Deep attentive end-to-end continuous breath sensing from speech

Alexis Deighton MacIntyre, Georgios Rizos, Anton Batliner, Alice Baird, Shahin Amiriparian, Antonia Hamilton, Björn W. Schuller

Deep Attentive End-to-End Continuous Breath Sensing from Speech

Alexis Deighton MacIntyre[1]*, *Georgios Rizos*[2]*, *Anton Batliner*[3], *Alice Baird*[3],
Shahin Amiriparian[3], *Antonia Hamilton*[1], *Björn W. Schuller*[2,3]

[1]Institute of Cognitive Neuroscience, University College London, UK
[2]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[3]EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

`a.macintyre.17@ucl.ac.uk, georgios.rizos12@imperial.ac.uk`

Abstract

Modelling of the breath signal is of high interest to both healthcare professionals and computer scientists, as a source of diagnosis-related information, or a means for curating higher quality datasets in speech analysis research. The formation of a breath signal gold standard is, however, not a straightforward task, as it requires specialised equipment, human annotation budget, and even then, it corresponds to lab recording settings, that are not reproducible in-the-wild. Herein, we explore deep learning based methodologies, as an automatic way to predict a continuous-time breath signal by solely analysing spontaneous speech. We address two task formulations, those of continuous-valued signal prediction, as well as inhalation event prediction, that are of great use in various healthcare and Automatic Speech Recognition applications, and showcase results that outperform current baselines. Most importantly, we also perform an initial exploration into explaining which parts of the input audio signal are important with respect to the prediction.

Index Terms: breath prediction from speech, end-to-end deep learning, neural attention, biological signal monitoring

1. Introduction

Breathing patterns provide medical doctors and speech therapists with vital information about an individual's physical health state, as well as insight into human affective states [1, 2] and cognitive and neurological circumstances [3, 4] more broadly. In the case of speech, respiratory activity reflects important motor planning processes [5]. The ability to rapidly and flexibly sequence chest movements to produce speech breathing is considered unique to modern humans (and potentially neandertals) [6], and the loss of fine respiratory control during vocalisation can be an early and acute symptom of neurodegenerative motor disorders, such as Parkinson's disease (PD) [7, 8]. Recent work, for example, demonstrates that patients with PD are more likely to breathe between syntactic boundaries [9] than neurotypical speakers, and that both patients and their healthy ageing counterparts need to breathe more often than younger adults to meet metabolic demands during speech [10].

Although the latter work was focused on the sentence-structural aspects of speech breathing, relatively little is known of the sub-second temporal dynamics of the breath signal, in part because respiratory recording is relatively intrusive and the manual annotation of breathing patterns is laborious, resulting in a high cost to benefit ratio of data collection. For example, the authors of the study performed in [1] identified and annotated the breath events manually by listening to the speech recordings. This approach is certainly limited to the research domain,

is not scalable, and is dependent on variable human annotator skill, as well as being vulnerable to bias. Established methods for measuring respiratory activity include the application of electromyography [11], chest pneumograph [12], and inductive plethysmography, which entails subjects being fitted with an elastic belt that shrinks or expands with breathing movements [13]. These devices require direct application to the human body, which could affect the natural and spontaneous expression of the speaker [1]. A possible non-contact method would be the usage of a thermal camera [14], which is, however, a less cost-effective approach, and furthermore also requires human annotation. Computational methods that automatically detect breathing events purely by analysing recorded speech should facilitate the aforementioned healthcare applications [15, 13]. Although 'laboratory speech' (e. g., formulaic texts that are read aloud) forms the basis of much linguistic and speech sciences research [16], a deeper challenge arises in the case of **spontaneous speech**, where greater cognitive effort is required in comparison to reading, as well as the adaptation of speech rhythm to accommodate breathing [17, 5].

