Identifying Surgical-Mask Speech Using Deep Neural Networks on Low-Level Aggregation

Xinzhou Xu[1], Jun Deng[2], Zixing Zhang[3], Chen Wu[1], and Björn Schuller[3,4]

[1]School of Internet of Things, Nanjing University of Posts and Telecommunications, P.R. China
[2]Agile Robots AG, Germany
[3]GLAM–Group on Language Audio and Music, Imperial College London, UK
[4]Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
xinzhou.xu@njupt.edu.cn

ABSTRACT

The task of Mask-Speech Identification (MSI) aims at judging whether a chunk of speech is pronounced when the speaker is wearing a facial mask or not. Most of the existing related research focuses on investigating the influence of wearing a mask, which only adapts in some certain cases to speech analysis. Thus in order to generalise the research on MSI, we propose an MSI approach using deep networks on Low-Level Aggregation (LLA) for speech chunks. The proposed approach benefits from data augmentation on Low-Level Descriptors (LLDs), resulting in more adaptation to deep models through inputting much more samples in training without employing pre-trained knowledge. Experiments are performed on the dataset of Mask Augsburg Speech Corpus (MSC) used in the INTERSPEECH 2020 ComParE challenge, considering the influence from employing different strategies. The experimental results show effectiveness of the proposed approach compared with the ComParE challenge baselines.

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**; *Life and medical sciences*; • **Human-centered computing** → *Human computer interaction (HCI)*; • **Computing methodologies** → Machine learning;

KEYWORDS

Mask-speech identification, low-level aggregation, computational paralinguistics, deep neural networks

ACM ISBN 978-1-4503-8104-8
https://doi.org/10.1145/3412841.3441938

ACM Reference Format:
Xinzhou Xu, Jun Deng, Zixing Zhang, Chen Wu, and Björn Schuller. 2021. Identifying Surgical-Mask Speech Using Deep Neural Networks on Low-Level Aggregation. In *The 36th ACM/SIGAPP Symposium on Applied Computing (SAC '21),March 22–26, 2021, Virtual Event, Republic of Korea.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3412841.3441938

1 INTRODUCTION

Computational paralinguistics make it possible to extract latent knowledge in audio signals (i.e., spoken signals) from human beings or animals [1–3]. Typical paralinguistics-related topics include emotion and personality recognition [4–6], autism diagnosis [7], native-speaker identification [8], or eating classification [9]. As an emerging topic in paralinguistics, *Mask-Speech Identification* (MSI) attempts to automatically distinguish whether a spoken utterance is pronounced by its speaker with or without a surgical mask [4]. Through taking effective measures, the research of this topic can make sense in detecting mask wearing in public areas to prevent the spread of epidemics, for example, the spreading of COVID-19 [10, 11]. In addition to the usage of surgical-mask detection, as a preprocessing step, MSI helps improve robustness when cascading modules of linguistic (e.g., speech understanding) [12, 13] or paralinguistic tasks (e.g., speaker identification and emotion recognition) [14].

Nevertheless, current research in relation to the MSI task relies on two aspects as follows. First, most existing works focus on exploring the influence to analysing speech when the speaker wears a mask [12, 14]. This leads to designing processing strategies only for specific conditions on audio signals. Furthermore, existing works achieved their best recognition performance through transferring information from pre-trained models [4, 15, 16], requiring additional computational and storage cost for local processors, in the case of not choosing to pass all the raw features or digital signals to remote processing units.

In view of the existing research, we propose an MSI approach using deep neural networks on *Low-Level Aggregation* (LLA) for each utterance, without employing external information, e.g., from image datasets. The proposed LLA approach includes three steps: low-level-descriptor extraction, deep-network training, and decision on aggregation. We

utilise the *Low-Level Descriptors* (LLDs) [4, 7, 17] to obtain low-level features without knowing any pre-trained framework, in order to reduce the workload of local processors when providing limited bandwidth for transmission. Note that our strategy takes a much less complicated pipeline for feature extraction compared with Ref. [18–20]. The contributions of this paper can be summarised as making use of low-level aggregation to achieve data augmentation and avoiding using external pre-trained models.

