, while Snapshot Ensembles are the weights from 10, 15, 20, 25, 30 epochs. We can think of the fusion of those two predictions as a weighted ensemble of different epochs in one model. While there is a possibility that the baseline single best model had diversity compared with Snapshot Ensembles, this result might confirm that PANNs plus mixup can be improved with further training.

However, we still lack explanations about the best model. In future studies, we will explore the learnt higher representations from the proposed models. Furthermore, we will explore more advanced data augmentation methods such as generative adversarial networks (GANs) [33, 34]. We will consider exploring the capacity of the proposed methods on other audio based health applications like snore sound classification task [35, 36].

6. Conclusions

In this paper, we introduced various techniques used to tackle the ComParE 2020 competition, especially for the Mask sub-challenge. We augmented a log-Mel spectrogram with SpecAugment, and trained the model in a mixup manner to increase the data volume and its variety. Furthermore, we adopted the DNN model pre-trained on the AudioSet and fine-tuned it with reference to PANNs.

We found that PANNs with mixup augmentation improved the score while SpecAugment did not. Furthermore, the best model was enough to reach robustness on the test set without Snapshot Ensembles and surpassed the baseline fusion by scoring 76.2 % on the test set. The late fusion of the proposed models and the best baseline model scored 77.5 % on the test set. Future efforts beyond those named will also include analysis of generalisability onto other paralinguistic tasks.

7. Acknowledgements

This work was partially supported by the Zhejiang Lab's International Talent Fund for Young Professionals (Project HANAMI), P. R. China, the JSPS Postdoctoral Fellowship for Research in Japan (ID No. P19081) from the Japan Society for the Promotion of Science (JSPS), Japan, and the Grants-in-Aid for Scientific Research (No. 19F19081 and No. 17H00878) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

8. References

[1] B. W. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. New York, NY, USA: Willey, 2013.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition

with a back-propagation network," in *Proc. NIPS*, Denver, CO, USA, 1989, pp. 396–404.

[4] 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, 2010.

[5] 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 *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3512–3516.

[6] M. Freitag, S. Amiriparian, N. Cummins, M. Gerczuk, and B. W. Schuller, "An 'end-to-evolution' hybrid approach for snore sound classification," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3507–3511.

[7] Z. Ren, N. Cummins, V. Pandit, J. Han, K. Qian, and B. W. Schuller, "Learning image-based representations for heart sound classification," in *Proc. DH*. Lyon, France: ACM, 2018, p. 143–147.

[8] Z. Ren, V. Pandit, K. Qian, Z. Yang, Z. Zhang, and B. W. Schuller, "Deep sequential image features for acoustic scene classification," in *Proc. DCASE Workshop*, Munich, Germany, 2017, pp. 113–117.

[9] Z. Ren, K. Qian, Z. Zhang, V. Pandit, A. Baird, and B. W. Schuller, "Deep scalogram representations for acoustic scene classification," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 3, pp. 662–669, 2018.

[10] T. Koike, K. Qian, Q. Kong, M. D. Plumbley, B. W. Schuller, and Y. Yamamoto, "Audio for audio is better? An investigation on transfer learning models for heart sound classification," in *Proc. EMBC*, Montréal, Canada, 2019, pp. 1–4, in press.

[11] 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 *Proc. INTERSPEECH*, Shanghai, P. R. China, September 2020, pp. 1–5, to appear.

[12] 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.

[13] K. Qian, X. Li, H. Li, S. Li, W. Li, Z. Ning, S. Yu, L. Hou, G. Tang, J. Lu, F. Li, S. Duan, C. Du, Y. Cheng, Y. Wang, L. Gan, Y. Yamamoto, and B. W. Schuller, "Computer audition for healthcare: Opportunities and challenges," *Frontiers in Digital Health*, vol. 2, no. 5, pp. 1–4, 2020.

