Interspeech 2022
18-22 September 2022, Incheon, Korea

Distinguishing between pre- and post-treatment in the speech of patients with chronic obstructive pulmonary disease

Andreas Triantafyllopoulos[1], *Markus Fendler*[3], *Anton Batliner*[1], *Maurice Gerczuk*[1], *Shahin Amiriparian*[1], *Thomas M. Berghaus*[3,4], *and Björn W. Schuller*[1,2]

[1]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]GLAM – Group on Language, Audio, & Music, Imperial College, UK
[3]Department of Cardiology, Respiratory Medicine and Intensive Care, University Hospital Augsburg, University of Augsburg, Germany
[4]Ludwig-Maximilians-University Munich, Germany
andreas.triantafyllopoulos@uni-a.de

Abstract

Chronic obstructive pulmonary disease (COPD) causes lung inflammation and airflow blockage leading to a variety of respiratory symptoms; it is also a leading cause of death and affects millions of individuals around the world. Patients often require treatment and hospitalisation, while no cure is currently available. As COPD predominantly affects the respiratory system, speech and non-linguistic vocalisations present a major avenue for measuring the effect of treatment. In this work, we present results on a new COPD dataset of 20 patients, showing that, by employing personalisation through speaker-level feature normalisation, we can distinguish between pre- and post-treatment speech with an unweighted average recall (UAR) of up to 82 % in (nested) leave-one-speaker-out cross-validation. We further identify the most important features and link them to pathological voice properties, thus enabling an auditory interpretation of treatment effects. Monitoring tools based on such approaches may help objectivise the clinical status of COPD patients and facilitate personalised treatment plans.

Index Terms: digital health, pathological speech, COPD, personalisation, feature interpretation

1. Introduction and related work

Chronic obstructive pulmonary disease (COPD) is a respiratory disease characterised by a chronically inflamed and obstructed airway. The long-term exposure to damaging irritants, especially cigarette smoke, is the predominant cause of COPD [1] – the third leading cause of death after ischemic heart disease and stroke [2]. Its prevalence in Europe ranges between 15 and 20 % for adults aged over 40 years [3]. Thus, the socio-economic burden of COPD is immense and results in a high consumption of clinical resources and overall costs to society [4]. While the clinical appearance of COPD varies, the main symptoms are chronic and progressive dyspnea, as well as coughing. Sometimes, patients suffer from a severe deterioration of their respiratory symptoms, called exacerbation, which leads to a need of intensified therapy, often under in-patient or even intensive care treatment. Besides treatment and prevention of exacerbation, the main objective of COPD therapy is to provide symptomatic relief. For this purpose, patients receive (mainly inhalative) medication, such as inhaled corticosteroids (ICS), long-acting muscarinic antagonists (LAMA) or long-acting beta-2-agonists (LABA) – often administered in combination. Sometimes, further systemic medication, such as systemic corticosteroids, is needed. The intensity of treatment depends on the severity of COPD, which is classified according to the recommendations of the Global Initiative of Chronic Obstructive Lung Disease. This classification is based on the severity of airflow limitations and takes into account symptoms and risks of exacerbation [5]. Spirometric examinations and clinical findings are still the relevant diagnostic tools for COPD. Especially for acute exacerbation, clinical practitioners must rely on clinical examination to monitor treatment success or failure. So far, there are no technical tools to objectivise clinical symptoms of COPD, especially during an exacerbation episode. In order to accelerate diagnosis and distinguish different respiratory illnesses, new artificial intelligence (AI)-assisted monitoring tools can be helpful. For instance, new data shows the potential use of AI speech analysis for respiratory diseases, such as COVID-19 [6]. It is intuitive to expand these findings to COPD, since this obstructive airway disease inevitably affects speech and non-linguistic vocalisations, especially during exacerbation.

