Computer Speech and Language 75 (2022) 101377

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Automatic screening of mild cognitive impairment and Alzheimer's disease by means of posterior-thresholding hesitation representation[☆]

José Vicente Egas-López [a,*], Réka Balogh [b], Nóra Imre [b,1], Ildikó Hoffmann [c,d], Martina Katalin Szabó [a,e], László Tóth [a], Magdolna Pákáski [b], János Kálmán [b], Gábor Gosztolya [a,f,2]

[a] University of Szeged, Institute of Informatics, Szeged, Hungary
[b] University of Szeged, Department of Psychiatry, Szeged, Hungary
[c] University of Szeged, Department of Linguistics, Szeged, Hungary
[d] Hungarian Research Centre for Linguistics, ELRN, Budapest, Hungary
[e] Centre for Social Sciences, Computational Social Science, ELRN, Budapest, Hungary
[f] MTA-SZTE Research Group on Artificial Intelligence, ELRN, Szeged, Hungary

ARTICLE INFO

Keywords:
Mild cognitive impairment
Alzheimer's disease
Dementia
Automatic speech recognition
Feature extraction
Temporal speech parameters

ABSTRACT

Dementia is a chronic or progressive clinical syndrome, characterized by the deterioration of problem-solving skills, memory and language. In Mild Cognitive Impairment (MCI), which is often considered to be the prodromal stage of dementia, there is also a subtle deterioration of these cognitive functions; however, it does not affect the patients' ability to carry out simple everyday activities. The timely identification of MCI could provide more effective therapeutic interventions to delay progression, and to postpone the possible conversion to dementia. Since language changes in MCI are present even before the manifestation of other distinctive cognitive symptoms, a non-invasive way of early automatic screening could be the use of speech analysis. Earlier, our research team developed a set of temporal speech parameters that mainly focus on the amount of silence and hesitation, and demonstrated its applicability for MCI detection. However, for the automatic extraction of these attributes, the execution of a full Automatic Speech Recognition (ASR) process is necessary. In this study we propose a simpler feature extraction approach, 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 experimentally demonstrate that this approach, operating directly on the frame-level output of a HMM/DNN hybrid acoustic model, is capable of extracting attributes as useful as the ASR-based temporal parameter extraction workflow was able to. That is, on our corpus consisting of 25 healthy controls, 25 MCI and 25 mild AD subjects, we achieve a (three-class) classification accuracy of 70.7%, an F-measure score of 89.6 and a mean AUC score of 0.804. We also show that

[☆] This study was partially funded by the National Research, Development and Innovation Office of Hungary via contract NKFIH FK-124413, by grant NKFIH-1279-2/2020 of the Hungarian Ministry of Innovation and Technology, and by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program (MILAB).

[*] Corresponding author.
E-mail address: egasj@inf.u-szeged.hu (J.V. Egas-López).

[1] Nóra Imre was supported by the New National Excellence Program of the Ministry for Innovation and Technology, Hungary (ÚNKP-21-4-I) and by the Faculty of Medicine, University of Szeged (EFOP-3.6.3-VEKOP-16-2017-00009).

[2] Gábor Gosztolya was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences, and by the ÚNKP-21-5 New National Excellence Program by the Hungarian Ministry of Innovation and Technology.

https://doi.org/10.1016/j.csl.2022.101377
Received 1 March 2021; Received in revised form 5 November 2021; Accepted 7 March 2022
Available online 18 March 2022

this approach can be applied on simpler, context-independent acoustic states with only a slight degradation of MCI and mild Alzheimer's detection performance. Lastly, we investigate the usefulness of the three speaker tasks which are present in our recording protocol.

1. Introduction

Mild cognitive impairment (MCI) is a heterogeneous clinical syndrome, often considered as the transitional phase between normal cognitive aging and dementia (Petersen et al., 2014). Besides the heterogenity of its underlying pathology (Schneider et al., 2009), the outcome of MCI also varies across patients: symptoms of the condition may remit, fluctuate, but there is also a high risk of MCI to progress to dementia (Kaduszkiewicz et al., 2014). These progressive types of MCI are most of the time precursor conditions to Alzheimer's disease (AD), but they may be also due to vascular or other neurodegenerative diseases (Petersen, 2003).

MCI affects both memory- and non-memory-related domains. It is often characterized by the deterioration of memory, language, and problem-solving skills; however, in contrast to those of dementia, the cognitive impairments that occur in MCI are not severe enough to affect the patients' ability to carry out simple everyday activities (Petersen et al., 2014; Association, 2020). According to estimates, the prevalence of MCI ranges from 15% to 20% in individuals of 60 years and older, while the annual progression rate from MCI to dementia is between 8% and 15% (Petersen, 2016). MCI may be present up to 15 years before the clinical manifestation of dementia (Laske et al., 2015). This wide time window offers a chance to detect the subtle signs of cognitive impairment, which is crucial since it can provide an opportunity to reduce the rate of cognitive decline (Hahn and Andel, 2011). Despite this, a substantial percentage of MCI and even dementia cases remain undetected (Lang et al., 2017), highlighting the need for effective methods that can aid the screening of the disease.

Changes in language performance can act as an early and valuable indicator of MCI, since language-related alterations can appear before the manifestation of other distinctive cognitive symptoms (McCullough et al., 2018). Spoken language can reliably reflect cognition, as speech production requires the parallel functioning of several domains, which gradually deteriorate during the course of MCI (such as lexical-semantic abilities, memory and executive functions Beltrami et al., 2018). It has been shown that changes in language production are related to subclinical declines in memory, e.g. the fluency of spontaneous speech has been shown to deteriorate in the case of people with early MCI (Mueller et al., 2018). Compared to healthy controls, MCI patients tend to have a lower speech rate, and an increased number and length of hesitations (Szatlóczki et al., 2015). Speech contains an increasing amount of pauses and disfluencies with the progression of the disease (de Ipiña et al., 2018), attributable to the word retrieval difficulties of the patients (Szatlóczki et al., 2015). These characteristics can have a strong effect on the overall time course of the speech; therefore, analyzing the temporal aspects of speech allows the indirect investigation of cognition.

