Interspeech 2024
1-5 September 2024, Kos, Greece

Multimodal Fusion for Vocal Biomarkers Using Vector Cross-Attention

Vladimir Despotovic[1], *Abir Elbéji*[2], *Petr V. Nazarov*[1,3], *Guy Fagherazzi*[2]

[1]Bioinformatics & AI, Department of Medical Informatics, Luxembourg Institute of Health
[2]Deep Digital Phenotyping, Department of Precision Health, Luxembourg Institute of Health
[3]Multi-Omics Data Science, Department of Cancer Research, Luxembourg Institute of Health
{vladimir.despotovic,abir.elbeji,petr.nazarov,guy.fagherazzi}@lih.lu

Abstract

Vocal biomarkers are measurable characteristics of person's voice that provide valuable insights into various aspects of their physiological and psychological state, or health status. The use of standardized voice tasks, such as reading, counting, or sustained vowel phonation are common in vocal biomarker research, but semi-spontaneous tasks where the person is instructed to talk about a particular topic, or spontaneous speech are also increasingly used. However, limited efforts were made to combine multiple voice modalities. In this paper, we propose a simple, yet efficient approach of fusing multiple standardized voice tasks based on vector cross-attention, showing improved predictive capacity for derived vocal biomarkers in comparison to single modalities. The multimodal approach is tested on the assessment of respiratory quality of life from reading and sustained vowel phonation recordings, outperforming single modalities up to 4.2% in terms of accuracy (relative increase of 7%).

Index Terms: vocal biomarker, multimodal fusion, attention mechanism

1. Introduction

A vocal biomarker is a feature, or a combination of features from the audio signal of the human voice that represents a distinctive signature of a clinical outcome, and holds potential applications in patient monitoring, disease diagnosis, or the assessment of its severity [1]. While the standard protocol for voice recordings used to identify vocal biomarkers does not exist, they can be roughly classified into standardized vocal tasks (isolated words, reading a sentence or a passage, sustained vowel phonation, diadochokinetic task, counting), semi-spontaneous tasks (story narration, picture description) and spontaneous speech. Standardized vocal tasks provide a controlled and reproducible environment, ensuring that the same task is performed consistently across participants. Still, vocal biomarkers identified in this way may not fully represent an individual's natural vocal behavior, missing important vocal features that emerge in more dynamic and unconstrained communication [2]. Spontaneous speech does not suffer from such constraints, but introduces a high degree of variability and complexity in data, making data analysis and interpretation more challenging [3]. Sustained vowel phonation is less susceptible to the speaker's dialect, language, speaking rate, stress, or intonation [4, 5]. On the other hand, reading tasks or spontaneous speech can capture articulatory movements, which can influence vocal characteristics [6], and are not represented by sustained vowel phonation.

Leveraging multiple voice tasks for the identification of vocal biomarkers allows for capturing diverse voice characteristics that might be overlooked when relying solely on a single voice modality, thereby enhancing overall performance. However, efforts of the community were mostly focused on combining voice with other types of digital data, such as handwriting and gait for assessment of Parkinson's disease [7], chest X-ray images [8] and self-reported clinical data [9] for Covid-19 diagnosis, or facial video for monitoring of schizophrenia [10]. Martinc and Pollak combined audio features extracted from speech recordings with linguistic features extracted from text transcripts for recognition of Alzheimer's dementia in [11], whereas Escobar-Grisales et al. fused speech and language representations using gated multimodal units to distinguish between Parkinson's disease patients and healthy subjects [12].

