Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 53 (2019) 181–197

www.elsevier.com/locate/csl

Identifying Mild Cognitive Impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features[☆]

Gábor Gosztolya[a,*], Veronika Vincze[a,b], László Tóth[c], Magdolna Pákáski[d], János Kálmán[d], Ildikó Hoffmann[e,f]

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

Received 7 December 2017; received in revised form 21 March 2018; accepted 26 July 2018
Available online 6 August 2018

Abstract

Alzheimer's disease (AD) is a neurodegenerative disorder that develops for years before clinical manifestation, while mild cognitive impairment is clinically considered as a prodromal stage of AD. For both types of neurodegenerative disorders, early diagnosis is crucial for the timely treatment and to decelerate progression. Unfortunately, the current diagnostic solutions are time-consuming. Here, we seek to exploit the observation that these illnesses frequently disturb the mental and linguistic functions, which might be detected from the spontaneous speech produced by the patient. First, we present an automatic speech recognition based procedure for the extraction of a special set of acoustic features. Second, we present a linguistic feature set that is extracted from the transcripts of the same speech signals. The usefulness of the two feature sets is evaluated via machine learning experiments, where our goal is not only to differentiate between the patients and the healthy control group, but also to tell apart Alzheimer's patients from those with mild cognitive impairment. Our results show that based on only the acoustic features, we are able to separate the various groups with accuracy scores between 74–82%. We attained similar accuracy scores when using only the linguistic features. With the combination of the two types of features, the accuracy scores rise to between 80–86%, and the corresponding F_1 values also fall between 78–86%. We hope that with the full automation of the processing chain, our method can serve as the basis of an automatic screening test in the future.

Keywords: Mild cognitive impairment; Alzheimer's disease; Automatic screening; Automatic speech recognition; Natural language processing; Classifier combination

[☆] This paper has been recommended for acceptance by R. K. Moore.

[*] Corresponding author.
E-mail address: ggabor@inf.u-szeged.hu (G. Gosztolya).

https://doi.org/10.1016/j.csl.2018.07.007
0885-2308/ 2018 Elsevier Ltd. All rights reserved.

1. Introduction

The number of patients suffering from different types of dementia is expected to multiply in the next few decades (Prince et al., 2013). The most frequent reason behind dementia is Alzheimer's disease (AD), which is difficult to diagnose due to the lack of sensitive screening methods. The most frequently used therapy is most effective in the early, mild stage of AD (mAD), hence the timely recognition of AD patients in this stage of the disease would delay the progress of the disease, which could improve the life conditions of the patients (Galvin and Sadowsky, 2012; Nelson and Tabet, 2015). However, the early recognition of dementia is difficult, while the ratio of early diagnosed cases is assessed to be rather low, partly because subjects visit mental clinics only when the dementia is already in an advanced state.

The process of dementia may start around the age of forty with a mild cognitive impairment (MCI). There are several screening tests in the international practice that target early detection, but they are either too time-consuming or cannot diagnose preclinical states. Recognizing MCI is one of the most difficult tasks of gerontopsychiatry. There are test procedures that are designed to help the diagnosis, but their effectiveness is limited and the different tests often yield contradictory results. The currently used dementia filter tests (MMSE, Clock Drawing, ADAS-COG) are not able to accurately recognize MCI (Folstein et al., 1975; Janka et al., 1988; Kálmán et al., 1995; Patocskai et al., 2014; Rosen et al., 1984). Of course, their results can be verified via other diagnostic tools such as volumetric MRI (Scheltens et al., 2002; Zimny et al., 2011; Yin et al., 2013) or diffusion tensor imaging (Stricker et al., 2009; Nakata et al., 2009; Matsuda et al., 2017); these, however, are very time-consuming and costly techniques to utilize for early screening.

In MCI, only very subtle changes (e.g. short term memory, word finding or attention disturbances) can be detected. Dementia is a persisting cognitive deficit in at least three areas of the following mental functions: memory, language, spatial visual skills, abstraction, counting, judgment, emotional state/personality (Cummings and Mega, 2003; Reichman and Cummings, 1999). Recent studies point out that certain types of linguistic impairment can be detected in MCI and mAD (for a summary, see Szatlóczki et al., 2015), such as temporal changes in spontaneous speech, word finding and word retrieval difficulties, verbal fluency difficulties, and reduction in productive and receptive discourse-level processing. (For a short list of such studies, see Table 1.)

Recently, several studies sought to identify different types of dementia with Natural Language Processing (NLP) and Automatic Speech Recognition (ASR) techniques. For instance, automatic speech recognition tools were employed in detecting aphasia (Fraser et al., 2013b; 2013a; 2014), mild cognitive impairment (Lehr et al., 2012) and Alzheimer's disease (Baldas et al., 2010; Satt et al., 2014). Jarrold et al. (2014) relied on speech rate, mean and standard deviation of vowels and consonants in spontaneous speech samples. The lexical analysis of spontaneous speech may also indicate different types of dementia (Holmes and Singh, 1996; Bucks et al., 2000; Lunsford and Heeman, 2015) and it may be exploited in the automatic detection of patients suffering from dementia (Thomas et al., 2005; Shibata et al., 2016). dos Santos et al. (2017) made use of word embeddings to identify MCI patients based on their speech transcripts. As for analyzing written language, changes in the writing style of people may also indicate

Table 1
A short list of studies pointing out various types of linguistic impairments in MCI and mild AD.

Linguistic changes	Publications
Temporal changes in spontaneous speech	Forbes-McKay and Venneri (2005), Hoffmann et al. (2010), Roark et al. (2011), Meilán et al. (2012), de Ipiña et al. (2013), Satt et al. (2014), Jarrold et al. (2014), Laske et al. (2015), Tóth et al. (2015), and Vincze et al. (2016)
Word finding and word retrieval difficulties	Smith et al. (1989), Bayles (1993), Light (1993), Kempler and Zelinski (1994), Kemper et al. (2001), Garrard et al. (2005), Taler and Phillips (2008), Santos et al. (2011), Cardoso et al. (2014), Fraser et al. (2014), Laske et al. (2015), and Garrard et al. (2014)
Verbal fluency difficulties	Barth et al. (2005), Juncos-Rabadán et al. (2010), Hoffmann et al. (2010), Santos et al. (2011), Roark et al. (2011), Satt et al. (2014), and Jarrold et al. (2014)
Reduction in productive and receptive discourse-level processing	Hodges et al. (1992), Ripich (1994), Taler and Phillips (2008), Weiner et al. (2008), Juncos-Rabadán et al. (2010), Rapp and Wild (2011), Tsantali et al. (2013), and Cardoso et al. (2014)

dementia (Garrard et al., 2005; Hirst and Wei Feng, 2012; Le et al., 2011). Finally, eye-tracking data has also been used for detecting MCI in a scenario where patients were required to read short texts (Fraser et al., 2017).

Besides English, there are studies that aim to identify dementia in native speakers of e.g. Portuguese (dos Santos et al., 2017), Japanese (Shibata et al., 2016) and Swedish (Kokkinakis et al., 2017; Fraser et al., 2017). Concerning the automatic detection of MCI in Hungarian subjects, Tóth et al. (2015) experimented with speech recognition techniques, while Vincze et al. (2016) sought to identify MCI patients based on linguistic features gained from the transcripts of spontaneous speech recordings.

Here, we focus on the analysis of the temporal characteristics of speech, which allows us to examine the cognitive processes during operation, and it helps us to explore the connections between language and memory. In our investigations, we mainly focus on the various forms of hesitations, since MCI is known to manifest itself in longer hesitations and a lower speech rate (Roark et al., 2011; Jarrold et al., 2014). While hesitation is defined as an absence of speech, it can be divided into two categories: silent pauses and filled pauses, the latter being vocalizations like uhm, er, etc. Our overall goal here is to create a software application that can perform a pre-filtering of the possible patients, which could then be followed by a diagnosis by a medical expert. We would like to add that we do not wish to create the final diagnosis of the subjects, as this is the task of medically trained personnel.

