Adaptation of Speaker and Speech Recognition Methods for the Automatic Screening of Speech Disorders using Machine Learning

Ph.D. Thesis

José Vicente Egas López
Supervisor: Gábor Gosztolya, Ph.D.

Doctoral School of Computer Science
Department of Computer Algorithms and Artificial Intelligence
Faculty of Science and Informatics
University of Szeged

Szeged, October 16, 2022

1 Introduction

This PhD thesis presents methods for exploiting the non-verbal communication of individuals suffering from specific diseases or health conditions aiming to reach an automatic screening of them. More specifically, we employ one of the pillars of non-verbal communication, paralanguage, to explore techniques which could be utilized to model the speech of subjects. Paralanguage is a non-lexical component of communication that relies on intonation, pitch, speed of talking, and others, which can be processed and analyzed in automatic manners. This is called *Computational Paralinguistics*, which can be defined as the study of modelling non-verbal latent patterns within the speech of a speaker by means of computational algorithms; these patterns go beyond the *linguistic* approach. By means of machine learning, we present models from distinct scenarios of both *paralinguistics* and *pathological speech* which are capable of estimating the health status of a given disease such as Alzheimer's, Parkinson's, Depression, among others, in a automatic manner.

The dissertation consists of four major parts, in the sections below we will summarize the concepts, experiments, and results of Chapters 3-6. A brief review of the first chapters, is described next. Chapter 1 introduces the reader to the concepts of non-verbal communication and paralanguage. Also, we briefly cover concepts on Speech and Speaker Recognition. The same chapter continues with a more in depth explanation of paralanguage and computational paralinguistics, covering contemporary early works in the mentioned field. Chapter 2 describes the concepts of the machine learning algorithms used for producing ways of automatic screening of a given speech-pathology, as well as definitions of pathological speech, and the type of features we employed for processing the speech samples.

2 Front-End Factor Analysis

Some of the symptoms caused by Alzheimer's Disease are linked to the speech difficulties of the subject. More in particular, the inability to recall vocabulary, which makes the patient's speech different. Mild Cognitive Impairment (MCI), which is considered as a prodromal neuro-degenerative state of AD, also carries these type of symptoms but in moderate levels. The key to mitigate the progress of both disorders is achieving an early diagnosis. However, actual ways of diagnosis are costly and quite time-consuming.

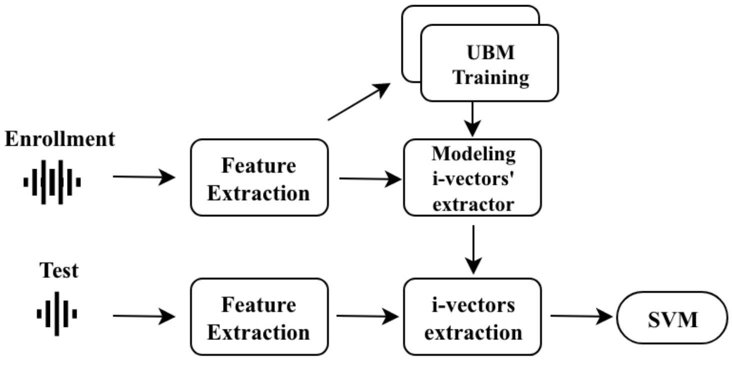

Figure 1: *The generic methodology applied for Alzheimer's screening by means of the speech.*

There is a scarcity of screening techniques for Alzheimer's which makes it complex to diagnose. The importance of an early diagnosis may be the key to a find more efficient manners that can slow down the development of the disease. Seeking for a non-invasive tool to help with the screening of AD, we employ a method intended to extract meaningful speaker traits. We rely on the *i-vector* approach which was an early state-of-the-art method for speaker recognition [see more in 13, 16].

In Chapter 3, we proposed the i-vector approach for the extraction of features from speech of subjects. These features were able to model the speech pattern of the three mental conditions from the speakers. These i-vector representations were extracted from Mel-Frequency Cepstral Coefficients, and were given to a SVM classifier in order to classify the speech in one of the following manners: AD - Alzheimer Disease, MCI - Mild Cognitive Impairment, HC - Healthy Control. We tested these i-vector features by performing a 5-fold cross-validation and measured performances relying on F1-score.

The experiments were executed in the following manner: (1) MFCCs features were extracted separately from 225 (i.e. dementia dataset) and 44 speech recordings (i.e., BEA dataset) (2) the UBM was trained using the MFCCs obtained from the BEA dataset (3) the i-vector extractor model was trained using the UBM of the previous step, and MFCCs from the dementia dataset (4) MFCCs from the dementia dataset were processed to extract a set of 225 i-vectors, and lastly, (5) a Support Vector Machines (SVM) performed the classification process. These stages are outlined in Figure 1.

