Automatic Detection of Profile Features

John Mendonça
Instituto Superior Técnico, University of Lisbon
Email: john.mendonca@tecnico.ulisboa.pt

***Abstract*—Speech corpora collected via crowdsourcing typically require costly validation to verify certain characteristics of speakers, or submission correctness. Moreover, this validation should also exclude recordings corresponding to multiple speakers sharing the same account or multiple accounts for the same speaker. This thesis focus on the use of speech pattern recognition techniques to perform this automatic validation. This is accomplished by training an x-vector based system in a large open-source corpus, and enrolling the first utterance from each speaker in a crowdsourcing corpora collection job which is then compared to subsequent task completions. The resulting speaker embeddings are also used for gender verification. As a proof-of-concept, we used this approach to validate different datasets in 3 languages, adopting score normalisation techniques. Results show an EER below the 4% mark on all experiments, indicating the possibility to adopt the same threshold in different datasets without substantial loss of performance. This enables the validation of crowdsourced task completions immediately after submission.**

This thesis also involved the participation in an international Computational Paralinguistics Challenge, where we studied the automatic prediction from conversational speech of breath signals obtained from respiratory belts. We analysed both original and predicted signals and identified the subsets of most irregular belt signals which yield the worst performance, showing how they affect results. We proposed several variants of an end-to-end baseline system, such as BiLSTM, and AM/FM decomposition as input. We showed that these models can predict breathing patterns and clinically relevant parameters, such as breathing rate, in simulated video-conferencing sessions.

Index Terms: Crowdsourcing; Paralinguistics; Speaker Verification; Gender Recognition; Breath Detection.

I. Introduction

Speech technology has significantly influenced the lives of everyday users, impacting the way people find, consume, and act on information. Starting with the widespread adoption of mobile devices such as smartphones, more recent technological advancements have led to a larger use of voice search and Intelligent Virtual Assistants. Recent studies indicate that over 50% of web searches will be conducted through voice by 2020, and that 55% of U.S. households will possess an intelligent virtual assistant [1]. Fuelling this sharp increase is the growing consumer demand for online self-service, self-reliance, and rapid query resolution, while at the same time helping companies enhance operational efficiency and reduce costs [2]. With speech technologies trending towards a more predominant use, the need for efficient and effective interactions with users has become increasingly important. The large amount of data collected resulting from the interaction with speech-based systems allows Artificial Intelligence (AI) to adapt and improve over time. AI enables the continuous improvement of speech systems by including collected speech data in the training of Machine Learning models that tackle common speech applications such as ASR or Speech Synthesis. However, more and more use cases and industry applications that use speech to obtain interpretable speaker information have surged.

The human voice conveys substantial amounts of information related to the speaker. For instance, information including physical traits (age [3], gender[4]), language (nationality, nativeness) [5] [6], health (speech affecting diseases) [7] and mood [8] can be obtained from voice. Such profile information can be extracted directly from speech (using the raw-time waveform or spectrum), or from speech derived features (intensity, voice quality features, speech rate, breathing rate) [9]. The automatic detection of profile features enables the development of smarter user interfaces and an enhanced user experience, especially when using devices or applications where this information is required. Additionally, it can assist in more sensitive applications such as identity verification, where speaker verification or facial reconstructions from voice [10] may be of value.

In this work, we will investigate how the automatic detection of profile features and other metadata extracted from voice can be used. More specifically, we intend to apply this extracted information in crowdsourced speech data collections and on breathing pattern estimation.

This paper is organized as follows: In Chapter 2, the use of voice pattern recognition techniques is explored in the context of fraud detection in a crowdsourced speech data collection environment, namely speaker and gender verification. Chapter 3 provides an analysis of breathing pattern recognition from voice, where we propose a system that automatic predicts these patterns and related metrics. Chapter 4 is the final chapter, where conclusions pertaining this work are drawn, together with some topics for future work.

II. Voice Profiling for Crowdsourcing

The advent of complex models such as Deep Neural Networks raises the need for large amounts of labelled data [11]. Instead of using experts to label a dataset, crowdsourcing platforms enable a more scalable labelling process by breaking down large datasets into small tasks. These well-defined micro-tasks are performed by the crowd with similar quality results [12]. This technique is often used by companies and universities by providing the required data to create accurate models at a lower cost.

The required user base for a given dataset is obtained by rewarding users for each completed task. On the one hand, this invites a larger pool of willing workers to complete these tasks. On the other hand, users are also encouraged to produce low quality work as it often blends in with the crowd [13]. As a result, several methods to detect low quality work have been developed, namely agreement between users or with a gold standard, or more complex behavioural capturing techniques to predict outcome measures such as work quality, errors, and the likelihood of cheating [13].

Validating speech corpora collected via crowdsourcing raises particular challenges, as it is not possible to establish a gold-standard. In this case, validation tasks are typically set up to validate certain characteristics of speakers (nativeness, gender) or submission correctness (prompt matching the audio, for example). However, this validation process adds to the costs of the dataset, and does not solve demographic problems such as incorrect profile labels (due to mistakes made when users fill out their profile), multiple accounts for the same speaker, or multiple speakers sharing the same account. The detection of incorrect profile labels can be partly automatised by using gender/age classifiers [14] [15] or nativeness classifiers [6], for instance. The detection of multiple accounts for the same speaker, on the other hand, may be addressed by speaker verification

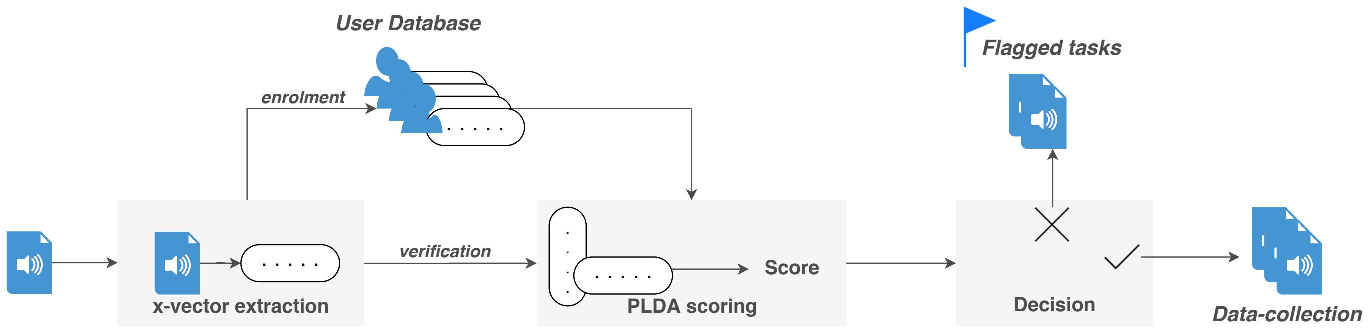

Fig. 1. Proposed Speaker Verification pipeline.

techniques [16] [17]. This work focuses on the detection of gender and multiple speakers sharing the same account, describing the use of pattern recognition techniques to flag profile errors or personification attempts as early as possible, during the data collection process.

