INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Towards reference speech characterization for health applications

Catarina Botelho[1], *Alberto Abad*[1], *Tanja Schultz*[2], *Isabel Trancoso*[1]

[1]INESC-ID/Instituto Superior Técnico, University of Lisbon, Portugal
[2]Cognitive Systems Lab (CSL), University of Bremen, Germany

catarina.t.botelho@tecnico.ulisboa.pt, {alberto.abad, isabel.trancoso}@inesc-id.pt,
tanja.schultz@uni-bremen.de

Abstract

Speech has been used as a biomarker for the binary classification of multiple diseases, with promising results. However these speech affecting diseases often co-exist in the same individual and produce similar manifestations in the speech signal. Thus we propose to characterize normative speech using reference intervals for interpretable speech features (acoustic and linguistic), as a first step towards the adoption of speech analysis for multidisease screening in health applications. We discuss the impact of demographics and speech tasks. Finally, we compare the reference intervals with subjects suffering from Parkinson's disease, Alzheimer's disease and depression.

Index Terms: Reference intervals, pathological speech, trustworthy biomarker

1. Introduction

Speech is a rich biosignal that encodes a great deal of information about the speaker, including hints for the presence of a plethora of diseases (e.g. neurological, psychiatric, and respiratory disorders). Many researchers have leveraged this to propose systems to perform the binary classification of healthy controls and single diseases (e.g. Alzheimer's disease (AD) [1, 2], depression [3], Parkinson's disease (PD) [4, 5], and obstructive sleep apnea (OSA) [6, 7]). In real scenarios however, these diseases are not isolated from each other, and often co-exist. The coexistence of two or more chronic conditions in the same individual, or *multimorbidity*, is common and has been rising in prevalence over recent years [8], both in in developed countries [9], and low- and middle-income countries [10]. The problem of multimorbidity gains special importance in the context of an aging population, where the coexistence of multiple diseases tends to be the norm and not the exception [8].

PD and AD, for instance, are a risk factor for depression, as well as the converse [11]. OSA has also been associated with affective disorders, and decline of cognitive functions. Although evidences for these interrelationships vary and further researcher is needed, several studies report an association between OSA and depression, possibly mediated by disturbed sleeping patterns, and/or obesity, which is a major risk factor for OSA; and between OSA and cognitive impairments, possibly mediated through repetitive hypoxemia [12].

Besides the fact that these diseases often co-exist, it is also important to note that their manifestations on the speech signal partly overlap with each other. For example, both depression ad PD have been associated with psychomotor retardation which, in turn, causes disordered articulation, reduced speech rate and affected the laryngeal control. For these two reasons, we hypothesize that a speech-based tool to support medical diagnosis and monitoring of chronic conditions, as is the case of AD, PD, and OSA, should address health from a holistic perspective, and allow an interpretative assessment of multiple diseases, rather than providing a binary classification between a given disease and healthy controls. However, existing datasets for disease detection are often small and labeled for individual diseases. Naive combination of different datasets containing individuals with a single specific disease to perform a cross-corpora study for multi-disease classification would most likely result in unreliable results that would not properly generalize to unseen recording conditions [13, 14].

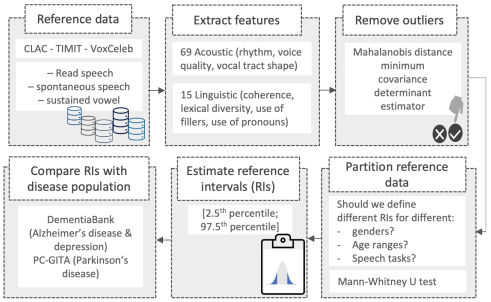

Figure 1: *Pipeline for characterizing reference speech.*

With this in mind, we claim that a valuable step towards the adoption of speech and language technologies in real health applications would be to obtain a definition of control healthy speech that could be used independently of the dataset of origin, and later be applied to identify disease signatures. We propose that this definition is based on reference intervals (RIs), which describe the typical values (usually the central 95% interval) of certain speech and language features within a reference healthy population. RIs are a concept typically applied in the context of laboratory results for medical diagnosis and monitoring. Here, we explore that concept in the context of pathological speech.

