ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888421

Phoneme-Aware Acoustic Analysis of Natural Speech for Lung Function Assessment

Sejal Bhalla[1], Tien Han[1], Andrea Gershon[2], Robert Wu[3], Eyal de Lara[1], Alex Mariakakis[1]

[1]*University of Toronto, Toronto, Canada*
[2] *Sunnybrook Health Sciences Centre, Toronto, Canada*
[3] *University Health Network, Toronto, Canada*

***Abstract*—Techniques that use speech analysis for tasks like health monitoring and emotion recognition usually operate on moderately sized windows with little regard for what the individual is saying. In this work, we argue that isolating specific phonemes within speech offers greater nuance that leads to more consistent yet natural sounds for analysis. We examine this hypothesis in the context of lung function estimation. We recruited 11 patients with chronic obstructive pulmonary disease (COPD) to read from a script and perform spirometry to quantify their lung function. After segmenting their audio recordings into discrete phonemes, we extracted various phonation, prosodic, and spectral features to summarize their acoustic qualities. We then examined the correlation between those audio features and measurements from spirometry, observing that certain combinations of features and phonemes led to higher correlations than the best-performing phoneme-agnostic baseline for our dataset.**

***Index Terms*—speech analysis, phonemes, COPD, lung function**

I. Introduction

Chronic obstructive pulmonary disease (COPD) is a prevalent and debilitating respiratory condition characterized by persistent airflow limitation. Its high prevalence, morbidity, and mortality make it a significant global public health issue [1]–[3]. As COPD progresses, patients experience deteriorating lung function that can be measured using pulmonary function tests. While spirometry is considered the gold standard for assessing lung function, there is growing interest in utilizing speech analysis as a non-invasive, low-effort, complementary approach.

COPD affects voice production and quality because of its direct relationship with lung function and indirect effects from symptoms such as cold and cough [4]–[6]. The disease leads to reduced lung volume, diminished subglottic pressure, and compromised glottal closure during phonation [7], [8]. Thus, prior studies have explored the use of temporal and spectral features from continuous speech to detect lung function deterioration. Many of these approaches employ machine learning models that process brief speech segments using a fixed-length window [9]–[11]. While this method allows models to be used whenever a person speaks, it creates a complex and heterogeneous feature space that requires an equally diverse training dataset to achieve robust performance. In contrast, some approaches instruct users to produce sustained phonemes, such as /s/ in "sit" or /aa/ in "father" to generate a more homogeneous feature space [12]–[15]. These controlled exercises simplify the modeling process, but they often feel unnatural and may not accurately reflect everyday speech patterns, thus limiting their practical application.

Our work bridges these two approaches by leveraging phonemes extracted from natural speech to assess lung function. We collected data from 11 COPD patients who read scripted passages and completed spirometry. Using a forced alignment tool, we segmented the phonemes from the recordings and filtered out the ones that were uttered less frequently. After segmentation, we extracted a range of phonation, prosodic, and spectral features to characterize the acoustic properties of the remaining phonemes. We then performed correlation analyses between these acoustic features and spirometry metrics. Our findings reveal that spectral features extracted from certain phonemes, such as /l/, /ah/, and /iy/, exhibit strong correlations with lung function metrics and can even outperform traditional speech analysis approaches. By adopting a structured phoneme-based segmentation approach, we uncovered nuanced relationships between acoustic phoneme features and lung function metrics, paving the way for more precise and effective respiratory health monitoring.

II. Methods

A. Data Collection

Data for this investigation was drawn from a larger study involving remote COPD patient monitoring [16], [17] that received approval from the Research Ethics Board at the University of Toronto under Protocol #41568. The most relevant exclusion criteria to this investigation was that patients had to be able to read and speak in English. It is also worth noting that patients were recruited during periods of stable symptoms rather than at the time of a severe exacerbation. This paper focuses on 11 patients from the study (6 females, mean age: 70.3 $\pm$ 8.2) who were able to complete the full onboarding procedure consisting of scripted speech recordings and spirometry sessions.

This study was funded in part by Canadian Institutes of Health Research (CIHR) Grant #155458, Natural Sciences and Engineering Research Council of Canada (NSERC) Grants #RGPIN-2021-03457 and #RGPIN-2017-06618, and Samsung Research America (SRA).