Speaker embeddings extracted from deep neural networks (DNNs) such as x-vectors [16] have become the state-of-the-art in text-independent speaker recognition tasks, surpassing older systems, namely i-vectors [18]. These embeddings are then paired with a probabilistic linear discriminant analysis (PLDA) [19] classifier which is used to compare embeddings, allowing same-or-different speaker decisions. In [20], the authors explored the extracted i-vectors and x-vectors from a speaker-verification trained system to probe additional information. This information included gender, speaking rate and session related information such as word and phoneme recognition. Our solution is based on training an x-vector system using a large open-source corpus, and enrolling the first utterance from each speaker which is then compared to his/hers subsequent task completions. We also take advantage of the extracted embedding to predict gender.

A. Datasets

In this experiment, three data sources were used. The out-of-domain corpus used to train our speaker verification system was Voxceleb (VC) [21] [22], which is a large, multi-lingual collection of YouTube videos from celebrities. Details of all datasets are shown in Table I.

1) DC: Crowdsourced data collections: The speech datasets obtained in a real crowdsourcing environment (DC) are a collection of multilingual prompt reading tasks, recorded in a mobile application environment. For this work, we selected 3 datasets of American English, Hebrew and Mexican Spanish. Several validation steps have been previously applied to these recordings, however none included a biometric evaluation step, meaning speaker labels were not validated. Considering users are paid for each completed task, there is motivation to enrol additional speakers on the same account in order to expedite completed tasks.

2) CV: Common Voice : The Common Voice (CV) project [23] is an open-source crowdsourced speech data collection. It includes almost seven thousand hours of validated recorded speech from 56 languages and dialects. The collection is conducted by untrained volunteers who read sentences from original contributions and public domain texts. Each user has a unique ID, but additional profile information such as gender and age is not required. For this work, we used English and German multi-accent datasets and filtered users without profile information. Given that the data collection process is similar to for-profit crowdsourcing, CV datasets were used to evaluate the performance of our validation systems, with the assumption that each speaker corresponds to a single account, since the collection was done on a volunteer basis.

TABLE I
DATASET SIZE AND DETECTED FRAUD.

	Size		Fraud	
Dataset	# Utt	# Spk	#Utt	# Spk
VC1	4,878	40	-	-
VC2	1,092,009	5,994	-	-
CV_EN	5,848	2,467	0	0
CV_DE	6,680	1,191	0	0
DC_EN	2,745	277	0	0
DC_HE	2,144	147	13	5
DC_ES	8,333	65	8	3

B. Proposed speaker verification architecture

Figure 1 represents an overview of the speaker verification pipeline used in the experiments. Its front end consists of an embedding extraction network, which condenses information related to the speaker to a fixed sized feature vector from a variable length audio signal. The back-end system consists of a scoring procedure, followed by a decision step. A score is attributed to a pair of embeddings using PLDA scoring [19]. An utterance is considered verified if its score is higher than a given threshold, when evaluated against the enrolled utterance.

Decision thresholds are a by-product of minimising speaker recognition performance metrics. When using the Equal Error Rate (EER) as a metric, the threshold is chosen as to equate the False Rejection Rate (FRR) with the False Acceptance Rate (FAR). In the context of crowdsourcing, False Acceptances occur when the system validates fraudulent task completions from speakers other than the enrolled one. On the other hand, False Rejections occur when the system erroneously flags tasks that were completed by the enrolled speaker.

In a live production environment, utterances are submitted 'on-the-fly', meaning a decision threshold must be decided beforehand. If the new trials belong to unseen, out-of-domain data (different language or channel conditions), the previously computed threshold must be adapted in order to achieve the same performance [24]. Score space normalisation techniques can be used to tackle this problem, by reducing variability in the scores. The Adapted Symmetric Scoring normalisation [25] normalises scores according to the mean and standard deviation of impostor (different speaker) distributions. This normalisation is calculated from the N_t closest files from a subset of enrolment/test called cohort list. Typical cohort lists have sizes (N_c) of thousands, making them able to experiment with N_t. For instance, in [26], the authors reported a minDCF minimum by using a N_t set to between 200 and 500, in a cohort list with N_c over 2,000 files. Other authors have also suggested a random selection of utterances [25]. In an online setting where there is no prior enrolment, the use of score normalisation techniques requires a waiting period to allow for a number of utterances to be submitted and be used in the cohort list. In our experiments, we opted, for each dataset, to select a smaller cohort list containing random utterances, and using the full list for normalisation calculation (i.e., $N_t = N_c$).

1) Experimental Set-up: Our embedding extraction and decision-making followed the Kaldi Speech Recognition Toolkit [27] recipe for *VoxCeleb*. We experimented with i-vectors, but due to space limitations, and considering the overall superior x-vector performance [16], we only present results for the latter. Training was conducted on the dev set of *VoxCeleb2*, augmented with reverberation and music, babble and noise from the MUSAN corpus [28].

The features were 30 dimensional MFCCs obtained every 10ms with a frame-length of 25ms, mean-normalised over a sliding window of up to 3 seconds. An energy-based VAD module filtered out non-speech frames. The x-vector was extracted from the last layers of the pre-trained DNN model (before the softmax layer), outputting 512-dimensional embeddings which were centred, dimensionality reduced to 200 using LDA, and length normalised.

In our experiments, we assume users have no previous submitted work: the system's performance is evaluated on a dataset level only, meaning there is no enrolment information available. As such, our trial setting differs from typical speaker verification evaluations because the enrolment set used in our experiments consists of only the first completed task. This decision-making process follows a production setting that compares the initially completed tasks to all subsequent tasks from a given user. This allows for the assessment of identity as early as possible. Impostor rejection is evaluated by having all utterances belonging to other speakers compared to any given enrolled speaker, generating impostors trials.

To assess and remove the occurrence of fraudulent behaviour, the DC datasets were manually validated by a single annotator. Considering the size of the datasets, only a subset of trials were selected according to the following steps: **1)** A speaker verification task using x-vectors on the full dataset is used to obtain PLDA scores; **2)** For each user, all flagged utterances (that failed automatic verification) were manually validated, together with the automatically verified utterance with the lowest PLDA score. If the lowest verified utterance was a false acceptance, we proceeded to the next verified utterance, up until the first true acceptance. We assumed all utterances with a higher score than the first true acceptance of each user were also valid; **3)** Inter-speaker comparisons were used to check whether speakers were using multiple accounts. Only the utterances with the lowest PLDA score were validated, as we assumed utterances with higher scoring were correctly verified.

