POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Machine learning-based analysis of spontaneous speech to detect and monitor decline of cognitive function in elderly people

Laurea Magistrale in Biomedical Engineering - Ingegneria Biomedica

Author: Chiara Giangregorio

Advisor: Prof.ssa Simona Ferrante

Co-advisors: Prof.ssa Emilia Ambrosini, Ing. Eugenio Lomurno

Academic year: 2020-2021

1. Introduction

Dementia is a category of neurodegenerative disease that entails a long-term and usually gradual decrease in cognitive functioning. It is characterized by a set of symptoms including memory loss, thought difficulties, poor executive functions (e.g. problem-solving, decision-making), language impairment, motor problems, and emotional distress. Current diagnostic procedures require a thorough examination by medical specialists, which are too cost- and time-consuming to be provided on a large scale. Since speech and language capacity is a well-established early indicator of cognitive deficits, including dementia, speech processing methods offer great potential to automatically screen for prototypical indicators in real-time. In recent years, voice has been one of the most studied digital biomarkers [1]. It is widely employed since it allows an ecological and rapid assessment of several aspects linked to the health status of a subject, such as the respiratory system, cognitive decline, emotions, and heart dysfunctions that would usually be assessed in presence of clinicians, thus allowing these trials to be carried out during everyday activities [1]. Concerning the cognitive functions, several acoustic parameters have been proposed in the literature as potential indicators of decline. The most popular ones are those described by:

- the *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS)
- the *emobase* feature set
- the *ComParE* feature set

These feature sets comprise several Low-Level Descriptors (LLD), e.g. pitch, MFCC, Loudness, jitter, and several related statistics, extracted via the openSmile toolbox. Feature extraction is usually carried out via processing of audio recordings of vocal signals obtained through the accomplishment of several tasks, such as interviews, movie recalls, and day descriptions, the most common one being the Picture description task, because of its great test-retest reliability. Moreover, it overcomes some difficulties arousing from memory, enabling even the subjects with severe memory loss to effectively carry out the task.
There exists a variety of pictures to be administered, the most employed being the "Cookie Theft" of the Boston Diagnostic Aphasia Examination test.
The use of voice to detect mental disease and cognitive decline from the analysis of acoustic

features has been eased by the advent of artificial intelligence. Indeed, on recognition of cognitive decline from spontaneous speech, several works have been reported in literature. In [2], Authors analysed the temporal parameters of reading fluency to discriminate between Spanish-speaking asymptomatic subjects and those with Alzheimer's Disease (AD). The algorithms applied to the recordings were capable of differentiating between AD patients and controls with an accuracy of 80% (specificity 74.2%, sensitivity 77.1%) based on speech rate.
Moreover, in [3] it has been demonstrated that it is possible to differentiate between several kinds of dementia and Mild Cognitive Impairment both in binary and multiclass scenarios, through free speech tasks with high classification accuracy. In [4], it was showed that acoustic parameters such as speech rate, hesitation ratio, number of pauses and articulation rate yield significant results in the discrimination between MCI and healthy subjects in the movie recall task, obtaining an F1-score of 78.8%. Moreover, in [5] Authors were able to discriminate between controls and MCI subjects with Random Forest and Support Vector with a high F1-score of around 75% with the nested-leave-one-subject-out cross-validation.
The current work focuses on the analysis of acoustic features of speech by means of machine learning techniques to support the early identification of decline of cognitive functions. The long-term goal is to develop a mobile app for large-scale remote monitoring, hence there is the need to automatically extract the acoustic features. For this reason, features must be computed on smaller time-lengths, i.e. at most on 15s, to reduce computational cost, and hence avoiding storage of recordings. Therefore, the same features have been computed in Matlab at different time scales of 5-10-15 seconds. The current approach has been tested on two datasets of Latin languages, Italian and Spanish, which come from a previous European project to monitor and contrast decline and social exclusion in elderly people.
Therefore, the main contribution of this work can be summarized in:

- Optimization of feature extraction algorithm with the addition of new features
- Evaluation of optimal duration of speech segments for feature extraction
- Models evaluation via nested 10-fold cross-validation

2. Methods

To classify cognitive decline, the feature extraction code was optimized from a previous work [6] with the introduction of new features. Moreover, while in the aforementioned work, features were computed on the whole length of the audio recordings, in the current work different time segments (5-10-15s) have been taken into account to analyze the dependence of features with time and to find the optimal duration length for features extraction.

