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

Analyzing Breath Signals for the Interspeech 2020 ComParE Challenge

John Mendonça[1,2], *Francisco Teixeira*[1,2], *Isabel Trancoso*[1,2], *Alberto Abad*[1,2]

[1]INESC-ID, Portugal
[2]Instituto Superior Técnico, Universidade de Lisboa, Portugal
{john.mendonca, francisco.s.teixeira, isabel.trancoso, alberto.abad}@tecnico.ulisboa.pt

Abstract

This paper presents our contribution to the INTERSPEECH 2020 Breathing Sub-challenge. Besides fulfilling the main goal of the challenge, which involves the automatic prediction from conversational speech of the breath signals obtained from respiratory belts, we also analyse both original and predicted signals in an attempt to overcome the main pitfalls of the proposed systems. In particular, we identify the subsets of most irregular belt signals which yield the worst performance, measured by the Pearson correlation coefficient, and show how they affect the results that were obtained by both the baseline end-to-end system and variants such as a Bidirectional LSTM. The performance of this type of architecture indicates that future information is also relevant when predicting breathing patterns.

We also study the information retained from the AM-FM decomposition of the speech signal for this purpose, showing how the AM component significantly outperforms the FM component on all experiments, but fails to surpass the prediction results obtained using the original speech signal.

Finally, we validate the system's performance in video-conferencing conditions by using data augmentation and compare clinically relevant parameters, such as breathing rate, from both the original belt signals and the ones predicted from the simulated video-conferencing signals.

Index Terms: Breath Detection, ComParE, Paralinguistics

1. Introduction

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 [1]. 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 [2] [3] or to evaluate cognitive and neurological health [4][5]. Furthermore, information extracted from breathing patterns during speech can be used to assist speech therapists in identifying speech impediments resulting from unfavourable respiratory planning [6]. Breathing monitoring in this context is often conducted using wearable sensors, namely, face masks and/or respiratory belts [7]. 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. [8], [9]). In [10], 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 [11], 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 [12], 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.

This paper is organized as follows: Section 2 describes the datasets employed. Section 3 introduces the methodology used for the experiments. In this section, results are analyzed for the baseline and AM-FM decomposition experiments, as well as for the augmented dataset. Section 4 presents methodology and results of breathing rate estimation.Section 5 draws conclusions and presents directions for future work.

2. Datasets

The experiments for the Breathing Sub-challenge [11] 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).

http://dx.doi.org/10.21437/Interspeech.2020-2778

An analysis of the belt signals in these datasets shows considerable variability, as illustrated in Figure 1: 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.

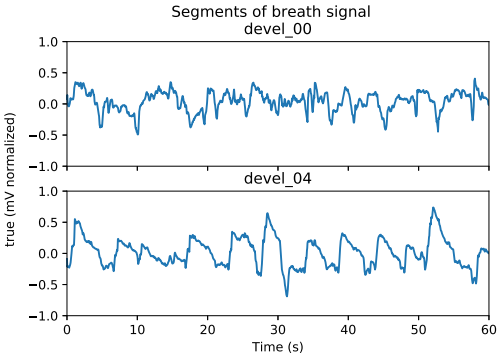

Figure 1: *Segments of breath signals from sessions 00 and 04.*

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 [13]. 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 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}).

3. Prediction of Breathing Patterns

3.1. Model Architectures

The official provided end-to-end baseline architecture was used as a base for all experiments[1]. 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.

[1]https://github.com/glam-imperial/ComParE2020-Breathing-End2End

3.2. 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 2. 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 1, top line). The sessions showing less regular patterns corresponded to much lower values of r, and were therefore excluded from the reduced development set, $dev2$. 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 2, showing original and predicted breath signals.

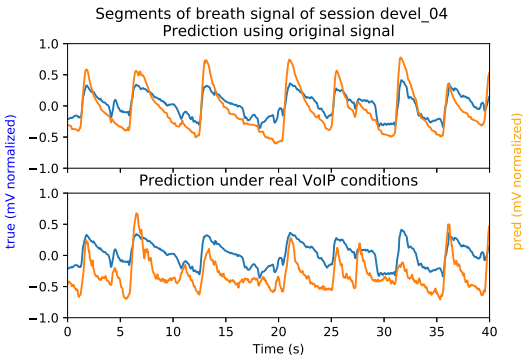

Figure 2: *Segments of breath signals from session devel_04. Reference breath signal in blue, predicted signal in orange; above with the original signal, bottom under VoIP conditions.*

3.3. Results on the Augmented dataset

The results on the augmented dataset, also presented in Table 2, do not show consistent differences in performance when compared to the challenge dataset. The results on the VoIP-modified sessions are presented in Table 1 (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.

The bottom part of Figure 2 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 [16] audio compression and codec.