It is apparent then that there is great importance in **breath sensing of spontaneous speech**, so we focus on developing a computational, deep-learning methodology for transforming speech into breath signal. The breath signal we use to form our ground-truth is recorded by elastic piezoelectric belts worn by the participants. We focus on two different task formulations of the speech-based, breath sensing problem: a) Predicting the **continuous-time, real-valued breath belt signal (BELT)**, and b) detecting **maximum inhalation events (MAX)**. The former task works as a proof of concept towards the development of a breath sensing solution that does not require specialised equipment, other than a microphone. The latter task is of more interest in studies that are based on the detection and localisation of inhalation events, either for removal [15, 18, 19, 4, 20, 21, 22], or further processing [23, 24, 25, 26] (see Section 2).

Herein, we propose two computational improvements for breath sensing: a) the eschewing of the feature extraction step, in favour of a fully end-to-end approach [27], which involves learning to extract features from the speech waveform using Convolutional Neural Networks (CNNs), b) the application of attention mechanisms [28] for enhanced performance, as well as a gateway towards continuous-time interpretation through the analysis of the attention weights (see Figure 2) according to the high standards of recent research [29], and c) the application thereof on two tasks with real-world applications.

2. Related work

In the study performed in [15], the authors have proposed a methodology for identifying and removing breath sound seg-
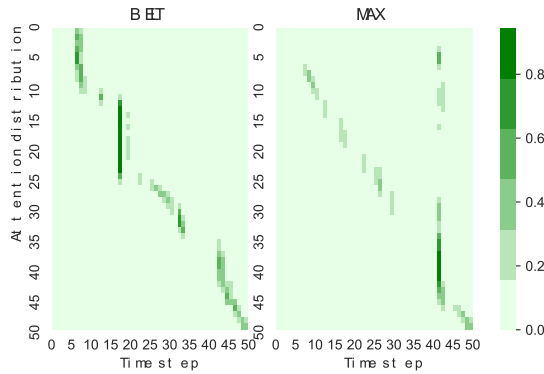

Figure 1: *Attention maps for the same 2 second segment (50 time steps), averaged across 10 trials. **Left** corresponds to the weights learnt for the BELT task, and **Right** for the MAX task. Deviations from a diagonal indicate unequal importance allocation to all sequence elements. Specifically, for the MAX task, we observe that a lot of weight is placed at an element within the max inhalation event, as this is the minority class of interest.*

Figure 2: *The attention importance weights (red) plotted against the breath signal (teal) for both the BELT task (upper), and the MAX task (middle). The attention curve is an average across 10 trials, followed by standardisation. We also show the speech waveform volume (gray, lower) for reference.*

ments in speech recordings. This is done by utilising recorded breath examples to train a breath sound template based on the extraction and analysis of Mel Frequency Cepstral Coefficients (MFCCs), and then using the latter within a pattern matching framework on longer songs and narrated speech recordings. In a recent extension on this method that was applied on a database of news reports, a Voice Activity Detection (VAD) step is also applied such that the template matching is only focused on non-speech segments [21]. Elaborate post-processing, requiring domain knowledge, was also used for unifying closely spaced breath segments, as well as discarding small breaths. The goal of such studies is to *identify and remove* sharp inhalation sounds towards the improvement of the recording quality.

This approach has also been used for the removal of breath segments in order to *curate* a clean speech corpus for speech synthesis [22], Automatic Speech Recognition (ASR) in Japanese [20], as well as classification of speech as being produced by subjects with schizophrenia [18], breathing problems due to lung cancer [19], or rapid eye movement sleep behaviour disorder and PD [4]. On the other hand, *acoustic analysis of the breathing sounds themselves* has also been applied for a variety of problems, such as improving speaker recognition [24, 25], and detecting obstructive sleep apnea [26] or major respiratory diseases (i. e., flu, pneumonia, and bronchitis) [23]. Our study is not only useful in detecting inhalation events, which is of great interest to the aforementioned studies, but at predicting continuous-time breath signals, towards the provision of a *more holistic understanding* of the breathing activity of a subject.