There are several ways that speech samples can be obtained from participants, which, being a relatively quick, non-invasive and cheap way of gathering data, is a promising method for MCI screening. In the most widely used methods, spontaneous (or unstructured) speech tasks are included. Speech samples can be obtained in different ways: by requesting the subjects for executing spoken tasks, e.g., reading, counting backwards, or sentence repeating (König et al., 2018; Fraser et al., 2019); and by inducing the subjects to perform unstructured or spontaneous speech. In the latter case, the speakers may be asked to do narrative recall tasks or asked to speak about a specific topic. For example, Beltrami et al. state that one of the main keys for the adequate discrimination of MCI patients from those of healthy controls relies on the indicators got from the spontaneous speech of the subjects; namely, when they are asked to talk about a given topic such as their previous day or their hobbies (Beltrami et al., 2018).

Research on the automatic screening of different types of dementia was carried out in several previous studies which utilized techniques taken from Natural Language Processing (NLP), Automatic Speech Recognition (ASR), and Speaker Recognition for their purposes. For instance, Natural Language Processing is frequently exploited for the screening of Alzheimer's Disease by means of the transcribed utterances of subjects (see e.g., Balagopalan et al., 2020; Yuan et al., 2020; Martinc and Pollak, 2020). Also, ASR methods were utilized for detecting diseases like aphasia (Fraser et al., 2013, 2014), mild cognitive impairment (Lehr et al., 2012) and Alzheimer's (Baldas et al., 2010; Satt et al., 2014). Likewise, former and current state-of-the-art approaches for speaker recognition (i-vectors Dehak et al., 2011 and x-vectors Snyder et al., 2018) have been largely applied for dementia screening tasks. E.g., linguistic features along with acoustic features (i.e., i-vectors) were employed by Weiner and Schultz for the classification of Alzheimer's using the speech of subjects (Weiner and Schultz, 2018). Also, x-vectors were utilized for the automatic screening of pathological speech related to Alzheimer's (Botelho et al., 2020; R'mani Haulcy, 2020). These types of techniques have produced high performances at capturing meaningful and specific speaker traits from the speech recordings.

In previous studies, our team developed a set of temporal parameters that characterize the hesitation contained in the spontaneous speech of the subjects (Hoffmann et al., 2010; Tóth et al., 2015; Gosztolya et al., 2016; Tóth et al., 2018; Gosztolya et al., 2019). 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. Mattys et al., 2005; Fraser et al., 2013; Igras-Cybulska et al., 2016; Al-Ghazali and Alrefaee, 2019; Sluis et al., 2020). 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 relying on a phone-level Automatic Speech Recognition framework.

In this study we propose a feature set which, similarly to our team's previous studies, describes the amount of hesitation in the spontaneous speech of the subject. However, instead of using an Automatic Speech Recognition system and analyzing its output

Table 1
The instructions to the patients when recording the three speech recordings.

(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?*”

as we did before (see e.g. Tóth et al., 2018; Gosztolya et al., 2019), we will focus directly on the frame-level output of the Deep Neural Network acoustic model. This, in contrast with our previous studies, has the advantage that we do not need the overhead of an ASR decoder, but we can perform the subject classification step right after evaluating the acoustic neural network. According to our experimental results, this approach yields the same (or even better) classification performance as the original version of the S-GAP test, while it is more resource-efficient.

The rest of our paper is structured as follows. In Section 2, we describe the dataset we used in our experiments. Then, in Section 3, we explain the technique employed for the feature extraction phase, i.e., the Posterior-Thresholding Hesitation Representation (PTHR) approach. In Section 4, the experimental setup is described. Namely, the details of the training of the DNN acoustic model, the way PTHR was configured, and an explanation of the evaluation metrics employed. Then in Section 5 we present the results of the experiments. Lastly, in Section 6, we perform an analysis of the various speaker tasks which constitute our recording protocol.

2. Data

Our speech clips 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 Folstein et al., 1975), the Clock Drawing Test (CDT Freedman et al., 1994) and the Alzheimer’s Disease Assessment Scale – Cognitive Subscale (ADAS-Cog Rosen et al., 1984). In the case of MCI, Petersen’s criteria (Petersen et al., 1999), while for AD, internationally used guidelines (McKhann et al., 2011) were followed. The possibility of depression was assessed using the 15-item version of the Geriatric Depression Scale (GDS Yesavage and Sheikh, 1986): 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. Taler and Phillips, 2008; Roark et al., 2011; Satt et al., 2014)). 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 . For more details about our experimental setup for recording, see the study of Hoffmann et al. (2010). 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; these recordings were not split any further. 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 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

Table 2
Demographic data (i.e. age and education) and the results of the MMSE, CDT and ADAS-Cog tests of the three subject groups. (HC = Healthy Control, MCI = Mild Cognitive Impairment, mAD = Mild Alzheimer's Disease; MMSE = Mini-Mental State Examination; CDT = Clock Drawing Test; GDS = Geriatric Depression Scale).