Concerning the automatic detection of MCI, our previous study experimented with speech recognition techniques (Tóth et al., 2015), while later we attempted to identify MCI patients based on linguistic features acquired from the transcripts of spontaneous speech recordings (Vincze et al., 2016). Here, we extend our experiments to involve both mAD and MCI patients, and we are interested to see whether the proposed techniques are able to distinguish between these two categories as well. Furthermore, we also combine the low-level acoustic features extracted from the speech signal with the higher-level linguistic features extracted from the speech transcripts in the hope that these two groups of features are complementary, and hence their combination will improve the performance of our machine learning-based classifiers.

The structure of the paper is as follows. In Section 2, we present the MCI-mAD database we used in our experiments. Then, in Section 3, we describe the acoustic markers we extracted from the spontaneous speech of the subjects, and show the test results using these acoustic features. Next, in Section 4, we present the linguistic features used, and perform dementia identification experiments using these attributes. Lastly, in Section 5, we fuse the two approaches and show that the acoustic and linguistic markers can support each other: we achieved our best results by combining the two different types of features.

2. The Hungarian MCI-mAD database

Our database was recorded at the Memory Clinic at the Department of Psychiatry of the University of Szeged, Hungary. The study was approved by the Ethics Committee of the University of Szeged, and it was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants. We collected utterances from three categories of subjects: those suffering from MCI, those affected by early-stage AD, and those having no cognitive impairment at the time of recording (i.e. the control group). The three categories of subjects were matched for age, gender and education. 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 concurrently supported by CT, MRI, and cognitive tests (Mini-Mental State Examination (MMSE, Folstein et al., 1975), the Clock Drawing Test (CDT, Freedman et al., 1994) and ADAS-Cog (Rosen et al., 1984)).

All our previous studies (Hoffmann et al., 2010; Tóth et al., 2015; Gosztolya et al., 2016b) and studies performed by other groups (e.g. Taler and Phillips, 2008; Roark et al., 2011; Satt et al., 2014) found that MCI and AD affect the *spontaneous* speech of the patients more than their planned speech. This is because in the case of planned speech, speakers usually have some time in advance to think about what they would like to say, hence difficulties in word finding (due to memory decline) cannot be reliably detected. However, in the case of spontaneous speech, speakers are required to speak on the spot, i.e. without any time to prepare their speech, which might truly reflect their difficulties in word finding. Therefore, our aim was to record spontaneous speech. This is why our experimental setup for recording was as follows (for the details, see Hoffmann et al., 2010). After the presentation of a specially designed one-minute-long animated film, the subjects were asked to talk about the events seen on the film (*immediate recall*).

After the presentation of a second film, the subjects were asked to talk about their previous day (*previous day*). As the last task, the subjects were asked to talk about the second film (*delayed recall*). (For the instructions to the patients, see Table 2.) Each recording was edited: parts before the subject started to speak and after his last phoneme uttered were removed. Hence, we had three recordings for each subject, each containing spontaneous speech, but the tasks performed were different. Of course, it may turn out that some tasks are less useful for detecting MCI or mAD than others, but this cannot be known in advance.

Our database of MCI and AD patients is continuously growing; at the time of writing we had recordings taken from more than 150 persons. For various reasons (poor sound quality, controversial diagnosis, etc.) we had to filter out some patients; furthermore, since we insisted on matching the three groups of speakers by age, gender and level of education, we could not use some of the recordings, which otherwise fulfilled our requirements of having a clear diagnosis and an acceptable sound quality. Therefore, in the end we used the recordings of 25 speakers for each speaker group, resulting in a total of 75 speakers and 225 recordings. We applied one-way ANOVA to check if there were significant differences among the different groups. F and *p*-values can be seen in Table 3.

2.1. Subject classification by machine learning

We used the Weka tool (Hall et al., 2009), which is a free, open-source collection of machine learning algorithms. We applied Support-Vector Machines (SVM Schölkopf et al., 2001) with a linear kernel, utilizing the *SMO* implementation in Weka. Each speaker was characterized by one feature vector, i.e. the acoustic markers calculated based on the three recordings containing the speech of the subject were concatenated.

From a machine learning perspective, having only 75 examples (i.e. subjects) is an extremely small dataset. However, the number of diagnosed MCI and mAD patients is limited, collecting recordings of their speech and obtaining a medical diagnosis is time-consuming. The similar studies we found involved fewer than 100 patients (Satt et al., 2014; Jarrold et al., 2014; Lehr et al., 2012; Roark et al., 2011; Fraser et al., 2013b; Weiner et al., 2016).

Having so few examples, we did not create separate training and test sets, but opted for cross-validation (CV). In order to guarantee that each fold had the same number of speakers from each speaker group, we used 5-fold cross-validation: we always trained on the features extracted from the speech of 60 speakers, from which 20 had MCI, 20 had mAD and 20 were control subjects. In the next step, this machine learning model was evaluated on the remaining 15 speakers. We repeated this for all speakers, and we then aggregated the results into one final score.

The *C* complexity meta-parameter was determined by a technique called *nested cross-validation* (Cawley and Talbot, 2010). That is, in each case we trained on the data of 4 folds (consisting of 60 subjects), we performed *another* cross-validation session. In this 4-fold cross-validation, we chose the *C* value which led to the highest average AUC score of the MCI and/or mAD classes; then we trained an SVM model with the selected complexity meta-parameter on the data of all the 60 speakers, and this model was evaluated on the remaining 15 speakers. This way we ensured that there was no peeking, which would have created a bias in our scores if we had used standard cross-validation.

2.2. Evaluation

The choice of evaluation metric is not a clear-cut issue for this task. First of all, we can use the traditional classification accuracy score, since the class distribution is balanced for this dataset. However, besides indicating how well

Table 2
The instructions to the patients when recording the three utterances.

(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) "*Now, I would like to ask you to tell me about your day yesterday in details.*"

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

Table 3
Demographic data (i.e. age and education) and the results of the MMSE, CDT and Adas-Cog tests of the three subject groups.

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

the subjects were identified as the members of each category, this task can also be viewed as a detection task, where we are interested in whether the speaker has *any* sort of cognitive disorder, i.e. treating the MCI and mAD categories together. As in this case the class distribution becomes imbalanced (25 control subjects and 50 subjects having some kind of cognitive disorder), we will also report (two-class) classification accuracy scores, but using the Unweighted Average Recall score (UAR, calculated as the mean of the class-wise recall scores) also makes sense. We can also use the standard Information Retrieval metrics of *precision* and *recall*. As there is evidently a trade-off between these two scores, they are usually aggregated together by the *F-measure* (or F_1*-score*), which is the harmonic mean of precision and recall. Medical studies tend to report *sensitivity* and *specificity* instead, sensitivity being equivalent to recall, while specificity being practically the recall of the negative class (in this case, healthy controls). In the experiments we will present (3-class) accuracy scores and all the five 2-class scores (i.e. accuracy, UAR, precision, recall and F-measure).

2.3. Two-class evaluation

Of course, in a real application not all kinds of mis-classifications are of equal importance. It may be worth investigating how efficiently the different speaker groups can be differentiated from each other. Although analyzing the confusion matrix of the classifier may serve as a basis of such an investigation, we think that binary (class-wise) performance can be evaluated most reliably by the performance of binary classifiers specifically trained to distinguish the two appropriate classes. Therefore, in our experiments, we also trained binary classifiers in four variations. In the first case we used the data of all the 75 speakers, but patients diagnosed with either MCI or mAD were treated as members of the same class. In the other three cases, we used only the 25−25 speakers of two speaker groups. Note that when we sought to differentiate MCI from mAD, we calculated precision, recall and F-measure by considering mAD as the positive class.