The remainder of this paper is organised as follows. Section 2 introduces the MSI data investigated in this paper, while Section 3 provides the methodology used for data processing. Then, experimental results and their corresponding analysis are presented in Section 4 to show the performance.

2 THE DATASET

We employ the ComParE-challenge dataset of the *Mask Augsburg Speech Corpus* (MASC) [4] in investigating the task of MSI, containing German speech chunks from 32 native speakers (16 female), with the ages ranging from 21 to 40 years. The unpaired speech chunks in the dataset cover the cases speaking with and without a surgical mask during the tasks of answering questions, reading words (mainly used in the condition of medical operation rooms), and describing pictures. The dataset contains the samples with a fixed length of one second for each chunk [4]. The unpaired chunks imply that it is unavailable to directly build relationships each pair of chunks only different on whether the speakers wear a mask or not. The audio signals have been processed with a sampling rate of 16 kHz, stored in the mono format. The training set includes 10 895 chunks (5 542 with mask), while the validation/test sets contain 7 323 samples (4 014 with mask) for the validation set, and 7 324 samples (3 967 with mask) for the test set, both from the original development set in the challenge [4].

3 METHODOLOGY

As shown in Fig. 1, the proposed approach makes use of raw speech to obtain low-level descriptors to reconstruct a new training set, instead of the original set consisting of chunk-level features or representations. Then, a deep neural network containing a one-dimensional-convolutional layer and mutiple fully-connected layers is trained on these descriptors with pre-processing. Finally, for the aggregation step, we collect the corresponding outputs to get low-level decisions and infer the predicted label for a chunk sample in the validation or test set.

3.1 Low-Level Descriptors

For the steps of processing raw speech signals, we consider the acoustic low-level descriptors provided in ComParE, including 65 LLDs and their delta descriptors, without considering their functionals. The LLDs contain the categories of loudness, energy, *Zero-Crossing Rate* (ZCR), *RelAtive Spectral TrAnsform* (RASTA) auditory band, *Mel Frequency Cepstrum Coefficient* (MFCC), spectral features, and F0-related

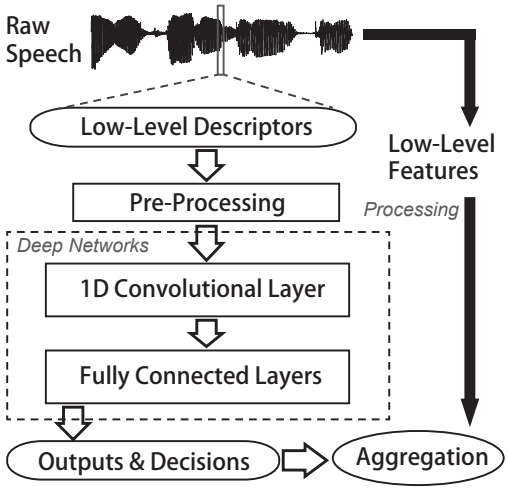

Figure 1: A diagrammatic overview of the proposed approach for mask-speech identification.

features [7, 21]. This leads to $n = 130$ original and delta LLDs in total for one low-level cell. Afterwards, we perform the pre-processing steps of shuffling and normalisation. For an arbitrary speech chunk, we note its chunk-level features as $x = \{\tilde{x}_{\cdot 1}, \tilde{x}_{\cdot 2}, \ldots, \tilde{x}_{\cdot m_x}\} \subset \Re^n$ with their low-level labels replicated from the corresponding chunk-level label y, where m_x is the number of low-level samples for the chunk x.

3.2 The Deep Network

We design the deep network including two modules of a one-dimensional convolutional layer and multiple fully connected layers, both considering dropout to reduce overfitting. The size of the convolutional layer's perceptive field is set as 1 to perform feature mapping. Then, we add a five-layer fully-connected intermediate structure of 2 048-2 048-1 024-512-256 [22, 23] and an output layer with binary outputs. All the intermediate layers employ a *Rectified Linear Unit* (ReLU) as the activation, while the output layer employs the mapping of softmax.