	Subject groups			Test statistics	
	HC (n = 25)	**MCI** (n = 25)	**mAD** (n = 25)		
Gender (male/female)	7/18	10/15	12/13	$\chi^2 = 2.136$	$p = 0.344$
Age (mean ± SD)	69.24 ± 5.118	70.68 ± 6.725	72.56 ± 6.192	$F = 1.894$	$p = 0.158$
Years of education (mean ± SD)	11.40 ± 2.041	10.80 ± 2.102	10.92 ± 2.515	$H = 2.437$	$p = 0.296$
MMSE score (mean ± SD)	29.12 ± 0.526	27.04 ± 0.841	24.16 ± 2.135	$H = 59.596$	$p < 0.001$
CDT score (mean ± SD)	8.48 ± 2.064	7.00 ± 2.646	6.12 ± 3.032	$H = 11.828$	$p = 0.003$
ADAS-Cog score (mean ± SD)	7.75 ± 2.543	10.75 ± 2.924	18.08 ± 5.994	$F = 32.824$	$p < 0.001$
GDS score (mean ± SD)	3.32 ± 3.024	5.04 ± 3.272	4.08 ± 2.629	$F = 2.082$	$p = 0.132$

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.: Pan et al., 2021; Lehr et al., 2012; Pérez-Toro et al., 2021; Satt et al., 2014; Wang et al., 2019).

To ensure that there were no statistically significant differences among the speaker groups in their age, gender and education, we applied one-way ANOVA, the Kruskal–Wallis H test (when the normality assumption was violated) or Chi-squared test (for categorical values). The test statistics along with the mean and standard deviation for each speaker category are presented in Table 2. From these values, it is clear that the subjects of the three groups do not differ significantly in terms of gender, age and level of education ($p > 0.05$ in all three cases), while all three screening tests show a statistically significant difference.

3. Posterior-thresholding hesitation representation

Next, we will describe the proposed feature extraction approach, focusing on measuring the amount of hesitation (i.e. silent and filled pauses) in the spontaneous speech of the subjects. More precisely, we defined hesitation as the absence of speech for at least 30 ms; we distinguished two subtypes of hesitation: silent pauses and filled pauses (i.e. vocalizations such as 'er', 'umm' etc.). Furthermore, in the following, we will use the standard term "utterance" in the sense of "a speech clip processed at once", which, in our case, is the whole audio clip (i.e. the response of the subject).

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. 1 for the architecture of the proposed approach.

3.1. 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 recordings 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.

3.2. 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) modeling (Hinton et al., 2012), where the phonetic labeling also takes the (left and right) neighbors of the actual phone into consideration. As in this