There has been limited study on combining multiple voice tasks for building vocal biomarkers. Even though collecting multiple types of voice recordings is not uncommon [13, 14, 15, 16, 17, 18, 19], they are rarely used in the multimodal setup. To the best of our knowledge, the only attempt to fuse voice, cough and breathing recordings was presented by Dang et al. [20], but it was realized by a simple concatenation of individual embeddings, disregarding potential correlations between modalities. We bridge this gap by introducing an approach for multimodal fusion of vocal tasks using vector cross-attention, an adapted version of the standard single-head and multi-head cross-attention, which instead of representing dependencies between token embeddings belonging to two sequences, captures relationships between different elements of two embeddings corresponding to two voice modalities, i.e. sustained vowel phonation and reading task. By determining how sustained vowel phonation attends to reading and vice versa, we are able to capture complex dependencies between these modalities, thereby providing valuable insights into the broader mechanisms of influence of different voice characteristics on predicting the outcome of interest, in our case, the respiratory quality of life (RQoL).

2. Materials and methods

2.1. Data

The data set was collected within the Colive Voice study[1], with the goal of identifying vocal biomarkers for screening and monitoring various chronic diseases and common health symptoms. A multilingual audio databank in English, French, German, and Spanish languages is acquired with voice recordings of diverse vocal tasks, such as sustained vowel phonation, reading, counting, coughing and breathing. Voice recordings are further linked to annotated clinical and socio-demographic information, collected via self-reported validated disease-specific questionnaires on symptoms, treatments, and quality of life. The National Research Ethics Committee in Luxembourg (CNER)

[1]https://www.colivevoice.org

10.21437/Interspeech.2024-156

Table 1: *Study population characteristics*

	Total				**Normal RQoL** ($VQ11 < 22$)				**Impaired RQoL** ($VQ11 \geq 22$)			
Participants	1842				921 (50%)				921 (50%)			
Mean VQ11 score	21.60 (8.20)				14.94 (3.03)				28.25 (6.06)			
Gender	F 1240 (67.3%)	M 582 (31.6%)	O 20 (1.1%)		F 620 (67.3%)	M 291 (31.6%)	O 10 (1.1%)		F 620 (67.3%)	M 291 (31.6%)	O 10 (1.1%)	
Age	42.29 (14.16)				42.27 (14.15)				42.31 (14.18)			
Language	EN 1180 (64.1%)	FR 604 (32.8%)	DE 32 (1.7%)	ES 26 (1.4%)	EN 590 (64.1%)	FR 302 (32.8%)	DE 16 (1.7%)	ES 13 (1.4%)	EN 590 (64.1%)	FR 302 (32.8%)	DE 16 (1.7%)	ES 13 (1.4%)
Audio duration		/a/ 8:30:17	Reading 14:00:11			/a/ 4:26:33	Reading 6:56:37			/a/ 4:03:44	Reading 7:03:34	

F - Female; M - Male; O - Other; EN - English; FR - French; DE - German; ES - Spanish.

granted approval for the study (reference number 202103/01). Participants were required to be at least 15 years old and were provided an electronic informed written consent. Additionally, the study protocol for Colive Voice is registered on ClinicalTrials.gov (reference number NCT04848623).

A part of the Colive Voice data set focused on examining the RQoL in the general population was extracted for this study, with voice recordings of reading and sustained vowel phonation tasks (vowel /a/) and annotations obtained through the self-administered VQ11 questionnaire [21]. The VQ11 questionnaire is composed of 11 items rated on a scale of five categories (not at all, a little, moderately, much, extremely), and assigned a value from 1 to 5. The total score is calculated by summing individual item scores, resulting in a value ranging from 11 to 55, with lower values indicating better RQoL [21]. The cutoff value of 22 was used to stratify the participants into impaired RQoL (VQ11 $\geq$ 22), and normal RQoL (VQ11 $<$ 22) [22, 23].

Due to a substantial imbalance in the number of participants with impaired RQoL compared to those with normal RQoL, a balanced subset was created matched by age, gender and language to minimize the influence of potential confounding factors. This subset comprises a total of 1842 recordings of sustained vowel phonation and 1842 recordings of the reading task, evenly distributed between two groups. Sustained vowel phonations were produced as long as possible at a comfortable pitch and loudness, whereas the reading task assumed reading the 25$^{\text{th}}$ article from the Human Rights Declaration in the language of participant's choice (English, French, German or Spanish). Study population characteristics are summarized in Table 1. Note that no publicly available data sets for vocal biomarkers with multiple vocal tasks and similar characteristics were available on a similar scale.