For the binary-class loss function, focal loss without α-balance is employed to focus on marginal samples [24], represented as

$$\mathcal{L}_{FL}(p_t) = -(1 - p_t)^{\gamma} \log(p_t), \tag{1}$$

where $p_t = p$ when the sample's label is positive, while $p_t = 1 - p$ otherwise, where p represents the network's estimated probability for positive cases. $\gamma \geq 0$ is the tunable focusing parameter. Note that we omit the α-balance variant (or equivalently $\alpha = 0.5$) in view of the balanced number of samples for each class in the training set.

3.3 Aggregation

We implicitly regard the deep neural networks approximatively as probabilistic models using the decision rule of *Maximum A Posteriori* (MAP). Thus, the deep network outputs its low-level decisions using the data mapping $f(\cdot)$ predicted

through approximately maximising the probability as

$$\hat{f} = \arg\max_{f} P(\mathcal{Y}|\{f(x_1), f(x_2), \ldots, f(x_N)\}), \qquad (2)$$

where the training chunk x_i $(i = 1, 2, \ldots, N)$ yields its deep-network mapping $f(x_i) = \{f(\tilde{x}_{i1}), f(\tilde{x}_{i2}), \ldots, f(\tilde{x}_{im_i})\}$ for low-level samples and the label set of the training samples $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$. m_i represents the number of low-level samples in the training chunk x_i.

Afterwards, we induce the operator $g(\cdot)$ for low-level aggregation, which outputs the final decision for a chunk through aggregating all the low-level predictions. Note that we perform majority voting [25] on binary low-level decisions, or average the network's softmax outputs [26] as the aggregation methods, with a threshold of 0.5. Note that the first aggregating method provides more robustness due to its binary decisions, while the softmax method makes the aggregation more accurate through using real values. Using the aggregation $g(\cdot)$, the predicted label of the arbitrary chunk sample x can be

$$\hat{y} = g(\{\hat{f}(\tilde{x}_{\cdot 1}), \hat{f}(\tilde{x}_{\cdot 2}), \ldots, \hat{f}(\tilde{x}_{\cdot m_x})\}). \qquad (3)$$

4 EXPERIMENTS

4.1 Experimental Setup

We keep the measure *Unweighted Accuracy* (UA) [27] or equivalently the unweighted average recall in the experiments, as in the ComParE challenge [4].

The operation of extracting LLDs in the experiments contains the frame size of 60 ms (for F0-related LLDs) or 20 ms (for the remaining LLDs) with a frame rate of 10 ms, using the openSMILE toolkit (version 2.3.0) [4, 28, 29]. We note the LLDs using the ComParE setup as 'LLDComParE', due to their usage in the feature set of ComParE [4, 7]. Accordingly, this procedure generates over one million low-level samples for training.

We utilise *Adaptive moment estimation* (Adam) as the optimiser in our deep structures, with the initial learning rate set to 8×10^{-6} and the maximal number of epochs as 30. The batch size is set to 1 024. The number of filters in the single one-dimensional convolutional layer is set to 512. We add dropout layers for the convolutional layer (dropout rate 0.2) and the first to the third fully connected layer (dropout rate 0.5). For the loss function, we set $\gamma = 2$ as in [24]. We observed the performance per two epochs in the experiments to save the best results, considering ten trials due to the randomisation in training.

4.2 Results and Analysis

Influence of Different Strategies:

We present experimental results for different strategies of aggregation, low-level filtering, and convolutional filters, in order to explore optimal setups of the proposed approach for the current MSI task.

First, we examine the UA performance using different aggregation types and low-level filtering strategies. The aggregation types consist of majority voting and softmax combination, while we consider filtering the low-level samples with or without a *Voice Activity Detection* (VAD) module for pre-processing. Note that the VAD used here is designed through simply setting a small threshold of 10^{-8} on the LLD of F0 via *SubHarmonic Summation* (SHS) [28, 30]. Table 1 lists the chunk-level UAs (including their average and standard deviation) of the validation and test set, for the four cases of the voting/softmax and with/without VAD. The results indicate that the best performance of UAs appears in the case of majority voting without VAD. Thus, we employ this setup in our following experiments. On the aspect of aggregation types, the better performance of the majority voting without VAD may be due to its cutoff for reducing the noise from the low-level predictions. As to the VAD pre-processing, the results of Table 1 partially imply that unvoiced segments may contain valuable information for MSI, e.g., unvoiced sound.