Data collection took place in a quiet lab environment. For the speech recordings, participants either read the Rainbow Passage [18] or a passage from Harry Potter and the Sorcerer's Stone [19]. Their voice was recorded using a Samsung Galaxy Watch with a microphone that had a sampling rate of 44.1 kHz, and the average duration of each recording was 97.9 ± 13.2 seconds. Patients also went through three rounds of spirometry under researcher supervision. Our analyses focus on two measurements from the device: forced expiratory volume (FEV1), the volume of air that a person can forcefully exhale within 1 second; and forced vital capacity (FVC), the total volume of air that a person can forcefully exhale. FEV1 and FVC readings were averaged across the three trials to produce two outcome variables for our analysis.

B. Data Processing

To prepare the data for analysis, we first removed silence from the audio recordings to isolate speech content. We then used the Montreal Forced Aligner (MFA) [20] to identify precise phoneme boundaries; in short, MFA uses a phonetic model to automatically align an audio signal with a known transcript. Phoneme segments shorter than 50 milliseconds were excluded from the analysis, as phonemes in the English language typically exceed this duration [21], [22]. Following segmentation, we extracted audio features using the Parselmouth library [23], which builds upon the Praat software [24]. The features are listed below:

- **Phonation features:** These features capture the temporal and amplitude variations within an audio frame [25].
  - *Jitter* metrics assess variations in pitch frequency. Local jitter measures short-term frequency variations, absolute jitter assesses overall frequency stability, RAP (relative average perturbation) and PPQ5 (5-point pitch perturbation quotient) provide additional frequency stability metrics, and DDP (difference of differences) measures the variability in pitch changes.
  - *Shimmer* metrics quantify amplitude fluctuations. Local shimmer, APQ3 (amplitude perturbation quotient), and APQ5 measure short-term, 3-point, and 5-point amplitude perturbations respectively.
- **Prosodic features:** These features capture the overall pitch and loudness of an audio signal [26].
  - *Fundamental frequency* summarizes the rate of vocal fold vibration as the pitch of voiced speech.
  - *Formant frequencies* describe the resonant frequencies of the vocal tract. Fi^j denotes a specific fractional position relative to the formant i's bandwidth. For instance, F$1^{1/2}$ refers to the center frequency of the F1 formant.
- **Spectral Features**: Mel-frequency cepstrum coefficients (MFCCs) provide a compact spectral representation of a speech signal's power spectrum [27]. We compute 12 MFCCs to trade off richness with analytical efficiency.

TABLE I: Prevalence of different phonetic sounds in our dataset and the English Language.

Phoneme	Instances per Patient (mean ± std)	Prevalence in English
/ae/	15.5 ± 4.0	2.1%
/ah/	38.6 ± 7.4	11.5%
/d/	30.9 ± 6.0	4.2%
/eh/	15.5 ± 3.7	2.9%
/er/	22.6 ± 9.0	6.9%
/ih/	31.6 ± 6.0	6.3%
/iy/	23.1 ± 5.8	3.6%
/k/	17.9 ± 7.4	3.2%
/l/	21.9 ± 5.1	4.0%
/m/	12.4 ± 2.9	2.8%
/n/	36.1 ± 6.7	7.1%
/r/	19.8 ± 5.9	6.9%
/s/	35.4 ± 12.3	4.8%
/t/	35.1 ± 7.8	6.9%
/z/	19.1 ± 5.1	2.8%

We averaged phoneme-level features across all instances of the same phoneme uttered by each patient, resulting in a single feature vector with 33 values per phoneme.

To generate phoneme-agnostic baselines, we applied sliding window segmentation using two different window lengths: 2 seconds to align with prior studies [9], [11] and 120 milliseconds to approximate the average phoneme duration in our dataset. Similar to phoneme-level data processing, feature vectors were computed for each baseline by averaging the feature values across all speech windows.

C. Analysis

Given the modest size of our dataset, we conducted a correlation analysis to examine the association between audio features and lung function. We calculated the Pearson correlation between all possible combinations of audio features and spirometry metrics to identify which features were most indicative of poor lung function. We ran this procedure using both baseline phoneme-agnostic windows and individual phonemes to identify the most effective speech segmentation method for this task.

