1.5em

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

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The ability to exploit verbal and non-verbal forms of communication is said to be of great importance within an individual's personal and professional life. In fact, the success of the human species, historically, constantly depends on these forms of communication for survival. Nowadays, communicating verbally and non-verbally is still of central importance as we deal with daily routines. These two types of communication are an influential factor in an individual's degree of 'success' in both interpersonal relationships and business aspects; and they can define, to a great extent, our physical and psychological states.

With the goal of achieving an effective and enhanced state of communication, the first step may be to understand the verbal and non-verbal aspects and to determine their roles within an individual's interactions with others. While verbal communication involves the literal content of a structured message, whether it is delivered as spoken, written, or signed words, non-verbal communication relates to how the message is interpreted.

Examples of non-verbal signals include body language, posture, eye contact, facial expressions. Likewise, the sound of voice, as another non-verbal communication trait, is capable of transmitting the message by means of the volume or the tone of the voice. These signals are able to project an individual's real intentions and feelings during a social interaction, and may directly affect the perceptions of the people receiving a message

The non-verbal communication in our species has been relevant since time immemorial. Indeed, it is profoundly associated with the evolution of the human language; namely, current forms of languages may be the result of an evolved system of non-verbal communication [140]. Furthermore, over 80% of the way humans communicate is primarily non-verbal [192]. Non-verbal communication can be grouped into the following categories: haptics (touch), kinesics (body movement), **paralanguage** (vocalics), and chronemics (structure of time) [96, 188].

In this book, *paralanguage*, also known as vocalics, is of particular interest. Paralanguage can be defined as a non-lexical component of communication by means of the speech, for example, intonation, pitch and speed of speaking, hesitation noises, emotions. An analysis of these characteristics contributes to the exploration of automatic ways for predicting speaker states and traits such as the age, gender, and even the health condition, among others, which can be catalyzed via Artificial Intelligence (AI) algorithms. Consequently, all the mentioned facts fall into the branch known as *Computational Paralinguistics*. However, the above description might also be suitable for a sub-field named *Pathological Speech Processing*, which covers speech that displays pathological signs that may link to underlying diseases.

Different approaches can be exploited to extract features that represent speech patterns such as acoustic and phonological features, or speech properties. This thesis, however, will present research findings involving Speaker and Speech Recognition techniques that have been utilized to build feature representations, along with the AI methods that were applied using them.

Computational Paralinguistics can offer a wide variety of real-life applications including the analysis and monitoring of speech phenomena, the enhancement of the human-computer interaction experience, and provide tools for screening different diseases (see more in [20, 170]).

1.1 Speech and Speaker Recognition: A Brief Review

Before going into speech and speaker recognition, we should first examine Speech and Signal Processing (SSP). Put simply, speech can be defined as the way humans communicate using language. A signal can be thought as of a function that can transmit information about a specific event; and, a digital signal is a data representation of a sequence of discrete values. Speech signals (e.g., a waveform) can be handled using digital signal processing (DSL) to represent them as either analog or digital forms. Back in the late 19th and the early 20th centuries, when the *radio, telephone*, and *phonograph* were invented, audio signal processing field attracted the interest of researchers. Davis et al. [32] reported studies that attempted, for example, to synthesize the human voice in the 1880s; and a hundred of years later, the use of ASR for automatizing call centers.

1.1.1 Automatic Speech Recognition

Automation has been one of the key features for the emergence of new technology in human history. In the speech processing context, Automatic Speech Recognition (ASR) and speech synthesis are examples of automation tasks that have had great attention since the last century. In essence, the main goal of ASR is to recognize and

translate spoken language into text; while speech synthesis seeks to produce spoken language. These have a variety of real-life applications such as in in-car systems, health care, education, robotics, and home automation.

Using speech signals, researchers initially performed experiments on simple phonetic elements like vowels, attempting early speech processing and recognition techniques. In the early 1950s, systems that could recognize digits [32] or 10 syllables [143] spoken by a single speaker were developed and contributed to the progress of the field. However, it was not until the mid 1980s that the speech recognition met one of its most prevalent and significant methods, namely the Hidden Markov model (HMM) framework [94].

Hidden Markov models can be used to construct general-purpose speech recognition systems based on statistics. HMMs are used to predict a sequence of hidden variables from a set of observed ones. A common example of the use of a HMM is on weather forecasting based on the states of mood a subject. Here, the weather is the *hidden* variable, and mood states are the *observed* variables. Let us define the variables as shown in Figure 1.1: the set of states $X = \{Happy, Grumpy\}$, the set of hidden states $Q = \{Sunny, Rainy\}$, and the set of observed states $O = \{Happy, Grumpy, Grumpy, Happy\}$.

Based on Figure 1.1, we see that the emission probabilities depend on the observed and hidden variables. More specifically, 80% and 60% are the emission probabilities in this case. For instance, a person has an 80% probability of being Happy if the climate at that point of observation is Sunny. Likewise, the same person would have a 60% chance of being Grumpy given that the climate is Rainy.

The same figure depicts the transition probabilities, which are those that account for the transition from and to hidden states. For example, as the weather is a hidden state that influences the observed variables, there exists a correlation between Sunny days in a row and alternate Rainy days. There is 80% and 60% of probability that Sunny and Rainy weather will occur in consecutive days, respectively. HMMs are or were employed in speech recognition due to the fact that a speech signal can be viewed as a short-time stationary signal, and also because they are simple and computationally feasible in practice.

In the past few years, approaches involving Neural Networks and Deep Learning have started to dominate the ASR field, whether it is combined with HMMs [30, 214], or as standalone 'end-to-end' approaches [14, 74]. In contrast with HMMs, DNNs make less explicit assumptions about feature statistical variables, and tend to estimate the probability values of a speech feature segment in a discriminative fashion more efficiently. Conventional DNNs lack the ability to model temporal parameters, but they could be used as a feature pre-processing technique prior to using a HMM. However, there are more sophisticated networks that can efficiently handle this task such as Recurrent Neural Networks (RNN), and Long-Short Term Memory (LSTM).

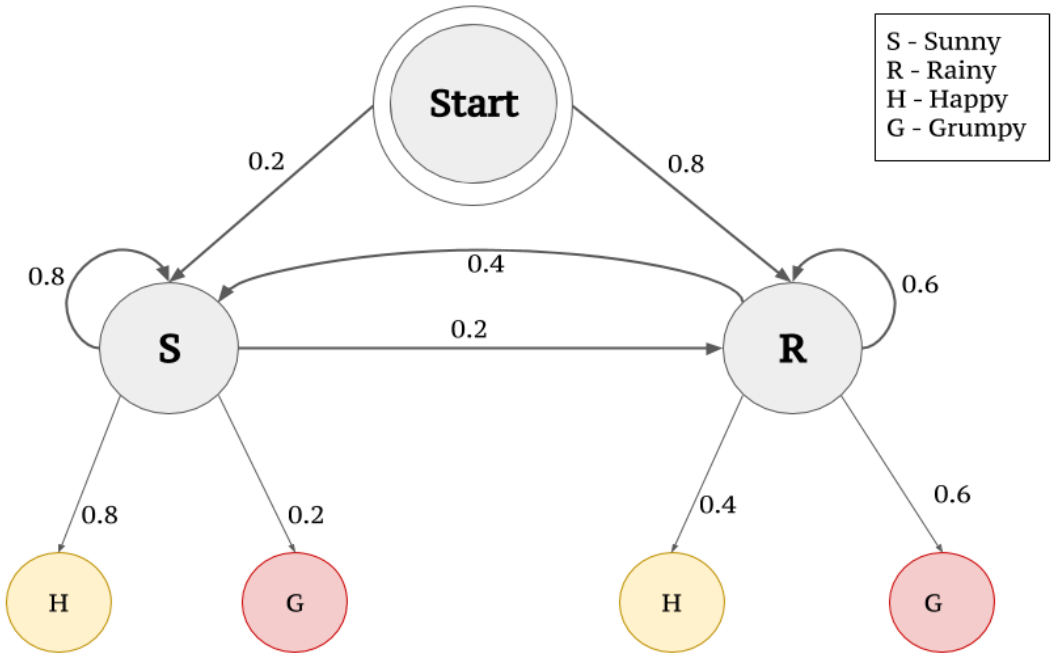


Figure 1.1: *Markov Model as Finite State Machine.*[1]

1.1.2 Speaker Recognition

In a similar way, Speaker Recognition (SR) has also been a focus of interest over the past few decades. One of the first SR systems was reported in [107] as "Voiceprint Verification" in 1962. Then, in 1977, researchers used spectral analysis to build the first autonomous *text-dependent* SR system [40, 57]. In contrast, *text-independent* approaches only saw major advances when the cepstrum was introduced to SR in 1981 [56]. One of the biggest advances in SR took place in the early 90s with the use of Gaussian Mixture Models (GMM) for modelling voice features [160]. Later, this paradigm was improved by adding the so-called Universal Background Model (UBM) [159].

The adoption of GMMs (and later UBMs) led to significant advancements in SR. And, almost a decade after, the field experienced another milestone based on Joint Factor Analysis (JFA) [103] and GMM supervectors [21]. Representations called the *i-vectors* [34] built using a Front-end Factor Analysis approach, achieved the best performances in SR at that time. Later in 2018, the emergence of a new method that relies on Deep Neural Networks (DNN) for feature extraction, and probabilistic linear discriminant analysis (PLDA) for classification demonstrated state-of-the-art results in speaker recognition (see *x-vectors* in [186]). Then in 2020 improvements on the x-vector architecture were carried out by means of emphasized channel attention,

[1] Source: `Source:https://vivekvinushanth.medium.com/`

propagation and aggregation over the same TDNN. This method is called ECAPA-TDNN [38] and is the current state-of-the-art for the Speaker Recognition.

1.2 Paralanguage

One of the most important channels of non-verbal communication is that of paralanguage. It is a non-linguistic category mainly related to the speech and it is capable of modifying the meaning of a message via paralinguistic properties like volume, pitch, rhythm, intonation, among others [96]. Every speech signal or utterance contains a voice, which has these paralinguistic properties that can reveal information about a speaker's state then the age, sex, gender, geographic origin, the emotional state, and even their state of health. The latter concerns the evaluation of the speech properties like phonation, fluency, and intonation, which may be affected when a subject suffers from a pathological speech condition [167].

As stated in Knapp and Hall [110], the voice qualities that comprise the paralanguage involve the pitch, rhythm, tempo, articulation, and resonance. There are acoustic cues that correlate with specific emotions. Some examples of this are: the pitch can affect social meanings; the silent and the filled pauses may interfere with the quality of the delivered message; speech disfluency affects the flow of the conversation; and the volume may influence the state of a person.

Pitch is basically the perceived frequency of the sound. Pause is nothing but a temporary stop, a hesitation, or a rest; while the filled pauses may be viewed as simple hesitation sounds. Speech disfluency concerns irregularities, breaks or non-lexical vocables occurrence in a person's speech. As for volume, it is related to emotions like excitement and fear [110, 167]. These and other traits are present within the speech of every individual and they can be used to automatically analyze a speaker trait or state.

1.2.1 Computational Paralinguistics

The term 'computational' refers to a computer processing and analysis of a specific phenomenon in an automatic way. *Computational Paralinguistics*, can thus be defined as the study of modelling non-verbal latent patterns within the speech of a speaker by means of a computer; and these patterns go beyond the *linguistic* approach. As mentioned above, the paralinguistic properties of the voice convey information about the speaker such as age, gender, and personality. These are defined as speaker *traits*. These voice properties may also contain information about the emotions of a speaker, and they are called speaker *states* [167].

Traits are long-term and states are short-term; however, in between there are other kinds of short-term states such as the sleepiness, having a cold, and being

intoxicated by alcohol. Computational Paralinguistics, which here is also referred as to ***CP***, investigates these states and traits which are latent in the speech signal of a given individual, and it seeks to identify non-verbal patterns by means of Speech and Signal Processing along with **Machine Learning (ML)** discrimination algorithms.

1.2.2 Contemporary Research

Over the last decade, studies have addressed tasks for the automatic detection of speech disorders [25, 46, 77], marking significant advancements in the area of CP. Perhaps one of the most relevant and continuous impulses to the field is the so-called Computational Paralinguistic Challenge (ComParE) [168] as part of the Interspeech Conference [2] since 2009; organized every year ever since. It is an open challenge series that concentrates on certain areas and communities of CP such as audio, speech and signal processing, and affective and behavioural computing. Moreover, it also relates with *Pathological Speech Processing*, which involves health, medicine, and psychology related tasks.

The area of computational paralinguistics differs from Automatic Speech Recognition (ASR), which focuses on the actual *content* of the speech of an audio signal. Here, computational paralinguistics may provide the necessary tools for determining the *way* speech is spoken. Various studies have offered promising results in this area, e.g., diagnosing neuro-degenerative diseases using the speech of the patients [41, 42, 68], the classification of crying sounds and heart beats [70], estimating the sincerity of apologies [64], and determining the state of depression in a subject [27].

Usually, studies focused on acoustic, articulatory, or phonological approaches for modelling latent representations of a given utterance [63, 183, 198]. These kinds of representations are much easier to interpret and are not computationally expensive. Speaker Recognition methods can be applied to exploit these features and obtain more meaningful traits. To name a few, the i-vector approach showed great potential for recognizing emotions from the speakers [210], for age estimation [76], and even for discovering things in forensics [125]. More studies on both computational paralinguistics and pathological speech processing fields will be discussed in Chapter 2, Section 2.2.

1.3 Structure of the Dissertation

Chapter 2: This chapter covers the use of machine learning methods and algorithms in the field of computational paralinguistics. More specifically, we provide

[2] https://www.isca-speech.org/iscaweb/index.php/conferences

an overview of ML and supervised learning, as well as the classification/estimation algorithms employed in our methodology. Also, we go through the definitions of pathological speech processing and the feature representations utilized.

Chapter 3: Here, we outline the use of front-end factor analysis as a means for the automatic screening of Alzheimer's Disease. In more detail, we introduce the utilization of the so-called i-vector features for modelling the speech of speakers suffering from the mentioned neuro-degenerative disease. We propose a pipeline that achieves better performance scores on the specific corpus presented in this chapter.

Chapter 4: With the aim of introducing an automatic assessment approach for different types of paralinguistic tasks based on the speech, here we explain the use of the Fisher vector method for Parkinson's Disease, cold speech, escalation in the dialogue, and primate species detection. Originally intended for image recognition, we show that Fisher vectors can also capture meaningful speaker traits from the speech of subjects.

Chapter 5: This chapter describes deep neural network embeddings applied to distinct types of tasks such as the excessive daytime sleepiness, clinical depression, escalation of the dialogue, and primates species chatter. Specifically, we use x-vector embeddings to model the speech samples of individuals from the above-mentioned tasks. We demonstrate that training x-vector extractors from scratch and with in-domain corpora both lead to improvements in the general classification performances.

Chapter 6: Here, we describe how automatic speech recognition methods can be leveraged to automatically assess both Mild Cognitive Impairment and Alzheimer's Disease. We show that temporal speech parameters as well as posterior-thresholding representations are able to produce robust features at the moment of modelling the speech of subjects. On the one hand, the first method shows that the language for training the ASR system is of secondary importance in terms of overall performance. On the other hand, we show that a full ASR system is not required to have a robust set of features that model the speech of patients.

Chapter 2

Machine Learning and Pathological Speech Processing

Thus far, this thesis has described the principles, evolution, and early studies comprising Speaker Recognition, Speech Recognition, and Computational Paralinguistics. Carrying out data processing for achieving a proper feature extraction phase is only one part of the story, nonetheless. In order to be able to produce an automatic screening for a given speech-pathology, there is the necessity to rely on discrimination algorithms using the processed information. Machine Learning (ML) approaches come to play a relevant role in this scenario.

This chapter focuses on detailing the ML methods used for addressing the discrimination tasks that will be introduced later in this thesis. Here, a brief explanation of ML will be provided, along with the related algorithms employed to handle the experiments conducted. Also, a review and the progress of Pathological Speech Processing as well as its connection to ML will be discussed.

2.1 Machine Learning

Humanity's ambition for automation is one of the most relevant driving forces that has encouraged technological innovation over time. Although the earliest instances of modern automation dates back to the 1980s, automation has been around since ancient times. For instance, automata were employed in the ancient Greece to animate statues of divinities for religious purposes; and even for the invention of the first cuckoo clock which served to keep track of time [18]. In the modern world, automation has countless applications, such as in consumer electronics, automobiles, kitchen tools, medical equipment, and more. Seeking to transcend from 'simple' automation to more sophisticated scenarios, people began to explore the possibility of simulating *human thinking*, which today is known as Artificial Intelligence (AI).

Artificial Intelligence is a broader term that involves Machine Learning as an interdisciplinary sub-field of study. To put it simply, Machine Learning can be thought of as having a computer to take actions, and make decisions or predictions automatically, that is, without it being explicitly programmed. This can be achieved by means of algorithms, which, relying on historical data, have the capacity to improve via experience. These are also called learning algorithms. They learn from a provided input (*training data)* and build a model based on it; consequently, they are able produce outputs of new inputted information (*test data)* that generally relates to the context of the data they learned from.

Generally speaking, machine learning is divided into three paradigms: *supervised learning, unsupervised learning,* and *reinforcement learning*. In a 'typical' ML task, the training data comes with labeled instances. That is, suppose there is a set of N training examples of the form $\{(x_1, y_1), ..., (x_N,\ y_N)\}$, where x_i is a feature vector of the i-th example, and y_i is the class label corresponding to that example. This is the case for supervised machine learning; that is, when the label y_i is actually included within the training data. In contrast, unsupervised learning lacks this y_i label, so the learning algorithm itself must discover the patterns present in the training data. On the other hand, reinforcement learning is a rewarding-wise approach where the algorithm interacts with a dynamic environment where it seeks to take actions and maximize the rewards in order to achieve a specific objective.

Although there exists a vast number of research studies on machine learning and related areas [5, 12, 113, 120, 130], the target of this thesis is supervised machine learning as all of the investigations presented involve labeled datasets. In the next section we shall discuss supervised learning and focus on the key parts of this dissertation.

2.1.1 Supervised Machine Learning

Again, let N be a set of training samples of the form $\{(x_1, y_1), ..., (x_N,\ y_N)\}$, such that x_i is the feature vector of the i-th sample and y_i is its label. The learning algorithm looks for a function $g : X \to Y$, where X is the input space and Y the output space. Here, g is an element of the possible functions G, and it can be defined using a scoring function $f : X \times Y \to \mathbb{R}$. Thus, f can be thought of as a target function that best maps an input vector X to an output variable Y, that is, $Y = f(X)$. g is defined as the value returned by y that provides the highest score, namely, $g(x) = \arg\max_y f(x, y)$.

As discussed, the existence of the label y_i makes the machine learning paradigm a *supervised* one. In a nutshell, the learning algorithm *trains* itself by employing the (*labeled*) training data, which serves as a feedback for optimization. Afterwards, the algorithm is put to the test by providing it with *unseen* (labeled) test data in order to measure its discrimination performance.

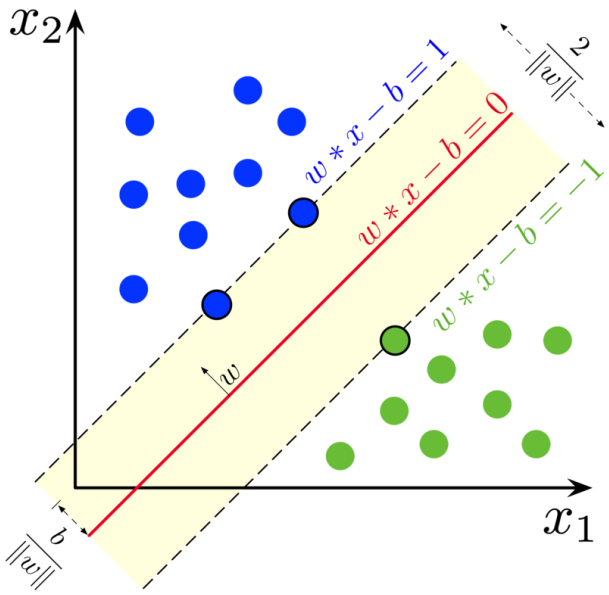


Figure 2.1: *Hyperplane and margins for a Support Vector Machines (two-class problem). The samples on the margin are the support vectors.*

Two types of discrimination tasks exist called **classification** and **regression**. The former seeks to predict *discrete* values such as *assessing the health state of an individual*, while the latter handles the prediction of *continuous* values, for instance, *estimating the level of sleepiness of a subject*. Actually, these predicted values are nothing but the output variables given by the learning algorithm.

Depending on the nature of the task, there may be a different number of possible outputs. This divides classification into two main cases: *binary* classification, and *multi-class* classification. A binary task (also called a two-class problem) involves identifying a particular category of an instance or sample, while a multi-class problem seeks to determine the instance of the particular category which may belong to three or more class labels.

Computational Paralinguistics and Pathological Speech Processing often involve discrimination problems as those described above. For instance, in speech emotion recognition the task might be the classification of the emotional states of a speaker; or the states of depression of individuals in pathological speech tasks. Both are multi-class problems. Differently, examples of binary-class problems may be determining the health condition subjects such as the presence of a neuro-degenerative disease like Alzheimer's or Parkinson's, or the existence of a common cold.

2.1.2 Support Vector Machines

Support Vector Machines (SVM) is a supervised learning algorithm used for classification or regression. Stated simply, SVM seeks to classify a previously unseen sample into one of its two possible categories. The main goal of SVM is to map the training samples to points in the plane in such a way that the width gap between the two classes is maximized. When a new instance arrives, it is mapped to this plane and, depending on which side of the gap it was mapped, the prediction of belonging to one of the two categories is performed.

More specifically, let's suppose a training set of n samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$,, where y_i is 1 or -1 that indicates the category of the sample x_i, which is a p-dimensional real vector. The goal is to find the maximum-margin hyper-plane which splits the samples x_i that belong to $y_i = 1$ from those that $y_i = -1$. The distance between this hyper-plane and the nearest point x_i from any of the categories is maximized. The hyper-plane can be defined by a set of samples (points) x that satisfy $\mathbf{w}^T\mathbf{x} - b = 0$. Here, w is the normal vector to the hyper-plane [80]. (See Figure 2.1.)

SVMs are widely employed in many classification tasks because it is memory efficient, effective in high dimensional spaces (as long as the dimensions are not much greater than the number of instances), and it provides different kernel functions (i.e., linear, RBF) for mapping the data into feature space.

2.1.3 XGBoost Algorithm

eXtreme Gradient Boosting or XGBoost is an implementation based on Gradient Boosting Machines (GBM) [53]. GBM is a regression/classification algorithm that makes use of an ensemble of *weak* models, i.e. small decision trees, to make predictions. A decision tree ensemble in XGBoost is a set of CARTs (Classification and Regression Trees). Put simply, GBM sequentially adds *decision tree* models to correct the predictions made by the previous models, and based on gradient descent, it minimizes the loss function. This is iterated until the objective function (training loss and regularization) finds that no further improvement can be made [137]. Both XGBoost and GBM, basically act in the same manner; however, the main difference between them is that XGBoost, in order to control over-fitting, employs a more regularized model than GBM does.

This algorithm is widely used in machine learning mostly due to its scaling capability and model performance; it was designed to exploit the limits of the computational resources for GBM algorithms [26]. Our decision to use XGBoost was also influenced by its advanced capability for performing model tuning. We see the performance of XGBoost in [121, 202, 203], where the authors report high scores in speech-related classification tasks. In the experiments described in the following chapters, we will employ the Python implementation of XGBoost [26].

2.2 Pathological Speech Processing

Speech disorder is what the American Speech-Language Hearing Association (ASHA) categorizes as one of the five types of pathological disorders; being language, social communication, cognitive communication, and swallowing the other four types of disorders [11]. Pathological speech can be defined as a set of specific disorders that hinders a subject's ability to communicate as a consequence of some underlying disease. Such disorders may include communication disorders (speech and language), voice disorders, and swallowing disorders [10].

In the last decades, studies have proposed a sound theoretical framework based on the normal process of speech and language production that seeks to contribute in the research and management of communication disorders related to pathological speech. Based on the traditional three-level speech production model in [90], an alternate framework is proposed for speech disorders in [15, 129]. This framework contains a hierarchy that characterizes pathological speech in a four-level approach, including linguistic encoding (conceptualization), programming (formulation), *motor planning*, and execution (more details in [129]).

Speech production and pathological speech are said to be strongly correlated [102, 105, 199]. In this context, speech-language pathologists have undertaken studies on specific speech and language disorders involving dysarthia (difficulty on controlling the muscles used for speech production) [115, 213], apraxia (a problem with the motor coordination of speech) [106], dysphonia (involuntary sounds affecting the pitch and volume of the voice) [189], aphasia (inability to communicate) [205], and even neuro-degenerative diseases like Parkinson's [39] and Alzheimer's [16].

Besides Computational Paralinguistics, *Pathological Speech Processing* (PSP) was also included in the ComParE Challenges mentioned in Chapter 1, Section 1.2.2. PSP involves health related areas such as medicine and psychology. To name a few instances, one of the former challenges was to determine the intelligibility of a speaker [174]. Another task involved autism classification from the speakers as a type of pathology [175]. Later on, a task related to cognitive psychology was addressed in [171] where the objective was to predict the levels of cognitive load from the subjects. And, in a classification problem, Parkinson's Disease had to be told apart from healthy speakers [173]. A more recent task related to health was conducted in 2021, where the problem was to discriminate between healthy and COVID-19 speakers [176]. Apart from early contributions (see e.g., [9, 29, 79, 208]), the research produced by holding these kinds of challenges has proved relevant to the development of both Computational Paralinguistics and Pathological Speech Processing. In Chapter 2, an overview of the use of Artificial Intelligence algorithms for catalyzing tasks from both fields of study will be discussed.

2.2.1 Corpora and Feature Representations

The quantity and quality of training instances are relevant aspects that have a direct impact on a machine learning task. Small datasets, due to their nature, tend to produce lower performances than their large counterparts at the moment of recognizing patterns [158]. Different methods need to be applied when dealing with small datasets since the scarcity of the samples may affect the generalization ability of the algorithm. Moreover, an evaluation on small datasets may result in optimistic estimation performance scores, which means that the model is not able to assess new instances in a proper way.