2.4. Demographic attributes

Gender and age are both the most influential risk factors of MCI (Sachdev et al., 2012). These two attributes were also available for our training set, so we added them to the feature set along with the number of years of education, resulting in 27 and 84 features for the basic and extended feature sets, respectively. While fairly reliable techniques exist to automatically assess the age and gender from the speech signal (see e.g. Kockmann et al., 2010; Meinedo and Trancoso, 2011; Kumar et al., 2016; Grzybowska and Kacprzak, 2016), in the planned application we can simply ask the subjects to provide these data when commencing the test.

To provide reference scores, we performed classification experiments by using only these three demographic attributes; these tests followed our experimental setup described in Section 2.1 in every way. In this experiment, 3-class accuracy turned out to be 40%, only slightly exceeding 33.3% achievable by random guessing in a (balanced) 3-class task.

From the resulting scores (see Table 4) we can see that the demographic attributes help to differentiate the mild AD patients from healthy controls the most. When we try to determine if the speaker has any sort of dementia (i.e. *Control vs. MCI + mAD* case), the accuracy and F_1 scores look fine at first glance; notice, however, the 59% UAR and the 53.3% specificity values which barely exceed 50%, which is the straightforward baseline scores in a 2-class setup. This also indicates that the 68% accuracy was mainly achieved due to the imbalanced class distribution in this

Table 4
The various accuracy scores obtained by using only the demographic information (i.e. age, gender and education) in the 2-class machine learning cases. The 3-class accuracy score was 40.0%.

Speaker groups	Accuracy	UAR	Precision	Recall	Specificity	F_1
Control vs. MCI + mAD	68.0%	59.0%	71.7%	86.0%	53.3%	78.2
Control vs. MCI	60.0%	60.0%	61.9%	52.0%	58.6%	56.5
Control vs. mAD	68.0%	68.0%	73.7%	56.0%	64.5%	63.6
MCI vs. mAD	52.0%	52.0%	52.2%	48.0%	51.9%	50.0

case. As for the remaining two cases, in our opinion the results show that these demographic features allow the identification of MCI patients (versus healthy controls) only at a low rate, while we cannot distinguish MCI patients from mAD ones at all.

3. Extracting acoustic markers from spontaneous speech

The analysis of the time course of speech has been shown to be an especially sensitive neuropsychological method for investigating cognitive processes such as speech production and planning (Hoffmann et al., 2010). Investigating the temporal parameters of spontaneous speech is vital because it can provide sensitive measures of a subject's speech and language skills (Baum et al., 1990; Illes, 1989).

In a study for Hungarian, the following parameters of speech were measured for AD patients and a normal control group: articulation rate, speech tempo, hesitation ratio, and rate of grammatical errors. The results indicated that these parameters of speech may have a diagnostic value for mild-stage AD and therefore could be a useful aid in medical practice (Hoffmann et al., 2010). Other scientific studies have also confirmed that speech analysis could be a useful method in examining, or even diagnosing mild AD (Roark et al., 2011; Satt et al., 2014; Jarrold et al., 2014; Illes, 1989; de Ipiña et al., 2013; Meilán et al., 2014). In addition, lexical decision reaction time studies showed a longer overall latency in AD and MCI patients than in normal controls (Taler and Jarema, 2006; Cuetos et al., 2003; Walla et al., 2005). These results also confirm that speech analysis can contribute to the effective diagnosis of dementia. Many studies (e.g. Barth et al., 2005; Juncos-Rabadán et al., 2010) noted that the presence of dementia increases the amount of hesitation present in spontaneous speech. Following our previous studies (Tóth et al., 2015; Gosztolya et al., 2016b), our experts first calculated eight acoustic markers manually by using the Praat software. These speech-related markers were: articulation rate (1), speech tempo (2), length of utterance (3), duration of silent and filled pauses (hesitation) (4−5), number of silent and filled pauses (6−7) and hesitation rate (8). Hesitation was defined as the absence of speech for more than 30 ms (Gósy, 1998). We should add that the absence of speech does not necessarily mean silence, but includes filled pauses as well. Table 5 summarizes the eight basic acoustic markers and how they were calculated.

3.1. Automatic acoustic marker extraction using ASR

As calculating the above acoustic markers manually is quite expensive and requires skilled labor, it would be beneficial to automate this step. One way of automating it is to use signal processing techniques. For example, Satt et al. employed the Praat software to segment the utterance into voice/silent and periodic/aperiodic parts (Satt et al., 2014). However, these simple techniques cannot extract all the features of Table 5; most importantly, they cannot

Table 5
A description of the eight acoustic markers found to correlate with MCI by Hoffmann et al. (2010).

(1) Articulation rate was calculated as the number of phones per second during speech (excluding hesitations).
(2) The speech tempo (phones per second) was calculated as the number of phones per second divided by the total duration of the utterance.
(3) The length of utterance, given in milliseconds.
(4−5) The duration of silent and filled pauses was calculated as the total duration of filled and silent pauses.
(6−7) The number of silent and filled pauses is calculated as the absolute occurrence of silent and filled pauses, respectively.
(8) The hesitation rate characterizes the ratio of pauses and speech, which was calculated by dividing the length of the utterance by the total duration of pauses (both silent and filled).

distinguish filled pauses from speech. The second option is to apply ASR techniques. However, an off-the-shelf ASR tool (like the one employed by Fraser et al. (2013a)) may be suboptimal. This is because standard speech recognizers are trained to minimize the transcription errors at the word level, while here we seek to extract non-verbal acoustic features like the rate of speech or the duration of silent and filled pauses. Furthermore, while the filled pauses do not explicitly appear in the output of a standard ASR system, our feature set specifically requires them to be found. And lastly, by examining the speech of dementia patients it was observed that the amount of agrammatical sentences and incorrect word inflections increases (Fraser et al., 2014). It is almost impossible to build a standard ASR system to handle these unpredictable errors.

However, notice that the basic speech-related markers listed in Table 5 do not require the correct identification of the phonemes: we need only to *count* them. The only phenomena we need to take special care of are the two forms of hesitation: silent and filled pauses. For these reasons we decided to use a speech recognizer that just provides a phone sequence as output (including filled pause as a special ‘phoneme’). This allows us the automatic extraction of acoustic markers, which can then be utilized to perform automatic subject categorization via machine learning techniques. For the scheme of our workflow, see Fig. 1.

Unfortunately, recognizing the spontaneous speech of elderly people is known to be difficult (Ramabhadran et al., 2003). Doing this without a vocabulary, only at the phonetic level clearly increases the number of errors. However, as we pointed out, not all types of phone recognition errors harm the extraction of our acoustic markers. In our previous experiments (Tóth et al., 2015) we found that this kind of automatic feature extraction was useful for distinguishing speakers having MCI from healthy controls. In the current study we will use the same acoustic features and the same automatic extraction method. However, here were are interested to see whether the same approach is useful in distinguishing three speaker categories instead of just two.

3.2. *Extending the set of speech-related markers*

In the study that served as our starting point, we examined the eight acoustic features shown in Table 5. The reason for this was that calculating and evaluating the features manually required an expensive workload. However, in Section 3.1 we introduced an approach to automatically obtain the time-aligned phoneme sequence of the utterances, serving as the basis of speech marker extraction. Hence, we can readily extend the feature set by using other features that can be calculated via the phone labels, as it can be done quite cheaply at this point. Therefore, we looked for further features that we assumed could support the machine learning method applied in the second phase. This *extended* feature set was calculated as follows.