III. Results

A. Phoneme Prevalence

Table I shows the average number of times each phoneme was detected by MFA across all patients. It also shows the prevalence of each phoneme in the English language, calculated by triangulating the CMU Pronouncing Dictionary [28] with the lemmatized frequency list[1] for the British National Corpus [29]. To prioritize robust results, we focused our analyses on phonemes that have a prevalence of 2% in the English language and were detected an average of 15

[1]https://www.kilgarriff.co.uk/bnc-readme.html

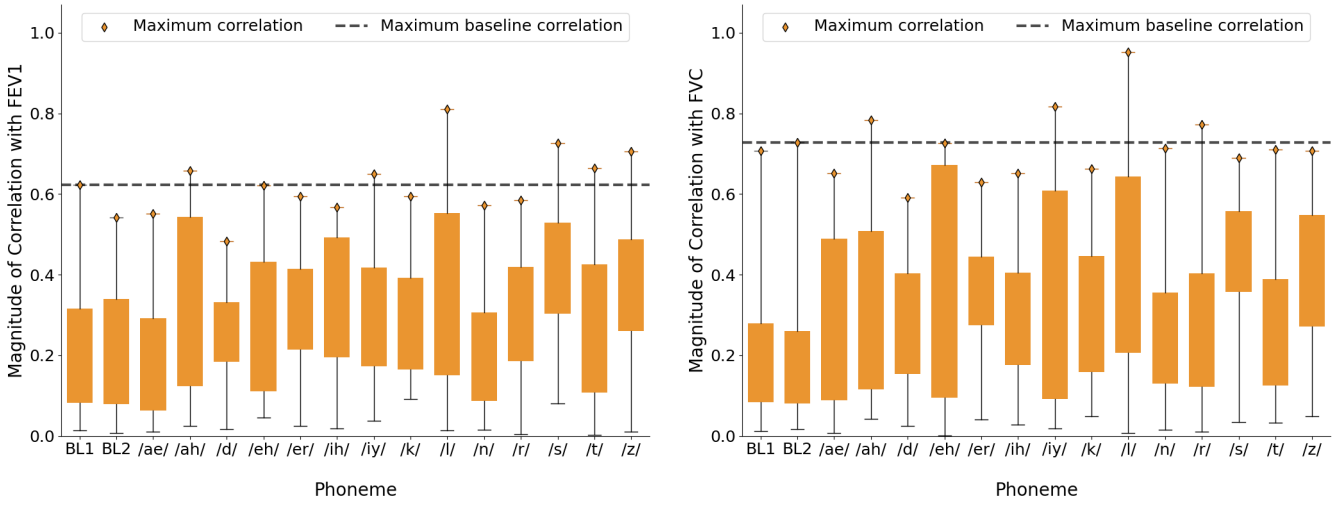

Fig. 1: Box-and-whisker plots showing how the distribution of correlation magnitudes between audio features and either (left) FEV1 or (right) FVC vary across different phonemes. The two leftmost distributions in each plot correspond to phoneme-agnostic baselines BL1 and BL2 with window lengths of 2 seconds and 120 milliseconds, respectively.

times per patient. Note that the latter quantity was not expected to be consistent across patients because they read one of two scripts. Patients also enunciated the script differently due to variations in dialects or respiratory support, so phonemes might have been missed or filtered out by the duration requirement we implemented in our pipeline.

B. Distribution of Correlations

Fig. 1 shows the distribution of correlation magnitudes between phonemes' audio features and spirometry outcome variables. Distributions that exceed the maximum correlation observed in the baselines indicate instances when creating a more homogenous set of audio windows resulted in stronger correlations for at least one combination of audio features and outcome variables. While the range of correlation values reflects the variability in the predictive power of different features, the maximum values are critical for identifying promising phonemes for lung function assessment.

The baseline distribution for correlation magnitudes against FEV1 ranged from 0.01 to 0.62 (median: 0.18, IQR: [0.08, 0.32]). Six of the fourteen phonemes in our analyses extended above the baseline distribution for at least one audio feature. For FVC, the baseline distribution for correlation magnitudes ranged from 0.01 to 0.71 (median: 0.17, IQR: [0.08, 0.28]). Four of the fourteen phonemes extended above the baseline distribution for at least one audio feature. We observed that phonemes /ah/, /iy/, and /l/ exceeded the maximum baseline correlations for both FEV1 and FVC. The prominence of /ah/ and /iy/ aligns with prior studies that utilize vowel sounds to characterize lung condition, with the similar sounding /aa/ phoneme being the most commonly employed sound in sustained phonation for a variety of speech analysis tasks [14], [30].