The size of a corpus, both in medical and paralinguistic tasks, is usually small due to the inherent nature of data collection. Gathering instances for these kind of datasets often incurs in a high cost of data collection and relies merely on humans most of the time. The amount of samples is scarce due to several factors that involve the data collection process, which sometimes turns the mechanism of gathering data into a time-consuming or complex task. For instance, the type of procedure, the kind of participants, the instruments employed, or even the physical environment where the samples are collected. A real-life example of our specific case is the collection of speech samples taken from patients required to build a specific kind of corpus. For this, the sampling scenario must have good conditions (i.e., a proper microphone, silent rooms, willingness of the patient) in order to obtain a high-quality recording where the background noise is as low as possible and the quality of the voice is acceptable. Most of the time, however, this is not the case and there is even a scarcity of available patients to take the samples from, and real-life recording conditions are usually far from optimal.

To overcome these issues, methods such as Cross-Validation (CV) come into play. This statistical technique is employed to divide datasets into two distinct segments to get an adequate evaluation and comparison of using learning algorithms.

Furthermore, experimenting with small corpora may have a direct impact in the generalization quality of a model at the moment of evaluation. Splitting it into training, development, and test sets would lead to even smaller sub-corpora and cause the model to show *optimistic* performance in the evaluation phase. The reason for this owes to training and testing the model using limited amounts of instances. This may cause the model to *underfit* the data, which means that it will not generalize well on both seen and unseen samples. However, using the entire corpora as-is would naturally lead to *overfitting*. This occurs when the model is presented with similar or the same examples at the moment of both training and evaluation. An *appropriate* scenario would be to have an adequate amount of data to train the model, so that it is capable of learning the patterns necessary to successfully predict unseen instances (see Figure 2.2).

As we said before, these are the main limitations for building a medical (speech)

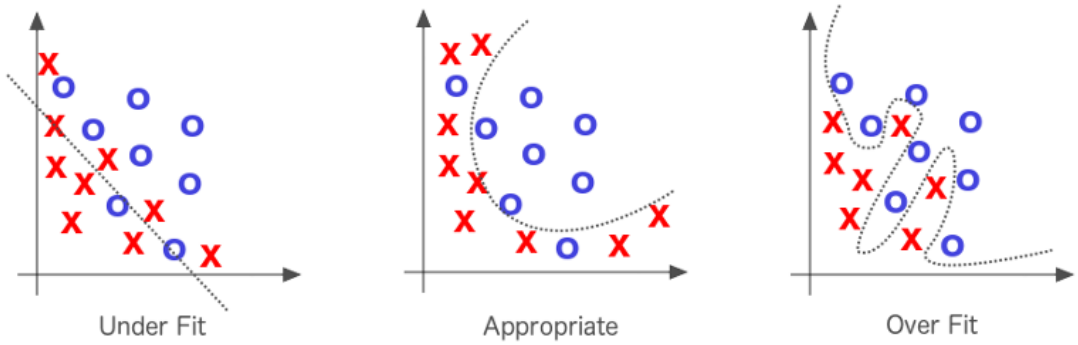


Figure 2.2: *Types of model generalization.*

dataset, which have consequences in both the quality of the data and the size of it. Nevertheless, another issue comes with the ratio between the categorical subjects (e.g., healthy and non-healthy patients), which is unbalanced most of the times. This ratio, unfortunately, is greatly dependant on the availability and willingness of the patients, as well as on the number of existing patients. This will affect the number of samples for each class distribution.

In the case of imbalanced datasets, the estimation algorithms tend to produce optimistic performances and give a bad generalization to unseen instances. One of the methods used to mitigate this scenario is to even the number instances belonging to each category. One way to achieve this is called *undersampling*, which seeks to reduce the number of samples associated with all the classes, to the number of samples of the minority class. This way, the samples become more balanced with respect to their class distribution. Of course, this procedure must be carried out before inputting the data to the final classifier. Some discrimination models such as SVM provide alternative features that can be used for imbalanced corpora as well. A built-in function for balancing the weights of the classes can be used in SVM where there exists the possibility of inputting the weights for each class or just to have SVM to automatically compute them.

2.2.2 Frame-level Features

In order to produce representations that characterize audio signals, there are two approaches that may be applied: the extraction of short-term *frame-level features* (the goal of the studies presented in this thesis), and long-term clip level features. Frames are nothing but very short time intervals that are analyzed when processing an utterance. Frame-level features are usually split into time-domain (calculated using the waveforms), and frequency-domain features (derived from the Fourier transfor-

mation of samples over a given frame). This section describes the different kinds of frame-level features that were utilized to carry out audio-signal processing in the experiments detailed in the later chapters. Specifically, we cover filterbanks, spectrograms, Mel-Frequency Cepstral Coefficients, and Perceptual Linear Predictions.

Filterbanks are basically arrays of digital bandpass filters utilized to analyze an input signal by splitting it into multiple signals with non-overlapping frequency content. A filter bank is classified into two forms: 1) filter bank analysis with a series of analysis filters; 2) filter bank synthesis with a series of synthesis filters. The former separates the input broadband signal into multiple components, each carrying a subband of the original signal. The latter unites these sub-bands into a single broadband signal, which is just a reconstruction of the original input signal.

A filter bank can be employed in bandwidth reduction, spectral composition and decompositions of signals, sample rate modification, among others. These type of frame-level representations are relevant in modern signal and image processing applications such as audio and image coding; and have been applied successfully in speech processing [201], as well as in speech recognition [182].

A **spectrogram** is nothing but an 'image' that represents a particular waveform. More in specific, the spectrogram shows the signal strength (loudness) of the audio-signal over time at various frequency values. In this way, the presence of energy levels as well as how the energy varies over time is easier to perceive. Typical real-life applications of spectrograms include in medicine, the study of phonetics, speech synthesis, among others.

Figure 2.3 shows the form of a spectrogram. It is basically a two-dimensional (frequency in kHz, and time in seconds) graph with a third dimension represented by colors (intensity). The frequency can be thought of as the pitch or tone of the utterance, having its lowest and highest frequencies at the bottom and at the top, respectively. The amplitude, or the 'loudness' of a particular frequency at a specific time is represented by the color. Darker tones mean lower amplitudes and brighter ones mean higher (louder) amplitudes. Spectrograms are of great value in the speech processing area as they are able to capture robust representations at the frame-level [75, 209].

Among the most popular short-term acoustic features are the **Mel-Frequency Cepstral Coefficients (MFCC)**, which can be obtained by implementing the following operations on the utterances: power spectrum, logarithm, and Discrete Cosine Transform (DCT). These deliver the first coefficients along with another coefficient associated with the energy of the frame. Velocity and acceleration (first and second derivatives) are affixed to the MFCCs together with their energy coefficients. In our experimental framework, we use MFCCs beacuse this technique has proved to be one of the most effective when it comes to creating speaker models [59, 78].

Another popular frame-level feature method is the **Perceptual Linear Predictions (PLP)**, which is very similar to the MFCC method described previously as both

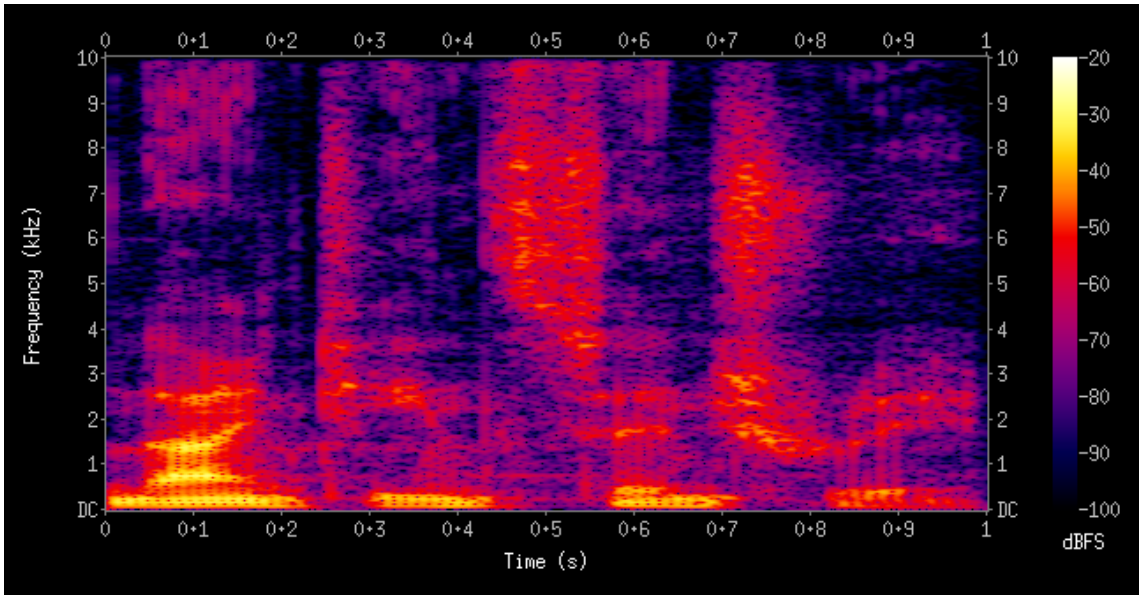


Figure 2.3: *A spectrogram representing two spoken words.*

attempt to model the human auditory system. The PLP consists of a combination of critical bands, intensity-to-loudness compression and equal loudness pre-emphasis obtained from the speech information. In contrast with MFCC, PLP relies on cube-root compression instead of log compression. However, the main difference between the two arises from the filter-banks, equal-loudness pre-emphasis, the intensity-to-loudness conversion, and the application of linear predictions. For instance, the shape of the filter from PLP is trapezoidal while that of the MFCC is triangular, so the number and width of the filters vary [215]. The use of PLP features has proved effectiveness at the moment of modeling frame-level representations for speech recognition [31].

Chapter 3

Front-End Factor Analysis

This chapter introduces the use of Front-End Factor Analysis for the automatic assessment of Alzheimer's Disease (AD) by examining the speech of the subjects. We will cover the current ways of assessing the disease, which are sometimes inefficient. Then we will provide an approach based on the so-called i-vector features that capture meaningful speaker characteristics able to model the speech of patients suffering from Alzheimer's. This could provide the basis for a tool to help clinicians when screening the disease. Similar to Alzheimer's, depression can also be assessed automatically by means of the speech of subjects in a non-invasive manner.

In this chapter, i-vectors can be utilized as a baseline for the discrimination of clinical depression. This, however, will be covered in detail in Chapter 5.

3.1 Introduction

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

The speech difficulties of patients suffering from Alzheimer's Disease become noticeable in the moderate stage of the disease. Such adversities are characterized by the incapacity to recall vocabulary, leading to constant incorrect word substitutions, also known as paraphasias [47]. The vocabulary of the patient is limited to a simple set of phrases or single words; progressively, the patient may entirely lose their speech, resulting in a substantial decrease in the quality of life [47, 49]. In most cases, these factors create the structure and the patterns of speech of these kinds of patients, which is generally formed by syntactic complexity, insufficient speech

fluency, and vocabulary limitation.

There is scarcity of techniques for Alzheimer's screening which makes it hard to diagnose the disease at an early stage. 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. This early stage of diagnosis is generally not easy to achieve [58, 138]. Namely, patients arrive at the clinic when Alzheimer's is already in an advanced state, which lowers the ratio of early AD detection cases. MCI (Mild Cognitive Impairment), as part of the process of dementia, may begin around the age of 40. Screening tests for detecting MCI take a long time, they shortage of pre-clinical state diagnosis and they require a high budget to fund them [73]. Seeking for a non-invasive tool to help with the screening of AD, we will employ a method intended to extract meaningful speaker trait: the *i-vector* approach. This technique was once a state-of-the-art method for speaker recognition [see more in 61, 88].

Likewise, a mental condition such as clinical depression is considered to affect a significant part of society, and it has detrimental effects in both personal and professional life. As a part of our study in [43], we make use of the i-vector approach to establish a robust baseline for the estimation of the degrees of clinical depression.

3.2 Related Works

The i-vector technique makes use of a GMM-UBM (Universal Background Model) for collecting sufficient statistics along with a total variability matrix that contains speaker and channel factors. This matrix is used to extract fixed-sized *i-vector* features from the variable-length utterances. The similarity between two i-vectors is computed by using Probabilistic Linear Discriminant Analysis (PLDA) [61, 62, 88, 104]. In computational paralinguistics, i-vectors have been widely employed as features; for example, Dehak et al. employed i-vectors for language recognition [37], while Grzybowska and Kacprzak utilized i-vectors to determine speaker age [76]. There have been several such works in medicine as well: Garcia et al. performed an evaluation of Parkinson's Disease and the neurological consequences in patients using i-vectors [60], Cummins et al. utilized i-vectors to determine depression from speech [27].

At the moment of writing, no studies have applied i-vector features specifically to predict AD from speech before. We think that, owing to the nature of factor analysis, which is used to obtain information about speaker and channel variabilities, i-vector features are able to capture efficiently the information needed to model an AD subject's speech in a proper way.

3.3 The i-vector Approach

GMM (Gaussian Mixture Model) supervectors [21] and JFA (Joint Factor Analysis) [103] are successful approaches that were once the state-of-the-art systems for robust speaker recognition. In an attempt to combine of both techniques, JFA speaker factors were used as features for SVM classifiers [36]. It was found that the channel factors estimated with JFA not only contain channel effects but speaker-dependent information as well; hence, speaker and channel factors were combined into a single space. Factor Analysis (FA), which is used as a feature extractor, defines a new low-dimensional *total variability space* in which a speech utterance is defined by a new vector called *i-vector* [34] that contains the estimates of the *total factors*:

$$M = m + Tw, \tag{3.1}$$

where M is the Gaussian Mixture Model (GMM) speaker supervector for a given signal; m is the speaker/channel-independent component, namely, the UBM supervector; T is the Total Variability matrix (TV); and w is a standard normal distributed hidden variable, i.e. the i-vector. This vector can be thought of as a representation of a given recording in a lower-dimension space.

In contrast to JFA, i-vectors do not make any distinction between speaker and channel; here, each utterance is assumed to be acquired from a different speaker. The i-vector approach is, in plain words, a dimensionality reduction technique of the GMM supervector.

3.4 The Corpora

The data for the experiments comprises 225 speech signals recorded from 75 subjects (*dementia dataset*), and 44 recordings taken from generic speakers (*subset of the BEA Hungarian Spoken Language Database [139]*). The speech utterances of the dementia dataset were recorded at the Memory Clinic, University of Szeged, Hungary. Three categories of utterances were recorded, namely, subjects suffering from MCI, subjects affected by the early-stage of AD, and subjects having no cognitive impairment at the time of recording. Such categories were matched for age, gender and education. We worked with the utterances of 25 speakers for each speaker group, resulting in a total of 75 speakers and 225 recordings. Table 3.1 lists the clinical characteristics of the control, the MCI and the mAD group. The subset of the BEA corpus consisted of a 120 minute-long set spontaneous speech similar to the recordings collected from the patients.

Table 3.1: The characteristics of the three groups of the study participants. *Groups: MCI = mild cognitive impairment; mAD = mild Alzheimer's Disease. Tests: MMSE = Mini-Mental State Examination; CDT = Clock Drawing Test; ADAS-Cog = Alzheimer's Disease Assessment Scale. Values are given as mean ± standard deviation.*

	Subject groups			**Statistics**	
	Control (n = 25)	**MCI** (n = 25)	**mAD** (n = 25)	F(2;74)	p
Age	70.72 ± 5.004	72.4 ± 3.594	73.96 ± 6.846	2.321	$p = 0.105$
Years of education	12.08 ± 2.326	10.84 ± 2.304	10.76 ± 2.818	2.202	$p = 0.118$
MMSE score	29.24 ± 0.523	27.16 ± 0.898	23.92 ± 2.488	76.213	$p < 0.001$
CDT score	8.88 ± 2.007	6.44 ± 3.429	5.88 ± 3.244	7.254	$p = 0.001$
Adas-COG score	8.575 ± 2.374	12.044 ± 3.205	18.675 ± 5.818	38.35	$p < 0.001$

3.5 The Experiments

Next, we will describe the audio pre-processing steps performed and the type of features utilized before fitting the classification algorithm. We also describe the corpora utilized in our experiments and the way they were carried out. Finally, we present and analyze the results obtained.

3.5.1 Feature Extraction

In all of our experiments, we relied on MFFCs for the pre-processing of the audio signals before training the i-vector models. We extracted 20-dimensional coefficients from the audio signals. We had a frame length of 25 ms, and time-shift of 10 ms. For this, we relied on the Kaldi Speech Recognition Toolkit [155].

3.5.2 The i-vector Training

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 3.1.

Kaldi [155] was used to perform the i-vectors extraction process. Here, two cases were considered when extracting these features: with normalization and without normalization of the audio samples. The values of the following parameters were adjusted in order to train the UBM and get a universal model of the speakers: the

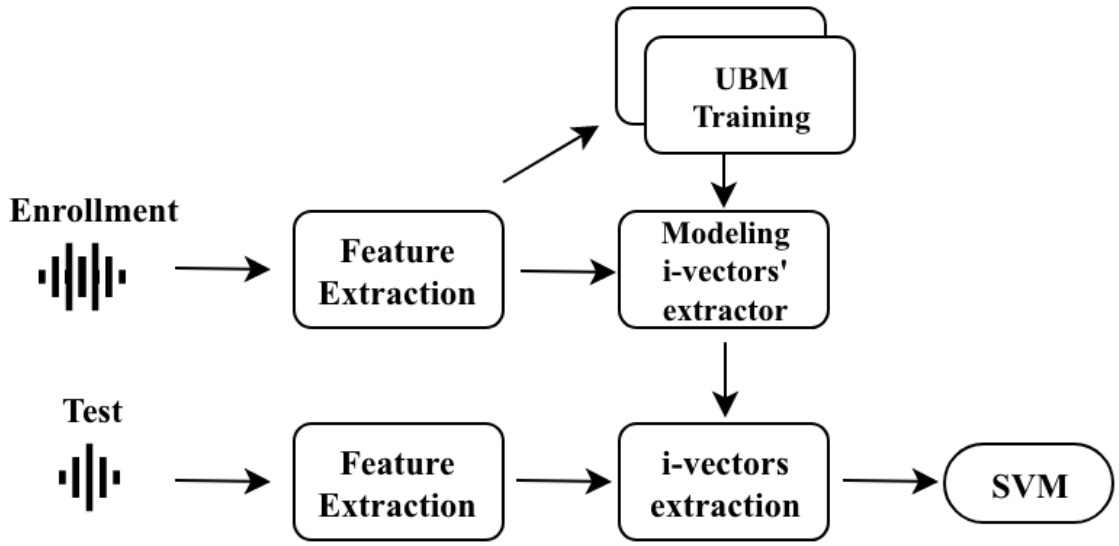


Figure 3.1: *The generic methodology applied for Alzheimer's screening by means of the speech.*

Table 3.2: *Scores obtained when SVM classifies with i-vectors.*

Used Recording(s)	UBM size	Performance (%)			
		Acc.	Prec.	Rec.	F_1
Immediate recall	32	42.7%	82.6%	76.0%	79.2%
Previous day	32	41.3%	72.2%	78.0%	75.0%
Delayed recall	4	46.7%	78.7%	74.0%	76.3%
All utterances	16	56.0%	80.9%	76.0%	78.4%

number of Gaussian components, C, from 2 to 256; and the number of Gaussians to keep per frame, C_f, was given by $log_2(C)$.

3.5.3 Evaluation

We performed our classification with the use of Support-Vector Machines [165]. To avoid overfitting due to having a large number of meta-parameters, we applied a linear kernel; the value of complexity (C) was set in the range $10^{\{-5,-4,...,0,1\}}$. The subjects were classified using 5-fold cross-validation. Each fold contained the utterances of 5 healthy controls, 5 speakers having AD, and 5 speakers suffering from MCI. Each SVM model was trained on the utterances of 60 speakers.

The evaluation was carried out in four ways, where we measured the performance

of the recordings: immediate recall, previous day, delayed recall, and all utterances together, respectively. Table 4.7 lists the results got in terms of F1-scoring and accuracy. The best F1-score outcome belongs to the immediate recall measurement. However, the best accuracy score was obtained when using all the utterances. It can be seen that Immediate Recall and Previous Day recordings performed the best with 32 Gaussian components in the UBM; but this is not true for Delayed recall, and All utterances evaluations, they performed the best when the size of the UBM was 4 and 16, respectively.

3.5.4 Results

In Figure 3.2 we see there is a big difference between the values of accuracy for the set 'All tasks' and the accuracy scores from the other set of tasks (i.e. Immediate recall, Previous day, and Delayed recall). This was because the accuracy score was measured as a 3-wise set, that is, it was obtained in terms of the AD, MCI, and HC classifications. This means that SVM had a 3-class classification with an accuracy score of 56%. In contrast, a 2-wise set used in the rest of the scores, that is, AD and MCI were treated as one class, while HC was the other class, which allowed the classifier to perform better. Thus here the evaluation was basically whether the subject has dementia (AD or MCI) or the subject is healthy (HC). The same figure describes the number of Gaussian components required to get the best results in terms of accuracy, it turns out that the best configurations were obtained when using the number of Gaussian components was less than 32 in the case of Immediate Recall and Previous Day tasks. For Delayed Recall just 4 components were needed. When all the utterances were combined, it was enough to use 16 Gaussian components so as to achieve the best accuracy scores with less computation time. Thus, i-vector features in these experiments performed better when using smaller number of Gaussian components.

It should be added that the best configuration of the number of components *C* in the SVM classifier depended on the type of recordings used, i.e. for the best F1-score (Immediate recall) $C = 10^{-2}$, while for the best accuracy (All utterances) $C = 10^{-3}$. A complexity constant value that is too large may lead to overfit the model; however, a value that is too small may result in over-generalization. Here, the best SVM complexity constant values, which define the tolerance for misclassification, were low in the two best cases, which means that C just needed 'hard' boundaries of tolerance to perform well, and over-fitting was controlled by the cross-validation.

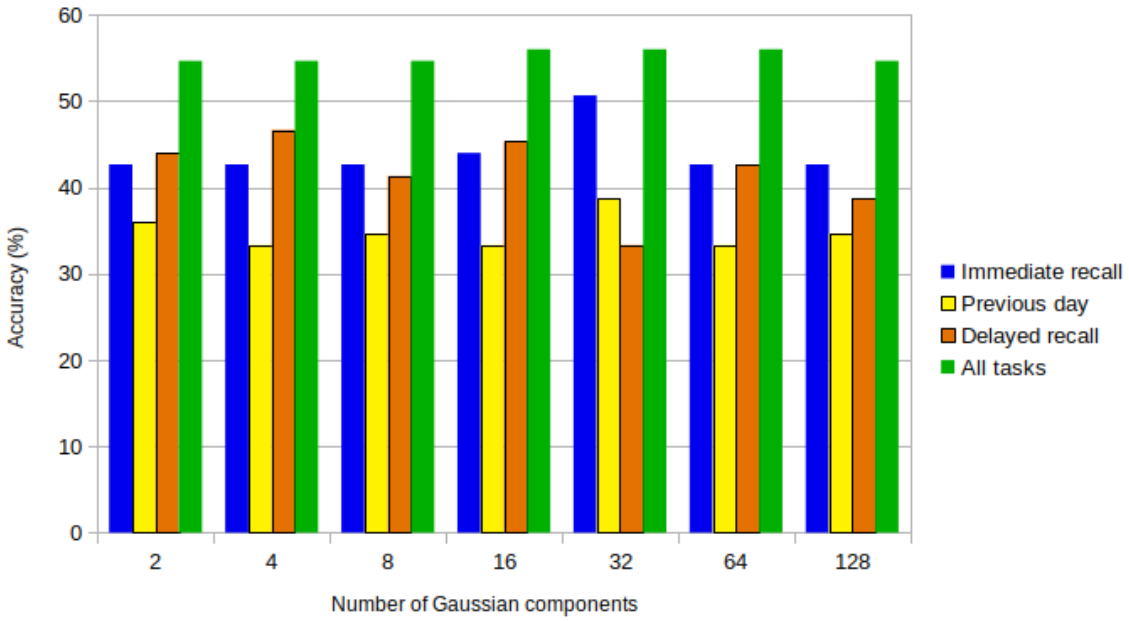


Figure 3.2: *Achieved accuracy scores in terms of the number of Gaussian components.*

3.6 Concluding remarks

Alzheimer's Disease is currently very difficult to accurately diagnose, and the methods of diagnosis generally comprise several costly and time-consuming tasks that the patient may be asked to repeat many times. A successful and precise diagnosis may just depend on the expertise of the physician. Mild Cognitive Impairment is commonly viewed as a prodromal stage of Alzheimer's, it induces a gentle-yet-noticeable decline in cognitive abilities (i.e. memory and thinking). Generally speaking, a person with MCI has a relatively high risk of developing AD or some other type of dementia. Unfortunately, the successful diagnosis of MCI greatly depends on the doctor's experience and judgement which may not be the best. MCI diagnosis is also based on costly biomaker tests (e.g. brain imaging and cerebrospinal fluid tests).

Here, it was 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). The i-vector features were classified using a SVM with 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).

The author of this thesis is responsible for the following contributions presented in this chapter:

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.

Chapter 4

The Fisher Vector

In this chapter we show how the Fisher Vector, a method intended for image classification, can be employed for computational paralinguistic and pathological speech tasks. It turns out that Fisher vector representations are able to capture relevant speaker features that we call FV encodings, which can be applied to examine different speaker states using speech recordings or audio signals. In particular, we experiment with (1) the automatic assessment of Parkinson's Disease, (2) Cold speech, (3) contextual escalation (the level of escalation inferred by the dialogue), and (4) the classification of species of primates. Each sub-section will cover and describe each above task separately.

4.1 Introduction

Seeking to develop a quick, reliable and non-invasive manner to screen, assess and detect the events described in the tasks on the later on, we opt for a technique that enables the use of the speech to this purpose: the Fisher Vector (FV) encoding. We will utilize this method to extract features from the audio-signals available for each task. FV was originally designed as an image representation procedure that pools local image descriptors. The FV can be compared with the Bag-of-Visual-Words (BoV, [148]) technique, both approaches assign a local descriptor to elements in a visual dictionary. However, instead of just storing visual word occurrences, FV representations take into account the difference between dictionary elements and pooled local features, and store their statistics. An advantage of the FV representation is that, regardless of the number of local features (i.e. SIFT for images or frame-level features for audio-signals), it extracts a *fixed-sized* feature representation from each sample. We will show that the proposed approach gives a better performance than for instance, using i-vectors, and provides a simple-yet-effective way of combining the predictions with other methods.

