2023 IEEE 17th International Symposium on Applied Computational Intelligence and Informatics (SACI) | 979-8-3503-2110-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/SACI58269.2023.10158671

Using Custom X-vectors for the Automatic Screening of COVID-19 Based on Coughing Audio Samples

José Vicente Egas-López
ELRN-SZTE
Research Group on Artificial Intelligence
Szeged, Hungary
egasj@inf.u-szeged.hu

Gábor Gosztolya
Institute of Informatics
University of Szeged
Szeged, Hungary
ELRN-SZTE
Research Group on Artificial Intelligence
Szeged, Hungary
ggabor@inf.u-szeged.hu

***Abstract*—A lot of effort has gone into eradicating the pandemic caused by the COVID-19 outbreak. One initiative in the efficient control of the spread of it lies in the methods for its diagnosis. Numerous techniques for screening the disease have emerged to date, which, combined with social measures, have helped to diminish the spread. Nevertheless, two years after the outbreak, the virus continues to propagate and claim victims worldwide. Therefore, there is a need for inexpensive, efficient, and real-time screening methods. In this scenario, the use of coughing samples as audio signals is a potential way to provide clinicians with an automatic tool for pre-diagnosing COVID-19 using AI techniques. This study investigates the use of cough-utterances of subjects for the automatic detection of COVID-19. Relying on x-vector embeddings obtained from custom-trained deep neural network extractors on cough audio recordings, we were able to get highly competitive classification performance. Furthermore, we analyze the sensitivity of the extractors to domain dependence; and the quality of the embeddings produced in this context.**

***Keywords*—cough analysis, COVID-19, computational paralinguistics, x-vectors**

I. Introduction

As of March 2020, the World Health Organization (WHO) formally declared a worldwide epidemic of the novel coronavirus called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), now simply known as COVID-19. Two years on (March 2022), the virus has taken 6,074,234[1] lives so far (officially). Experts and epidemiologists have made efforts to find ways to carry out massive COVID-19 screenings in an attempt to control the spread. A large number of tools and methods for screening COVID-19 are now available and these have helped to control the pandemic [1], [2]. The different ways of diagnosis (e.g. viral and serology tests), although effective, have limitations; also, the lack of complementary pre-screening techniques that can efficiently decide who should be tested makes it difficult to efficiently limit the spread. The current COVID-19 tests also require time to get the results, and depending on the country, they might not be economically feasible for a massive-scale deployment [3].

Hence, there is still the need for a less expensive, non-invasive, and readily accessible form of pre-diagnosis that is capable of giving real-time results. Although COVID-19 symptoms vary from subject to subject, the most common ones are a dry cough, fever, nasal congestion, breathing difficulties, a sore throat, and, in certain cases, the subject might not even display any symptoms at all [4]. Moreover, symptoms like nasal congestion and breathing difficulties may have a straightforward effect on the way the subjects produce the speech. The automatic analysis of coughing audio samples could be a potential way for pre-screening and even monitoring the disease, and could be extensively applied using e.g. smartphone devices, to prevent or slow the spread of the disease [5].

Automatic cough discrimination based on utterances is not a new approach. The literature reports studies on classifying pneumonia and asthma [6] and diagnosing pertussis [7] from coughing data samples. In this context, COVID-19 has also been examined, as recent studies carried out investigations on screening the disease using cough recordings, relying on deep learning methods (e.g., CNNs) [8], [9] and on standard machine learning algorithms [10]. Naturally, the analysis and classification of COVID-19 speech has gained attention recently. For instance, Bartl-Pokorny et al. found that voice acoustic correlates with a COVID-19 infection based on a set of acoustic parameters [11]; and Udhaya Sankar et al identified regular and irregular speech/voice patterns for the detection of the disease [12].