A smaller validation was also conducted on the CV datasets to confirm the absence of fraud. In this validation, we only validated the 50 worst performing utterances of the full dataset.

2) Results and Discussion: The results of the manual validation are reported in the rightmost columns of Table I, for each dataset. Table II summarises verification results obtained on the different crowdsourced datasets. Performance was measured on reduced datasets that resulted from the removal of all fraud: all flagged utterances belonging to the same user were individually removed, while all utterances belonging to users found to participate in other accounts were removed, together with the flagged user. We also present results for the "baseline" *VoxCeleb1* dataset (VC1). It is possible to observe that overall EER(%) results on the crowdsourced datasets with the trial setting indicated in Section 4 are similar to the results on VC1 (less than 1% absolute increase in DC_ES). This is a promising result, considering the enrolment data is a single utterance per speaker. Furthermore, we note that the Decision Thresholds (DT) that yielded the reported EER values are variable (mean absolute difference of 2.62). This confirms the need for a normalisation step in order to use the same threshold.

TABLE II
RESULTS OBTAINED ON DIFFERENT DATASETS.

	None		AS-Norm	
Dataset	EER(%)	DT	EER(%)	DT
VC1	3.128	-3.26	-	-
CV_EN	2.319	7.08	2.432	1.94
CV_DE	2.915	7.67	2.860	0.83
DC_EN	2.083	8.09	2.492	1.47
DC_HE	0.599	14.74	1.549	1.93
DC_ES	4.082	11.40	3.676	-0.30

We experimented with several values for the size of the cohort list N_c. Considering the size of the datasets, we present results on each one as the average of five random samples using $N_c = 50$. The latter value was obtained through experimentation, being the lowest size of the cohort list with no significant performance loss in terms of EER. We did not observe a large variation of EER, except for the DC_HE dataset, where an absolute increase of 1% was noted. The Decision threshold shifts resulted in a mean absolute difference of 0.71. This means a single decision threshold can be applied to these datasets with a performance loss of less than 1%. The resulting score distributions can be visualised in Figure 2, for DC_EN.

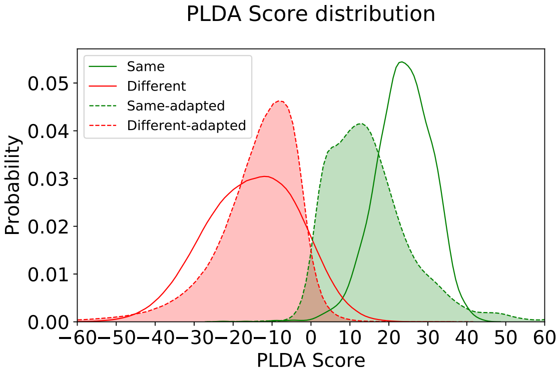

Fig. 2. Score distributions (DC_EN) for same and different speakers with and without score normalisation.

The Detection Error Trade-off (DET) curves in Figure 3 show relevant differences between the AS-norm adapted and non-adapted curves, namely the progression of False Acceptance probabilities when decreasing False Rejection probability. However, False Rejection probabilities are lower on the adapted scores when minimising False Acceptance probabilities for the DC_HE and DC_EN datasets. This can be explained by a normalisation that agglomerates scores to the opposite decision region, instead of making them more separable. We hypothesise this is a consequence of the size of the cohort list. Unlike the DET curves for DC_HE and DC_EN, the curves for the DC_ES dataset do not show substantial differences, with AS-norm achieving better performance near the EER point. We hypothesise this is due to the having 65 speakers in this dataset (contrasting with 147 and 277 speakers in DC_HE and DC_EN, respectively), which leads to a normalisation that reflects the original score distributions.

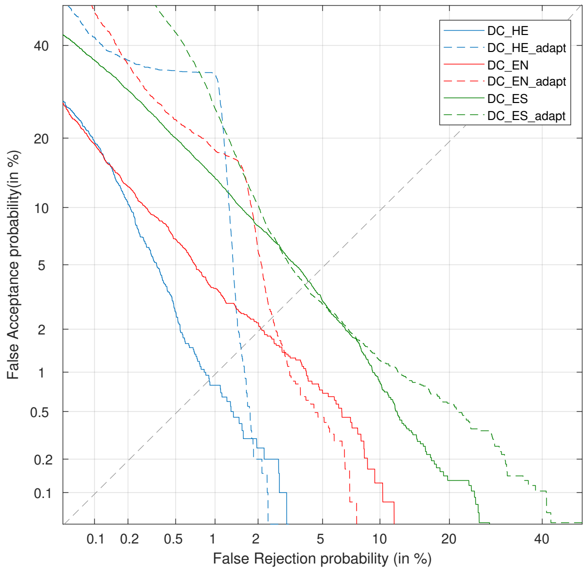

Fig. 3. DET curve for DC datasets.

Finally, fraud detection results for our system on the original datasets are presented in Table III. We note that while the number of False Acceptances (accepts fraud) is very low, it also presents a high number of False Rejections (rejects good submissions). A higher DT would alleviate this, at the cost of a higher FAR.

TABLE III
FRAUD DETECTION RESULT.

Dataset	FA(#)	FR (#)
DC_EN	-	52
DC_HE	8	8
DC_ES	0	356

C. Gender Verification

The gender extraction model from the speaker trained embedding followed the architecture proposed in [20]. The model is an MLP with a single hidden layer and a ReLU activation for the first layer and a sigmoid activation for the output layer. The hidden layer size was fixed at 500. Binary cross entropy loss was used together with Adam [29] as the optimizer, with a learning rate of 0.001. Two separate models were trained using extracted i-vectors and x-vectors using *VoxCeleb2 dev* as the training dataset and *Voxceleb2 test* as the development set.

1) Experimental Setup: To compare the performance of the model that predicts the gender from the speaker-trained embedding we used a dedicated gender recognition model as baseline. The network was based on the M5 network architecture [30]. Both models were implemented and trained in Python, using the PyTorch deep learning framework. The *Voxceleb2 dev* subset was used for training, and the *Voxceleb2 test* for development.

