Challenges of using longitudinal and cross-domain corpora on studies of pathological speech

Catarina Botelho1,2, Tanja Schultz2, Alberto Abad1, 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, tanja.schultz@uni-bremen.de, {alberto.abad, isabel.trancoso}@inesc-id.pt

Abstract

Several promising works have reported very exciting results in the field of speech in health, however there are still issues to address before deploying such systems into clinical applications. One of such issues is to ensure the generalisability and reliability of results. With this in mind, in this work, we perform a comparative analysis of healthy speech in two scenarios: (1) collected for six different datasets spoken in the same language, and (2) collected across different times in a single longitudinal corpus. We show that feature sets typically used for disease detection from speech (eGeMAPS, ComParE, ECAPA-TDNN embeddings and i-vectors) encode much information about the dataset or about changing recording conditions over time, in longitudinal studies. We support our results with classification results largely above chance level for both scenarios, and through unsupervised clustering experiments, where we observe that data naturally clusters according to dataset.

Index Terms: healthy speech, cross-corpora, clustering

1. Introduction

Recent technological advances in the machine learning and deep learning field, allied with non-invasive biomarkers such as speech, have the potential to provide a route for mass screenings of multiple diseases, potentially contributing towards a scalable improvement in global healthcare. In fact, several research studies have proposed to detect a plethora of diseases from the speech signal of patients, including speech and language disorders (e.g. sigmatism, stuttering); neurodegenerative diseases such as Parkinson’s Disease (PD) [1, 2]; Alzheimer’s Disease (AD) [3, 4]; Huntington’s disease [5] and Amyotrophic Lateral Sclerosis [6]; mood and anxiety-related disorders such as depression [2] and bipolar disease [7]; and diseases that concern respiratory organs, such as obstructive sleep apnea, [8, 9], tuberculosis, and COVID-19 [10]. The references provided only cite a few examples among many existing studies.

Notwithstanding the large potential of the machine learning systems allied with noninvasive biomarkers, and the promising results reported by many researchers [11] [12], one observes that most of these systems have not yet transitioned to commercial products. Arguably, the most important issue to address before such transition takes place is to ensure the generalisability and reliability of the obtained results. Often the datasets used to train such systems have two strong limitations. The first limitation is that datasets are small, and may not be representative of an entire population. The second limitation is that they focus on a single disease versus healthy controls, whereas in real-life, and namely in the context of an aging population, the coexistence of multiple diseases in the same patient, or multimorbidity, tends to be the norm and not the exception [13].

One possibility to tackle these two limitations is cross-corpora studies. In one hand, it would allow to perform out-of-domain evaluation of the models trained in each small dataset. On the other hand it could enable the simultaneous detection of multiple diseases. However, cross-corpora studies have some challenges. The differences in languages, speech tasks and recording conditions may compromise the conclusions reached. In fact, in our recent work [14], we have explored the screening of dementia from speech in a longitudinal conversational corpus – the ILSE corpus [15] – and compared our findings with the ADReSS corpus [16]. Although we were able to achieve promising results for each dataset individually, we observed that the best feature sets for one dataset do not match the best for the other one, and also that models trained on one dataset did not perform well on the other. While the differences in terms of language, speech task, and recording conditions may explain the different results, we argue that it is crucial to discuss how to translate results across different domains, and how to measure robustness and trust of the results achieved in the research area of speech for health. This discussion is also relevant in the context of long duration longitudinal studies, where it is only natural that recording conditions change over time.

In this work, we discuss whether it is possible to mix different datasets for disease detection. Although ideally acoustic-based systems for disease detection would be language independent, recent studies have found that language-specific differences in aspects such as phonation and prosody may influence the perception of speech impairment [17]. For this reason, we restrain the problem to mixing corpora of the same language. We then compare the speech of healthy subjects collected across different datasets, using features that are typically adopted in studies of pathological speech. Concretely, we design experiments to answer the following questions:

1. How easy is it for a machine learning classifier to distinguish between recordings of healthy people collected for different datasets (or measurement time, in the case of longitudinal studies), using acoustic features typically used for disease screening from speech?
2. Are embeddings obtained with neural networks trained with data augmentation more robust to dataset changes than knowledge based features?
3. Do the acoustic data represented by the typical acoustic features used for health classification clusters around dataset (or measurement time, in the case of longitudinal studies)?