4.1.1 Parkinson's Disease Screening

Some of the classic symptoms of Parkinson's include shaking, rigidity, slowness of movement, and speech difficulties. The motor system is directly affected by them, and this triggers a decrease in the number of dopamine-producing neurons [99]. Such pathologies are often related to one of the most common neuro-degenerative disorders, that is, Parkinson's Disease. A person suffering from Parkinson's is prone to develop changes and disorders in speech and swallowing. This can occur at any time during the disease, but it generally appears as the disease advances. Commonly, the speech of the patient is also affected in terms of its tone, volume, and rate, which might lead to dysprosody. Words comprising the speech of the subject may be slurred or mumbled. Additionally, typical articulatory problems exhibited by PD patients are referred to as dysarthia. Also, the speech can fade away at the end of the sentences; likewise, patients may speak slowly and with a breathy kind of speech [99, 154].

In order to detect diseases like Parkinson's, different kinds of features can be extracted from the speech of the subjects. Several feature types are task-specific, namely, they were designed to capture and reflect specific properties of the actual disease we try to detect. For instance, the articulatory aspects of the speech such as the vowel quality, speech timing, occlusion weakening, or speech coordination [141]. Nevertheless, due to the specific nature of these attributes, they are typically hand-crafted, and as such, they require domain-specific knowledge. Another drawback of such feature types is that they tend to be quite specific to the actual disease, which limits the research effort devoted to their development.

Relying on Computer Tomography (CT) and Magnetic Resonance Imaging (MRI), the brain scans of people can be harnessed to diagnose PD. However, their results may appear to be normal, which makes it difficult for physicians to give an accurate diagnosis. Currently, there are no existing standard blood or laboratory tests that can be utilized to diagnose PD. Hence, the diagnosis, which sometimes may not be the most accurate, is often made based on the medical history of the patient and/or a neurological examination. In some cases, signs and symptoms of PD may be catalogued as the result of normal aging. Limitations within the commonly used process to assess patients with PD include the high cost and the lack of efficiency when evaluating the disease. This process generally has two main drawbacks. Firstly, it greatly depends on the expertise of the clinician, which is subjective; and, secondly, the limitation of taking the patient to the clinic to try out exhaustive medical assessments and screenings [195].

4.1.2 Cold Speech Screening

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 [169].

4.1.3 Escalation in the Dialogue, and Primates Species Detection

Public security, human-computer interactions, or human-to-human conversations are some of the scenarios that can benefit 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. The same goes for public areas like airports and train/bus stations, where passengers frequently converse. Likewise, these automatic tools could be useful for maintaining the safety of the passengers in the above-mentioned areas and maintain public order. To this end, we make use of the **Escalation** Corpus described in [176].

Seeking to develop better tools for monitoring biodiversity, researchers have experimented with bio-acoustics, attempting to annotate or label the different sounds from nature. In our specific case, we are interested in the discrimination of vocalization from **Primates** species. This task was also introduced by Schuller et al. in [176].

4.2 Related Works

Automatic speech analysis has been utilized in many medical branches in order to tackle the above-mentioned obstacles by offering accurate and non-expensive solutions that are able to assess the diagnosis of different neuro-degenerative diseases by the use of speech recordings. Former studies have already addressed these matters for **Parkinson's Disease** screening [134, 216], where the performance of different speech processing techniques such as i-vectors or ASR-based features (e.g. speech tempo or hesitation ratio) are applied. After the initial applications of the FV approach in image classification, it has soon been applied in audio processing as well. Former studies using Fisher Vectors for human speech focused on tasks like categorizing audio files as speech, music and miscellaneous [133], emotion detection [67], and determining food type from eating sounds [101].

Earlier studies applied various approaches for **Cold Classification** from subjects using the speech of the patients. For instance, Gosztolya et al. employed Deep Neural Networks for feature extraction for this purpose [69]. Huckvale and Beke utilized four types of voice features for studying changes in health [87]. Furthermore, Kaya et

al. [100] introduced the application of a weighting scheme on instances of the corpus, making use of a Weighted Kernel Extreme Learning Machine in order to handle the imbalanced data that comprises the URTIC corpus. Like every other computational paralinguistic task, assessing a cold from the speech is a challenging issue. Finding out the latent patterns that characterize or represent a cold speech not only depends on the feature extraction phase but in the data itself too. These includes factors like a limited amount of data, data imbalance, quality of the recordings.

4.3 The Fisher Vector

As mentioned earlier, we rely on the FV approach [91] for our experiments in all the four tasks. In this particular scenario, this encoding technique represents audio-signals as gradients of a global generative GMM of low-level utterance descriptors.

The Fisher Vector technique seeks to represent an image through the extraction of local patch descriptors (e.g. a set of SIFT descriptors). This approach utilizes the same principles as those of the Fisher Kernel (FK) introduced in [91]. This section describes the FK as a method for statistical classification along with its principles applied to the FV. Likewise, it explains the use of FV representations for audio-signals.

4.3.1 The Fisher Kernel

Put it simply, the FK is a way to measure the similarity between two objects by means of their deviation from a generative model. More formally, let us define $X = \{x_t, t = 1, \ldots, T\}$ as a sample of T observations $x_t \in \mathcal{X}$; and υ_λ as the probability density function that models the generative process of the elements in $\mathcal{X}$. Here, $\lambda = [\,\lambda_1, \ldots, \lambda_M]' \in R^M$ represents the parameter vector of υ_λ. Hence, the gradient of the log-likelihood of the data X can be employed as a statistical score function ([91]):

$$G_\lambda^X = \bigtriangledown_\lambda \log \upsilon_\lambda(X), \tag{4.1}$$

which tells the way the parameter υ_λ should be changed in order to best fit the data X. Note that the dimensionality of $G_\lambda^X \in R^M$ is not related to the size of the sample T, instead, it depends on the number of parameters M in λ.

As mentioned before, the FK defines the similarity between two samples, say, X and Y. This can be expressed as:

$$K_{FK}(X, Y) = G_\lambda^{X\prime} F_\lambda^{-1} G_\lambda^Y. \tag{4.2}$$

Since F_λ is positive semi-definite, $F_\lambda = F_\lambda^{-1}$. Eq. (4.3) shows how the Cholesky decomposition $F_\lambda^{-1} = L_\lambda' L_\lambda$ can be utilized to rewrite the Eq. (4.2) in terms of the

dot product:

$$K_{FK}(X,Y) = \mathscr{G}_\lambda^{X\prime}\mathscr{G}_\lambda^{Y}, \tag{4.3}$$

where

$$\mathscr{G}_\lambda^{X} = L_\lambda G_\lambda^{X} = L_\lambda \bigtriangledown_\lambda \log \upsilon_\lambda(X). \tag{4.4}$$

Such a normalized gradient vector is the so-called *Fisher Vector* of X [163]. Both the FV $\mathscr{G}_\lambda^{X}$ and the gradient vector G_λ^{X} have the same dimension.

4.3.2 The Fisher Vector for audio-signals

The SIFT descriptors characterize occurrences of rotation- and scale-invariant primitives of an image [124]. Here, we use MFCC representations as *descriptors* (i.e. frame-level features) of the utterances. Let $X = \{X_t, t = 1 \dots T\}$ be the set of D-dimensional frame-level features extracted from an audio-signal and let the assumption of independent samples hold. Then Eq. (4.4) becomes:

$$\mathscr{G}_\lambda^{X} = \sum_{t=1}^{T} L_\lambda \bigtriangledown_\lambda \log \upsilon_\lambda(X_t). \tag{4.5}$$

The assumption of independence permits the FV to become a sum of normalized gradients statistics $L_\lambda \bigtriangledown_\lambda \log \upsilon_\lambda(x_t)$ calculated for each frame-level feature. That is:

$$X_t \rightarrow \varphi_{FK}(X_t) = L_\lambda \bigtriangledown_\lambda \log \upsilon_\lambda(X_t), \tag{4.6}$$

which describes an operation that can be thought of as a higher dimensional space embedding of the frame-level features X_t.

Simply stated, the FV approach extracts low-level local descriptors from the MFCCs. Then, utilizing a GMM with diagonal covariances, the distribution of the extracted features can be modeled. The log-likelihood gradients of the features modeled by the parameters of such GMM are encoded through the FV. This type of encoding stores the mean and covariance deviation vectors of the K components that form the GMM together with the elements of the frame-level features. The utterance is represented by the concatenation of all the mean and the covariance vectors that gives a final vector of length $(2D + 1)K$, for K quantization cells and D dimensional descriptors [150, 163].

The FV approach can be compared with the traditional encoding method called BoV (Bag of Visual Words), and with a first order encoding method like VLAD (Vector of Locally Aggregated Descriptors). In practice, BoW and VLAD are outperformed by FV due to its second order encoding property of storing additional statistics between codewords and local feature descriptors [178]. Here, we use FV features to encode the MFCC features extracted from audio-signals of HC and PD subjects. FV allows

us to give a complete representation of the sample set by encoding the count of occurrences and high order statistics associated with its distribution.

4.4 The Corpora

4.4.1 PC-GITA Corpus

For the Parkinson's Disease experiments, we used the PC-GITA speech corpus [144], which contains the recorded speech of 100 Colombian Spanish speakers (50 PD patients and 50 HC). All of the patients were evaluated by a neurologist. The corpus contains a total set of 100 recordings of Colombian Spanish speakers, and it is divided into 50 patients suffering from PD and 50 HC. An expert neurologist assessed the diagnosis for each of the patients. The audio-signals were sampled to 16 kHz. The subjects were asked to perform four different tasks during the recordings: six diadochokinetic (DDK) exercises (e.g. the repetition of the sequence of syllables /pa-ta-ka/), monologue speeches, text reading, and ten short sentences. Speech/non-speech and voiced/unvoiced segmentation were the types of segmentation used in this study, the former utilizes energy-threshold Voice Activity Detection (VAD) and the latter makes use of the auto-correlation method [17].

4.4.2 Upper Respiratory Tract Infection Corpus (URTIC)

The URTIC dataset consists of 382 male speakers, 248 female speakers, with a mean age of 29.5 years; and a sampling rate of 44.1kHz downsampled to 16kHz. This dataset was utilized in the cold assessment task. The corpus was provided by the Institute of Safety Technology, University of Wuppertal, Germany. The following tasks were performed by the participants: they had to read short stories (e.g. the well-known story in the field of phonetics *The North Wind and the Sun*, to produce voice commands (such as numbers from 1 to 40), and to narrate spontaneous speech (i.e. tell something about their last weekend or their best vacation). Note that the number of tasks varied for each speaker. The recordings were split into 28,652 chunks of 3 to 10 seconds in length. Specifically, the division of the chunks was carried out in a speaker-independent manner, each partition (i.e. train, development, and test) having 210 speakers. The training and development sets are both comprised by 37 subjects having a cold and 173 subjects not having a cold. The reader may see more details in [172]. The number of samples and classes for each subset is described in Table 4.1.

Table 4.1: *Upper Respiratory Tract Infection Corpus (URTIC).*

Class	Train	Development	Test	Total
Cold	970	1,011	895	2,876
Not-Cold	8,535	8,585	8,656	25,776
Total	9,505	9,596	9,551	28,652

Table 4.2: *The Escalation Corpus.*

Class	Train	Development	Test	Total
Low	156	69	260	485
Medium	75	34	191	300
High	64	16	50	130
Total	295	119	501	915

4.4.3 Escalation Corpus

This dataset consists of both the Dataset of Aggression in Trains (TR) [118] and the Stress at Service Desk Dataset (SD) [117]. The corpus presents unscripted interactions between actors, where friction is present as the speakers spontaneously react to each other based on short scenario descriptions. The TR dataset comprises 21 scenarios of unwanted behaviours in trains and train stations. Such scenarios are, for instance, harassment, theft, travelling without a ticket. The annotation was carried out relying on aggression levels on a 5 point scale by 7 raters. On the other hand, the SD corpus has scenarios of problematic interactions situated at a service desk. For example, a slow and incompetent employee while the customer has an urgent request. These were annotated for stress levels on a 5 point scale by 4 raters. The original labels were mapped onto a 3 point scale: SD classes 1 and 2 and TR class 1 onto Low, SD class 3 and TR class 2 onto Medium, and the rest of the data onto High escalation. The sample distribution is shown in Table 4.2. See more details in [176].

4.4.4 Primates Vocalisation Corpus

As described in [176], the corpus consists of recordings from different species of primates including Chimpanzees, Mandrills, Red-capped mangabeys, and a mixed group of Guenons. The annotation process relied on both manual and semi-automatic annotations: a) initial annotation based on spectrogram analysis and listening; b) vocalisation detection based on energy/variation in specific frequency sub-bands (150 Hz - 2 KHz); and c) final annotation relying on spectrogram analysis and listening, resulting in over 10,000 annotated vocalisations. Background utterances labeling

Table 4.3: *Primate Vocalisations Corpus*

Class	Train	Development	Test	Total
Chimpanzee	2,217	2,217	2,218	6,652
Mandrill	874	874	875	2,623
Red-capped mangabeys	208	209	210	627
Guenons	158	159	159	476
Background	3,458	3,459	3,461	10,378
Total	6,915	6,918	6,923	20,756

was taken based on the recording that were not labeled as vocalisation. Table 4.3 shows the class distribution of the datasets.

4.5 The Experiments

The pipeline carried out in the experiments on Parkinson's consisted of the following steps: (1) VAD-based segmentation, (2) feature extraction, (3) fitting a GMM to the local image features, (4) the 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. A similar workflow is employed for the rest of the tasks (i.e., Cold, Escalation, and Primates), with the difference that segmentation is not taken into consideration, and that for some tasks the kernel of the SVM may be distinct. (See Fig. 6.1). We used the Kaldi Speech Recognition Toolkit [155] for feature extraction. All the FV features were standardized by removing the mean and scaling to unit variance before training the SVM model.

4.5.1 Feature Extraction

For the **Parkinson's** task, the experiments were executed by 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. Before extracting the features we performed speech/non-speech segmentation by means of Voice Activity Detection (VAD), and also by voiced/unvoiced using the auto-correlation method from Praat [17]. For articulation evaluation, the first 22 Bark bands (BBE) in *voiced/unvoiced* and *unvoiced/voiced* transitions were treated as features [145]. Features obtained from phonation and articulation in *voiced segments* constitute a 14-dimensional vector with 30 ms of windows analysis and 5 ms of time shift. These features contained log-energy, pitch (F_0), first and second formants (F_1, F_2) together with their first and second derivatives, Jitter and Shimmer. Prosody in-

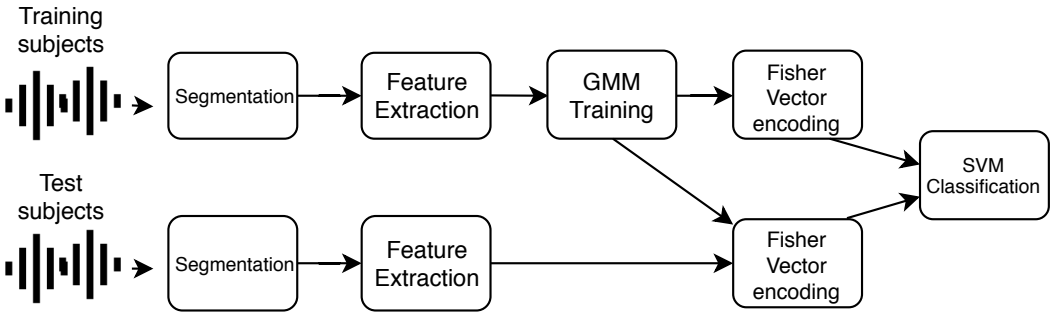


Figure 4.1: *Generic methodology applied in our experiments.*

formation was represented by means of the approach introduced in [35]; hence, we got a 13-dimensional feature vector formed by using the number of voiced frames per segment and the 12 coefficients.

For the **Cold** task, 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 each of the them, we employed a frame-length of 25 ms, and frame-shift of 10 ms.

4.5.2 Training and Evaluation Methods

To construct the FV representation, we experimented with $N = 2, 4, 8, 16, 32, 64$ and 128 Gaussian components in all the tasks. We relied on the VLFeat library in order to get the Fisher vectors [200]. As stated before, the classification was done using a Support Vector Machines algorithm. We employed the libSVM implementation [23] with a linear kernel and, as suggested in [91], the C complexity parameter was set in the range 10^{-5}, ..., 10^{1}.

For the Cold, Escalation, and Primates tasks the performance of the SVM classifier was measured via Unweighted Average Recall (UAR), which is a proper metric for these kinds of paralinguistic tasks; also it is commonly used when there is a need to handle class imbalance situations. As for Parkinson's screening, the metric used is described next.

Parkinson's

Owing to the limited size of the PC-GITA corpus, we conducted the experiments in a speaker-independent 10-fold *nested cross-validation* (CV) setting; each fold contained the utterances of 5 PD and 5 HC speakers. The classification was carried out using the SVM estimator fitted on 9 folds (i.e., 90 speakers). To get the right meta-parameters, we performed *another CV* over the 90 speakers of the training folds. After

determining the optimal N (number of Gaussian components for FV) and C (SVM complexity) meta-parameters, we trained a SVM classifier using 90 speakers based on meta-parameter values mentioned earlier. Hence, we obtained predictions for all speakers without relying on any kind of data or information about the given subject. The SVM's outputs were employed to compute the Area Under the Receiver Operating Characteristics Curve (AUC), which is a widely used statistic for summarizing the performance of automatic classification systems in medical applications. Moreover, we calculated the classification accuracy and F-measure (or F_1) scores. These metrics were calculated by choosing the decision threshold along with the Equal Error Rate (EER).

Cold

The data samples are highly imbalanced. The training dataset consists of 9505 utterances, where 8535 (89.8%) are labeled as *healthy* (not-cold) and the rest, 970 (10.2%), are labeled as *cold*. Likewise, the development dataset comprises 1011 *cold* and 8585 *not-cold* labels, which are 10.53% and 89.47%, respectively. Having a high class imbalance is more likely to affect the performance of the SVM classifier. Hence, we opted for random undersampling, which reduces the number of samples associated with all classes to that of the minority class (i.e., *cold*). In our first experiments we reduced the dimensions of the features via Principal Component Analysis (PCA) [98], keeping a variance of 0.95. Chatfield et al. [24] demonstrated that applying PCA before classification enhances the discrimination task with FV and reduces the memory consumption as well.

Moreover, the FV features were normalized with Power Normalization (PN) and *l2-Normalization*. Power Normalization was found to be helpful for FVs representations [163] as it reduced the impact of the features that became more sparse as the number of Gaussian components increased. In the following experiments, we applied these normalization techniques before reducing the dimensions using PCA. Likewise, we found that *l2-Norm.* helped to alleviate the effect of having different utterances with distinct amounts of background information projected into the extracted features, which attempts to improve the prediction performance. In order to search for the best complexity C value of the SVM, Stratified Group k-fold Cross Validation (CV) was applied over the training and development sets. Stratified Group k-fold CV allowed us to avoid having the same speaker in more than one specific fold while simultaneously preserving the percentage of samples for each class within each fold.

Escalation and Primates

Since both datasets in the Escalation and in the Primates tasks were imbalanced, we performed downsampling by randomly discarding training examples from the more

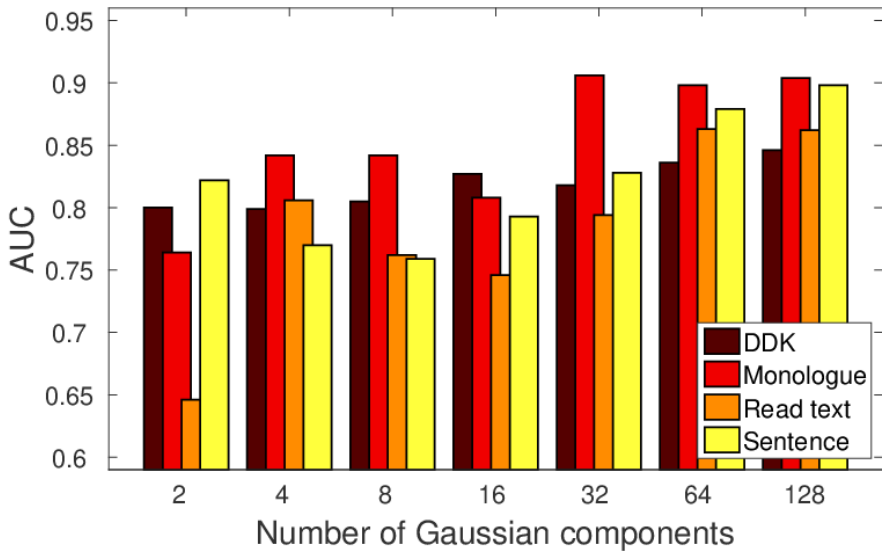


Figure 4.2: *Achieved AUC values as a function of N for the four speaker tasks, when using the MFCC feature set for the **Parkinson's** task.*

frequent classes during training. Since this process introduces a further random factor into the training phase, we decided to fit several models and average out the resulting posterior values. For the **Escalation** task, we repeated the model training 100 times; while we trained 10 models in each case for the **Primates** task. The predictions from the test set were obtained by fitting on the training and development sets combined. Furthermore, we trained independent SVM models for the different types of frame-level features (i.e., MFCCs, FBANKs, and spectrograms), and combined the predictions in the second step by taking the weighted mean of the posterior estimates. Additionally, in the **Escalation** and **Primates** tasks we experimented with the so-called *x-vector* approach, which will be covered more in depth in Chapter 5.

4.5.3 Results and Discussion

Parkinson's results

The results for the different speaker tasks and the various frame-level feature sets are shown in Table 4.4. The best scores in each case were gotten with the MFCC feature set (except for the 'Read text' task, where the accuracy and F_1-scores appeared to be higher with articulatory features along with an identical AUC score). This is probably because the FV approach assumed that the frame-level feature values could be modeled along with a diagonal covariance matrix. This assumption is quite realistic for MFCCs and, perhaps, for the filter bank values, of the voiced/unvoiced transitions (i.e. the articulatory features), but it may not be true for the phonational and prosodic attributes.

Table 4.4: *Results obtained for the various tasks and feature sets for the **Parkinson's** task.*

Task	Features	Acc.	F_1	AUC
DDK	MFCC	**78%**	**0.78**	**0.834**
	Articulation	70%	0.70	0.782
	Phon./Artic.	62%	0.62	0.737
	Prosody	62%	0.62	0.666
Monologue	MFCC	**80%**	**0.80**	**0.880**
	Articulation	76%	0.76	0.847
	Phon./Artic.	70%	0.70	0.749
	Prosody	58%	0.58	0.621
Read text	MFCC	76%	0.76	**0.848**
	Articulation	**80%**	**0.80**	**0.848**
	Phon./Artic.	72%	0.72	0.758
	Prosody	78%	0.78	0.798
Sentences	MFCC	**80%**	**0.80**	**0.891**
	Articulation	76%	0.76	0.834
	Phon./Artic.	76%	0.76	0.804
	Prosody	62%	0.62	0.684

Next, we focused on the trends in the optimal number of Gaussian components (i.e. N) for the tasks. We tried out all the possible N values, and then just the C complexity parameter was determined in a nested CV. (Of course, this was not a completely fair setup from a machine learning perspective. Still, in our opinion, this small amount of 'peeking' was both necessary and acceptable in this scenario. Then we could focus on classification performance as a function of N.) Fig. 4.2 shows the AUC scores for the MFCC features.

In general, using fewer GMMs ($N \leq 16$) led to a sub-optimal performance, excepting the DDK task, where we can see a close-to-optimal AUC value even for $N = 16$. For the Monologue task, $N = 32$ components were needed for optimal performance, while $N = 64$ and $N = 128$ were enough for the Read text and the Sentences tasks, respectively. AUC scores were above 0.8 for three tasks even for $N = 4$; as it meant 104-176 attributes for each subject, we achieved relatively high classification performance even with this compact representation.

Moreover, aiming to improve the classification performance, we experimented with *late fusion* [172]. The class-wise posterior estimates produced by the SVM algorithm were employed for taking the mean of two or more posterior vectors and achieve a 'feature set combination'. We combined the different *feature sets* and *tasks*, and applied late fusion by taking the weighted mean of the posterior estimates with

Table 4.5: *Results obtained when combining the different feature sets for the 'Monologue' task (for **Parkinson's**)*

Features	Acc.	F_1	AUC
MFCC	80%	0.80	0.880
MFCC + Articulation	**84%**	**0.84**	**0.908**
MFCC + Phon./Artic.	78%	0.78	0.871
MFCC + Prosody	78%	0.78	0.878
MFCC + Artic. + Phon./Artic	**82%**	**0.82**	**0.897**
MFCC + Artic. + Prosody	**84%**	**0.84**	**0.900**
All feature sets	**82%**	**0.82**	**0.895**

an increment of 0.05; the weights are determined in a nested cross-validation process.

The results of 'task set combination' (for the Monologue task) are shown in Table 4.5. Note that the results regarding the MFCC feature set improve when articulatory features are added: the classification accuracy rose from 80% to 84%, the corresponding F_1 value went up from 0.8 to 0.84, and the AUC value of the PD class also rose from 0.880 to 0.908. However, adding more feature sets proved futile: although the accuracy and F-measure values remained constant even after utilizing the prosodic features as well, the AUC score fell to 0.900. Still, the 0.908 score achieved by fusing the predictions got from the first two feature sets brought an improvement of 20% in terms of the RER.

As per the 'feature set combination' experiments, the results are displayed in Table 4.6, the highest results correspond to the 'Read text' task. It matched the performance of MFCCs in terms of the AUC, while the accuracy and F_1 values appeared to be higher. Besides 'Read text', using the 'Monologue' task resulted in a performance improvement, while incorporating the 'DDK' task as well increased the AUC value even further (although the classification accuracy and F_1 dropped slightly), leading to a 29% of RER score.