Firstly, we kept all the original features of Table 5. However, features (6) and (7) were altered slightly: instead of calculating the raw number of silent and filled pauses, we normalized them by dividing them by the total number of phones in the utterance. Furthermore, as we already have the length of each occurrence of silent/filled pauses, it was easy to extend the feature set with the mean and standard deviation of the lengths for these label occurrences. In addition, we observed that the ASR system often confused filled pauses with certain phones. For example, the most frequent sound uttered during hesitation is a schwa, which is easily confused with the vowel [ø]. Another example is substituting the hesitating word “hmm” with the phone [m]. Thus, we conjectured that an increase in the number and cumulative duration of these phones in the ASR output might indicate the presence of mis-recognized filled pauses.

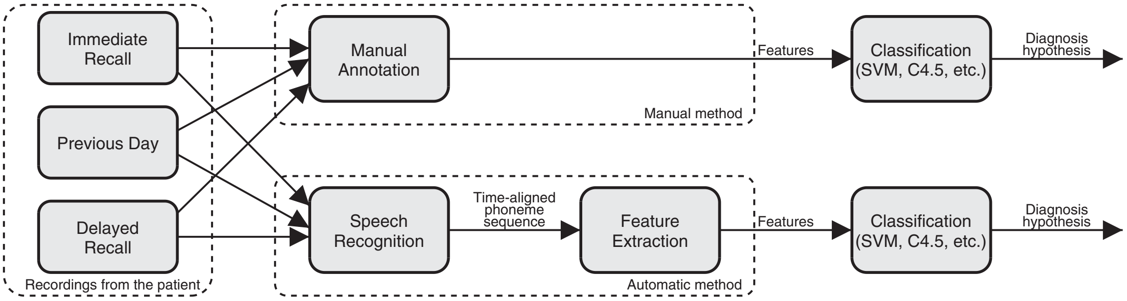

Fig. 1. The steps of MCI detection using manual (upper path) or ASR-based (lower path) acoustic marker extraction.

Table 6
The four additional statistical descriptors (Tóth et al., 2015).

(1) The number of occurrences of the given phoneme divided by the total number of phoneme occurrences.
(2) The total duration of occurrences of the given phoneme divided by the duration of the utterance.
(3) The mean length of the occurrences of the given phoneme.
(4) The standard deviation of the length of the occurrences of the given phoneme.

This led us to extend our feature set with features that describe the distribution of these phones in the utterance. More precisely, for the phones [m], [n] and [ø] we added the following four statistical features to the feature set: cumulative duration (divided by the duration of the utterance), the number of occurrences (divided by the number of phonemes in the utterance), and the mean and standard deviation of the phone duration (see Table 6.) With these extensions we obtained a set of 81 features, which will be referred to as the 'extended' feature set in the experiments.

3.3. ASR parameters

Our automatic speech recognizer was trained on the BEA Hungarian Spoken Language Database (Gósy, 2012). This database contains spontaneous speech that is similar to the recordings collected from our patients. We used roughly seven hours of speech data from the BEA corpus, mainly recordings from elderly persons, in order to match the age group of the targeted audience. To make sure the annotation suited our needs, we added filled pauses, breath intakes and exhales, laughter, coughs and gasps to the transcriptions in a consistent manner.

The ASR system was trained to recognize the phones in the utterances, where the phone set included the special non-verbal labels listed above (i.e. filled pauses, coughs, breath intakes etc.). We used a workflow based on HTK (Young et al., 2006); for acoustic modeling we used a standard feed-forward Deep Neural Network (DNN) (Tóth, 2015). The DNN had 3 hidden layers, each consisting of 1000 ReLU neurons, while we used softmax neurons in the output layer. We used 39 MFCC+$\Delta + \Delta\Delta$ vectors as features on the frame level; to improve model accuracy, we evaluated our model on a sliding window of feature vectors with a width of 15 frames. As a language model we employed a simple phone bigram (again, including all the above-mentioned non-verbal audio tags).

The output of the ASR system is the phonetic segmentation and labeling of the input signal, which includes both silent and filled pauses. Based on this output, the acoustic markers described in Sections 3.1 and 3.2 can be easily extracted via simple calculations.

3.4. Results

Table 7 shows the subject classification accuracy scores obtained using the various acoustic feature sets. The 3-class accuracy scores might seem somewhat low at first glance, but recall that we had a three-class machine learning model trained with equal class distribution, therefore chance level is equivalent to an accuracy score of 33.3%; furthermore, these values significantly exceed the 40% obtained when using only the demographic information. Noting this point, we consider the 61.3% 3-class accuracy score obtained by manual feature calculation a good result. When translating this performance into a 2-class detection problem, we can see that we got quite high scores, indicating that perhaps the main source of error in the 3-class case was the confusion between the MCI and mAD classes.

Calculating the same feature set automatically led to much lower metric scores both in the 2-class and in the 3-class cases. Turning to the extended feature set, however, increased all the metric values: most values are slightly

Table 7
The various accuracy scores obtained using the different acoustic feature sets in the 3-class case.

		3-class	2-class					
Feature extraction	Feature set	Accuracy	Accuracy	UAR	Precision	Recall	Specificity	F_1
Manual	Basic	61.3%	76.0%	74.0%	83.3%	80.0%	63.0%	81.6
Automatic	Basic	50.7%	64.0%	60.0%	73.5%	72.0%	46.2%	72.7
	Extended	58.7%	73.3%	74.0%	85.7%	72.0%	57.6%	78.3

Table 8
The various accuracy scores obtained using the "extended" acoustic feature set in the 2-class machine learning cases.

Speaker groups	Accuracy	UAR	Precision	Recall	Specificity	F_1
Control vs. MCI + mAD	74.7%	72.0%	81.6%	80.0%	61.5%	80.8
Control vs. MCI	78.0%	78.0%	85.0%	68.0%	73.3%	75.6
Control vs. mAD	82.0%	82.0%	78.6%	88.0%	86.4%	83.0
MCI vs. mAD	76.0%	76.0%	72.4%	84.0%	81.0%	77.8

lower than in the manual case, but the same UAR score of 74% indicates that the difference is mainly due to the different types of mis-classifications. In our opinion, these scores indicate that the automatic feature extraction process proved to be less precise than the manual one; however, the additional attributes calculated from the ASR output with a negligible computational cost could counter this effect, allowing practically the same machine learning performance.

Table 7 suggests that the extended feature set works the best among the two tested automatic ones, therefore we used only these speech-related markers when training binary classifiers. Table 8 shows the results obtained. When using the data of all the 75 speakers, we got results similar to when we trained a 3-class SVM model and merged the predictions of the two classes associated with dementia: besides a slightly higher accuracy score (74.7% instead of 73.3%), we got slightly lower UAR and precision, while recall and F_1 rose by a small amount. The following rows give the performance of the classifiers trained for each class pair: control subjects and those having mild Alzheimer's are the easiest to distinguish, being on the two endpoints of the cognitive disorder scale in our study. Determining whether the subject has MCI or he is a healthy control, and distinguishing the two types of dementia turned out to be similarly difficult, although the accuracy scores of 76−−78% reflect a good performance.

3.5. *Summary*

Based on our previous studies of MCI detection, we created an acoustic feature set which focuses on articulation rate, speech tempo, and other descriptors of silent and filled pauses present in the speech of the subject. We showed that using this set of indicators is useful for detecting both MCI and mild AD, and that they can be calculated in an automatic way using ASR techniques, which allows for a similar quality distinction between the speaker groups examined. We also demonstrated that this feature set can be extended with other indicators with a negligible additional computational cost, and that using these indicators is beneficial for MCI and mAD detection performance. Next, we will use completely different kinds of markers: those of linguistic features.