This paper presents our first efforts in this direction of characterizing reference speech. First, we define a knowledge-based feature set informed by previous literature on the expected manifestations in different diseases, including linguistic and acoustic features. A key aspect of these features is that they should be explainable and have some physical meaning, to allow meaningful discussion with the medical community. Then, we define RIs for these speech features using three English corpora (CLAC [15], VoxCeleb [16], and TIMIT [17]). In this pilot work, we compare the derived RIs with the Pitt Corpus of Dementiabank [18], which is annotated for dementia and depression, and with the subset of sustained vowels of PC-GITA [19],

This work was funded by Portuguese national funds through Fundação para a Ciência e a Tecnologia, with references UIDB/50021/2020 and SFRH/BD/149126/2019, and by the Recovery and Resilience Plan and Next Generation EU European Funds, with reference C644865762-00000008 Accelerat.AI.

10.21437/Interspeech.2023-1435

annotated for Parkinson's disease. Future work should extend the reference population, the feature set, and the validation of the RIs with corpora defined for other diseases.

2. Related Work

RIs describe the typical values of certain parameters within a reference healthy population. When a result falls out of the RI, it does not indicate the presence of a disease, but rather the necessity for medical follow up and investigation of the possible underlying reasons. On the other hand, results within the RI do not necessarily imply minimal risk of an adverse clinical outcome [20]. There are two main approaches for deriving RIs: the *direct approach* and the *indirect approach*. The *direct approach* refers to the traditional method that starts with selecting a population and collecting the samples. The *indirect approach* consists of performing data mining of results that already integrate existing routine pathology databases, that were collected as part of routine clinical testing. In this work, we propose a method parallel to the indirect approach, where we leverage speech recordings from existing corpora designed for different purposes. The guidelines for RI estimation recommend a minimum of 400 subjects in each partition.

3. Method

The general pipeline is summarized in figure 1. A more detailed description follows bellow[1].

3.1. Feature set

We defined a feature set of 84 interpretable features (acoustic and linguistic) that have been associated with the manifestation of different diseases in speech. To extract the linguistic features, we first generated automatic transcriptions using Whisper [21]. Five out of the 15 linguistic features, extracted with the BlaBla toolkit [22], capture the lexical diversity or the use of fillers (content density, idea density, Honore statistic, Brunet index, type-token-ratio, and discourse marker rate). Four features capture the coherence using cosine similarity between sentence embeddings [23] of consecutive sentences. Two features capture the use of ambiguous pronouns, by computing the ratio of the number of referenciation chains that start with a 3rd person pronoun. The referenciation chains were computed with the model proposed in [24]. Finally, one feature captures the use of first person pronouns. Regarding acoustic features, we used Praat (using the python package praat-parselmouth) and openSMILE [25] (eGeMAPS configuration [26]). We consider 17 features that capture speech-rhythm information (e.g. articulation rate, speech rate, and average syllable duration), 32 features to capture voice quality (e.g. *F0*, jitter and shimmer features), and 20 features that capture the vocal tract shape (formant features).

3.2. Reference population

For selecting the corpora for the reference population, we took CLAC, a dataset collected on purpose to serve as a speech corpus of healthy English speakers and two other corpora totally unrelated to the study of diseases from speech: TIMIT and VOXCELEB. The criteria for choosing the last two corpora was the availability of gender and age information.

CLAC includes speech from 1,832 speakers almost all located in the USA, recruited via a crowdsourcing platform. The subsets of the corpus used in this study include two read passages, two picture description task (Cookie Theft picture and a picnic picture, typically used in the diagnosis of cognitive impairment), and a sustained vowel /a/ from each speaker that claims to have no health-related symptoms that might affect their speech.

[1]Code available at `https://github.com/mcatarinatb/reference-speech-characterization`.

