Pathological speech detection using *x-vector* embeddings

M. Catarina Botelho*, Francisco S. Teixeira*, Thomas Rolland, Alberto Abad, Isabel Trancoso
INESC-ID/Instituto Superior Técnico, University of Lisbon, Portugal
{catarina.t.botelho, francisco.s.teixeira}@tecnico.ulisboa.pt,
{thomas.rolland, alberto.abad, isabel.trancoso}@inesc-id.pt

***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.**

***Index Terms*—Speech, Speaker embeddings, Parkinson's disease, Obstructive sleep apnea**

I. 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 Alzheimers [1], Parkinsons disease (PD) [2] and Amyotrophic Lateral Sclerosis [3]; mood disorders, such as depression, anxiety [4] and bipolar disorder [5]; respiratory diseases, such as obstructive sleep apnea (OSA) [6]. 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. [2] proposed the use of *i-vectors* for PD classification and Laaridh et al. [7] 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 [2].

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 [8]–[10]. 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 [11], age and gender classification [12], the detection of obstructive sleep apnea [13] and as a complement to the detection of Alzheimer's Disease [1]. 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 [9]. 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

*Both authors contributed equally to this work. This work was supported by national funds through FCT, Fundao para a Ciłncia e a Tecnologia, with grant numbers SFRH/BD/149126/2019, BD2018 ULisboa, project UIDB/50021/2020 and partially funded by the TAPAS project under Marie Sklodowska-Curie grant agreement No 766287.

of mid-to-late life after Alzheimers disease [14], 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 [15] [16].

OSA is a sleep-concerned breathing disorder characterized by a complete stop or decrease of the airflow, despite continued or increased inspiratory efforts [17]. This disorder has a prevalence that ranges from 9% to 38% through different populations [18], 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 [19], [20]. It is also associated with diabetes, hypertension and cardiovascular diseases [17], [21]. Moreover, undiagnosed sleep apnea can have a serious economic impact, having had an estimated cost of $\$150$ billion in the U.S, in 2015 [22]. 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 [23].

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

II. 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 [24] obtained from MAP adapted GMM-UBM [25] and *i-vectors* [26].

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 [24], [26], [27]. In their work, Hauptman et. al. [2] 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 I, *x-vectors* are deep neural network-based speaker embeddings that were originally proposed by [9] 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 [9].

The *x-vector* system, described in detail in [8], 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. [28]. 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.

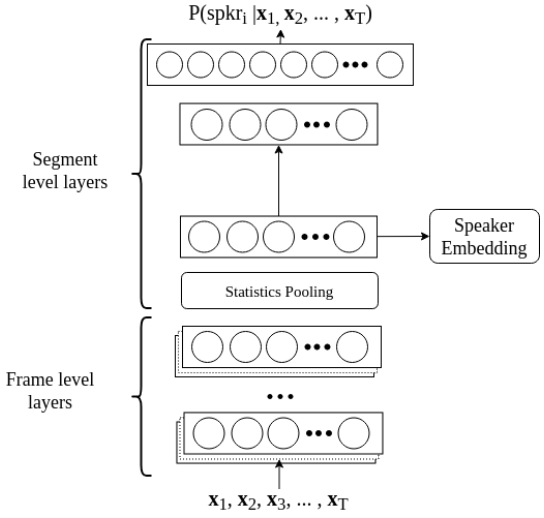

Fig. 1. *X-vector* network (adapted from [9]).

III. 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.

A. Corpora

1) Speaker Recognition - Portuguese (PT-EASR) corpus: This corpus is a subset of the EASR (Elderly Automatic Speech Recognition) corpus [29]. 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.

2) PD detection - Portuguese PD (PPD) corpus: The PPD corpus corresponds to a subset of the FraLusoPark corpus [30], 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.

3) 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 [23]. 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).

4) 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 [31]). 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 I.

TABLE I
CORPORA DESCRIPTION.