2.2. Data Preprocessing

Participants were instructed to record voice in a quiet environment to maintain the quality by minimizing ambient noise. A correct example for each voice task was provided in the language of their choice. Nevertheless, to cope with the challenges of data collection in the wild with variations in devices, microphones and recording conditions, audio preprocessing was implemented to harmonize the recordings. This includes DC removal, resampling to 16 kHz, converting stereo to mono, removing leading and trailing silences, and peak normalization.

2.3. Feature Extraction

Deep audio embeddings were extracted from each recording using a self-supervised pretrained model BYOL-A (Bootstrap Your Own Latent for Audio) [24, 25]. BYOL-A processes normalized 96x64 bin log-mel spectrograms and generates two augmented versions of each spectrogram by applying pitch shifting and time stretching. These augmented versions are then fed into two parallel networks — the online and target networks. The online network predicts the output representation of the target network, and this prediction is iteratively refined as an exponential moving average of the parameters of the online network. The model produces 2048-dimensional feature vectors at its output [24].

2.4. Multimodal Fusion with Vector Cross-Attention

2.4.1. Single-Head Vector Cross-Attention

To fuse the audio embeddings coming from two voice tasks (sustained vowel phonation and reading task) and generated via the BYOL-A feature extractor, we adapt the standard cross-attention mechanism that works on sequences of embeddings to estimate the importance of each time-step embedding in a target modality using every time-step embedding of a source modality [26, 27]. Given that we do not operate on sequences of embeddings, but rather have a single embedding per modality for each instance in our dataset, we want to determine the relevance of each position in the first modality embedding (sustained vowel phonation) to every position in the second modality embedding (reading task) and vice versa.

Let us define the queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ of the attention mechanism, which are determined via matrix multiplication of the learnable attention weight matrices $\mathbf{W_q} \in \mathbb{R}^{d_m \times d_m}$, $\mathbf{W_k} \in \mathbb{R}^{d_m \times d_m}$, and $\mathbf{W_v} \in \mathbb{R}^{d_m \times d_m}$ and the input embedding $\mathbf{x} \in \mathbb{R}^{d_m}$ [28]:

$$\begin{aligned} \mathbf{Q} &= \mathbf{W_q} \cdot \mathbf{x} \\ \mathbf{K} &= \mathbf{W_k} \cdot \mathbf{x} \\ \mathbf{V} &= \mathbf{W_v} \cdot \mathbf{x} \end{aligned} \tag{1}$$

Note here that as opposed to standard cross-attention that operates on matrices, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are vectors of length d_m. Therefore, we name the approach the vector cross-attention.

To determine the cross-modal interactions by adapting the source modality to the target modality, the source modality is

given by keys $\mathbf{K}$ and values $\mathbf{V}$, whereas a target modality is defined by a query $\mathbf{Q}$. The vector cross-modal module then maps modality 1 (sustained vowel phonation) to modality 2 (reading) and outputs the sustained vowel phonation embedding adapted to reading embedding $\mathbf{x}_{1\to2}$, and vice versa.

$$\mathbf{x}_{1\to2} = \mathrm{Att}(\mathbf{Q_2}, \mathbf{K_1}, \mathbf{V_1}) = \mathrm{Softmax}\left(\frac{\mathbf{Q_2}{\mathbf{K_1}}^T}{\sqrt{d_m}}\right)\cdot \mathbf{V_1}$$
$$\mathbf{x}_{2\to1} = \mathrm{Att}(\mathbf{Q_1}, \mathbf{K_2}, \mathbf{V_2}) = \mathrm{Softmax}\left(\frac{\mathbf{Q_1}{\mathbf{K_2}}^T}{\sqrt{d_m}}\right)\cdot \mathbf{V_2} \quad (2)$$