Cold results

As shown in Table 4.7, for the baseline we utilized the ComParE functionals (i.e. Bag-of-Audio-Words features) that were originally presented and described in [169]. According to the results outlined in Table 4.7, these representations achieved an UAR score of 67.3% on the test set. This score was slightly outperformed by two of our configurations: when PCA was applied (67.65%), and when PN was applied along with PCA (67.81%). Table 1 shows the results obtained when using Fisher Vectors with their complete number of features as a function of their reduced number of

Table 4.6: *Results obtained when combining the different tasks for the articulatory features for the **Parkinson's** task.*

Features	Acc.	F_1	AUC
Read text	80%	0.80	0.848
Read text + DDK	76%	0.76	**0.860**
Read text + Monologue	**84%**	**0.84**	**0.878**
Read text + Sentences	74%	0.74	**0.862**
Read text + Monol. + DDK	**82%**	**0.82**	**0.892**
Read text + Monol. + Sentences	76%	0.76	**0.867**
All tasks	80%	0.80	**0.877**

Table 4.7: *UAR scores obtained for the **Cold** task.*

Features	GMM size	Performance (%)	
		Cross-Val	Test
ComParE (BoAW-baseline)	-	64.54%	67.30%
Fisher Vectors	64	63.98%	66.12%
Fisher Vectors + PCA	64	64.72%	67.65%
Fisher Vectors + PN + PCA	64	64.92%	67.81%
ComParE + Fisher Vectors (+PN+PCA)	-	63.01%	70.17%

features. As can be seen, when the classifier learned the *raw* Fisher Vector features it achieved a UAR score of 63.98% in the CV. On the test set the performance was higher (66.12%). PCA proved to be useful here by contributing to a better classification performance in both CV and test phases (64.72% and 67.65%, respectively). However, we found that applying PN before PCA was effective as the CV and test UAR scores increased to 64.92% and 67.81%, respectively. Afterwards, we used the ComParE BoAW [172] feature set posterior probabilities and we combined them with those of the (power-normalized and reduced) Fisher Vectors, that is, we carried out a *late fusion*. The UAR score rose to 70.17% of UAR score on test set, which outperformed the BoAW baseline.

Escalation and Primates results

It should be noted that tables for both tasks are sparse due to the challenge rules. Table 4.8 shows the results obtained for the **Escalation** task. We can see that the ensemble x-vector approach performed well, considering that it is a 3-class classification task: the UAR values are in the range $62.6 \ldots 72.5\%$, the last being just as

Table 4.8: *The results obtained for the **Escalation** Task.*

Feature Set	Dev	Test
ComParE functionals	72.8%	—
Ensemble x-vectors (MFCC)	62.6%	—
Ensemble x-vectors (FBANK)	68.0%	—
Ensemble x-vectors (spectrogram)	72.5%	—
Fisher vectors (FBANK + Δ + ΔΔ)	74.3%	—
ComParE + x-vectors (spectr.)	74.5%	61.5%
ComParE + FV (FBANK + Δ + ΔΔ)	77.8%	63.2%
Official ComParE baseline	—	59.8%

effective as ComParE functionals (72.8%). By combining the two feature types, we achieved a slight improvement (74.3%). Fisher vectors were slightly better (note that, due to the lack of space, we only reported the best FV configuration); in the end, we achieved the best results with the combination of ComParE functionals and FVs. Our two test set submissions achieved similar results to the scores on the development set: FVs slightly outperformed the ensemble x-vectors. However, both approaches scored above the official Challenge baseline (obtained via Bag-of-Audio-Words).

Table 4.9 showcases the results got for the **Primates** task. For this task, FBANK-based and MFCC-based (ensemble) x-vectors turned out to be better than the spectrogram-based one; and although even the best one, relying on FBANKs, performed below the standard ComParE functionals attribute set (78.3% and 81.1%, respectively), they could be combined effectively, as the UAR score on the development set improved to 82.6% in this case. Just like that for the Escalation corpus, we achieved even better scores with the Fisher vectors (although now Δs and ΔΔs proved to be redundant); this UAR score of 82.7%, measured on the development set, could further be improved to 87.5% by a combination with the ComParE functionals. Regarding the test set scores, the combination of the ComParE feature set with ensemble x-vectors resulted in a test set UAR value below the Challenge baseline. However, we still managed to surpass the ComParE functionals score reported in the baseline paper (see [176]), while with the ComParE + FV method we even exceeded the official baseline score of 87.5%, which was a fusion of five methods itself. This value was further exceeded by incorporating the auDeep features as well.

Table 4.9: *The results obtained for the **Primates** Task*

Feature Set	Dev	Test
ComParE functionals	81.1%	—
Ensemble x-vectors (MFCC)	75.7%	—
Ensemble x-vectors (FBANK)	78.3%	—
Ensemble x-vectors (spectrogram)	70.7%	—
Fisher vectors (FBANK)	82.7%	—
ComParE + x-vectors (FBANK)	82.6%	83.3%
ComParE + FV (FBANK)	87.5%	88.8%
ComParE + FV (FBANK) + auDeep	88.2%	89.8%
Official ComParE baseline	—	87.5%

4.6 Concluding remarks

Parkinson's Disease

The Parkinson's states are often difficult to diagnose accurately by doctors. This creates many limitations in terms of time and costs for the patients that possess the pathology of the disease and need to be assessed. A non-invasive and promising procedure for assessing and diagnosing Parkinson's is the automatic analysis of speech of the subject. Our study showed how useful FV are 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. Samples comprising such dataset were segmented, and cepstral, articulatory, phonological and prosodic features were extracted from the voiced parts. These features were represented by FV-encoding and they were classified using Support-Vector Machines. This workflow produced a high-precision classification performance.

The first experiments revealed that MFCC features performed the best in three of the four tasks. The task 'Sentences' came first 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, while using the articulatory features, but incorporating the predictions for the tasks 'Read text', 'Monologue' and 'DDK', also led to a significant improvement over relying on the 'Read text' task only. Using different feature sets and/or tasks is not the only possible combination approach possible.

Cold

Compared with studies conducted by other teams using the same dataset [87, 172], 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 [172]. PCA in combination with the SVM allowed us to carry out a better classification of the actual data while monitoring the memory consumption. PN helped to reduce the impact of the features that increase their sparsity as the number of Gaussian components increase. Furthermore, L2-normalization was applied before fitting the data. This helped to alleviate the effect of having different utterances with distinct amounts of background information projected into the extracted features, which attempts to improve the prediction performance.

Escalation and Primates

Although this chapter focuses on Fisher vectors, our main contribution for this specific task relied on the x-vector technique (covered in more detail in Chapter 5). Our UAR scores on the development set demonstrated the superiority of the ensemble classifiers over the independent x-vector-based ones. However, the Fisher vector approach, was more successful at the moment of modelling these particular corpora. Our experiments managed to overpass all the official baselines presented in [176]. 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].

The author of this PhD thesis is responsible for the following contributions presented in this chapter:

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. I constructed a machine learning model capable of discriminating cold from the speech of individuals using Fisher vectors. I demonstrated the superiority

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

of XGBoost over SVM when of employing the above-mentioned features for cold speech classification.

II / 3. As part of the procedures conducted in this study, 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 the automatic assessment of the given tasks.

II / 4. I designed a pipeline for 'cold' speech feature extraction based on Fisher vector encodings. I demonstrated that these types of features are capable of accurately modelling the speech of subjects having a cold.

Chapter 5

Deep Neural Network Embeddings

Here, we shall introduce deep neural network embeddings for pathological speech processing and paralinguistics tasks. More precisely, we will examine the use of the x-vector approach (a method originally intended for speaker recognition) as a feature extraction technique for audio-signals containing pathological speech. We experiment with the automatic screening of the following tasks: the degree of **sleepiness**, **depression**, the classification of **primate** sounds, and the levels of **escalation** in speech. Here, we will employ x-vectors to extract meaningful representations from the speech of subjects in the above-mentioned tasks. With the methodology outlined in [186], we will adopt the DNN architecture described there. We train the network from scratch employing distinct corpora and use its encoded embeddings for estimation. We show that x-vector features are able to produce high quality speaker models for tasks not related to speaker recognition.

5.1 The x-vector Method

The x-vector approach can be thought as of a neural network feature extraction technique that provides fixed-dimensional embeddings corresponding to variable-length utterances. Such a system can be viewed as a feed-forward Deep Neural Network (DNN) that computes such embeddings. Below, we will describe the architecture of the DNN (based on [186]) and the embeddings that are extracted using it.

5.1.1 DNN structure

Table 5.1 outlines the architecture of the DNN. The *frame-level* layers have a time-delay architecture [187], let's assume that t is the actual time step. Then, at the input, the frames are spliced together; namely, the input to the current layer is the spliced output of the previous layer (i.e., input to layer *frame3* is the spliced output of layer *frame2*, at frames $t-3$ and $t+3$). Next, the *stats pooling* layer gets the

T frame-level output from the last frame-level layer (*frame5*), aggregates over the input segment, and computes the mean and standard deviation. The mean and the standard deviation are concatenated and used as input for the next *segment* layers; from any of these layers the *x-vectors* embeddings can be extracted. And finally, the *softmax* output layer (which is discarded after training the DNN) [185, 186, 187].

Instead of predicting frames, the network is trained to predict speakers from variable-length utterances. Namely, it is trained to classify the N speakers present in the train set utilizing a multi-class cross entropy objective function (see Eq. 5.1). Let K be the number of speakers in N training segments. Then, the probability of the speaker k given T input frames $(x_1^{(n)}, x_2^{(n)}, ..., x_{1:T}^{(n)})$ is given by: $P(spkr_k|x_{1:T}^{(n)})$. If the speaker label for segment n is k, then the quantity of d_{nk} is 1, and 0 otherwise [185].

$$E = -\sum_{n=1}^{N}\sum_{k=1}^{K} d_{nk} \ln P(spkr_k|x_{1:T}^{(n)}). \tag{5.1}$$

5.1.2 Embeddings

The embeddings produced by the network described above capture information from the speakers over the whole audio-signal. Such embeddings are called *x-vectors* and they can be extracted from any *segment* layer; that is, either *segment6* or *segment7* layers (see Table 5.1). Normally, embeddings from the *segment6* layer give a better performance than those from *segment7* [186]. In this study, these type of representations can capture meaningful information from recordings of speakers suffering from Parkinson's, which may help to discriminate better the utterances; these characteristics are acquired at utterance level rather than at frame level. For both training the extractors and extracting the embeddings, we used the Kaldi Toolkit [155].

5.2 Excessive Daytime Sleepiness Detection

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) [97]. The National Sleep Foundation of the United States, in their Sleep in America Poll for 2020 [1], found that almost half of Americans report feeling sleepy between three and seven days per week. Thus, the detection and monitoring of sleepiness crucial for reducing the risks of having fatal accidents (e.g., when operating machinery or driving vehicles). Moreover, it may be beneficial for the early detection of specific

[1] https://www.sleepfoundation.org/wp-content/uploads/2020/03/SIA-2020-Q1-Report.pdf?x90960

Table 5.1: *DNN architecture of the x-vector system. It comprises five frame-level layers, a statistics pooling layer, two segment-layers and a final softmax layer as output. N represents the number of training speakers in the softmax layer. The DNN structure here is the same as that given in Snyder et al. [185].*

Layer	Layer context	Tot. context	In, Out
frame1	[t-2, t+2]	5	120, 512
frame2	{t-2, t, t+2}	9	1536, 512
frame3	{t-3, t, t+3}	15	1536, 512
frame4	{t}	15	512, 512
frame5	{t}	15	512, 1500
stats pooling	[0, T}	T	1500T, 3000
segment6	{0}	T	3000, 512
segment7	{0}	T	512, 512
softmax	{0}	T	512, N

neurological problems. Sleepiness is seen as a symptom caused by an underlying problem such as a neurological disease [136, 146] rather than a condition. Here, we propose a non-invasive way to monitor and control the degree of sleepiness by analyzing the speech of the subjects. This could be of help in automatic risk detection while driving and in similar scenarios.

5.2.1 SLEEP (Dusseldorf Sleepy Language) Corpus

This corpus was built by the Institute of Psychophysiology, Duesseldorf, and the Institute of Safety Technology, University of Wuppertal, Germany. The dataset comprises the recordings of 915 German speakers, 364 females, and 551 males, from 12 to 84 years of age, with a mean age of 27.6. The utterances were recorded with 44.1 kHz and downsampled to 16 kHz, using a quantisation of 16 bit. The audios were made in quiet rooms with similar acoustic conditions. The subjects were asked to read passages and carry out speaking tasks. Likewise, the subjects were asked to speak about, for example, their last weekend or to describe a picture; this resulted in spontaneous narrative speech. It contains 5564, 5328 and 5570 utterances, training, development and test sets, respectively; all three subsets contain recordings of just below six hours, leading to 17 hours and 35 mins of speech overall.

The degree of sleepiness of the subjects was assessed using the Karolinska Sleepiness Scale (KSS) [181]. Each subject reported their sleepiness level on the Karolinska Sleepiness Scale (KSS): from 1 (extremely alert) to 9 (very sleepy). At the same time, two observers assigned posthoc observer KSS ratings. The average of both

scores was the reference sleepiness value [177]. Later, this corpus was included in the Interspeech Computational Parainguistic Challenge in 2019 [177].

5.2.2 Related Works

The task, introduced by Schuller et al. [177], was already addressed by various early studies. For instance, in [66], a combination of Fisher Vectors, BoAW, and the ComParE functionals is carried out for their final CC scores (.383). (Note that this score was improved to .387 by training ensembles of classifiers [66].) In [211], the authors employ attention networks and adversarial augmentation, in the end, their best results (.369 of CC on test) are achieved by a fusion of neural network models. In [8], a .367 of CC was obtained by an early fusion of the learnt representations from attention and sequence to sequence autoencoders. Fisher Vector encodings were fused with the outputs of the ComParE Functionals in [207] to get a CC score of .365. In both [44] and [55], CNNs were exploited in an end-to-end deep learning approach: no fusion techniques are executed in the former study to get a .335 of CC; in the latter, a fusion of their CNN models was made to get a .325 of CC score.

5.3 Clinical Depression Screening

Noting James et. al [92], 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 in various aspects. Although it is a frequent and curable disease, estimating its occurrence is hard due to the specific clinical expertise needed [54]. The subject's speech is a biomarker containing information about a wide variety of traits (e.g., the mental status of the speaker).

According to a 2012 survey by the WHO, depression is the third most frequent mental disorder in the population [156]. The fact that there may be a connection between depression and speech was pointed out by Kraepelin [114], one of the founders of modern psychology. Early examinations dealt with the analysis of individual speech features, and reported a decrease in the mean and dynamics of pitch values, slower articulation tempo [6] along with monotonous and lifeless dynamics [28, 123].

5.3.1 Hungarian Depressed Speech Dataset (HDSDb)

In the Hungarian Depressed Speech Dataset [108], the degree of severity of depression was recorded using the Beck Depression Inventory II (BDI) scale [1]. The BDI-II scale ranges from 0 (healthy state), to 63 (severe condition). This scale uses the following rating: 0-13 healthy, 14-19 mild depression, 20-28 moderate depression,

29-63 severe depression. The corpus consists of 222 speakers, 116 patients suffering from depression (mean BDI for males and females are 24.9 ($\pm$7.4) and 27.8 ($\pm$9.2), respectively) and 106 healthy control speakers (mean BDI for males and females are 4.3 ($\pm$3.4) and 4.2 ($\pm$3.0), respectively) with a balanced age distribution.

The speakers were asked to read a short tale ('The North Wind and the Sun'). The patients were diagnosed with depression by the Psychiatric and Psychotherapeutic Clinic of Semmelweis University, Budapest. The recordings were sampled at 16 kHz and 16-bit.

5.3.2 Related Works

Various former studies investigated the possibility of assessing depression from speech. For instance, using CNNs for the enhancement of the detection of depression [86]; the analysis of gender and identity issues from the patients [122]; or feature extraction from the motor incoordination [206]. Prior studies made use of the HDSDb but with fewer samples, e.g.: CNNs and a speech correlation structure were used in [95] (accuracy of 84.1% with 188 samples). Also, the use of a special feature acoustic parameter selection approach in [109] (8.10 of RMSE with 127 samples). Both studies relied on Leave-One-Out Cross-Validation (LOOCV).

5.4 Escalation and Primates

As already stated in Chapter 4, Section 4.1.3, **Escalation** detection have real-life applications such as in public security, conversations in public places, and even human-computer interactions. Similarly, the automatic detection of **Primate** species by means of audio-signals can help to maintain the control and monitoring of biodiversity.

5.4.1 Escalation and Primates Corpora

More details about both the Escalation and the Primate corpora can be found in sections 4.4.3 and 4.4.4.

5.5 The Experiments

In experiments we trained x-vector DNN models (i.e., extractors) from scratch. For this, we computed different types of frame-level features from the audio-signals. We employed the *segment6* layer of the DNN to compute the 512-dimensional neural network embeddings (*x-vectors*).

5.5.1 Feature Extraction

We relied on Mel-Frequency Cepstral Coefficients as frame-level features in all of our experiments. For the **Sleepiness** trials, we extracted 20 cepstral coefficients; we used 23 for the **Depression**, **Escalation**, and **Primates** tasks. In each case we used a frame-length of 25ms and a window step size of 10ms.

Additionally, we 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 [131, 179]. 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 [13].

5.5.2 Training and Evaluation Methods

The classification and regression procedures were done using a Support Vector Machines algorithm with a linear kernel and, the C complexity parameter was set in the range 10^{-5}, ..., 10^{1}, for all the tasks in question.

Moreover, to experiment with the independence of the x-vectors from different recording and speaking conditions (e.g., language), as well as to deal with limited amounts of training data, we fitted the DNN extractors on another (larger) 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 [139]). This corpus was used for the experiments in all the tasks in question.

It is a standard practice to employ **data augmentation** when training x-vector DNNs in order to improve the noise robustness of the model. We applied this strategy in the **Sleepiness** and **Depression** tasks, respectively. From the original training data, two augmented versions were added. From additive noises and reverberation, two of the following types of augmentation were chosen randomly: babble, music, noise, and reverberation. The first three types correspond to adding or fitting noise to the original utterances. The fourth one involves a convolution of room impulse responses with the audio, i.e. reverberation. The reader can peruse [186] for more details about the augmentation strategies. Our goal here was to evaluate the contribution of the augmentation techniques to the overall performance scores. The augmentation process increased the BEA corpus to a total of 52,636 utterances (293 hours).

Additionally, we utilized the publicly available, pre-trained x-vector model described by Snyder et al. [186]. The model was fitted on English speech, specifically, employing a combination of a portion of Switchboard (SWBD) with a subset of the NIST SRE corpus. We aim to discover the differences amongst the DNN performances

when using corpora that differ in both duration and language from the original corpora.

Sleepiness

Besides the BEA corpus, we also **trained** different x-vector Deep Neural Network models (i.e., extractors) using the training and development sets of the SLEEP corpus combined (10,892 utterances, 11 hours and 39 mins). Also, we carried out **data augmentation** which increased that corpus training sets to 52,982 utterances (over 56 hours). This resulted in five x-vector training variations described above (*SLEEP Corpus train-dev*, *SLEEP Corpus train-dev (augmented)*, *BEA Corpus* and *BEA Corpus (augmented)*), and the *pre-trained x-vector DNN* model. The evaluation metric was the Spearman's Correlation Coefficient, which is typical for these types of tasks [177].

Depression

As the HDSDb corpus size is quite limited, we did not make use of it to **train** any extractor. Our training configurations resulted in the following types based on the corpora: BEA, BEA augmented, and the pre-trained x-vector model.

As for the **evaluation** approach, in contrast to former studies on the same corpus, and, seeking to avoid an optimistically-biased evaluation of the model, we chose speaker-wise Nested Cross-Validation. The metrics employed were the Pearson's Correlation Coefficient of the ground truth and predicted BDI scores of the subjects, along with the Root Mean Square Error (RMSE). Additionally, we evaluated our models from a binary class problem perspective. Thus, the subjects were automatically categorized as having depression or not by binarizing the labels based on their BDI values, where: if $BDI \geq 13.5$, the patient was cataloged as depressed; healthy control otherwise. This way, the class distribution resulted in 116 patients and 106 healthy controls. To this aim, we selected various metrics that provided a broader picture of the performance of the transformed predictions. As in most medical research, we used sensitivity and specificity, F1-score, along with Unweighted Average Recall (UAR, being the mean of specificity and sensitivity), and Area Under the Receiver Operating Characteristics Curve (AUC).

As an approach for the **baseline**, we opted for a former state-of-the-art speaker recognition method: the i-vector approach, which is known to capture speech, speaker and utterance meta information [34]. Akin to x-vectors, i-vectors also contain relevant information within the channel factor, which was used to classify emotion before [93]. Moreover, i-vectors have been successfully adapted to depression screening tasks giving good performances [2, 180]. Here, we trained the GMM-UBM model utilizing the same corpus (i.e., the BEA) that was employed for training the *first* x-vector

extractor. The GMM-UBM was fitted with 256 Gaussian components, which was used to extract i-vector representations from the HDSDb dataset.

Escalation and Primates

As mentioned in section 4.5.2, besides Fisher vectors, we also relied on **x-vectors** for these tasks. More specifically, we experimented with ensemble learning at extractor level. The basic principle of ensemble learning is to train several different, but similar machine learning models, and combine their outputs in some way. In this study we build an ensemble based on the x-vector feature extractors. That is, we propose training several x-vector neural network models on the same data, but each time using a different random seed during random DNN weight initialization. By calculating the embeddings for each of them, we get a number of different representations of the same training data. Although in theory concatenating these feature vectors and training only one classifier model might lead to a more robust performance than relying on any of the individual representations, we would end up with an unfeasibly large feature vector. Therefore we chose to train separate machine learning (e.g. SVM) models on these x-vector representations in the next step. To make the predictions more robust (and thus, making hyperparameter selection more reliable), we suggest simply averaging out the prediction scores got after evaluation in an unweighted manner.

More formally, we calculate the posterior estimate provided by the ensemble model as

$$P_e(c_i|X) = \frac{1}{m}\sum_{j=1}^{m} P_j(c_i|X) = \frac{1}{m}\sum_{j=1}^{m} P_j(c_i|H^j), \tag{5.2}$$

where c_i denotes the ith class ($1 \leq i \leq K$), X is the frame-level feature sequence of the actual utterance, H^j is the representation of X calculated by the jth x-vector extractor DNN, and the P_j value is the individual posterior estimate provided by the jth SVM model.

Since speaker ID is required to train x-vectors, and it was not available for either corpora, we trained our x-vector extractor DNN models using the SLEEP Corpus[177].The number of models in the ensemble (m) was set to 10. The evaluation metric utilized was Unweighted Average Recall (UAR).

5.6 Results and Discussion

5.6.1 Sleepiness

Table 5.2 outlines the Spearman's correlation coefficient scores got by the x-vectors embeddings. Overall, x-vector features extracted employing the SLEEP train-dev

Table 5.2: *Results of the experiments on the SLEEP Corpus given in Spearman's Correlation Coefficient. We show the results of former studies as well. The * means that the scores were achieved by a fusion of the best configurations. In contrast, the rest of the scores were obtained by applying a single approach. The x-vectors scores are given in accord with the corpus used to train the DNN they were extracted with.*

ComParE 2019 Features [177]	**Dev**	**Test**
ComParE Functionals	.251	.314
Bag-of-Audio-Words (BoAW)	.250	.304
AuDeep	.261	.310
Three-wise fusion*	—	.343
Former Studies		
Gosztolya* [66]	.367	.383
Yeh et al.* [211]	.373	.369
Amiriparian et al.* [8]	.320	.367
Wu et al.* [207]	.326	.365
Elsner et al. [44]	.290	.335
Fritsch et al.* [55]	.317	.325
DNN Embeddings (x-vectors)		
SLEEP Corpus train-dev (12h)	.303	**.365**
SLEEP Corpus train-dev (augmented)	.275	.324
BEA Corpus (60h)	.287	.313
BEA Corpus (augmented)	.256	.301
SWBD + SRE (pre-trained model, [186])	.300	.355

model gave better performances. These features achieved a .303 and a .365 of CC score on dev and test, respectively. However, using the augmented version of this model resulted in a decrease of the CC scores in both dev and test sets (.275 and .324). A similar situation occurred in the BEA Corpus model, namely, its augmented version led to a decrease in the CC scores. On dev, CC decreased slightly from .287 to .256; and from .313 (no augmentation) to .301 (augmented) on the test set. Although augmentation gives more diversity to the original data and attempts to make the models more robust. Here, the results indicate that the DNN was able to capture more meaningful information from the non-augmented versions than from their noise-robust counterparts. That is, adding noises and reverberation to this particular datasets could have caused the DNN to learn from non-relevant information, resulting in a poorer mapping (i.e., x-vector embeddings) for the specified task.

While the x-vector pre-trained model produced better results (.355 of CC on test),

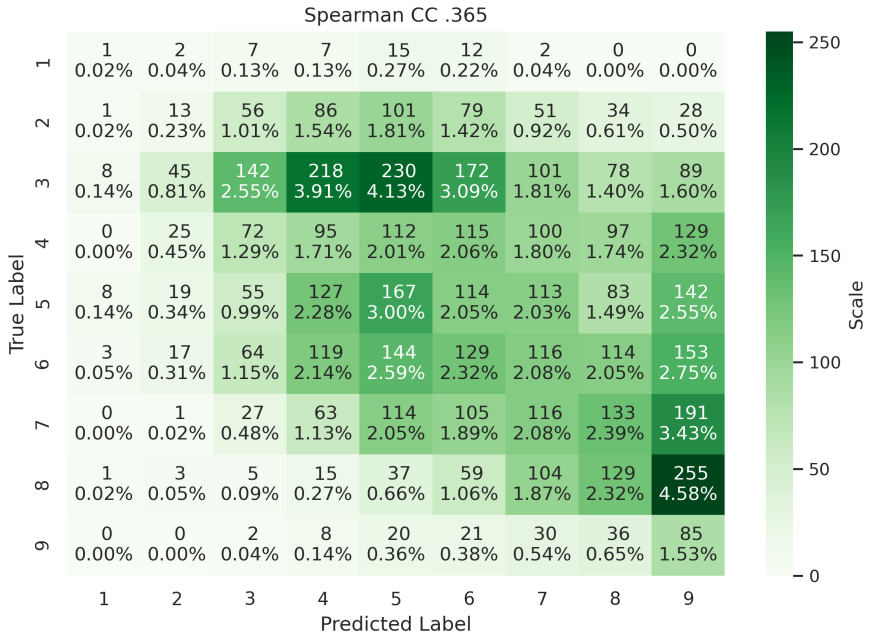


Figure 5.1: *Confusion matrix for the best results on the test set on the Sleepiness task.*

its performance could not reach that of the SLEEP train-dev extractor. This could be attributed to a language-dependant situation (i.e., the pretrained model was fitted using English corpora). It appears that, in this case, the model trained with in-domain data (i.e. using the SLEEP corpus) was able to generate better representations than the pre-trained model that was trained with huge data amounts of different domain data.

