Pathological speech detection using x-vector embeddings

Abstract

The potential of speech as a non-invasive biomarker to assess a speaker's health has been repeatedly supported by the results of multiple works, for both physical and psychological conditions. Traditional systems for speech-based disease classification have focused on carefully designed knowledge-based features. However, these features may not represent the disease's full symptomatology, and may even overlook its more subtle manifestations. This has prompted researchers to move in the direction of general speaker representations that inherently model symptoms, such as Gaussian Supervectors, i-vectors and, x-vectors. In this work, we focus on the latter, to assess their applicability as a general feature extraction method to the detection of Parkinson's disease (PD) and obstructive sleep apnea (OSA). We test our approach against knowledge-based features and i-vectors, and report results for two European Portuguese corpora, for OSA and PD, as well as for an additional Spanish corpus for PD. Both x-vector and i-vector models were trained with an out-of-domain European Portuguese corpus. Our results show that x-vectors are able to perform better than knowledge-based features in same-language corpora. Moreover, while x-vectors performed similarly to i-vectors in matched conditions, they significantly outperform them when domain-mismatch occurs.

Introduction

Recent advances in Machine Learning (ML) and, in particular, in Deep Neural Networks (DNN) have allowed the development of highly accurate predictive systems for numerous applications. Among others, health has received significant attention due to the potential of ML-based diagnostic, monitoring and therapeutic systems, which are fast (when compared to traditional diagnostic processes), easily distributed and cheap to implement (many such systems can be executed in mobile devices).

Furthermore, these systems can incorporate biometric data to perform non-invasive diagnostics.

Among other data types, speech has been proposed as a valuable biomarker for the detection of a myriad of diseases, including: neurological conditions, such as Alzheimer's BIBREF0, Parkinson's disease (PD) BIBREF1 and Amyotrophic Lateral Sclerosis BIBREF2; mood disorders, such as depression, anxiety BIBREF3 and bipolar disorder BIBREF4; respiratory diseases, such as obstructive sleep apnea (OSA) BIBREF5. However, temporal and financial constraints, lack of awareness in the medical community, ethical issues and patient-privacy laws make the acquisition of medical data one of the greatest obstacles to the development of health-related speech-based classifiers, particularly for deep learning models. For this reason, most systems rely on knowledge-based (KB) features, carefully designed and selected to model disease symptoms, in combination with simple machine learning models (e.g. Linear classifiers, Support Vector Machines). KB features may not encompass subtler symptoms of the disease, nor be general enough to cover varying levels of severity of the disease. To overcome this limitation, some works have instead focused on speaker representation models, such as Gaussian Supervectors and i-vectors. For instance, Garcia et al. BIBREF1 proposed the use of i-vectors for PD classification and Laaridh et al. BIBREF6 applied the i-vector paradigm to the automatic prediction of several dysarthric speech evaluation metrics like intelligibility, severity, and articulation impairment. The intuition behind the use of these representations is the fact that these algorithms model speaker variability, which should include disease symptoms BIBREF1.

Proposed by Snyder et al., x-vectors are discriminative deep neural network-based speaker embeddings, that have outperformed i-vectors in tasks such as speaker and language recognition BIBREF7, BIBREF8, BIBREF9. Even though it may not be evident that discriminative data representations are suitable for disease detection when trained with general datasets (that do not necessarily include diseased patients), recent works have shown otherwise. X-vectors have been successfully applied to paralinguistic tasks such as emotion recognition BIBREF10, age and gender classification BIBREF11, the detection of

obstructive sleep apnea BIBREF12 and as a complement to the detection of Alzheimer's Disease BIBREF0. Following this line of research, in this work we study the hypothesis that speaker characteristics embedded in x-vectors extracted from a single network, trained for speaker identification using general data, contain sufficient information to allow the detection of multiple diseases. Moreover, we aim to assess if this information is kept even when language mismatch is present, as has already been shown to be true for speaker recognition BIBREF8. In particular, we use the x-vector model as a feature extractor, to train Support Vector Machines for the detection of two speech-affecting diseases: Parkinson's disease (PD) and obstructive sleep apnea (OSA).