In this study, we present our methodology based on *x*-vectors, which is the current state-of-the-art in speaker recognition [13]. Applied as feature extractors, x-vectors have been shown to capture meta-information from the human voice such as the gender of the speaker, as well as their speech

[1]"COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)". 20 March 2022.

rate (long-term speech traits). This functionality has been exploited especially in text-independent speaker recognition approaches [14]–[17]. X-vectors have been widely applied in a variety of studies related to text-independent speaker recognition (see e.g. [14]–[17]). Moreover, x-vectors are used in the field of computational paralinguistics. For instance, studies reported good performances for classifying emotions [18], Alzheimer's Disease [19], the age and gender [20], and the sleepiness of subjects [21].

In this paper, taking advantage of the availability of a large-sized corpus (COUGHVID [22]) related to the domain of the actual tasks, we present custom x-vector DNN models trained from scratch. These models are then used for feature extraction in the next step. We validated our approach for the standardized subset of the Cambridge COVID-19 Sound database, containing the coughing sounds of individuals having and not having COVID-19. It is a binary classification task, where the goal is to predict whether a subject has COVID-19 based on their coughing samples. The proposed approach gives a competitive performance on the cough subset of the Cambridge COVID-19 Sound corpus. Our findings, which are also in accord with those from a previous study [21], indicate that the extractors fitted with in-domain data achieve a better performance than the standard pre-trained models (such as the pre-trained x-vector extractor described by Snyder et al. [13]).

II. Data

We performed our experiments on a standardized subset of the Cambridge COVID-19 Sound database. It comprises data on the diagnosis of COVID-19 based primarily on voice, breathing, and coughing. The Cough subset consists of 725 recordings (1.63 hours) from 343 subjects. Each cough recording consists of one to three forced coughs. For each recording, a COVID-19 test result was available which was self-reported by the participant: positive or negative. Although the corpus was crowd-sourced, efforts were made to ensure its good quality. After manually checking the audio quality, all the recordings were converted to a 16 kHz sampling rate and a mono, 16-bit resolution. The standardized train, development, and test sets contain mutually different speakers, but within each set, the same speaker can appear more than once. Later, this subset was used in the Interspeech 2021 Computational Paralinguistic Challenge (ComParE) [23].

III. Deep Neural Network Embeddings

The x-vector approach can be thought as of a feed-forward neural network feature extraction method that provides fixed-dimensional embeddings for variable-length utterances.

A. DNN Architecture

Table I outlines the structure of the DNN. The *frame-level* layers have a time-delay architecture. Let us assume that t is the actual time step. At the input, the frames are spliced together; namely, the input to the current layer is the spliced output of the previous layer (i.e., input to layer *frame3* is the spliced output of layer *frame2*, at frames $t-3$ and $t+3$).

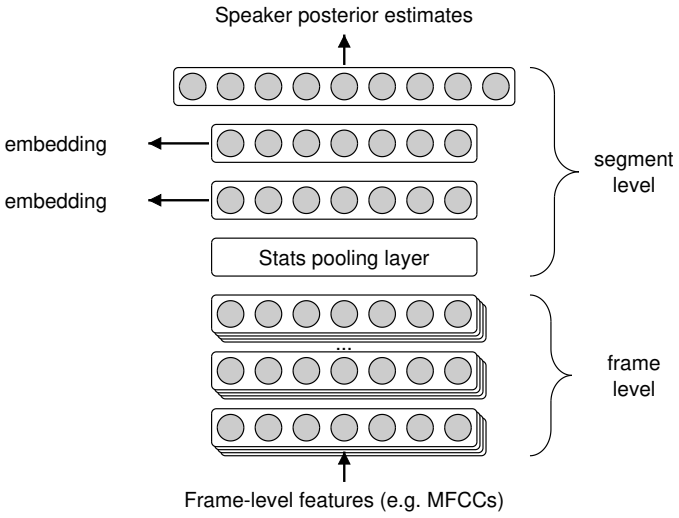

Fig. 1. The DNN structure of the x-vector feature extractor, following the work of Snyder et al. [13].