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 → $m$ do
    $cnt := 0$
    $th := i \cdot s$
    for j := 1 → $N$ do
      if $likelihoods(j) \geq th$ then
        $cnt := cnt + 1$
      end if
    end for
    $features(i) := cnt/N$
  end for
  return $features$
```

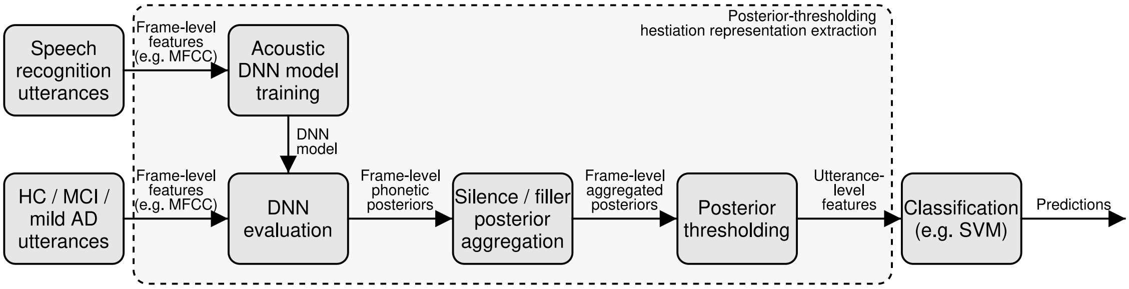

Posterior-thresholding
hestiation representation extraction
Speech recognition utterances
Frame-level features (e.g. MFCC)
Acoustic DNN model training
DNN model
HC / MCI / mild AD utterances
Frame-level features (e.g. MFCC)
DNN evaluation
Frame-level phonetic posteriors
Silence / filler posterior aggregation
Frame-level aggregated posteriors
Posterior thresholding
Utterance-level features
Classification (e.g. SVM)
Predictions

Fig. 1. The general workflow of the applied DNN-based feature extraction process.

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 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.

3.3. 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. 2 illustrates the mechanism of this step.

Note that extracting the posterior thresholding feature set is equivalent to calculating the *cumulative histogram* (Schowengerdt, 2006) of the frame-level posterior estimates. These types of histograms were employed in former ASR techniques (Molau et al., 2001) as well as in numerous other tasks like texture classification (Hiremath and Shivashankar, 2008), handwritten character recognition (Heutte et al., 1998), analog-to-digital converter testing (Alegria and da Cruz Serra, 2001) and in computational paralinguistics (Gosztolya, 2019). 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 detail with a fixed-size vector.

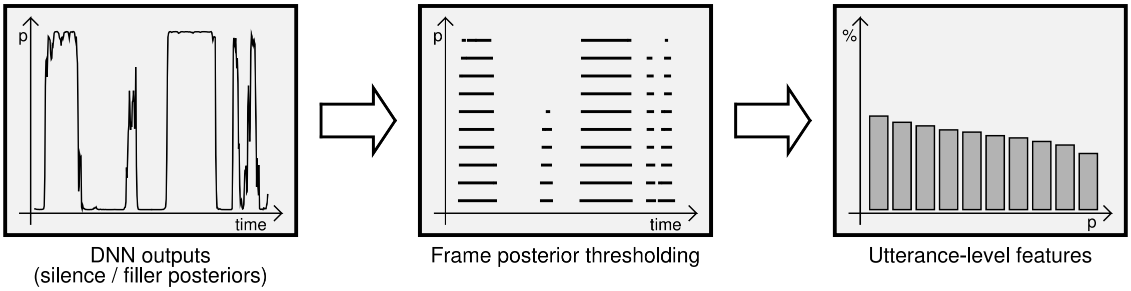


Fig. 2. The schema of the posterior-thresholding feature extraction step.

4. Experimental setup

Now we shall describe in detail how we set up and conducted our experiments. Here, we will cover the procedures carried out for the frame-level feature extraction phase, the DNN acoustic model training and evaluation, the whole Posterior-Thresholding mechanism, and the final classification.

4.1. The DNN acoustic model

Our Deep Neural Network acoustic models were trained on a subset of the BEA Hungarian corpus (Neuberger et al., 2014); we trained the DNN on the speech of 116 subjects (44 h of recordings overall in 9.7k recordings (mean duration of 16.4 s, median duration of 13.3 s)). 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 30 ms in the annotation of this corpus; mean durations were 535 ms and 234 ms, while median durations were 410 ms and 180 ms, 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 accomplished 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 (Odell, 1995); the criterion used during state tying was a Kullback–Leibler divergence-based one (Gosztolya et al., 2015), leading to 911 tied states.

4.2. Posterior-thresholding hesitation representation

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 also sighs 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 recording. When calculating duration, we first omitted the beginning and ending frames where the likelihood of the silent pause category exceeded 0.9.

4.3. Utterance-level classification

As is common in medical speech processing tasks, we relied on the Support Vector Machine (SVM) algorithm for the classification phase. SVM is better suited to datasets with a limited amount of data and it provides a flexible decision boundary; we applied the libSVM implementation (Chang and Lin, 2011) 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

Table 3
The various accuracy scores obtained with the S-GAP temporal speech parameters, following the approach of Tóth et al. (2018) and Gosztolya et al. (2019). (Acc. = classification accuracy, Prec. = precision, Spec. = specificity; HC = Healthy Control, MCI = Mild Cognitive Impairment, mAD = Mild Alzheimer's Disease).

Features	Speaker task	Classification metrics					Area-Under-Curve			
		Acc.	Prec.	Recall	Spec.	F_1	HC	MCI	mAD	Mean
Silence	Immediate recall	38.7%	75.0%	60.0%	60.0%	66.7	0.637	0.474	0.705	0.605
	Previous day	41.3%	68.4%	52.0%	52.0%	59.1	0.502	0.646	0.562	0.570
	Delayed recall	41.3%	68.4%	52.0%	52.0%	59.1	0.558	0.536	0.774	0.623
	All three tasks	42.7%	77.5%	62.0%	64.0%	68.9	0.619	0.374	0.689	0.561
All	Immediate recall	40.0%	71.8%	56.0%	56.0%	62.9	0.580	0.566	0.743	0.630
	Previous day	42.7%	72.5%	58.0%	56.0%	64.4	0.622	0.611	0.569	0.601
	Delayed recall	50.7%	77.5%	62.0%	64.0%	68.9	0.673	0.592	0.805	0.690
	All three tasks	60.0%	83.7%	72.0%	72.0%	77.4	0.728	0.600	0.780	0.705

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 a technique called *nested cross-validation* (Cawley and Talbot, 2010). That is, each time we trained on the data of 72 (i.e. 3 × 24) subjects, we performed *another* (24-fold) cross-validation session, looking for the C meta-parameter value that led to the highest AUC score. Afterwards, 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.

4.4. 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. Gosztolya et al., 2019; Gosztolya, 2019), 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.

4.5. Temporal speech parameters (S-GAP)