The baseline network consisted of four convolutional layers, each followed by a batch normalization layer and a maxpooling layer. The first layer receptive field receives a time-domain 16000-length vector that represents a waveform of 2 seconds, at a sampling rate of 8 kHz. This layer possesses a receptive field size of 80, with 256 filters with stride 4. This offers a receptive field that covers 10ms of speech, which is comparable to window lengths of other feature extractors. The following convolutional layers have a fixed receptive field of size 3, with increasing filter length of 128-258-512. The number of feature maps doubles as temporal resolution decreases by a factor of 4 in the max pooling layers. Batch Normalization is used on the output of each convolutional layer, before applying ReLU non-linearity. This alleviates the problem of exploding and vanishing gradients. The classification step is conducted using an average pooling layer, paired with a fully connected layer of length 512, and a sigmoid layer for the output.

2) Results and Discussion: We reports gender verification results obtained on DefinedCrowd and Common Voice speech data collections and includes Precision, F1-score and Recall for the i-vector, x-vector and End-to-End models. The best results for each metric and dataset is marked in bold.

The obtained results show significant performance variations in between datasets and genders. In [14], the authors reported a Recall of 98.04 and 95.05 for 'Male' and 'Female', respectively, which is similar to the performance detected on the DefinedCrowd datasets. Typically, 'Male' recall outperforms 'Female' recall, due to the fact that many speech corpora are unbalanced in terms of gender. This is also the case of $Voxceleb$, to a smaller extent, but obtained results do not show a consistent out-performance for 'Male' labels on the DefinedCrowd datasets. We note, however, that for the Common Voice dataset, performance metrics for 'Female' are much lower than for 'Male' (20% absolute difference in precison on CV_EN), which is beyond what is expected due to gender unbalance during training. Unlike the DefinedCrowd datasets, which were manually validated, we presented results on Common Voice datasets under the assumption that gender labels were correct. Considering these results, a manual validation step was conducted by one annotator, obtaining the true gender label of the worst performing utterances. These are characterized by having network outputs close to 0 or 1, indicating strong predictions. In CV_EN, out of the 5,847 utterances under test, 508 were miss classified, with 56 of these have strong predictions. Meanwhile in CV_DE, out of the 6,680 utterances under test, 376 were miss classified with 46 of these having strong predictions.

As a result of this manual validation, we detected that a majority of these instances (over 80%) had in fact the wrong gender label. Furthermore, all of the erroneous labels were female and were attributed to male speakers. While we have no concrete explanation for the reason why a substantial amount of male speakers had 'Female' labels, we believe this is due to error during profile registration, as there is no incentive to provide 'Female' labels other than the fact the datasets themselves lack female representation.

It can be observed that the performance obtained using the speaker embeddings as input is comparable to the End-to-End model, with the added benefit that the model is much simpler, an MLP. In fact, the end-to-end model failed to outperform the embedding-based models on the majority of metrics, something we believe is due to the nature of the embedding extraction, which is able to convey information related to the full embedding, unlike the end-to-end model, which is restricted to exactly 2 seconds of the embedding. This means utterances with duration lower than 2 seconds are padded with zeros before being fed to the network, and utterances longer

TABLE IV
RESULTS OBTAINED ON CROWDSOURCED DATASETS.

		Male			Female		
Dataset	Architecure	Precision	Recall	F1 Score	Precision	Recall	F1 Score
CV_EN	i-vector	**0.98**	0.91	0.95	0.66	0.89	0.76
	x-vector	**0.98**	**0.94**	0.96	**0.72**	**0.90**	**0.80**
	End2End	0.97	**0.94**	0.95	0.71	0.83	0.76
CV_DE	i-vector	**0.98**	**0.96**	0.97	**0.80**	**0.90**	**0.85**
	x-vector	**0.98**	0.94	0.96	0.72	**0.90**	0.80
	End2End	**0.98**	**0.96**	0.97	**0.80**	0.86	0.83
DC_EN	i-vector	0.93	0.96	0.94	**0.97**	0.94	0.95
	x-vector	**0.94**	**0.97**	0.95	**0.97**	**0.95**	**0.96**
	End2End	0.90	**0.97**	0.93	**0.97**	0.92	0.95
DC_HE	i-vector	**0.99**	**0.98**	**0.99**	0.97	**0.99**	**0.98**
	x-vector	**0.99**	**0.98**	0.98	**0.98**	**0.99**	**0.98**
	End2End	**0.99**	0.97	0.98	0.97	**0.99**	**0.98**

than 2 seconds are cropped, possibly discarding relevant information pertaining gender.

III. AUTOMATIC PREDICTION OF BREATHING PATTERNS

The production of speech is highly dependent on organs that are shared with the respiratory system: the lungs and the diaphragm are responsible for the pressure production required for speech; the upper vocal tract (which includes the nose, mouth, pharynx and larynx) is responsible for producing speech [31]. As such, human respiratory and speech parameters provide important cues to physicians and first-responders in determining a wide range of cardiac and respiratory diseases [32] [33] or to evaluate cognitive and neurological health [34][35]. Furthermore, information extracted from breathing patterns during speech can be used to assist speech therapists in identifying speech impediments resulting from unfavourable respiratory planning [36]. Breathing monitoring in this context is often conducted using wearable sensors, namely, face masks and/or respiratory belts [37]. The installation of these sensors requires the presence of trained medical assistants and is frequently time-consuming, negating their usefulness in emergency situations, or when the patient cannot be physically reached. A typical example of the latter scenario occurs during medical virtual online consultations, with the patient at home, where breathing information could be of use for diagnosis or monitoring. As such, automated methods based on recorded speech alone that are able to predict breathing events and parameters such as breathing rate and tidal volume may be of substantial value.

Previous studies on this topic have focused mainly on automatic recognition of breathing patterns and events directly from a processed signal (e.g. [38], [39]). In [40], the authors studied the automatic detection of the breathing signal using Deep Neural Networks (DNNs). They reported a correlation coefficient between the predicted signal and the original one of .47, with error rates pertaining breathing rate of 4.3%.

The dataset for the current work is part of the INTERSPEECH 2020 Computational Paralinguistics Challenge [41], entitled Breathing Sub Challenge. This dataset includes recordings of spontaneous speech and associated breathing patterns.

Besides describing the submitted systems aiming at the automatic prediction of breath signals from conversational speech, we also analyse both original and predicted signals in an attempt to overcome the main pitfalls of the proposed systems.

As part of this analysis, and motivated by previous work on the carrier nature of the speech signal [42], we investigate the use of the Amplitude Modulated (AM) and Frequency Modulated (FM) components of the speech signal for predicting breathing signals. The AM component only contains information related to the message, while the FM component contains information related to the speaker. As such, by using only the message component of the speech signal, we investigate if the separation of information improves overall prediction.