Table 1: *Number of audio files (F) and speakers (spk) per speech task and dataset, by gender and age range.*

		$\text{CLAC}_{picture}$		CLAC_{read}		CLAC_{vowel}		TIMIT		VoxCeleb		All	
		F	Spks	F	Spks	F	Spks	F	Spks	F	Spks	F	Spks
M	<50	1041	740	1478	763	761	761	415	415	851	287	4546	1489
M	≥50	129	97	199	104	97	97	13	13	686	193	1124	312
F	<50	1061	742	1504	757	739	739	181	181	794	277	4279	1216
F	≥50	183	137	273	137	137	137	10	10	277	90	880	241
All		2414	1716	3454	1761	1734	1734	619	619	2608	826	10829	3237

TIMIT contains speech from 630 native speakers of American English each reading 10 phonetically-rich sentences. The speakers were screened by a professional speech pathologist. One subject was excluded for lack of age information.

VoxCeleb includes short clips from interviews uploaded to YouTube. A subset of VoxCeleb contains annotations with age, gender, and nationality [27]. In this study, we included only the 840 subjects with available age information, that were from the USA, to avoid the interviews that are not spoken in English.

In VoxCeleb and TIMIT, we concatenated several turns of each participant, such that each audio segment would have roughly one minute of audio. In the reference data, we have five pairs of *dataset–task*: *CLAC–read_speech*, *CLAC–picture*, *CLAC–vowel_a*, *TIMIT–read_speech*, and *VoxCeleb–interview_segments*. Both *CLAC–picture* and *VoxCeleb–interview_segments* are examples of spontaneous speech, but still different tasks. In this document, to distinguish between the sustained vowels and all other speech tasks, we refer to vowel- and sentence-based tasks. It is also important to mention that the linguistic features will only be analysed in the context of the picture description and not the interview segments, nor any other tasks. The reason for this is that the segments are only short snippets of the interview, concatenated together, without enforcing any consecutive order, or any meaningful linguistic content. For the vowel task, we will use only the acoustic features, excluding speech-rhythm based features.

Outlier Removal

After extracting the features, we performed outlier removal using the Mahalanobis distance, and the Minimum Covariance Determinant estimator to determine the covariance matrix. Data samples with a Mahalanobis distance higher than a threshold, defined using the interquartile range method, were considered outliers. We only used a subset of 34 acoustic features for this analysis, to avoid considering features that were expected to be highly correlated. The outliers for the vowel task were estimated separately from the sentence-based tasks. This outlier removal method excluded a total of 464 (60 vowels and 404 of the sentence-based) samples, out of the original pool of 11,293 audio samples. Table 1 shows the number of audio files and speakers in each speech task and dataset, after outlier removal.

Partitioning the reference population

Most features in the feature set are strongly impacted by different factors, which may include the speech task, noise, recording conditions, and speaker dependent factors, such as gender, age, body mass index, education, smoking habits, etc. Because we do not have information on all these variables, in this work we focus on gender and age which are the most important variables to consider for RI estimation [28]. We also explore the impact of speech tasks and source dataset. Thus, to determine whether

we should partition the reference population data to define different RIs, we formulate the following questions:

- **Q1**: Is there a statistically significant difference between the speech features of male and female speakers, that justifies estimating different RIs for each gender group?
- **Q2**: Is there a statistically significant difference between the speech features of speakers at different age ranges, that justifies estimating different RIs for each age range?
- **Q3**: Is there a statistically significant difference between the speech features extracted for different speech tasks?
- **Q4**: Is there a statistically significant difference between speech features of two dataset, that would impede combining two datasets under the same RI?

To answer these questions, we perform a Mann-Whitney U test that evaluates whether the two subgroups are likely to be derived from the same population (null hypothesis). For a *p-value* < 0.01, we rejected the null hypothesis, i.e., we conclude the two subgroups are not likely to be derived from the same population, and thus distinct RIs should be derived for the two subgroups. In Q2, we simplified the problem to two age ranges: subjects below 50 years old, and subjects with 50 years or more. We acknowledge that this analysis should be carried with finer age ranges, but such analysis was not possible considering the size of the reference population, namely for the older subset.