Due to the widespread prevalence of COPD and its negative effect on public health, voice-based digital detection and monitoring tools have recently attracted increased attention [7, 8, 9, 10, 11, 12, 13]. These utilise different types of vocalisations, such as breathing [10], coughing [8], sustained vowels [11], or read/free speech [7, 11, 12] to distinguish between COPD patients and healthy individuals and different states of COPD (such as 'stable' vs exacerbation [11]). Yet, most of these studies are merely identifying acoustic descriptors that are correlated with COPD and do not build an automatic detection tool. Instead, we focus on developing a machine learning (ML)-based voice evaluation tool that distinguishes between pre- and post-treatment states of patients after exacerbation, based on read speech. We investigate a set of different feature sets, segmentation strategies, and normalisation procedures. It turns out that speaker-level feature normalisation is crucial for obtaining good performance – demonstrating that personalisation is key for this application. This paves the way for more advanced personalisation techniques which adapt to specific speakers using either speaker-dependent models [14] or test-time adaptation [15]. Furthermore, we try to interpret the most important features, which helps characterise the effect of exacerbation and the corresponding benefits of treatment on the speakers' voice.

The remainder of our contribution is organised as follows: Section 2 describes the dataset used in this study. Section 3 outlines our experimental protocol, followed by our results and accompanying discussion and interpretation in Section 4. The work ends with some concluding remarks in Section 5.

10.21437/Interspeech.2022-10333

2. Dataset

Our COPD dataset was recorded at the University Hospital Augsburg between October 2020 and December 2021. Patients were recruited soon after hospitalisation, if possible, already in the emergency department or the intensive care unit. Only patients older than 18 years, able to sit and read a short text without any need of ventilation during the time of recording, were included, resulting in: 20 (11 male / 9 female) subjects, aged 48 to 82 (median: 70), with a median of 37.5 pack years[1] and an improvement (post- to pre-treatment) between 0 and 7 (median: 2.39) on a modified Borg scale, a subjective assessment of dyspnea from 0 (worst) to 10 (best).

The local ethics committee approved the study on June 24, 2020 (BKF 2020-34). Two recordings – one pre-treatment, one post-treatment – with a distance of 1 to 16 days (median: 5) took place bedside with the portable recorder H5 from Zoom® and a Sony lapel mic ECM-144. All patients obtained their standard and individual home medication, mainly LABA and LAMA, sometimes in addition to ICS and systemic corticosteroids. Some patients needed non-invasive ventilation, which was paused for the time of recording.

Patients were required to produce: a) a set of sustained vowels (*/a:/*, */e:/*, */i:/*, */o:/*, */u:/*); b) a few spontaneous utterances; c) (forced) coughing; d) breathing; e) reading *Der Nordwind und die Sonne* [*The Northwind and the Sun*] (NuS). We assume no or a negligible habituation effect, as the second recording took place days after the first one. Here, we deal only with NuS: it is longer than other sound types (median: 53 s), thus, resulting in more samples after segmentation, which is beneficial for ML algorithms. This was confirmed by preliminary experiments with the other sounds.

3. Experimental setup

SEGMENTATION: the North Wind and the Sun [*Der Nordwind und die Sonne*] (NuS) is too long as a unit of analysis for most speech processing applications, which typically operate on shorter segments. For a segmentation yielding shorter units, we use the Munich AUtomatic Segmentation (MAUS) system [16, 17]. It utilises forced alignment to derive word and phone boundaries from the text transcriptions, which in the case of read speech is trivially available. We experiment with two different types of segmentation resulting in two different units of analysis: **Word units**, where we keep the original word boundaries returned by MAUS, resulting in a total of $18 \times 2 \times 188 = 3888$ segments; and **Phrase units**, where we segmented the story into 20 prosodic phrases for a total of $18 \times 2 \times 20 = 720$ segments.

FEATURES: We extract a set of acoustic descriptors per unit (for both words and phrases) that can be used to distinguish between pre- and post-treatment COPD speech, employing both expert, handcrafted features, and learnt representations of deep neural networks, thus contrasting the two dominant ongoing trends in speech processing applications:

eGeMAPS: The extended Geneva minimalistic acoustic parameter set (eGeMAPS) [18] is a small set of (88) interpretable acoustic parameters that has previously been shown to contain relevant information for respiratory diseases, such as COVID-19 [19]. eGeMAPS is extracted using openSMILE [20].