4. Detecting MCI and mAD by linguistic features

In order to distinguish MCI, mAD and healthy subjects, we also made use of the transcripts of the recordings. For our investigations, we relied on the methodology described in our previous study (Vincze et al., 2016); the main difference is that earlier we sought to distinguish only control subjects and those having MCI, while now we used all three speaker groups.

The recordings were manually transcribed by linguists. These transcripts reflect several characteristics of spontaneous speech. They contain several forms of hesitations and silent pauses, also marked in the transcripts. Moreover, they abound in phenomena typical of spontaneous Hungarian speech such as phonological deletion (*mer* instead of the standard form *mert* "because" or *ement* instead of the standard form *elment* "(he) left") and lengthening (*utánna* instead of the standard form *utána* "then"). There are duplications (*ez ezt* "this this-ACC") and neologisms created by the speaker (*feltkáva*, which probably means *főtt kávé* "brewed coffee").

Fillers also deserve special attention when studying transcripts. Besides hesitations, we treated words and phrases referring to some kind of uncertainty together with indefinite pronouns as fillers such as *ilyen* "such", *olyan* "such", *izé* "thing, gadget", *és aztán* "and then", *valamilyen* "some kind of", *valahogy* "somehow", *valamerre* "somewhere". Thus, MCI patients often seem to substitute content words with fillers or indefinite pronouns, and they also appear to use lots of paraphrases, which also indicate uncertainty just like *egy ilyen bagolyszerűség* a kind of owl-likeness “something similar to an owl” or *az olyan délelőtt volt* that such morning was “it happened some time in the morning”.

Transcripts were first morphologically and syntactically analysed with magyarlanc, a linguistic preprocessing toolkit developed for Hungarian (Zsibrita et al., 2013). For classification, we exploited morphological, syntactic and semantic features extracted from the output of magyarlanc.

Recall that in our recording protocol, each subject was asked to produce spontaneous speech in three distinct ways. As both MCI and mAD are strongly related to memory deficit, we think that the order of the tasks might also influence performance, hence we processed each transcript separately. Thus, for each person, features were calculated separately for the three transcripts and all of them were exploited in the machine learning system.

4.1. Experiments

In our experiments, we employed features of spontaneous speech and morphological and semantic features derived from the transcripts and their automatic linguistic analyses. When defining our features, we took into account the fact that the speech of MCI patients may contain more pauses and hesitations than those of healthy controls (Tóth et al., 2015) and they are also supposed to have a restricted vocabulary due to cognitive deficit, which may affect the choice of words and the frequency of parts of speech (Croot et al., 2000) and might even produce neologisms. We also made use of demographic features that were at our disposal.

Our feature set contained the following features:

Morphological features: number of tokens and words; number and rate of lemmas; number of punctuation marks; number and rate of nouns, verbs, adjectives, pronouns and conjunctions; number of first person singular verbs; number and rate of unanalyzed words, i.e. those with an “unknown” POS tag.

Spontaneous speech-based features: number of filled and silent pauses; number and rate of hesitations compared to the number of tokens; number of pauses that follow an article and precede content words; number of lengthened sounds.

Semantic features: number and rate of uncertain words compared to the number of all tokens; number and rate of words/phrases related to memory activity (e.g. *nem emlékszem* not remember-1SG “I can’t remember”); number of negation words; number and rate of content words and function words; number of thematic words related to the content of the films, based on manually constructed lists.

Demographic features: gender; age; education. We always included these features in our feature set.

When detecting uncertain phrases, we applied list matching methods, based on the lists defined by Vincze (2014). Lists for memory activity were compiled by linguists. The lists of related words to the films were also constructed by linguists who watched the films and collected terms (together with their synonyms) that describe their content. Each mention was counted separately. Of course, this feature was not employed for the open ended task, i.e. recalling the events of the previous day.

Our machine learning set-up was quite similar to those used in Section 3 and described in details in Section 2.1. As above, we again compared results of differentiating all the three classes of patients and of aggregated results for the MCI and mAD patients, contrasted with healthy controls. We also conducted experiments to study which group

Table 9
The various accuracy scores obtained using the different linguistic feature sets in the 3-class case.

	3-class	2-class					
Feature set	Accuracy	Accuracy	UAR	Precision	Recall	Specificity	F_1
Morphological	56.0%	70.7%	66.0%	76.9%	80.0%	56.5%	78.4
Speech-based	60.0%	70.7%	69.0%	80.4%	74.0%	55.2%	77.1
Semantic	56.0%	76.0%	74.0%	83.3%	80.0%	63.0%	81.6
Morph. + Speech-based	66.7%	80.0%	79.0%	87.2%	82.0%	67.9%	84.5
Morph. + Semantic	58.7%	73.3%	69.0%	78.9%	82.0%	60.9%	80.4
Speech-based + Semantic	62.7%	81.3%	81.0%	89.1%	82.0%	69.0%	85.4
All	60.0%	77.3%	74.0%	82.4%	84.0%	66.7%	83.2

Table 10
The various accuracy scores obtained using all linguistic features in the 2-class machine learning cases.

Speaker groups	Accuracy	UAR	Precision	Recall	Specificity	F_1
Control vs. MCI + mAD	80.0%	78.0%	85.7%	84.0%	69.2%	84.9
Control vs. MCI	76.0%	76.0%	84.2%	64.0%	71.0%	72.7
Control vs. mAD	82.0%	82.0%	80.8%	84.0%	83.3%	82.4
MCI vs. mAD	68.0%	68.0%	68.0%	68.0%	68.0%	68.0

of linguistic features seems to have the most added value to the identification task, i.e. what the most effective features are.

4.2. Results

Table 9 contains our results obtained by using the different linguistic feature subsets. Making use the subsets independently led to mixed results: 3-class accuracy turned out to be highest for the speech-based attributes, but when measuring performance by 2-class accuracy or F_1, we got the lowest values with these features as well. When combining two feature subsets, combining either morphological or semantic features with the speech-based ones led to even higher scores; however, all 3-class scores were better when we used only the semantic features than when we also included the morphological attributes in our feature set (with the exception of precision). All in all, each group of features seems to considerably outperform the baseline results. Thus, various linguistic features are able to effectively distinguish MCI and mAD patients and healthy controls, which might reflect that there are linguistic differences among the three groups at several linguistic layers, involving phonetics and phonology, morphology and semantics.

What we find particularly interesting is the good performance of the semantical attributes. Semantic-based features include the usage of words referring to memory activities, e.g. “I can’t remember”, “I have forgotten”. Cognitive abilities decline over time, which leads to the unability of remembering and recalling events as well as to problems with word finding. Patients often verbalize their mental efforts to find words or to remember things, which manifests in the usage of memory-related terms. Moreover, issues with word finding can also be found in using vague and uncertain words (e.g. “maybe”, “I guess”, “something like a bird”): when the patient is unsure about his/her memories due to cognitive decline, s/he often avoids exact and specific phrasing and signals the possibility of recalling his/her memories incorrectly. Healthy controls usually have no such problems, which explains why semantic features can effectively distinguish the groups.

Tables 9 and 10 show our results for 2-way and 3-way classifications, respectively. As can be seen from comparing the last row in Table 9 and the first row in Table 10, 2-way classification proves to be slightly more efficient. This result is in harmony with our previous experiments, hence it is highly probable that our method’s main contribution is to distinguish healthy controls from patients with different stages of dementia. It is also justified by our results that the cognitive gap is bigger between healthy people and mAD patients than healthy people and MCI patients as we could achieve higher scores for distinguishing the former groups than the latter ones. We should also mention that the linguistic features tested performed the worst when the task was to distinguish MCI patients from mAD patients.