Given the potential interest of breathing pattern prediction in telehealth applications, we conduct additional experiments transforming the challenge dataset to emulate Voice over Internet (VoIP) conditions.

A. Datasets

The experiments for the Breathing Sub-challenge [41] are conducted using a subset of the UCL Speech Breath Monitoring (UCL-SBM) database. The dataset includes speech recorded from a head-mounted condenser microphone and normalized linear voltage readings from two piezoelectric respiratory belts that respond to changes to the thoracic circumference. All speech recordings were spontaneous, as reading tasks may introduce some bias, forcing stops that do not necessarily coincide with the breathing rhythm. The recordings were produced by native English speakers of ages ranging from 18 to 55 years old. To the best of our knowledge, all speakers were healthy. The data set contains 49 sessions, each 4 minutes in length. The corpus is split into training, development and test sets (17, 16, and 16 sessions, respectively).

An analysis of the belt signals in these datasets shows considerable variability, as illustrated in Figure 4: while most of the signals in the training set have quite regular breath patterns, this was not observed in almost half of the signals in the development set. This was the motivation for also experimenting with a reduced development set, $dev2$, from which 7 sessions were excluded, since the training material did not include sufficient examples of such irregular patterns (only 2 out of 17 sessions). The objective exclusion criteria was based in experimental results, as explained in the next Section.

In order to emulate the video-call consultation with a physician, the provided challenge dataset was augmented. The augmentation consists in passing the original, down-sampled (8 kHz) speech signal by an ITU-T G.723.1 dual rate speech coder and decoder [43]. The G.723.1 audio codec, part of the ITU-T recommendation H.324, is a Code-Excited Linear Prediction Coder widely used in VoIP applications. It compresses voice audio in 30 ms frames and operates with a sampling frequency of 8 kHz/16-bit. In this implementation in

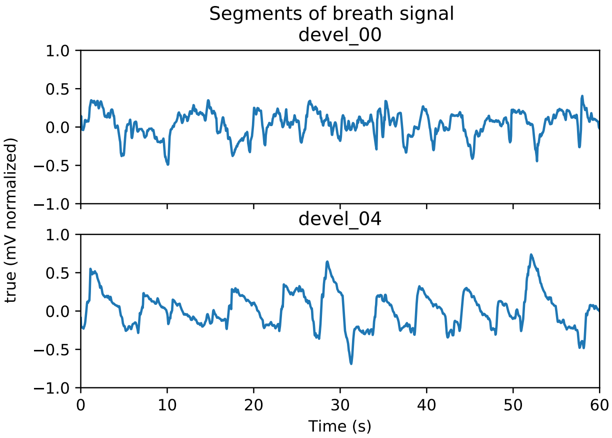

Fig. 4. Segments of breath signals from sessions *00* and *04*.

particular, MPC-MLQ (Multi-pulse Coding) mode is used, operating at 6.3 kb/s. After the decoding, the signal is up-sampled back to 16 kHz and is used in training alongside with the challenge data. This augmentation results in the doubling of the training and development data (dev_{aug}).

B. Prediction of Breathing Patterns

1) Model Architectures: The official provided end-to-end baseline architecture was used as a base for all experiments. This architecture follows typical sequence labelling models by combining a CNN for character-level representation with an RNN (in this case an LSTM) for obtaining context. The output of these layers is then fed to a dense layer for final prediction. The training loss used is the Pearson correlation coefficient r, calculated between the true and predicted belt signals.

In an effort to model respiratory planning, we replaced the original LSTM with a Bidirectional LSTM. Each RNN layer is composed of 256 hidden units with the depth-concatenated forward and backward outputs being fed to the dense layer for prediction.

2) Results on the Challenge dataset: A summary of the results obtained for the model with the best development performance of the 100 epochs of training is presented in Table VI. Results on dev did not indicate any improvement of the BiLSTM approach when compared to the baseline.

Considering the fact that overall, our development set results were much lower when compared to those obtained for the training set and those that were reported in the official baseline for the test set led us to inspect the individual results of the Pearson correlation coefficient r for each session of the development set (Table V, top line). The sessions showing less regular patterns corresponded to much lower values of r, and were therefore excluded from the reduced development set, $dev2$ (marked in bold). As expected, average results are considerably higher for this dataset (absolute improvement of .2). Additional models were also trained, combining $train$ with dev and $dev2$. Our best models were submitted to $test$. An example of the performance of the systems is illustrated in the top plot of Figure 5, showing original and predicted breath signals.

C. Results on the Augmented dataset

The results on the augmented dataset, also presented in Table VI, do not show consistent differences in performance when compared to the challenge dataset. The results on the VoIP-modified sessions are presented in Table V (bottom row), showing no significant differences either, which indicates that there is no information loss regarding breathing events when passing speech signals through the G.723.1 audio codec.

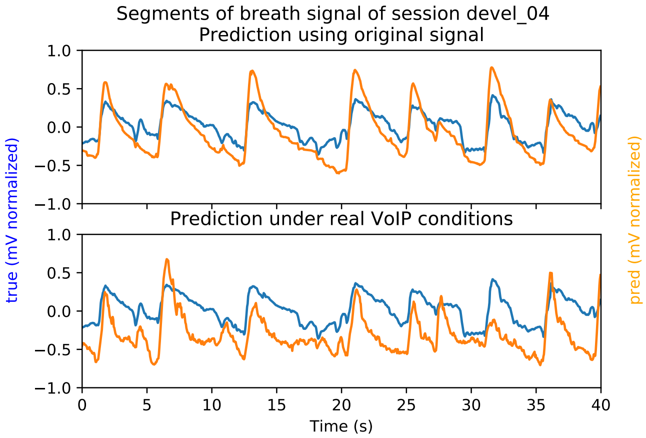

Fig. 5. Segments of breath signals from session *devel_04*. Reference breath signal in blue, predicted signal in orange; above with the original signal, bottom under VoIP conditions.

The bottom part of Figure 5 illustrates the system's ability to correctly predict breathing patterns in VoIP conditions. The true breathing signal is compared with the one predicted from a signal obtained by passing a session of the UCL dataset through a real VoIP scenario. The audio recording is transmitted over-the-air using a mobile phone and recorded using Skype platform, which uses the SILK [46] audio compression and codec.

1) AM-FM decomposition: The rationale behind the AM-FM decomposition is that speech is generated by a source (FM component containing speaker information), which is modulated by the vocal tract (AM component containing the message) [42]. Previous work [47] conducting AM-FM decomposition have shown only a small loss in performance (4.8% WER absolute increase) when using the AM component in an HMM-GMM ASR system. This contrasted with the WER obtained using only the FM component (43.8% absolute increase).