TABLE I
THE DNN ARCHITECTURE OF THE X-VECTOR SYSTEM, CONSISTING OF FIVE FRAME-LEVEL LAYERS, A STATISTICS POOLING LAYER, TWO SEGMENT LAYERS AND A FINAL SOFTMAX LAYER. N REPRESENTS THE NUMBER OF TRAINING SPEAKERS IN THE SOFTMAX LAYER. THIS ARCHITECTURE IS BASED ON THE ONE DESCRIBED BY SNYDER ET AL. [24]

Layer	Layer context	Tot. context	In, Out
frame1	[t-2, t+2]	5	120, 512
frame2	{t-2, t, t+2}	9	1536, 512
frame3	{t-3, t, t+3}	15	1536, 512
frame4	{t}	15	512, 512
frame5	{t}	15	512, 1500
stats pooling	[0, T}	T	1500T, 3000
segment6	{0}	T	3000, 512
segment7	{0}	T	512, 512
softmax	{0}	T	512, N

Next, the *stats pooling* layer gets the T frame-level activations of the last frame-level layer (*frame5*), aggregates over the input segment, and computes segment-level statistics, i.e. the mean and standard deviation. These statistics are concatenated and used as input for the next *segment6* and *segment7* layers, respectively. The last layer is the *softmax* output layer, which is discarded after training the DNN. Instead of predicting frames, the DNN is trained to predict speakers from variable-length utterances. Namely, it is trained to classify speakers present in the train set utilizing a multi-class cross-entropy objective function [24].

B. The x-vector

The embeddings produced by the network described above capture information from the speakers over the whole audio-signal. This type of embedding may help us to better discriminate the utterances as their characteristics here are acquired at the utterance level rather than at the frame level. Such embeddings are called *x-vectors* and they can be extracted from any *segment* layer; that is, either *segment6* or *segment7*

layers (see Table I) [13], [24]. Normally, embeddings from the *segment6* layer give a better performance than those from *segment7* [13].

IV. Custom x-vector Extractors

Most standard feature extractor approaches provide pre-trained models that were fitted on huge amounts of data and these models are usually applied to similar or distinct domain tasks. Of course, this is the case for the x-vector approach as well. However, pre-trained models do not always perform the best in particular cases, e.g., when the data differs from their original domain. A way to overcome this issue might be to train the whole feature extractor model (in our case, an x-vector DNN) from scratch. Following the findings of a previous study [21], where we demonstrated the benefits for the extractor when it learns from in-domain data compared with those of the pre-trained models, here we apply the same strategy. This way, the extractor might compute DNN representations of a higher quality than those got with out-of-domain data. Commonly, training like this requires a significant amount of data, while the COVID-19 Cough dataset is quite limited in size: the x-vector extractor DNN tends to under-fit due to the small size of the dataset, if we would use the training set (only 286 utterances, i.e. 31 minutes) for training our x-vector extractors on. Owing to this, we will present custom x-vector extractors fitted from scratch on a corpus of a relevant size that is related to the domain of the given task.

A. COVID-19 Cough Extractor

As stated above, the DNN-extractor might perform better when trained on data related to the corpus in question. With this in mind, for the COVID-19 Cough dataset, we employed an in-domain corpus. Namely, we used the COUGHVID [22] database to train the DNN to extract features from the COVID-19 Cough corpus. The COUGHVID database is an extensive dataset of (COVID-19 related) coughing sounds that were collected via a web app by the École Polytechnique Fédérale de Lausanne (EPFL) in Switzerland. It consists of more than 20,000 crowdsourced cough recordings that were partially validated by expert pulmonologists [22]. Since the COUGHVID corpus is crowdsourced, it suffers from data contamination, i.e. it comprises samples that are unrelated to the content in question. However, this corpus carries compiled metadata which can be used to filter the data and find the utterances that have a high probability of being coughing sounds (see [22]). We selected the recordings with $p > 0.95$. After this step, the number of samples was 10,966 utterances (26 hours).

V. Experimental Setup

Snyder et. al introduced a pre-trained x-vector model in their study [13] that was fitted using a combination of a portion of Switchboard (SWBD) and a subset of the NIST SRE corpus. Here, we also used this model as an additional DNN-extractor in our experiments for computing the embeddings from the Cambridge COVID-19 Sound database. The motivation for using this model lies in the comparison and analysis of the quality of the features produced by a *(standard)* model that was trained using data that had a different context from the actual task. Below, we describe the models we employed for the x-vector feature extraction step on the different corpora. The representations were extracted using the *segment6* affine layer in each experiment. We utilized the Kaldi Toolkit [25] both for training our x-vector extractors and for extracting the x-vector embeddings.