The softmax function is used here to calculate the attention scores, while the notation $[1 \to 2]$ denotes that modality 1 attends to modality 2. Finally, to fuse the adapted features concatenation of the attended embeddings is performed:

$$\mathbf{x_{fused}} = \mathrm{Concat}(\mathbf{x}_{1\to2}, \mathbf{x}_{2\to1}) \quad (3)$$

where $\mathbf{x_{fused}} \in \mathbb{R}^{2\cdot d_m}$. This defines a single-head cross-attention (SHCA).

2.4.2. *Multi-Head Vector Cross-Attention*

In multi-head cross attention (MHCA), instead of performing a single attention function, the queries, keys, and values are linearly projected n times, i.e. split across n heads, and the attention function is then performed in parallel on each of these projected versions of queries, keys and values. After concatenating the outputs from all heads, they are once again linearly projected, resulting in the final attended embeddings [28]:

$$\begin{aligned}\mathbf{x}_{1\to2} &= \mathrm{MultiHead}(\mathbf{Q_2}, \mathbf{K_1}, \mathbf{V_1}) \\ &= \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head}_n)\cdot \mathbf{W_o}\end{aligned} \quad (4)$$

where $\mathbf{W_o} \in \mathbb{R}^{d_m\times d_m}$ is a learnable output weight matrix. In the similar way one can determine $\mathbf{x}_{2\to1}$, whereas $\mathbf{x_{fused}}$ is obtained according to equation 3 by a simple concatenation of $\mathbf{x}_{1\to2}$ and $\mathbf{x}_{2\to1}$, as in the SHCA.

2.5. Prediction and Evaluation

The training process employed a single GPU (NVIDIA Quadro RTX 6000) for 10 epochs with a learning rate set at 10^{-4}. We used Python (v3.9.16) and PyTorch (v2.0.1, CUDA v11.8) for all experiments and analyses in this study. The model was trained with AdamW optimizer, cosine annealing learning rate scheduler and binary cross-entropy loss function.

To mitigate overfitting and ensure robust performance estimation and selection of optimal hyperparameters, nested 5×5 stratified cross-validation was used for hyperparameter optimization in the inner loop (maximum accuracy was the criterion for selection), whereas the outer loop was utilized to evaluate the generalization performance on 5 held-out test folds. Stratified cross-validation preserves the class distribution across different folds [29]. Hyperparameters were selected by grid search with bounds shown in Table 2. Features were standardized to zero mean and unit variance with the statistics calculated over each training cross-validation subset to prevent data leakage.

Standardized deep audio embeddings extracted using BYOL-A from the sustained vowel phonation and the reading task were fed to a classification head for prediction, which was composed of one hidden layer with the number of nodes selected by hyperparameter tuning, rectified linear unit (ReLU) activation function and dropout, and an output layer with the sigmoid activation function for the binary classification task. The same classification head was added on top of multimodal fusion module for single-head and multi-head vector attention, where its parameters are jointly learned with the parameters of the attention module.

Table 2: *Hyperparameter optimisation*

Hyperparameter	**Range**
Number of hidden layer nodes	64, 128, 256
Mini-batch size	64, 128, 256
Dropout rate	0.2, 0.3, 0.4
Number of heads	8, 16, 32

The task was to predict whether the participant belongs to a class of impaired RQoL (VQ11 $\geq$ 22) or normal RQoL (VQ11 $<$ 22) based solely on voice recordings. The evaluation was performed with accuracy, sensitivity (recall, true positive rate), specificity (true negative rate), precision, F1 score, and area under the receiver operating characteristic curve (AUROC). All performance metrics were provided as mean values over five outer cross-validation folds and 95% confidence intervals (in parenthesis) determined via bootstrapping approach with 1000 bootstrap samples [30].