4.3. Summary

In our experiments we utilized automatically extracted linguistic features from the manual transcripts of the recordings. We also analyzed the usefulness of different feature sub-types for the separation of MCI and mAD patients and control subjects. We found that our feature set, developed to detect mild cognitive impairment, is also useful for detecting mild Alzheimer’s disease. Next, we are going to investigate whether the combination of acoustic and linguistic features can improve our results in identifying patients with dementia.

5. Combining acoustic and linguistic features

Previously in Section 3 we found experimentally that hesitation can be observed in the speaker's spontaneous speech, and that our automatically extracted, hesitation-related acoustic features can be utilized for MCI and mAD detection. In Section 4 we analyzed the speakers' speech by extracting linguistic features from the transcripts of the recordings. Since the two approaches are different by nature, we expect that using both of them may reinforce their strong points, leading to more accurate MCI and mAD detection. Therefore next we will present the results of our experiments with fusing the two types of feature sets.

From the acoustic feature sets we will use the extended one, as this led to the highest-quality predictions. However, among the linguistic features, we will test all subsets and possible subset combinations (described in Section 4). The main reason for this is that the type and applicability of the linguistic feature groups is quite different. The speech-based indicators, for example, describe similar phenomena present in the patient's spontaneous speech as the acoustic indicators do (i.e. they focus on silent and filled pauses), therefore using both feature sets might turn out to be unnecessary. Still, the semantic features focus on the presence of specific words and expressions, which indicate the speaker's uncertainty. Although our current study relies on the manual transcription of the recordings, these keyword occurrences might even be detected automatically by some form of spoken term detection technique (Junkawitsch et al., 1996; Gosztolya and Tóth, 2011; Lee et al., 2016), without having to obtain the complete transcription of the utterance.

5.1. Classifier fusion

To combine the predictions achieved via two or more feature sets, two basic approaches are available. The first possible way (called *early fusion* (Snoek et al., 2005)) is to merge the *feature vectors* of each example, and then train a common classifier model. However, it is often more beneficial to train separate machine learning models for different types of features, as these may require different meta-parameter settings for optimal performance. Therefore we utilized the second approach called *late fusion* (Snoek et al., 2005), where we train separate machine learning methods for each feature set. To combine the outputs of the two models, we suggest taking the weighted mean of the posterior probabilities, which we found to be a simple-yet-robust technique (see e.g. Gosztolya et al., 2016a; Gosztolya et al., 2017). Since, in general, the linguistic features led to higher accuracy scores than the acoustic features did, we will compare the scores obtained via feature set fusion with the scores obtained by using the linguistic attributes (see Tables 9 and 10).

5.2. Results

Table 11 shows the various accuracy values obtained using late classifier fusion. We can see that 3-class classification accuracy improved in every case. When we combined all the linguistic features with the acoustic indicators, we obtained the highest scores: the 3-class accuracy score of 69.3%, in our opinion, reflects a good classification performance, while the binary F_1 value of 86.3% is also quite high. Note, however, that a very similar performance

Table 11
The various accuracy scores obtained by combining the "extended" feature set with the different linguistic feature subsets via late classifier fusion, in the 3-class task.

Feature set		3-class	2-class					
Acoustic	Linguistic	Accuracy	Accuracy	UAR	Precision	Recall	Specificity	F_1
Extended	Morphological	61.3%	74.7%	74.0%	84.4%	76.0%	60.0%	80.0
	Speech-based	66.7%	78.7%	79.0%	88.6%	78.0%	64.5%	83.0
	Semantic	62.7%	80.0%	81.0%	90.7%	78.0%	65.6%	83.8
	Morph. + Speech-based	68.0%	80.0%	79.0%	87.2%	82.0%	67.9%	84.5
	Morph. + Semantic	61.3%	76.0%	74.0%	83.3%	80.0%	63.0%	81.6
	Speech-based + Semantic	68.0%	82.7%	83.0%	91.1%	82.0%	70.0%	86.3
	All	69.3%	82.7%	83.0%	91.1%	82.0%	70.0%	86.3

Table 12
Accuracy values obtained by combining the “extended” feature set with all linguistic features in the 2-class machine learning tasks.

Speaker groups	Accuracy	UAR	Precision	Recall	Specificity	F_1
Control vs. MCI + mAD	81.3%	80.0%	87.5%	84.0%	70.4%	85.7
Control vs. MCI	80.0%	80.0%	85.7%	72.0%	75.9%	78.3
Control vs. mAD	86.0%	86.0%	84.6%	88.0%	87.5%	86.3
MCI vs. mAD	80.0%	80.0%	75.9%	88.0%	85.7%	81.5

Table 13
Accuracy values obtained by combining the “extended” feature set with the semantic linguistic features in the 2-class machine learning tasks.

Speaker groups	Accuracy	UAR	Precision	Recall	Specificity	F_1
Control vs. MCI + mAD	80.0%	74.0%	80.7%	92.0%	77.8%	86.0
Control vs. MCI	86.0%	86.0%	87.5%	84.0%	84.6%	85.7
Control vs. mAD	84.0%	84.0%	81.5%	88.0%	87.0%	84.6
MCI vs. mAD	78.0%	78.0%	73.3%	88.0%	85.0%	80.0

could be achieved by using the acoustic, speech-based and semantic attributes, i.e. omitting morphological features. This, in our opinion, indicates that from all the feature types tested in our current study, morphological attributes are the least useful ones for detecting MCI and mAD; this hypothesis is also supported by the other rows of Table 11.

Table 12 lists the results obtained via training binary classifiers, combining the output of the models got using all the linguistic features and the extended acoustic feature set. These scores display similar tendencies as those presented earlier: the proposed method of classification can distinguish subjects having some sort of dementia from control subjects with quite high efficiency, achieving a 81.3% accuracy score, 80% UAR score and a 85.7% F_1 score. Subjects having mild Alzheimer’s and those belonging to the control group can be distinguished relatively easily, which is reflected in the 86% accuracy and 86.3% F-measure scores; the other two class pairs are somewhat harder to distinguish, but we consider the accuracy score of 80% for both cases to be quite high.

Repeating the same series of experiments but with the semantic features only (see Table 13) led to similar, although usually slightly lower scores: control subjects and those being in the early stages of Alzheimer’s were the easiest to differentiate, but the accuracy scores were above 78% and the F-measure scores around 85% in most cases. Notice that, even though we used just the semantic attributes from the linguistic ones, our scores still exceeded both those obtained by using the acoustic features (see Table 8) and those got by all the linguistic markers, but without the help of the acoustic attributes (see Table 10). We find this result especially interesting, because the semantic attributes mostly consisted of the number of occurrences of specific keywords and expressions that reflect the uncertainty of the speaker. By using spoken term detection techniques (see e.g. Junkawitsch et al., 1996; Gosztolya and Tóth, 2011; Lee et al., 2016), such keywords can be located with high accuracy, allowing the automatic extraction of these attributes. This, however, is the subject of future work.

6. Conclusions

Alzheimer’s disease is a very distinct neurodegenerative disorder that develops for years before clinical manifestation, while mild cognitive impairment is clinically considered as a prodromal stage of AD. For both types of neurodegenerative disorders, early diagnosis is crucial in order to allow timely treatment to decelerate progression. In this study, extending our previous studies, we sought to differentiate the three speaker groups (i.e. healthy controls, those having MCI and those having mild AD) by relying on automatically extracted acoustic markers from the spontaneous speech of the subjects. We also utilized various morphological, speech-based and semantic linguistic features, calculated from the transcription of the spontaneous utterances. However, we got the best results when we combined the two different feature types. One promising finding of our study was that, by making use of only the semantic linguistic attributes, the accuracy scores obtained decreased only slightly relative to the case of using all the linguistic features. Since these semantic attributes express the presence of specific keywords or key expression in the utterance, we may expect that they may be easily estimated by spoken term detection techniques in the future.