For comparison, we also tested the temporal speech parameters (S-GAP) described in our previous studies (e.g. Tóth et al., 2018). This attribute set comprises of utterance duration, speech rate, articulation rate, the total length of pauses/duration ratio, pause rate, and the average length of pauses. These attributes are calculated from the output of a phone-level ASR system after the phonetic decoding step; we used the same context-dependent DNN acoustic model as that employed in our other experiments. For comparison, we also report classification results when using only the silence-related attributes of the S-GAP attribute set: besides the duration of the utterance, it included silence occurrence rates (how much of the actual utterance duration and the phones found by the HMM/DNN hybrid model corresponded to silence), average silence length (s) and silence frequency (1/s).

4.6. Evaluation

As is usual in most medical speech processing research, here 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 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), known as the F-measure or F_1-score, was also employed as a metric to assess the performances of our models. 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, which is practically equivalent to the recall of the healthy control speaker group. These metrics were calculated by setting the decision threshold along with the Equal Error Rate (EER).

Table 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).

Speaker task	Features	Classification metrics					Area-Under-Curve			
		Acc.	Prec.	Recall	Spec.	F_1	HC	MCI	mAD	Mean
Immediate recall	Silence	52.0%	77.5%	62.0%	64.0%	68.9	0.587	0.614	0.759	0.653
	Filler	33.3%	68.4%	52.0%	52.0%	59.1	0.533	0.561	0.498	0.530
	All hesit.	46.7%	75.0%	60.0%	60.0%	66.7	0.570	0.687	0.746	0.668
	All	50.7%	78.0%	64.0%	64.0%	70.3	0.580	0.643	0.769	0.664
Previous day	Silence	50.7%	77.5%	62.0%	64.0%	68.9	0.613	0.684	0.594	0.630
	Filler	38.7%	78.0%	64.0%	64.0%	70.3	0.734	0.522	0.510	0.589
	All hesit.	40.0%	64.9%	48.0%	48.0%	55.2	0.415	0.657	0.574	0.549
	All	50.7%	78.0%	64.0%	64.0%	70.3	0.665	0.680	0.610	0.652
Delayed recall	Silence	60.0%	86.4%	76.0%	76.0%	80.9	0.755	0.670	0.746	0.724
	Filler	33.3%	68.4%	52.0%	52.0%	59.1	0.455	0.497	0.455	0.469
	All hesit.	68.0%	88.9%	80.0%	80.0%	84.2	0.857	0.706	0.802	0.788
	All	68.0%	89.1%	82.0%	80.0%	85.4	0.842	0.700	0.775	0.773
All three tasks	Silence	61.3%	86.4%	76.0%	76.0%	80.9	0.773	0.683	0.734	0.730
	Filler	41.3%	81.0%	68.0%	68.0%	73.9	0.758	0.566	0.526	0.617
	All hesit.	68.0%	88.9%	80.0%	80.0%	84.2	0.854	0.712	0.806	0.791
	All	70.7%	93.5%	86.0%	88.0%	89.6	0.911	0.709	0.794	0.804

5. Experimental results

5.1. Results using the temporal speech parameters (S-GAP)

Table 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 have led 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 the Equal Error Rate.) Due to these values, F_1 was around $63-69$, while the mean AUC fell 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 using the S-GAP temporal speech parameters described by Tóth et al. (2018).

5.2. Results using the posterior-thresholding hesitation representation with context-dependent states

Table 4 shows the results 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 4 –, we can see similar values to those in the silent pause case. Using all three types of attributes together (the 'All' case) 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 those 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.

Table 5
The various accuracy scores obtained with the Posterior-Thresholding Hesitation Representation using Context-Independent states. (Acc. = classification accuracy, Prec. = precision, Spec. = specificity; HC = Healthy Control, MCI = Mild Cognitive Impairment, mAD = Mild Alzheimer's Disease).

Speaker task	Features	Classification metrics					Area-Under-Curve			
		Acc.	Prec.	Recall	Spec.	F_1	HC	MCI	mAD	Mean
Immediate Recall	Silence	50.7%	77.5%	62.0%	64.0%	68.9	0.577	0.590	0.758	0.642
	Filler	30.7%	64.9%	48.0%	48.0%	55.2	0.467	0.509	0.553	0.510
	All hesit.	48.0%	75.0%	60.0%	60.0%	66.7	0.574	0.693	0.769	0.678
	All	50.7%	77.5%	62.0%	64.0%	68.9	0.580	0.597	0.759	0.645
Previous day	Silence	46.7%	77.5%	62.0%	64.0%	68.9	0.618	0.648	0.550	0.605
	Filler	40.0%	81.0%	68.0%	68.0%	73.9	0.749	0.516	0.454	0.573
	All hesit.	33.3%	58.3%	42.0%	40.0%	48.8	0.373	0.640	0.557	0.523
	All	49.3%	80.5%	66.0%	68.0%	72.5	0.659	0.645	0.561	0.622
Delayed recall	Silence	58.7%	86.4%	76.0%	76.0%	80.9	0.772	0.690	0.750	0.738
	Filler	33.3%	67.6%	50.0%	52.0%	57.5	0.478	0.517	0.488	0.494
	All hesit.	62.7%	86.0%	74.0%	76.0%	79.6	0.778	0.684	0.798	0.753
	All	62.7%	86.4%	76.0%	76.0%	80.9	0.786	0.685	0.794	0.755
All three tasks	Silence	62.7%	88.9%	80.0%	80.0%	84.2	0.786	0.693	0.746	0.742
	Filler	40.0%	81.4%	70.0%	68.0%	75.3	0.680	0.519	0.540	0.580
	All hesit.	64.0%	86.4%	76.0%	76.0%	80.9	0.772	0.696	0.802	0.757
	All	69.3%	91.3%	84.0%	84.0%	87.5	0.866	0.703	0.770	0.780

In general, we see 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 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.

5.3. Results using the posterior-thresholding hesitation representation with context-independent states

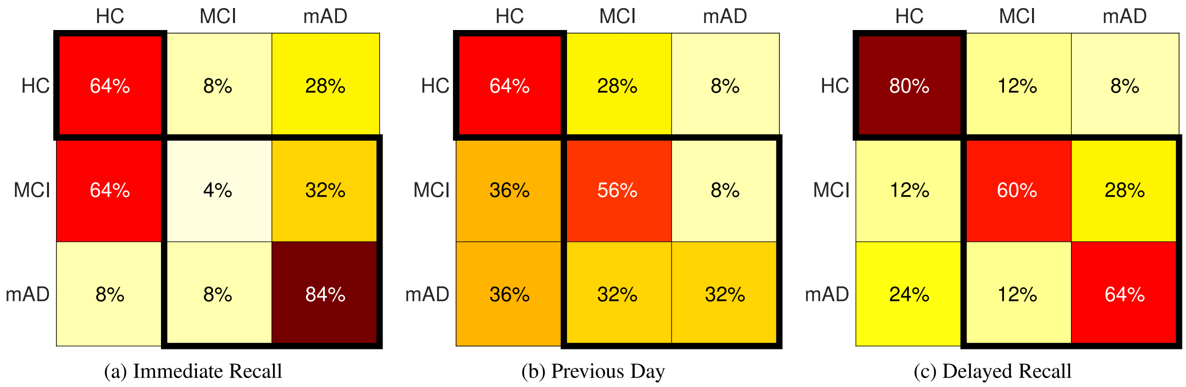
We found that with the PTHR approach we achieved competitive scores for detecting MCI and mAD subjects using a context-dependent DNN acoustic model. However, we wanted to find out whether a context-independent Deep Neural Network component might be enough to express the likelihood of silent and filled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

In our last series of experiments, we examine the behavior 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. 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. 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. 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. 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



