Predicting Mini-Mental Status Examination Scores through Paralinguistic Acoustic Features of Spontaneous Speech

Citation for published version:

Digital Object Identifier (DOI):
10.1109/EMBC44109.2020.9175379

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Predicting Mini-Mental Status Examination Scores through Paralinguistic Acoustic Features of Spontaneous Speech

Ziyang Fu¹, Fasih Haider¹, Saturnino Luz¹

Abstract—Speech analysis could provide an indicator of cognitive health and help develop clinical tools for automatically detecting and monitoring cognitive health progression. The Mini Mental Status Examination (MMSE) is the most widely used screening tool for cognitive health. But the manual operation of MMSE restricts its screening within primary care facilities. An automatic screening tool has the potential to remedy this situation. This study aims to assess the association between acoustic features of spontaneous speech and assess whether acoustic features can be used to automatically predict MMSE score. We assessed the effectiveness of paralinguistic feature set for MMSE score prediction on a balanced sample of DementiaBank’s Pitt spontaneous speech dataset, with patients matched by gender and age. Linear regression analysis shows that fusion of acoustic features, age, sex and years of education provides better results (mean absolute error, MAE = 4.97, and $R^2 = 0.261$) than acoustic features alone (MAE = 5.66 and $R^2 = 0.125$) and age, gender and education level alone (MAE of 5.36 and $R^2 = 0.17$). This suggests that the acoustic features of spontaneous speech are an important part of an automatic screening tool for cognitive impairment detection.

Clinical relevance We hereby present a method for automatic screening of cognitive health. It is based on acoustic information of speech, a ubiquitous source of data, therefore being cost-efficient, non-invasive and with little infrastructure required.

I. INTRODUCTION

Cognitive ability, also known as cognitive functioning or cognitive intelligence, refers to brain-based capabilities of processing and applying information. To be more specific, these capabilities involve learning, abstract thinking, reasoning, remembering, problem solving, decision making, attention, comprehending complex ideas, and so on ([11], [2], [3]. Cognitive impairment is a great threat to public health. According to Sachdev et Al., the overall prevalence of Mild Cognitive Impairment (MCI) ranges from 5.9% to 12% based on different diagnostic criteria [4]. Regarding the aspect of dementia, it is the fifth largest cause of death, with 2.4 million deaths (4.4% of total death) in 2016. Furthermore, the dementia population is expanding rapidly. In 2016, 43.8 million people were living with dementia, with the number doubling during the past 25 years.

The Mini Mental Status Exam (MMSE) is an examination for evaluating the cognitive status of patients, which consists of multiple test items originally grouped into “orientation, memory, attention, naming, follow verbal and written commands, write a sentence spontaneously, and copy a complex polygon” [5]. The performance in each test item is scored and these scores are added together. The total achievable score of MMSE is 30 points. Patients with a MMSE score of under 23 are conventionally considered to have cognitive impairment [6]. However, some studies have suggested a higher threshold (e.g. 26 point) may be better for sensitivity and specificity of identification of cognitive impairment [7], [8].

Changes in speech and language among people with cognitive impairment have been noticeable in patients with MCI and dementia [9], [10]. Appell et al. reported word finding to be difficult, represented by a high incidence of circumlocutions [11]. In addition, Ahmed et al. found that two-thirds of patients with MCI suffered various degrees of decline in connected speech [12]. Several studies provide evidence that speech rate (i.e. phonemes per second) and the amount of pausing during speech had a significant correlation with the severity of cognitive impairment both in narrative and dialogical speech [13], [14], [10], [15]. Furthermore, the findings of Bayles et al. suggested that impairment of language ability may have a linear declining trend and becomes severe with the development of cognitive impairment [16]. As a result, language impairment is considered an important symptom contributing to the diagnosis of cognitive impairment [17], [18].

Picture description is a common way to generate connected speech and is widely used in epidemiological studies to assess language ability. For example, The Cookie Theft picture has been a commonly used picture in many studies [12], [19]. A similar way to picture description, aiming at generating connected speech, is asking patients to recall and describe a past experience, e.g. one of their happiest experiences [20]. This group of tasks mainly focuses on semantic content or syntactic complexity, which have a good ability of capturing the global progression of linguistic impairment through successive screening [12]. Many studies based on picture description have suggested different words’ selection between patients with cognitive impairment and healthy controls. It was found that patients with cognitive impairment produce more meaningless content [21], [19], which can account for finding that patients need a larger number of words to describe a picture [22]. As well as the number of words, the speech produced by patients tends to contain a larger proportion of nouns, pronouns, and shorter and more commonly used words [21], [23]. However, language (content) based diagnosis have methodological limitations and do not always generalise well across languages [24], [25].