It could be demonstrated that speech analysis offers a non-intrusive, non-expensive and faster way to perform the diagnosis of Alzheimer's by means of the utterances (i.e. speech recordings) of subjects. Here, we presented the advantage of i-vectors as features to model the particular speech of an Alzheimer's sufferer. Two groups of speech signals were represented via MFFCs features, one for the BEA Hungarian Spoken Language Database and the other got from the *dementia* dataset. Next, i-vector modeling was performed over these features with the goal of extracting their total factors (i.e., i-vector features). SVM utilized these i-vectors and classified them using a linear kernel. It achieved an F1 score of 79.2% for the three groups, namely, Alzheimer Disease (AD), Mild Cognitive Impairment (MCI), and Healthy Control (HC).

3 The Fisher Vector

In Chapter 4 we cover how the Fisher Vector, a method intended for image classification, can be employed for distinct computational paralinguistic and pathological speech tasks. Turns out that Fisher vector representations are able to capture relevant speaker features that we call FV encodings, which can be applied to screen different speaker states using speech recordings or audio signals. We represented the utterances of subjects having different kinds of pathological speech conditions modeled by the mentioned approach.

More in detail, we experiment with (1) the automatic assessment of Parkinson's Disease, (2) Cold speech, (3) textual escalation (the level of escalation in the dialogue), and (3) the classification of species of primates. Each sub-section will cover and describe every mentioned task separately. For the Cold, Escalation, and Primates tasks to be described below, the performance of the classifier was measured via Unweighted Average Recall (UAR), which is a proper metric for these kinds of paralinguistic tasks; also because it is commonly

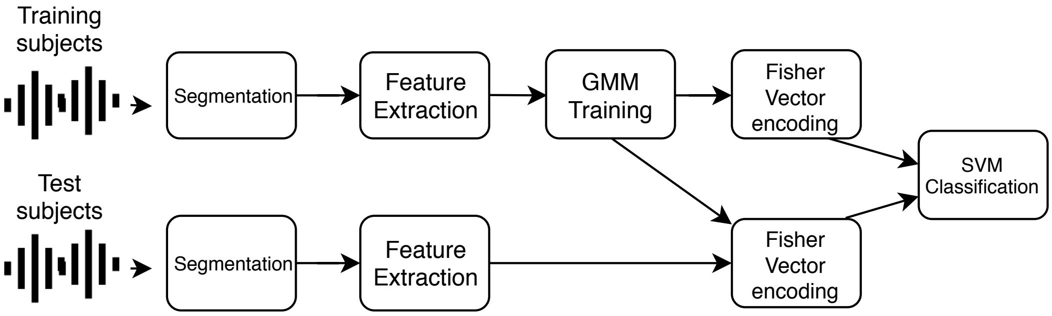

Figure 2: *Generic methodology applied in our experiments.*

used when there is the need to handle class imbalance situations.

3.1 Parkinson's Disease Screening

Some of the classic symptoms of Parkinson's include shaking, rigidity, slowness of movement, and speech difficulties. Commonly, the speech of the patient is also affected in terms of its tone, volume, and rate. The automatic speech analysis has been utilized in many medical branches to offer accurate and non-expensive solutions that are able to assess the diagnosis of different neuro-degenerative diseases such as Parkinson's by means of speech recordings.

The pipeline carried out for the experiments about Parkinson's is comprised by the following steps: (1) VAD-based segmentation, (2) feature extraction, (3) fitting a GMM to the local image features, (4) construction of the (audio) word dictionary by means of the GMM, that is, the encoded FV that now represents the global descriptor of the original spectrum, and (5) SVM classification. (See Fig. 2) The experiments were executed relying on four different feature sets. The first consisted of 20 MFCCs, obtained from 30 ms wide windows; and the rest of the feature sets were built by articulation, phonation, and prosody, respectively.

MFCC features performed the best in our experiments, which were divided according to tasks as per the audio recordings from the subjects: ’DDK’, ’Monologue’, ’Read Text’, and ’Sentences’. The task ’Sentences’ became the leader in terms of the AUC, with a score of 0.891. In the subsequent experiments, we showed that the predictions obtained for the different frame-level feature sets and tasks could be combined, allowing an even higher classification performance. This way, our AUC scores improved even further, and we got 0.908 with the combination of MFCCs with articulatory features for the ’Monologue’ task. Overall, incorporating the predictions for the tasks ‘Read text’, ‘Monologue’ and ‘DDK’, also led to a significant improvement.