In Table 5.2 we also compare our performance scores with those of previous studies and official baselines on the same task. It can be seen that the proposed DNN embeddings were capable of outperforming all the baseline scores of the Interspeech 2019 ComParE Challenge [177]. Moreover, it is evident that most of the former studies achieved their best results by relying on a *fusion* of the scores. Actually, in [66], a combination of Fisher Vectors, BoAW, and the ComParE functionals is carried out for their final CC scores (.383). (Note that this score was improved to .387 by training ensembles of classifiers [66].) In [211], the authors employ attention networks and adversarial augmentation, in the end, their best results (CC of .369 on test) are achieved by a fusion of neural network models.

In [8], a CC of .367 was obtained by an early fusion of the learnt representations from attention and sequence to sequence autoencoders. Fisher Vector encodings were fused with the outputs of the ComParE Functionals in [207] to get a .365 of CC. In both [44] and [55], CNNs were exploited in an end-to-end deep learning approach: no fusion techniques are executed in the former study to get a .335 of CC; in

Table 5.3: *Results of the experiments for the **Depression** task using all the feature dimensions. Each row corresponds to a different x-vector extractor in function of the data used to train it.*

	Regression		Classification				
	Pearson's	**RMSE**	**UAR**	**SPEC**	**SENS**	**AUC**	**F1**
i-vector baseline	.608	10.45	80.44	89.39	82.00	0.920	89.36
BEA (FBANK)	.625	10.26	88.65	86.79	90.51	0.920	89.36
BEA (MFCC)	.615	10.36	82.64	77.35	87.93	0.904	84.29
BEA-aug. (FBANK)	**.684**	**9.54**	**89.00**	**84.90**	**93.10**	**0.940**	**90.00**
BEA-aug. (MFCC)	.635	10.16	80.75	73.58	87.93	0.908	82.92
Pre-trained [186]	.675	9.64	82.99	75.47	90.51	0.935	85.02

the latter, a fusion of their CNN models was made to get a .325 of CC score. However, in our study, x-vector representations are still competitive and even outperform some of the former studies without the need for any kind of fusion strategy.

Fig. 5.1 displays the confusion matrix of our best configuration. The figure tells us that categories $3, 5, 6, 7, 8$ had similarly high accuracies. This means that the model was capable of distinguishing a large variety of categories including one of the extreme labels (8), the slightly extreme classes (3 and 7), as well as the middle categories (5 and 6). As for the extreme labels 1, 2 and 9, the scores are much lower. Perhaps this is due to the number of samples for these classes: these three categories represent approximately 13% of the number of samples in the dataset. Moreover, it seems that the model tends to overestimate the sleepiness level of the speaker, as we got higher values mostly above the main diagonal.

5.6.2 Depression

Table 5.3 presents the results of our experiments along with the i-vector baseline, which was surpassed by the methods based on x-vectors. In general, the DNN embeddings could model better speaker traits for depression screening than the i-vectors. The augmented extractor produced better embeddings than their non-augmented counterparts, and demonstrated the effectiveness of data-augmentation when using x-vectors on this specific corpus. In particular, the extractor trained with the BEA-augmented corpus (with FBANKs) gave the highest Pearson's CC: .684, and the lowest RMSE: 9.54. As for the *binary* classification evaluation, the same configuration gave the highest scores: a UAR of 89.00%, an specificity of 84.90%, a sensitivity of 93.10%, a AUC-score of 0.940, and an F1-score of 90. We got quite a low number

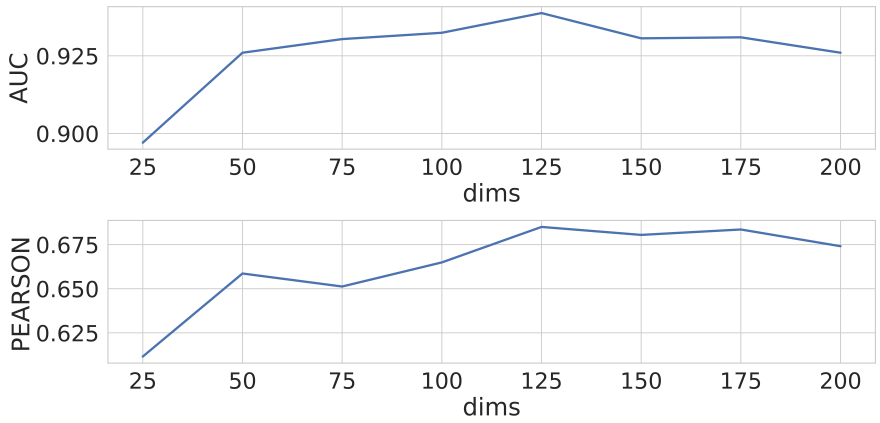


Figure 5.2: *CC and AUC scores of the feature selection process from the **BEA-augmented Extractor (FBANK)** on the Depression task.*

of false negative and false positive cases, which indicates the potential feasibility of the model for screening. Moreover, the AUC-score value suggests a considerably high discriminating ability of the model.

The extractors fitted with log-energies (except for the non-augmented version) outperformed their cepstra counterparts in every case. This may be due to the fact that MFCCs attempt to eliminate unimportant variations for recognition, and lead to a reduction in the input-signal dimension (less information). Meanwhile, the FBANKs contain a more integral representation as they produce a less pre-processed input-signal with a larger set of filter-bank coefficients (more information); it appears that DNNs are able to better exploit these type of representations. The embeddings from the *pre-trained model (MFCCs)*, however, achieved better scores than the *BEA Extractor (FBANKs)* configuration. Although our custom extractors used in-domain language data, a possible reason for this might be the significant difference between their corresponding amounts of training corpora.

The *Pre-trained Model [186]*, although competitive, could not surpass the results of the best custom model, we got a lower CC (.675). The results may confirm an existing *data-domain* dependence of the x-vector architecture. More specifically, we experimented with models that learned from data closer to the actual task domain (language-related in this specific case), and they produced better quality representations than the pre-trained model did.

Correlation-Based Feature Selection

In practice, the x-vector features will have a bigger number of dimensions than the total number of samples of the dataset. Eventually, this might lead to a decay in the performance due to the regularization bias growth towards the training data. Hence,

Table 5.4: *Results of the experiments using the **correlation-based feature selection** in the Depression task. The best feature selection configurations are presented only. N denotes the number of features from the automatic feature selection process. Each row corresponds to a different x-vector extractor in function of the data used to train it. **CC** stands for Pearson's Correlation Coefficient.*

	Regression		**Classification**					
	CC	**RMSE**	**UAR**	**SPEC**	**SENS**	**AUC**	**F1**	**N**
BEA (FBANK)	.632	10.14	87.19	83.01	91.34	0.915	88.33	150
BEA (MFCC)	.586	10.59	78.39	68.86	87.93	0.891	81.27	175
BEA-aug. (FBANK)	**.685**	**9.50**	**88.61**	**85.84**	**91.37**	**0.938**	**89.45**	125
BEA-aug. (MFCC)	.603	10.41	81.11	71.69	90.51	0.914	83.66	175
Pre-trained [186]	.672	9.70	83.89	76.41	**91.37**	0.934	85.82	200

before training, we carried out an automatic feature selection, seeking to reduce the number of features. More precisely, we computed the CC for each feature-column with respect to the BDI labels; from these values, we selected those that had the highest CC scores. The final selection of the number of dimensions (N) was based on a step size of 25 (i.e., $N = 25, 50, \ldots, 200$ selected dimensions). The procedure was carried out within the speaker-wise Nested-CV to avoid peaking. Consequently, besides dimensionality reduction, it also meant that we just had relevant features (those that contribute the most to the final predictions), and thus speeded up the BDI estimation step.

The results of this approach are given in Table 5.4. Similar to the previous experiments, the augmented extractors fitted with FBANKs also outperformed the rest of the configurations in this case. Moreover, the CC increased slightly to .685, while the RMSE decreased to 9.50. These results were achieved just using 125 of the 512 available original features after the feature selection process. Also, the classification metrics for the same configuration changed slightly: while the specificity score experienced an increment, the sensitivity, AUC, UAR, and F1 scores only decreased a small amount. Again, FBANKs features gave more efficient performances based on the number of selected features. Fewer dimensions were needed for the model to provide a better generalization; that is, FBANK-based embeddings actually contained more meaningful information than those from MFCCs.

Figure 5.2 depicts the AUC and the Pearson's CC scores obtained using the different N feature selection values for the corresponding dimension size. The line plots display a tendency where the CC values increase as the number of dimensions increment as well, and they both start to decrease after dimension 150. In general,

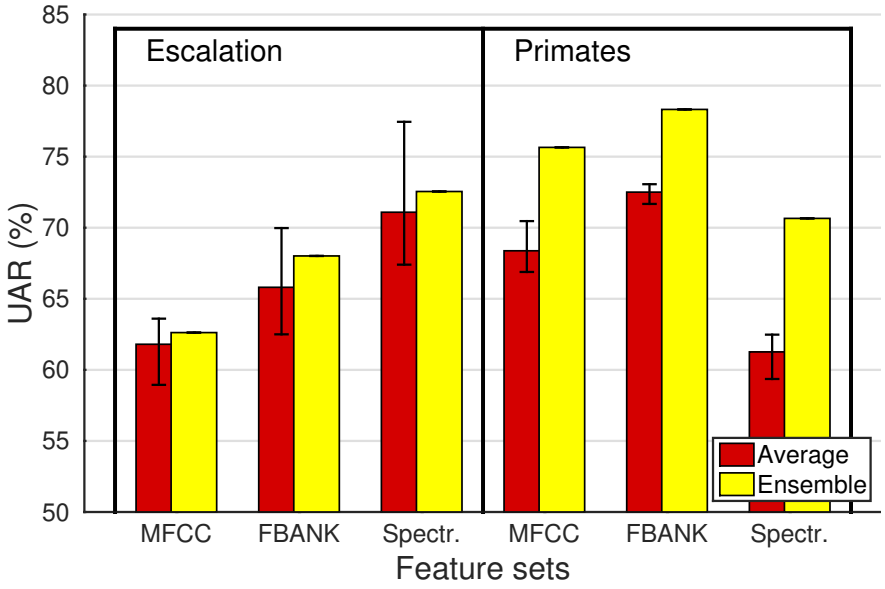


Figure 5.3: *The UAR scores of the individual and the ensemble x-vector approaches obtained on the development set; the error bars indicate minimum and maximum values. Escalation and Primates tasks.*

both metrics suggest quite similar trends over the number of dimensions. Overall, the correlation-based feature selection, besides discarding irrelevant information and helping to reduce the computation times, also helped to increase the CC and reduce the RMSE in most of the cases. Furthermore, all the configurations necessitated only less than the half of the original number of dimensions and produced better or competitive results.

5.6.3 Escalation and Primates

Although the results and analysis for the first stage of the experiments with these tasks was already shown in section 4.5.3, here we will adopt the ensemble x-vector approach to be consistent with the current chapter of the book. The reader may return to the this section if necessary. Moreover, we describe another experiment conducted on the **Escalation** corpus.

Fig. 5.3 lists the results obtained on the development sets with the individual x-vectors and the ensemble x-vectors for both Escalation and Primates. Notice that the ensemble approach always outperformed the average of the individual models. In the Escalation task, there was a large difference (4-10%) between the performance of the best model and the worst model, probably because of the limited amount of data. For the Primates sub-challenge, this variance was smaller (although still significant: between 1.4% and 3.6%); however, in this case, the ensemble model outperformed

Table 5.5: *The results obtained for the Escalation Sub-Challenge with the SSPNet Conflict Corpus-based approaches*

Feature Set	Dev	Test
ComParE functionals	72.8%	—
SSPNet Conflict Corpus-based	62.6%	—
ComParE + SSPNet Conflict	73.8%	62.4%
ComParE + x-vectors + FV + SSPNet	79.8%	63.9%
Official ComParE baseline	—	59.8%

even the best of the 10 individual x-vector models. This, in our opinion, confirms that ensemble x-vectors is a viable approach.

Table 5.5 shows the UAR values obtained via the proposed approach. Although the UAR score of 62.8% on the development set might seem low compared to the ComParE functionals case, for a 3-class (and cross-corpus, as the test set of the Escalation sub-challenge is comes a different dataset than its training and development sets) task it is realistic, as it significantly exceeds the 33.3% value achievable via random guessing. Furthermore, we did not want to utilize this approach on its own, but we sought to use it to aid the other classification methods; and by combination, we achieved an UAR value of 73.8%. On the test set we attained 62.4% with this approach, which was improved to 63.9% by combining all four methods. Both values exceed the official baseline of the Escalation sub-challenge, which, in our opinion, demonstrates the usefulness of this cross-corpus method.

5.7 Concluding Remarks

5.7.1 Sleepiness

We experimented with five different DNN models to map utterances to fixed-sized representations (i.e., x-vectors). The SLEEP and BEA corpora were employed to fit two different DNN models using augmented data, and two with no augmentation. The fifth model used was the pre-trained DNN from [186]. 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 particular, 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 performance scores of various previous works, also produced the highest Spearman's CC score by a standalone (single) method for this particular task.

5.7.2 Depression

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.

5.7.3 Escalation and Primates

In our experimental setup, we built an ensemble x-vector classifier by training 10 independent x-vector extractor neural networks on the same data. The ensemble was constructed aiming to improve 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. As stated in Chapter 4 Section 4.6, a more traditional feature extractor (i.e., the Fisher vector) was even more successful. Our last technique, which used the SSPNet Conflict Corpus in the Escalation sub-challenge, also led to promising UAR values. Overall we outperformed the official baselines from [176] for both tasks, which supports the efficacy of the applied techniques.

The author of this PhD thesis is responsible for the following contributions presented in this chapter:

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, originally intended for speaker verification, were 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 was capable of surpassing the performance scores of those that rely on more elaborate techniques like ensemble machine learning and classifier combination.

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

Chapter 6

Automatic Speech Recognition Methods

This chapter introduces the use of different speech analysis and ASR-based techniques for the automatic screening of Alzheimer's and Mild Cognitive Impairment via the speech of subjects. First, we present a method that extracts a set of **temporal parameters** that characterize hesitation in the spontaneous speech of patients suffering from MCI in a cross-lingual environment. These attributes are computed based on a ASR framework. It should be added that the study concerning **temporal parameters** is not the main contribution by the author of this thesis.

On the other hand, and similar to temporal speech parameters, we demonstrate another feature set that can be used to describe the amount of hesitation present in the speech, with the difference being that these do not rely on any ASR system; we call these **posterior thresholding hesitation representations**.

6.1 Introduction

Dementia is a chronic or progressive clinical syndrome, affecting mainly elderly people worldwide. It is characterized by the deterioration of memory, language and problem-solving skills, which are severe enough to adversely affect the patients' ability to carry out everyday activities [7]. According to the estimates, the number of affected individuals, which at present exceeds 46.8 million, may double by 2050 [157].

The most widely used term to describe the preclinical stage of dementia is Mild Cognitive Impairment (MCI), which condition is often considered to be the borderline between normal aging and dementia [152]. This syndrome shows similar characteristics to dementia, although in the case of MCI the symptoms do not interfere with the patients' activity of daily living [48]. However, given its high conversion rate to dementia [2-31% annually, see e.g. 19], MCI should be regarded as a severe condition. As the transition phase from MCI to dementia can last even 15 years [116],

there is a wide time window in which the subtle signs of cognitive decline could be detected. Since the timely identification of MCI could provide more effective therapeutic interventions to delay progression, the importance of developing methods that allow early recognition has been emphasized in the recent years.

These progressive types of MCI are most of the time precursor conditions to Alzheimer's disease (AD), but they can be also due to vascular or other neurodegenerative diseases [151]. Changes in language performance can act as an early and valuable indicator of MCI or Alzheimer's, since language-related alterations can appear before the manifestation of other distinctive cognitive symptoms [127].

It has been shown that changes in language production are associated with subclinical declines in memory, e.g. the fluency of spontaneous speech has been proven to deteriorate in people with early MCI [135]. During the course of the disease, filled pauses (i.e. vocalizations like 'uhm' and 'er') and disfluencies become evermore frequent in the subject's speech [33], corresponding to the word-finding or word-retrieval difficulties of patients [190]. Earlier studies also indicated that compared to healthy controls, MCI patients tend to have lower speech rate, and an increased number and length of hesitations [190]. The above-mentioned characteristics can strongly influence the overall time course of the speech; therefore, the analysis of such temporal aspects can help us explore the relationships between language and memory.

6.2 Related Works

In the last decade, numerous attempts have been made to distinguish cognitively healthy control (HC) subjects from people with MCI or with Alzheimer's disease (AD) using different speech analysis techniques. In the earlier studies, analyzed speech features were extracted mainly from manually transcribed data, which is rather labor-intensive. In more recent studies the goal was to find out whether extraction by automated techniques could produce similar results. In the past few years, several such automatic speech analysis studies have been published [e.g. 33, 51, 73, 111, 112, 116, 184, 193, 194, 197].

There exist previous studies that developed a set of temporal speech parameters which characterize the hesitation contained in the spontaneous speech of the subjects [72, 73, 84, 196, 197]. Hesitation is defined as an absence of speech. It can be divided into two categories: silent pauses and filled pauses. Measuring the amount of silent pauses in human speech is quite common [see e.g. 3, 50, 89, 126, 184]. The attribute set developed by our team, besides silent pauses, also summarizes the amount of *filled* pauses (i.e. vocalizations such as 'er', 'umm' etc.) in the speech of the subject in the temporal attribute set. This set of temporal attributes (the Speech Gap Test or S-GAP test) can be calculated by using speech processing tools, i.e. by

(1) **Articulation rate** was calculated as the number of phones per second during speech (excluding hesitations).

(2) **Speech tempo** (phones per second) was calculated as the number of phones per second divided by the total duration of the utterance.

(3) **Duration of utterance**, given in seconds.

---

(4) **Pause occurrence rate** was calculated by dividing the number of pause occurrences by the number of phones in the utterance.

(5) **Pause duration rate** was calculated by dividing the total duration of pauses by the length of the utterance.

(6) **Pause frequency** was calculated by dividing the number of pause occurrences by the length of the utterance.

(7) **Average pause duration** was calculated by dividing the total duration of pauses by the number of pauses.

Table 6.1: *The examined temporal speech parameters, based on our previous studies [85, 197].*

relying on a phone-level ASR framework.

6.3 Temporal Speech Parameters

To investigate the spontaneous speech of MCI patients and HC subjects, we calculated specific temporal parameters from their spontaneous speech.

For this purpose, we employ an ASR system trained to recognize phones in the utterances, where the phone set included the special non-verbal labels listed above (i.e. filled pauses, coughs, breath intakes etc.).

This set of temporal parameters can be seen in Table 6.1. The articulation rate and speech tempo (i.e. parameters (1) and (2)) both describe how fast the subject speaks (although in a slightly different manner), while the duration of the utterance (parameter (3)) is related to the amount the subject could remember about his / her previous day. The remaining parameters ((4)–(7)) all describe the amount of hesitation in the spontaneous speech of the subject by focusing on the number or on the duration of pauses in some way. We defined hesitation as the absence of speech for at least 30ms; we distinguished two sub-types of hesitation: silent pauses and filled pauses (i.e., vocalizations such as ‘er’, ‘umm’ etc.).

6.4 Posterior-Thresholding Hesitation Representation

Similar to the approach from Section 6.3, here we also focus on measuring the amount of hesitation (i.e. silent and filled pauses) in the spontaneous speech of the subjects. However, we design our pipeline dispensing with the ASR system.

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. Next, we will describe these steps in a more detailed manner. Please see Fig 6.1 for the architecture of the proposed approach.

6.4.1 Frame-level DNN Evaluation

In hybrid HMM/DNN ASR systems the role of the Deep Neural Network component is to estimate the likelihood of the Hidden Markov model states for each frame of the speech signal (typically at 100 frames/sec). It is then the task of the HMM component to perform the sentence-level decoding by combining these local, frame-level estimates. The first stage of our approach corresponds to evaluating this DNN acoustic model on the utterances of the subjects. For this, we have only one special requirement: this DNN must be trained on an audio corpus that contains occurrences of filled pauses both in the audio and in the transcription. This is so because our approach focuses on both pause types, and while it is common to have (and annotate) silent pauses, several ASR corpora do not contain filled pauses (or their occurrences are just not marked), because it is not a requirement of a standard ASR system to locate such vocalizations and include them in its output (i.e. in the automatic transcription).

The result of this step is the sequence of frame-level posterior estimate vectors of all the phonetic states of the ASR system.

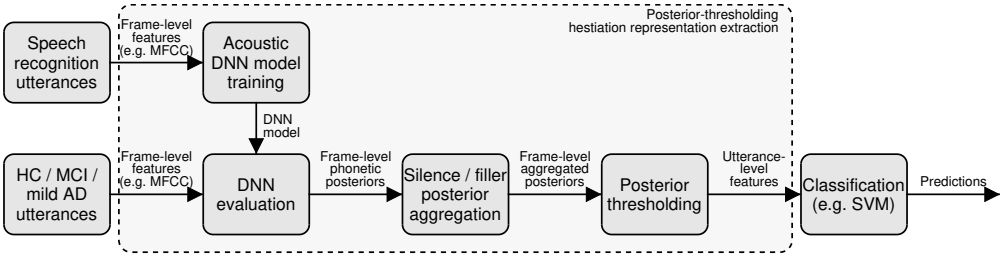


Figure 6.1: *The general workflow of the applied DNN-based feature extraction process.*

Algorithm 1 Posterior-Thresholding Feature Extraction

Require: N: the number of frames in the utterance
Require: $likelihoods$: the frame-level aggregated posterior estimates (with a length of N)
Require: s: the step size ($s < 1$)
  $m := \lfloor 1/s \rfloor$
  for i := $1 \rightarrow m$ **do**
    $cnt := 0$
    $th := i \cdot s$
    for j := $1 \rightarrow N$ **do**
      if $likelihoods(j) \geq th$ **then**
        $cnt := cnt + 1$
      end if
    end for
    $features(i) := cnt/N$
  end forreturn $features$

6.4.2 Hesitation Posterior Estimation

The states of the HMM system are related to the phone set of the given language, but usually there is no direct one-to-one correspondence, as the states typically represent a finer resolution. First, we model several acoustic phenomena like filled pauses, noises, breathing, gasps and coughs by assigning special models to them. Second, the phones are traditionally divided into three production states, as it is known to improve recognition performance. Third, instead of working with such simple, context-independent (CI) phone labels, even better speech recognition results can be achieved by context-dependent (CD) modelling [82], where the phonetic labeling also takes the (left and right) neighbors of the actual phone into consideration. As in this HMM/DNN hybrid model, the role of the DNN acoustic model is to estimate the local (i.e. frame-level) posteriors of the HMM states, the number of the DNN outputs should be the same as the number of HMM states. Therefore, to obtain the frame-level posterior estimates of the silent or the filled pauses, we have to add up

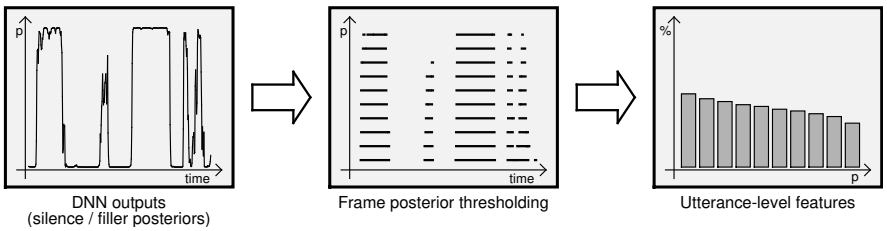


Figure 6.2: *The schema of the posterior-thresholding feature extraction step.*

the likelihoods of all the phonetic states which correspond to silence and to filler events for each frame. In the second step of our feature extraction approach, this is what is done. Therefore, the result of this step is a sequence of the (aggregated) frame-level posterior estimates of silence and filler events.

6.4.3 Posterior-Based Utterance-Level Feature Extraction

Even though, at this point, we have the posterior estimates of silence and hesitation, we cannot utilize them directly in the classification step. The reason for this is that these posterior estimates are present at the frame level; therefore, the size of their vector is proportional to the length of the given utterance. However, for utterance-level classification we need a fixed-size representation. This last step of the proposed feature extraction method provides a way to fill this gap; that is, to describe the frame-level posterior sequence for the whole utterance in a fixed-size form.

More specifically, for a given threshold value $0 \leq th \leq 1$, we count the number of frames where the corresponding posterior estimate is greater than or equal to th. Since the number of such frames is also affected by the duration of the utterance, we divide this sum by the total number of frames (i.e. we normalize them). This value will be used as a newly extracted feature. To adequately describe the posterior sequence, this process is repeated for the values $s, 2 \cdot s, 3 \cdot s, \ldots, 1$ as the th threshold, where s is a step size parameter of the method. The reader should take a look at Algorithm 1 to see the pseudo-code of our approach; furthermore, Fig. 6.2 illustrates the mechanism of this step.

Note that extracting the posterior thresholding feature set is equivalent to calculating the *cumulative histogram* [166] of the frame-level posterior estimates. These types of histograms were employed in former ASR techniques [132] as well as in numerous other tasks like texture classification [83], handwritten character recognition [81], analog-to-digital converter testing [4] and in computational paralinguistics [65]. Our motivation for employing this feature representation is that, this way, we can describe the distribution of the posterior estimates of the whole utterance in finer details with a fixed-size vector.

6.5 The Hungarian MCI-AD Corpus

The utterances were recorded at the Memory Clinic at the Department of Psychiatry of the University of Szeged, Hungary. The study, conducted in accordance with the Declaration of Helsinki, was approved by the Regional Human Biomedical Research Ethics Committee of the University of Szeged. The recordings were collected from three categories of subjects: those suffering from MCI, those affected by early-stage AD (mild AD or mAD), and those having no cognitive impairment at the time of recording (i.e. healthy controls, HC). All the participants signed a consent prior the recording phase. The exclusion criteria were drugs or alcohol consumption, being under pharmacological treatment affecting cognitive functions, and visual or auditory deficits. Anyone who had previously suffered from head injuries, depression or psychosis was also excluded.

MCI and mAD patients were selected after a medical diagnosis. Diagnosis was based on the consensus of a clinical expert panel consisting of a psychiatrist, a neurologist and a psychologist, who reviewed neuroimaging scans (CT, MRI) when available, and also the results of three cognitive screenings tests: the Mini-Mental State Examination [MMSE 45], the Clock Drawing Test [CDT 52] and the Alzheimer's Disease Assessment Scale – Cognitive Subscale [ADAS-Cog 162]. In the case of MCI, Petersen's criteria [153], while for AD, internationally used guidelines [128] were followed. The possibility of depression was assessed using the 15-item version of the Geriatric Depression Scale [GDS 212]: participants scoring above 10 on the test were excluded from the study.