The study that is closest to our BELT task is the one performed in [13], as it utilises deep learning for continuous-time breath signal regression, as recorded from elastic breath belts. In this paper, the authors employ an approach that consists of an MFCC feature extraction step, followed by the application of a stacked Recurrent Neural Network (RNN) model on speech segment sizes between 4 and 8 seconds. We instead follow our work in [30], and consider uninterrupted spontaneous speech segments of 4 minutes, the predictive modelling of which we improve, and *take first steps towards interpreting the temporal patterns that are potentially informative in breath prediction*.
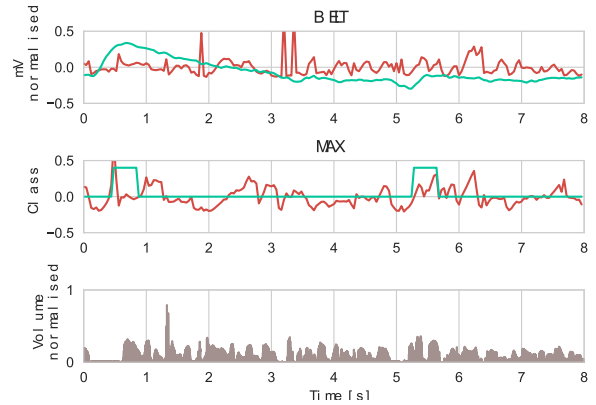
We additionally experiment with a categorical, continuous-time prediction task, i. e., max-breath event detection. The authors of [22] extracted speech segments from a podcast conversation, annotated as being either speech or breath produced by one of two speakers, and applied a classification mechanism based on [31] in a multi-class framework. In contrast, we treat the max breath event detection task as *time-continuous*, where a sequence element is either within an event, or not, and furthermore focus on the more general task of *spontaneous speech* with *speaker-independent* partitioning.

2.1. Attention for Explanation?

Attention mechanisms [28, 32] have extensively been used to provide a dimension of explainability as to what the model believes is an important part of the input, however, they have recently received criticism on that account [33] for generating inconsistent attentive explanations, for example across different trials. More recently, the authors of [29] have addressed the criticisms by claiming that the existence of alternative explanations is not indicative of lack of explanatory power, as there may be *multiple explanations for the predictions of a model*, something also indicated by the success of models that utilise multi-head self-attention [34]. They further quantify this attention distribution variability by using the Jensen-Shannon divergence. In all our visualisations of attention maps, weights, and discussions thereupon, we utilise *averages across multiple trials*, and we further report *attention distribution correlations*.

3. From speech to breath

In the context of this study, we denote by $x_i \in \mathbb{R}^{T^x \times d^x}$ the i-th sample utterance, regardless of the model to be used. Each input sample is sequential, with T^x being the number of time steps for the inputs, and d the dimensionality. For example, if we are working on the raw audio waveform, d corresponds to 1, whereas if we are working on MFCCs, d^x, T^x correspond to the number of MFCCs, and the length of the spectrogram, respectively. We denote our model by M, that receives x_i and outputs the corresponding prediction $y_i \in \mathbb{R}^{T^y \times d^y}$. The number of time steps for our label sequences for both considered tasks is denoted by T^y. In the continuous-valued breath signal prediction

Table 1: *Summary of the gender balanced (F – Female, M – Male), speaker independent UCL-SBM database partitions.*

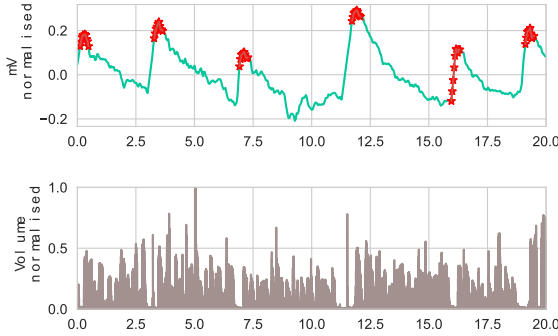| #	Train	Dev	Test	Σ
F	9	10	10	29
M	8	6	6	20
Σ	17	16	16	49

Figure 3: *The upper plot depicts the continuous-valued breath signal (teal), along with the parts that we assume constitute maximum breath events (starred red). The lower plot depicts the speech waveform volume (gray) for reference.*

d^y equals to 1, signifying the prediction of the continuous breath belt signal. As for the max-breath event detection, the model outputs 2 logits per time step, one per class.