To our knowledge only a very few studies have tried to predict MMSE scores particularly using only the acoustic

¹ Z. Fu, F. Haider and S. Luz are with the Usher Institute, Edinburgh Medical School, The University of Edinburgh, UK. {z1818665, fasih.haider, s.luz}@ed.ac.uk
information. Only one study used acoustic features in combination with Lexicosyntactic and semantic features (along with manually transcribed content) [26], obtaining a MAE of 3.83 but no results are reported separately for acoustic features only. In addition to linguistic variables, Mendes et al. established a prediction equation with age, sex and education as independent variables. However, the accuracy of this prediction equation has not been reported. The result of this study showed that all three variables are important predictors, because 38% of the total variance of the MMSE scores can be explained by these three variables [27]. We have highlighted the global burden of cognitive impairment and the benefits of screening for cognitive impairments. An automatic screening tool based on speech might provide an economical and scalable screening method. Speech-based methods have shown good performance on Alzheimer dementia classification [15], [10], but only one study used speech-related features to predict the results of the MMSE [26], which is the most widely-used screening tool in primary care facilities. However the limitation of previous study [26] is that it relies on manual transcription of data.

The work presented in this paper contributes to research into MMSE prediction by evaluating and demonstrating the potential of acoustic features based on the ComParE feature set [28] and their fusion with age, gender and level of education, for automatic MMSE prediction. This is, to the best of our knowledge, the first empirical attempt to use acoustics feature sets as “digital biomarkers” for automatic MMSE score prediction.

II. EXPERIMENTATION

A. Dataset

The Pitt corpus was gathered longitudinally between 1983 and 1988 on a yearly basis as part of the Alzheimer Research Program at the University of Pittsburgh [29]. Participants are categorised into three groups such as dementia, control (i.e. healthy), and unknown. All participants were required to be above 44 years of age, have at least seven years of education, no history of nervous system disorders or be taking neuroleptic medication, have an initial MMSE score of 10 or more and be able to provide informed consent. Extensive neuropsychological and physical assessments conducted on the participants are also included [30].

The Pitt Corpus contains participants’ speech data collected by the Alzheimer and Related Dementia Study at the University of Pittsburgh School of Medicine on the following tasks: a picture description task in which the participant is asked to describe, verbally in their own words, a picture, a word fluency task, a story recall task, and a sentence construction task.

We specifically chose the picture (shown in Figure 1) description task sample for the present study, as it encompasses spontaneously generated narrative speech. The data include MMSE scores for control (cognitively normal) participants, patients diagnosed with MCI, and patients diagnosed with probable Alzheimer’s dementia (AD). The selected participants are matched for age and gender (Table I). The resulting dataset was segmented for voice activity using a signal energy threshold. We set the energy threshold parameter to 65dB with a maximum duration of 10 seconds for a speech segment. The segmented dataset contains 2033 speech segments from 82 non-AD subjects and 2043 speech segments from 82 AD subjects. The average number of speech segments produced by each participant in their descriptions was 24.86 (standard deviation sd = 12.84). Audio volume was normalized across all speech segments to control for variation caused by recording conditions, such as microphone placement.

<table>
<thead>
<tr>
<th>Age Interval</th>
<th>AD/MCI Male</th>
<th>Female</th>
<th>Control Male</th>
<th>Female</th>
<th>Mean (SD) MMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(50, 55)</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>27.8 (2.6)</td>
</tr>
<tr>
<td>(55, 60)</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td>23.8 (6.6)</td>
</tr>
<tr>
<td>(60, 65)</td>
<td>4</td>
<td>9</td>
<td>4</td>
<td>9</td>
<td>23.7 (7.3)</td>
</tr>
<tr>
<td>(65, 70)</td>
<td>10</td>
<td>14</td>
<td>10</td>
<td>14</td>
<td>24.7 (6.2)</td>
</tr>
<tr>
<td>(70, 75)</td>
<td>9</td>
<td>11</td>
<td>9</td>
<td>11</td>
<td>22.9 (7.4)</td>
</tr>
<tr>
<td>(75, 80)</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>23.78 (6.7)</td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
<td>46</td>
<td>36</td>
<td>46</td>
<td></td>
</tr>
</tbody>
</table>

Mean (SD) 19.9 (6.0) 18.3 (6.1) 29.0 (1.0) 29.1 (1.3) 24.0 (6.7)