Our study showed how useful are FV over i-vectors as features in the assessment of PD via the analysis of speech. We used the PC-GITA dataset to do experiments and classify PD and HC subjects. We used four frame-level feature sets as the input of the FV method, and applied (linear) Support Vector Machines for classifying the speech of subjects. Our findings showed that our approach offers superior performance compared to classification based on the i-vector and cosine distance approach, and it also provided an efficient combination of machine learning models trained on different feature sets or on different speaker tasks.

3.2 Screening of Cold

The so-called upper respiratory tract infection (URTI) is an infectious process for any of the components of the upper airway. For instance, it includes the common cold, a sinus infection, amongst others. The automatic assessment whether a subject has a cold may be relevant when trying to prevent the spread of it by predicting its patterns of propagation. We focus on finding specific voice patterns latent in the speech of subjects having a *cold* utilizing the Upper Respiratory Tract Infection Corpus (URTIC) which was the dataset of one of the Sub-Challenges in the ComParE Challenge from Interspeech 2017 [23].

A similar workflow as in Fig. 2 is employed for the Cold task as well. With the difference that segmentation is not taken into consideration, and that we utilized two classifiers: SVM and XGBoost. We employed MFCCs with a dimension of 13 coefficients along with their first and second order derivatives. We experimented with 23-dimensional MFCCs, 40-dimensional FBANKs and spectrograms for the **Escalation** and **Primates** tasks, respectively. For all of the them, we set the frame-length to 25 ms, and the frame-shift to 10 ms.

Compared with works done by other teams using the same dataset [15, 24], our performance is competitive, and our feature extraction pipeline seems to be simpler than those studies given that we utilized one single type of feature representation for training a model. We found that SVM gave better results when the feature pre-processing step was applied before executing the training phase. Thus, we demonstrated how applying Power Normalization along with dimension reduction via Principal Component Analysis on the Fisher Vector features improved the classification performance.

Combining Power Normalization with PCA gave better UAR scores on test set. These results are higher compared to those got using the Bag-of-Audio-Words approach described in [24]. Overall, SVM achieved a final score of 67.81% of Unweighted Average Recall on the test set. On the other hand, XGBoost achieved better results on the test set by just using raw (non-standardized) Fisher vectors, achieving a UAR score of 70.43%.

3.3 Escalation in the Dialogue, and Primates Species Detection

Public security, human-computer interactions, or human-to-human conversations are a few scenarios that can be benefited from the automatic detection of the levels of escalation in the dialogue. The acoustic-based escalation assessment include real life applications such as e-commerce customer service systems to alert and prevent potential conflicts before they take place. To this end, we make use of the **Escalation** Corpus described in [25]. Aiming to have better tools for monitoring biodiversity, researchers have experimented with bio-acoustics, seeking to annotate or label the different sounds from the nature. In our specific case, we are interested in the discrimination of vocalization from **Primates** species. The task was also introduced by Schuller et al. in [25].

We employ a similar workflow as in Fig. 2 for this task. With the difference that segmentation is not taken into consideration. We experimented with 23-dimensional MFCCs, 40-dimensional FBANKs and spectrograms for the **Escalation** and **Primates** tasks, respectively. For all of the them, we employed a frame-length of 25 ms, and frame-shift of 10 ms.

The Fisher vector approach was successful at the moment of modelling this particular corpora. Our experiments managed to overpass all the official baselines presented in [25] for both tasks. Moreover, the results achieved in the **Escalation** task positioned our paper as the winner of the ComParE Primates Sub-Challenge from the Interspeech Conference of 2021[1]. We achieved an UAR value of 73.8%. On the test set we reached 62.4% with this approach, which was improved to 63.9% by a 'late-fusion'.

4 Deep Neural Network Embeddings

The aim of this section is to introduce deep neural network embeddings to pathological speech and paralinguistics tasks. More in specific, here we cover the use of the x-vector approach (a method originally intended for speaker recognition) as a feature extraction method for audio-signals. The x-vector method is a state-of-the-art technique for speaker recognition. A handful of previous studies exploited x-vector embeddings in computational paralinguistic tasks; for instance, to classify emotions from the speech of subjects [22], and to screen neuro-degenerative diseases like Alzheimer's [28].

In Chapter 5 of the PhD thesis, based on the methodology outlined in [27], we adopt the DNN architecture described there. For all the tasks, we train custom networks from scratch employing different corpora. The DNN extracts the final neural network embeddings, which are utilized by a Support Vector Regression (SVR) or Support Vector Machines (SVM) for the estimation. We experiment with the automatic screening of four main tasks: the degree of **sleepiness**, **depression**, the classification of **primates** sounds, and the levels of **escalation** in the speech. The classification or regression procedures were done using a Support Vector Machines algorithm with a linear kernel and, as suggested the C complexity parameter was set in the range 10^{-5}, …, 10^{1}, for all the tasks in question. Unless specified the contrary, the employed evaluation metric was Unweighted Average Recall (UAR). We show that x-vector features are able to produce high quality speaker models for tasks not related to speaker recognition.