This work was supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020 and grant number SFRH/BD/149126/2019.
2. Method

As explained above, this study compares the speech of healthy subjects (i.e., subjects for which there is no reported presence of disease) across different corpora. The non-longitudinal study involves six corpora in American English: ADReSS [16], DAIC-WOZ [18], CLAC [19], TIMIT [20], VoxCeleb [21], and WSM [2]. The longitudinal study involves four measurement times of the ILSE corpus in German. The comparison was made on the basis of feature sets that are typically used in studies of disease detection from the acoustic signal, namely: eGeMAPS [22], ComParE feature set [23], i-vectors [24] and ECAPA-TDNN embeddings [25]. This section briefly describes the datasets and the set up of each of our experiments.

2.1. Corpora

In the selection of the English datasets for this study, we involved spontaneous speech datasets, read speech datasets and datasets including both. All datasets are roughly balanced in terms of gender. As a starting point, we took three datasets corresponding to the healthy control subjects of corpora collected for detecting diseases from speech: AD (ADReSS), PD (WSM), and Depression (DAIC-WOZ and WSM). To complement these subsets 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 age information that could be explored in later studies.

ADReSS includes speech recordings of 78 healthy control subjects from the Pitt corpus [26], also known as DementiaBank. Our study was restricted to the subset of this corpus used in the ADReSS challenge. Although the version used in the challenge was acoustically enhanced, we opted for using the original version in order to avoid the tonal noise.

CLAC includes speech from 1,832 speakers almost all located in the USA (15 speakers from other countries). The speakers were recruited via a crowdsourcing platform and claim to have no health-related symptoms that might affect their speech. The subsets of this corpus that were used in this study include two read passages (totaling 9 sentences) and two picture description task (Cookie Theft picture and a picnic picture, typically used in the diagnosis of cognitive impairment) from each speaker.

DAIC-WOZ includes data from Wizard-of-Oz interviews, conducted by an animated virtual interviewer controlled by a human interviewer in another room. The healthy control subset includes 145 sessions of interactions ranging between 5-20 minutes each. All subjects were fluent American English speakers.

TIMIT contains speech from 630 native speakers of American English representing 8 major dialect divisions, 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 taken from interviews uploaded to YouTube. [27] have annotated a subset of VoxCeleb 2 with age, gender, nationality, among other information. In our study, we included only the interviews of subjects from the USA with available age information, to avoid including interviews that are not spoken in English. This results in a subset of 840 subjects.

WSM consists of videos collected from online multimedia repositories (YouTube). The subset of videos in which only one subject appears in the entire video, and the speaker did not claim to suffer from PD or depression, consisted of 587 subjects, all fluent speakers of American English. The audio recordings were segmented using an in-house VAD [28].

ILSE is a longitudinal corpus of biographic interviews, where the participants gave elaborate answers to open-ended questions. All participants were born in Germany, in two distinct cohorts: 1930-1932 and 1950-1952. Although each participant engaged in up to four measurements, conducted in intervals of approximately 5 years, we used a single interview from each speaker, amounting to 356 speakers equally divided into 4 subsets. Naturally, the age of the subjects in the longitudinal corpus increases across the measurement times, but because the corpus includes two cohorts with distinct birth dates, there is always some age overlap between each of the measurement times.

We respected the train/dev/test set partitions proposed for these corpora, unless the partitions were not made available, in each case we randomly split the data to define the partitions. In ADReSS and DAIC, we excluded speech from interviewers. In interview/vlog datasets and TIMIT, we concatenated several participants turns such that each audio segment would have roughly one minute of audio. Table 1 summarizes the number of subjects and segments, as well as age and duration information for each of the datasets used. We normalized the eGeMAPS, and ComParE feature sets with zero mean and unit variance, based on the train set, for all experiments except when a different normalization strategy is explicitly described.

2.2. Experiments

Experiment A: Supervised classification of healthy speakers from distinct datasets. We perform classification using Support Vector Machines (SVMs) and each of the feature sets described above. The SVM hyperparameters (kernel, C, gamma and degree) were chosen with a grid search on the development set. The goal of this experiment is to discuss how easy is it for a learning system to detect the dataset. In experiment Experiment A-1, we perform 6-class classification, using the 6 English datasets. In experiment Experiment A-2, we perform 4-class classification to distinguish between the 4 measurement times of the longitudinal corpus ILSE. Although we expect that this experiment is harder than experiment A-1, it is just as relevant because it is natural that recording conditions change over the course of large longitudinal corpus, and the participants of the studies may only develop diseases at later measurement times. Hence, recording conditions will be a strong confounder when trying to detect diseases from speech in such data.