PD is the second most common neurodegenerative disorder of mid-to-late life after Alzheimer’s disease BIBREF13, affecting 1% of people over the age of 65. Common symptoms include bradykinesia (slowness or difficulty to perform movements), muscular rigidity, rest tremor, as well as postural and gait impairment. 89% of PD patients develop also speech disorders, typically hypokinetic dysarthria, which translates into symptoms such as reduced loudness, monoloudness, monopitch, hypotonicity, breathy and hoarse voice quality, and imprecise articulation BIBREF14BIBREF15.

OSA is a sleep-concerned breathing disorder characterized by a complete stop or decrease of the airflow, despite continued or increased inspiratory efforts BIBREF16. This disorder has a prevalence that ranges from 9% to 38% through different populations BIBREF17, with higher incidence in male and elderly groups. OSA causes mood and personality changes, depression, cognitive impairment, excessive daytime sleepiness, thus reducing the patients' quality of life BIBREF18, BIBREF19. It is also associated with diabetes, hypertension and cardiovascular diseases BIBREF16, BIBREF20. Moreover, undiagnosed sleep apnea can have a serious economic impact, having had an estimated cost of $\$150$ billion in the U.S, in 2015 BIBREF21. Considering the prevalence and serious nature of the two diseases described above, speech-based technology that tests for their existence has the potential to become a key tool for early detection, monitoring and prevention of these conditions BIBREF22.

The remainder of this document is organized as follows. Section SECREF2 presents the background concepts on speaker embeddings, and in particular on x-vectors. Section SECREF3 introduces the experimental setup: the corpora, the tasks, the KB features and the speaker embeddings employed. The results are presented and discussed in section SECREF4. Finally, section SECREF5 summarizes the main conclusions and suggests possible directions for future work.

Background - Speaker Embeddings

Speaker embeddings are fixed-length representations of a variable length speech signal, which capture relevant information about the speaker. Traditional speaker representations include Gaussian Supervectors BIBREF23 obtained from MAP adapted GMM-UBM BIBREF24 and i-vectors BIBREF25.

Until recently, i-vectors have been considered the state-of-the-art method for speaker recognition. An extension of the GMM Supervector, the i-vector approach models the variability present in the Supervector, as a low-rank total variability space. Using factor analysis, it is possible to extract low-dimensional total variability factors, called i-vectors, that provide a powerful and compact representation of speech segments BIBREF23, BIBREF25, BIBREF26. In their work, Hauptman et. al. BIBREF1 have noted that using i-vectors, that model the total variability space and total speaker variability, produces a representation that also includes information about speech disorders. To classify healthy and non-healthy speakers, the authors created a reference i-vector for the healthy population and another for the PD patients. Each speaker was then classified according to the distance between their i-vector to the reference i-vector of each class.

As stated in Section SECREF1, x-vectors are deep neural network-based speaker embeddings that were originally proposed by BIBREF8 as an alternative to i-vectors for speaker and language recognition. In contrast with i-vectors, that represent the total speaker and channel variability, x-vectors aim to model

characteristics that discriminate between speakers. When compared to i-vectors, x-vectors require shorter temporal segments to achieve good results, and have been shown to be more robust to data variability and domain mismatches BIBREF8.

The x-vector system, described in detail in BIBREF7, has three main blocks. The first block is a set of five time-delay layers which operate at frame level, with a small temporal context. These layers work as a 1-dimensional convolution, with a kernel size corresponding to the temporal context. The second block, a statistical pooling layer, aggregates the information across the time dimension and outputs a summary for the entire speech segment. In this work, we implemented the attentive statistical pooling layer, proposed by Okabe et al. BIBREF27. The attention mechanism is used to weigh frames according to their importance when computing segment level statistics. The third and final block is a set of fully connected layers, from which x-vector embeddings can be extracted.

Experimental Setup

Four corpora were used in our experiments: three to determine the presence or absence of PD and OSA, which include a European Portuguese PD corpus (PPD), a European Portuguese OSA corpus (POSA) and a Spanish PD corpus (SPD); one task-agnostic European Portuguese corpus to train the i-vector and x-vector extractors. For each of the disease-related datasets, we compared three distinct data representations: knowledge-based features, i-vectors and x-vectors. All disease classifications were performed with an SVM classifier. Further details on the corpora, data representations and classification method follow bellow.