ComParE: The Interspeech Computational Paralinguistics ChallengE feature set (ComParE) is a large-scale feature set (6373) that has been successfully used for several computational paralinguistics tasks, beginning with the 2013 Interspeech Computational Paralinguistics Challenge [21], also extracted using openSMILE [20].

w2v2-xlsr: Substantial progress has been seen through the use of models trained on vast amounts of data with self-supervised methods. We use a variant of WAV2VEC2.0 [22], pre-trained on 53 languages – including German [23]. The model operates on raw audio and returns contextualised representations roughly corresponding to 25 ms of audio with a stride of 20 ms, which we subsequently average over the time dimension to obtain the final 1024-dimensional embeddings.

NORMALISATION: We experiment with three different normalisation procedures. In all cases, z-score normalisation on each feature is performed separately; what changes is the set over which we compute and apply statistics.

Global: As a standard baseline, we normalise the data on a global basis – for each fold in our cross-validation setup, we compute feature statistics on the training set, and subsequently use those to normalise the development and testing partitions. This is the prevailing type of normalisation.

Word/Phrase-level: We experiment with a normalisation procedure targeted at the respective *unit of analysis*. Using read speech recordings, we collect identical audio content for each speaker. This content, however, is influenced by non COPD-related factors, which can be abstracted away by normalising each word or phrase unit independently (using data from all speakers).

Speaker-level: Given our expectation that there are individual differences in the manifestation of COPD in human vocalisations, we employ a speaker-level normalisation procedure in an attempt to abstract away from them. This is based on computing (and applying) mean and standard deviation normalisation *independently* for each speaker, that is, using all their data to compute parameters irrespective of whether they are part of the training, development, or test partition. As such, this form of normalisation assumes oracle knowledge of the identity of each speaker. We consider this a realistic assumption for *personalised* digital health applications in controlled conditions.

CLASSIFIER: We use support vector machines (SVMs) where we optimise the cost parameter ($\{.0001, .0005, .001, .005, .01, .05, .1, .5, 1\}$) and kernel function ($\{$linear, polynomial, radial basis function (RBF)$\}$) in a grid search manner. These parameters are always optimised on the development partition.

EVALUATION PROTOCOL: As the size of our dataset is limited (20 speakers), we use leave-one-speaker-out cross-validation, whereby data from every speaker is used exactly once for testing, each time using the data of all other speakers for training. For each fold, we perform *nested cross-validation* for optimising SVM parameters by further splitting the training speakers into two speaker-disjoint sets. Once the optimal set of parameters has been identified (based on development set performance), we train a final model on the entire training data for each fold.

METRICS: We use unweighted average recall (UAR), the mean of the diagonal cells in the confusion matrix in percent. This balances the sensitivity and specificity of both classes, which in our case are both equally important (i. e., we have no 'positive' and 'negative' class)[2]. We differentiate between three different

[1] Pack years are calculated by multiplying the number of packs of cigarettes smoked per day by the number of years the person has smoked.

[2] As we always have the same number of instances for both classes, UAR is identical with accuracy, as well as with (sensitivity + specificity)

Table 1: *U_{UAR} and ST_{UAR} using leave-one-speaker-out cross-validation with 95 % CIs.*

Normalisation	Features	Word units U_{UAR}[%]	Word units ST_{UAR}[%]	Phrase units U_{UAR}[%]	Phrase units ST_{UAR}[%]
Global	eGeMAPS	56 (54-57)	60 (45-75)	56 (52-59)	52 (37-68)
	ComParE	52 (51-54)	60 (44-76)	53 (49-56)	50 (34-66)
	w2v2-xlsr	54 (52-55)	45 (29-61)	59 (55-62)	65 (50-80)
Word/Phrase	eGeMAPS	56 (55-58)	60 (46-75)	59 (56-62)	57 (42-73)
	ComParE	53 (52-55)	62 (47-78)	53 (49-57)	55 (39-70)
	w2v2-xlsr	54 (52-55)	50 (34-65)	55 (51-59)	57 (42-72)
Speaker	eGeMAPS	60 (58-62)	68 (53-82)	63 (60-66)	**80 (67-92)**
	ComParE	53 (52-54)	62 (48-79)	55 (51-58)	55 (40-70)
	w2v2-xlsr	59 (58-61)	**82 (70-94)**	66 (63-70)	78 (66-90)