A. Frame-level Representations

In the x-vector approach, it is standard practice to employ Mel-Frequency Cepstral Coefficients (MFCCs) as features. However, as the x-vector extractors are neural networks, frame-level representations like the spectrograms and Mel-frequency filter-banks ("FBANKs") might provide high-quality features as well. Both kinds of representations have proved to be useful in deep learning studies related to speech analysis. The former, for instance, was utilized for emotion recognition research [26], and in speech enhancement studies [27]; while the latter was used in speech recognition [28].

This is why in our experiments we used all three types of frame-level features extracted from the utterances (i.e. MFCCs, FBANKs and spectrograms). All three were computed with the standard values of a 25ms frame length and a step size of 10ms. For the MFCCs, we extracted 23-dimensional coefficients, while for the FBANKs we computed 40 mel bins. Regarding the spectrograms, we used a window-size of 25ms, and a step size of 10ms, along with the energy computation.

B. Extractors with Data Augmentation

In order to increase the variance of the training data and make the extractors noise-robust, we applied data augmentation on the COUGHVID corpus in the following way. From additive noises and reverberation, two of the following types of augmentation were selected randomly: babble, music, noise, and reverberation. The first three types correspond to adding or fitting noise to the original utterances. The fourth one involves a convolution of room impulse responses with the audio (reverberation). The final size of the augmented COUGHVID corpus comprised 20,000 samples (67 hours). Besides training on the original COUGHVID corpus, we trained further x-vector extractors on this augmented version as well.

C. Evaluation Methods

Support Vector Machines (SVM) was the algorithm utilized to perform the classification relying on the x-vector representations. We employed the libSVM implementation [29] with a linear kernel and the C complexity parameter was set in the range 10^{-5}, ..., 10^{1}, based on the performance on the development set. As for the metrics, we employed Unweighted Average Recall (UAR), which is the de facto standard on the COVID-19 Cough corpus for evaluation [23].

TABLE II
UAR SCORES OF THE EXPERIMENTS GOT USING THE EMBEDDINGS FROM THE EXTRACTORS FITTED ON THE COUGHVID DATA; AND GOT USING THE PRE-TRAINED X-VECTOR EXTRACTOR EMBEDDINGS [13].

Feature Set	Dev	Test
COUGHVID (standard) (MFCC)	57.6	60.1
COUGHVID (standard) (FBANK)	63.9	70.9
COUGHVID (standard) (spectrogram)	63.6	71.8
COUGHVID (augmented) (MFCC)	61.5	52.8
COUGHVID (augmented) (FBANK)	67.4	62.8
COUGHVID (augmented) (spectrogram)	60.4	70.5
Pre-trained Extractor [13]	59.7	56.4
ComParE functionals	57.4	70.6

VI. RESULTS AND DISCUSSION

Table II lists the performances (UAR percentage scores) we obtained for the custom x-vector extractors, and also when employing the pre-trained standard x-vector model. The same table shows the results obtained with the 6373-sized 'ComParE functionals' attribute set [23]. In general, we can see that MFCCs do not really perform well, leading to 2-class UAR scores of $52.8 \ldots 61.5\%$. This is understandable, though, as they were originally intended to represent the spoken content of speech; furthermore, deep networks (such as x-vector extractors) were shown to perform better on raw features such as FBANKs [30]. As expected, FBANKs and spectrogram features helped to produce better results: the scores lay between 60.4% and 67.4% on the development set, while on the test set we measured over 70% in three out of the four cases.

It is also apparent that the augmented extractors did not gain any benefits from the noisy-augmented data. This may be due to the quality of the utterances in the COUGHVID corpus; as they are part of a crowdsourced process, a significant number of the recordings suffer from background noise and are of poor quality. Thus, applying augmentation techniques based on the addition of *further* noise became counterproductive when training the extractor, as the model might have learned irrelevant information from the augmented utterance variants.