3.1. Breath sensing task formulation

Breath belt signal prediction (BELT): In this sequential regression problem, the task is to predict a signal (in mV) that resembles the piezoelectric breath belt output. We thus train a model that is able to provide a breath belt-like signal to be used in absence of the required instruments, as a proxy for them. We evaluate this task with the Pearson Correlation Coefficient (r).

Max-breath event detection task (MAX): Maximum belt extension events have been localised in time using *peak detection*, as proxies of maximum air volume inhalation events. The task would be the accurate detection of max-breath events along time. We detected peaks with topographic prominence more than 0.1, and assumed that a time window of 440 ms centred around the peak corresponds to a max-breath event. The evaluation measure used here is the Macro averaged F1 score (Macro-F1). The prediction labels of the two tasks are depicted in Figure 3, alongside the speech waveform, for reference.

3.2. Attentive end-to-end deep learning

Both the BELT and the MAX tasks are sequential in nature. Thus, we adopt a common two-layer stack of Long Short-Term Memory (LSTM) RNNs to be applied on the various feature baselines, where each layer has 256 hidden units. For our proposed method, we engage an end-to-end training method from the raw speech waveform [27] by the usage of a stacked CNN. We use three stacked CNN layers with 64, 128, and 256 hidden units, respectively; each layer was followed by a max-pooling operation, with corresponding rates 10-8-8. We then segment the hidden state sequence into sub-segments of 2 seconds (i. e.,

50 time steps), and apply to each of them in such a localised manner the Luong dot product attention [32]. This produces a 50×50 map, where each element is the corresponding dot product. A 1-dimensional softmax function is used to produce an *attention map* comprising 50 discrete probability distributions of size 50, one per hidden state, where the probabilities signify the relation of another hidden state towards the one under examination. Each probability distribution is used to perform a weighted average of the hidden state sequence per time step. Finally, a fully connected layer is applied to the new sequence to extract the final numerical prediction, or logit. In Figure 1, we show attention maps for the same 2 second segment, for both tasks. By adding all the probabilities related to a hidden state found on all distributions, we get a measure of importance for this hidden state, hereby **an importance weight**. In Figure 2, we show for both tasks the attention weights plotted against the labels for an 8 second segment. Both attention maps and importance weights are averaged across 10 trials.

4. Speech-breath database

We utilise the **UCL-SBM** database, which was the basis of the Breathing Sub-challenge of Interspeech 2020 ComParE [30]. The partitions are summarised in Table 1. Here, we use only spontaneous speech recordings that pose a greater challenge in terms of respiratory planning [5], and data from one of the two piezoelectric respiratory belts worn by the subjects. The belt was positioned approximately four centimetres below the collarbone to record chest breathing, and produces a linear voltage reading in response to changes in thoracic circumference associated with respiration. All signals were sampled at 40 kHz; speech was downsampled to 16 kHz and breath belts to 25 Hz in post-processing. The breath signal was further normalised by dividing each value by the maximum recorded value across the dataset. All 49 speakers[1] (29 f, 20 m) reported English as a primary language, but ranged in regional accent (e. g., American, Irish, etc.), as well as sociolect; ages range from 18 to approximately 55 years old (mean age 24 years; std. dev. = ~10 years). Each participant contributed five minutes of spontaneously generated speech, following instructions to imagine having a conversation with a new acquaintance in a polite, yet informal situation. The recordings were edited at the four minute mark to a common duration for conformity, as well as to avoid background noise or the experimenter's instructions.

