Towards reference speech characterization for health applications

Catarina Botelho¹, Alberto Abad¹, Tanja Schultz², Isabel Trancoso¹

¹INESC-ID/Instituto Superior Técnico, University of Lisbon, Portugal
²Cognitive 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 multiscreen disease 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 depres-
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 below.

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, Homonote 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 type-token-ratio, and discourse marker rate. The outliers for the vowel task were estimated using the interquartile range method, were considered outliers.

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 sub-

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>F</th>
<th>spk</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLACicture</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLACread</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLACvowel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIMIT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VoxCeleb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M &lt;30 1041</td>
<td>740</td>
<td>1716</td>
</tr>
<tr>
<td>F &gt;30 10829</td>
<td>3237</td>
<td>3237</td>
</tr>
<tr>
<td>M ≥30 1299</td>
<td>971</td>
<td>2111</td>
</tr>
<tr>
<td>F &gt;30 10829</td>
<td>3237</td>
<td>3237</td>
</tr>
<tr>
<td>M ≥30 1299</td>
<td>971</td>
<td>2111</td>
</tr>
<tr>
<td>F &gt;30 10829</td>
<td>3237</td>
<td>3237</td>
</tr>
<tr>
<td>M ≥30 1299</td>
<td>971</td>
<td>2111</td>
</tr>
<tr>
<td>F &gt;30 10829</td>
<td>3237</td>
<td>3237</td>
</tr>
</tbody>
</table>

<https://github.com/mcatarinatb/reference-speech-characterization>
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 speakers at different age ranges, that justifies estimating different RIs for each gender group?

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 ± 1.96 × 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 be made. In the case of a non-parametric estimation, the limits of the RI correspond to the 2.5th and 97.5th 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 bootstrapping. 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 bootstrapping, 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, 

<table>
<thead>
<tr>
<th>Total feature count</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4–dataset normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>84</td>
<td>84</td>
<td>69</td>
<td>69</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Features with</th>
<th>Female</th>
<th>Male</th>
<th>F &amp; M</th>
<th>F &amp; M</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value ≥ 0.01</td>
<td>57</td>
<td>24</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>58</td>
<td>22</td>
<td>65</td>
<td></td>
</tr>
</tbody>
</table>

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

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).

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.

<table>
<thead>
<tr>
<th>(i) Samples</th>
<th>(ii) Features out of RI</th>
<th>(iii) [%] samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>PD</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>AD</td>
<td>77</td>
<td>84</td>
</tr>
<tr>
<td>Dep+AD</td>
<td>77</td>
<td>46</td>
</tr>
</tbody>
</table>

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 supersposition 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, 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 assist 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.

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.

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 variability. 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.
6. References