ways of computing UAR: unit-, story-, and speaker-level. The first (**U_{UAR}**) quantifies how well the model works over individual **units** (instances); the second (**ST_{UAR}**), how well it classifies speakers into pre- and post-treatment states after aggregating all individual predictions for each **story** (the major focus of our work); the third (**SP_{UAR}**), how well the model works for an individual **speaker** by computing performance over only their instances. Given a set of speakers $\{s_1, ..., s_S\}$ (with $S = 20$ being the total number of speakers), each producing the NuS story twice (corresponding to the two classes {(b)efore, (a)fter}), with each story segmented into N_u units (i. e., words/phrases) resulting in a total of N units (3888/720) overall and N_s units per speaker (216/40), we generate a total of N unit-level predictions $\hat{y}_i$ using the setup outlined above. We define the different evaluation protocols as follows:

$$\mathrm{U_{UAR}} = \frac{1}{2} \sum_{c \in \{a,b\}} \frac{|i \in [N] : y_i = c, \hat{y}_i = c|}{|i \in [N] : y_i = c|}$$

$$\mathrm{ST_{UAR}} = \frac{1}{2} \sum_{c \in \{a,b\}} \frac{|i \in [S] : y_i = c, \underset{j \in [N_s]}{\mathrm{maxvote}}(\hat{y}_j) = c|}{|i \in [S] : y_i = c|}$$

$$\mathrm{SP_{UAR}} = \frac{1}{2} \sum_{c \in \{a,b\}} \frac{|i \in [N_s] : y_i = c, \hat{y}_i = c|}{|i \in [N_s] : y_i = c|}$$

4. Results and discussion

Overall results are presented in Table 1 with $\mathrm{U_{UAR}}$ and $\mathrm{ST_{UAR}}$ computed over all instances and corresponding 95 % CIs computed over 1000 bootstrap samples; $\mathrm{ST_{UAR}}$ shows a bigger range over $\mathrm{U_{UAR}}$, because it is computed with far less samples (36 vs 3888/720). Story-level performance is highest for w2v2-xlsr and word-level segmentation with a UAR of 82 % (CI: 70 %-94 %) followed by eGeMAPS and phrase-level segmentation ($\mathrm{ST_{UAR}}$: 80 %; CI: 67 %-92 %) – both using speaker-level normalisation, with phrase-level w2v2-xlsr features trailing close behind ($\mathrm{ST_{UAR}}$: 78 %; CI: 66 %-90 %). The best result without subject-level normalisation is obtained with ComParE with word-level segmentation and normalisation with a $\mathrm{ST_{UAR}}$ of 62 % (CI: 47 %-78 %) – a large drop over subject-level normalisation which showcases the need for personalisation.

Interpretability is a necessary requirement for digital health applications in order to explain the outcomes to patients and medical practitioners. As eGeMAPS yields near-top performance, with the features also being easily interpretable due to their expert-designed nature, we focus our subsequent analysis on this setting. The 95 % CI for males (52 %-75 %) showed substantial overlap to that of females (53 %-72 %), indicating that the classifier performs approximately equal for both genders. We further analysed the performance w. r. t. the available subject metadata. We first divided speakers into two subsets: those whose individual performance exceeds the $\mathrm{U_{UAR}}$ performance of 63 % (the phrase-level $\mathrm{U_{UAR}}$ for eGeMAPS when using instances from all speakers, see Table 1), and those whose performance falls below that threshold, and subsequently compared the 95 % CIs of the different metadata for those two speaker groups. This comparison revealed that models work better for subjects which have a higher post-to-pre-treatment difference in the BORG scale ([2.12-4.75] vs [1.92-3.16]), are of a higher age ([63-73] vs [58-67] years), and have smoked more pack years ([39-75] vs [28-44]). All these factors might contribute to a worse clinical condition with subjects subsequently gaining more from treatment, thus, accordingly leading to bigger changes in their voice characteristics and making it easier to distinguish between their pre- and post-treatment states.