Acknowledgments

Ildikó Hoffmann was supported by the EU-funded Hungarian grant EFOP-3.6.1-16-2016-00008, Gábor Gosztolya was supported by the National Research, Development and Innovation Office of Hungary via contract NKFIH FK-124413, László Tóth was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences, and Veronika Vincze was supported by the ÚNKP-17-4 New National Excellence Program of the Ministry of Human Capacities.

References

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.

Barth, S., Schönknecht, P., Pantel, J., Schröder, J., 2005. Mild cognitive impairment and Alzheimer's disease: an investigation of the CERAD-NP test battery. Fortschritte der Neurologie-Psychiatrie 73 (10), 568–576.

Baum, S.R., Blumstein, S.E., Naeser, M.A., Palumbo, C.L., 1990. Temporal dimensions of consonant and vowel production: An acoustic and CT scan analysis of aphasic speech. Brain Lang. 39 (1), 33–56.

Bayles, K.A., 1993. Pathology of language behaviour in dementia. In: Blanken, G., Dittmann, J., Grimm, H., Marshall, J.C., Wallesch, C.-W. (Eds.), Linguistic Disorders and Pathologies. de Gruyter Berlin/New York, pp. 388–409.

Bucks, R., Singh, S., Cuerden, J., Wilcock, G., 2000. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: evaluation of an objective technique for analysing lexical performance. Aphasiology 14 (1), 71–91.

Cardoso, S., Silva, D., Maroco, J., de Mendonca, A., Guerreiro, M., 2014. Non-literal language deficits in Mild Cognitive Impairment. Psychogeriatrics 14 (4), 222–228.

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.

Croot, K., Hodges, J.R., Xuereb, J., Patterson, K., 2000. Phonological and articulatory impairment in Alzheimer's disease: a case series. Brain Lang. 75 (2), 277–309.

Cuetos, F., Martinez, T., Martinez, C., Izura, C., Ellis, A.W., 2003. Lexical processing in Spanish patients with probable Alzheimer's disease. Cognit. Brain Res. 17 (3), 549–561.

Cummings, L., Mega, S., 2003. Neuropsychiatry and Behavioral Neuroscience. Oxford University Press.

dos Santos, L.B., Corrêa Jr., E.A., Oliveira Jr., O.N., Amancio, D.R., Mansur, L.L., Aluísio, S.M., 2017. Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. In: Proceedings of ACL. Vancouver, Canada, pp. 1284–1296.

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

Forbes-McKay, K., Venneri, A., 2005. Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. Int. J. Geriatric Psychiatry 2005 (26), 243–256.

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.

Fraser, K.C., Fors, K.L., Kokkinakis, D., Nordlund, A., 2017. An analysis of eye-movements during reading for the detection of mild cognitive impairment. In: Proceedings of EMNLP. Copenhagen, Denmark, pp. 1027–1037.

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.C., Rudzicz, F., Rochon, E., 2013. Using text and acoustic features to diagnose progressive aphasia and its subtypes. In: Proceedings of Interspeech. Lyon, France, pp. 25–29.

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

Galvin, J.E., Sadowsky, C.H., 2012. Practical guidelines for the recognition and diagnosis of dementia. J. Am. Board Family Med. 25 (3), 367–382 .

Garrard, P., Maloney, L.M., Hodges, J.R., Patterson, K., 2005. The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. Brain 128 (2), 250–260.

Garrard, P., Rentoumi, V., Gesierich, B., Miller, B., Gorno-Tempini, M.L., 2014. Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. Cortex 55, 122–129.

Gósy, M., 1998. The paradox of speech planning and production. Magyar Nyelvőr 12 (1), 3–15.

Gósy, M., 2012. BEA: A multifunctional Hungarian spoken language database. Phonetician 105 (106), 50–61.

Gosztolya, G., Busa-Fekete, R., Grósz, T., Tóth, L., 2017. DNN-based feature extraction and classifier combination for child-directed speech, cold and snoring identification. In: Proceedings of Interspeech. Stockholm, Sweden, pp. 3522–3526.

Gosztolya, G., Grósz, T., Szaszák, Gy., Tóth, L., 2016. Estimating the sincerity of apologies in speech by DNN rank learning and prosodic analysis. In: Proceedings of Interspeech. San Francisco, CA, USA, pp. 2026–2030.

Gosztolya, G., Tóth, L., 2011. Spoken term detection based on the most probable phoneme sequence. In: Proceedings of SAMI. Smolenice, Slovakia, pp. 101–106.

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.

Grzybowska, J., Kacprzak, S., 2016. Speaker age classification and regression using i-vectors. In: Proceedings of Interspeech. San Francisco, CA, USA, pp. 1402–1406.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. ACM SIGKDD Explor. Newslett. 11 (1), 10–18.

Hirst, G., Wei Feng, V., 2012. Changes in style in authors with Alzheimer's disease. Engl. Stud. 93 (3), 357–370.

Hodges, J., Salmon, D., Butters, N., 1992. Semantic memory impairment in Alzheimer's disease: failure of access or degraded knowledge? Neuropsychologia 30 (4), 301–314.

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.

Holmes, D.I., Singh, S., 1996. A stylometric analysis of conversational speech of aphasic patients. Literary Linguist. Comput. 11 (3), 133–140.

Illes, J., 1989. Neurolinguistic features of spontaneous language production dissociate three forms of neurodegenerative disease: Alzheimer's, Huntington's, and Parkinson's. Brain Lang. 37 (4), 628–642.

de Ipiña, K.L., Alonso, J.-B., Travieso, C.M., Sol-Casals, J., Egiraun, H., Faundez-Zanuy, M., Ezeiza, A., Barroso, N., Ecay-Torres, M., Martinez-Lage, P., de Lizardui, U.M., 2013. On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. Sensors 13 (5), 6730–6745.

Janka, Z., Somogyi, A., Maglóczky, E., Pákáski, M., Kálmán, J., 1988. Dementia szűrővizsgálat cognitív gyorsteszt segítségével. Orvosi hetilap 129, 297–299.

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M.L., Ogar, J., 2014. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In: Proceedings of CLPsych. Baltimore, Maryland, USA, pp. 27–37.

Juncos-Rabadán, O., Pereiro, A.X., Facaly, D., Rodríguez, N., 2010. Una revisión de la investigación sobre lenguaje en el deterioro cognitivo level. Revista de Logopedia, Foniatría y Audiología 30 (2), 73–83.

Junkawitsch, J., Neubauer, L., Höge, H., Ruske, G., 1996. A new keyword spotting algorithm with pre-calculated optimal thresholds. In: Proceedings of ICSLP, 4. Philadelphia, PA, USA, pp. 2067–2070.

Kálmán, J., Maglóczky, E., Janka, Z., 1995. Óra Rajzolási Teszt: gyors és egyszerű dementia szűrőmódszer. Psychiatria Hungarica 10 (3), 11–18.

Kemper, S., Marquis, J., Thompson, M., 2001. Longitudinal change in language production: effects of aging and dementia on grammatical complexity and propositional concent. Psychol. Aging 16 (4), 600–614.

Kempler, D., Zelinski, E.M., 1994. Language in dementia and normal aging. In: Huppert, F.A., Brayne, C., O'Connor, D.W. (Eds.), Dementia and Normal Aging. Cambridge University Press Cambridge, pp. 331–365.

Kockmann, M., Burget, L., Cernocký, J., 2010. Brno University of Technology system for Interspeech 2010 Paralinguistic Challenge. In: Proceedings of Interspeech. Makuhari, Chiba, Japan, pp. 2822–2825.