3.3. Computing reference intervals

If the underlying distribution of the data is Gaussian, the RI corresponds to the $mean \pm 1.96 \times std$, in which std stands for standard deviation. If such assumption cannot be made, which is frequently the case for the studies of RI estimation, either data must be first transformed to a Gaussian distribution, e.g. using a Box-Cox power transformation, or a non-parametric estimation can me made. In the case of a non-parametric estimation, the limits of the RI correspond to the 2.5^{th} and 97.5^{th} percentile [29]. [29] found that both the non-parametric approach and the power transformation of data followed by the parametric approach provide similar results, and the non-parametric method is recommended by [30]. Thus in this work, we chose the non-parametric approach. To provide a confidence measure on the estimated RI, we derive 99% confidence intervals (CIs) for both the lower and upper limits of the RI via boostrapping. Data was resampled 1000 times to estimate the CIs. If the CI for any of the reference limits is larger than 20% of the RI, then the RI is not considered valid. After the bootsrapping, we fixed each RI as the outer bounds of the CI.

3.4. The diseased population

After defining the RIs for the reference population, we compare them to data from two datasets designed for disease detection:
DementiaBank, for the analysis of AD and depression. In this English corpus, also known as Pitt corpus [18], each speaker describes the Cookie Theft picture (we removed the interviewer's turns). The DementiaBank contains several annotations, including different forms of dementia and depression. In this study, to define the subjects with AD, we relied on the "Probable AD" tag; and to define the subjects with depression we used the Hamilton scale (depressed for a value ≥ 8) [31]. Thus, we define four subgroups: control subjects, subjects that suffer from AD but not depression, subjects that suffer from depression but not AD, and subjects that suffer from both AD and depression.
PC-GITA, the Parkinson's disease Corpus from the Applied Telecommunications Group at Universidad de Antioquia, Colombia (PC-GITA) [19], for the analysis of PD. This corpus, fully spoken in Spanish, includes recordings of 50 PD patients and 50 controls matched by age and gender performing several speech tasks. To allow a fair comparison with RI determined for English, in this work we only explore the enunciation of a sustained vowel /a/ (three repetitions per subject).

Table 2: *Number of features with p-value* ≥ 0.01 *in the Mann-Whitney U test applied to questions Q1-Q4.*

		Q1	Q2	Q3	Q4	Q4–dataset normalization
Total feature count		84	84	69	69	69
features with *p-value* ≥ 0.01	All	15	–	–	–	–
	Female	–	57	24	12	65
	Male	–	58	22	9	65
	F & M	–	41	17	4	64

To allow the comparison of the disease datasets to the estimated RIs, we normalize each dataset separately. However, we use only the controls subjects to fit the scaler. Table 3 shows the number of audio samples in each dataset, per diagnosis and gender group.

4. Results

Partitioning the reference population: The results for the analysis on how to partition the RIs are summarized in Table 2. A detailed description follows below.

To answer Q1, we analyse data for a single speech task from a single dataset, to minimize other potential sources of variability: *CLAC–picture_description*. After normalizing data with 0-mean and unit variance, we compute the Mann-Whitney U test between male and female individuals, and conclude that there is no need to partition the RI for only 15 out of the 84 features. These 15 features are mostly linguistic (8), but also rhythm-related (5), and F0-related (falling slope, and max F0). Given that for the majority of the features, the recommendation is to partition, henceforth, we always compute two separate models: one for each gender, each normalized separately.

Regarding Q2, we observe that we can combine the data of both age ranges to estimate one single RI for 57 and 58 features, in the subpopulation of females, and males, respectively. The intersection of the subgroups of features that passed the test correspond to 41 features. Nevertheless, considering that partitioning by age after partitioning by gender would result in small sample sizes for each group (often smaller than the 400 subjects suggested in the guidelines), in this pilot work, we opt to not partition the data according to age.

To investigate Q3, we compare *CLAC–picture_description* and *CLAC–read_speech*, normalizing all data simultaneously. Note that we excluded the linguistic features from this analysis. We observe that we could only keep the subpopulations together for 17 out of the 69 features. Thus, we will keep the analyses of different speech tasks separate.

Regarding Q4, we compare *TIMIT–read_speech* and *CLAC–read_speech*, normalizing all data simultaneously. In this scenario, we could compute the RIs using all data together for only 4 out of the 69 features. This would mean that we could not employ the RIs estimated with one dataset to unseen datasets, even under the same gender and speech task. To tackle this, we perform dataset-dependent normalization. With this approach, we make the distributions of the two subgroups more similar, and thus can assume that the samples come from the same distribution for 64 out of the 69 features.