5. Experiments

5.1. Baselines

We apply the common LSTM RNN architecture described in Sub-section 3.2 on all the following baseline feature sets, both with the attention step (denoted by **ATT**) and without. r was optimised directly using Adam [35] with initial learning rate .001. We ran our experiments for 100 epochs, validating every 5, and report test measures using the model that yields the best validation performance. We execute 10 trials of each method.

WAV – Raw Audio Waveform: In this case, we utilise an additional stacked CNN (described in Subsection 3.2) model at the beginning that processes the raw waveform and learns to extract shift-invariant features in an End-to-End manner [27, 30].

---

[1]Participants were recruited via word of mouth and an online psychology subject pool database. Informed written consent was obtained prior to testing, and the project received approval from the UCL research ethics committee.

Table 2: *Test performance results for the UCL-SBM database.*

| Method | BELT (r) | | MAX (Macro-F1 %) | |
	No ATT	ATT	No ATT	ATT
ComParE+RNN	.721	.712	74.643	74.721
MFCC+RNN	.721	.730	72.818	74.148
WAV+CRNN	.728	.731	74.743	75.469

MFCC – Mel Frequency Cepstral Coefficients: We calculate 80 MFCCs using a 25 ms raised cosine Fast Fourier window, with 10 ms stride. An RNN processing MFCCs corresponding to a 4-8 second segment was the approach recently utilised in [13], without attention. MFCCs were also the features of choice in [22], albeit in a non speaker independent, utterance-level prediction task.

COMPARE – Low-level Descriptor Acoustic Feature Set: 65 COMPARE feature set low-level descriptors (LLDs) were extracted at a 40 ms hop size, as well as their first derivation (delta), resulting in a 130 dimensional LLD feature set. A full description of the feature set can be found in [36].

5.2. Predictive performance results

Comparison results are summarised in Table 2. The COMPARE features, but mostly the utilisation of end-to-end learning yield some computational improvement over the MFCC baseline, especially for the MAX task. Furthermore, the utilisation of attention brings an improvement across the board, again more noticeable in the MAX task. We hypothesise this is because it is easier to focus the attention on the minority positive class.

5.3. Discussion – Where does the network attend to?

In Figure 2, there is no easily discernible pattern for attention with respect to the BELT task. This might be due to the continuous valued nature of the task, where many different locations in the sequence are important. For the MAX task, we observe that there is a tendency for high attention weights to concentrate at the beginning of the max breath event (the r between importance weights and the continuous binary MAX labels is a non-trivial 0.141). This makes sense, as the positive class is the minority, and the network learns to properly focus on the corresponding segments. Finally, low importance weights appear to correspond to segments with lack of speech, perhaps a consequence of the network realising that there is no useful signal there. A notable exception, in terms of high attention during the absence of speech, occurs immediately before an inhalation; *this may be attributable to the presence of inhalation sounds.*

Towards a more quantitative examination, attentional weight vectors were resampled to 1 kHz, low-pass filtered at 10 Hz using a 4th order Butterworth filter, and rescaled between -1 and 1. Attentional peaks of peak prominence more than .25 were detected (determined by piloting to strike a balance between humanly-discernible peaks and noise). Inter-peak intervals (IPI) were calculated to ascertain the relative degree of periodicity and any underlying patterns in terms of the temporal structure of attentional weighting. Concerning the end-to-end method, the median BELT attentional IPI was calculated on a speaker-by-speaker basis, with a group average of 1 743.75 (SD 723.4) ms, and an inter-quartile range of 1 991.69 (SD 972.86) ms. For the MAX task, the corresponding group average of median attentional IPI was 1 314.13 (SD 304.74) ms, with an inter-quartile range of 1 684.38 (SD 405.59) ms. This suggests that attention