We further analyse the features that have the largest impact on classifier decisions in order to characterise the impact of treatment on patients' voices. To that end, SHAP (SHapley Additive exPlanations) [24] has emerged as a powerful tool for extracting feature importance values for individual predictions. SHAP is based on Shapley values, whose theoretical definition for each feature relies on building surrogate models on all potential feature subsets, and taking the expectation of model output differences for all subsets including the target features vs the same subsets but excluding that feature [24]. These values can then be aggregated over an entire (test) dataset to derive global feature importance values that can be used to interpret model behaviour. We focus on the ten most important features, defined by their average SHAP values. We computed the mean of each feature separately (normalised using subject-level normalisation) for each subject and phrase for pre- and post-treatment. We then plotted the resulting $18 \times 2 = 36$ points relative to the $\mathrm{SP_{UAR}}$ corresponding to each speaker. This allows us to compare how these individual features change after treatment, but also to relate this change to speakers for which the prediction fails. To provide a better understanding of the effectiveness of each individual feature, we additionally used each of them in isolation to train a (new) model with the same experimental setup discussed before. Note that this might result in a different 'measure' of feature performance as features behave differently in isolation vs in the presence of other (potentially correlated) features [25]. In Figure 1, we show the $\mathrm{ST_{UAR}}$ (and 95 % CIs) obtained for each of them.

Due to space limitations, we only include 6 of those: As four of the original ten were merely functionals of pitch (mean, median, 20$^{\text{th}}$ and 80$^{\text{th}}$ percentiles – measured in semitones), and all of them showed a similar trend, we only provide the best performing one, the mean. Moreover, we exclude the worst-performing feature, the **3$^{\text{rd}}$ MFCC** ($\mathrm{ST_{UAR}}$: 60 %), showing slightly lower values for low performing subjects.[3] Figure 1 shows the remaining six features.

The **bandwidth of the 2$^{\text{nd}}$ formant F2** ($\mathrm{ST_{UAR}}$: 63 %) and the **bandwidth of the 3$^{\text{rd}}$ formant F3** ($\mathrm{ST_{UAR}}$, 68 %), above left and second left in Figure 1, are computed from the roots of the Linear Predictor (LP) coefficient polynomial. Dysphonic speakers display a broader formant bandwidth [27] meaning higher formant dispersion and mutual masking of neighbouring formants and by that, vowels [28].

/ 2, and differs from $\mathrm{F_1}$ score throughout just by ± 1 percent.

[3] We know from other types of atypical speech [26] that the vowel space is centralised, due to a less tense and less precise articulation. This might be the case here as well.