The standard pre-trained x-vector extractor did not give better results than its counterparts. These results give us a clue about the domain sensitivity that the x-vector approach may be subject to. Although x-vector is considered a *data-greedy* approach [13], [31], and this pre-trained x-vector model was fitted on several hundreds of hours of data, we saw that employing a significantly smaller amount of data still produced good performances (at least for this particular task). That is, with 26 hours for the *non-augmented* cough-extractors we were able to outperform the several hundreds of hours of the pre-trained extractor, most likely due to training with in-domain data.

A. Combination Experiments

Next, we performed combination experiments, where we fused each x-vector based prediction with the 'ComParE functionals' feature set. For this, we opted for late fusion by taking the mean of the corresponding posterior scores. Since we did this in an unweighted manner, no meta-parameters were tuned in this step, which, by our expectations, should increase the robustness of the predictions. Table III shows the UAR scores obtained this way. Clearly, this combination improved the performance of the x-vector-based models in *each* case. In the end, we achieved UAR scores up to 73.5% which, although not exceeding the highest score published on this corpus (75.9% [9]), is still a quite competitive performance score.

TABLE III
UAR SCORES GOT ON THE TEST SET WITH AN UNWEIGHTED LATE FUSION OF THE PREDICTIONS WITH THE COMPARE FUNCTIONALS FEATURES. THE IMPROVED VALUES ARE SHOWN IN **bold**.

Feature Set	Standard	Augmented
COUGHVID (MFCC)	**60.2**	**57.3**
COUGHVID (FBANK)	**73.1**	**66.8**
COUGHVID (spectrogram)	**73.5**	**73.1**
Pre-trained Extractor [13]	—	**63.8**

B. Comparison with the Literature

Finally, to place our results in context, we compare them with results reported in the literature. Table IV shows the notable scores published on the same corpus. We see that the performance scores of the individual methods reported in the ComParE Challenge baseline paper (i.e. in [23], see the first block) are outperformed by both the standalone and the proposed combined methods, with one exception. However, we should mention that even the result of the Bag-of-Audio-Words approach was obtained by tuning its hyperparameter (the number of the audio words) on the test set; the configuration which gave best development-set score (i.e. 500 audio words instead of 2000) led to an UAR value of 67.6% on the test set, which falls below our scores of 71.8% and 73.1% which were obtained in a scientifically sound manner (i.e. tuning all parameters on the development set).

The second block of Table IV shows the results of standalone methods published by other research teams on the Cambridge COVID-19 Sound (Cough) corpus. Using the PASE+ features [32] proved to be inferior to our COUGHVID x-vectors approach (UAR scores of 64.1% and 71.8%, respectively). Illium et al. experimented with four different CNN architectures [33]; we were able to outperform their Vision Transformer and Vertical Vision Transformer networks (employing blocks of multi-head self-attentions followed by fully-connected layers), while we were on par with their Sub-SpectralClassifier network (utilizing four small CNNs trained on different, non-overlapping Mel bands and aggregated via a classifier sub-network). Casanova et al. also employed several different CNNs; their most successful model utilized transfer learning from the PANN CNN14 model [34]. Surprisingly, though, the model trained on 5-fold cross-validation performed

TABLE IV
UAR SCORES GOT ON THE TEST SET REPORTED IN THE LITERATURE. "*" DENOTES AN APPROACH INVOLVING METHOD FUSION.

Approach	Test
ComParE functionals [23]	65.5
Bag-of-Audio-Words [23]	72.9
DenseNet121 [23]	64.1
AuDeep (-60 dB) [23]	67.6
PASE+ features [32]	64.1
Vertical Vision Transformer CNN [33]	68.9
Vision Transformer CNN [33]	69.9
Sub Spectral Classifier CNN [33]	72.0
Transfer learning CNN (5-fold ensemble) [9]	69.6
Transfer learning CNN (simple holdout) [9]	75.9
CNN + TDNN-F + PASE+ features* [32]	69.3
ComParE Challenge baseline (fusion on test)* [23]	73.9
COUGHVID x-vectors (spectrogram)	71.8
ComParE functionals + COUGHVID x-vectors*	73.1

significantly worse than the one trained using a simple holdout set.