2.1. Dataset

Prediction of cognitive decline was carried out on two already existing datasets of Latin-derived-language speaking subjects (Italian and Spanish). The two datasets are composed of 153 Italian and 150 Spanish-speaking subjects, respectively. Participants have been divided into three groups based on the Mini-Mental State Examination (MMSE), a test which is extensively used in clinical and research settings to evaluate cognitive impairment, according to the achieved score:

- Group 1: healthy subjects (MMSE$>$ 26)
- Group 2: mild cognitive impairment ($20 \leq$ MMSE ≤ 26)
- Group 3: severe cognitive impairment (MMSE$<$ 20)

Each participant carried out the following tasks, for a total of 4 recordings:

- 3 story-telling tasks
- Picture description task

For the first 3 tasks, subjects were asked to tell three short stories in an interrupted way for approximately two minutes each: a positive and a negative story, and an episodic one, in neutral tone. To avoid the inclusion of depressed subjects, those from Group 1 filled out the Geriatric Depression Test (GDS) which is a self-report assessment used to identify depression in the elderly population. Those with a GDS score over 9 are considered mildly depressed and a score over 20 corresponds to severe depression. Therefore, only the subjects with scores below the threshold of mild depression have been included. A summary of the composition of the two datasets is shown in Table 1.

Table 1: Dataset Composition

Dataset		Group 1	Group 2	Group 3
ITALY	Numerosity	45	44	44
	Age(years)*	76,6 (4,9)	82,8 (4,6)	86 (5,7)
	Men/Female	6/39	11/33	7/37
	Years of education*	12,4 (3,6)	8,7 (4,3)	7,4 (4,5)
	MMSE*	29 (1)	24 (2)	16 (3)
SPAIN	Numerosity	43	45	45
	Age(years)*	79,7 (7,5)	82,4 (6,9)	85,5 (6,6)
	Men/Female	22/21	9/36	17/28
	Years of education*	6 (4,2)	5 (3,2)	6,3 (3,9)
	MMSE*	28 (1)	23 (2)	7 (2)

*. Mean (standard deviation)

2.2. Optimization of Matlab algorithms for features extraction

Changes in machine learning performances when extracting features at different time scales were analyzed, to find the optimal time interval. indeed, the former code [6] has been optimized by allowing segmentation of audio recordings in smaller chunks. Therefore, features were computed on segments of pre-defined length, 5, 10 and 15s. Moreover, further analysis has been carried out, by considering datasets in which the features computed at different scales were considered altogether (5-10-15s). Finally, the new Matlab Audio Toolbox allowed to add the following features to the original set:

- **Pauses**: The number of pauses and their mean duration inside each segment has been computed with the function *detectSpeech*. The function uses a thresholding algorithm based on energy and spectral spread on each frame to highlight the indices corresponding to the boundaries of speech signals.
- **Spectral centroid**: it represents the center of 'gravity' of the spectrum
- **Mel-Frequency Cepstral Coefficients (MFCC)**: they are a type of cepstral representation of the signal, where the frequency bands are distributed according to the mel-scale, instead of the linearly spaced approach. They are the coefficients making up a Mel-Frequency Cepstrum (MFC), a representation of a short-term power spectrum of a sound [7].
- **Speech temporal regularity**: it is computed from the first 16 coefficients of the Mel-Frequency Spectrum (MFCC), capturing the temporal structure of speech.

The final set of acoustic features computed by the Matlab algorithm is shown in Table 2, with their trends in subjects with dementia with respect to the healthy values.

Table 2: Summary table of extracted features

Type	Feature	Description	Trend with dementia
Voice periodicity	Unvoiced percentage	Percentage of aperiodic parts in the audio segment	Increasing
	Voiced and unvoiced parts	Mean, median, 15 and 85 percentile of the parts of the signal with and without periodic nature	Decreasing voiced and Increasing unvoiced
	Pitch	Contour of fundamental frequency F0	Decreasing
	Shimmer	Random cycle-to-cycle temporal changes of the amplitude of the vocal fold vibration	Decreasing
Glottal pulses	Total voice breaks	Percentage of distances between consecutive glottal periods	Increasing
Formants	Standard deviation of the 3rd formant	degree of tonal modulation of the voice	Increasing
Syllables	Speech Rate	Number of syllables per second	Decreasing
	Phonation Percentage	Percentage of syllables throughout the speech signal	Decreasing
	Articulation Rate	Number of syllables over the phonation time	Decreasing
	Intersyllabic time	Duration between syllables	Increasing
	Intrasyllabic time	Duration of syllables	Increasing
Pauses	Pauses	Number of pauses for each audio segment	Increasing
	Duration of pauses	Mean duration of pauses for each audio segments	Increasing
Spectral features	MFCC	First 16 Mel-Frequency Cepstral Coefficients	Decreasing
	Speech Temporal Regularity	Temporal structure of speech segments	Decreasing
	Centroid	Location of the center of mass of the spectral signal	Increasing

2.3. Data Pre-processing

To reduce noise among the segments, for each feature computed on the specified segment mean and standard deviation or median and interquartile range have been computed over the recordings, depending on whether the distribution was normal or not. In this way, each subject was represented by a single entry in the final dataset. Therefore, a normality distribution check for each feature was carried out through Matlab by performing the Anderson-Darling test (*adtest*). Finally, to reduce computational cost and improve models' performances, data normalization was performed in Python, by rescaling each feature in the range [0,1].