3. Experimental results

We evaluated the RQoL from two types of voice recordings, i.e. sustained vowel phonation and reading task. We furthermore applied three multimodal fusion strategies to combine these two voice inputs: early (feature-level) fusion where features extracted from two separate modalities were simply concatenated before being fed to a classifier, and fusion with single-head and multi-head vector cross-attention, as explained in Section 2.4.

The performance evaluation is shown in Table 3. Vector SHCA has shown the best overall performance, with accuracy equal to 66.1%, F1 score of 63.7% and AUROC of 0.71. Sensitivity was slightly higher for the vector MHCA (60.4%), leading to a better balance between sensitivity/specificity and precision/recall. However, this improvement of vector cross-attention over the individual voice tasks and early fusion comes at the cost of increased complexity and runtime.

Figure 1 presents the confusion matrix for the best performing model (vector SHCA), showing that the performance was in general better for the normal RQoL than for the impaired RQoL. Figure 2 shows the average ROC curve across five outer nested

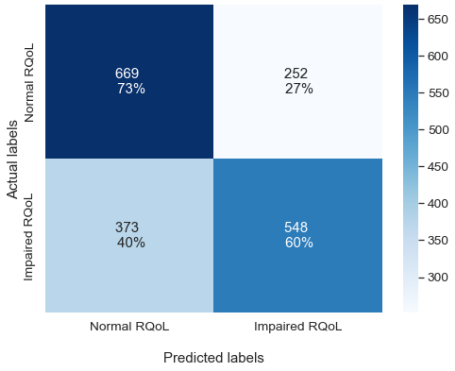

Figure 1: *Confusion matrix of the best model (multimodal fusion with single-head vector cross-attention)*

Table 3: *RQoL assessment based on sustained vowel phonation, reading task and various multimodal fusion strategies. Mean values and 95% confidence intervals over all folds are reported.*

Modality	Accuracy [%]	Sensitivity [%]	Specificity [%]	AUROC	Precision [%]	F1 score [%]	Runtime
Vowel phonation (/a/)	63.8 (59.0-68.7)	55.4 (48.4-62.3)	72.3 (65.7-78.7)	0.69 (0.64-0.74)	66.6 (58.8-74.0)	60.3 (54.0-66.3)	0.2 s/epoch
Reading	61.9 (56.9-66.8)	58.3 (51.2-65.6)	65.5 (58.5-72.0)	0.65 (0.59-0.70)	63.0 (55.7-69.9)	60.4 (54.2-66.2)	0.2 s/epoch
Early fusion	65.2 (60.3-70.0)	59.7 (52.8-66.6)	70.7 (64.0-77.0)	0.70 (0.65-0.75)	67.4 (59.9-74.6)	63.2 (57.2-69.0)	0.3 s/epoch
Fusion (SHCA)	**66.1 (61.3-70.8)**	59.5 (52.5-66.5)	**72.6 (66.1-78.9)**	**0.71 (0.66-0.76)**	**68.5 (61.0-75.7)**	**63.7 (57.7-69.4)**	14.6 s/epoch
Fusion (MHCA)	65.4 (60.5-69.9)	**60.4 (53.3-67.3)**	70.4 (63.7-76.5)	0.70 (0.65-0.75)	67.1 (59.9-74.0)	63.5 (57.4-69.0)	17.7 s/epoch

cross-validation folds, with shaded area denoting the standard deviation.