C. Most Promising Correlations

Fig. 2 highlights the correlations that exceeded baseline values in magnitude for all possible combinations of selected phonemes and audio features. Notably, MFCC6 for the phoneme /l/ exhibited the strongest negative correlation for both FEV1 ($r = -0.81$, $p < .01$) and FVC ($r = -0.95$, $p < .001$). On the other hand, $F3^{2/3}$ for the phoneme /l/ and MFCC4 for the phoneme /iy/ demonstrated the highest positive correlations with FEV1 ($r = 0.77$, $p < .01$) and FVC ($r = 0.82$, $p < .01$), respectively. MFCCs and formant frequencies generally yielded stronger correlations than phonation features.

The disparate combinations of phonemes and audio features surfaced by our analysis illustrate the utility of restricting analysis to individual phonemes. Among the selected features, DDP jitter and MFCC12 for FEV1 and MFCC4 for FVC showed the highest number of strong correlations, suggesting that these features may be particularly informative in predicting lung function outcomes. The direction of the correlations also varied, even within the same category of features and outcome variables. As an example, DDP jitter was positively correlated with FEV1 for the phoneme /s/ ($r = 0.66$, $p < .05$), while a negative correlation was observed for the phoneme /ah/ ($r = -0.66$, $p < .05$).

IV. Discussion

Speech analysis for health monitoring and emotion recognition has traditionally focused on either natural speech [9]–[11] or sustained phonation of phonemes [12]–[15]. Our work provides additional evidence in favor of merging these approaches with respect to lung function assessment. We showed that there is useful information in phonemes extracted from natural speech,

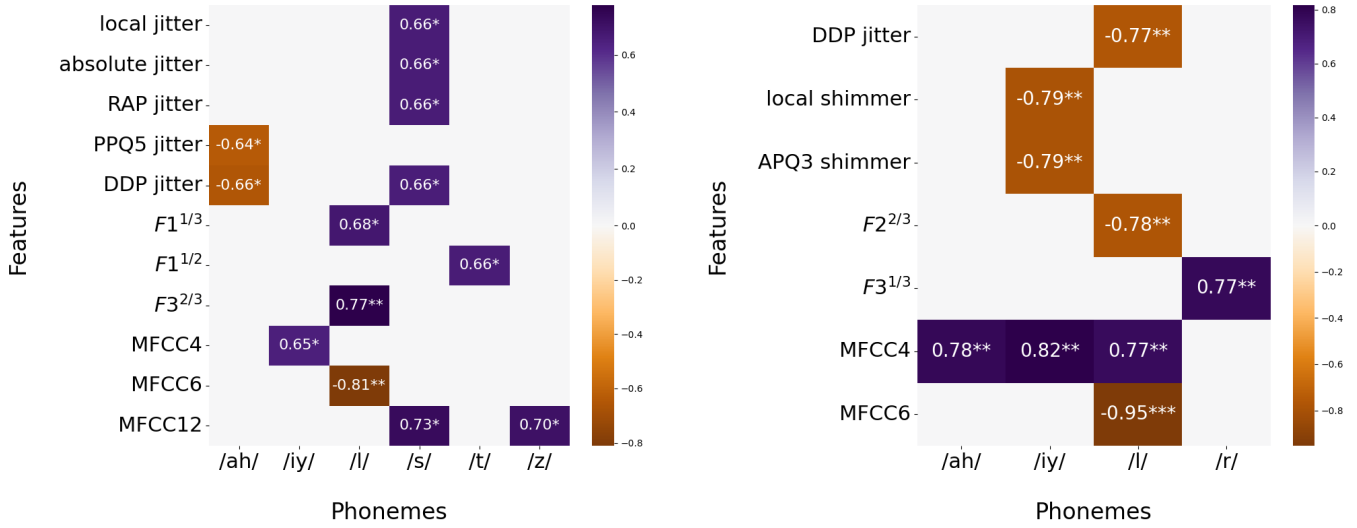

Fig. 2: Heatmaps showing the correlations between audio features and FEV1 (left) or FVC (right). Only the correlations that were statistically significant and greater in magnitude than the largest baseline correlation ($|r| > 0.62$ for FEV1, $|r| > 0.71$ for FVC) are shown. Phonemes that did not exhibit any notable correlations across any audio features are not shown and vice versa. Significant correlations are indicated as follows: $*p < .05$, $**p < .01$, $***p < .001$.

requiring less effort compared to vocal exercises from patients who may already be experiencing respiratory distress. Through our comprehensive analysis of varied phonetic sounds, we confirmed the utility of phonemes used in prior work (e.g., /s/, /z/, vowel sounds) [14], [15] and revealed even more potential indicators of poor lung function that could easily be overlooked when examining heterogeneous natural speech segments.