Several studies found that MCI and AD affect the *spontaneous* speech of the subjects more than their planned speech [see e.g. 161, 164, 191]). Therefore, we decided to record spontaneous speech as well. After the presentation of a specially designed one-minute-long animated film, the subjects were asked to talk about the events seen in the film (*immediate recall*). Afterwards, the subjects were asked to talk about their previous day (*previous day*). In the last task, the subjects watched a second film, and they were asked to talk about this film after a one-minute break (*delayed recall*). Details about the actual instructions given to the patients are shown in Table 6.2. For more details about our experimental setup for recording, see the study of Hoffmann et al. [84]. Unfortunately, our ethical agreement does not allow the sharing of these speech recordings. Each recording was edited; namely, parts before the subject started to speak and after his last phoneme uttered were manually removed. Hence, we had three recordings for each subject, each containing spontaneous speech with a different speaker task. In a real application scenario (e.g. within a mobile phone application), this step could be automated at the time of recording; for example by using a specific sound (e.g. beep) to mark the start of the recording, and apply voice activity detection with a larger time threshold to detect the exact end

(1) "*I am going to show you a silent movie lasting about a minute. Try to remember the story, the actors, the objects and the places, paying attention to the details.*"

(2) "*Please tell me about your previous day in as much detail as you can.*

(3) "*Now, I am going to show you another clip. Try to remember the story, the actors, the objects and the places, paying attention to the details. OK, I am going to start it now.*"
The Patient watches the clip. If he starts talking about it, he is reminded that he is not yet allowed to talk about it. When the clip ends:
"*Now we will take a one-minute break.*"
If the Patient starts talking during the break, he is reminded that it is still break-time, and he has to wait until the minute is over. After the one-minute break is over:
"*Right, could you please tell me what you saw in the clip?*"

Table 6.2: *The instructions to the patients when recording the three utterances.*

of the response of the subject.

This corpus comprises recordings taken from more than 150 subjects. Due to technical issues like poor sound quality and controversial diagnosis (i.e. when our clinical expert panel could not reach a consensus), some subjects were filtered out. Furthermore, we insisted on performing our experiments on data where the demographic properties of the speaker groups did not differ significantly, which also reduced the number of subjects. Therefore, in the end we used the recordings of 25 speakers for each speaker group, resulting in a total of 75 speakers and 225 recordings. Although at first glance this number might seem low, having 75 subjects is considered significant in this area, as most studies involve fewer than 200 subjects [see e.g.: 119, 147, 149, 164, 204].

To ensure that there were no statistically significant differences among the speaker groups in their age, gender and education, we applied one-way ANOVA, Kruskal-Wallis H test (when the normality assumption was violated) or Chi-squared test (for categorical values).

6.6 Posterior-Thresholding Hesitation Representation: The Experiments

Our Deep Neural Network acoustic models were trained on a subset of the BEA Hungarian corpus [139]; we trained the DNN on the speech of 116 subjects (44 hours of recordings overall in 9.7k recordings (mean duration of 16.4s, median duration of 13.3s)). We made sure that the annotation suited our needs, i.e. filled pauses, breathing sounds, laughter, coughs and gasps were marked in a consistent manner. The minimum duration of both silent and filled pauses were 30ms in the annotation of this corpus; mean durations were 535ms and 234ms, while median durations were 410ms and 180ms, silent and filled pauses, respectively.

Although context-dependent models have been shown to achieve better performance in ASR in terms of Word Error Rate (WER) than their simpler context-independent counterparts, for our Posterior-Thresholding Hesitation Representation approach we only need to distinguish silent and filled pauses from everything else. We wanted to find out whether this could be solved at the same (or a very similar) level of performance with simple CI phone states as with the more complex CD ones. To ascertain whether there is a difference in subject classification performance, we experimented both with context-dependent and context-independent phonetic mappings.

We used a quite traditional DNN structure in our acoustic model: we utilized 40 Mel-frequency filter banks along with raw energy as frame-level features, and included the first- and second-order derivatives (i.e. the Δ and $\Delta\Delta$ values). To improve model accuracy, our model used a sliding window with a width of 15 frames (1845 frame-level features overall). Following this, we utilized 5 hidden layers, each consisting of 1024 ReLU neurons. Lastly, we included a softmax layer that had as many neurons as the number of states. Since we had 57 phones (including silence and filled pauses as special 'phones'), the Context-Independent DNN acoustic model had 171 output neurons. In the Context-Dependent case, we employed the standard tree-based clustering method for state tying [142]; the criterion during state tying was a Kullback-Leibler divergence-based one [71], leading to 911 tied states.

6.6.1 Feature Extraction

To extract the Posterior-Thresholding Hesitation Representation, we employed a step size s of 0.02, hence we had 50 features for each hesitation type. We experimented with using the silent pauses as input (treating gasps, breath intakes and sighs also as silent pauses) as well as using the filled pauses (treated as a special phone). Furthermore, we experimented with a setup where all HMM states were considered which corresponded to either the silent or the filled pauses during the posterior summing step (i.e. step (2) of the feature extraction process), which practically means that we

measure the amount of all pauses; we will refer to this case as 'all hesitation'.

These feature sets were extended with one further feature: the duration of the utterance. When calculating duration, we first omitted the beginning and ending frames where the likelihood of silent pause exceeded 0.9.

6.6.2 Utterance-level Classification

As is common in medical speech processing tasks, we relied on the Support Vector Machine (SVM) algorithm for the classification phase; we applied the libSVM implementation [23] with a linear kernel. Since we had a relatively low number of examples (75, as each subject corresponded to only one example), we employed a 25-fold cross-validation, where each fold consisted of 1 HC, 1 MCI and 1 mAD subject. Therefore, each classifier model was trained on the speech of 72 subjects. The C complexity parameter was set in the range $10^{-5}, 10^{-4}, \ldots, 10^{2}$.

The complexity C meta-parameter of the SVM was set by nested cross-validation [22]. Each time we trained on the data of 72 (i.e., 3×24) subjects, we performed *another* (24-fold) cross-validation process, looking for the C meta-parameter value that led to the highest AUC score. After this, we trained an SVM model with the selected meta-parameters on the data of all 72 speakers, and this model was evaluated on the remaining speaker. This way we ensured that we avoided any form of peeking, which would have created a bias in our scores, had we used standard cross-validation.

6.6.3 Prediction Combination

Besides training an SVM classifier for the silence-related and filler-related feature sets, we were also interested in what could be achieved with a combination of two or more attribute sets. To do this, we combined our predictions obtained from the previous classification experiments. Following our previous studies [see e.g. 65, 73], we decided to take the weighted mean of the posterior probability estimates produced by the individual classifier models, which we found to be a simple-yet-robust technique. This combination allowed us to measure the classification performance for all three speaker tasks (i.e. immediate recall, previous day and delayed recall) and/or all three feature subsets (i.e. silent pauses, filled pauses and all hesitation) as well.

6.6.4 Evaluation

We evaluated our models by utilizing the Area Under the Receiver Operating Characteristics Curve (AUC) score. This statistic is widely employed for summarizing the performance of automatic classification systems in medical applications. In our experiments, we computed the AUC score for all three speaker categories (i.e., for

Table 6.3: *The various accuracy scores obtained with the S-GAP temporal speech parameters, following the approach of Tóth et al. [197] and Gosztolya et al. [73]. (Acc. = classification accuracy, HC = Healthy Control, MCI = Mild Cognitive Impairment, mAD = Mild Alzheimer's Disease). On the task column: IR = Immediate Recall, PD = Previous Day, DR = Delayed Recall, All-3 = Delayed Recall*

		Classification		Area-Under-Curve			
Feat.	**Task**	**Acc.**	F_1	**HC**	**MCI**	**mAD**	**Mean**
Silence	IR	38.7%	66.7	0.637	0.474	0.705	0.605
	PD	41.3%	59.1	0.502	0.646	0.562	0.570
	DR	41.3%	59.1	0.558	0.536	0.774	0.623
	All-3	42.7%	68.9	0.619	0.374	0.689	0.561
All	IR	40.0%	62.9	0.580	0.566	0.743	0.630
	PD	42.7%	64.4	0.622	0.611	0.569	0.601
	DR	50.7%	68.9	0.673	0.592	0.805	0.690
	All-3	60.0%	77.4	0.728	0.600	0.780	0.705

healthy controls, for MCI and for mAD speakers), and we also report the mean of the three AUC scores. Since our dataset had a balanced class distribution, we also made use of the traditional classification accuracy score. Likewise, Information Retrieval metrics such as precision and recall scores were also added to our metrics. Moreover, the harmonic mean of these two (precision and recall), that is, F_1-score, was also employed. In these cases we combined the MCI and mAD speaker categories to form the positive class, while the HC category was treated as the negative one. Lastly, we report the specificity value as well. These metrics were calculated by setting the decision threshold along with the Equal Error Rate (EER).

6.7 Results and Discussion

6.7.1 Results Using the Temporal Speech Parameters (S-GAP)

Table 6.3 shows the accuracy metrics obtained by using our temporal speech parameters developed in our previous investigations. Although the scores do not seem high, recall that we treated this task as a three-class classification one, therefore random guessing would lead to a classification accuracy of 33.3%, AUC scores of 0.500, etc. Otherwise, there was no significant difference between the three speaker tasks: classification accuracy fell in the range $40.0\% \ldots 50.7\%$, precision in the range $71.8\% \ldots 77.5\%$, while recall and specificity lay between 56% and 64%. (Of course, the latter two metrics were very similar, because we presented our results using Equal Error Rate.) Due to these values, F_1 was around $63 - 69$, while the mean AUC fell

Table 6.4: *The various accuracy scores obtained with the Posterior-Thresholding Hesitation Representation using Context-Dependent states. (Acc. = classification accuracy, Prec. = precision, Spec. = specificity; HC = Healthy Control, MCI = Mild Cognitive Impairment, mAD = Mild Alzheimer's Disease). On the task column: IR = Immediate Recall, PD = Previous Day, DR = Delayed Recall, All-3 = Delayed Recall*

		Classification		**Area-Under-Curve**			
Task	**Feat.**	**Acc.**	F_1	**HC**	**MCI**	**mAD**	**Mean**
IR	Sil.	52.0%	68.9	0.587	0.614	0.759	0.653
	Filler	33.3%	59.1	0.533	0.561	0.498	0.530
	All hes.	46.7%	66.7	0.570	0.687	0.746	0.668
	All	50.7%	70.3	0.580	0.643	0.769	0.664
PD	Sil.	50.7%	68.9	0.613	0.684	0.594	0.630
	Filler	38.7%	70.3	0.734	0.522	0.510	0.589
	All hes.	40.0%	55.2	0.415	0.657	0.574	0.549
	All	50.7%	70.3	0.665	0.680	0.610	0.652
DR	Sil.	60.0%	80.9	0.755	0.670	0.746	0.724
	Filler	33.3%	59.1	0.455	0.497	0.455	0.469
	All hes.	68.0%	84.2	0.857	0.706	0.802	0.788
	All	68.0%	85.4	0.842	0.700	0.775	0.773
All-3	Sil.	61.3%	80.9	0.773	0.683	0.734	0.730
	Filler	41.3%	73.9	0.758	0.566	0.526	0.617
	All hes.	68.0%	84.2	0.854	0.712	0.806	0.791
	All	70.7%	89.6	0.911	0.709	0.794	0.804

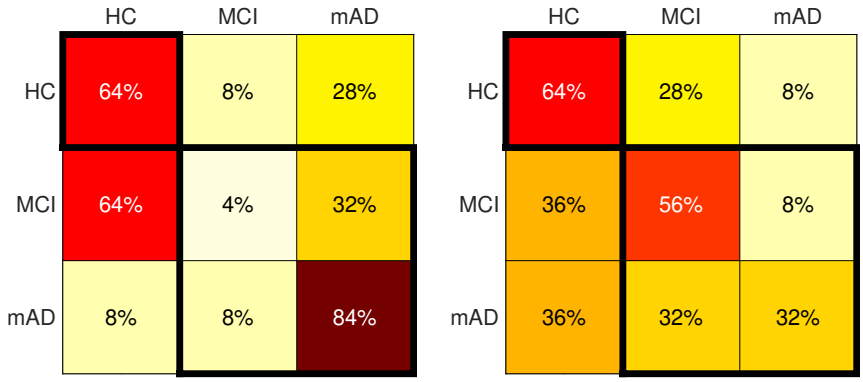
between 0.601 (previous day) and 0.690 (delayed recall). Judging from the individual values, the immediate recall and delayed recall tasks were the best for detecting mild AD speakers (AUC values of 0.743 and 0.805, respectively). Overall, the delayed recall task seems to be the most efficient one, although for the MCI speaker category, the previous day task seems to be more useful. The combined predictions, which relied on all three speaker tasks, were much better though: accuracy rose to 60%, precision to 83.7%, while the recall and sensitivity scores were both 72%, leading to an F-score of 77.4. Of course, these scores serve as a kind of baseline in this study, since they were achieved via the S-GAP temporal speech parameters described by Tóth et al. [197].

6.7.2 Results Using the Posterior-Thresholding Hesitation Representation with Context-Dependent States

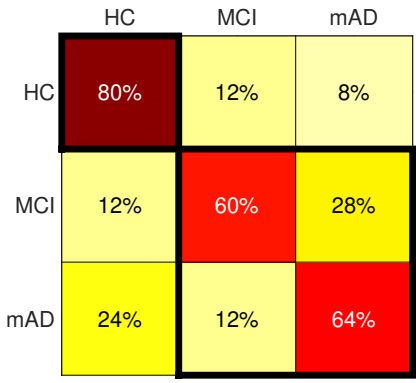
Table 6.4 lists the results got when applying the proposed Posterior-Thresholding Hesitation Representation as features, relying on the context-dependent (CD) DNN acoustic model. Regarding the speaker task of **immediate recall**, we found that relying on the silent pause-related attributes led to an acceptable performance: the classification accuracy score of 52% and the F_1 value of 68.9 are definitely above what could be achieved by random guessing, and the mean AUC score of 0.653 is fine as well; still, this score is the mean of a good AUC value for the mAD speaker category (0.759), while the values for the HC and MCI classes are much lower. Examining the classification metrics for the filler events case, we observe much lower values, which suggests that they are not useful for detecting MCI and mAD for the immediate recall speaker task. When we added the posterior estimates of the silent and filled pauses together before applying the posterior-thresholding step (i.e. step (3)) – that is, the 'All hesitation' case in Table 6.4 –, we can see similar values to those in the silent pause case. Using all three types of attribute together (case 'All') brought a slight improvement in all metric scores.

Using the recordings obtained from the **previous day** speaker task, we obtained similar scores for the silence-based attributes as before, with the exception of a higher AUC value for the MCI category. However, with filled pauses we measured higher scores than for immediate recall, which, in our opinion, indicates that this type of hesitation had different patterns for the three subject types for this particular speaker task. When we merged the phonetic states of both pause types (the 'All hesitation' case), though, our classification results fell. In the case of the **delayed recall** speaker task we found that filled pauses were not really useful; however, silent pause-related attributes led to good scores, and focusing on all hesitations was actually even (slightly) more successful: we obtained an F_1 score of 85.4 and an AUC value of 0.773 this way. Of course, the best scores were achieved by fusing the predictions for all three speaker tasks; but the improvement was only slight in most cases.

In general, we can observe that the results achieved are similar or just slightly better than those obtained with the S-GAP temporal parameters for the *immediate recall* and *previous day* speaker tasks; however, considering the fact that the proposed Posterior-Thresholding Hesitation Representation approach can be realized without a Hidden Markov model, we consider this a promising finding. For the *delayed recall* task, however, we actually obtained higher metric values: the classification accuracy score of 68.0%, the precision score of 89.1%, the recall and specificity values of $80 - 82\%$ and the F-score of 85.4 are all quite high values, all significantly exceeding those achieved via the S-GAP parameters. When utilizing the speech samples of all three



(a) *Immediate Recall* **(b)** *Previous Day*



(c) *Delayed Recall*

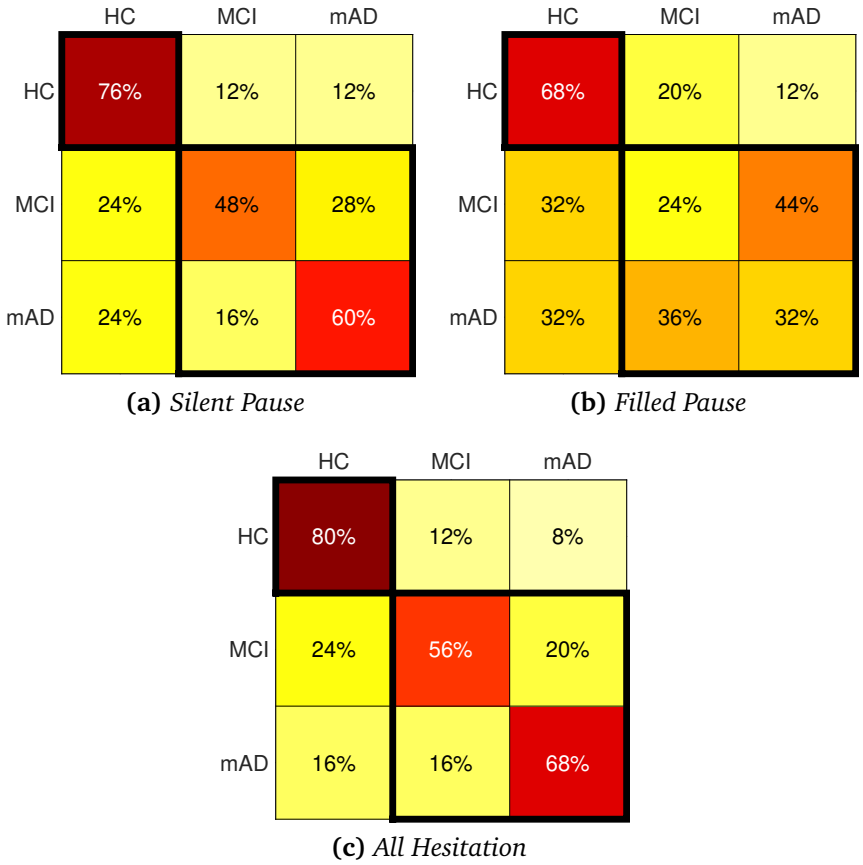
Figure 6.3: *The confusion matrices obtained for the three speaker tasks (rows: ground truth speaker categories, columns: predictions. (HC = Healthy Control, MCI = Mild Cognitive Impairment, mAD = Mild Alzheimer's Disease)*

tasks, these scores were slightly higher (with the exception of classification accuracy). Overall, we managed to achieve a mean AUC score of 0.804 as well.

6.7.3 Results Using the Posterior-Thresholding Hesitation Representation with Context-Independent States

We found that with the PTHR approach we achieved competitive scores for detecting MCI and mAD subjects using a context-dependent DNN acoustic model. However, we wanted to find out whether a context-independent Deep Neural Network component might be enough to express the likelihood of silent and filled pauses, which has the advantage that it is a much more compact model. Table 6.5 lists the results obtained using such a CI DNN acoustic model.

In general, we observe very similar tendencies to those we found in the context-dependent case. In the immediate recall speaker task, silent pauses were more use-



(a) *Silent Pause* **(b)** *Filled Pause*

(c) *All Hesitation*

Figure 6.4: *The confusion matrices obtained for the hesitation types (ground truth: real speaker categories, columns: predictions.*

ful than filled pauses, and mAD subjects were identified more precisely than either healthy controls or subjects with MCI were. In the previous day task, silent and filled pauses were similarly useful, the latter leading to a high AUC score (0.749) for the HC subject category. The most useful speaker task was again delayed recall, when we relied on silent pauses and on all hesitations. Besides these tendencies, the metric scores were quite similar as well: in most cases, using the simpler context-independent DNN hybrid acoustic models led to only a slight fall in the scores, or none at all.

6.7.4 The Performance of Speaker Tasks and Feature Subsets

In our last sequence of experiments, we examine the behaviour of the classifiers for various speaker tasks (i.e. immediate recall, previous day and delayed recall) and feature subsets (i.e. attributes based on silent pauses only, on filled pauses only, and on all hesitation). To do this, we calculated the confusion matrix for each approach.

Table 6.5: *The various accuracy scores obtained with the Posterior-Thresholding Hesitation Representation using Context-Independent states. (Acc. = classification accuracy, HC = Healthy Control, MCI = Mild Cognitive Impairment, mAD = Mild Alzheimer's Disease). On the task column: IR = Immediate Recall, PD = Previous Day, DR = Delayed Recall, All-3 = Delayed Recall*

		Classification Metrics		Area-Under-Curve			
Task	**Features**	**Acc.**	F_1	**HC**	**MCI**	**mAD**	**Mean**
IR	Silence	50.7%	68.9	0.577	0.590	0.758	0.642
	Filler	30.7%	55.2	0.467	0.509	0.553	0.510
	All hesit.	48.0%	66.7	0.574	0.693	0.769	0.678
	All	50.7%	68.9	0.580	0.597	0.759	0.645
PD	Silence	46.7%	68.9	0.618	0.648	0.550	0.605
	Filler	40.0%	73.9	0.749	0.516	0.454	0.573
	All hesit.	33.3%	48.8	0.373	0.640	0.557	0.523
	All	49.3%	72.5	0.659	0.645	0.561	0.622
DR	Silence	58.7%	80.9	0.772	0.690	0.750	0.738
	Filler	33.3%	57.5	0.478	0.517	0.488	0.494
	All hesit.	62.7%	79.6	0.778	0.684	0.798	0.753
	All	62.7%	80.9	0.786	0.685	0.794	0.755
All-3	Silence	62.7%	84.2	0.786	0.693	0.746	0.742
	Filler	40.0%	75.3	0.680	0.519	0.540	0.580
	All hesit.	64.0%	80.9	0.772	0.696	0.802	0.757
	All	69.3%	87.5	0.866	0.703	0.770	0.780

For the sake of readability, we expressed the number of subjects as percentages of the cardinality of the given (actual) speaker groups. (The columns show the hypotheses, while the rows show the correct speaker categories.)

Fig. 6.3 shows the normalized confusion matrices obtained for the various speaker tasks. That is, in these cases we limited our features to one task only, but we used the fusion of the predictions for all three feature types. Examining the matrix for the **immediate recall** task (see Fig. 6.3 (a)) we notice that the mAD speakers were identified with a high recall rate (84%); yet, the majority of the MCI subjects were classified as healthy controls, while only 4% of them (that is, one speaker) was classified correctly. In our opinion, this indicates that the immediate recall task is not quite suited for detecting mild cognitive impairment, as the symptoms of the MCI subjects are probably too subtle to distinguish them from healthy controls when recalling recent events. Regarding **previous day** (see Fig. 6.3 (b)), we see that the MCI speakers were identified with a much higher confidence than with the immediate recall task,

while the HC subjects were detected with the same accuracy. However, the mAD speakers were almost completely missed: roughly one-third of them were classified as controls, subjects having MCI, and subjects having mAD. From this figure we might draw the conclusion that this specific part of our protocol is useful for detecting Mild Cognitive Impairment, but not for identifying early Alzheimer's Disease.

Examining Fig. 6.3 (c), corresponding to **delayed recall**, we can see that perhaps this task proved to be the most effective of the three involved in our protocol: the recall rate the HC speaker category was high (80%), while MCI and mAD speakers were detected with 60% and 64% rate, respectively. But even the majority of the misclassified MCI subjects (28%) were classified as mild AD, which is a more tolerable mistake than confusing them with healthy subjects. Overall, 88% of the MCI subjects were assigned to either the MCI or the mAD category; likewise, 76% of the actual mAD subjects were classified in this way, which are rather high scores.

Fig. 6.4 shows similar confusion matrices for the attributes extracted from the posteriors of the different pause types (when using all the speaker tasks). **Silent pauses** (see Fig. 6.4 (a)) seem to be useful for distinguishing healthy controls from the other two speaker categories (with a recall rate of 76%); however, MCI and mAD subjects were detected at a lower rate (48% and 60%, respectively). However, most confusion occurred between the latter categories, and only 24% of the subjects were classified as healthy controls. Relying on **filled pauses** seems to be less effective (see Fig. 6.4 (b)): based on them, only 24% of MCI and 32% of mAD subjects were classified correctly. Still, most mistakes again arose from confusing MCI and mAD speakers, and only $32 - 32\%$ of these subjects were considered as healthy controls, while 68% of the HC speakers were classified correctly. We obtained the best values with the combination of the two pause types (see Fig. 6.4 (c), the **all hesitation** case). (Note that, as previously, silent and filled pauses were merged by adding up their frame-level posterior estimates, before the actual thresholding step; i.e. in step (2) of the PTHR method (see 6.4.2).) In this case, the percentage of correctly classified subjects was higher for all three subject categories than either for the silent or for the filled pause cases; and even the (relatively) low number of correctly identified MCI subjects (56%) was mainly due to the high number of MCI-mAD confusion instances (20%).

6.8 Concluding remarks

Alzheimer's and MCI early diagnosis might allow timely treatment to delay progression. We presented a feature extraction approach which describes the amount of hesitations without the need for a whole speech recognition workflow by relying only on the Deep Neural Network acoustic model of a standard HMM/DNN hybrid model. We calculated our features directly from the DNN outputs corresponding to the HMM

states associated with silent and/or filled pauses. Based on our experimental results, this representation allows the automatic detection of MCI and mild AD with the same (or even higher) accuracy as the temporal speech parameters developed earlier.

Our best accuracy score was 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 seem competitive with the works of other research groups. For instance, Themistocleous et al. relied on Deep Sequential Neural Networks for the classification of MCI/HC (i.e. a binary problem) using a Swedish Alzheimer's corpus; where, based on a 5-fold CV, they reported an accuracy score of 83% [193]. Fraset et al., on the same dataset, presented 0.88 and 83% of AUC and accuracy scores, respectively (also for a binary class problem) [51]; these where achieved by a multimodal language data and cascaded classifiers approach. Also, König et al. focused on the same task and extracted vocal makers from a French corpus for an automatic speech analysis approach. The authors report classification scores for HC, Alzheimer's, and MCI, however, the task was evaluated as pairwise combination of the three classes (two-class problem) [112].

As our proposed approach first adds up the frame-level likelihoods of all HMM states which were regarded as silent and/or filled pauses, it may not be necessary to employ a context-dependent (CD) neural network only to support these aggregated posterior estimates. Therefore, in our next experiment, we investigated whether using a simpler and computationally cheaper context-independent (CI) acoustic model would lead to the same subject classification performance. Our findings showed that, although there were slight drops in the various evaluation metrics, we were able to achieve the same level of performance with CI neural networks as we could with the CD ones, which might justify their application.