Further, it remains unknown on how to best choose the number of convolutional layers. We therefore carries out additional experiments with results given in Fig. 2, showing the UAs for the proposed approach employing 512, 256, and 128 filters in the convolutional layer of the deep model. Note that we utilise the majority-voting setup without VAD in view of the results in Table 1. We perform a one-way *ANalysis Of VAriance* (ANOVA) with *Scheffé*'s posthoc method [31]

Table 1: The UA (%) comparison between the cases of the proposed LLD based LLA using the aggregation types including majority voting and softmax, with and without using VAD filtering.

Approaches \ Sets	Validation Set	Test Set
LLD (Majority Voting)	**68.6** $\pm$ 0.3	**68.3** $\pm$ 0.3
LLD (Softmax)	68.4 $\pm$ 0.2	68.0 $\pm$ 0.2
LLD-VAD (Majority Voting)	64.9 $\pm$ 0.3	64.4 $\pm$ 0.2
LLD-VAD (Softmax)	65.2 $\pm$ 0.3	64.7 $\pm$ 0.4

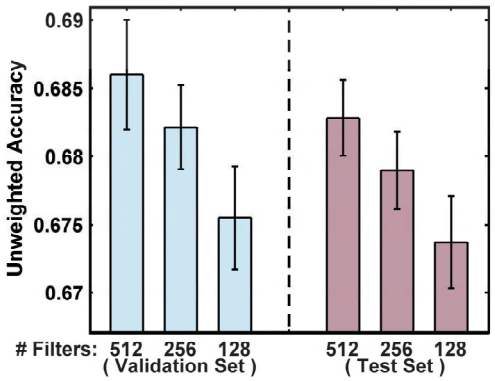

Figure 2: The UAs of the proposed approach using majority voting for its aggregation, when considering 512, 256**, and** 128 **(left to right, each) convolutional filters on the set of validation and test.**

Table 2: The chunk-level and low-level UAs (%) for the cases of using LLDComParE, MFCC, and PLP as the low-level features, including the average UAs and their standard deviations across the ten-time repetition.

Features	Validation Set	Test Set	Development Set
Chunk-Level UAs:			
LLDComParE	**68.6** $\pm$ 0.3	**68.3** $\pm$ 0.3	**68.5** $\pm$ 0.3
MFCC	67.4 $\pm$ 0.5	66.2 $\pm$ 0.6	66.8 $\pm$ 0.5
PLP	62.6 $\pm$ 0.3	62.3 $\pm$ 0.4	62.5 $\pm$ 0.3
Low-Level UAs:			
LLDComParE	**59.35** $\pm$ 0.12	**59.35** $\pm$ 0.08	**59.35** $\pm$ 0.09
MFCC	58.34 $\pm$ 0.14	58.13 $\pm$ 0.17	58.23 $\pm$ 0.16
PLP	57.39 $\pm$ 0.08	57.25 $\pm$ 0.07	57.32 $\pm$ 0.07

Table 3: The best UAs (%) in the existing literatures and the proposed LLA approaches on the development set.

Approaches	Development Set
ComParE functionals [4]	62.6
BoAW [4]	64.2
ResNet50 [4]	63.4
S2SAE [4]	64.4
SpectralNet [19]	68.2
LLA-MFCC	67.6
LLA-PLP	63.0
LLA-LLDComParE (average)	68.5
LLA-LLDComParE (maximum)	69.0
LLA-LLDComParE (fusion)	**69.1**

on the test set, which indicates significantly better results ($p < 0.05$) between the 512-filter and the other two cases.

Comparison:

We make a comparison between our proposed system and state-of-the-art approaches on the development set.

