PHONEME BASED RESPIRATORY ANALYSIS OF READ SPEECH

Venkata Srikanth Nallanthighal[1,2] *Aki Härmä*[1] *Helmer Strik*[2] *Mathew Magimai Doss*[3]

[1]Philips Research, Eindhoven, The Netherlands
[2]Centre for Language Studies (CLS), Radboud University Nijmegen
[3]Idiap Research Institute, Martigny, Switzerland

ABSTRACT

Recent work shows that it is possible to use deep learning techniques to sense the speaker's respiratory parameters directly from a speech signal. This can be a beneficial option for future telehealth services. In this paper, we dive deeper and study how respiratory effort depends on the linguistic content of the speech utterance. This is obtained by analysis of respiratory belt sensor data and phoneme-aligned speech data. The results show, for example, that the respiratory effort was highest for fricatives, compared to other broad phonetic classes, and especially high for the glottal consonants. The insights may help to develop more efficient protocols for respiratory health monitoring in telehealth applications.

Index Terms— breathing signal, phonetics, Respiratory effort, speech technology, signal processing.

1. INTRODUCTION

Understanding the relationship between respiration and speech production has been an important field of research and the current pandemic situation of Covid-19 accelerates this need [1]. Our previous research showed that respiratory or breathing signals could be obtained automatically from the speech signal using deep learning models [2, 3]. In this study, we explore the relationship between breathing signal and speech production at the phoneme level.

Earlier studies on respiratory effort and articulation activity during the speech were performed on sustained vowels, selected consonants, and single utterances, and they showed that respiratory flow volume measurements are relevant for speech production [4, 5]. However, the current authors are not aware of an earlier systematic studies on relations of respiratory effort and phonetic classes in normal conversational or read speech content, possibly due to the difficulties in aligning speech data with other sensor data at the phoneme level. In the current paper the alignment is performed using a HMM/GMM acoustic model trained with a large speech database(Section 2.1).

Studying the respiratory effort at the phoneme level in normal speech can provide deeper insights into the mechanism of speech production. This would also have interesting applications in diagnosis and monitoring pathological speech conditions [6, 7]. For example, some *dysarthric* speakers have problems producing bilabials, especially stops. By comparing respiratory effort for bilabials to previous measurements of the same speaker, it is possible to determine whether their condition improves or deteriorates. Similar measurements and comparisons can be carried out for non-native speakers for assessment and training by providing instantaneous feedback, especially while learning a new language. A better understanding of the relation between linguistic content and respiration effort can give us more insight into breathing planning involved during the speech. Therefore, we have studied the relationship between phonemes and their respiratory effort. We analyze the respiratory effort characteristics of individual phonemes and different classes of phonemes such as (1) Broad Phonetic Classes, (2) Voiced and Unvoiced phonemes (3) Phonemes with the same place of articulation.

2. METHOD

Our speech database was collected at Philips Research[1], Eindhoven. The data was collected using the following setup: two NeXus respiratory inductance plethysmography belts (RIP belts) placed over the ribcage and abdomen. The belt sensor data corresponds to the changes in the cross-sectional area of ribcage and abdomen at the sample rate of 2 kHz. The sum of ribcage and abdomen expansions measured by the respiratory belt transducers is considered as the measure for the respiratory or breathing signal [8]. This breathing signal is down sampled to 100 hertz and mapped with the phoneme aligned speech. Speech was captured at 48kHz sampling rate using an Earthworks microphone M23 which is placed at a distance of 1 meter from the speakers. The data consists of recordings from 30 healthy subject's data (15 male, 15 female) with ages in 20 to 40 years. Each subject was asked to read out a phonetically balanced paragraph 'The Rainbow Passage' for approximately 2 minutes with regular speed and controlled loudness. This allows us to assume that respiratory effort involved in articulating a particular phoneme is stable across the subjects which is also observed in our analysis. Using same text for read speech, we receive similar

[1] with the approval of the Internal Committee of Biomedical Experiments.

linguistic content from each talker. Also, in read speech, the breath events are largely organised around phrase and sentence boundaries[9], while in spontaneous speech, higher percentage of breath events are placed in ungrammatical locations [10]. Thus limiting our initial exploration for read speech with controlled setup helps us make conclusive results in this study.