Experiment B: Clustering We perform agglomerative clustering to assess whether, even without supervision, data would naturally cluster around dataset. We hypothesize that the gender of the subjects may also play a relevant role when forming the clusters, thus we compare three different numbers of clusters [2, number of classes, 2 × number of classes], where number of classes corresponds to the number of datasets or measurement times. In Experiment B-1, we perform clustering over a
Figure 1: t-SNE projections of each of the feature sets, for the 6 English datasets. ■ for male and ▲ for female.

Figure 2: t-SNE projections of each of the feature sets, for the 4 measurement times of the ILSE. ■ for male and ▲ for female.

pool of data that includes all 6 English datasets. In Experiment B-2, we perform clustering over ILSE data.

3. Results

3.1. Exploring the features

Figures 1 and 2 show the t-SNE projections of each of the feature sets, for the six English datasets and ILSE, respectively. We observe that eGeMAPS and ComParE feature sets allow the projection of the different datasets/measurement times to clearly separate regions of the space. On the other hand, ECAPA-TDNN appears to provide a less clear separation between datasets. It is also visible that gender plays an important role. For example, TIMiT (in brown) forms two clear clusters, one for male and one for female speakers, in all feature sets.

3.2. Classification results

Supervised classification results are reported in terms of accuracy, and the unweighted average of recall (UAR), precision (UAP) and f1-score (UF1) across the classes. We emphasise UAR analysis, since it is the standard metric used for disease classification from speech, because the vast majority of datasets have an imbalanced class ratio [29]. The results of the classification experiments A-1, using the six English datasets are presented on the top part of Table 2. All UAR results are above 82%, and the ComParE feature set reaches 95% UAR. These results show that dataset classification is fairly easy, using these standard feature sets. We expected that ECAPA-TDNN embeddings would contain less information relevant for dataset classification, because they are discriminative representations trained with data augmentation, thus should be more robust to small domain shifts. The results confirm this expectation, i.e., ECAPA-TDNN embeddings achieve the lowest results over all metrics, but the performance gap is not as large as could be expected. i-vectors, on the other hand, model total variability, which can justify the fact that they contain more information specific to the dataset conditions, such as the task or recording conditions.

The bottom part of Table 2 shows the results after performing dataset dependent normalization of the features. We hypothesize that this normalization may remove some dataset-specific information, and make the feature sets more dataset agnostic. In fact, the performance drop after normalization confirms that hypothesis. This performance drop is particularly evident for ECAPA-TDNN embeddings and i-vectors. Nevertheless, this normalization does not remove all dataset-specific information, since results for all feature sets are highly above chance level, which would correspond to an UAR of 16.6%.

Table 3 shows the results for experiment A-2, the classification of ILSE corpus into the four measurement times, before and after measurement time-dependent feature normalization. Similarly to what was observed in experiment A-1, the ECAPA-TDNN feature set performs worse than the remaining feature sets, when trying to classify measurement time. Regarding the effects of the measurement time feature normalization, we observe that after the normalization, results are similar or worse for i-vectors, ECAPA-TDNN and ComParE feature sets, as expected. However, the reverse trend is observed for eGeMAPS, which supports the claim that they contain more information specific to the dataset conditions, such as the task or recording conditions.

The bottom part of Table 2 shows the results after performing dataset dependent classification of six distinct datasets using features often used for disease detection from speech. The results confirm this expectation, i.e., ECAPA-TDNN embeddings achieve the lowest results over all metrics, but the performance gap is not as large as could be expected. i-vectors, on the other hand, model total variability, which can justify the fact that they contain more information specific to the dataset conditions, such as the task or recording conditions.

The bottom part of Table 2 shows the results after performing dataset dependent normalization of the features. We hypothesize that this normalization may remove some dataset-specific information, and make the feature sets more dataset agnostic. In fact, the performance drop after normalization confirms that hypothesis. This performance drop is particularly evident for ECAPA-TDNN embeddings and i-vectors. Nevertheless, this normalization does not remove all dataset-specific information, since results for all feature sets are highly above chance level, which would correspond to an UAR of 16.6%.

Table 3: Experiment A-2: classification of six distinct datasets using features often used for disease detection from speech.