4.1 Excessive Daytime Sleepiness Detection

The excessive lack of sleep may lead to poor performance in daily activities, can contribute to accidents, and eventually lead to mortality. The most common causes of excessive daytime sleepiness (hypersomnia) are sleep deprivation and disorders like apnea (cessation of breathing) and insomnia (the inability to stay or fall asleep) [18]. The detection and monitoring of sleepiness crucial for reducing the risks of having fatal accidents; We propose a non-invasive way to monitor and control the degree of sleepiness by using the speech of the subjects. The task using Dusseldorf Sleepy Language Corpus for sleepiness screening was first introduced by Schuller et al. [25], and has been already addressed earlier by various studies. For instance, Gosztolya [14] achieved the best score at the time (Pearson's Correlation Coefficient of .387) using a combination of Fisher Vectors, BoAW, the ComParE functionals, and training ensembles of classifiers.

For the trials, we extracted 20 MFCCs using a frame-length of 25ms and a window step size of 10ms. We trained different x-vector Deep Neural Network models (i.e., extractors) using two distinct datasets. First, we used the data of the training and development sets of the SLEEP corpus combined (10,892 utterances, 11 hours and 39 mins). Second, to

[1] http://www.compare.openaudio.eu/winners/

experiment with the independence of the x-vectors from different recording and speaking conditions (e.g., language), we trained the extractor (DNN) on another corpus (also for speaker recognition). We used a subset of 60 hours (10,636 utterances) of the BEA Corpus, which contains Hungarian spontaneous speech (for more details, see [21]). This corpus has a relevant size (in comparison with the SLEEP Corpus), which is convenient when training DNNs. Moreover, we also experimented with data augmentation on these datasets, we relied on additive noises and reverberation.

Our findings indicate that the augmentation strategies applied on both corpora did not give any improvements: the quality of the embeddings extracted using the augmented models only reduced the final scores. Furthermore, it appears that making use of in-domain data causes the extractors (DNN models) to generate more meaningful features than just using out-of-domain data. In specific, we achieved the best performance employing the x-vector features computed via the SLEEP Corpus model. Moreover, in contrast to former studies, we did not rely on fusion strategies yet the results are competitive. More generally, we demonstrated that our methodology, besides surpassing the performances of various previous works, also produce the highest Spearman's CC score via a standalone (single) method for this particular task.

4.2 Clinical Depression Screening

As per James et. al [17], depression is a common mental disorder that affects globally to more than 264 million people of all ages. It is described as a psychiatric disorder affecting the patient on a wide scale. Although it is a frequent and curable disease, estimating its occurrence is hard due to the specific clinical expertise needed [12]. The speech is a biomarker containing information about a wide variety of traits (e.g., the mental status of the speaker). The fact that there may be a connection between depression and speech was pointed out by Kraepelin [19], one of the founders of modern psychology. To this end, we propose DNN embeddings for a non-invasive and automatic screening of depression from the speech of individuals.

We fitted two different x-vector extractors using distinct corpora: *first*, we employed a subset of 60 hours (10,636 utterances) of the *BEA Corpus* [21] (Hungarian spontaneous speech). And *second*, we utilized the pre-trained x-vector model [27] that was fitted on English speech corpora (Switchboard (SWBD) plus NIST SRE). Besides investigating the usefulness of the pre-trained model on a different type of task, we also sought to discover the difference in quality of x-vector representations extracted using distinct models, which differ in both amount and language in terms of their training data. Furthermore, seeking to improve the diversity of the data and the noise robustness of the model, we carried out data augmentation on the BEA corpus. The augmented dataset was used to fit two additional extractors.

We demonstrated that x-vector embeddings contain information that is predictive of the levels of clinical depression via the speech of subjects. Our custom x-vector extractors learned from distinct frame-level features acquired from corpora matching the language of the actual task. Also, we found an improvement of the quality of the embeddings when computing them using *augmented* x-vector models. In this context, we spotted a slight language-domain dependence of the x-vector method as our best tailored extractor surpassed the performance of the pre-trained model even after the feature selection process.

Furthermore, our findings confirmed that log-energies appear to be a robust alternative of cepstra coefficients for x-vector training as they provide larger (and more informative) input representations. We showed how our correlation-based feature selection approach produced similar performance scores using only a quarter of the features. Finally, we presented highly competitive CC and RMSE scores compared to those from former studies that used the same corpus and based their evaluations using optimistic methods (i.e, LOOCV), which proves the effectiveness of our approaches.