Lang.	Task	Group	Speakers	Segments	Duration (h)
PT	Spk. Rcg.	-	919	290,690	171.81
	PD	Patient	75	1,838	1.24
		Control	65	1,527	1.07
	OSA	Patient	30	1,793	1.10
		Control	30	1,702	1.05
SP	PD	Patient	50	661	0.49
		Control	50	655	0.50

B. Knowledge-based features

1) Parkinson's disease: Proposed by Pompili et al. [14], the KB feature set used for PD classification contains 36 features common to eGeMAPS [32] 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.

2) Obstructive sleep apnea: For this task, we use the KB feature set proposed in [31], 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 [33].

C. Speaker representation models

1) i-vectors: Following the configuration of [2], 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 [34] over the PT-EASR corpus.

2) x-vectors: The architecture used for the *x-vector* network is detailed in Table II, 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 [34] 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.

TABLE II
X-vector NETWORK ARCHITECTURE

Layer	Layer Ctx	Total Ctx	In x Out
frame1	[t-2, t+2]	5	5F×256
frame2	{t-2, t, t+2}	9	768×256
frame3	{t-3, t, t+3}	15	768×256
frame4	{t}	15	256×256
frame5	{t}	15	256×512
stats pooling	[0,T)	T	512T×1024
segment 6	{0}	T	1024×512
segment 7	{0}	T	512×512
softmax	{0}	T	512×S

D. 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 III, 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.

TABLE III
SVM MODEL PARAMETERS

Lang.	Task	Features	Kernel	C	Gamma
PT	PD	KB	Linear	1000	-
		i-vectors	RBF	10000	0.01
		x-vectors	Linear	0.01	-
	OSA	KB	RBF	10	0.01
		i-vectors	RBF	10	0.001
		x-vectors	RBF	10	0.00001
SP	PD	KB	Linear	0.001	-
		i-vectors	RBF	10	0.001
		x-vectors	RBF	1000	0.0001

IV. 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 IV, V and VI represent the best results, both at the speaker and segment levels.

A. Parkinson's disease - Portuguese corpus

Results for PD classification with the PPD corpus are presented in Table IV. 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 [9]. 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.

B. Obstructive sleep apnea

Table V 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 ∼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.

TABLE IV
RESULTS FOR THE PORTUGUESE PD CORPUS

Features		Precision	Recall	F1 Score
KB	Seg	64.5	64.6	64.5
	Spk	72.2	72.3	72.1
i-vectors	Seg	66.6	66.6	66.6
	Spk	**75.6**	**75.7**	**75.6**
x-vectors	Seg	**66.7**	**66.8**	**66.7**
	Spk	74.4	74.5	74.3

TABLE V
RESULTS FOR THE PORTUGUESE OSA CORPUS

Features		Precision	Recall	F1 Score
KB	Seg	64.8	64.9	64.8
	Spk	**82.0**	**81.7**	81.6
i-vectors	Seg	65.6	65.6	65.6
	Spk	72.3	75.0	75.0
x-vectors	Seg	**73.3**	**73.3**	**73.3**
	Spk	81.7	**81.7**	**81.7**

C. Parkinson's disease: Spanish PD corpus

Table VI 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.

TABLE VI
RESULTS FOR THE SPANISH PD CORPUS

Features		Precision	Recall	F1 Score
KB	Seg	**79.0**	**79.0**	**79.0**
	Spk	**87.1**	**87.0**	**87.0**
i-vectors	Seg	75.7	75.7	75.7
	Spk	85.1	85.0	85.0
x-vectors	Seg	77.2	77.2	77.1
	Spk	86.0	86.0	86.0

V. 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 [35].

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] Y. Hauptman et al., "Identifying distinctive acoustic and spectral features in parkinsons disease," *Proc. Interspeech 2019*, pp. 2498–2502, 2019.

[3] P. Gomez-Vilda, A. R. M. Londral, V. Rodellar-Biarge, J. M. Ferrandez-Vicente, and M. de Carvalho, "Monitoring amyotrophic lateral sclerosis by biomechanical modeling of speech production," *Neurocomputing*, vol. 151, pp. 130–138, 2015.