The spectrograms of Figure 6 illustrate the contents of the two components in the presence of a breathing event. The FM carrier signal clearly shows a breath signal between two words whose voicing patterns are visible. The AM signal containing the linguistic information exhibits longer pauses between the corresponding words. This was the motivation for a set of experiments on predicting breath signals from the raw time wave representation of the envelope, the carrier, or combinations of these with and without the original signal.

The AM-FM decomposition is conducted using a frequency domain linear prediction (FDLP) approach. FDLP proposes to model the speech in critical bands as a modulated signal with the AM component obtained using Hilbert envelope estimate and the FM component obtained from the Hilbert carrier. In the implementation followed [48], the input speech was decomposed into 32 conventional quadrature mirror filter (QMF) bands with an analysis window of 1 second. FDLP was then applied on each band to model the sub-band temporal envelopes (AM components). The LP residual represents the FM in the sub-band signal. The reconstruction of the signal from the QMF bands was done by reversing the above-mentioned steps. The resulting envelope signal contains the re-synthesized

TABLE V
PEARSON CORRELATION COEFFICIENT USING OUR BEST REPORTED SYSTEM ON THE CHALLENGE DEVELOPMENT SET.

Session	00	01	02	**03**	**04**	05	**06**	**07**	08	**09**	**10**	**11**	**12**	13	**14**	15
r	.000	.610	.566	.768	.833	.668	.837	.781	.262	.753	.760	.820	.889	.291	.784	.321
r_{aug}	.005	.613	.569	.777	.834	.655	.845	.770	.262	.788	.734	.822	.887	.263	.794	.327

TABLE VI
EXPERIMENTAL RESULTS FOR ALL SYSTEMS

	r		
	dev	$dev2$	$test$
Baseline Approaches - Challenge dataset			
openSMILE [44]	.244	-	.442
openXBOW [45]	.226	-	.366
End2End	.507	.769	.731
Proposed Approaches - Challenge Dataset			
End2End FM	.442	.657	-
End2End AM	.490	.722	-
BiLSTM Original	.507	.787	.720
BiLSTM FM	.441	.696	-
BiLSTM AM	.500	.742	-
End2End Org+AM+FM	.476	.749	-
Proposed Approaches - Augmented Dataset			
	dev_{aug}	$dev2_{aug}$	$test$
End2End Original	.509	.784	-
End2End FM	.424	.621	-
End2End AM	.482	.740	-
BiLSTM Original	.514	.767	.728
BiLSTM FM	.432	.657	-
BiLSTM AM	.515	.755	-
End2End Org+AM+FM	.500	.742	-
BiLSTM Org+AM+FM	.506	.765	-
BiLSTM AM+FM	.488	.744	-

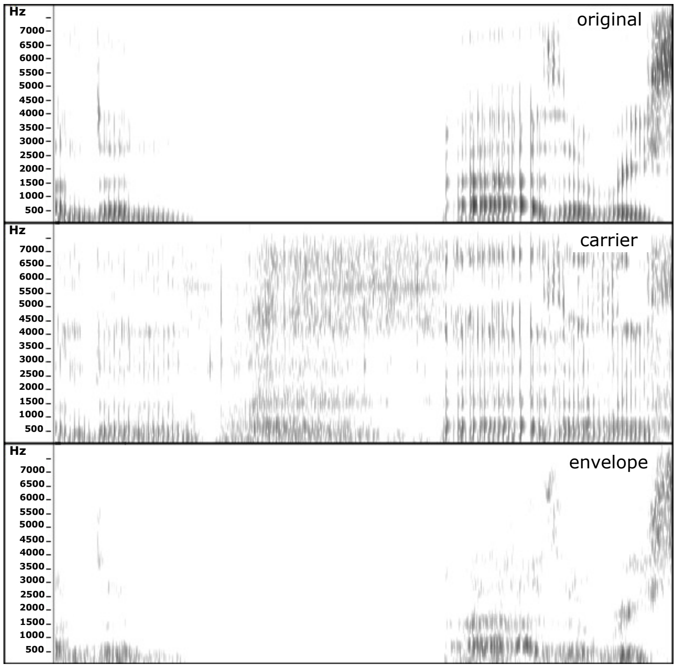

Fig. 6. Spectrograms of speech signal showing a breathing event in between two words.

signal with the intact message, but with whispered speech. With the carrier information alone, the synthesized signal sounds message-less, but with identifiable speaker cues, namely pitch and voice quality features, such as creakiness.

2) Results with AM and FM components: Compared with the results of the original signal, as seen in Table VI, no improvements were detected when using only the carrier or the envelope signal (the performance gain of the BiLSTM AM model when compared to the BiLSTM Original is residual). Furthermore, all experiments indicate the performance using only the AM signal yield the best results when compared to the FM signal. This can be explained by the fact that the AM component retains most of the information relevant for detecting breathing patterns, which is the message. The performance degradation on the AM component, when compared to the original signal, can be explained by the fact that relevant information is carried by the Hilbert FM carrier instead, such as voiced breathing events, that appear on the envelope as silence.

The combination of the AM and FM components, or even when including the original speech signal, failed to outperform the BiLSTM system with the original audio, and the challenge's baseline. This indicates that the availability of the various representations during training does not improve results.

D. Estimation of Breathing Rate

Breathing events are characterized in the breathing signal as a peak value (local maxima), as shown in Figure 7. Previous attempts to detect these events typically include the detection of zero-crossings and thresholding of the signal (using its first and second derivatives) [38] [49]. In this work, we used a slightly different approach: Considering breath is a quasi-periodic signal (the typical respiratory rate for a healthy adult at rest is 12–18 breaths per minute [50]), the resulting cyclic characteristics of the auto-correlation will be equal to the original signal. As such, the peaks of the auto-correlation are found and the average time differences between them report the short period of the signal, which roughly corresponds to the periodicity of breath. This period will then be used as the stride of a window that will detect the local maxima of the original signal.

The *findpeaks* detection algorithm of *MATLAB ver. R2019a* was used to detect both the peaks in the auto-correlation and the breath signal. The obtained short period of the auto-correlation was then used for minimum peak separation in the breath signal. A peak detection threshold of 0.1 mV was added to filter out noise. The corresponding breathing rate is then calculated by dividing the number of detected breath events by the duration of the signal in seconds. An example of this detection is illustrated in Figure 7.