The variability in correlation direction across different phonemes highlights that distinct sounds may capture unique aspects of lung function. Each phoneme engages different articulatory and respiratory mechanisms, which likely influences its effectiveness as an indicator of lung health [31]. This reinforces the need for phoneme-level analysis, as combining phonemes in unstructured speech segments could obscure these critical relationships. Different features of the same phoneme also exhibited varying correlation strengths and directions with lung function, emphasizing the multidimensionality of speech as a physiological signal. Thus, speech-based health assessments must leverage a diverse set of features to accurately reflect the complex interactions in vocal production, enabling more precise and reliable monitoring of respiratory health.

Our work has its limitations, the most prominent being the size of our dataset. Participant recruitment was rate-limited by the broader study protocol used for recruitment, and not all patients were able to perform spirometry at the time of enrollment. This is why we opted for a correlation analysis instead of machine learning, but with more data, other evaluation techniques may be viable. On that note, another limitation of our work is the fact that we examined associations between individual features and outcome variables. Most applications of speech analysis rely on multiple features simultaneously using machine learning on either handcrafted features or spectrograms [9]–[11]. With more data and machine learning, we may be able to combine audio features from multiple phonemes to generate more information for respiratory health prediction.

V. Conclusion

This study demonstrates that isolating phonemes within natural speech provides a more accurate method for analyzing respiratory function compared to traditional speech segmentation techniques. By focusing on individual phonemes, our analysis revealed significant correlations between speech features and lung health that might be obscured when using heterogeneous speech segments. For instance, MFCC features of the phoneme /l/ exhibit strong correlations with lung function metrics, with a linear correlation as high as -0.95 with FVC and -0.81 with FEV. This phoneme-based approach not only enhances the precision of non-invasive lung function monitoring but also represents a notable advancement in speech-based health assessment. In the future, we hope to validate these findings with a larger and more diverse dataset. We also aim to explore additional speech analysis tasks that would further support the utility of phoneme-based segmentation in various health monitoring contexts.

References

[1] S. Suissa, S. Dell'Aniello, and P. Ernst, "Long-term natural history of chronic obstructive pulmonary disease: severe exacerbations and mortality," *Thorax*, vol. 67, no. 11, pp. 957–963, Nov. 2012.

[2] S. Chen, M. Kuhn, K. Prettner, F. Yu, T. Yang, T. Bärnighausen, D. E. Bloom, and C. Wang, "The global economic burden of chronic obstructive pulmonary disease for 204 countries and territories in 2020–50: a health-augmented macroeconomic modelling study," *The Lancet Global Health*, vol. 11, no. 8, pp. e1183–e1193, 2023.

[3] O. E. L. Jeetvan G Patel, Anna D Coutinho and A. A. Dalal, "COPD affects worker productivity and health care costs," *International Journal of Chronic Obstructive Pulmonary Disease*, vol. 13, pp. 2301–2311, 2018, pMID: 30104870.

[4] A. Shastry, R. K. Balasubramanium, and P. R. Acharya, "Voice analysis in individuals with chronic obstructive pulmonary disease," *International Journal of Phonosurgery & Laryngology*, vol. 4, no. 2, pp. 45–49, Dec. 2014.

[5] E. E. Mohamed and R. A. El maghraby, "Voice changes in patients with chronic obstructive pulmonary disease," *Egyptian Journal of Chest Diseases and Tuberculosis*, vol. 63, no. 3, pp. 561–567, 2014.

[6] G. dos Anjos Palagi da Silva, T. D. Feltrin, F. dos Santos Pichini, C. A. Cielo, and A. S. Pasqualoto, "Quality of life predictors in voice of individuals with chronic obstructive pulmonary disease," *Journal of Voice*, 2022.

[7] J. Iwarsson and J. Sundberg, "Effects of lung volume on vertical larynx position during phonation." *Journal of Voice*, vol. 12 2, pp. 159–65, 1998.

[8] T. D. Feltrin, M. da Silva Packaeser Gracioli, C. A. Cielo, J. A. Souza, D. A. de Oliveira Moraes, and A. S. Pasqualoto, "Maximum phonation times as biomarkers of lung function," *Journal of Voice*, 2024.