B. Feature Extraction and Selection

We have extracted ComParE feature set from the speech segments. The ComParE 2013 [28] feature set includes energy, spectral, Mel-Frequency Cepstral Coefficients (MFCC), and voicing related Low-Level Descriptors (LLDs). LLDs include logarithmic harmonic-to-noise ratio, voice quality features, Viterbi smoothing for F0, spectral harmonicity and psychoacoustic spectral sharpness. Statistical functionals are also computed, bringing the total to 6,373 features. Pearson’s correlation test was performed on the whole dataset to remove acoustic features that were significantly correlated with duration (when $R > 0.2$). “Dummy variables” caused by homogeneity of LLD groups were also detected and removed. Detection of dummy variables employed the Variance Inflation Factor (VIF) to detect the multiple collinearity in a linear model. As a result, a linear model was established...
to detect the dummy features. The equation of this model is shown below:

$$F_i = \beta_0 + \sum_{i=1}^{n} \beta_i F_i + \epsilon \quad (1)$$

where $n$ is the number of features after removing duration related features, $F_i$ is the $i$th acoustic features after removing duration-related features, $\beta$ is the estimated regression coefficients, $\epsilon$ is the random error. The value of the VIF was also calculated. According to Bollinger, a rule of thumb is that multicollinearity exists when the VIF is larger than 10 [31]. In other words, a feature can be considered as a dummy feature when VIF is larger than 10.

**III. RESULTS AND DISCUSSION**

In this study, we repeat 55 times the hypothesis tests for all 52 selected acoustic features and 3 covariates (i.e. age, sex and years of education). As a result, 7 out of 52 acoustic features are considered to have a statistically significant association ($p < 0.001$, adjusted with Bonferroni correction) with MMSE scores. As shown in Table II, these features are all MFCC-related features that belong to the spectral LLD. This suggests MFCCs can be considered as the major group of features that are associated with MMSE scores. MFCCs are a set of acoustic features representing the timbre of a sound, and can be considered as a representation of sound quality. This means that MFCCs association with MMSE scores might be the association between the sound produced by participants and their MMSE scores. In other words, people with different cognitive statuses might have differences in their vocal timbre. More specifically, the features selected (Table II) describe functionals smoothed by a moving average filter (sma), comprising interquartile range between the 50% and the 70% percentiles (iqr2-3), percentage of time above 90% of range plus minimum (up-leveltime90), spectral flatness (flatness), root quadratic mean (rqmean), and quadratic regression coefficient 3 (qregc3).

**TABLE II**

<table>
<thead>
<tr>
<th>Feature</th>
<th>$r$</th>
<th>SE</th>
<th>t.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>mfcc_sma</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$[\text{iqr}2-3]$</td>
<td>-0.07</td>
<td>0.02</td>
<td>-3.46</td>
</tr>
<tr>
<td>$[\text{up-leveltime90}]$</td>
<td>-3.13</td>
<td>0.95</td>
<td>-3.31</td>
</tr>
<tr>
<td>$[\text{flatness}]$</td>
<td>4.72</td>
<td>1.16</td>
<td>4.07</td>
</tr>
<tr>
<td>$[\text{rqmean}]$</td>
<td>0.08</td>
<td>0.02</td>
<td>3.82</td>
</tr>
<tr>
<td>$[\text{qregc3}]$</td>
<td>0.03</td>
<td>0.01</td>
<td>4.89</td>
</tr>
<tr>
<td>$[\text{rqmean}]$</td>
<td>-0.15</td>
<td>0.04</td>
<td>4.00</td>
</tr>
<tr>
<td>$[\text{qregc3}]$</td>
<td>0.05</td>
<td>0.01</td>
<td>4.73</td>
</tr>
</tbody>
</table>

The result of the linear regression model that only includes acoustic features is $R^2 = 0.125$ (Table III), suggesting that the 52 acoustic features that were selected by LASSO account for 12.5% of the variance of MMSE scores. After including sex, age and education, $R^2$ increased to 0.261. This indicates that inclusion of sex, age and education can improve the prediction of MMSE. However, the value of $R^2$ is still low, indicating that only about one fourth of the variance of MMSE scores can be explained with the additional consideration of age, sex and years of education. The mean absolute error of the model with only acoustic value was 5.66. After including age, sex and years of education into the model the mean absolute error decreased to 4.97. To further contextualise these results, we classified all speech segments in relation to a clinical criterion for diagnosis of cognitive impairment, setting those with predicted MMSE < 23 as positive instances for cognitive impairment and taking the clinical diagnosis provided by clinicians as ground truth. Table III shows that 70.98% accuracy, 57% sensitivity and 82% specificity can be achieved when all features are used.
Yancheva et al. also predicted MMSE scores with speech-related features [26] using full dementia bank dataset with longitudinal observations. However we used a sub-corpora of dementia bank dataset without longitudinal observations to avoid bias toward age and gender in our results. Yancheva et al. reported a mean absolute error (MAE) of 3.83 in predicting MMSE. However, they employed lexicosyntactic and semantic features derived from manual transcription, rather than automatically extracted acoustic features as we used in our analysis. In [26], those linguistic features were the main features selected from a group of 477, with acoustic features typically not being among the most relevant. Therefore no quantitative results were reported for acoustic features.