3.4. 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) [12]. Previous work [17] 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

Table 1: *Pearson correlation coefficient using our best reported system on the challenge development set. Top line shows results on this set and bottom line on the augmented set. The 9 sessions included in the reduced development set, $dev2$, are marked in bold.*

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 2: *Experimental Results for all systems on the Breathing Sub-challenge*

	r		
	dev	$dev2$	$test$
Baseline Approaches - Challenge dataset			
openSMILE [14]	.244	-	.442
openXBOW [15]	.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	-

with the WER obtained using only the FM component (43.8% absolute increase).

The spectrograms of Figure 3 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 [18] [2], 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 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.

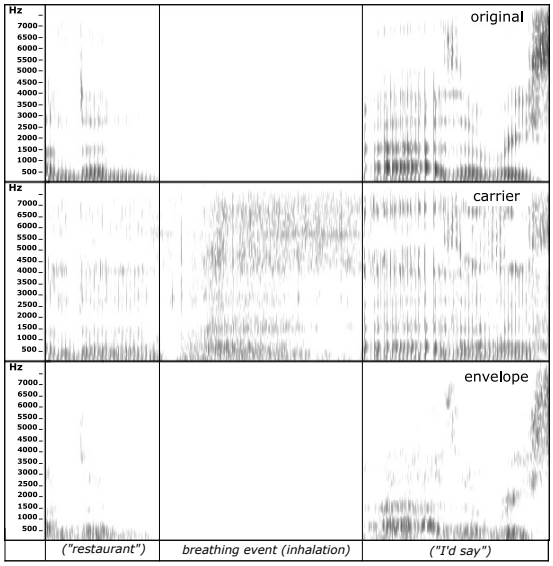

Figure 3: *Spectrograms of speech signal showing a breathing event in between two words.*

3.5. Results with AM and FM components

Compared with the results of the original signal, as seen in Table 2, 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.

4. Estimation of Breathing Rate

Breathing events are characterized in the breathing signal as a peak value (local maxima), as shown in Figure 4. Previous attempts to detect these events typically include the detection of zero-crossings and thresholding of the signal (using its first and second derivatives) [8] [19]. In this work, we used a slightly different approach: Considering breath is a quasi-periodic signal

[2] https://github.com/iiscleap/SignalAnalysisUsingAm-FM

(the typical respiratory rate for a healthy adult at rest is 12–18 breaths per minute [20]), 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 4.

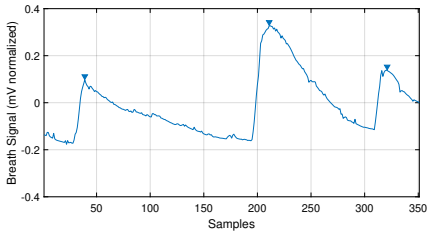

Figure 4: *Sample of a breathing signal. The automatically identified peaks indicate maximum intake of air during inspiration.*

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 [21]. The acoustic model is a TDNN and the language model was trained on a mix of broadcast transcriptions and web news corpora [22]. An example of the output is shown in Figure 5. This segment was chosen in particular as it shows the limitations of the use of the speaker noise event detection for 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.

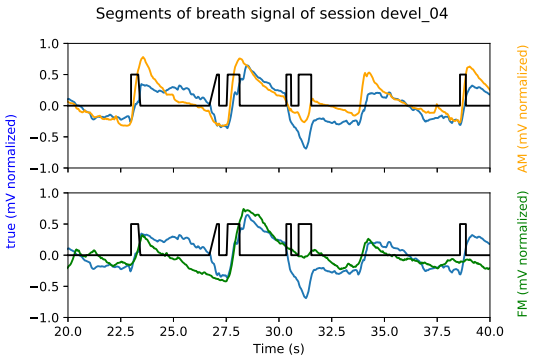

Figure 5: *Segments of true and predicted breath signals with breathing detection algorithm using ASR (in black).*

4.1. Results

The breathing rate estimation results are shown in Figure 6. 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 [10] [19], 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.

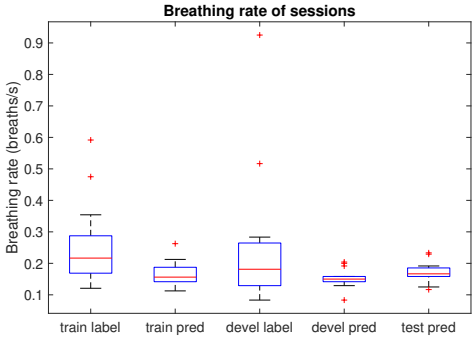

Figure 6: *Average breathing rates (breaths per second) for the different datasets. The reported distributions of the predictions were obtained using our best model in* $dev2_{aug}$.

5. Conclusions

In this work we 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.

A short term future goal 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.

6. Acknowledgements

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

7. References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

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

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.