Table 3: *Cross-trial attention distribution correlations. Based on r, the attention distributions are well correlated across trials, which is an indication that they are neither arbitrary, nor conditional on chance, as hypothesised in [33].*

Method	BELT	MAX
ComParE+RNN	.583	.534
MFCC+RNN	.536	.559
WAV+CRNN	.624	.466

Table 4: *Cross-method attention distribution correlations. The non-trivial correlation scores indicate that regardless of input level feature representation, there exist universal temporal patterns that are of use towards breath prediction.*

Method Pair	BELT	MAX
WAV & MFCC	.626	.553
WAV & ComParE	.462	.242
MFCC & ComParE	.504	.198

in the MAX task operated on a faster and less variable timescale in comparison with attention in the BELT task. In both cases, the data were largely positively skewed, with modal peaks at approximately 940 ms for the BELT task, and 555 for the MAX task, and inter-speaker variability.

5.4. Discussion – Are the attention weights brittle?

We try to quantify the attention distribution distances in a manner inspired by [29]. For each cross-trial pair, we calculate the r values for corresponding attention distributions, in each 2 sec attention segment, and each speaker. We report the averages corresponding to each attentive method and task in Table 3. In Table 4, we perform a similar r calculation, this time across attention distributions by different methods, after trial averaging.

6. Conclusions & future work

We have shown that learning audio features in an end-to-end manner is beneficial towards breath sensing from speech, and that attention mechanisms help identify useful patterns from the speech signal, that persist across choice of method[2]. An important next step is to validate the possibility of cross-corpus breath sensing, thus verifying that our method can stand in place of more specialised measurement equipment. Towards a deeper understanding of the relation between speech and breath, a further exploration of the explanatory potential of attention weights should be attempted via more powerful attentional models [34].

7. Acknowledgements

---

[2]Codebase: `https://github.com/glam-imperial/Deep-Breath-From-Speech`

8. References

[1] F. Goldman-Eisler, "Speech-breathing activity-a measure of tension and affect during interviews," *British Journal of Psychology*, vol. 46, no. 1, p. 53, 1955.

[2] E. Heim, P. H. Knapp, L. Vachon, G. G. Globus, and S. J. Nemetz, "Emotion, breathing and speech," *Journal of Psychosomatic Research*, vol. 12, no. 4, pp. 261–274, 1968.

[3] A. I. Gillespie, "The relationship between voice and breathing in the assessment and treatment of voice disorders," Perspectives of the ASHA special interest groups, 2016.

[4] 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, p. 12, 2017.

[5] A. Capellan and S. Fuchs, "The interplay of linguistic structure and breathing in German spontaneous speech," in *Proc. Interspeech*, Lyon, France, 2013, pp. 2014–2018.

[6] A. M. MacLarnon and G. P. Hewitt, "The evolution of human speech: The role of enhanced breathing control," *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, vol. 109, no. 3, pp. 341–363, 1999.

[7] G. E. Tzelepis, F. D. McCool, J. H. Friedman, and F. G. Hoppin Jr, "Respiratory muscle dysfunction in Parkinson's disease1-3," *Am Rev Respir Dis*, vol. 138, pp. 266–271, 1988.

[8] N. P. Solomon and T. J. Hixon, "Speech breathing in Parkinson's disease," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 294–310, 1993.

[9] J. E. Huber, M. Darling, E. J. Francis, and D. Zhang, "Impact of typical aging and Parkinson's disease on the relationship among breath pausing, syntax, and punctuation," *American Journal of Speech-Language Pathology*, 2012.

[10] J. E. Huber, "Effects of utterance length and vocal loudness on speech breathing in older adults," *Respiratory physiology & neurobiology*, vol. 164, no. 3, pp. 323–330, 2008.