The behaviour of the breathing patterns of the AM and FM components was compared to a breathing event detection algorithm based on an ASR system. This system was trained on the English HUB-4 dataset using Kaldi [27]. The acoustic model is a TDNN and the language model was trained on a mix of broadcast transcriptions and web news corpora [51]. An example of the output is shown in Figure 8. This segment was chosen in particular as it shows the limitations of the use of the speaker noise event detection for

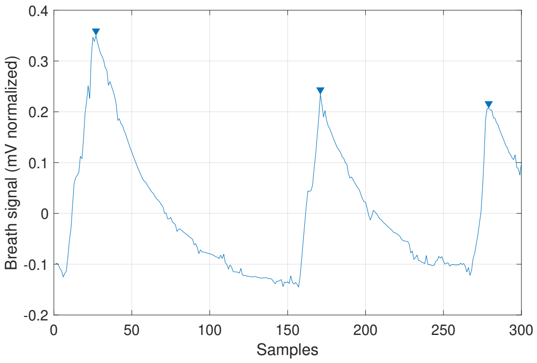

Fig. 7. Sample of a breathing signal. The automatically identified peaks indicate maximum intake of air during inspiration.

breathing detection. We note that by using the generic labels the system is unable to differentiate between voiced exhalation and voiced inhalation and that it does not detect unvoiced inhalation. Furthermore, the system trained with the FM component is unable to detect these voiced exhalations.

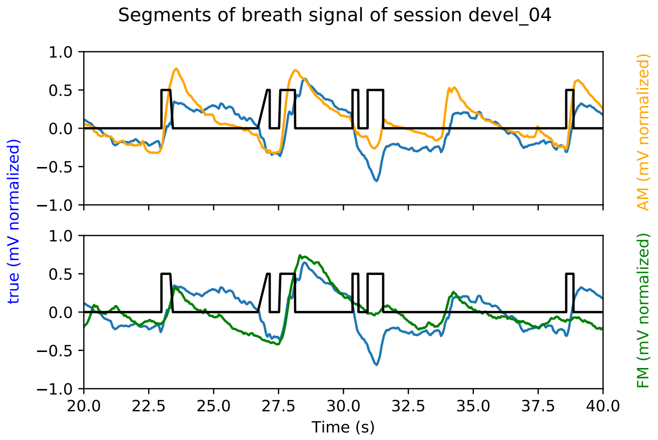

Fig. 8. Segments of true and predicted breath signals with breathing detection algorithm using ASR (in black).

E. Results

The breathing rate estimation results are shown in Figure 9. Considering no actual breathing rates were provided for each session, the results obtained from the predicted signals are compared against the breathing rate estimations of the true signals. The breathing rates for the test set are also provided.

We note that the range of values of breathing rate for the labels is much higher than the ones estimated using the predicted breath signal. Additionally, the presence of outliers in the true signals is much more spread apart when compared to the predicted signals, which indicates some of the sessions have noisy or otherwise disrupted breath signals. While this had already been shown for the development set, the data presented here shows that some sessions of the training data also share the same problem.

Rates of under 0.2 were reported in [40] [49], for conversational speech, which is in agreement with the results obtained from the predicted signals. A Mean Absolute Error of 0.0664 and 0.1232 was obtained on training and dev sets, respectively.

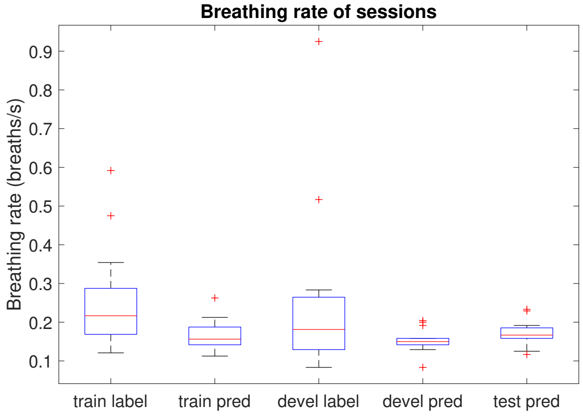

Fig. 9. Average breathing rates (breaths per second) for the different datasets. The reported distributions of the predictions were obtained using our best model in $dev2_{aug}$.

IV. Conclusions

In this work we discussed the employment of voice profile metadata for speech corpora and how an automatic system that is able to detect such metadata can assist in the development of machine learning models. This paper demonstrates that the use of extracted speaker embeddings can provide the needed crowdsourcing submission control by providing a single-dimensional vector capable of verifying speakers and their gender. Additionally, it also shows that voice pattern recognition techniques can be used to predict breathing patterns and breathing-related parameters.

We first presented a speaker verification task in the context of quality control for crowdsourced speech data collections . Noting the various combinations of different languages and conditions that occur during data collection, our proposed speaker verification system is pre-trained on an out-of-domain dataset and adapted to each dataset automatically. Evaluation results on crowdsourced datasets indicate an EER with or without score normalisation within the values of other speaker verification benchmarks. The possibility of using a single DT enables the deployment of an online fraud detection system.

We then analyzed and automatically predicted breathing patterns from speech, using signals extracted from respiratory belts as ground truth. Moreover, we studied the applicability of the AM-FM decomposition of speech to this same task. We found that while the decomposed components did not surpass the performance of the original signal, our experiments support the hypothesis that the breathing rate is dependent on the message, since, individually, the results obtained with the AM component were able to outperform those obtained with just the FM component. In order to simulate the conditions of medical consultations over the internet, the challenge dataset was augmented by passing it through a VoIP coder-decoder. Overall, our experiments also indicate that future information modelled by the Bidirectional LSTM improves results.

For future work, we plan on expanding experiments to include more datasets with different languages, channel conditions and task domains (e.g. free speech). A larger explicit enrolment, or one that uses previous, validated, tasks from other datasets could also improve current results. The use of unsupervised agglomerate clustering, besides also solving the problem of detecting multiple speakers using a single account, may also help detecting speakers using multiple accounts.

A short term future goal in breathing pattern predition is to explore additional parameters that can be extracted from breathing patterns such as volumetric information (e.g. tidal volume). Additionally, given how breathing provides important markers to several medical conditions, such as cardiac, respiratory and neurological diseases, we plan to explore speech derived breathing patterns for assisting in the automatic detection of these conditions.

Acknowledgements

This work was supported by national funds through *Fundação para a Ciência e a Tecnologia* (FCT) with reference UIDB/50021/2020.

References

[1] AppDynamics, "The Future of Voice Technology In the Enterprise," AppDynamics, Tech. Rep., 2018.

[2] S. G. Preeti Wadhwani, "Intelligent Virtual Assistant (IVA)," Global Market Insights, Tech. Rep., 2018.

[3] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Muller, R. Huber, B. Andrassy, J. G. Bauer, and B. Littel, "Comparison of four approaches to age and gender recognition for telephone applications," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 4, 2007, pp. IV–1089–IV–1092.