4.3 Escalation in the Speech, and Primates Species detection

Speech is the primary communication channel of humans. Evidently, human speech not only encodes the actual words spoken, but it incorporates a wide range of non-verbal content as well, transmitting a variety of information about the physical and mental state of the speaker. As already stated in Section 3.3, **Escalation** detection can provide real-life applications such as in public security, conversations in public places, and even human-computer interactions. Similarly, the automatic detection of **Primates** species by means of audio-signals can help to maintain the control and monitoring of the biodiversity.

For the experimental setup, we utilize a similar x-vector pipeline as in our previous configurations. Besides MFCCs, we also experimented with filter-banks of size 40 for the Depression, Escalation, and Primates corpus. While MFCCs are the standard for fitting x-vector models, FBANKs have proved to be effective in deep learning studies related to speech analysis, e.g., in speech recognition tasks [20, 26]. Moreover, for Escalation and Primates we also computed spectrograms. These have been proved to be useful in research related to computational paralinguistics such as in emotion recognition [11].

Building an ensemble x-vector classifier by training 10 independent x-vector extractor neural networks on the same data improved both the robustness and the performance of the x-vectors embeddings. Our UAR scores on the development set demonstrated the superiority of the ensemble classifiers over the independent x-vector-based ones. The ensemble x-vectors seem to be an effective approach for modelling Escalation's dialogue and estimating Primate sounds from different chimpanzees sounds. Our last technique, which used the SSPNet Conflict Corpus in the Escalation sub-challenge, also led to promising UAR values. Overall we over-passed the official baselines from [25] for both tasks, which supports the efficacy of the applied techniques.

5 Automatic Speech Recognition Methods

As stated in Section 2, dementia and MCI are progressive clinical syndromes which could provide more effective therapeutic interventions to delay progression if identified with anticipation. Since language changes in MCI are present even before the manifestation of other distinctive cognitive symptoms, a non-invasive way of early automatic screening could be the use of speech analysis. In Chapter 6 we present a set of temporal speech parameters that mainly focus on the amount of silence and hesitation, and demonstrated its applicability for MCI and dementia detection. For the automatic extraction of these attributes, we rely on an Automatic Speech Recognition (ASR) system. However, the main focus of the mentioned chapter of the thesis is to omit the necessity of execution of a full ASR process for temporal parameters extraction while keeping the amount of silence and hesitation in the speech of the subject quantified. We experimentally demonstrate that this approach, operating directly on the frame-level output of a HMM/DNN hybrid acoustic model, is capable of extracting attributes as useful as the ASR-based temporal parameter extraction workflow was able to.

5.1 Temporal Speech Parameters

To investigate the speech of MCI patients and HC subjects, we calculated specific temporal parameters from their spontaneous speech by means of ASR acoustic models in both English and Hungarian languages. We investigate the language-dependence of our speech processing workflow, developed for distinguishing between mild cognitive impairment (MCI) and healthy controls (HC) subjects. The set of temporal speech parameters consists of the articulation rate, speech tempo, utterance duration, and attributes describing various characteristics of hesitation present in the speech of the patient.

Due to the bilingual nature of our study, we employed two datasets to train the DNN acoustic models of our two phone-level ASR systems (i.e., English and Hungarian). The datasets contain spontaneous speech. The ASR system was trained to recognize the phones in the utterances, where the phone set included the special non-verbal labels listed above (i.e. filled pauses, coughs, breath intakes, etc.).

For the English subjects, we achieved high classification scores (in the range 80.0-85.7%, an Area-Under-Curve score of 0.932 and $\min C_{llr} = 0.305$), while for Hungarian, the classification performance was acceptable (with classification metrics falling into the range 66.7-76.9% and with an AUC value of 0.727). By only using specific subsets of our temporal parameters, we noticed that filled pauses were more useful for both speaker groups (i.e. English and Hungarian) than silent pauses; surprisingly, silent pauses were not useful at all for distinguishing the Hungarian subjects. Overall, it appears that the differences we found in the temporal parameters appear to be the effect of a difference in the *training databases*, and they have little to do with the difference in the languages.

5.2 Posterior-Thresholding Hesitation Representation

This technique proposed a feature set which, similarly to the previous method, describes the amount of hesitation in the spontaneous speech of the subject. However, instead of using an Automatic Speech Recognition system and analyzing its output, we focused directly on the frame-level output of the Deep Neural Network acoustic model.

The feature extraction approach is divided into three steps. These are:

(1) A Deep Neural Network acoustic model is evaluated on the utterances, using frame-level features (e.g. MFCCs).

(2) Based on the outputs provided by the DNN, we estimate the local posterior probability of silence and filler events. This step is still performed at the frame level.