The third block of Table IV shows values obtained by combinations of methods; we see that even our standalone approach outperformed the three-wise combination of Solera-Ureña et al. [32], while our combined method performed only slightly worse than the four-wise baseline combination, fine-tuned on the test set. Overall, our proposed approach (see the last block of Table IV) led to competitive UAR values compared to those achieved by other research groups (mostly obtained by deep learning approaches).

VII. Conclusions

In this study we applied deep neural network models for the extraction of x-vector embeddings, in order to discriminate COVID-19 cough recordings as an automatic tool for screening the disease. We described custom x-vector extractors built upon distinct frame-level representations. Unlike standard approaches, our DNN models were trained from scratch by utilizing data related to the domain of the actual task (i.e. classification of cough sounds). We demonstrated the efficiency of our extractors by producing competitive scores for the COVID-19 Cough dataset. Our findings indicate that spectrograms and FBANKS may be a powerful alternative as frame-level features for the x-vector DNN architecture type. We also found that the standard x-vector pre-trained model did not produce better representations than its customized extractor counterparts. This could be attributed to *domain-sensitivity*, i.e. that this model was fitted on a different data domain (although using huge amounts of it). We found that, in this particular task, training with a significantly smaller amount of data allowed the extractors to remain competitive and they even outperformed the pre-trained model. Furthermore, our custom DNN models may be useful for transfer learning approaches in future studies related to COVID-19 screening based on cough audio recordings.

VIII. Acknowledgements

This study was supported by the NRDI Office of the Hungarian Ministry of Innovation and Technology (grant no. TKP2021-NVA-09), and within the framework of the Artificial Intelligence National Laboratory Program (RRF-2.3.1-21-2022-00004).

References

[1] D. Vinh, X. Zhao, K. Kiong, T. Guo, Y. Jozaghi, C. Yao, J. Kelley, and E. Hanna, "Overview of COVID-19 testing and implications for otolaryngologists," *Head & neck*, vol. 42, no. 7, pp. 1629–1633, 2020.

[2] N. Beeching, T. Fletcher, and M. Beadsworth, "Covid-19: testing times," *The BMJ*, vol. 369, 2020.

[3] E. Surkova, V. Nikolayevskyy, and F. Drobniewski, "False-positive COVID-19 results: Hidden problems and costs," *The Lancet Respiratory Medicine*, vol. 8, no. 12, pp. 1167–1168, 2020.

[4] A. Carfì, R. Bernabei, and F. Landi, "Persistent symptoms in patients after acute covid-19," *Journal of the American Medical Association*, vol. 324, no. 6, pp. 603–605, 2020.

[5] 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," *Frontiers in Digital Health*, vol. Mar, 2021.

[6] Y. Amrulloh, U. Abeyratne, V. Swarnkar, and R. Triasih, "Cough sound analysis for pneumonia and asthma classification in pediatric population," in *Proceedings of ISMS*, Kuala Lumpur, Malaysia, 2015, pp. 127–131.

[7] R. Pramono, S. Imtiaz, and E. Rodriguez-Villegas, "A cough-based algorithm for automatic diagnosis of pertussis," *PloS one*, vol. 11, no. 9, 2016.

[8] J. Laguarta Soler, F. Hueto, and B. Subirana, "Covid-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 275–281, 2020.

[9] E. Casanova, A. Candido Jr., R. C. Fernandes Jr., M. Finger, L. R. Stefanel Gris, M. A. Ponti, and D. P. Pinto da Silva, "Transfer learning and data augmentation techniques to the COVID-19 identification tasks in ComParE 2021," in *Proceedings of Interspeech*, Brno, Czechia, 2021, pp. 436–440.

[10] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data," in *Proceedings of ACM SIGKDD*, virtual, 2020, pp. 3474–3484.