[4] B. Andreeva, G. Demenko, B. Möbius, F. Zimmerer, J. Jügler, and M. Oleskowicz-Popiel, "Differences of pitch profiles in germanic and slavic languages," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[5] J. E. Flege, "Second language speech learning: Theory, findings, and problems," *Speech perception and linguistic experience: Issues in cross-language research*, vol. 92, pp. 233–277, 1995.

[6] J. Lopes, I. Trancoso, and A. Abad, "A nativeness classifier for ted talks," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5672–5675.

[7] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients," *Journal of Speech and hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[8] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, G. Parker *et al.*, "From joyous to clinically depressed: Mood detection using spontaneous speech." in *FLAIRS Conference*. Citeseer, 2012, pp. 141–146.

[9] R. Singh, *Profiling humans from their voice*. Springer, 2019.

[10] Y. Wen, R. Singh, and B. Raj, "Reconstructing faces from voices," 2019.

[11] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.

[12] T. S. Behrend, D. J. Sharek, A. W. Meade, and E. N. Wiebe, "The viability of crowdsourcing for survey research," *Behavior research methods*, vol. 43, no. 3, p. 800, 2011.

[13] J. M. Rzeszotarski and A. Kittur, "Instrumenting the crowd: using implicit behavioral measures to predict task performance," in *UIST*, 2011.

[14] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "An open-source speaker gender detection framework for monitoring gender equality," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5214–5218.

[15] 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 *Proc. Interspeech 2018*, 2018, pp. 277–281.

[16] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[17] N. Brummer, A. Mccree, S. Shum, D. Garcia-Romero, and C. Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Odyssey 2014*, 2014, pp. 260–264.

[18] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[19] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition," in *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2012, pp. 4257–4260.

[20] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 726–733.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Interspeech 2018*, Sep 2018.

[22] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Interspeech 2017*, Aug 2017.

[23] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[24] D. A. Reynolds, "The effects of handset variability on speaker recognition performance: experiments on the switchboard corpus," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1, 1996, pp. 113–116 vol. 1.

[25] S. Cumani, P. Batzu, D. Colibro, C. Vair, P. Laface, and V. Vasilakakis, "Comparison of speaker recognition approaches for real applications." 01 2011, pp. 2365–2368.

[26] P. Matějka, O. Novotný, O. Plchot, L. Burget, M. D. Sánchez, and J. Černocký, "Analysis of score normalization in multilingual speaker recognition," in *Proc. Interspeech 2017*, 2017, pp. 1567–1571.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[28] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 421–425.

[31] P. Lieberman, S. Fecteau, H. Théoret, R. R. Garcia, F. Aboitiz, A. MacLarnon, R. Melrose, T. Riede, I. Tattersall, and P. Lieberman, "The evolution of human speech: Its anatomical and neural bases," *Current anthropology*, vol. 48, no. 1, pp. 39–66, 2007.

[32] C. G. Gallagher and M. Younes, "Breathing pattern during and after maximal exercise in patients with chronic obstructive lung disease, interstitial lung disease, and cardiac disease, and in normal subjects," *American Review of Respiratory Disease*, vol. 133, no. 4, pp. 581–586, 1986.

[33] J. A. Hirsch and B. Bishop, "Respiratory sinus arrhythmia in humans: how breathing pattern modulates heart rate," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 241, no. 4, pp. H620–H629, 1981.

[34] I. Homma and Y. Masaoka, "Breathing rhythms and emotions," *Experimental Physiology*, vol. 93, no. 9, pp. 1011–1021, 2008.

[35] J. Hlavnička, R. Čmejla, T. Tykalová, K. Šonka, E. Růžička, and J. Rusz, "Automated analysis of connected speech reveals early biomarkers of parkinson's disease in patients with rapid eye movement sleep behaviour disorder," *Scientific reports*, vol. 7, no. 1, pp. 1–13, 2017.

[36] A. Rochet-Capellan and S. Fuchs, "The interplay of linguistic structure and breathing in German spontaneous speech," in *14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, Lyon, France, Aug. 2013, p. 1228.

[37] K. Konno and J. Mead, "Measurement of the separate volume changes of rib cage and abdomen during breathing," *Journal of applied physiology*, vol. 22, no. 3, pp. 407–422, 1967.

[38] J. Korten and G. Haddad, "Respiratory waveform pattern recognition using digital techniques," *Computers in biology and medicine*, vol. 19, no. 4, pp. 207–217, 1989.

[39] Y. Cho, N. Bianchi-Berthouze, and S. J. Julier, "Deepbreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 456–463.

[40] V. S. Nallanthighal, A. Härmä, and H. Strik, "Deep Sensing of Breathing Signal During Conversational Speech," in *Proc. Interspeech 2019*, 2019, pp. 4110–4114.

[41] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly emotion, Breathing & Masks," in *Proceedings of Interspeech*, Shanghai, China, September 2020, p. 5 pages, to appear.

[42] H. Dudley, "The carrier nature of speech," *Bell System Technical Journal*, vol. 19, no. 4, pp. 495–515, 1940.

[43] P. Kabal, "ITU-T G. 723.1 speech coder: A MATLAB implementation," 2004.

[44] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 835–838.

[45] M. Schmitt and B. Schuller, "OpenXBOW: Introducing the passau open-source crossmodal bag-of-words toolkit," *J. Mach. Learn. Res.*, vol. 18, no. 1, p. 3370–3374, Jan. 2017.

[46] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, "High-quality, low-delay music coding in the opus codec," *arXiv preprint arXiv:1602.04845*, 2016.

[47] P. Motlicek, H. Hermansky, S. Madikeri, A. Prasad, and S. Ganapathy, "AM-FM decomposition of speech signal: Applications for speech privacy and diagnosis," Idiap, Rue Marconi 19, Idiap-RR Idiap-RR-01-2020, 1 2020.

[48] S. Ganapathy, P. Motlicek, and H. Hermansky, "Autoregressive models of amplitude modulations in audio compression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1624–1631, 2010.

[49] S. Fuchs, U. D. Reichel, and A. Rochet-Capellan, "Changes in speech and breathing rate while speaking and biking," in *ICPhS*, 2015.

[50] S. Fleming, M. Thompson, R. Stevens, C. Heneghan, A. Plüddemann, I. Maconochie, L. Tarassenko, and D. Mant, "Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies," *The Lancet*, vol. 377, no. 9770, pp. 1011–1018, 2011.

[51] A. Abad, P. Bell, A. Carmantini, and S. Renais, "Cross lingual transfer learning for zero-resource domain adaptation," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.