In summary, we conclude that we need to estimate different RIs for each gender, and each speech tasks. To combine differ-

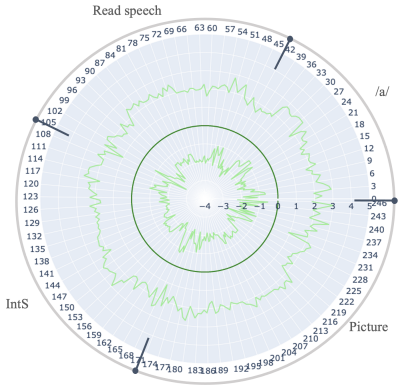

Figure 2: *Radar plot with normalized RIs for female speakers. /a/ – vowel /a/; RS – read speech; IntS – concatenated interview segments of voxceleb ; Picture – picture description only.*

Table 3: *(i) Number of audio samples. (ii) Average number of features outside of the RIs, per sample. (iii) [%] of samples with 0 features outside the RIs. Dep – depression.*

	(i) #samples				(ii) #features out of RI				(iii) [%] samples			
	Female		Male		Female		Male		Female		Male	
	C	D	C	D	C	D	C	D	C	D	C	D
PD	75	75	75	75	2.1	6.2	2.3	4.3	37.3	6.7	44.0	24
AD	77	84	46	46	4.3	5.8	3.8	7.2	11.7	2.4	8.7	0
Dep	77	3	46	7	4.3	1.3	3.8	4.3	11.7	0	8.7	0
Dep+AD	77	45	46	18	4.3	5.8	3.8	5.9	11.7	2.2	8.7	0

ent datasets under the same RI, we perform dataset-dependent normalization. This is consistent with results described in [32]. However, even under dataset-dependent normalization, 5 out of the 69 features could not be combined, and thus were excluded from further analysis.

These are simplifying assumptions, that we believe to be reasonable in this proof-of-concept exploring the feasibility of defining RIs for speech. Future work should not only study a larger reference population, but also consider other methods for partitioning the RIs, such as the Lahti criteria [33], or the Ichiahara method [29, 34].

Reference intervals: Figure 2 shows the normalized RIs for all features, for female speakers, as an example. Each region corresponds to a speech task. The light green shows the RI, while the dark green shows the mean. A similar plot was derived for male speakers. For 12 features, the RI was considered invalid because the confidence interval on at least one of the reference limits was larger than 20% of the RI estimated. Thus, those features were excluded from the analyses. This resulted in a total of 247 speech features.

Comparing RIs with datasets for disease detection: Figure 3 shows the overlap of all female subjects in the disease corpora over the RIs. Each colored line corresponds to one audio sample. Lines with a stronger intensity correspond to the superposition of multiple samples on top of each other. The figures compare control samples to disease samples separately. By visually inspecting PC-GITA (top of Figure 3), it appears that PD samples are more often outside the RIs, than control samples. The same observation is not so evident for DementiaBank, when comparing AD patients, or AD+depression patients to controls. There are only three female patients with depression only, thus they are omitted from the figure.

Table 3 summarizes the results by reporting (i) the number of samples under analysis, (ii) the average number of features outside the corresponding RI, per audio sample, and (iii) the percentage of samples with no features outside the corresponding RI. The Table shows that the average number of features outside the RI per sample is higher for patients, than for control subjects. This observation is verified for all diseases, with the exception of female depressed subjects. However, given the very small size of this subset (3), we argue that further analysis should be conducted. The Table also shows that the percentage of samples that have no features outside the corresponding RI is always much larger in controls than in diseased subjects.

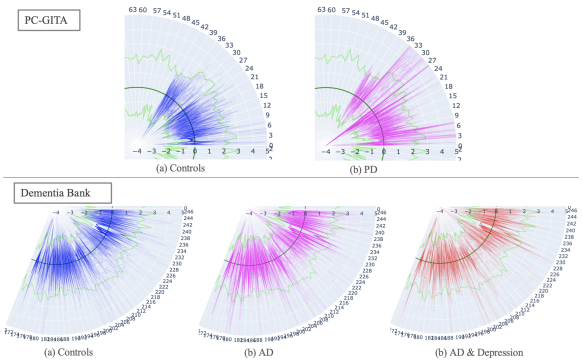

Figure 3: *Female controls and patients, over the RIs.*

The differences between patients and controls are more noticeable in PC-GITA than in DementiaBank. This may suggest that the task of enunciating a sustained vowel may be more suitable for this RI analysis, as it allows less source of variablity. It is also possible that the noisy recording conditions in DementiaBank play a strong role.