2.4. Classification

Classification was evaluated with the following two strategies: binary and multiclass classification. Binary classification was performed to distinguish between Group 1 (no cognitive decline) and Group 2 (Mild cognitive decline), to evaluate the capability of the models to detect early stages of dementia. Multiclass classification was implemented to discriminate between the 3 groups.

To perform classification, the following standard Machine Learning classifiers were applied:

- Logistic regression (LR)
- Support Vector Machines (SVM)
- CatBoostClassifier (CATBOOST)

Each classifier was implemented in Python using the Scikit-learn libraries and catboost library. The performances of the different classifiers were compared in terms of F1-score, since it summarizes the contribution of precision and recall, as:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

Moreover, accuracy, recall, and precision were measured on the test set. To analyze the contribution of the features, SHAP, a method based on cooperative game theory, was employed at the end of each classification task, through summary plots, which combine feature importance with feature effects. Indeed, each point of the plot corresponds to a SHAP value for a feature and an instance, whereas the color represents the value of the feature from low (blue) to high (red). Moreover, features are ordered based on their importance on the y-axis in descending order. Finally, the position along the x-axis gives indications of the impact on the prediction.
Due to the relatively small number of subjects, to obtain more robust testing, both classification and regression algorithms were validated with nested 10-fold cross-validation procedure.

3. Results

3.1. Binary Classification

F1-scores obtained by the models in binary classification when dividing subjects between group 1 (MMSE> 26) and group 2 (20 ≤MMSE≤ 26) are shown in Table 3, respectively for the Italian and Spanish datasets. For the best models, accuracy, precision, recall, and F1-score on the test set are shown in Table 4. Moreover, feature rankings with SHAP are shown in Figures 1 and 2.

Table 3: F1-score (± standard deviation) on Validation Set of Italian and Spanish Datasets

	Time Scale (s)	CATBOOST	SVM	LR
ITALY	**5**	0,76 (±0,03)	0,71 (±0,03	0,71 (±0,04)
	10	0,76 (±0,03)	0,67 (±0,04)	0,69 (±0,04)
	15	**0,77** (±0,03)	0,63 (±0,05)	0,68 (±0,04)
	5-10-15	0,74 (±0,03)	0,68 (±0,05)	0,69 (±0,03)
SPAIN	**5**	0,74 (±0,03	0,62 (±0,04	0,66 (±0,04)
	10	0,74 (±0,03)	0,62 (±0,04)	0,62 (±0,03)
	15	**0,76** (±0,03)	0,55 (±0,06)	0,60 (±0,05)
	5-10-15	0,72 (±0,03)	0,60 (±0,05)	0,61 (±0,05)

Table 4: Performance metrics on test sets of Italian and Spanish Dataset CatBoost, best model

	Time Scale (s)	Accuracy	Precision	Recall	F1-score
ITALY	15	0,71	0,74	0,63	0,66
SPAIN	15	0,69	0,72	0,70	0,69

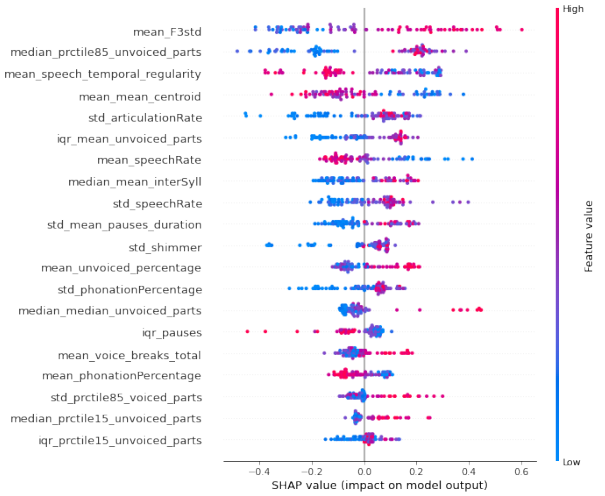

Figure 1: *Feature ranking of binary classification - Italy*

3.2. Multiclass Classification

F1-score for multiclass classification models is shown in Table 5 for the Italian and Spanish datasets. For the best models, accuracy, precision, recall, and F1-score on the test set are shown in Table 6. Moreover, to have a better view of how predictions are distributed among the 3 classes, the confusion matrices on the test sets are shown in Figures 3 and 4, respectively for the Italian and Spanish datasets.