Regarding the feature subsets examined, we found that silent pauses were the most suitable for distinguishing mild Alzheimer's speakers from healthy controls, while the MCI detection performance was fair. Filled pauses were less effective for all three speaker groups; however, we achieved our best results when we expressed the amount of pauses regardless of their type. In this last case, only 20% of control subjects were classified as either MCI or mAD speakers, and likewise, only $16 - 24\%$ of the MCI and mAD subjects were identified as healthy. Of course, in a practical screening application there is no need to limit the input of our classifier model to just one type of pause. This is especially true as our feature set is a quite compact one, consisting only of 51 utterance-level attributes; even when merging the features corresponding to silent pauses, to filled pauses, and to all hesitations, we still have only 151 attributes for each utterance, which is significantly smaller than, say, 512-long x-vectors.

The author of this PhD thesis is responsible for the following contributions pre-

sented in this chapter (this being the first item not his main contribution):

VI/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.

VI/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.

Bibliography

[1] JRZ Abela and DU D'Alessandro. A test of the diathesis-stress and causal mediation components of beck's cognitive theory of depression. *British Journal of Clinical Psychology*, 41(1):1, 2002.

[2] A. Afshan, J. Guo, S. J. Park, V. Ravi, J. Flint, and A. Alwan. Effectiveness of voice quality features in detecting depression. *Proceedings of Interspeech*, 2018.

[3] Abeer Al-Ghazali and Yasser Alrefaee. Silent pauses in the speech of Yemeni EFL learners. *ELS Journal on Interdisciplinary Studies on Humanities*, 2(1), 2019.

[4] F.A.C. Alegria and A.M. da Cruz Serra. Influence of frequency errors in the variance of the cumulative histogram [in ADC testing]. *IEEE Transactions on Instrumentation and Measurement*, 50:461–464, 2001.

[5] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[6] Murray Alpert, Enrique R Pouget, and Raul R Silva. Reflections of depression in acoustic measures of the patient's speech. *Journal of affective disorders*, 66(1):59–69, 2001.

[7] Alzheimer's Association. 2020 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 16(3):391–460, 2020.

[8] S. Amiriparian, P. Winokurow, V. Karas, S. Ottl, M. Gerczuk, and B. W. Schuller. A Novel Fusion of Attention and Sequence to Sequence Autoencoders to Predict Sleepiness From Speech. *arXiv preprint*, 2020.

[9] Meisam Khalil Arjmandi and Mohammad Pooyan. An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine. *Biomedical Signal Processing and Control*, 7(1):3–19, 2012.

[10] American Speech-Language-Hearing Association et al. Scope of practice in speech-language pathology. 2016.

[11] American Speech-Language-Hearing Association et al. Council for clinical certification in audiology and speech-language pathology, 2019.

[12] Taiwo Oladipupo Ayodele. Types of machine learning algorithms. *New advances in machine learning*, 3:19–48, 2010.

[13] 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.

[14] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.

[15] Kirrie J Ballard. Response generalization in apraxia of speech treatments: Taking another look. *Journal of Communication Disorders*, 34(1-2):3–20, 2001.

[16] Renée L Beard. In their voices: Identity preservation and experiences of alzheimer's disease. *Journal of Aging studies*, 18(4):415–428, 2004.

[17] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *IFA Proceedings 17*, pages 97–110, 1993.

[18] Gerard Brett. The automata in the byzantine" throne of solomon". *Speculum*, 29(3):477–487, 1954.

[19] Maddalena Bruscoli and Simon Lovestone. Is MCI really just early dementia? A systematic review of conversion studies. *International Psychogeriatrics*, 16(2):129–140, 2004.

[20] Felix Burkhardt, Richard Huber, and Anton Batliner. Application of speaker classification in human machine dialog systems. In *Speaker Classification I*, pages 174–179. Springer, 2007.

[21] William M Campbell, Douglas E Sturim, and Douglas A Reynolds. Support Vector Machines using GMM supervectors for speaker verification. *IEEE signal processing letters*, 13(5):308–311, 2006.

[22] Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul):2079–2107, 2010.

[23] Chih-Chung Chang and Chih-Jeh Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:1–27, 2011.

[24] Ken Chatfield, Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman. The devil is in the details: An evaluation of recent feature encoding methods. In *British Machine Vision Conference*, volume 2, pages 76.1–76.12, 11 2011.

[25] Lim Sin Chee, Ooi Chia Ai, M Hariharan, and Sazali Yaacob. MFCC based recognition of repetitions and prolongations in stuttered speech using k-nn and lda. In *2009 IEEE Student Conference on Research and Development (SCOReD)*, pages 146–149. IEEE, 2009.

[26] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, abs/1603.02754:785–794, 2016.

[27] N. Cummins, J. Epps, V. Sethu, and J. Krajewski. Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech. In *Proceedings of ICASSP)*, pages 970–974, 2014.

[28] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

[29] Nicholas Cummins, Julien Epps, Vidhyasaharan Sethu, and Jarek Krajewski. Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 970–974. IEEE, 2014.

[30] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained Deep Neural Networks for large-vocabulary Speech Recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2011.

[31] Namrata Dave. Feature extraction methods lpc, plp and mfcc in speech recognition. *International journal for advance research in engineering and technology*, 1(6):1–4, 2013.

[32] Ken H Davis, R Biddulph, and Stephen Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 1952.

[33] Karmele López de Ipiña, Unai Martinez de Lizarduy, Pilar M. Calvo, Blanca Beitia, J. García-Melero, Elsa Fernández, Mirian Ecay-Torres, Marcos Faundez-Zanuy, and P. Sanz. On the analysis of speech and disfluencies for automatic detection of mild cognitive impairment. *Neural Computing and Applications*, 9, 2018.

[34] 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*, 19(4):788–798, 2011.

[35] Najim Dehak, Pierre Dumouchel, and Patrick Kenny. Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2095–2103, 2007.

[36] Najim Dehak, Patrick Kenny, Reda Dehak, Ondrej Glembek, Pierre Dumouchel, Lukas Burget, Valiantsina Hubeika, and Fabio Castaldo. Support vector machines and joint factor analysis for speaker verification. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4237–4240. IEEE, 2009.

[37] Najim Dehak, Pedro A. Torres-Carrasquillo, Douglas Reynolds, and Reda Dehak. Language recognition via i-vectors and dimensionality reduction. In *Proceedings of Interspeech*, pages 857–860, Florence, Italy, 2011.

[38] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Proceedings of Interspeech*, 2020.

[39] P Di Benedetto, M Cavazzon, F Mondolo, G Rugiu, A Peratoner, and E Biasutti. Voice and choral singing treatment: A new approach for speech and voice disorders in parkinson's disease. *European journal of physical and rehabilitation medicine*, 45(1):13–19, 2009.

[40] George R Doddington. Speaker recognition—identifying people by their voices. *Proceedings of the IEEE*, 73(11):1651–1664, 1985.

[41] José Vicente Egas López, Juan Rafael Orozco-Arroyave, and Gábor Gosztolya. Assessing parkinson's disease from speech using fisher vectors. 2019.

[42] José Vicente Egas-López, László Tóth, Ildikó Hoffmann, János Kálmán, Magdolna Pákáski, and Gábor Gosztolya. Assessing Alzheimer's Disease from Speech Using the i-vector Approach. In *Proceedings of SPECOM*. Springer, 2019.

[43] José Vicente Egas-López, Gábor Kiss, Dávid Sztahó, and Gábor Gosztolya. Automatic assessment of the degree of clinical depression from speech using x-vectors. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8502–8506, 2022.

[44] D. Elsner, S. Langer, F. Ritz, R. Mueller, and S. Illium. Deep Neural Baselines for Computational Paralinguistics. *arXiv preprint arXiv:1907.02864*, 2019.

[45] M.F. Folstein, S.E. Folstein, and P.R. McHugh. Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198, 1975.

[46] Everthon Silva Fonseca, Rodrigo Capobianco Guido, Paulo Rogério Scalassara, Carlos Dias Maciel, and José Carlos Pereira. Wavelet time-frequency analysis and least squares support vector machines for the identification of voice disorders. *Computers in Biology and Medicine*, 37(4):571–578, 2007.

[47] H. Förstl and A. Kurz. Clinical features of Alzheimer's disease. *European Archives of Psychiatry and Clinical Neuroscience*, 249(6):288–290, Dec 1999.

[48] Norman L. Foster, Mark W. Bondi, Rohit Das, Mary Foss, Linda A. Hershey, Steve Koh, Rebecca Logan, Carol Poole, Joseph W. Shega, Ajay Sood, Niranjan Thothala, Meredith Wicklund, Melissa Yu, Amy Bennett, and David Wang. Quality improvement in neurology. *Neurology*, 93(16):705–719, 2019.

[49] EM. Frank. Effect of Alzheimer's Disease on Communication function. *IJournal of the South Carolina Medical Association.*, 9(90):417–23, 1994.

[50] Kathleen Fraser, Frank Rudzicz, Naida Graham, and Elizabeth Rochon. Automatic speech recognition in the diagnosis of primary progressive aphasia. In *Proceedings of SLPAT*, pages 47–54, Grenoble, France, 2013.

[51] Kathleen C. Fraser, Kristina Lundholm Fors, Marie Eckerström, Fredrik Öhman, and Dimitrios Kokkinakis. Predicting MCI status from multimodal language data using cascaded classifiers. *Frontiers in Aging Neuroscience*, 11, 2019.

[52] M. Freedman, L. Leach, E. Kaplan, G. Winocur, K.I. Shulman, and D. Delis. *Clock Drawing: A Neuropsychological Analysis*. New York: Oxford University Press, 1994.

[53] Jerome H Friedman. Greedy function approximation: a Gradient Boosting Machine. *Annals of statistics*, pages 1189–1232, 2001.

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

[55] J. Fritsch, S.P. Dubagunta, and M. Magimai-Doss. Estimating the Degree of Sleepiness by Integrating Articulatory Feature Knowledge in Raw Waveform Based CNNS. In *Proceedings of ICASSP*, pages 6534–6538. IEEE, 2020.

[56] Sadaoki Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2):254–272, 1981.

[57] Sadaoki Furui. 50 years of progress in speech and speaker recognition. In *Proc. SPECOM*, pages 1–9. Citeseer, 2005.

[58] James E. Galvin and Carl H. Sadowsky. Practical guidelines for the recognition and diagnosis of dementia. *The Journal of the American Board of Family Medicine*, 25(3):367–382, 2012.

[59] Todor Ganchev, Nikos Fakotakis, and George Kokkinakis. Comparative evaluation of various mfcc implementations on the speaker verification task. In *Proceedings of the SPECOM*, volume 1, pages 191–194, 2005.

[60] Nicanor García, Juan Rafael Orozco, Luis D'Haro, Najim Dehak, and Elmar Noeth. Evaluation of the neurological state of people with parkinson's disease using i-vectors. pages 299–303, 08 2017.

[61] 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.

[62] Daniel Garcia-Romero, Xinhui Zhou, and Carol Y Espy-Wilson. Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4257–4260. IEEE, 2012.

[63] JI Godino-Llorente, S Shattuck-Hufnagel, JY Choi, L Moro-Velázquez, and JA Gómez-García. Towards the identification of idiopathic parkinson's disease from the speech. new articulatory kinetic biomarkers. *PloS one*, 12(12):e0189583, 2017.

[64] G. Gosztolya, T. Grósz, G. Szaszák, and L. Tóth. Estimating the sincerity of apologies in speech by DNN rank learning and prosodic analysis. In *Proceedings of Interspeech*, pages 2026–2030, San Francisco, CA, USA, Sep 2016.

[65] Gábor Gosztolya. Posterior-thresholding feature extraction for paralinguistic speech classification. *Knowledge-Based Systems*, 186, 2019.

[66] 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.

[67] Gábor Gosztolya. Using the Fisher vector representation for audio-based emotion recognition. *Acta Polytechnica Hungarica*, page to appear, 2020.

[68] Gábor Gosztolya, Anita Bagi, Szilvia Szalóki, István Szendi, and Ildikó Hoffmann. Identifying schizophrenia based on temporal parameters in spontaneous speech. In *Proceedings of Interspeech*, pages 3408–3412, Hyderabad, India, Sep 2018.

[69] Gábor Gosztolya, Róbert Busa-Fekete, Tamás Grósz, and László Tóth. DNN-based feature extraction and classifier combination for child-directed speech, cold and snoring identification. In *Proceedings of Interspeech*, pages 3522–3526, Stockholm, Sweden, Aug 2017.

[70] Gábor Gosztolya, Tamás Grósz, and László Tóth. General utterance-level feature extraction for classifying crying sounds, atypical & self-assessed affect and heart beats. In *Proceedings of Interspeech*, pages 531–535, Hyderabad, India, Sep 2018.

[71] Gábor Gosztolya, Tamás Grósz, László Tóth, and David Imseng. Building context-dependent DNN acousitc models using Kullback-Leibler divergence-based state tying. In *Proceedings of ICASSP*, pages 4570–4574, Brisbane, Australia, Apr 2015.

[72] Gábor Gosztolya, László Tóth, Tamás Grósz, Veronika Vincze, Ildikó Hoffmann, Gréta Szatlóczki, Magdolna Pákáski, and János Kálmán. Detecting Mild Cognitive Impairment from spontaneous speech by correlation-based phonetic feature selection. In *Proceedings of Interspeech*, pages 107–111, San Francisco, CA, USA, Sep 2016.

[73] Gábor Gosztolya, Veronika Vincze, László Tóth, Magdolna Pákáski, János Kálmán, and Ildikó Hoffmann. Identifying Mild Cognitive Impairment and mild Alzheimer's disease based on spontaneous speech using asr and linguistic features. *Computer Speech & Language*, 53:181 – 197, 2019.

[74] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. PMLR, 2014.

[75] Steven Greenberg and Brian ED Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1647–1650. IEEE, 1997.

[76] Joanna Grzybowska and Stanislaw Kacprzak. Speaker age classification and regression using i-vectors. In *Proceedings of Interspeech*, pages 1402–1406, San Francisco, CA, USA, Sep 2016.

[77] Rodrigo C Guido, Jose C Pereira, Everthon Fonseca, Fabrico L Sanchez, and Lucimar S Vieira. Trying different wavelets on the search for voice disorders sorting. In *Proceedings of the Thirty-Seventh Southeastern Symposium on System Theory, 2005. SSST'05.*, pages 495–499. IEEE, 2005.

[78] J. H. L. Hansen and T. Hasan. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, 32(6):74–99, Nov 2015.

[79] Muthusamy Hariharan, Lim Sin Chee, Ooi Chia Ai, and Sazali Yaacob. Classification of speech dysfluencies using lpc based parameterization techniques. *Journal of medical systems*, 36(3):1821–1830, 2012.

[80] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[81] L. Heutte, T. Paquet, J.V. Moreau, Y.Lecourtier, and C.Olivier. A structural/statistical feature based vector for handwritten character recognition. *Pattern Recognition Letters*, 19:629–641, 1998.

[82] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[83] P.S. Hiremath and S. Shivashankar. Wavelet based co-occurrence histogram features for texture classification with an application to script identification in a document image. *Pattern Recognition Letters*, 29:1182–1189, 2008.

[84] Ildikó Hoffmann, Dezső Németh, Cristina D Dye, Magdolna Pákáski, Tamás Irinyi, and János Kálmán. Temporal parameters of spontaneous speech in Alzheimer's disease. *International Journal of Speech-Language Pathology*, 12(1):29–34, 2010.

[85] Ildikó Hoffmann, László Tóth, Gábor Gosztolya, Gréta Szatlóczki, Veronika Vincze, Eszter Kárpáti, Magdolna Pákáski, and János Kálmán. Beszédfelismerés alapú eljárás az enyhe kognitív zavar automatikus felismerésére spontán beszéd alapján. *Általános nyelvészeti tanulmányok*, 29(1):385–405, 2017.

[86] Z. Huang, J. Epps, D. Joachim, B. Stasak, J. Williamson, and T. Quatieri. Domain adaptation for enhancing Speech-based depression detection in natural environmental conditions using dilated CNNs. *Proceedings of Interspeech 2020*, pages 4561–4565, 2020.

[87] 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.

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

[89] Magdalena Igras-Cybulska, Bartosz Ziółko, Piotr Żelasko, and Marcin Witkowski. Structure of pauses in speech in the context of speaker verification and classification of speech type. *EURASIP Journal on Audio, Speech, and Music Processing*, 2016(1):18, 2016.

[90] Motonobu Itoh. Articulatory movements in apraxia of speech. *Apraxia of speech: Physiology, acoustics, linguistics, management*, pages 135–165, 1984.

[91] Tommi S. Jaakkola and David Haussler. Exploiting Generative Models in Discriminative Classifiers. In *Proceedings of NIPS*, pages 487–493, Denver, CO, USA, 1998.

[92] 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.

[93] Igor Jauk. *Unsupervised learning for expressive speech synthesis*. Universitat Politècnica de Catalunya, 2017.

[94] Frederick Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556, 1976.

[95] Attila Zoltán Jenei and Gábor Kiss. Possibilities of recognizing depression with convolutional networks applied in correlation structure. In *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*, pages 101–104. IEEE, 2020.

[96] Swati Johar. *Emotion, affect and personality in speech: The Bias of language and paralanguage*. Springer, 2015.

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

[98] Ian. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[99] Lorraine V. Kalia and Anthony E. Lang. Parkinon's disease. *The Lancet*, 386(4):896–912, 2015.

[100] Heysem Kaya and Alexey A Karpov. Introducing weighted kernel classifiers for handling imbalanced paralinguistic corpora: Snoring, addressee and cold. In *INTERSPEECH*, pages 3527–3531, 2017.

[101] Heysem Kaya, Alexey A. Karpov, and Albert Ali Salah. Fisher Vectors with cascaded normalization for paralinguistic analysis. In *Proceedings of Interspeech*, pages 909–913, 2015.

[102] Daniel Kempler and Diana Van Lancker. Effect of speech task on intelligibility in dysarthria: A case study of parkinson's disease. *Brain and language*, 80(3):449–464, 2002.

[103] Patrick Kenny. Joint factor analysis of speaker and session variability: Theory and algorithms. *CRIM, Montreal,(Report) CRIM-06/08-13*, 14(28-29):2, 2005.

[104] Patrick Kenny. Bayesian speaker verification with heavy-tailed priors. In *Odyssey*, volume 14, 2010.

[105] Ray D Kent. Research on speech motor control and its disorders: A review and prospective. *Journal of Communication disorders*, 33(5):391–428, 2000.

[106] Raymond D Kent and John C Rosenbek. Acoustic patterns of apraxia of speech. *Journal of Speech, Language, and Hearing Research*, 26(2):231–249, 1983.

[107] Lawrence George Kersta. Voiceprint identification. *The Journal of the Acoustical Society of America*, 34(5):725–725, 1962.

[108] G. Kiss, M. Tulics, D. Sztahó, A. Esposito, and K. Vicsi. Language independent detection possibilities of depression by speech. In *Recent advances in nonlinear speech processing*, pages 103–114. Springer, 2016.

[109] G. Kiss and K. Vicsi. Mono-and multi-lingual depression prediction based on speech processing. *International Journal of Speech Technology*, 20, 2017.

[110] Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[111] Alexandra König, Aharon Satt, Alex Sorin, Ran Hoory, Alexandre Derreumaux, Renaud David, and Phillippe H. Robert. Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. *Current Alzheimer Research*, 15(2):120–129, 2018.

[112] Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H. Robert, and Renaud David. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's Disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):112–124, 2015.

[113] Sotiris B Kotsiantis, I Zaharakis, P Pintelas, et al. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24, 2007.

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

[115] Kaitlin L Lansford and Julie M Liss. Vowel acoustics in dysarthria: Speech disorder diagnosis and classification. 2014.

[116] Christoph Laske, Hamid R. Sohrabi, Shaun M. Frost, Karmele López de Ipiña, Peter Garrard, Massimo Buscema, Justin Dauwels, Surjo R. Soekadar, Stephan Mueller, Christoph Linnemann, Stephanie A. Bridenbaugh, Yogesan Kanagasingam, Ralph N. Martins, and Sid E. O'Bryant. Innovative diagnostic tools for early detection of Alzheimer's disease. *Alzheimer's & Dementia*, 11(5):561–578, 2015.

[117] Iulia Lefter, Gertjan J Burghouts, and Leon JM Rothkrantz. An audio-visual dataset of human–human interactions in stressful situations. *Journal on Multimodal User Interfaces*, 8(1):29–41, 2014.

[118] Iulia Lefter, Léon JM Rothkrantz, and Gertjan J Burghouts. A comparative study on automatic audio–visual fusion for aggression detection using meta-information. *Pattern Recognition Letters*, 34(15):1953–1963, 2013.

[119] Maider Lehr, Emily Prud'hommeaux, Izhak Shafran, and Brian Roark. Fully automated neuropsychological assessment for detecting Mild Cognitive Impairment. In *Proceedings of Interspeech*, pages 1039–1042, Portland, OR, USA, 2012.

[120] Feng Li. Textual analysis of corporate disclosures: A survey of the literature. *Journal of accounting literature*, 29(1):143–165, 2010.

[121] Jie-Min Long, Zhang-Fa Yan, Yu-Lin Shen, Wei-Jun Liu, and Qing-Yang Wei. Detection of Epilepsy using MFCC-Based Feature and XGBoost. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–4. IEEE, 2018.

[122] P. Lopez-Otero and L. Docio-Fernandez. Analysis of gender and identity issues in depression detection on de-identified speech. *Computer Speech & Language*, 65:101118, 2021.

[123] Daniel M Low, Kate H Bentley, and Satrajit S Ghosh. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1):96–116, 2020.

[124] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[125] Miranti Indar Mandasari, ML McLaren, and David A van Leeuwen. Evaluation of i-vector speaker recognition systems for forensic application. 2011.

[126] Sven L. Mattys, Christopher W. Pleydell-Pearce, James F. Melhorn, and Sharron E. Whitecross. Detecting silent pauses in speech: A new tool for measuring on-line lexical and semantic processing. *Psychological Science*, 16(12):958–964, 2005.

[127] Kim C. McCullough, Kathryn A. Bayles, and Erin D. Bouldin. Language performance of individuals at risk for mild cognitive impairment. *Journal of Speech, Language, and Hearing Research*, 62(3):706–722, 2018.

[128] Guy M. McKhann, David S. Knopman, Howard Chertkow, Bradley T. Hyman, Clifford R. Jack Jr., Claudia H. Kawas, William E. Klunk, Walter J. Koroshetz, Jennifer J. Manly, Richard Mayeux, Richard C. Mohs, John C. Morris, Martin N. Rossor, Philip Scheltens, Maria C. Carrillo, Bill Thies, Sandra Weintraub, and Creighton H. Phelps. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging – Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3):263–269, 2011.

[129] Malcolm Ray McNeil. *Clinical management of sensorimotor speech disorders*. Thieme, 2009.

[130] Tom Michael Mitchell. *The discipline of machine learning*, volume 9. Carnegie Mellon University, School of Computer Science, Machine Learning . . . , 2006.

[131] 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.

[132] Sirko Molau, Michael Pitz, and Hermann Ney. Histogram based normalization in the acoustic feature space. In *Proceedings of ASRU*, pages 1–4, Madonna di Campiglio, Italy, Dec 2001.

[133] Pedro J. Moreno and Ryan Rifkin. Using the Fisher kernel method for web audio classification. In *Proceedings of ICASSP*, pages 2417–2420, Dallas, TX, USA, 2010.

[134] Laureano Moro-Velázquez, Jorge Andrés Gómez-García, Juan Ignacio Godino-Llorente, Jesús Villalba, Juan Rafael Orozco-Arroyave, and Najim Dehak. Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinon's disease. *Applied Soft Computing*, 62:649–666, 2018.

[135] Kimberly D. Mueller, Rebecca L. Koscik, Bruce P. Hermann, Sterling C. Johnson, and Lyn S. Turkstra. Declines in connected language are associated with very early mild cognitive impairment: Results from the Wisconsin registry for Alzheimer's prevention. *Frontiers in Aging Neuroscience*, 9, 2018.

[136] B.J. Murray. A Practical Approach to Excessive Daytime Sleepiness: a Focused Review. *Canadian respiratory journal*, 2016, 2016.

[137] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7:21, 2013.

[138] Lucy Nelson and Naji Tabet. Slowing the progression of Alzheimer's Disease; what works? *Ageing Research Reviews*, 23(B):193–209, 2015.

[139] 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.

[140] J Neuliep. The nonverbal code in intercultural communication: A contextual approach (pp. 285-332), 2011.

[141] Michal Novotnỳ, Jan Rusz, Roman Čmejla, and Evžen Ržička. Automatic evaluation of articulatory disorders in parkinson's disease. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(9):1366–1378, 2014.

[142] J.J. Odell. *The Use of Context in Large Vocabulary Speech Recognition*. PhD thesis, University of Cambridge, 1995.

[143] Harry F Olson and Herbert Belar. Phonetic typewriter. *The Journal of the Acoustical Society of America*, 28(6):1072–1081, 1956.

[144] Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesús Francisco Vargas-Bonilla, María Claudia Gonzalez-Rátiva, and Elmar Nöth. New Spanish speech corpus database for the analysis of people suffering from Parkinson's Disease. In *Proceedings of LREC*, pages 26–31, Reykjavik, Iceland, May 2014.

[145] Juan Rafael Orozco-Arroyave, JC Vásquez-Correa, Florian Hönig, Julián D Arias-Londoño, JF Vargas-Bonilla, Sabine Skodda, Jan Rusz, and Elmar Nöth. Towards an automatic monitoring of the neurological state of Parkinon's patients from speech. In *Proceedings of ICASSP*, pages 6490–6494. IEEE, 2016.

[146] J.F. Pagel. Excessive Daytime Sleepiness. *American Family Physician*, 79(5):391–396, 2009.

[147] Yilin Pan, Venkata Srikanth Nallanthighal, Daniel Blackburn, Heidi Christensen, and Aki Härmä. Multi-task estimation of age and cognitive decline from speech. In *Proceedings of ICASSP*, pages 7258–7262, Toronto, Canada (online), Jun 2021.

[148] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, 150:109–125, 2016.

[149] P.A. Pérez-Toro, J.C. Vásquez-Correa, T. Arias-Vergara, P. Klumpp, M. Sierra-Castrillón, ME Roldán-López, D Aguillón, L Hincapié-Henao, CA Tóbon-Quintero, T Bocklet, et al. Acoustic and linguistic analyses to assess early-onset and genetic Alzheimer's Disease. In *Proceedings of ICASSP*, pages 8338–8342, Toronto, Canada (online), Jun 2021.