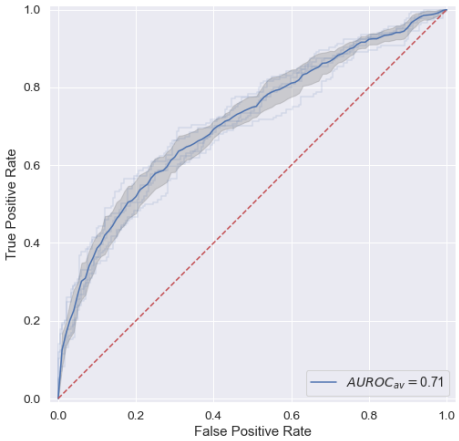

Figure 2: *ROC curve of the best model (multimodal fusion with single-head vector cross-attention). Light blue lines denote the ROC curves across five outer nested cross-validation folds, whereas a thick blue line represents the average ROC curve. Standard deviation is highlighted with the shaded area.*

4. Discussion

Sustained vowel phonation is commonly used in vocal biomarkers research, since it provides a stable and reproducible acoustic signal, allowing to capture consistent vocal characteristics over time, without the variability introduced by speech, semantics, or emotional content [4, 5]. This makes it practical for monitoring and diagnosing certain pathologies, including neurological disorders, laryngeal pathologies or respiratory conditions [1, 2]. On the other hand, reading task assumes natural speech production, allowing to capture aspects of vocal behavior that are more representative of everyday communication, and represent subtle changes related to articulatory precision and speech fluency [6]. As it involves linguistic content and semantics, the reading task may be a preferred choice for investigating vocal biomarkers associated with cognitive disorders, emotional or psychological states [31]. However, in many cases, all of these aspects turn out to be highly relevant for monitoring or identifying certain healthcare conditions. Therefore, using multiple voice tasks at the same time can provide a more comprehensive and clinically relevant approach for understanding vocal behavior, ensuring that a broader range of vocal characteristics is considered.

In this work we employ different multimodal fusion strategies to combine two voice tasks, showing that multimodal fusion substantially outperforms each of the individual voice tasks. Even with a simple concatenation of the feature vectors for sustained vowel phonation and reading tasks (early fusion), an improvement of 1.4% up to 3.3% in terms of accuracy, and approximately 3% measured by F1 score are obtained. This proves that sustained vowel phonation and reading tasks are indeed complementary for the evaluation of RQoL from voice.

Multimodal fusion with vector SHCA reflects more closely the relationships between two voice tasks, allowing for further improvement of all metrics in comparison to early fusion. Even though the audio embeddings are not inherently sequential, cross-attention mechanism facilitates the alignment of information across modalities, enabling the model to capture meaningful dependencies. By allowing modalities to attend to each other, the model learns to selectively combine information from each modality, resulting in more informative and discriminative features. Given the high average Pearson correlation between the corresponding embeddings for sustained vowel phonation and reading tasks ($r = 0.72$), single attention head was sufficient to effectively capture the main relationships between the embeddings, which may explain the slightly lower performance of vector MHCA.

Sensitivity is, in general, in all setups lower than specificity, i.e. the models are better at predicting instances of the negative class (normal RQoL) than the positive class (impaired RQoL). Combining information from multiple voice tasks via vector cross-attention slightly increased specificity, but more substantially improved sensitivity, achieving a more optimal trade-off between them (see Figure 2), and leading to more robust vocal biomarkers. However, RQoL evaluated from patient reported outcomes can be influenced by a wide range of factors, including subjective perceptions of health, or comorbidities, which can be challenging to predict accurately using voice only.

The main limitation of the vector cross-attention is the increased computational cost, caused by quadratic complexity of the scaled dot-product attention with respect to the input length, limiting its use in resource-constrained computational settings. Future work will focus on reducing the vector cross-attention to linear complexity to decrease computational costs, as well as incorporating health metadata (age, body mass index, symptoms, comorbidities, etc.) as an additional modality to provide contextual information for more accurate predictions.

5. Conclusions

This study explores the efficacy of integrating sustained vowel phonation and reading voice tasks through various multimodal fusion strategies for the evaluation of respiratory quality of life from voice. Our findings reveal that multimodal fusion with single-head and multi-head vector cross-attention substantially enhances the performance of individual modalities, emphasizing the importance of capturing inter-task relationships, and yielding more robust vocal biomarkers. Integrating these two vocal tasks ensures a more nuanced and comprehensive representation of the patient's vocal characteristics, allowing the model to selectively focus on relevant information in one modality by leveraging insights derived from another modality.