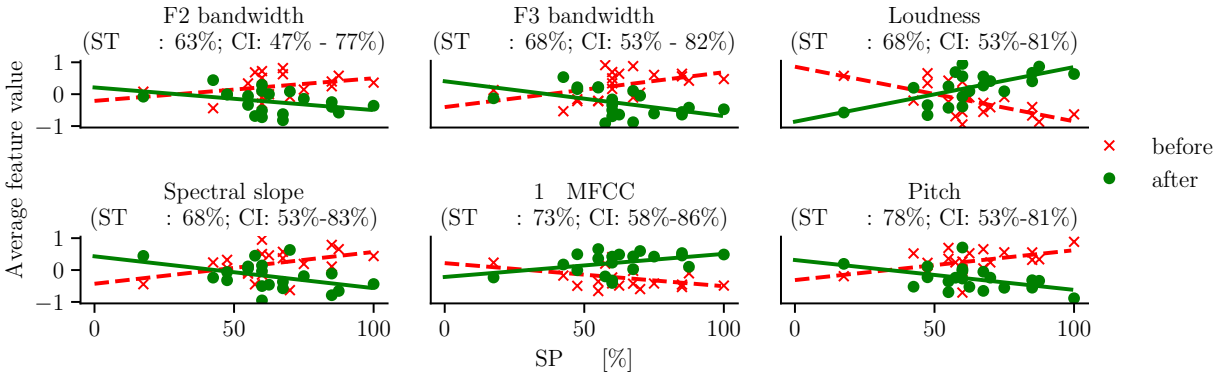

Figure 1: *Average normalised feature values (taken over all utterances of a speaker) vs SP_{UAR} for 6 out of 10 most important features as computed by SHAP values; values before (red, dashed, cross) and after (green, continuous, point) treatment; identical position per speaker on the x-axis across all plots, with crosses and points sharing the same x-coordinate corresponding to the same speaker; thus, the rightmost points on the x-axis show the speaker normalised mean feature values corresponding to the best performing speaker with 98 % (39/40) of their phrases classified correctly with eGeMAPS and speaker-level normalisation. Subtitles show ST_{UAR} (total and 95 % CI) when using single features for classification (see text for discussion). Linear regression lines fitted to better highlight trends.*

Loudness (ST_{UAR}: 68 %, above right) is partially controlled by transglottal airflow [29], which is constricted by COPD [5]; thus, patients before treatment produce on average utterances of lower loudness; this can be seen when we compare the declining line for before with the rising line for after in Figure 1.[4]

The **spectral slope** (ST_{UAR}: 68 %), below left in Figure 1, is computed by fitting an OLS estimator to the logarithmic power spectrum and is thus steeper when the higher frequencies have more energy than the lower ones. This is opposite to the **1st MFCC** (ST_{UAR}: 73 %), below middle, that can be interpreted as the inverse spectral slope, as it is a weighted ratio of the lower to the higher frequencies. These features show opposing trends, with the 1st MFCC increasing and the slope decreasing after treatment – indicating that the ratio of higher to lower frequencies decreases after treatment. A higher ratio of higher to lower frequencies can be interpreted as higher breathiness [30], which is then reduced through treatment. Note that higher breathiness might go together with decreased loudness [31].

Pitch (measured in semitones, ST_{UAR}: 78 %) is the most effective *power feature* [25]. It shows a strong lowering trend after treatment – in contrast to Merkus *et al.* [11], who found that F0 increases for subjects with stable COPD compared to those with exacerbation (mean: 190 Hz vs 154 Hz). A potential confounder is that irregular phonation caused by exacerbated COPD can lead to more errors in pitch estimation by missing some voiced segments, especially in laryngealised (creaky) parts [32, 33]. By default, openSMILE excludes segments with a value of 0 in its calculation of the mean; thus, the lower pitch values post-treatment compared to pre-treatment (mean: 150 Hz vs 160 Hz) might be due to less irregular phonation. When including 0-valued segments in the calculation of the mean, we obtain higher values of F0 post-treatment (mean: 89 Hz vs 85 Hz), similar to Merkus *et al.* [11]. We further investigated this hypothesis by comparing the average voiced segments per second before (mean: 24 %) and after treatment (mean: 28 %); indeed the proportion of segments detected as voiced increases. As the read text is identical in both conditions, (unaccounted) pitch estimation errors remain a plausible explanation for the difference between our findings and those of Merkus *et al.* [11]. Yet, there might be a 'cocktail' of intervening factors: especially before treatment, patients are more unsettled and stressed, thus both speech pathology and psychological state might result in strained voice and higher pitch, and at the same time, in irregular voice partly (mis-) recognised as unvoiced. In the post-stage, speech pathology is weakened and at the same time, the patients are relieved and more relaxed, and all this might result in a less strained voice and lower pitch.