Table 5: F1-score (± standard deviation) for multiclass classification on the validation set of Italian and Spanish datasets

	Time Scale (s)	CATBOOST	SVM	LR
ITALY	**5**	0,63 (±0,02	0,50 (±0,03	0,51 (±0,03)
	10	0,63 (±0,02)	0,49 (±0,04)	0,51 (±0,03)
	15	0,62 (±0,02)	0,48 (±0,02)	0,51 (±0,02)
	5-10-15	**0,64** (±0,03)	0,50 (±0,05)	0,52 (±0,04)
SPAIN	**5**	0,62 (±0,02	0,50 (±0,03	0,54 (±0,02)
	10	0,61 (±0,03)	0,50 (±0,03)	0,51 (±0,02)
	15	**0,63** (±0,03)	0,49 (±0,03)	0,50 (±0,02)
	5-10-15	0,62 (±0,03)	0,49 (±0,05)	0,50 (±0,05)

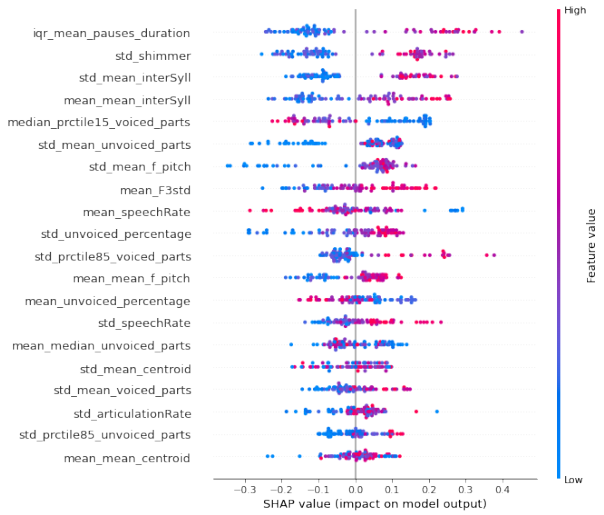

Figure 2: *Feature ranking of binary classification - Spain*

Table 6: Performance metrics for multiclass classification on test sets of Italian and Spanish Dataset with best model - CatBoost

	Time Scale (s)	Accuracy	Precision	Recall	F1-score
ITALY	5-10-15	0,66	0,68	0,66	0,63
SPAIN	15	0,55	0,56	0,54	0,53

4. Discussion

Similar results were achieved regardless of the time scale used for the computation of features. This result demonstrates that there is no need for long audio recordings and it allows to speed up computational time. Results from SHAP in Figure 1 and 2 highlighted that different sets of features are relevant depending on the considered language. Overall, an increase in speech rate is noticed in healthy subjects in the two datasets. The addition of new features such as *speech temporal regularity*, mainly for the Italian dataset, and *mean duration of pauses*, for the Spanish dataset, highly contribute to the detection of cognitive impairment. Indeed, as stated in literature, regarding speech temporal regularity, in normal voices, the duration of the contiguous speech segments tends to be longer and more "regular", resulting in higher values of the first 16 MFCCs. Conversely, for AD cases, the duration of the contiguous speech segments tends to be shorter and less "regular", which usually results in lower average values of the 16 MFCCs [8]. For the Italian dataset, the most discriminative features are the standard deviation of the 3^{rd}