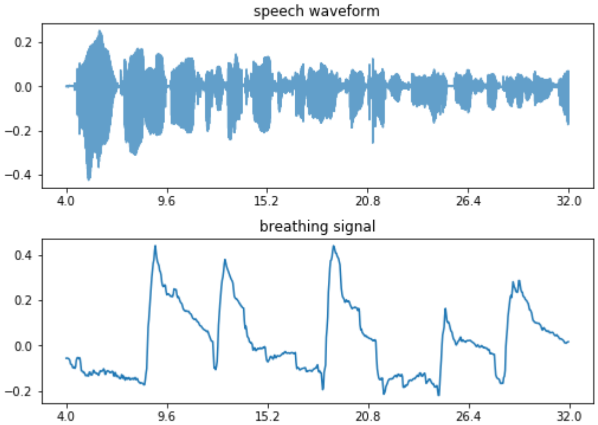

Fig. 1. Respiratory belt signal with speech signal

2.1. Phoneme alignment

The read speech of each subject was phonetically aligned using an HMM/GMM acoustic model trained on 960 hours of Librispeech data set using the Kaldi toolkit[11]. The features are extracted at the frame rate of 10ms. The dictionary is based on CMU-DICT, augmented with pronunciations for new words obtained through grapheme-to-phoneme conversion using Sequitur [12]. After obtaining the alignments, we removed position markers and lexical stress markers from the phones for our analysis. This yielded a phone set of 39 context-independent phones presented in Table 1.

2.2. Respiratory effort parameters

Respiratory effort information is measured by analyzing the breathing signal with the phoneme aligned speech. The Phoneme alignment gives the phoneme for each window size of 10ms. Counting the number of consecutively repeating phonemes(n), we calculate the duration(t) of a particular phoneme (t = 10ms*n). As the speech and breathing signal are aligned, we calculate the change in the breathing signal over this time 't' to get the lung volume change over the phoneme utterance.

As the breathing signal obtained based on RIP belts' expansion and contraction estimates lung volume changes(ΔLV) [13], it has quantitative significance though not calibrated. Breathing signal is normalized between 0 and 1 and the normalized value can be used for comparison.

1. **Lung volume change(ΔLV)**: This is measured by measuring the change over the breathing signal measured using RIP belt sensors for the window equal to the length of the phoneme. Unit: Volume(cm^3).

2. **Lung volume change rate(ΔLV/Δt)**: This is measured by dividing lung volume change(ΔLV) with time duration(Δt) of a phoneme. This can also be seen as the velocity of lung volume change during speech. Unit: Volume/time(cm^3/s).

3. ANALYSIS

Using phone alignment, we extract the phoneme information for each recording for all 30 subjects and calculate the duration and respiratory effort parameters corresponding to the phonemes. We initially grouped phonemes of each subject according to Broad phonetic class (BPC), Voiced-Unvoiced, and place of articulation and calculated the average time duration and average respiratory effort parameters for each group. As each subject is asked to read out the same Phonetically balanced paragraph 'The Rainbow Passage' with regular speed and controlled loudness, we observed similar trend in respiratory effort parameters among all subjects for individual phonemes and also groups of phonemes, which is evident from the low standard deviation values in Figure 2. This enables us to group phonemes of all the subjects together for further analysis. Thus all the results reported in the following sections are averaged values for individual phoneme for all the 30 subjects. We also compared median and mean measurements to understand the data distribution with respect to different speakers and classes and found no difference; which suggests that there aren't many outliers.

3.1. Broad Phonetic Classes

Broad phonetic classes (BPC's) have widely been used in speech recognition research as, for instance, automatic language identification [14]; speaking rate estimation [15]; multi lingual systems [16] and, especially, in phone recognition [17]. In this study, we consider the following BPC's, see Table 1: vowels, fricatives, nasals, approximants, and stops or plosives.