[11] J. M. Clair-Auger, L. S. Gan, J. A. Norton, and C. A. Boliek, "Simultaneous measurement of breathing kinematics and surface electromyography of chest wall muscles during maximum performance and speech tasks in children: Methodological considerations," *Folia Phoniatrica et Logopaedica*, vol. 67, no. 4, pp. 202–211, 2015.

[12] B. Conrad and P. Schönle, "Speech and respiration," *Archiv für Psychiatrie und Nervenkrankheiten*, vol. 226, no. 4, pp. 251–268, 1979.

[13] V. S. Nallanthighal, A. Härmä, and H. Strik, "Deep sensing of breathing signal during conversational speech," *Proc. Interspeech 2019*, pp. 4110–4114, 2019.

[14] A. Basu, A. Routray, R. Mukherjee, and S. Shit, "Infrared imaging based hyperventilation monitoring through respiration rate estimation," *Infrared Physics & Technology*, vol. 77, pp. 382–390, 2016.

[15] D. Ruinskiy and Y. Lavner, "An effective algorithm for automatic detection and exact demarcation of breath sounds in speech and song signals," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 838–850, 2007.

[16] P. Wagner, J. Trouvain, and F. Zimmerer, "In defense of stylistic diversity in speech research," *Journal of Phonetics*, vol. 48, pp. 1–12, 2015.

[17] W. J. Levelt, *Speaking: From intention to articulation*. MIT press, 1993, vol. 1.

[18] V. Rapcan, S. D'Arcy, and R. B. Reilly, "Automatic breath sound detection and removal for cognitive studies of speech and language," in *IET Irish Signals and Systems Conference (ISSC 2009)*. IET, 2009, pp. 1–6.

[19] A. Abushakra and M. Faezipour, "Acoustic signal classification of breathing movements to virtually aid breath regulation," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 2, pp. 493–500, 2013.

[20] T. Fukuda, O. Ichikawa, and M. Nishimura, "Detecting breathing sounds in realistic Japanese telephone conversations and its application to automatic speech recognition," *Speech Communication*, vol. 98, pp. 95–103, 2018.

[21] M. I. Y. A. K. and A. Routray, "Automatic Detection of Breath Using Voice Activity Detection and SVM Classifier with Application on News Reports," in *Proc. Interspeech 2019*, 2019, pp. 609–613.

[22] É. Székely, G. E. Henter, and J. Gustafson, "Casting to corpus: Segmenting and selecting spontaneous dialogue for tts with a cnn-lstm speaker-dependent breath detector," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6925–6929.

[23] B. Lei, S. A. Rahman, and I. Song, "Content-based classification of breath sound with enhanced features," *Neurocomputing*, vol. 141, pp. 139–147, 2014.

[24] S. H. Dumpala and K. R. Alluri, "An algorithm for detection of breath sounds in spontaneous speech with application to speaker recognition," in *International Conference on Speech and Computer*. Springer, 2017, pp. 98–108.

[25] L. Lu, L. Liu, M. J. Hussain, and Y. Liu, "I sense you by breath: Speaker recognition via breath biometrics," *IEEE Transactions on Dependable and Secure Computing*, 2017.

[26] R. M. Simply, E. Dafna, and Y. Zigel, "Obstructive sleep apnea (osa) classification using analysis of breathing sounds during speech," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1132–1136.

[27] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. ICASSP*, Shanghai, China, 2016, pp. 5200–5204.

[28] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[29] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019, pp. 11–20.

[30] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen *et al.*, "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," *Proceedings INTERSPEECH. Shanghai, China: ISCA*, 2020.

[31] S. Amiriparian, N. Cummins, S. Julka, and B. Schuller, "Deep convolutional recurrent neural network for rare acoustic event detection," in *Proc. DAGA*, 2018, pp. 1522–1525.

[32] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[33] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 3543–3556.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

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

[36] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. R. Scherer, "On the Acoustics of Emotion in Audio: What Speech, Music and Sound have in Common," *Frontiers in Emotion Science*, vol. 4, pp. 1–12, 2013.