[11] K. Bartl-Pokorny, F. Pokorny, A. Batliner, S. Amiriparian, A. Semertzidou, F. Eyben, E. Kramer, F. Schmidt, R. Schönweiler, M. Wehler, and B. Schuller, "The voice of COVID-19: Acoustic correlates of infection," *Journal of the Acoustical Society of America*, vol. 149, no. 6, pp. 4377–4383, 2021.

[12] S. Udhaya Sankar, R. Ganesan, J. Katiravan, M. Ramakrishnan, and R. Ruhin Kouser, "Mobile application based speech and voice analysis for COVID-19 detection using computational audit techniques," *International Journal of Pervasive Computing and Communications*, 2020.

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker verification," in *Proceedings of ICASSP*, Calgary, Alberta, Canada, 2018, pp. 5329–5333.

[14] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker Recognition for Multi-speaker conversations using x-vectors," in *Proceedings of ICASSP*, Brighton, UK, 2019, pp. 5796–5800.

[15] A. Silnova, N. Brummer, D. Garcia-Romero, D. Snyder, and L. Burget, "Fast variational Bayes for heavy-tailed PLDA applied to i-vectors and x-vectors," in *Proceedings of Interspeech*, Hyderabad, India, 2018, pp. 72–76.

[16] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep Speaker Recognition," in *Proceedings of Interspeech*, Hyderabad, India, 2018, pp. 1086–1090.

[17] O. Novotnỳ, O. Plchot, P. Matejka, L. Mosner, and O. Glembek, "On the use of x-vectors for robust speaker recognition." in *Proceedings of Odyssey*, Les Sables d'Olonne, France, 2018, pp. 168–175.

[18] S. Zargarbashi and B. Babaali, "A multi-modal feature embedding approach to diagnose Alzheimer's disease from spoken language," *arXiv preprint arXiv:1910.00330*, 2019.

[19] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker verification," in *Proceedings of ICASSP*, Barcelona, Catalunia, Spain (online), 2020, pp. 7169–7173.

[20] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *Proceedings of ASRU*, Singapore, Dec 2019, pp. 726–733.

[21] J. V. Egas López and G. Gosztolya, "Deep neural network embeddings for the estimation of the degree of sleepiness," in *Proceedings of ICASSP*, Toronto, Ontario, Canada (online), Jun 2021.

[22] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, 2021.

[23] B. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, L. Rothkrantz, J. Zwerts, J. Treep, and C. Kaandorp, "The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates," in *Proceedings of Interspeech*, Brno, Czechia, 2021, pp. 431–435.

[24] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network embeddings for text-independent speaker verification," in *Proceedings of Interspeech*, Stockholm, Sweden, Aug 2017, pp. 999–1003.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proceedings of ASRU*, Big Island, HI, USA, Dec 2011.

[26] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proceedings of PlatCon*, Busan, Korea, 2017, pp. 1–5.

[27] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, "A fully convolutional neural network for complex spectrogram processing in speech enhancement," in *Proceedings of ICASSP*, Brighton, UK, 2019, pp. 5756–5760.

[28] H. Seki, K. Yamamoto, and S. Nakagawa, "A deep neural network integrated with filterbank learning for speech recognition," in *Proceedings of ICASSP*, New Orleans, LA, USA, 2017, pp. 5480–5484.

[29] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.

[30] A.-R. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using Deep Belief Networks," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 14–22, 2011.

[31] P. Ghahremani, P. S. Nidadavolu, N. Chen, J. Villalba, D. Povey, S. Khudanpur, and N. Dehak, "End-to-end deep neural network age estimation," in *Proceedings of Interspeech*, Hyderabad, India, 2018, pp. 277–281.

[32] R. Solera-Ureña, C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso, "Transfer learning-based cough representations for automatic detection of COVID-19," in *Proceedings of Interspeech*, Brno, Czechia, 2021, pp. 431–435.

[33] S. Illium, R. Müller, A. Sedlmeier, and C.-L. Popien, "Visual transformers for primates classification and Covid detection," in *Proceedings of Interspeech*, Brno, Czechia, 2021, pp. 451–455.

[34] 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," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.