Broad Phonetic Class (BPC)	**Phones**	**Count**
Vowels	UH, IH, AH, OY, OW AE, AO, AY, IY, AA AW, UW, EY, ER, EH	15
Fricatives	V, TH, Z, ZH, SH S, JH, F, DH, HH, CH	11
Nasals	M, N, NG	3
Approximants	R, W, Y, L	4
Stops or Plosives	G, B, P, K D, T	6

Table 1. Broad Phonetic classes

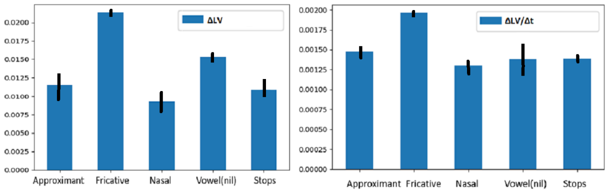

Fig. 2. ΔLV and ΔLV/Δt for Broad Phonetic classes

1. Among the BPCs, we observed the highest ΔLV and ΔLV/Δt values for fricatives (Figures 2). This is justified as fricatives are produced, not through a complete closure, such as for stops, but by creating a partial obstruction of the airstream forced through a greatly narrowed channel at the point of obstruction. This forced airflow results in a higher rate of lung volume change.

2. If we leave out the fricatives and look at the remaining four BPCs, we observe the highest ΔLV values for vowels and the lowest values for nasals. However, for these four BPCs, the ΔLV/Δt seems to be very similar. This is probably caused by the differences in average duration for the phonemes in these BPCs; vowels on average are for a longer duration, and nasals are for a shorter duration.

In figure 2, The standard deviation is plotted for the BPC's respiratory effort parameters. However, not any significant difference is observed for any meaningful conclusions. Thus standard deviation is not included in figures 3, 4, and 5 in this paper for simplicity.

3.2. Voiced vs Unvoiced phonemes

Speech is produced by the vocal cords and the vocal tract which includes the mouth and the lips. Voiced signals are produced when the vocal cords vibrate during the pronunciation of a phoneme and such phonemes are called voiced phonemes. Unvoiced phonemes, by contrast, do not entail the use of the vocal cords. Voiced phonemes tend to be louder like vowels and unvoiced phonemes tend to be more abrupt like stop consonants.

Analysis of ΔLV and ΔLV/Δt of voiced and unvoiced phonemes based on broad phonetic classes and place of articulation is very interesting in understanding the respiratory effort in speech production.

From Figures 2, 3, and 4, it can be observed that :

1. Unvoiced phonemes have higher respiratory effort compared to voiced phonemes (Figure 3). Voiced phonemes are produced when the vocal cords vibrate, which results in regular pulses from the glottal modulation of the airstream. This constriction at the glottis by voiced phonemes results in smaller lung volume changes compared to unvoiced phonemes.

	Phonemes
Voiced Phonemes	All vowels, All nasals All approximants Fricatives : V, Z, S, JH, DH Stops : G, D, B
Unvoiced Phonemes	Stops : T, P, K Fricatives : TH, ZH, SH, F, CH, HH

Table 2. Voiced and Unvoiced phonemes

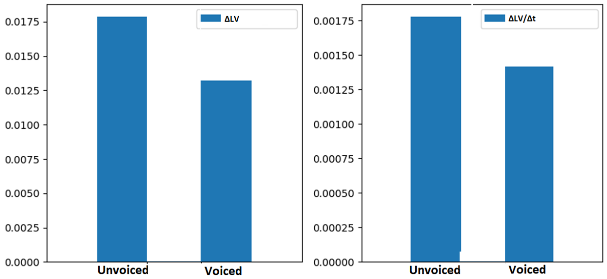

Fig. 3. ΔLV and ΔLV/Δt of voiced and unvoiced phonemes.

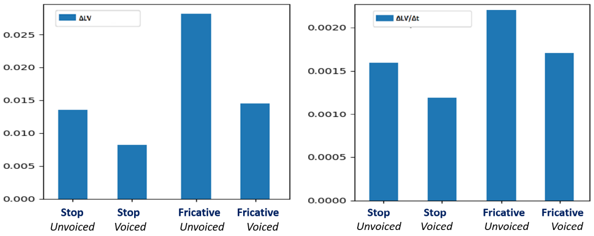

Fig. 4. ΔLV and ΔLV/Δt of voiced and unvoiced phonemes grouped as per Broad Phonetic Classes for fricatives and stops.

2. Among fricatives, unvoiced fricatives have higher ΔLV and ΔLV/Δt (Figure 4). Also among stops, unvoiced stops have higher ΔLV and ΔLV/Δt than voiced stops (Figure 4).

3. Nasals and voiced stops follow a similar trend for ΔLV and ΔLV/Δt (Figures 2, 4). The nasal consonants are the nasal counterparts of the voiced plosives in English, the main difference being that the air does not escape through the mouth, but rather the nasal tract [18].