Summing up our interpretation: Subjects after treatment show improved articulatory precision (as shown by the 3rd MFCC and F2/F3 bandwidth), decreased airflow blockage (as shown by the increase in loudness), decreased breathiness (as shown by the spectral slope and 1st MFCC), and more regular phonation (thus less pitch errors) and, by that, lowered 'regular' pitch – findings which are consistent with the expected decrease in symptomatology; as for a comparable voice quality spectrum for Parkinson's disease, see [34]. Naturally, although our present interpretation is consistent with previous phonetic research and with medical expectations, it should be evaluated on a larger sample size and tested against human evaluations.

5. Conclusion

We demonstrated that read passages can be successfully utilised to distinguish between pre- and post-treatment states of COPD patients. Using a variety of handcrafted and learnt features, we were able to achieve a top UAR of 82 % (95 % CI: 70 %-94 %) in a leave-one-speaker-out setup. Speaker-level normalisation proved to be crucial, as it removes speaker-related effects, which prevents generalisation; without it, performance reached a maximum of 62 %. Future work could be directed to more data efficient personalisation techniques, which do not require patient data to normalise with, as well as to detecting COPD in the presence of other respiratory diseases, such as COVID-19.

6. Acknowledgements

This work has received funding from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS) and from the EU's Horizon 2020 grant agreement No. 826506 (sustAGE).

[4]Note that we used a lapel microphone that prevents varying distances within the same recording session; yet, there might be slight differences across sessions which can constitute an intervening factor that cannot be fully controlled.

7. References

[1] T. Yoshida and R. Tuder, "Pathobiology of cigarette smoke-induced chronic obstructive pulmonary disease," *Physiol Rev.*, vol. 87, pp. 1047–1082, 2007.

[2] WHO, *The top 10 causes of death*, Published by World Health Organization (WHO) 2020 Dec 9, accessed 28 June 2021, 2020.

[3] C. Rycroft, A. Heyes, L. Lanza, and K. Becker, "Epidemiology of chronic obstructive pulmonary disease: a literature review," *Int J Chron Obstruct Pulmon Dis*, vol. 7, pp. 457–494, 2012.

[4] A. Guarascio, S. Ray, C. Finch, and T. Self, "The clinical and economic burden of chronic obstructive pulmonary disease in the USA," *Clinicoecon Outcomes Res*, vol. 5, pp. 235–245, 2013.

[5] D. Singh, A. Agusti, A. Anzueto, P. Barnes, J. Bourbeau, B. Celli, G. Criner, P. Frith, D. Halpin, M. Han, *et al.*, "Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease: the GOLD science committee report 2019," *Eur Respir J*, vol. 53, p. 1 900 164, 2019.

[6] 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," *CoRR*, vol. abs/2003.11117, 2020.

[7] E. E. Mohamed and R. A. E. maghraby, "Voice changes in patients with chronic obstructive pulmonary disease," *Egyptian Journal of Chest Diseases and Tuberculosis*, vol. 63, no. 3, pp. 561–567, 2014.

[8] M. G. Crooks, A. Den Brinker, Y. Hayman, J. D. Williamson, A. Innes, C. E. Wright, P. Hill, and A. H. Morice, "Continuous cough monitoring using ambient sound recording during convalescence from a copd exacerbation," *Lung*, vol. 195, no. 3, pp. 289–294, 2017.

[9] V. Nathan, K. Vatanparvar, M. M. Rahman, E. Nemati, and J. Kuang, "Assessment of chronic pulmonary disease patients using biomarkers from natural speech recorded by mobile devices," in *16th Int. Conf. on Wearable and Implantable Body Sensor Networks (BSN)*, Chicago, IL, USA, 2019, pp. 1–4.

[10] O. Ashraf, E. Rabold, K. Schlichtkrull, A. Singh, S. Venneti, M. M. D. A. Khan, R. Kulshreshtha, and P. P. Naval, "Voice-based screening and monitoring of chronic respiratory conditions," *Chest*, vol. 158, no. 4, A1687, 2020.

[11] J. Merkus, F. Hubers, C. Cucchiarini, and H. Strik, "Digital eavesdropper – acoustic speech characteristics as markers of exacerbations in copd patients," in *Proc. RaPID workshop of the 12th LREC*, Marseille, France, 2020, p. 78.