[4] J. Gratch et al., "The Distress Analysis Interview Corpus of human and computer interviews." in *LREC*. European Language Resources Association, 2014.

[5] F. Ringeval et al., "Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 3–13.

[6] M. C. Botelho, I. Trancoso, A. Abad, and T. Paiva, "Speech as a biomarker for obstructive sleep apnea detection," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5851–5855.

[7] I. Laaridh, W. Kheder, C. Fredouille, and C. Meunier, "Automatic prediction of speech evaluation metrics for dysarthric speech," in *Interspeech*, 2017.

[8] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5329–5333.

[9] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017, pp. 999–1003.

[10] D. Snyder et al., "Spoken language recognition using x-vectors." in *Odyssey*, 2018, pp. 105–111.

[11] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," *arXiv preprint arXiv:2002.05039*, 2020.

[12] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," 2019.

[13] J. M. Perero-Codosero, F. Espinoza-Cuadros, J. Anton-Martin, M. A. Barbero-Alvarez, and L. A. Hernandez, "Modeling obstructive sleep apnea voices using deep neural network embeddings and domain-adversarial training," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[14] A. Pompili et al., "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 (TSD)*, Sep. 2017, pp. 411–419.

[15] J. Vsquez-Correa, J. R. Orozco-Arroyave, and E. Nth, "Convolutional neural network to model articulation impairments in patients with parkinsons disease," in *Proc. Interspeech 2017*, 2017, pp. 314–318.

[16] L. V. Kalia and A. E. Lang, "Parkinson's disease." *Current neurology and neuroscience reports.*, Aug 2015.

[17] J. Arnold et al., "Obstructive sleep apnea," *Journal of pharmacy & bioallied sciences*, vol. 9, no. Suppl 1, p. S26, 2017.

[18] C. Senaratna et al., "Prevalence of obstructive sleep apnea in the general population: a systematic review," *Sleep Medicine Reviews*, vol. 34, pp. 70–81, 2017.

[19] N. Punjabi, "The epidemiology of adult obstructive sleep apnea," *Proceedings of the American Thoracic Society*, vol. 5, no. 2, pp. 136–143, 2008.

[20] T. Paiva, M. Andersen, and S. Tufik, "Sono e a medicina do sono. 1ª edição," 2014.

[21] R. F. Pozo et al., "Assessment of severe apnoea through voice analysis, automatic speech, and speaker recognition techniques," *EURASIP J. Adv. Signal Process*, vol. 2009, Jan. 2009.

[22] F. Sullivan, "Hidden health crisis costing america billions: Underdiagnosing and undertreating obstructive sleep apnea draining healthcare system," *American Academy of Sleep Medicine*, 2016.

[23] J. R. Orozco-Arroyave et al., "New Spanish Speech Corpus Database for the Analysis of People Suffering from Parkinson's Disease." in *LREC*, 2014, pp. 342–347.

[24] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[25] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[26] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[27] A. Abad, E. Ribeiro, F. Kepler, R. F. Astudillo, and I. Trancoso, "Exploiting phone log-likelihood ratio features for the detection of the native language of non-native english speakers," in *INTERSPEECH*, 2016.

[28] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[29] A. Hämäläinen et al., "The easr corpora of european portuguese, french, hungarian and polish elderly speech." in *LREC*, 2014, pp. 1458–1464.

[30] S. Pinto et al., "Dysarthria in individuals with parkinson's disease: a protocol for a binational, cross-sectional, case-controlled study in french and european portuguese (fralusopark)," *BMJ open*, vol. 6, no. 11, p. e012885, 2016.

[31] 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.

[32] F. Eyben 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, 4 2016, open access.

[33] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich Open-source Multimedia Feature Extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13, 2013, pp. 835–838.

[34] D. Povey et al., "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[35] J. Correia, B. Raj, I. Trancoso, and F. Teixeira, "Mining multimodal repositories for speech affecting diseases," *Interspeech*, 2018.