6. References

[1] G. Fagherazzi, A. Fischer, M. Ismael, and V. Despotovic, "Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice," *Digit. Biomark.*, vol. 5, no. 1, pp. 78–88, 2021.

[2] J. D. S. Sara, D. Orbelo, E. Maor, L. O. Lerman, and A. Lerman, "Guess what we can hear—novel voice biomarkers for the remote detection of disease," *Mayo Clin. Proc.*, vol. 98, no. 9, pp. 1353–1375, 2023.

[3] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task," *J. Acoust. Soc. Am.*, vol. 134, no. 3, pp. 2171–2181, 2013.

[4] B. R. Gerratt, J. Kreiman, and M. Garellek, "Comparing measures of voice quality from sustained phonation and continuous speech," *J. Speech Lang. Hear. Res.*, vol. 59, no. 5, pp. 994–1001, 2016.

[5] S. Arora, L. Baghai-Ravary, and A. Tsanas, "Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice," *J. Acoust. Soc. Am.*, vol. 145, no. 5, pp. 2871–2884, 2019.

[6] S. M. Tasko and M. D. McClean, "Variations in articulatory movement with changes in speech task," *J. Speech Lang. Hear. Res.*, vol. 47, no. 1, pp. 85–100, 2004.

[7] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth, "Multimodal assessment of Parkinson's disease: A deep learning approach," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 4, pp. 1618–1630, 2019.

[8] S. Unais, L. K. Gokul, S. Sanjana, K. Tarun, B. Rahul, P. Sunny, B. Kriti, and C. Anaghaa, "A deep-learning based multimodal system for Covid-19 diagnosis using breathing sounds and chest X-ray images," *Appl. Soft Comput. J.*, vol. 109, p. 107522, 2021.

[9] K. Nguyen-Trong and K. Nguyen-Hoang, "Multi-modal approach for COVID-19 detection using coughs and self-reported symptoms," *J. Intell. Fuzzy Syst.*, vol. 44, no. 3, p. 3501–3513, 2023.

[10] V. Richter, M. Neumann, H. Kothare, O. Roesler, J. Liscombe, D. Suendermann-Oeft, S. Prokop, A. Khan, C. Yavorsky, J.-P. Lindenmayer, and V. Ramanarayanan, "Towards multimodal dialog-based speech & facial biomarkers of schizophrenia," in *Proc ICMI 2022*, 2022, p. 171–176.

[11] M. Martinc and S. Pollak, "Tackling the ADReSS Challenge: A Multimodal Approach to the Automated Recognition of Alzheimer's Dementia," in *Proc. INTERSPEECH 2020*, 2020, pp. 2157–2161.

[12] D. Escobar-Grisales, T. Arias-Vergara, C. D. Ríos-Urrego, E. Nöth, A. M. García, and J. R. Orozco-Arroyave, "An Automatic Multimodal Approach to Analyze Linguistic and Acoustic Cues on Parkinson's Disease Patients," in *Proc. INTERSPEECH 2023*, 2023, pp. 1703–1707.

[13] D. Bhattacharya, N. K. Sharma, D. Dutta, S. R. Chetupalli, P. Mote, S. Ganapathy, C. Chandrakiran, S. Nori, K. K. Suhail, S. Gonuguntla, and M. Alagesan, "Coswara: A respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection," *Sci. data*, vol. 10, no. 1, 2023.

[14] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge," in *Proc. INTERSPEECH 2021*, 2021, pp. 3780–3784.

[15] J. W. Schwoebel, J. Schwartz, L. A. Warrenburg, R. Brown, A. Awasthi, A. New, M. Butler, M. Moss, and E. K. Pissadaki, "A longitudinal normative dataset and protocol for speech and language biomarker research," *medRxiv*, 2021.