[12] D. Cleres, F. Rassouli, M. Brutsche, T. Kowatsch, and F. Barata, "Lena: A voice-based conversational agent for remote patient monitoring in chronic obstructive pulmonary disease," in *Proc. ACM Conference on Intelligent User Interfaces*, 2021.

[13] M. Farrús, J. Codina-Filbà, E. Reixach, E. Andrés, M. Sans, N. Garcia, and J. Vilaseca, "Speech-based support system to supervise chronic obstructive pulmonary disease patient status," *Applied Sciences*, vol. 11, no. 17, p. 7999, 2021.

[14] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science Robotics*, vol. 3, no. 19, pp. 1–11, 2018.

[15] A. Triantafyllopoulos, S. Liu, and B. W. Schuller, "Deep speaker conditioning for speech emotion recognition," in *Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, Shenzhen, China: IEEE, 2021, pp. 1–6.

[16] F. Schiel, "Automatic Phonetic Transcription of Non-Prompted Speech," in *Proc. ICPhS14*, San Francisco, 1999, pp. 607–610.

[17] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, Virtual Special Issues, vol. 45, pp. 326–347, Sep. 2017, ISSN: 0885-2308. DOI: 10.1016/j.csl.2017.01.005.

[18] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[19] K. D. Bartl-Pokorny, F. B. Pokorny, A. Batliner, S. Amiriparian, A. Semertzidou, F. Eyben, E. Kramer, F. Schmidt, R. Schönweiler, M. Wehler, *et al.*, "The voice of covid-19: Acoustic correlates of infection in sustained vowels," *JASA*, vol. 149, no. 6, pp. 4377–4383, 2021.

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

[21] 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 *Proc. Interspeech, Lyon, France*, 2013.

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[23] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech*, 2021, pp. 2426–2430.

[24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, pp. 4768–4777, 2017.

[25] A. Batliner and B. Möbius, "Prosody in Automatic Speech Processing," in *The Oxford Handbook of Language Prosody*, C. Gussenhoven and A. Chen, Eds., Oxford, UK: Oxford University Press, 2020, pp. 633–645.

[26] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, "Automatic modelling of depressed speech: relevant features and relevance of gender," in *Proc. Interspeech 2014*, Singapore, 2014, pp. 1248–1252.

[27] K. Ishikawa and J. Webster, "The Formant Bandwidth as a Measure of Vowel Intelligibility in Dysphonic Speech," *J Voice*, vol. 20, 2020, Published online: October 31, 2020.

[28] A. De Cheveigné, "Formant bandwidth affects the identification of competing vowels," in *Proc. ICPhS14*, San Francisco, CA, USA, 1999, pp. 2093–2096.

[29] K. Baker, L. Ramig, and S. Sapir, "Control of vocal loudness in young and old adults.," *Journal of Speech, Language & Hearing Research*, vol. 44, no. 2, pp. 297–304, 2001.

[30] J. M. Hillenbrand and R. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of Speech and Hearing Research*, vol. 39, pp. 311–321, 1996.

[31] M. Södersten and P. A. Lindestad, "Glottal closure and perceived breathiness during phonation in normally speaking subjects," *J Speech Hear Res.*, vol. 33, pp. 601–611, 1990.

[32] A. Batliner, S. Burger, B. Johne, and A. Kießling, "MÜSLI: A Classification Scheme For Laryngealizations," in *Proc. ESCA Workshop on Prosody*, Lund, 1993, pp. 176–179.

[33] P. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice," in *Proc. ICPhS18*, Glasgow, Scotland, 2015, pp. 1–5.

[34] M. Cernak, J. Orozco-Arroyave, F. Rudzicz, H. Christensen, J. Vásquez-Correa, and E. Nöth, "Characterisation of voice quality of Parkinson's disease using differential phonological posterior features," *Computer Speech & Language*, vol. 46, pp. 196–208, 2017.