4. Approximants follow the same trend as of Vowels (Figure 2). Both being voiced phonemes, approximants are also referred to as semivowels because of their phonetic similarity.

3.3. Phonemes based on place of articulation

Place of articulation (Table 3) refers to the location where the constriction or obstruction of the vocal tract occurs, as well as to the active or passive articulators involved in the production of the consonant. Place of articulation generally is only used for consonants, and therefore the vowels here get the place attribute 'nil'. From Figure 5, it can be observed that :

1. All vowels, alveolars, palatals, retroflexs are voiced phonemes and thus have lower respiratory effort.
2. The glottal phoneme has the highest lung volume change and lung volume change rate. The glottal phoneme in English is the unvoiced fricative HH.
3. The palatal phoneme has the least respiratory effort. The palatal phoneme in English is the voiced approximant Y, as in the word 'yes.' We can observe that the airflow is well constricted during this pronunciation and hence lesser ΔLV and ΔLV/Δt.

Place of articulation	**Phones**	**Voiced or Unvoiced**
Alveolar	D, L, N, S, T, Z	Voiced
Bilabial	B, P M	Unvoiced Voiced
Dental	TH DH	Unvoiced Voiced
Glottal	HH	Unvoiced
Labiodental	F V	Unvoiced Voiced
Nil place	All Vowels	Voiced
Palatal	Y	Voiced
pos-alveolar	CH, SH, ZH JH	Unvoiced Voiced
Retroflex	ER, R	Voiced
Velar	K G, NG, W	Unvoiced Voiced

Table 3. Place of articulation

4. DISCUSSIONS

We present our research on respiratory effort for phonemes and groups of phonemes. The results are assuring, given what is known about lung volume changes and lung volume change rate during speech production, and generally are in line with previous findings. We observed higher ΔLV for unvoiced compared to voiced consonants, as in Fig 3. Unvoiced stops and fricatives have higher ΔLV compared to voiced stops and fricatives Fig. 4, which is in line with the findings of Stathopoulos [19]. This is plausible since the vocal folds do not vibrate, and air can pass more easily between the vocal folds for unvoiced consonants. For fricatives, ΔLV and ΔLV/Δt are higher. In fact, they show the highest respiratory effort among all the BPCs (Figs. 2 and 3), especially for unvoiced fricatives. This is plausible since to produce fricatives air has to keep flowing, and the velocity of the airflow has to be high enough to create frication (turbulence). Similarly, to produce frication (turbulence) at the glottal folds, the airflow and velocity of the airflow must be high enough. This explains why glottal consonants have the highest respiratory effort if we compare phoneme classes based on place of articulation.

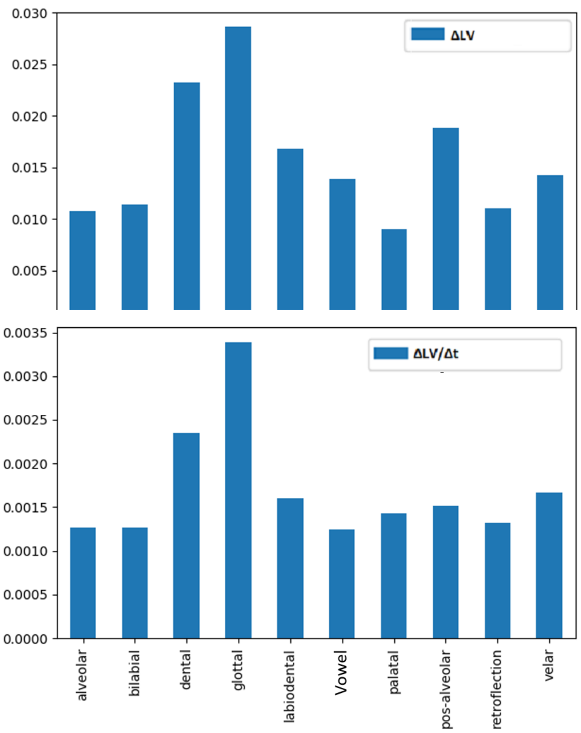

Fig. 5. ΔLV and ΔLV/Δt of phonemes grouped as per place of articulation

5. CONCLUSIONS AND FUTURE WORK

In this paper, we study the relationship between linguistic content and respiratory effort. Firstly, our research shows that it is possible to obtain respiratory information at the level of individual phonemes, which opens up novel possibilities to study respiratory activity during speech production for specific phonemes or groups of phonemes. Our previous research shows that the estimation of the respiration signal can be obtained automatically from the speech signal only [2, 3]. The combination of both approaches makes it possible to first estimate the breathing signal from speech, and next use this breathing signal and the same speech signal to obtain information at the phoneme level. It will then be feasible to obtain information on respiratory effort for phonemes using only speech signal, without using belts around chest and abdomen. This combined approach could then be used for various tasks such as diagnosis and monitoring for pathological speech, assessment, real-time feedback for non-native speakers.