The comparison is started by presenting the results of the proposed approach using LLDs from ComParE (noted as 'LLDComParE') on the development set, along with the case of using MFCC and *Perceptual Linear Predictive* (PLP) as the low-level features, in view of the usage of MFCC in paralinguistic tasks [32] and PLP in speech synthesis [33]. The experiments include using 39-dimensional MFCCs (MFCC 0 to 12 with their deltas and accelerations) and 18-dimensional PLP, both with the frame size of 25 ms and the rate of 10 ms. We set the maximal number of epochs for the two feature sets both as 70, with the period of two. The results show that as a type of low-level features, the LLDComParE performs better compare with MFCC and PLP on both of chunk-level and low-level aspects significantly ($p < 0.01$).

We now focus on examining the best UA results for the proposed LLA approaches compared with the baseline results (ComParE functionals or LLDs with functionals, BoAW, ResNet50 or the *Deep Spectrum*, and S2SAE or *auDeep*,

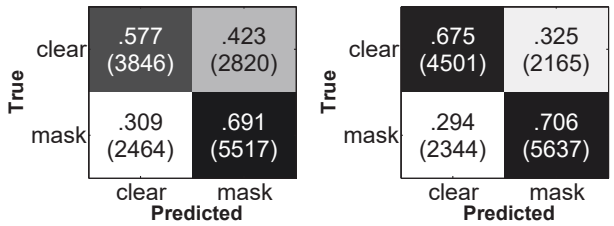

Figure 3: Confusion matrices (including recalls and the numbers) of (a) the baseline results (ResNet50) and (b) the proposed LLA-LLD with fusion.

Table 4: The estimated numbers of features and the corresponding pre-trained models for the proposed approach and the approaches in the existing references, employing pre-trained models.

References/Approaches	Pre-Trained Models	Est. # Feat.
Ref. [15]	ResNet18,34,50,101	256.5k
Ref. [20]	PANNs	100k
Ref. [4] (ResNet50)	ResNet50	8k
Raw Speech	–	16k
LLA-LLDComParE	–	6.5k

all using *Support Vector Machines* (SVMs)) [4] and the performance of the approaches without using pre-trained models [19], as shown in Table 3. The 'fusion' in the table denotes fusing the decision of the top-two best UAs for LLA-LLDComParE by simply operating 'and' on the 'mask' class. It can be learnt from the table that the proposed LLA-LLDComParE performs better compared to the results in these existing literatures.

For the purpose of making class-wise comparison, we analyse the confusion matrices for the baseline (ResNet50) and the proposed system (LLA-LLDComParE with fusion) in Figs. 3a and 3b, respectively. The results indicate that the proposed approach performs better, with the recalls of 67.5% (for 'clear') and 70.6% (for 'mask'), surpassing the baseline recalls 57.7% (for 'clear') and 69.1% (for 'mask'). This verifies the better performance of the proposed approach on both of the classes.

Finally, we consider the case of collecting a huge number of chunks on terminal devices and sending them to remote processing units, when using an extremely poor-quality communication channel. In this case, it is important to reduce the number of features. Thus, we make a comparison based on the estimated numbers of features in a one-second chunk (not truncating the low-level features at the end) between our proposed LLA-LLDComParE and typical existing works (without using pre-trained models on the terminal devices) in Table 4. It is drawn from the comparison that the proposed LLA-LLDComParE utilises a smaller number of features compared with the works in [4, 15, 20]. Note that 'Raw Speech'

represents the case of transmitting an original chunk. We tend to directly use the original signal instead of the extracted features when the dimensionality of features is extremely high. In addition, the proposed approach makes it possible to process the low-level units in accordance with low-bit-rate channels [34], in which only 130 features are required at most for each low-level cell.