(a) Immediate Recall

	HC	MCI	mAD
HC	64%	8%	28%
MCI	64%	4%	32%
mAD	8%	8%	84%

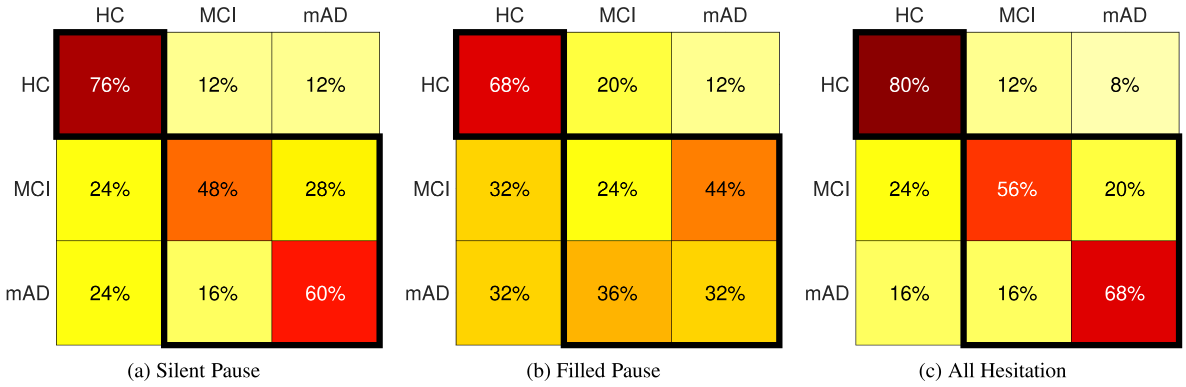
(b) Previous Day

	HC	MCI	mAD
HC	64%	28%	8%
MCI	36%	56%	8%
mAD	36%	32%	32%

(c) Delayed Recall

	HC	MCI	mAD
HC	80%	12%	8%
MCI	12%	60%	28%
mAD	24%	12%	64%

Fig. 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).



(a) Silent Pause

	HC	MCI	mAD
HC	76%	12%	12%
MCI	24%	48%	28%
mAD	24%	16%	60%

(b) Filled Pause

	HC	MCI	mAD
HC	68%	20%	12%
MCI	32%	24%	44%
mAD	32%	36%	32%

(c) All Hesitation

	HC	MCI	mAD
HC	80%	12%	8%
MCI	24%	56%	20%
mAD	16%	16%	68%

Fig. 4. The confusion matrices obtained for the hesitation types (ground truth: real speaker categories, columns: predictions).

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. 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. 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. 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. 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. 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 3.2).) In this case, the percentage of correctly classified subjects was higher for all three subject categories than that 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%).

7. Conclusions and discussion

Alzheimer's disease (AD) is a neurodegenerative disorder that might develop for years before its clinical manifestation, while mild cognitive impairment (MCI) is usually considered as a prodromal stage of AD. In the case of both disorders, early diagnosis might allow timely treatment to delay progression. In our previous studies our team developed an attribute set, focusing mainly on quantifying the amount of hesitations (both silent and filled pauses) present in the speech of the subjects by applying a phone-level ASR system. In this study we presented a feature extraction approach, which also describes the amount of hesitations, but which does not require the application of the whole speech recognition workflow. In contrast, we now only relied on the Deep Neural Network acoustic model of a standard HMM/DNN hybrid model, and 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%, and we achieved an F_1 value of 87.5 and a mean AUC score of 0.780. Although it is impossible to make a direct comparison with other values in the literature owing to the different corpora, experimental setup and evaluation metrics used, our results seem competitive with those 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% (Themistocleous et al., 2018). Fraset et al. on the same dataset, presented 0.88 and 83% of AUC and accuracy scores, respectively (also for a binary class problem) (Fraser et al., 2019); these where achieved via 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, but the task was evaluated as pairwise combination of the three classes (two-class problem) (König et al., 2015).

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. We found 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.

In our last investigation, we focused on the confusion matrices of specific approaches. Our recording protocol consisted of three different speaker tasks (immediate recall, previous day and delayed recall); first we examined the results of the classifier models for one task only. We found that *immediate recall* is best suited for detecting mild AD speakers, but not for distinguishing MCI and HC subjects; in contrast, *previous day* could not be used to identify mAD subjects. The speaker task that produced the best results was clearly that of *delayed recall*: besides allowing the vast majority of healthy control subjects to be identified, it also helped us to detect mAD speakers with a good performance. Furthermore, the majority of misclassifications occurred in the MCI-mAD relation, which is a more tolerable mistake than confusing MCI or mAD subjects with healthy controls and vice versa.

However, besides performance, there was another big difference among our speaker tasks. While *previous day* is quite easy to record, as it requires only spoken instructions, both *immediate recall* and *delayed recall* consisted of first watching a specially designed, one-minute-long film. Therefore the two recall tasks would require significant effort to incorporate them into an actual MCI/mAD screening application, while doing so with the previous day task would be quite straightforward. Regarding performance, although the previous day speaker task proved to be ineffective for obtaining spontaneous speech recordings that might be used to efficiently detect mild Alzheimer's, if we focus on early screening, the detection of MCI cases is more important. In this respect, the previous day speaker task proved to be effective.

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 performance of MCI detection 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 recording, which is significantly smaller than, say, 512-long x-vectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Al-Ghazali, A., Alrefaee, Y., 2019. Silent pauses in the speech of yemeni EFL learners. ELS J. Interdiscip. Stud. Humanit. 2 (1).

Alegria, F., da Cruz Serra, A., 2001. Influence of frequency errors in the variance of the cumulative histogram [in ADC testing]. IEEE Trans. Instrum. Meas. 50, 461–464.

Association, A., 2020. 2020 ALzheimer's disease facts and figures. Alzheimer's Dement. 16 (3), 391–460.