[16] W. Pan, F. Deng, X. Wang, B. Hang, W. Zhou, and T. Zhu, "Exploring the ability of vocal biomarkers in distinguishing depression from bipolar disorder, schizophrenia, and healthy controls," *Front. Psychiatry*, vol. 14, 2023.

[17] A. Elbéji, L. Zhang, E. Higa, A. Fischer, V. Despotovic, P. V. Nazarov, G. Aguayo, and G. Fagherazzi, "Vocal biomarker predicts fatigue in people with covid-19: results from the prospective predi-covid cohort study," *BMJ Open*, vol. 12, no. 11, 2022.

[18] J. Han, T. Xia, D. Spathis, E. Bondareva, C. Brown, J. Chauhan, T. Dang, A. Grammenos, A. Hasthanasombat, A. Floto, P. Cicuta, and C. Mascolo, "Sounds of COVID-19: exploring realistic performance of audio-based digital testing," *npj Digit. Med.*, vol. 5, no. 1, pp. 1–9, 2022.

[19] V. Despotovic, M. Ismael, M. Cornil, R. Mc Call, and G. Fagherazzi, "Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results," *Comput. Biol. Med.*, vol. 138, p. 104944, 2021.

[20] T. Dang, J. Han, T. Xia, D. Spathis, E. Bondareva, C. Siegele-Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, R. A. Floto, P. Cicuta, and C. Mascolo, "Exploring longitudinal cough, breath, and voice data for COVID-19 progression prediction via sequential deep learning: Model development and validation," *J Med Internet Res*, vol. 24, no. 6, p. e37004, Jun 2022.

[21] G. Ninot, F. Soyez, and C. Préfaut, "A short questionnaire for the assessment of quality of life in patients with chronic obstructive pulmonary disease: psychometric properties of VQ11," *Health Qual. Life Outcomes*, vol. 11, no. 179, 2013.

[22] J. Rubenstein, M. Zysman, F. L. Guillou, M.-H. Colson, C. Pochulu, L. Grassion, R. Escamilla, D. Piperno, J. Pon, and C. Raherison-Semjen, "COPD burden on sexual well-being," *Eur. Respir. J.*, vol. 56, no. suppl 64, 2020.

[23] I. Anane, F. Guezguez, H. Knaz, and H. Ben Saad, "How to stage airflow limitation in stable chronic obstructive pulmonary disease male patients?" *Am. J. Men's Health*, vol. 14, no. 3, 2020.

[24] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for Audio: Self-supervised learning for general-purpose audio representation," in *Proc. IJCNN 2021*, 2021.

[25] ——, "BYOL for Audio: Exploring pre-trained general-purpose audio representations," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 137—-151, 2023.

[26] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. ACL 2019*, A. Korhonen, D. Traum, and L. Màrquez, Eds., 2019, pp. 6558–6569.

[27] V. Rajan, A. Brutti, and A. Cavallaro, "Is cross-attention preferable to self-attention for multi-modal emotion recognition?" in *Proc. ICASSP 2022*, 2022, pp. 4693–4697.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS 2017*, vol. 30, 2017.

[29] A. A. H. de Hond, V. B. Shah, I. M. J. Kant, B. Van Calster, E. W. Steyerberg, and T. Hernandez-Boussard, "Perspectives on validation of clinical predictive algorithms," *npj Digit. Med.*, vol. 6, no. 1, pp. 1–3, 2023.

[30] L. Ferrer and P. Riera, "Confidence intervals for evaluation in machine learning [computer software]." [Online]. Available: https://github.com/luferrer/ConfidenceIntervals

[31] I. Martínez-Nicolás, F. Martínez-Sánchez, O. Ivanova, and J. Meilán, "Reading and lexical-semantic retrieval tasks outperforms single task speech analysis in the screening of mild cognitive impairment and Alzheimer's disease," *Sci Rep.*, vol. 13, no. 1, 2023.