5 CONCLUSION

This paper provided a novel approach for identifying mask-speech chunks using deep neural networks through aggregating on low-level descriptors. The proposed approach first employed the low-level features to generate new samples. Then, these low-level samples were input to a specifically designed deep network, outputting their low-level decisions for aggregation. We presented experimental results in order to show the performance of the proposed approach on the ComParE-challenge dataset. The results verified the better performance for the proposed approach, compared with existing research and certain strategies. Future work should focus on investigating more effective low-level features for the procedure of the low-level aggregation, including the aggregation on denser low-level sampling and pre-trained models on segments.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable help. This work was supported by the Natural Science Foundation of China under Grants No. 61801241 and No. 61801236, the Natural Science Foundation for Jiangsu Higher Education Institutions under Grants 18KJB510029, the Natural Science Foundation of Jiangsu under Grant BK20180746, and the NUPTSF under Grant NY217149.

REFERENCES

[1] Björn Schuller, Felix Weninger, Yue Zhang, Fabien Ringeval, Anton Batliner, Stefan Steidl, Florian Eyben, Erik Marchi, Alessandro Vinciarelli, Klaus Scherer, Mohamed Chetouani, and Marcello Mortillaro. Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge. *Computer Speech & Language*, 53:156–180, 2019.

[2] Fasih Haider, Sofia De La Fuente, and Saturnino Luz. An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):272–281, 2019.

[3] Björn W Schuller, Anton Batliner, Christian Bergler, Florian B Pokorny, Jarek Krajewski, Margaret Cychosz, Ralf Vollmann, Sonja-Dana Roelen, Sebastian Schnieder, Elika Bergelson, et al. The INTERSPEECH 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2378–2382, Graz, Austria, 2019.

[4] Björn Schuller, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, Harald Baumeister, Alexis Deighton MacIntyre, and Simone Hantke. The INTERSPEECH 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2042–2046, Shanghai, China, 2020. ISCA.

[5] Guozhen An, Sarah Ita Levitan, Rivka Levitan, Andrew Rosenberg, Michelle Levine, and Julia Hirschberg. Automatically classifying self-rated personality scores from speech. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1412–1416, San Francisco, CA, 2016.

[6] Xinzhou Xu, Jun Deng, Nicholas Cummins, Zixing Zhang, Li Zhao, and Björn Schuller. Autonomous emotion learning in speech: A view of zero-shot speech emotion recognition. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 949–953, Graz, Austria, 2019. ISCA.

[7] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 148–152, Lyon, France, 2013. ISCA.

[8] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity and native language. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2001–2005, San Francisco, CA, 2016. ISCA.

[9] Björn W. Schuller, Stefan Steidl, Anton Batliner, Simone Hantke, Florian Hönig, Juan Rafael Orozco-Arroyave, Elmar Nöth, Yue Zhang, and Felix Weninger. The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 478–482, Dresden, Germany, 2015. ISCA.

[10] Björn W. Schuller, Dagmar M. Schuller, Kun Qian, Juan Liu, Huaiyuan Zheng, and Xiao 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*, 2020.

[11] Trisha Greenhalgh, Manuel B Schmid, Thomas Czypionka, Dirk Bassler, and Laurence Gruer. Face masks for the public during the COVID-19 crisis. *British Medical Journal*, 369:m1435, 2020.

[12] Samuel R Atcherson, Lisa Lucks Mendel, Wesley J Baltimore, Chhayakanta Patro, Sungmin Lee, Monique Pousson, and M Joshua Spann. The effect of conventional and transparent surgical masks on speech understanding in individuals with and without hearing loss. *Journal of the American Academy of Audiology*, 28(1):58–67, 2017.

[13] Lisa Lucks Mendel, Julie A Gardino, and Samuel R Atcherson. Speech understanding using surgical masks: A problem in health care? *Journal of the American Academy of Audiology*, 19(9):686–695, 2008.

[14] Rahim Saeidi, Ilkka Huhtakallio, and Paavo Alku. Analysis of face mask effect on speaker recognition. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1800–1804. ISCA, 2016.

[15] Nicolae-Cătălin Ristea and Radu Tudor Ionescu. Are you wearing a mask? Improving mask detection from speech using augmentation by cycle-consistent GANs. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2102–2106, Shanghai, China, 2020. ISCA.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Proc. European Conference on Computer Vision (ECCV)*, pages 630–645, Amsterdam, The Netherlands, 2016. Springer.