Experimental Setup ::: Corpora ::: Speaker Recognition - Portuguese (PT-EASR) corpus

This corpus is a subset of the EASR (Elderly Automatic Speech Recognition) corpus BIBREF28. It

includes recordings of European Portuguese read sentences. It was used to train the i-vector and the x-vector models, for speaker recognition tasks. This corpus includes speakers with ages ranging from 24 to 91, 91% of which in the age range of 60-80. This dataset was selected with the goal of generating speaker embeddings with strong discriminative power in this age range, as is characteristic of the diseases addressed in this work. The corpus was partitioned as 0.70:0.15:0.15 for training, development and test, respectively.

Experimental Setup ::: Corpora ::: PD detection - Portuguese PD (PPD) corpus

The PPD corpus corresponds to a subset of the FraLusoPark corpus BIBREF29, which contains speech recordings of French and European Portuguese healthy volunteers and PD patients, on and off medication. For our experiments, we selected the utterances corresponding to European Portuguese speakers reading prosodic sentences. Only on-medication recordings of the patients were used.

Experimental Setup ::: Corpora ::: PD detection - Spanish PD (SPD) corpus

This dataset corresponds to a subset of the New Spanish Parkinson's Disease Corpus, collected at the Universidad de Antioquia, Colombia BIBREF22. For this work, we selected the corpus' subset of read sentences. This corpus was included in our work to test whether x-vector representations trained in one language (European Portuguese) are able to generalize to other languages (Spanish).

Experimental Setup ::: Corpora ::: OSA detection - PSD corpus

This corpus is an extended version of the Portuguese Sleep Disorders (PSD) corpus (a detailed description of which can be found in BIBREF30). It includes three tasks spoken in European Portuguese: reading a phonetically rich text; read sentences recorded during a task for cognitive load assessment;

and a spontaneous description of an image.

All utterances were split into 4 second-long segments using overlapping windows, with a shift of 2 seconds. Further details about each of these datasets can be found in Table TABREF8.

Experimental Setup ::: Knowledge-based features ::: Parkinson's disease

Proposed by Pompili et al. BIBREF13, the KB feature set used for PD classification contains 36 features common to eGeMAPS BIBREF31 alongside with the mean and standard deviation (std.) of 12 Mel frequency cepstral coefficients (MFCCs) + log-energy, and their corresponding first and second derivatives, resulting in a 114-dimensional feature vector.

Experimental Setup ::: Knowledge-based features ::: Obstructive sleep apnea

For this task, we use the KB feature set proposed in BIBREF30, consisting of: mean of 12 MFCCs, plus their first and second order derivatives and 48 linear prediction cepstral coefficients; mean and std of the frequency and bandwidth of formant 1, 2, and 3; mean and std of Harmonics-to-noise ratio; mean and std of jitter; mean, std, and percentile 20, 50, and 100 of F0; and mean and std of all frames and of only voiced frames of Spectral Flux.

All KB features were extracted using openSMILE BIBREF32.

Experimental Setup ::: Speaker representation models ::: i-vectors

Following the configuration of BIBREF1, we provide as inputs to the i-vector system 20-dimensional feature vectors composed of 19 MFCCs + log-energy, extracted using a frame-length of 30ms, with 15ms

shift. Each frame was mean-normalized over a sliding window of up to 4 seconds. All non-speech frames were removed using energy-based Voice Activity Detection (VAD). Utterances were modelled with a 512 component full-covariance GMM. i-vectors were defined as 180-dimensional feature vectors. All steps were performed with Kaldi BIBREF33 over the PT-EASR corpus.

Experimental Setup ::: Speaker representation models ::: x-vectors

The architecture used for the x-vector network is detailed in Table TABREF15, where F corresponds to the number of input features and T corresponds to the total number of frames in the utterance, S to the number of speakers and Ctx stands for context. X-vectors are extracted from segment layer 6. The inputs to this network consist of 24-dimensional filter-bank energy vectors, extracted with Kaldi BIBREF33 using default values for window size and shift. Similar to what was done for the i-vector extraction, non-speech frames were filtered out using energy-based VAD. The extractor network was trained using the PT-EASR corpus for speaker identification, with: 100 epochs; the cross-entropy loss; a learning rate of 0.001; a learning rate decay of 0.05 with a 30 epoch period; a batch size of 512; and a dropout value of 0.001.