6. ACKNOWLEDGEMENTS

This work is supported under the EU's H2020 Marie Skłodowska-Curie programme TAPAS; Grant No: 766287. The authors would like to thank Qingran Zhan, IDIAP Research institute for the help with phoneme alignment.

7. REFERENCES

[1] Bjorn W. Schuller, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, Harald Baumeister, Alexis Deighton MacIntyre, and Simone Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly emotion, Breathing & Masks," in *Proceedings of Interspeech*, Shanghai, China, September 2020, p. 5 pages, to appear.

[2] Venkata Srikanth Nallanthighal, Aki Härmä, and Helmer Strik, "Deep Sensing of Breathing Signal During Conversational Speech," in *Proc. Interspeech 2019*, 2019, pp. 4110–4114.

[3] V. S. Nallanthighal, A. Härmä, and H. Strik, "Speech breathing estimation using deep learning methods," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1140–1144.

[4] DH Klatt, KN Stevens, and J Mead, "Studies of articulatory activity and airflow during speech," *Annals of the New York Academy of Sciences*, vol. 155, no. 1, pp. 42–55, 1968.

[5] Nobuhiko Isshiki and Robert Ringel, "Air flow during the production of selected consonants," *Journal of Speech and Hearing Research*, vol. 7, no. 3, pp. 233–244, 1964.

[6] Leonard L La Pointe and Donnell F Johns, "Some phonemic characteristics in apraxia of speech," *Journal of Communication disorders*, vol. 8, no. 3, pp. 259–269, 1975.

[7] Hermann Ackermann, Susanne Gräber, Ingo Hertrich, and Irene Daum, "Phonemic vowel length contrasts in cerebellar disorders," *Brain and language*, vol. 67, no. 2, pp. 95–109, 1999.

[8] 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, PMID: 4225383.

[9] Alison L Winkworth, Pamela J Davis, Elizabeth Ellis, and Roger D Adams, "Variability and consistency in speech breathing during reading: Lung volumes, speech intensity, and linguistic factors," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 3, pp. 535–556, 1994.

[10] Yu-Tsai Wang, Jordan R Green, Ignatius SB Nip, Ray D Kent, and Jane Finley Kent, "Breath group analysis for reading and spontaneous speech in healthy adults," *Folia Phoniatrica et Logopaedica*, vol. 62, no. 6, pp. 297–302, 2010.

[11] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number CONF.

[12] Maximilian Bisani and Hermann Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech communication*, vol. 50, no. 5, pp. 434–451, 2008.

[13] Gerhard Wolf and John Arnold, "Noninvasive assessment of lung volume: Respiratory inductance plethysmography and electrical impedance tomography," *Critical Care Medicine*, vol. 33, no. 3, 2005.

[14] Timothy Kempton and Roger K Moore, "Language identification: insights from the classification of hand annotated phone transcripts.," in *Odyssey*, 2008, p. 14.

[15] Jiahong Yuan and Mark Liberman, "Robust speaking rate estimation using broad phonetic class recognition," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4222–4225.

[16] Lin Yang Jianping Zhang Yonghong Yan, "Acoustic units selection in chinese-english bilingual speech recognition," *NOn-LInear Speech Processing*, p. 96, 2007.

[17] Patricia Scanlon, Daniel PW Ellis, and Richard B Reilly, "Using broad phonetic group experts for improved speech recognition," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 803–812, 2007.

[18] W. Dodd , "J. c. catford, a practical introduction to phonetics (2nd edn.)," *Journal of the International Phonetic Association*, vol. 33, pp. 87 – 88, 06 2003.

[19] Elaine T Stathopoulos, "Oral air flow during vowel production," *Cleft Palate Journal*, vol. 21, no. 4, 1984.