Balagopalan, A., Eyre, B., Rudzicz, F., Novikova, J., 2020. To BERT or not to BERT: Comparing speech and language-based approaches for alzheimer's disease detection. In: Proceedings of Interspeech, pp. 2167–2171.

Baldas, V., Lampiris, C., Capsalis, C.N., Koutsouris, D., 2010. Early diagnosis of Alzheimer's type dementia using continuous speech recognition. In: Proceedings of MobiHealth, Ayia Napa, Cyprus, pp. 105–110.

Beltrami, D., Gagliardi, G., Favretti, R.R., Ghidoniand, E., Tamburini, F., Calza, L., 2018. Speech analysis by natural language processing techniques: A possible tool for very early detection of cognitive decline? Front. Aging Neurosci. 10.

Botelho, C., Teixeira, F., Rolland, T., Abad, A., Trancoso, I., 2020. Pathological speech detection using x-vector embeddings. arXiv:2003.00864.

Cawley, G.C., Talbot, N.L.C., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. 11 (Jul), 2079–2107.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) 2, 1–27.

Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. 19 (4), 788–798.

Folstein, M., Folstein, S., McHugh, P., 1975. Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. 12 (3), 189–198.

Fraser, K.C., Fors, K.L., Eckerström, M., Öhman, F., Kokkinakis, D., 2019. Predicting MCI status from multimodal language data using cascaded classifiers. Front. Aging Neurosci. 11.

Fraser, K.C., Meltzer, J.A., Graham, N.L., Leonard, C., Hirst, G., Black, S.E., Rochon, E., 2014. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. Cortex 55, 43–60.

Fraser, K., Rudzicz, F., Graham, N., Rochon, E., 2013. Automatic speech recognition in the diagnosis of primary progressive aphasia. In: Proceedings of SLPAT, Grenoble, France, pp. 47–54.

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

Gosztolya, G., 2019. Posterior-thresholding feature extraction for paralinguistic speech classification. Knowl.-Based Syst. 186.

Gosztolya, G., Grósz, T., Tóth, L., Imseng, D., 2015. Building context-dependent DNN acousitc models using Kullback-Leibler divergence-based state tying. In: Proceedings of ICASSP, Brisbane, Australia, pp. 4570–4574.

Gosztolya, G., Tóth, L., Grósz, T., Vincze, V., Hoffmann, I., Szatlóczki, G., Pákáski, M., Kálmán, J., 2016. Detecting mild cognitive impairment from spontaneous speech by correlation-based phonetic feature selection. In: Proceedings of Interspeech, San Francisco, CA, USA, pp. 107–111.

Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., Hoffmann, I., 2019. Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using ASR and linguistic features. Comput. Speech Lang. 53 (Jan), 181–197.

Hahn, E.A., Andel, R., 2011. Nonpharmacological therapies for behavioral and cognitive symptoms of mild cognitive impairment. J. Aging Health 23 (8), 1223–1245.

Heutte, L., Paquet, T., Moreau, J., Lecourtier, Y., Olivier, C., 1998. A structural/statistical feature based vector for handwritten character recognition. Pattern Recognit. Lett. 19, 629–641.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process. Mag. 29 (6), 82–97.

Hiremath, P., Shivashankar, S., 2008. Wavelet based co-occurrence histogram features for texture classification with an application to script identification in a document image. Pattern Recognit. Lett. 29, 1182–1189.

Hoffmann, I., Németh, D., Dye, C.D., Pákáski, M., Irinyi, T., Kálmán, J., 2010. Temporal parameters of spontaneous speech in Alzheimer's disease. Int. J. Speech-Lang. Pathol. 12 (1), 29–34.

Igras-Cybulska, M., Ziółko, B., Żelasko, P., Witkowski, M., 2016. Structure of pauses in speech in the context of speaker verification and classification of speech type. EURASIP J. Audio Speech Music Process. 2016 (1), 18.

de Ipiña, K.L., de Lizarduy, U.M., Calvo, P.M., Beitia, B., García-Melero, J., Fernández, E., Ecay-Torres, M., Faundez-Zanuy, M., Sanz, P., 2018. On the analysis of speech and disfluencies for automatic detection of mild cognitive impairment. Neural Comput. Appl. 9.

Kaduszkiewicz, H., Eisele, M., Wiese, B., Prokein, J., Luppa, M., Luck, T., Jessen, F., Bickel, H., Mösch, E., Pentzek, M., Fuchs, A., Eifflaender-Gorfer, S., Weyerer, S., König, H.-H., Brettschneider, C., van den Bussche, H., Maier, W., Scherer, M., Riedel-Heller, S.G., 2014. Prognosis of mild cognitive impairment in general practice: Results of the german AgeCoDe study. Ann. Fam. Med. 12 (2), 158–165.

König, A., Satt, A., Sorin, A., Hoory, R., Derreumaux, A., David, R., Robert, P.H., 2018. Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. Curr. Alzheimer Res. 15 (2), 120–129.

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P.H., David, R., 2015. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimer's Dement.: Diagnos. Assess. Dis. Monit. 1 (1), 112–124.

Lang, L., Clifford, A., Wei, L., Zhang, D., Leung, D., Augustine, G., Danat, I.M., Zhou, W., Copeland, J.R., Anstey, K.J., Chen, R., 2017. Prevalence and determinants of undetected dementia in the community: A systematic literature review and meta-analysis. BMJ Open 7 (2).

Laske, C., Sohrabi, H.R., Frost, S.M., de Ipiña, K.L., Garrard, P., Buscema, M., Dauwels, J., Soekadar, S.R., Mueller, S., Linnemann, C., Bridenbaugh, S.A., Kanagasingam, Y., Martins, R.N., O'Bryant, S.E., 2015. Innovative diagnostic tools for early detection of Alzheimer's disease. Alzheimer's Dement. 11 (5), 561–578.

Lehr, M., Prud'hommeaux, E., Shafran, I., Roark, B., 2012. Fully automated neuropsychological assessment for detecting mild cognitive impairment. In: Proceedings of Interspeech, Portland, OR, USA, pp. 1039–1042.