This tool also enables the identification of features that appear to be better markers of a disease, and simultaneously still robust to dataset shifts. Taking the example of PD females *vs* Control females (Figure 3 – at the top): there are 16 features for which more than 95% of the controls stay inside the RI and less than 95% of the diseased stay inside the RI. Particularly, for the feature harmonics-to-noise ratio computed with Praat (feature 25), only 3.5% of controls are out of the RI, compared to the 41.2% of patients out the RI. There are also 7 features, all related to shimmer measures, for which less than 5% of the controls fall outside the RI, and over 30% of PDs are out the RI.

5. Conclusions

This work establishes a proof-of-concept for the characterization of healthy speech using RIs. In fact, one can think of this speech analysis as a parallel to the usual blood tests. The radar plots, and particularly the RIs, appear to be useful for analysing multiple diseases from different datasets, as they enable the extraction of interpretable information which can guide further medical follow ups.

Possible future work directions include (i) replicating the method with a larger reference population, and other relevant speech tasks (e.g. diadochokinetic evaluation), (ii) exploring other statistical criteria to decide on RI partitioning, (iii) exploring alternatives to the dataset dependent normalization for allowing the comparison of RIs in different corpora, (iv) investigating how to evaluate the effectiveness of the method for individuals instead of group data, and (v) further compare the RIs with different diseases, and different datasets/speech tasks of the same disease to assess the generalizability of the results.

We believe that the RIs and the radar plot are a starting point to provide trustworthy and explainable insights to the use of speech as a biomarker for multidisease screening.

6. References

[1] S. Zargarbashi and B. Babaali, "A multi-modal feature embedding approach to diagnose Alzheimer disease from spoken language," *arXiv preprint arXiv:1910.00330*, 2019.

[2] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of Alzheimer's disease in conversational german." in *INTERSPEECH*, 2016, pp. 1938–1942.

[3] A. Afshan, J. Guo, S. J. Park, V. Ravi, J. Flint, and A. Alwan, "Effectiveness of voice quality features in detecting depression," *Interspeech 2018*, 2018.

[4] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect Parkinson's disease from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1155–1159.

[5] A. Pompili, A. Abad, P. Romano, I. P. Martins, R. Cardoso, H. Santos, J. Carvalho, I. Guimarães, and J. J. Ferreira, "Automatic detection of Parkinson's disease: An experimental analysis of common speech production tasks used for diagnosis," in *International Conference on Text, Speech, and Dialogue*. Springer, 2017, pp. 411–419.

[6] C. Botelho, I. Trancoso, A. Abad, and T. Paiva, "Speech as a biomarker for obstructive sleep apnea detection," in *ICASSP*. IEEE, 2019, pp. 5851–5855.

[7] J. M. Perero-Codosero, F. Espinoza-Cuadros, J. Antón-Martín, M. A. Barbero-Álvarez, and L. A. Hernández-Gómez, "Modeling obstructive sleep apnea voices using deep neural network embeddings and domain-adversarial training," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 240–250, 2019.

[8] WHO, "Multimorbidity: Technical series on safer primary care, https://apps.who.int/iris/bitstream/handle/10665/252275/9789241511650-eng.pdf," 2016. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/252275/9789241511650-eng.pdf

[9] K. Barnett, S. W. Mercer, M. Norbury, G. Watt, S. Wyke, and B. Guthrie, "Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study," *The Lancet*, vol. 380, no. 9836, pp. 37–43, 2012.

[10] H. H. Wang, J. J. Wang, S. Y. Wong, M. C. Wong, F. J. Li, P. X. Wang, Z. H. Zhou, C. Y. Zhu, S. M. Griffiths, and S. W. Mercer, "Epidemiology of multimorbidity in China and implications for the healthcare system: cross-sectional survey among 162,464 community household residents in southern China," *BMC medicine*, vol. 12, no. 1, pp. 1–12, 2014.