Experimental Setup ::: Model training and parameters

Nine classification tasks (three data representations for each of the three datasets) were performed with SVM classifiers. The hyper-parameters used to train each classifier, detailed in table TABREF17, were selected through grid-search.

Considering the limited size of the corpora, fewer than 3h each, we chose to use leave-one-speaker-out cross validation as an alternative to partitioning the corpora into train, development and test sets. This was done to add significance to our results.

We perform classification at the segment level and assign speakers a final classification by means of a weighted majority vote, where the predictions obtained for each segment uttered by the speaker were weighted by the corresponding number of speech frames.

Results

This section contains the results obtained for all three tasks: PD detection with the PPD corpus, OSA detection with the PSD corpus and PD detection with the SPD corpus. Results are reported in terms of average Precision, Recall and F1 Score. The values highlighted in Tables TABREF19, TABREF21 and TABREF23 represent the best results, both at the speaker and segment levels.

Results ::: Parkinson's disease - Portuguese corpus

Results for PD classification with the PPD corpus are presented in Table TABREF19. The table shows that speaker representations learnt from out-of-domain data outperform KB features. This supports our hypothesis that speaker discriminative representations not only contain information about speech pathologies, but are also able to model symptoms of the disease that KB features fail to include. It is also possible to notice that x-vectors and i-vectors achieve very similar results, albeit x-vectors present a small improvement at the segment level, whereas i-vectors achieve slightly better results at the speaker level. A possible interpretation is the fact that, while x-vectors provide stronger representations for short segments, some works have shown that i-vectors may perform better when considering longer segments BIBREF8. As such, performing a majority vote weighted by the duration of speech segments may be giving an advantage to the i-vector approach at the speaker level.

Results ::: Obstructive sleep apnea

Table TABREF21 contains the results for OSA detection with the PSD corpus. For this task, x-vectors outperform all other approaches at the segment level, most importantly they significantly outperform KB features by $\sim $8%, which further supports our hypothesis. Nevertheless, it is important to point out that both approaches perform similarly at the speaker level. Additionally, we can see that i-vectors perform worse than KB features. One possible justification, is the fact that the PSD corpus includes tasks - such as spontaneous speech - that do not match the read sentences included in the corpus used to train the i-vector and x-vector extractors. These tasks may thus be considered out-of-domain, which would explain why x-vectors are able to surpass the i-vector approach.

Results ::: Parkinson's disease: Spanish PD corpus

Table TABREF23 presents the results achieved for the classification of SPD corpus. This experiment was designed to assess the suitability of x-vectors trained in one language and being applied to disease classification in a different language. Our results show that KB features outperform both speaker representations. This is most likely caused by the language mismatch between the Spanish PD corpus and the European Portuguese training corpus. Nonetheless, it should be noted that, as in the previous task, x-vectors are able to surpass i-vectors in an out-of-domain corpus.

Conclusions

In this work we studied the suitability of task-agnostic speaker representations to replace knowledge-based features in multiple disease detection. Our main focus laid in x-vectors embeddings, trained with elderly speech data.

Our experiments with the European Portuguese datasets support the hypothesis that discriminative speaker embeddings contain information relevant for disease detection. In particular, we found evidence

that these embeddings contain information that KB features fail to represent, thus proving the validity of our approach. It was also observed that x-vectors are more suitable than i-vectors for tasks whose domain does not match that of the training data, such as verbal task mismatch and cross-lingual experiments. This indicates that x-vectors embeddings are a strong contender in the replacement of knowledge-based feature sets for PD and OSA detection.

As future work, we suggest training the x-vector network with augmented data and with multilingual datasets, as well as extending this approach to other diseases and verbal tasks. Furthermore, as x-vectors shown to behave better with out-of-domain data, we also suggest replicating the experiments with in-the-wild data collected from online multimedia repositories (vlogs), and comparing the results to those obtained with data recorded in controlled conditions BIBREF34.