(3) From the local posterior estimates calculated in step (2), new representations are computed at the utterance level.

Using the utterance-level feature vectors calculated in step (3), we can readily carry out the utterance-level (or, in our case, subject-level) classification, e.g. by using a Support Vector Machine (SVM) classifier.

According to our experimental results, we got an accuracy of 69.3%, while we achieved an F_1 value of 87.5 and a mean AUC score of 0.780. Although it is impossible to do a direct comparison with other values in the literature due to using different corpora, experimental setup and evaluation metrics, our results demonstrate that this approach, operating directly on the frame-level output of a HMM/DNN hybrid acoustic model, is capable of extracting attributes as useful the ASR-based temporal parameter extraction methods.

6 Contributions of the thesis

In the **first thesis group**, my contributions are related to the feature extraction by means of the i-vector approach as well as the classification procedures. Detailed discussion can be found in Chapter 3.

I / 1. My contribution relied on training i-vector models for the extraction speech representations of individuals suffering from Alzheimer's. I demonstrated that i-vector features are capable of extracting meaningful traits from this kind of speech.

I / 2. As a part of my proposals for the study in question, I employed i-vectors as a baseline approach for the automatic screening of the levels of clinical depression by means of the speech. Turns out that this method achieves comparable and even competitive performances compared with prior studies on the same corpus.

In the **second thesis group**, my contributions are related to the automatic assessment of Parkinson's Disease, the levels of escalation in speech, primate species sounds, and cold identification using speech features modelled by the Fisher vector approach. Detailed discussion can be found in Chapter 4.

II / 1. I developed a framework for the automatic assessment of Parkinson's Disease by means of the Fisher vector approach. My findings showed that these kind of features are capable of capturing meaningful information not only from images (as they were originally intended for) but from utterances as well.

II / 2. Built a machine learning model capable of discriminating cold from the speech of individuals using Fisher vectors. I demonstrated the superiority of XGBoost over SVM at the moment of employing the mentioned features for cold speech classification.

II / 3. As part of the procedures conducted in this scientific article, I modelled the levels of escalation in the speech of individuals using Fisher vectors; moreover, the same technique was employed to extract features from the sounds of primate species. I proved that such an approach is quite beneficial at the moment of automatic assessment of the tasks in question.

II / 4. I designed a pipeline for 'cold' speech feature extraction based on Fisher vector encodings. I proved that such type of features are capable of accurately modelling the speech of patients having a cold.

In the **third thesis group**, my contributions are related to the use of speech for the screening of the levels of sleepiness, the degree of clinical depression, the levels of escalation in speech, and primate species sounds. Detailed discussion can be found in Chapter 5.

III / 1. I proposed the use of deep neural network embeddings for the estimation the degree of sleepiness in an automatic manner by means of the speech. I showed that x-vectors, being originally intended for speaker verification, are capable of modelling speakers that suffer from day-time sleepiness with high accuracy.

III / 2. My proposal relied on the use of custom x-vector extractors for the assessment of the degree of clinical depression from the speech of patients. By training a handful of DNN models, I showed that a simple pipeline is capable of surpassing the performances of those that rely on more elaborated techniques like ensemble machine learning or classifier combination.

III / 3. Part of my contribution to this study comprised the training of various custom x-vector extractors. I proved that these deep neural network embeddings demonstrated competitive performances for both conflict escalation in the speech and primates species classification.

In the **fourth thesis group**, the contributions are related to the employment of temporal speech parameter as speaker features for the automatic screening of both Mild Cognitive Impairment and Alzheimer's Disease. Detailed discussion can be found in Chapter 6.

IV / 1. My main contribution to this study was the generation of temporal speech parameters via an ASR system on a frame-level approach. I showed that it is not needed to use the full ASR in order to obtain high-quality features comparable to those based on full ASR systems for both MCI and Alzheimer's screening.

IV / 2. My participation in this study was limited as I was not the main contributor. More in specific, I participated in the temporal speech parameters computation. This study demonstrated that the language on which the ASR system was trained only slightly affects the MCI classification performance; reducing the necessity for relying on a specific language-domain corpora.

Table 1 summarizes the relation between the thesis points and the corresponding publications.

Table 1: *Correspondence between the thesis points and my publications.*

Publication	Thesis point										
	I/1	I/2	II/1	II/2	II/3	II/4	III/1	III/2	III/3	IV/1	IV/2
[1]										●	
[2]											●
[3]	●										
[4]			●								
[5]				●							
[6]							●				
[7]						●					
[8]					●				●		
[9]		●						●			

The author's publications on the subjects of the thesis

Journal publications

[1] **Egas-López, J. V.**, Balogh, R., Imre, N., Hoffmann, I., Szabó, M. K., Tóth, L., ... & Gosztolya, G. Automatic screening of Mild Cognitive Impairment and Alzheimer's disease by means of posterior-thresholding hesitation representation. *Computer Speech & Language*, VOL(75), 2022.