[150] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *Proceedings of CVPR*, 2007.

[151] Ronald C. Petersen. *Conceptual Overview*, pages 1–14. Oxford University Press, 2003.

[152] Ronald C. Petersen, Barbara Caracciolo, Carol Brayne, Serge Gauthier, Vesna Jelic, and Laura Fratiglioni. Mild cognitive impairment: a concept in evolution. *Journal of Internal Medicine*, 275(3):214–228, 2014.

[153] Ronald C. Petersen, Glenn E. Smith, Stephen C. Waring, Robert J. Ivnik, Eric G. Tangalos, and Emre Kokmen. Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology*, 56(3):303–308, 1999.

[154] Serge Pinto, Rita Cardoso, Jasmin Sadat, Isabel Guimarães, Céline Mercier, Helena Santos, Cyril Atkinson-Clement, Joana Carvalho, Pauline Welby, Pedro Oliveira, 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*, 6(11):e012885, 2016.

[155] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Vesel. The Kaldi speech recognition toolkit. *Proceedings of ASRU*, 01 2011.

[156] Kymberlie Preiss, Leah Brennan, and David Clarke. A systematic review of variables associated with the relationship between obesity and depression. *Obesity Reviews*, 14(11):906–918, 2013.

[157] Martin Prince, Anders Wimo, Maëlenn Guerchet, Gemma-Claire Ali, Yu-Tzu Wu, and Matthew Prina. *World Alzheimer Report 2015. The Global Impact of Dementia*. Alzheimer's Disease International, London, UK, 2015.

[158] Sarunas J Raudys, Anil K Jain, et al. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on pattern analysis and machine intelligence*, 13(3):252–264, 1991.

[159] Douglas A Reynolds. Automatic speaker recognition using gaussian mixture speaker models. In *The Lincoln Laboratory Journal*. Citeseer, 1995.

[160] Douglas Alan Reynolds. *A Gaussian mixture modeling approach to text-independent speaker identification*. PhD thesis, Georgia Institute of Technology, 1992.

[161] Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. Spoken language derived measures for detecting mild cognitive impairment. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2081–2090, 2011.

[162] W.G. Rosen, R.C. Mohs, and K.L. Davis. A new rating scale for Alzheimer's disease. *Journal of Psychiatric Research*, 141(11):1356–1364, 1984.

[163] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the Fisher Vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[164] Aharon Satt, Ron Hoory, Alexandra König, Pauline Aalten, and Philippe H. Robert. Speech-based automatic and robust detection of very early dementia. In *Proceedings of Interspeech*, pages 2538–2542, Singapore, 2014.

[165] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alexander J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[166] Robert A. Schowengerdt. *Remote Sensing: Models and Methods for Image Processing*. Academic Press, Orlando, FL, USA, 2006.

[167] Björn Schuller and Anton Batliner. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[168] Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. 2009.

[169] 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.

[170] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39, 2013.

[171] Björn Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Yue Zhang. The interspeech 2014 computational paralinguistics challenge: Cognitive & physical load, multitasking. In *Proceedings INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore*, 2014.

[172] 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.

[173] Björn Schuller, Stefan Steidl, Anton Batliner, Simone Hantke, Florian Hönig, Juan Rafael Orozco-Arroyave, Elmar Nöth, Yue Zhang, and Felix Weninger. The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinson's & eating condition. 2015.

[174] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob Van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, et al. The interspeech 2012 speaker trait challenge. In *INTERSPEECH 2012, Portland, OR, USA*, 2012.

[175] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[176] 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.Rothkrantz 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.

[177] B.W. Schuller, A. Batliner, C. Bergler, F.B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S. Roelen, S. Schnieder, E. Bergelson, et al. The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. In *Proceedings of Interspeech*, pages 2378–2382, 2019.

[178] Marco Seeland, Michael Rzanny, Nedal Alaqraa, Jana Wäldchen, and Patrick Mäder. Plant Species Classification using Flower Images: A Comparative Study of Local Feature Representations. *PLOS ONE*, 12(2):1–29, 02 2017.

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

[180] M. Senoussaoui, M. Sarria-Paja, J. F Santos, and T. H Falk. Model fusion for multimodal depression classification and level detection. In *Proceedings of AVEC*, pages 57–63, 2014.

[181] A. Shahid, K. Wilkinson, S. Marcu, and C. M Shapiro. Karolinska Sleepiness Scale (KSS). In *STOP, THAT and One Hundred Other Sleep Scales*, pages 209–210. Springer, 2011.

[182] Ben J Shannon and Kuldip K Paliwal. A comparative study of filter bank spacing for speech recognition. In *Microelectronic engineering research conference*, volume 41, pages 310–12, 2003.

[183] Lawrence D Shriberg, Rhea Paul, Lois M Black, and Jan P Van Santen. The hypothesis of apraxia of speech in children with autism spectrum disorder. *Journal of autism and developmental disorders*, 41(4):405–426, 2011.

[184] Rachel A. Sluis, Daniel Angus, Janet Wiles, Andrew Back, Tingting (Amy) Gibson, Jacki Liddle, Peter Worthy, David Copland, and Anthony J. Angwin. An automated approach to examining pausing in the speech of people with dementia. *American Journal of Alzheimer's Disease & Other Dementias*, 35, 2020.

[185] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur. Deep Neural Network embeddings for text-independent speaker verification. In *Proceedings of Interspeech*, 2017.

[186] 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.

[187] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur. Deep Neural Network-based speaker embeddings for end-to-end speaker verification. In *Proceedings of SLT*, pages 165–170, 2016.

[188] Hickson St and NJ Moore. Nonverbal communication: studies and applications, 2004.

[189] Joseph C Stemple, Nelson Roy, and Bernice K Klaben. *Clinical voice pathology: Theory and management*. Plural Publishing, 2018.

[190] Gréta Szatlóczki, Ildikó Hoffmann, Veronika Vincze, János Kálmán, and Magdolna Pákáski. Speaking in Alzheimer's Disease, is that an early sign? Importance of changes in language abilities in Alzheimer's Disease. *Frontiers in Aging Neuroscience*, 7, 2015.

[191] Vanessa Taler and N.A. Phillips. Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. *Journal of Clinical and Experimental Neuropsychology*, 30(5):501–556, 2008.

[192] D Tannen. That's not what i meant: how communication style makes or breaks relationship, 1986.

[193] Charalambos Themistocleous, Marie Eckerström, and Dimitrios Kokkinakis. Identification of Mild Cognitive Impairment from speech in Swedish using Deep Sequential Neural Networks. *Frontiers in Neurology*, 9, 2018.

[194] Charalambos Themistocleous, Marie Eckerström, and Dimitrios Kokkinakis. Voice quality and speech fluency distinguish individuals with Mild Cognitive Impairment from Healthy Controls. *PloS one*, 15(7), 2020.

[195] Deborah G Theodoros, Gabriella Constantinescu, Trevor G Russell, Elizabeth C Ward, Stephen J Wilson, and Richard Wootton. Treating the speech disorder in Parkinon's disease online. *Journal of Telemedicine and Telecare*, 12(3_suppl):88–91, 2006.

[196] László Tóth, Gábor Gosztolya, Veronika Vincze, Ildikó Hoffmann, Gréta Szatlóczki, Edit Biró, Fruzsina Zsura, Magdolna Pákáski, and János Kálmán. Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In *Proceedings of Interspeech*, pages 2694–2698, Dresden, Germany, Sep 2015.

[197] László Tóth, Ildikó Hoffmann, Gábor Gosztolya, Veronika Vincze, Gréta Szatlóczki, Zoltán Bánréti, Magdolna Pákáski, and János Kálmán. A Speech Recognition-based solution for the automatic detection of Mild Cognitive Impairment from spontaneous speech. *Current Alzheimer Research*, 15(2):130–138, 2018.

[198] Athanasios Tsanas, Max A Little, Patrick E McSharry, Jennifer Spielman, and Lorraine O Ramig. Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. *IEEE transactions on biomedical engineering*, 59(5):1264–1271, 2012.

[199] Gwen Van Nuffelen, Marc De Bodt, Jan Vanderwegen, Paul Van de Heyning, and Floris Wuyts. Effect of rate control on speech production and intelligibility in dysarthria. *Folia Phoniatrica et Logopaedica*, 62(3):110–119, 2010.

[200] Andrea Vedaldi and Brian Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1469–1472. ACM, 2010.

[201] Martin Vetterli. A theory of multirate filter banks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3):356–372, 1987.

[202] Chen Wang, Chengyuan Deng, and Suzhen Wang. Imbalance-XGBoost: Leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *arXiv preprint arXiv:1908.01672*, 2019.

[203] Sheng-Hui Wang, Huai-Ting Li, En-Jui Chang, and An-Yeu Andy Wu. Entropy-assisted emotion recognition of valence and arousal using XGBoost classifier. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 249–260. Springer, 2018.

[204] Tianqi Wang, Chongyuan Lian, Jingshen Pan, Quanlei Yan, Feiqi Zhu, Manwa L. Ng, Lan Wang, and Nan Yan. Towards the speech features of mild cognitive impairment: Universal evidence from structured and unstructured connected speech of Chinese. In *Proceedings of Interspeech*, pages 3880–3884, Graz, Austria, Sep 2019.

[205] Kate E Watkins, Nina F Dronkers, and Faraneh Vargha-Khadem. Behavioural analysis of an inherited speech and language disorder: comparison with acquired aphasia. *Brain*, 125(3):452–464, 2002.

[206] J. Williamson, T. Quatieri, B. Helfer, R. Horwitz, B. Yu, and D. Mehta. Vocal biomarkers of depression based on motor incoordination. In *Proceedings of AVEC*, pages 41–48, 2013.

[207] H. Wu, W. Wang, and M. Li. The DKU-LENOVO systems for the INTERSPEECH 2019 Computational Paralinguistic Challenge. In *Proceedings of Interspeech*, pages 2433–2437, 2019.

[208] Jian-Da Wu and Bing-Fu Lin. Speaker identification using discrete wavelet packet transform technique with irregular decomposition. *Expert Systems with Applications*, 36(2):3136–3143, 2009.

[209] Lonce Wyse. Audio spectrogram representations for processing with convolutional neural networks. *arXiv preprint arXiv:1706.09559*, 2017.

[210] Rui Xia and Yang Liu. Using i-vector space model for emotion recognition. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[211] S.L. Yeh, G.Y. Chao, B.H. Su, Y.L. Huang, M.H. Lin, Y.C. Tsai, Y.W. Tai, Z.C. Lu, C.Y. Chen, T.M. Tai, et al. Using Attention Networks and Adversarial Augmentation for Styrian Dialect Continuous Sleepiness and Baby Sound Recognition. In *Proceedings of Interspeech*, pages 2398–2402, 2019.

[212] Jerome A. Yesavage and Javaid I. Sheikh. 9/geriatric depression scale (gds). *Clinical Gerontologist*, 5(1–2):165–173, 1986.

[213] Kathryn M Yorkston, Mark Hakel, David R Beukelman, and Susan Fager. Evidence for effectiveness of treatment of loudness, rate, or prosody in dysarthria: A systematic review. *Journal of Medical Speech-Language Pathology*, 15(2):xi–xi, 2007.

[214] Dong Yu, Li Deng, and George Dahl. Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world Speech Recognition. In *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[215] Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*, 2020.

[216] Luke Zhou, Kathleen C Fraser, and Frank Rudzicz. Speech recognition in Alzheimer's disease and in its assessment. In *INTERSPEECH*, pages 1948–1952, 2016.

Summary

This PhD thesis presented 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 employed 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 consisted of four major parts, in the sections below we will summarize the results of Chapters 3-6. 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 methods 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.

Work I.

In Chapter 3, we proposed the i-vector approach for the extraction of features from the 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 linear-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.

Work II.

In Chapter 4, we represented the utterances of subjects having Parkinson's Disease and those of healthy controls by means of the Fisher Vector approach. This method is common in image recognition, where it provides a representation of the local image descriptors via frequency and high order statistics. 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.

Moreover, we demonstrated that using Fisher vectors for the assessment of the levels of escalation in speech and primate species sounds leads to competitive or even better results than their x-vector embeddings counterparts for that specific corpus. Likewise, we presented models based on Fisher vector representations for the estimation of cold. We found that XGBoost algorithms were able to represent cold patterns in a better way than SVM for the given set of features.

Work III.

In Chapter 5, we employed the x-vector approach as a neural network feature extractor to detect the level of sleepiness of a speaker. We used different corpora for training the x-vector DNN from scratch, and also experimented with adding noise and reverberation to the audio samples. Using the publicly available Dusseldorf Sleepy Language Corpus, we demonstrated that our custom x-vector embeddings as features for Support Vector Regression consistently led to competitive performance scores in sleepiness detection. Our methods achieved the highest Spearman's correlation coefficient on the mentioned corpus that was achieved by a single method.

Furthermore, we introduced custom x-vector extractors and explored the performance of an out-of-domain pre-trained x-vector model for the estimation of the levels of depression. Our findings confirmed that x-vectors were able to capture meaningful speaker traits that contain information for depression discrimination. We demonstrated that the language of the extractor is of secondary importance compared to the frame-level feature set. Namely, our best model, which achieved an AUC score of 0.940 and an RMSE score of 9.54, was trained on log-energies (FBANKS) instead of MFCCs.

Lastly, we also presented an ensemble of classifiers from x-vector features for both conflict escalation estimation, and primate species sounds identification, respectively. We boosted the final performances by incorporating the SSPNet Conflict Corpus in the conflict escalation training workflow surpassing official baselines on the given task.

Work IV.

In Chapter 6, we presented a set of temporal speech parameters consisting of articulation rate, speech tempo and various other attributes describing the hesitation of a subject suffering from MCI. We showed the possibility of extracting these representations in a reliable way regardless of the language of the ASR system employed. Our experiments indicated that the language on which the ASR system was trained only slightly affects the MCI classification performance as multilingual (67-92%) and monolingual (67-92%) scores were similar.

On the other hand, we also introduced a similar feature extraction approach based on the same 'temporal speech parameters' which still quantifies the amount of silence and hesitation in the speech of the subject, but does not require the application of a full ASR system. We demonstrated 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 those from the ASR-based temporal parameter extraction approach for MCI and Alzheimer's detection.

Contributions of the thesis

In the **first thesis group**, the contributions are related to the automatic screening of Alzheimer's Disease by means of the i-vector approach using the speech of subjects. 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**, the 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 modeled 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**, the 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 6.6 summarizes the relation between the thesis points and the corresponding publications.

Table 6.6: *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]		•						•			

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 nemverbális 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.

1. téziscsoport

A 3. fejezet az i-vektoros megközelítés használatát tárgyalja a beszédből a jellemzők kinyerésére. Ezek a jellemzők képesek voltak modellezni a három beszélőcsoport (AD – Alzheimer-kór, MCI – enyhe kognitív zavar, HC – egészséges kontroll) beszédmintázatát. Az i-vektor reprezentációkat mel-frekvenciás együtthatókból (MFCC) számítottuk, majd egy lineáris SVM segítségével végeztük el a beszélők osztályozását. Az osztályozás során 5-szörös keresztvalidációt alkalmaztunk, a pontosságot pedig F1-értékkel mértük.

2. téziscsoport

A 4. fejezetben olyan kísérleteket mutatunk be, amelyekben Fisher vektorokat (FV) alkalmaztunk Parkinson-kórban szenvedő alanyok és egészséges kontrollok beszédfelvételeinek feldolgozása során. Ez a módszer a képfelismerésben elterjedt, ahol lokális deskriptorokat összegez gyakoriság és egyéb magasabbrendű statisztikák segítségével. Az FV eljárást négy különböző keretszintű jellemzőkészletre próbáltuk ki, míg az osztályozást lineáris SVM segítségével végeztük. Eredményeink azt mutatták, hogy megközelítésünkkel jobb osztályozási teljesítmény érhető el, mint az i-vektorokra vagy a koszinusz-távolságra támaszkodva, és jó alapot biztosított hogy különböző jellemzőkön vagy beszédfeladatokon tanított osztályozók kombinációjára is. A téziscsoportban ezen felül azt is megmutattuk, hogy a Fisher vektorok hatékony jellemzőkészletnek bizonyultak konfliktusok kibontakozásának detektálására, valamint különböző főemlős-fajok hangalapú megkülönböztetésére is (ezeken a feladatokon hatékonyabbnak bizonyultak, mint az x-vektorok). Hasonlóképpen, Fisher-vektor reprezentációkon alapuló modelleket tanítottunk a beszélő megfázásának megállapítására is. Azt találtuk, hogy az XGBoost osztályozási eljárás magasabb pontosságra volt képes, mint a lineáris SVM.

3. téziscsoport

Az 5. fejezetben az x-vektor mint jellemzőkinyerő eszköz alkalmazását tárgyaljuk különböző szűrési feladatok automatikus értékeléséhez. Első eredményünkben a beszélő álmosságának észlelésére alkalmaztuk. Az x-vektor jellemzőkinyerő modellt több korpuszon tanítottuk, és kísérleteztünk azzal is, hogy a tanítás során zajt adjunk a felvételekhez illetve visszhangosítsuk azokat. A nyilvánosan elérhető Dusseldorf Sleepy Language Corpuson saját x-vektor beágyazásaink mint jellemzők használatával (lineáris SVR mint regressziós módszer alkalmazásával) stabilan versenyképes teljesítményt értünk el. Módszereink az említett korpuszon a legmagasabb Spearman-féle korrelációs együtthatót érték el azon megközelítések közül, melyek egyetlen módszert (és nem különböző eljárások kombinációját) használták. A fejezetben emellett bemutattuk kísérleti eredményeinket saját és előtanított x-vektor jellemzőkinyerők használatával depresszió szintjének becslésére beszédjelből. Kísérleteink megerősítették, hogy az x-vektorok képesek olyan beszélőjellemzők megragadására, amelyek információt tartalmaznak a depresszió hatékony kimutatására. Megmutattuk, hogy az x-vektor modell nyelve másodlagos jelentőséggel bír a használt keretszintű jellemzőkészlethez képest. Nevezetesen, a legjobb modellünk, amelynek AUC pontszáma 0,940 és RMSE értéke 9,54 volt, keretszintű log-energiákon (FBANKs) lett tanítva MFCC-k helyett. Végezetül adtunk egy x-vektoros jellemzőkre épülő ensemble osztályozót is, konfliktus-kibontakozás becslésére, illetve főemlősfajok hangból

történő azonosítására. A konfliktus-kibontakozási feladatban a végső predikciók javítása érdekében az SSPNet Conflict korpuszt is fölhasználtuk, mellyel az adott feladaton a legmagasabb pontosságértéket értük el.

4. téziscsoport

Végül a 6. fejezetben bemutatunk egy időbeli beszédjellemző-készletet, amely az artikulációs sebességből, a beszédtempóból és további, az alany hezitációját leíró jellemzőkből áll. Megmutattuk, hogy ezek a reprezentációk megbízható módon kinyerhetők az alkalmazott beszédfelismerő rendszer nyelvétől függetlenül: az enyhe kognitív zavar fölismerésének pontossága csak kismértékben függött attól, hogy egynyelvű (67-92%) vagy keresztnyelvi (67-92%) jellemzőkinyerést végeztünk-e. Ezen felül bevezettünk egy, a korábban már bemutatott temporális beszédparaméterekhez hasonló jellemzőkészletet, amely továbbra is a néma és kitöltött szünetek mennyiségét számszerűsíti, de kiszámításához nem szükséges egy teljes fonetikai szintű beszédfelismerő rendszer használata. Megmutattuk, hogy ez az eljárás, amely közvetlenül a HMM/DNN hibrid modell akusztikai DNN modelljének keretszintű kimenetét használja bemenetként, hasonlóan informatív jellemzőket képes kinyerni a beszédből az enyhe kognitív zavar és az Alzheimer-kór felismerésére, mint a beszédfelismerésalapú, temporális paraméterek kinyerését végző megközelítés.

A disszertáció tézisei

Az **első téziscsoportban** a hozzájárulások az Alzheimer-kór beszédjelből történő automatikus szűrésével kapcsolatosak, i-vektor jellemzők használatával. A részletes kifejtés a 3. fejezetben található.

I/1. Hozzájárulásom az i-vektor modellek tanítására és az Alzheimer-kórral élők beszédéből jellemzők kinyerésére vonatkozott. Megmutattam, hogy az i-vektoros jellemzők képesek értelmes beszédtulajdonságok tárolására ilyen jellegű beszéd esetén.

I/2. Az alanyok beszédfelvételeire i-vektorokat alkalmaztam a klinikai depresszió szintjeinek automatikus szűrése céljából. Az eredményeim alapján ez a módszer hasonló eredményeket ér el, sőt versenyképes teljesítményt tesz lehetővé az ugyanezen témában végzett korábbi tanulmányokhoz képest.

A **második téziscsoportban** hozzájárulásaim a Parkinson-kór automatikus detektálásához, a konfliktusok kibontakozásának szintjeinek és főemlősfajok hangból történő meghatározásához, valamint a beszélő megfázásának automatikus detektálásá-

hoz kapcsolódnak, Fisher-vektor megközelítéssel modellezett beszédjellemzők segítségével. A részletes kifejtés a 4. fejezetben található.

II/1. Kidolgoztam egy keretrendszert a Parkinson-kór automatikus értékelésére Fisher-vektor jellemzőkészlet használatával. Eredményeim azt mutatták, hogy az ilyen jellegű jellemzők nemcsak a képekből képesek értelmes információt rögzíteni (ahogyan eredetileg szánták őket), hanem a beszédjelből is.

II/2. Olyan gépi tanulási modellt építettem, amely képes megkülönböztetni a megfázott alany beszédét az egészséges alanyétól, Fisher-vektor jellemzőkészlet használatával. Megmutattam, hogy az XGBoost osztályozó eljárás pontosabb osztályozást tesz lehetővé a lineáris SVM technikával szemben, amikor az említett jellemzőket alkalmazzuk megfázás automatikus detektálására.

II/3. A tudományos cikk keretében végzett kísérleti lépések részeként modelleztem a Fisher-vektorok segítségével az egyének beszédének eszkalációs szintjeit; továbbá ugyanezt a technikát alkalmaztam a főemlősfajok hangjainak jellemzőinek kinyerésére. Bebizonyítottam, hogy egy ilyen megközelítés hatékony a szóban forgó feladatok automatikus végrehajtása esetén.

II/4. Megterveztem és megvalósítottam egy FV beszédjellemzők kinyerésén és felhasználásán alapuló, a megfázott alanyok detektálására szolgáló workflow-t. Megmutattam, hogy az ilyen típusú jellemzők képesek a megfázott betegek beszédének hatékony modellezésére.

A **harmadik téziscsoportban** a hozzájárulások az álmosság szintjének, a klinikai depresszió mértékének, a depresszió szintjének, a konfliktus-kibontakozások szintjeinek, valamint a főemlősfajok audió-alapú meghatározásához kapcsolódnak. A részletes kifejtés az 5. fejezetben található.

III/1. Mély neurális hálókból kinyert beágyazások használatát javasoltam a beszélő álmossági fokának automatikus módon történő meghatározására. Megmutattam, hogy az x-vektorok, melyeket eredetileg a beszélő meghatározására fejlesztettek ki, alkalmasak a beszélők álmosságának nagy pontosságú modellezésére.

III/2. Egyedi x-vektor jellemzőkinyerő modelleket használtam a klinikai depresszió mértékének a betegek beszédéből történő becslésére. Számos x-vektor modell tanításával megmutattam, hogy egy egyszerű struktúrájú workflow is képes felülmúlni számos bonyolultabb módszer teljesítményét, mint például az ensemble osztályozók vagy osztályozó-kombinációk.

III/3. A tanulmányhoz való hozzájárulásom egy része különböző egyéni x-vektor jellemzőkinyerő modellek tanításából állt. Megmutattam, hogy ezek a mély neurális hálózati beágyazások versenyképes teljesítményt mutatnak mind a konfliktus-kibontakozások, mind a főemlős-fajok audió-alapú osztályozásában.

A **negyedik téziscsoportban** hozzájárulásaim az időbeli beszédparaméterek mint eszélőjellemzők alkalmazásával kapcsolatosak az enyhe kognitív zavar és az Alzheimer-kór vizsgálatában. A részletes kifejtés a 6. fejezetben található.

III/1. A fő hozzájárulásom a tanulmányhoz az időbeli beszédparaméterek generálása volt egy beszédfelismerő rendszer fölhasználásával, keretszintű megközelítésben. Megmutattam, hogy nem szükséges a teljes beszédfelismerő rendszer használata ahhoz, hogy hasonlóan jó minőségű jellemzőkhöz jussunk, mint a teljes ASR-rendszer használatával meghatározott temporális beszédparaméterek esetében (enyhe kognitív zavar és Alzheimer-kór szűrése esetén).

III/2. A tanulmányban való részvételem korlátozott volt, mivel nem én voltam a fő közreműködő. Hozzájárulásom az időbeli beszédparaméterek kiszámítása volt. Ez a tanulmány azt mutatta meg, hogy a beszédfelismerő rendszer nyelve csak kismértékben befolyásolja az enyhe kognitív zavar automatikus meghatározásának teljesítményét, csökkentve a nyelvfüggő beszédfelismerő rendszer használatának szükségességét.

Publications

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.

Acknowledgments

My apologies to whoever reading this section finds inconvenience with the language employed...

Quizá los individuos presentes en el área más abarrotada de la campana de Gauss encuentren engorrosa la manera en la que los senderos de esta travesía se construyen. Aquel bulevar al que ellos describen como pavoroso, mortífero, y hasta venenoso, para este escritor, por el contrario, se ha convertido en una travesía que además comprende de algarabía, dichas, y hasta hilaridad. Un recorrido en el que cada *valor de entrada* se suma a la construcción de la versión definitiva del individuo en la consumación de aquella travesía.

Este conjunto de letras dictadas por el alma, catalizadas con corazón y mente, y plasmadas por manos vehementes, van para el Creador, mi familia, y para mi eterna compañera de amor. A mi Madre y a mi abuela Carlota especialmente, que me han dado fuerzas y valentía, ¡para no caer nunca! No hace falta recalcar que esta dedicatoria se difunde a esos amigos que están latentes en la eternidad.

Dicha gratitud se extiende, asimismo, a quien tuvo el rol de guía académico en este camino ambivalente, mi supervisor; a Ecuador, Magyarországra.

Köszönöm szépen.