[9] S. Bhalla, S. Liaqat, R. Wu, A. S. Gershon, E. de Lara, and A. Mariakakis, "Pulmolistener: Continuous acoustic monitoring of chronic obstructive pulmonary disease in the wild," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 3, pp. 1–24, 2023.

[10] K. S. Chun, V. Nathan, K. Vatanparvar, E. Nemati, M. M. Rahman, E. Blackstock, and J. Kuang, "Towards passive assessment of pulmonary function from natural speech recorded using a mobile phone," in *2020 IEEE International Conference on Pervasive Computing and Communications*, 2020, pp. 1–10.

[11] T. Sedaghat, S. Liaqat, D. Liaqat, R. Wu, A. Gershon, T. Son, T. H. Falk, M. Gabel, A. Mariakakis, and E. de Lara, "Unobtrusive monitoring of copd patients using speech collected from smartwatches in the wild," in *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events*, 2022, pp. 818–823.

[12] B. Das, K. Daoudi, J. Klempir, and J. Rusz, "Towards disease-specific speech markers for differential diagnosis in parkinsonism," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 5846–5850.

[13] Y. Maryn, P. Corthals, P. Van Cauwenberge, N. Roy, and M. De Bodt, "Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels," *Journal of Voice*, vol. 24, no. 5, pp. 540–555, 2010.

[14] N. Saleheen, T. Ahmed, M. M. Rahman, E. Nemati, V. Nathan, K. Vatanparvar, E. Blackstock, and J. Kuang, "Lung function estimation from a monosyllabic voice segment captured using smartphones," in *International Conference on Human-Computer Interaction with Mobile Devices and Services*, 2020, pp. 1–11.

[15] M. Dogan, E. Eryuksel, I. Kocak, T. Celikel, and M. A. Sehitoglu, "Subjective and objective evaluation of voice quality in patients with asthma," *Journal of Voice*, vol. 21, no. 2, pp. 224–230, 2007.

[16] R. Wu, M. Calligan, T. Son, H. Rakhra, E. de Lara, A. Mariakakis, and A. S. Gershon, "Impressions and perceptions of a smartphone and smartwatch self-management tool for patients with copd: A qualitative study," *COPD: Journal of Chronic Obstructive Pulmonary Disease*, vol. 21, no. 1, p. 2277158, 2024.

[17] R. Wu, E. de Lara, D. Liaqat, S. Liaqat, J. L. Chen, T. Son, and A. S. Gershon, "Feasibility of a wearable self-management application for patients with copd at home: a pilot study," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 66, 2024.

[18] G. Fairbanks, "Voice and articulation drillbook," 1960.

[19] J. Rowling, *Harry Potter and the Sorcerer's Stone*. New York: Scholastic, 2001.

[20] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Interspeech*, 2017, pp. 498–502.

[21] L. Gwilliams, J.-R. King, A. Marantz, and D. Poeppel, "Neural dynamics of phoneme sequences reveal position-invariant code for content and order," *Nature Communications*, vol. 13, 11 2022.

[22] G. Ma, P. Hu, J. Kang, S. Huang, and H. Huang, "Leveraging phone mask training for phonetic-reduction-robust e2e uyghur speech recognition," in *Interspeech*, 2021, pp. 306–310.

[23] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[24] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, 2009.

[25] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis–jitter, shimmer and hnr parameters," *Procedia Technology*, vol. 9, pp. 1112–1122, 2013.

[26] R. D. Kent and H. K. Vorperian, "Static measurements of vowel formant frequencies and bandwidths: A review," *Journal of Communication Disorders*, vol. 74, pp. 74–97, 2018.

[27] V. Tiwari, "Mfcc and its applications in speaker recognition," *International Journal on Emerging Technologies*, vol. 1, no. 1, pp. 19–22, 2010.

[28] "CMU Pronouncing Dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict, accessed: 2024-09-08.

[29] "British national corpus," http://www.natcorp.ox.ac.uk/, accessed: 2024-09-08.

[30] H. Polat and I. Guler, "A simple computer-based measurement and analysis system of pulmonary auscultation sounds," *Journal of Medical Systems*, vol. 28, pp. 665–72, 01 2005.

[31] V. S. Nallanthighal, A. Härmä, H. Strik, and M. M. Doss, "Phoneme based respiratory analysis of read speech," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 191–195.