[2] **Gosztolya, G.**, Balogh, R., Imre, N., Egas-López, J. V., Hoffmann, I., Vincze, V., ... & Kálmán, J. Cross-lingual detection of Mild Cognitive Impairment based on temporal parameters of spontaneous speech. *Computer Speech & Language*, VOL(69), 2021.

Full papers in conference proceedings

[3] **Egas López, J. V.**, Tóth, L., Hoffmann, I., Kálmán, J., Pákáski, M., and Gosztolya, G. Assessing Alzheimer's disease from speech using the i-vector approach. In *International Conference on Speech and Computer (SPECOM)*, Springer, Cham., 289-298, 2019.

[4] **Egas López, J. V.**, Orozco-Arroyave, J. R. and Gosztolya, G. Assessing Parkinson's disease from speech using Fisher Vectors. In *Proceedings of Interspeech*, ISCA, 3063-3067, 2019.

[5] **Egas López, J. V.** and Gosztolya, G. Predicting a Cold from Speech Using Fisher Vectors; SVM and XGBoost as Classifiers. In *International Conference on Speech and Computer (SPECOM)*, Springer, Cham., 145-155, 2020.

[6] **Egas-López, J. V.**, and Gosztolya, G. Deep neural network embeddings for the estimation of the degree of sleepiness. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 7288-7292, 2021.

[7] **Egas-López, J. V.**, & Gosztolya, G. (2021). Using the Fisher Vector Approach for Cold Identification. In *Acta Cybernetica*, 25(2), 223-232.

[8] **Egas-López, J. V.**, Vetráb, M., Tóth, L., & Gosztolya, G. Identifying Conflict Escalation and Primates by Using Ensemble X-vectors and Fisher Vector Features. In *Proceedings of Interspeech*, ISCA, 476-480, 2021.

[9] **Egas-López, J. V.**, Kiss, G., Sztahó, D., & Gosztolya, G. Automatic Assessment of the Degree of Clinical Depression from Speech Using X-Vectors. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 8502-8506, 2022.

[10] **Vetráb, M.**, Egas-López, J.V., Balogh, R., Imre, N., Hoffmann, I., Tóth, L., Pákáski, M., Kálmán, J., Gosztolya, G. Using Spectral Sequence-to-Sequence Autoencoders to Assess Mild Cognitive Impairment. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 6467-6471, Singapore, 2022.

Other References

[11] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 International Conference on Platform Technology and Service (PlatCon)*, pages 1–5, 2017.

[12] Mary Jane Friedrich. Depression is the leading cause of disability around the world. *Jama*, 317(15):1517–1517, 2017.

[13] Daniel Garcia-Romero and Carol Y Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Twelfth annual conference of the international speech communication association*, 2011.

[14] Gábor Gosztolya. Using Fisher Vector and Bag-of-Audio-Words representations to identify Styrian dialects, sleepiness, baby & orca sounds. In *Proceedings of Interspeech*, pages 2413–2417, Graz, Austria, Sep 2019.

[15] Mark Huckvale and András Beke. It sounds like you have a cold! Testing voice features for the Interspeech 2017 Computational Paralinguistics Cold Challenge. In *Proceedings of Interspeech*, pages 3447–3451. International Speech Communication Association (ISCA), 2017.

[16] Noor Salwani Ibrahim and Dzati Athiar Ramli. I-vector Extraction for Speaker Recognition based on Dimensionality Reduction. *Procedia Computer Science*, 126:1534–1540, 2018.

[17] Spencer L James, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1789–1858, 2018.

[18] M.W. Johns. Daytime Sleepiness, Snoring, and Obstructive Sleep Apnea: the Epworth Sleepiness Scale. *Chest*, 103(1):30–36, 1993.

[19] Emil Kraepelin. Manic depressive insanity and paranoia. *The Journal of Nervous and Mental Disease*, 53(4):350, 1921.

[20] A. Mohamed, G. E Dahl, and G. Hinton. Acoustic modeling using deep belief networks. *IEEE transactions on audio, speech, and language processing*, 20(1):14–22, 2011.

[21] Tilda Neuberger, Dorottya Gyarmathy, Tekla Etelka Gráczi, Viktória Horváth, Mária Gósy, and András Beke. Development of a large spontaneous speech database of agglutinative Hungarian language. In *Proceedings of TSD*, pages 424–431, Brno, Czech Republic, Sep 2014.

[22] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak. X-vectors meet emotions: A study on dependencies between emotion and speaker verification. In *Proceedings of ICASSP*, pages 7169–7173, 2020.