[11] K. R. R. Krishnan, M. Delong, H. Kraemer, R. Carney, D. Spiegel, C. Gordon, W. McDonald, M. A. Dew, G. Alexopoulos, K. Buckwalter *et al.*, "Comorbidity of depression with other medical diseases in the elderly," *Biological psychiatry*, vol. 52, no. 6, pp. 559–588, 2002.

[12] J. Vanek, J. Prasko, S. Genzor, M. Ociskova, K. Kantor, M. Holubova, M. Slepecky, V. Nesnidal, A. Kolek, and M. Sova, "Obstructive sleep apnea, depression and cognitive impairment," *Sleep medicine*, vol. 72, pp. 50–58, 2020.

[13] C. Botelho, T. Schultz, and I. Abad, Albertoand Trancoso, "Challenges of using longitudinal and cross-domain corpora on studies of pathological speech," *Proc. Interspeech 2022*, pp. 2516–2520, 2022.

[14] A. Ablimit, C. Botelho, A. Abad, T. Schultz, and I. Trancoso, "Exploring dementia detection from speech: Cross corpus analysis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6472–6476.

[15] R. Haulcy and J. Glass, "CLAC: A Speech Corpus of Healthy English Speakers," in *Interspeech*, 2021, pp. 2966–2970.

[16] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "VoxCeleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[17] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1992.

[18] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[19] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease." in *LREC*, 2014, pp. 342–347.

[20] Y. Ozarda, K. Sikaris, T. Streichert, J. Macri, I. C. on Reference intervals, and D. L. (C-RIDL), "Distinguishing reference intervals and clinical decision limits–a review by the IFCC committee on reference intervals and decision limits," *Critical reviews in clinical laboratory sciences*, vol. 55, no. 6, pp. 420–431, 2018.

[21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[22] A. Shivkumar, J. Weston, R. Lenain, and E. Fristed, "BlaBla: Linguistic feature extraction for clinical analysis in multiple languages," in *Interspeech*, 2020.

[23] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[24] V. Dobrovolskii, "Word-level coreference resolution," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7670–7675. [Online]. Available: https://aclanthology.org/2021.emnlp-main.605

[25] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[26] F. Eyben, K. Scherer, and B. Schuller et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[27] K. Hechmi, T. N. Trong, V. Hautamaki, and T. Kinnunen, "VoxCeleb enrichment for age and gender recognition," 2021. [Online]. Available: https://arxiv.org/abs/2109.13510

[28] Y. Ozarda, "Reference intervals: current status, recent developments and future considerations," *Biochemia medica: Biochemia medica*, vol. 26, no. 1, pp. 5–16, 2016.

[29] K. Ichihara, J. C. Boyd *et al.*, "An appraisal of statistical procedures used in derivation of reference intervals," *Clinical chemistry and laboratory medicine*, vol. 48, no. 11, pp. 1537–1551, 2010.

[30] G. L. Horowitz, S. Altaie, J. Boyd, F. Ceriotti, U. Garg, P. Horn, A. Pesce, E. Harrison, and J. Zakowski, "EP28-A3C defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline," *San Diego: Clinical and Laboratory Standards Institute*, 2010.

[31] R. Sharp, "The Hamilton rating scale for depression," *Occupational Medicine*, vol. 65, no. 4, pp. 340–340, 2015.

[32] C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso, "Pathological speech detection using x-vector embeddings," *arXiv preprint arXiv:2003.00864*, 2020.

[33] A. Lahti, P. H. Petersen, J. C. Boyd, C. G. Fraser, and N. Jørgensen, "Objective criteria for partitioning gaussian-distributed reference values into subgroups," *Clinical chemistry*, vol. 48, no. 2, pp. 338–352, 2002.

[34] K. Ichihara, Y. Itoh, and C. W. Lam et al., "Sources of variation of commonly measured serum analytes in 6 Asian cities and consideration of common reference intervals," *Clinical chemistry*, vol. 54, no. 2, pp. 356–365, 2008.