Ambrosini et al. reported an accuracy of 80% while using acoustic (pitch, unvoiced duration, shimmer, pause duration, speech rate), age and educational level features for cognitive decline detection using an Italian dataset of an episodic story telling setting [2]. However, this dataset is less easily comparable to ours, as it is elicited differently, and is not age and gender balanced.

Greater accuracy in MMSE prediction can be obtained if medical imaging data are employed. Zhang et al., for instance, used whole-brain volumes as predictors, and the result of the prediction models with different statistical models showed 7% to 10% average error of prediction [32]. However, whole-brain volumetric data are obtained using magnetic resonance imaging (MRI) devices, which are not typically available in primary care. Therefore, this type of predictors is not helpful to establish an automatic screening tool.

At the other end of the spectrum of data availability, Mendes et al. established a generalized linear model based only on demographic data, namely: age, sex and education [27]. These three demographic variables together accounted for 38% of variance of the participants’ MMSE scores. In our model, we also included these three demographic variables, but, even together with acoustic variables, the variance that could be accounted for was still lower than 38%. The different results between the two studies might be because of the difference between the two studies’ population. However, both Mendes et al.s and our results support that age, sex and years of education are important predictors of MMSE scores.

As we have shown, the addition of automatically extracted speech data to demographic data improves MMSE prediction accuracy, and can therefore offer a scalable compromise to complement these easily available patient data for cognitive impairment screening.

A. Limitations

There is limited evidence of an association between the acoustic features and MMSE scores. In order to further interpret the association observed in our study, future research could focus on the relationship between acoustic features and each test item of the MMSE. Secondly, having been collected as part of a research programme that aimed to investigate the natural history of Alzheimer’s disease, our dataset contained an inadequate number of young participants. Therefore, future studies could test whether the association between acoustic features and MMSE scores is different among the young population. In our study, our prediction was based on a prediction equation, and its prediction performance required many assumptions, such as linear relationship, independent observation and homoscedasticity [33]. Methods that require fewer assumptions and provide greater generalisation capacity may improve upon the results of our current model [34]. However, no matter which method may be used in future studies, an informative dataset is the precondition of good performance of prediction. Fourthly, our study suggests the possibility of a speech-based automatic screening method for MMSE score prediction. Some researchers have reported the application of a smartphone to collect speech-related features [35]. As a result, future studies could test the feasibility of using a smartphone to automatically screen for cognitive impairments based on the speech-related features collected by the smartphone, and the feasibility of incorporating automatic screening of cognitive impairment into e-health services.

IV. CONCLUSION

Our results suggest that there is a significant association between acoustic features (mainly MFCCs) and MMSE scores. Although this association is not the dominant explanation for the variance of MMSE scores, acoustic features contribute greatly to MMSE score prediction when fused with age, gender and education level features.

In future work, we wish to extend this study further and apply it to spontaneous dialogue data, which we are currently collecting following the PREVENT-ED protocol [36]. PREVENT-ED participants are healthy adults with a comprehensive risk profile (genetics, cognitive assessments and family history of AD), imaging (PET, MRI) and biomarker data, as well as spontaneous speech, are collected longitudinally, thus offering a good opportunity for the investigation of possible relationships between acoustic features of spontaneous speech and clinical markers of cognitive changes.

ACKNOWLEDGMENTS

This research has received funding from the European Union’s Horizon 2020 research programme, under grant agreement No 769661, towards the SAAM project. We acknowledge B. MacWhinney (DementiaBank) for hosting and sharing the databases used in this research.

---

**Table III**

LINEAR REGRESSION ANALYSIS RESULTS: ACCURACY (ACC.), SENSITIVITY (SEN.) AND SPECIFICITY (SPE.) IN PERCENTAGE WITH RESPECT TO A CLINICAL CRITERION OF COGNITIVE IMPAIRMENT DETECTION (MMSE < 23) AND CLINICAL GROUND TRUTH USING AGE, GENDER AND EDUCATION LEVEL FOR SPEECH SEGMENTS.

<table>
<thead>
<tr>
<th>Feature</th>
<th>$R^2$</th>
<th>MAE</th>
<th>acc.</th>
<th>sen.</th>
<th>spe.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/G/EL</td>
<td>0.170</td>
<td>5.36</td>
<td>62.82</td>
<td>48.82</td>
<td>78.80</td>
</tr>
<tr>
<td>Acoustic</td>
<td>0.125</td>
<td>5.66</td>
<td>65.34</td>
<td>51.65</td>
<td>73.83</td>
</tr>
<tr>
<td>Fusion</td>
<td>0.261</td>
<td>4.97</td>
<td>70.98</td>
<td>57.59</td>
<td>82.13</td>
</tr>
</tbody>
</table>

---

REFERENCES