Martinc, M., Pollak, S., 2020. Tackling the ADReSS Challenge: A multimodal approach to the automated recognition of Alzheimer's Dementia. In: Proceedings of Interspeech, pp. 2157–2161.

Mattys, S.L., Pleydell-Pearce, C.W., Melhorn, J.F., Whitecross, S.E., 2005. Detecting silent pauses in speech: A new tool for measuring on-line lexical and semantic processing. Psychol. Sci. 16 (12), 958–964.

McCullough, K.C., Bayles, K.A., Bouldin, E.D., 2018. Language performance of individuals at risk for mild cognitive impairment. J. Speech Lang. Hear. Res. 62 (3), 706–722.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jr., C.R.J., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., Mohs, R.C., Morris, J.C., Rossor, M.N., Scheltens, P., Carrillo, M.C., Thies, B., Weintraub, S., Phelps, C.H., 2011. 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 Dement. 7 (3), 263–269.

Molau, S., Pitz, M., Ney, H., 2001. Histogram based normalization in the acoustic feature space. In: Proceedings of ASRU, Madonna di Campiglio, Italy, pp. 1–4.

Mueller, K.D., Koscik, R.L., Hermann, B.P., Johnson, S.C., Turkstra, L.S., 2018. Declines in connected language are associated with very early mild cognitive impairment: Results from the wisconsin registry for Alzheimer's prevention. Front. Aging Neurosci. 9.

Neuberger, T., Gyarmathy, D., Gráczi, T.E., Horváth, V., Gósy, M., Beke, A., 2014. Development of a large spontaneous speech database of agglutinative hungarian language. In: Proceedings of TSD. Czech Republic, Brno, pp. 424–431.

Odell, J., 1995. The Use of Context in Large Vocabulary Speech Recognition (Ph.D. thesis). University of Cambridge.

Pan, Y., Nallanthighal, V.S., Blackburn, D., Christensen, H., Härmä, A., 2021. Multi-task estimation of age and cognitive decline from speech. In: Proceedings of ICASSP, Toronto, Canada (online), pp. 7258–7262.

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

Petersen, R.C., 2003. Conceptual overview. In: Petersen, R.C. (Ed.), Mild Cognitive Impairment: Aging to Alzheimer's Disease. Oxford University Press, pp. 1–14.

Petersen, R.C., 2016. Mild cognitive impairment. Contin.: Lifelong Learn. Neurol. 22 (2 (Dementia)), 404–418.

Petersen, R.C., Caracciolo, B., Brayne, C., Gauthier, S., Jelic, V., Fratiglioni, L., 2014. Mild cognitive impairment: a concept in evolution. J. Intern. Med. 275 (3), 214–228.

Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E., 1999. Mild cognitive impairment: Clinical characterization and outcome. Arch. Neurol. 56 (3), 303–308.

R'mani Haulcy, J.G., 2020. Classifying Alzheimer's disease using audio and text-based representations of speech. Front. Psychol. 11.

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., Kaye, J., 2011. Spoken language derived measures for detecting mild cognitive impairment. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2081–2090.

Rosen, W., Mohs, R., Davis, K., 1984. A new rating scale for Alzheimer's disease. J. Psychiatr. Res. 141 (11), 1356–1364.

Satt, A., Hoory, R., König, A., Aalten, P., Robert, P.H., 2014. Speech-based automatic and robust detection of very early Dementia. In: Proceedings of Interspeech, Singapore, pp. 2538–2542.

Schneider, J.A., Arvanitakis, Z., Leurgans, S.E., Bennett, D.A., 2009. The neuropathology of probable Alzheimer's disease and mild cognitive impairment. Ann. Neurol. 66 (2), 200–208.

Schowengerdt, R.A., 2006. Remote Sensing: Models and Methods for Image Processing. Academic Press, Orlando, FL, USA.

Sluis, R.A., Angus, D., Wiles, J., Back, A., Gibson, T.A., Liddle, J., Worthy, P., Copland, D., Angwin, A.J., 2020. An automated approach to examining pausing in the speech of people with dementia. Amer. J. Alzheimer's Dis. Other Dement. 35.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S., 2018. X-Vectors: Robust DNN embeddings for speaker verification. In: Proceedings of ICASSP, pp. 5329–5333.

Szatlóczki, G., Hoffmann, I., Vincze, V., Kálmán, J., Pákáski, M., 2015. Speaking in Alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease. Front. Aging Neurosci. 7.

Taler, V., Phillips, N., 2008. Language performance in alzheimer's disease and mild cognitive impairment: A comparative review. J. Clin. Exp. Neuropsychol. 30 (5), 501–556.

Themistocleous, C., Eckerström, M., Kokkinakis, D., 2018. Identification of mild cognitive impairment from speech in Swedish using deep sequential neural networks. Front. Neurol. 9.

Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatlóczki, G., Biró, E., Zsura, F., Pákáski, M., Kálmán, J., 2015. Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In: Proceedings of Interspeech, Dresden, Germany, pp. 2694–2698.

Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatlóczki, G., Bánréti, Z., Pákáski, M., Kálmán, J., 2018. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. Curr. Alzheimer Res. 15 (2), 130–138.

Wang, T., Lian, C., Pan, J., Yan, Q., Zhu, F., Ng, M.L., Wang, L., Yan, N., 2019. Towards the speech features of mild cognitive impairment: Universal evidence from structured and unstructured connected speech of Chinese. In: Proceedings of Interspeech, Graz, Austria, pp. 3880–3884.

Weiner, J., Schultz, T., 2018. Selecting features for automatic screening for dementia based on speech. In: International Conference on Speech and Computer. Springer, pp. 747–756.

Yesavage, J.A., Sheikh, J.I., 1986. 9/Geriatric Depression scale (GDS). Clin. Gerontol. 5 (1–2), 165–173.

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., Church, K., 2020. Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease. In: Proceedings of Interspeech, pp. 2162–2166.