Table 3 shows the results for experiment A-2, the classification of ILSE corpus into the four measurement times, before and after measurement time-dependent feature normalization. Similarly to what was observed in experiment A-1, the ECAPA-TDNN feature set performs worse than the remaining feature sets, when trying to classify measurement time. Regarding the effects of the measurement time feature normalization, we observe that after the normalization, results are similar or worse for i-vectors, ECAPA-TDNN and ComParE feature sets, as expected. However, the reverse trend is observed for eGeMAPS, which supports the claim that they contain more information specific to the dataset conditions, such as the task or recording conditions.

3.3. Clustering results

Figures 3 and 4 show the results for the clustering experiments C-1 and C-2, respectively, using i-vectors as input features. We opted to display i-vectors since they seem to provide the most clear results. In each heat map displayed in figures 3 and 4, each row corresponds to one predicted cluster, and each column to a characteristic that we hypothesize could be associated with a cluster. The colors correspond to the normalized frequency of
observations of the given characteristic in each of the predicted clusters, thus the values in each column sum to 1. The most left heat map corresponds to clustering into two clusters; the center corresponds to the number of cluster equal to number of classes (six datasets in Figures 3, and four measurement times in figure 4); and the right heat map corresponds to 2 × number of classes.

Analysing Figure 3 (a) we observe that when clustering into two clusters, there is no predicted cluster that corresponds mostly to female (nor male) speaker. On the other hand, when clustering into six clusters, we do observe that each cluster corresponds roughly to each of the datasets. The exception occurs for WSM and VoxCeleb, which are placed into the same cluster – this makes sense considering that both WSM and VoxCeleb are collected from YouTube, although the former corresponds to vlog format and the latter to interview segments. TIMIT is split into two separate clusters, possibly one corresponding to male and the other to female subjects. In the the 12-cluster heat map, we observe that each cluster corresponds mostly to either male of female subjects of a single dataset, with the exception of WSM and VoxCeleb.

After dataset dependent normalization of i-vectors, Figure 3 (b), we observe that each cluster cannot be mapped anymore to single dataset. These results are consistent with the performance drop for i-vectors observed in table 2 after normalization. In Figure 4 (a), we observe that when the number of clusters for ILSE is two, there is a tendency to cluster according to gender. The 8-cluster heat map shows some mapping between predicted cluster and datasets. We observe that several of the female speakers of times 1 and 2 are grouped into cluster 0, and the male speakers of the same times are grouped into cluster 3. Most of the remaining clusters correspond to to a single gender, and a single measurement time. After measurement time dependent normalization of i-vectors, Figure 4 (b), any existing mapping between clusters and measurement times or gender vanishes. Although in Table 3 there is also a performance drop after i-vector normalization, Figure 4, (b) might suggest even worse supervised classification when applying normalization, than the results actually obtained.

It is important to highlight that the results in Figures 3 and 4 are completely unsupervised, and the fact that, in figures (a), data naturally clusters according to dataset, or measurement time shows just how much information about the recording conditions/task is encoded in the feature sets. Furthermore, the absence of defined clusters in figures (b) does not mean that dataset (or time measurement) specific information is completely removed by the normalization. It just means that it becomes less evident in the context of unsupervised learning.

Figure 5 shows the euclidean distances between the mean i-vector of each of the six English datasets, and of each of the four measurement times in ILSE (right). We observe that the smaller distances occur for the pairs WSM-VoxCeleb and measurement time 0 and 1, which is consistent to what we observed in the cluster figures. Generally, we also observe that the distances between measurement times in ILSE are smaller than distances between distinct datasets.

4. Conclusion

This work presents a comparative analysis of healthy speech in two scenarios, using feature sets typically employed for disease detection from speech: (1) collected for six different datasets spoken in the same language, and (2) collected across different times in a single longitudinal corpus. We show that all feature sets analysed encode much information about the dataset/recording conditions over time. We support this claim through supervised learning experiments with results largely above chance level, and through unsupervised experiments where data naturally clusters according to the dataset/time measurement. ECAPA-TDNN embeddings seem to be the data representation that contains less information, but they still provide results highly above chance level. We also show that a simple strategy such as dataset dependent normalization may, for some feature sets, make the classification task harder, but it does not come close to solving the problem.

In the future, we plan to further complete these experiments by analysing the importance of the speech task alone, from the same dataset. Another relevant direction is to define a proxy to measure the similarity between datasets – could the mean i-vector of a dataset be this proxy? Finally, we plan to compare the values of the features individually, across datasets and explore whether it is possible to define reference intervals of healthy speech, for certain features, that can hold across datasets, and flag the presence of a diseases in speech.
5. References