[17] Juliette Millet and Neil Zeghidour. Learning to detect dysarthria from raw speech. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5831–5835, Brighton, UK, 2019. IEEE.

[18] Tamás Grósz, Mittul Singh, Sudarsana Reddy Kadiri, Hemant Kathania, and Mikko Kurimo. Aalto's end-to-end DNN systems for the INTERSPEECH 2020 computational paralinguistics challenge. *arXiv preprint arXiv:2008.02689*, 2020.

[19] Steffen Illium, Robert Müller, Andreas Sedlmeier, and Claudia Linnhoff-Popien. Surgical mask detection with convolutional neural networks and data augmentations on spectrograms. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2052–2056, Shanghai, China, 2020. ISCA.

[20] Koike Tomoya, Kun Qian, Björn Schuller, and Yoshiharu Yamamoto. Learning higher representations from pre-trained deep models with data augmentation for the ComParE 2020 challenge mask task. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2047–2051, Shanghai, China, 2020. ISCA.

[21] Nicholas Cummins, Maximilian Schmitt, Shahin Amiriparian, Jarek Krajewski, and Björn Schuller. "You sound ill, take the day off": Automatic recognition of speech affected by upper respiratory tract infection. In *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3806–3809, Jeju Island, South Korea, 2017. IEEE.

[22] Daniel Garcia-Romero, Alan McCree, David Snyder, and Gregory Sell. JHU-HLTCOE system for the VoxSRC speaker recognition challenge. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7559–7563, Barcelona, Spain, 2020. IEEE.

[23] Aaron S Coyner, Ryan Swan, James M Brown, Jayashree Kalpathy-Cramer, Sang Jin Kim, J Peter Campbell, Karyn E Jonas, Susan Ostmo, RV Paul Chan, and Michael F Chiang. Deep learning for image quality assessment of fundus images in retinopathy of prematurity. In *Proc. American Medical Informatics Association Annual Symposium (AMIA)*, volume 2018, pages 1224–1232, San Francisco, CA, 2018. American Medical Informatics Association.

[24] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, Venice, Italy, 2017. IEEE.

[25] Stavros Petridis and Maja Pantic. Deep complementary bottleneck features for visual speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2304–2308, Shanghai, China, 2016. IEEE.

[26] Huy Phan, Lars Hertel, Marco Maass, and Alfred Mertins. Robust audio event recognition with 1-max pooling convolutional neural networks. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3653–3657, San Francisco, CA, 2016. ISCA.

[27] Xinzhou Xu, Jun Deng, Eduardo Coutinho, Chen Wu, Li Zhao, and Björn Schuller. Connecting subspace learning and extreme learning machine in speech emotion recognition. *IEEE Transactions on Multimedia*, 21(3):795–808, 2019.

[28] Florian Eyben and Björn Schuller. openSMILE:) The Munich open-source large-scale multimedia feature extractor. *ACM SIGMultimedia Records*, 6(4):4–13, 2015.

[29] Florian Eyben, Klaus R Scherer, Björn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

[30] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proc. International Conference on Multimedia*, pages 835–838, Barcelona, Catalunya, Spain, 2013. ACM.

[31] Gordon P Brooks and George A Johanson. Sample size considerations for multiple comparison procedures in ANOVA. *Journal of Modern Applied Statistical Methods*, 10(1):97–100, 2011.

[32] Gil Keren, Jun Deng, Jouni Pohjalainen, and Björn W Schuller. Convolutional neural networks with data augmentation for classifying speakers' native language. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2393–2397, San Francisco, CA, 2016. ISCA.

[33] Nirmesh J Shah, Bhavik B Vachhani, Hardik B Sailor, and Hemant A Patil. Effectiveness of PLP-based phonetic segmentation for speech synthesis. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 270–274, Florence, Italy, 2014. IEEE.

[34] Milos Cernak, Alexandros Lazaridis, Afsaneh Asaei, and Philip N Garner. Composition of deep and spiking neural networks for very low bit rate speech coding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(12):2301–2312, 2016.