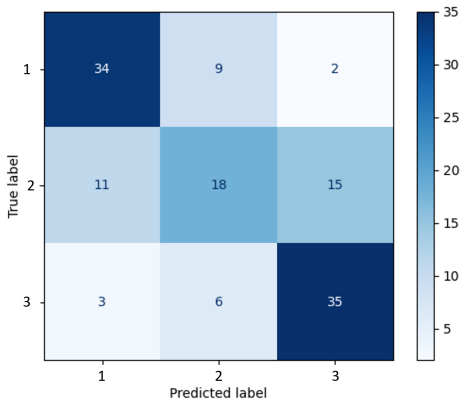

Figure 3: *Confusion matrix of multiclass classification on test set - Italy*

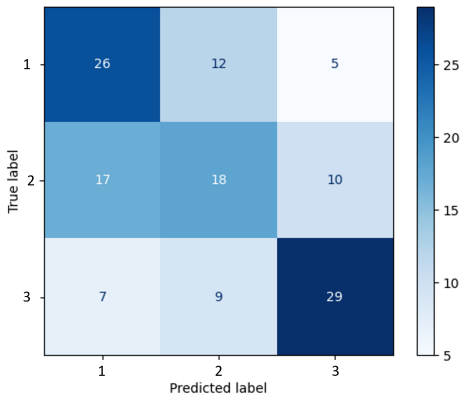

Figure 4: *Confusion matrix of multiclass classification on test set - Spain*

formant (the tonal modulation of the voice), the number of voiced parts, temporal regularity, and the speech and articulation rates. In particular, lower values of F3 and articulation rate suggest that a subject has no cognitive impairment. On the contrary, low values of speech temporal regularity suggest that the subject has mild cognitive impairment, which is in line with the literature. For the Spanish dataset, the most informant features are the variation of the length of pauses, the duration between syllables, and the variation of shimmer (changes of the amplitude of vibration of the vocal fold). Low values of these features suggest a preserved cognitive function, whereas higher values are noticed in subjects with mild cognitive impairment. For all the classification tasks, both for the Italian and Spanish datasets, CatBoostClassifier is the one that performs better in terms of F1-score on the valida-

tion sets. In particular, in binary classification, for the Italian dataset it achieves an F1-score of 77% (66% on the test set), while on the Spanish dataset, CatBoost yields an F1-score of 76% on the validation set (69% on the test set). From these results, the algorithm seems to overfit since there is a worsening of the metrics from validation to test set. In multiclass classification, the models mainly struggle to clearly detect subjects with mild cognitive impairment. As a matter of fact, in the Italian case the difficulty arises when discriminating between mild and severe cognitive decline, whereas for the Spanish dataset, the struggle is between the healthy and the mildly impaired subjects.

Results for the Italian dataset by considering only smaller segments for feature extraction have obtained comparable performances to those in [6] that evaluated instead the whole recordings of more than two minutes. Moreover, in this work, with nested 10-Fold CV, it was possible have an estimate on how the model performs on unseen data, whereas in the aforementioned work the Authors have only validated results with standard 5-fold Cross-Validation, without estimating the model performance on unseen data. With respect to [5], F1-score in binary classification was lower, but, in the case of the current work it was obtained without considering demographic features such as age and years of education, important indicators of cognitive decline, to evaluate the possibility to employ acoustic features for longitudinal monitoring. Regarding the Spanish dataset, classification was performed on a larger dataset than the one in [2], obtaining slightly worse performances on the binary classification. Still, SHAP analysis in Figure 2 confirmed that fluency is an important aspect of the evaluation of cognitive decline from spontaneous speech. The good performances obtained are promising for the development of an application for longitudinal monitoring of cognitive decline. Still, the models seem to overfit since performances on test set worsen. The problem may be the lack of generalization power of the model due to the high number of features, therefore it would be useful to implement a feature selection algorithm to keep only the most significant ones. For example, it would be interesting to implement a recursive feature elimination algorithm based on SHAP feature ranking.

In conclusion, the results of this work show that the extracted acoustic features from spontaneous speech provide a good discrimination power between healthy subjects and those with signs of cognitive decline, regardless of the spoken language. Furthermore, the analysis on the duration of segments suggests that it is feasible to design a mobile app for the extraction of acoustic features in real-time. Indeed, the use of small-time segments allows to compute the features in a faster way, thus this work represents a first step towards the implementation of applications for monitoring cognitive impairment in everyday activities, without directly involving clinical assessments and visits.

Further improvements would be to evaluate and predict the emotional state of the subjects since previous studies have widely shown the influence of emotion on acoustic features and the employment of deep learning methods for feature extraction.

References

1. Robin, J. *et al.* Evaluation of Speech-Based Digital Biomarkers: Review and Recommendations (2020).
2. Martínez-Sánchez, F. Oral reading fluency analysis in patients with Alzheimer disease and asymptomatic control subjects (2013).
3. König, A. *et al.* Use of Speech Analyses within a Mobile Application for the Assessment of Cognitive Impairment in Elderly People (2017).
4. Tóth, L. *et al.* A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech (2017).
5. Calzà, L. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia (2021).
6. Ambrosini, E. *et al.* Automatic speech analysis to early detect functional cognitive decline in elderly population (2019).
7. Xu, M. *et al. HMM-based audio keyword generation* in (2004).
8. Satt, A. *et al.* Evaluation of Speech-Based Protocol for Detection of Early-Stage Dementia (2013).