[23] Björn Schuller, Stefan Steidl, Anton Batliner, Elika Bergelson, Jarek Krajewski, Christoph Janott, Andrei Amatuni, Marisa Casillas, Amdanda Seidl, Melanie Soderstrom, et al. The Interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring. In *Proceedings of Interspeech*, pages 3442–3446, 2017.

[24] Björn Schuller, Stefan Steidl, Anton Batliner, Simone Hantke, Elika Bergelson, Jarek Krajewski, Christoph Janott, Andrei Amatuni, Marisa Casillas, Amanda Seidl, Melanie Soderstrom, Anne S. Warlaumont, Guillermo Hidalgo, Sebastian Schnieder, Clemens Heiser, Winfried Hohenhorst, Michael Herzog, Maximilian Schmitt, Kun Qian, Yue Zhang, George Trigeorgis, Panagiotis Tzirakis, and Stefanos Zafeiriou. The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, Cold & Snoring. In *Proceedings of Interspeech*, pages 1–5, 2017.

[25] Bjorn W. Schuller, Anton Batliner, Christian Bergler, Cecilia Mascolo, Jing Han, Iulia Lefter, Heysem Kaya, Shahin Amiriparian, Alice Baird, Lukas Stappen, Sandra Ottl, Maurice Gerczuk, Panaguiotis Tzirakis, Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, M.R̃othkrantz Leon J. Joeri Zwerts, Jelle Treep, and Casper Kaandorp. The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. In *Proceedings INTERSPEECH 2021, 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, September 2021. ISCA. to appear.

[26] H. Seki, K. Yamamoto, and S. Nakagawa. A deep neural network integrated with filterbank learning for speech recognition. In *Proceedings of ICASSP*, 2017.

[27] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. X-vectors: Robust DNN embeddings for speaker verification. In *Proceedings of ICASSP*, pages 5329–5333, 2018.

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

7 Összefoglalás

Jelen doktori értekezés olyan módszereket mutat be, amelyek bizonyos betegségekben vagy egészségi állapotban szenvedő egyének nemverbális kommunikációjának kiaknázását célozzák azok automatikus szűrésére. Konkrétabban, a non-verbális kommunikáció egyik pillérét, a paralingvisztikát alkalmaztuk olyan technikák feltárására, amelyek felhasználhatók az alanyok beszédének modellezésére. A paralingvisztika a kommunikáció egy nem lexikális összetevője, amely az intonáción, a hangmagasságon, a beszéd sebességén stb. alapszik, és amely automatikusan feldolgozható és elemezhető. Ezt Computational Paralinguistics-nak hívják, amely úgy definiálható, mint a beszélő beszédében lévő nemverbális látens minták számítási algoritmusok segítségével történő modellezése.

A gépi tanulás segítségével modelleket mutatunk be mind a paralingvisztikai, mind az orvosi célú beszédelemzés különböző forgatókönyveiből, amelyek alkalmasak egy adott betegséggel (például az Alzheimer-kór, Parkinson-kór, depresszió) élő alanyok egészségi állapotának automatikus becslésére.

A dolgozat négy nagy részből áll. Az 1. fejezet bevezeti az olvasót a nemverbális kommunikáció és a paranyelv fogalmába. Ezenkívül röviden ismertetjük a beszéd és a beszélőfelismerés fogalmait. Ugyanez a fejezet a számítógépes paralingvisztika mélyrehatóbb magyarázatával folytatódik, kitérve az említett terület szakirodalmára. A 2. fejezet ismerteti az orvosi célú beszédelemzés során használt gépi tanulási módszerek fogalmait, a patológiás beszéd definícióit, valamint a beszédminták feldolgozásához alkalmazott jellemzőket.

A 3. fejezet az i-vektoros megközelítés használatát tárgyalja a beszédből a jellemzők kinyerésére Alzheimer-kórban vagy enyhe kognitív károsodásban szenvedő alanyok esetében. A 4. fejezetben olyan kísérleteket mutatunk be, amelyekben Fisher vektorokat alkalmaytunk Parkinson-kórban szenvedő alanyok és egészséges kontrollok beszédfelvételeinek feldolgozása során. Az 5. fejezetben az x-vektor technika mint jellemzőkinyerő eszköz alkalmazását tárgyaljuk különböző szűrési feladatok automatikus értékeléséhez, mint például az álmosság észlelése, a depresszió, a konfliktusok léptékezése a beszédben és a főemlősfajok azonosítása hangból. Végül a 6. fejezetben bemutatunk egy időbeli beszédjellemzőkészletet, amely az artikulációs sebességből, a beszédtempóból és további, az alany hezitációját leíró jellemzőkből áll.