[14] B. W. Schuller, D. M. Schuller, K. Qian, J. Liu, H. Zheng, and X. Li, "COVID-19 and computer audition: An overview on what speech & sound analysis could contribute in the SARS-CoV-2 Corona crisis," *arXiv preprint arXiv:2003.11117*, pp. 1–7, 2020.

[15] J. Han, K. Qian, M. Song, Z. Yang, Z. Ren, S. Liu, J. Liu, H. Zheng, W. Ji, T. Koike, X. Li, Z. Zhang, Y. Yamamoto, and B. W. Schuller, "An early study on intelligent analysis of speech under COVID-19: Severity, sleep quality, fatigue, and anxiety," *arXiv preprint arXiv:2005.00096*, pp. 1–5, 2020.

[16] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition." *arXiv preprint arXiv:1912.10211*, pp. 1–13, 2019.

[17] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. INTERSPEECH*, Graz, Austria, 2019.

[18] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, pp. 1–13, 2017.

[19] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot Ensembles: Train 1, get m for free," in *Proc. ICLR*, Toulon, France, 2017.

[20] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson, A. Cristia, A. Seidl1, A. S. Warlaumont, L. Yankowitz1, E. Nöth, S. Amiriparian, S. Hantke, and M. Schmitt, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian dialects, Continuous sleepiness, Baby sounds & Orca activity," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 2378–2382.

[21] H. Wu, W. Wang, and M. Li, "The DKU-LENOVO systems for the INTERSPEECH 2019 Computational Paralinguistic Challenge," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 2433–2437.

[22] S.-L. Yeh, G.-Y. Chao, B.-H. Su, Y.-L. Huang, M.-H. Lin, Y.-C. Tsai, Y.-W. Tai, Z.-C. Lu, C.-Y. Chen, T.-M. Tai, C.-W. Tseng, C.-K. Lee, and C.-C. Lee, "Using attention networks and adversarial augmentation for styrian dialect continuous sleepiness and baby sound recognition," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 2398–2402.

[23] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, New Orleans, LA, USA, 2017, pp. 776–780.

[24] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better," in *Proc. CVPR*, Long Beach, CA, USA, 2019, pp. 2661–2671.

[25] A. Malte and P. Ratadiya, "Evolution of transfer learning in natural language processing." *arXiv preprint arXiv:1910.07370*, pp. 1–11, 2019.

[26] C.-X. Qin, D. Qu, and L.-H. Zhang, "Towards end-to-end speech recognition with transfer learning," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, pp. 1–9, 2018.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Las Vegas, NV, USA, 2016, pp. 770–778.

[28] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *Proc. ASRU*, 2013, pp. 309–314.

[29] A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales, "Data augmentation for low resource languages," in *Proc. INTERSPEECH*, Singapore, 2014, pp. 810–814.

[30] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, San Diego, CA, USA, 2015, pp. 1–15.

[31] T. Koike, K. Qian, B. W. Schuller, and Y. Yamamoto, "deepSELF: An open source deep self end-to-end learning framework," *arXiv preprint arXiv:2005.06993v1*, pp. 1–4, 2020. [Online]. Available: https://github.com/Tomoya-K-0504/deepSELF

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, Miami, FL, USA, 2009, pp. 248–255.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, Montréal, Canada, 2014, pp. 2672–2680.

[34] Z. Zhang, J. Han, K. Qian, C. Janott, Y. Guo, and B. W. Schuller, "Snore-GANs: Improving automatic snore sound classification with synthesized data," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 300–310, 2020.

[35] K. Qian, C. Janott, V. Pandit, Z. Zhang, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. W. Schuller, "Classification of the excitation location of snore sounds in the upper airway by acoustic multi-feature analysis," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1731–1741, 2017.

[36] K. Qian, C. Janott, M. Schmitt, Z. Zhang, C. Heiser, W. Hemmert, Y. Yamamoto, and B. W. Schuller, "Can machine learning assist locating the excitation of snore sound? a review," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–14, 2020, in press.