Kokkinakis, D., Fors, K.L., Björkner, E., Nordlund, A., 2017. Data collection from persons with mild forms of cognitive impairment and healthy controls − infrastructure for classification and prediction of dementia. In: Proceedings of NoDaLiDa. Gothenburg, Sweden, pp. 172–182.

Kumar, N., Nasir, M., Georgiou, P., Narayanan, S.S., 2016. Robust multichannel gender classification from speech in movie audio. In: Proceedings of Interspeech. San Francisco, CA, USA, pp. 2233–2237.

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 Dementia 11 (5), 561–578.

Le, X., Lancashire, I., Hirst, G., Jokel, R., 2011. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. Literary Linguist. Comput. 26 (4), 435–461.

Lee, S., Tanaka, K., Itoh, Y., 2016. Generating complementary acoustic model spaces in DNN-based sequence-to-frame DTW scheme for out-of-vocabulary Spoken Term Detection. In: Proceedings of Interspeech. San Francisco, CA, USA, pp. 755–759.

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.

Light, L.L., 1993. Language changes in old age. In: Blanken, G., Dittmann, J., Grimm, H., Marshall, J.C., Wallesch, C.-W. (Eds.), Linguistic disorders and pathologies. de Gruyter Berlin/New York, pp. 900–918.

Lunsford, R., Heeman, P.A., 2015. Using linguistic indicators of difficulty to identify mild cognitive impairment. In: Proceedings of Interspeech. Dresden, Germany, pp. 658–662.

Matsuda, H., Asada, T., Tokumaru, A.M., 2017. Neuroimaging Diagnosis for Alzheimer's Disease and Other Dementias. Springer.

Meilán, J.J., Martínez-Sánchez, F., Carro, J., López, D., Millian-Morell, L., Arana, J., 2014. Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia? Dement. Geriatric Cognit. Disord. 37 (5−6), 327–334.

Meilán, J.J.G., Martínez-Sánchez, F., Carro, J., Sánchez, J.A., Pérez, E., 2012. Acoustic markers associated with impairment in language processing in Alzheimer's disease. Span. J. Psychol. 15 (2), 487–494.

Meinedo, H., Trancoso, I., 2011. Age and gender detection in the I-DASH project. ACM Trans. Speech Lang. Process. 7 (4), 13.

Nakata, Y., Sato, N., Nemoto, K., Abe, O., Shikakura, S., Arima, K., Furuta, N., Uno, M., Hirai, S., Masutani, Y., Ohtomo, K., Barkovich, A.J., Aoki, S., 2009. Diffusion abnormality in the posterior cingulum and hippocampal volume: correlation with disease progression in Alzheimer's disease. Mag. Resonance Imag. 27 (3), 347–354.

Nelson, L., Tabet, N., 2015. Slowing the progression of Alzheimer's disease; what works? Ageing Res. Rev. 23 (B), 193–209.

Patocskai, A., Pákáski, M., Vincze, G., Fullajtár, M., Szimjanovszki, I., Boda, K., Janka, Z., Kálmán, J., 2014. Is there any difference between the findings of clock drawing tests if the clocks show different times? Int. J. Geriatric Psychiatry 39 (4), 749–757.

Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., Ferri, C., 2013. The global prevalence of dementia: a systematic review and metaanalysis. Alzheimer's Dementia 9 (1), 63–75.

Ramabhadran, B., Huang, J., Picheny, M., 2003. Towards automatic transcription of large spoken archives − English ASR for the MALACH project. In: Proceedings of ICASSP, pp. 216–219.

Rapp, A.M., Wild, B., 2011. Nonliteral language in Alzheimer dementia: a review. J. Int. Neuropsychol. Soc. 17 (2), 207–218.

Reichman, W., Cummings, J., 1999. Dementia. In: Duthie, E., Katz, P. (Eds.), Practice of Geriatrics. W.B. Saunders Philadelphia, PA. 395−304

Ripich, D., 1994. Functional communication with AD patient: a caregiver training program. Alzheimer Dis. Assoc. Disord. 8 (3), 95–109.

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

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

Sachdev, P., Lipnicki, D., Crawford, J., Reppermund, S., Kochan, N., Trollor, J., Draper, B., Slavin, M., Kang, K., Lux, O., Brodaty, K.M.H., 2012. Risk profiles for mild cognitive impairment vary by age and sex: the Sydney memory and ageing study. Am. J. Geriatric Psychiatry 20 (10), 854–865.

Santos, V.D., Thomann, P.A., Wüstenberg, T., Seidl, U., Essig, M., Schröder, J., 2011. Morphological cerebral correlates of CERAD test performance in mild cognitive impairment and Alzheimer's disease. J. Alzheimer's Dis. 23 (3), 411–420.

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.

Scheltens, P., Fox, N., Barkhof, F., Carli, C.D., 2002. Structural magnetic resonance imaging in the practical assessment of dementia: beyond exclusion. Lancet 1 (1), 13–21.

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., Williamson, R., 2001. Estimating the support of a high-dimensional distribution. Neural Comput. 13 (7), 1443–1471.

Shibata, D., Wakamiya, S., Aramak, E., 2016. Detecting Japanese patients with Alzheimer's Disease based on word category frequencies. In: Proceedings of ClinicalNLP. Osaka, Japan, pp. 78–85.

Smith, S., Murdoch, B.E., Chenery, H., 1989. Semantic abilities in dementia of the Alzheimer type 1: lexical semantics. Brain Lang. 36 (2), 314–324.

Snoek, C., Worring, M., Smeulders, A., 2005. Early versus late fusion in semantic video analysis. In: Proceedings of ACM International Conference on Multimedia. Singapore, Singapore, pp. 399–402.

Stricker, N., Schweinsburg, B., Delano-Wood, L., Wierenga, C., Bangen, K., Haaland, K., Frank, L., Salmon, D., Bondi, M., 2009. Decreased white matter integrity in late-myelinating fiber pathways in Alzheimer's disease supports retrogenesis. Neuroimage 45 (1), 10–16.

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 (195), 1–7.

Taler, V., Jarema, G., 2006. On-line lexical processing in AD and MCI: An early measure of cognitive impairment? J. Neurolinguist. 19 (1), 38–55 .

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

Thomas, C., Kešelj, V., Cercone, N., Rockwood, K., Asp, E., 2005. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In: Proceedings of the IEEE International Conference on Mechatronics and Automation, 3. IEEE, pp. 1569–1574.

Tóth, L., 2015. Phone recognition with hierarchical Convolutional Deep Maxout Networks. EURASIP J. Audio Speech Music Process. 2015 (25), 1–13.

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.

Tsantali, E., Economidis, D., Tsolaki, M., 2013. Could language deficits really differentiate mild cognitive impairment (MCI) from mild Alzheimer's disease. Arch. Gerontol. Geriatrics 57 (3), 263–270.

Vincze, V., 2014. Uncertainty detection in Hungarian texts. In: Proceedings of Coling. Dublin, Ireland, pp. 1844–1853.

Vincze, V., Gosztolya, G., Tóth, L., Hoffmann, I., Szatlóczki, G., Bánréti, Z., Pákáski, M., Kálmán, J., 2016. Detecting Mild Cognitive Impairment by exploiting linguistic information from transcripts. In: Proceedings of ACL. Berlin, Germany, pp. 181–187.

Walla, P., Püregger, E., Lehrner, J., Mayer, D., Deecke, L., Dal Bianco, P., 2005. Depth of word processing in Alzheimer patients and normal controls: a magnetoencephalographic (MEG) study. J. Neural Transmission 112 (5), 713–730.

Weiner, J., Herff, C., Schultz, T., 2016. Speech-based detection of Alzheimer's Disease in conversational German. In: Proceedings of Interspeech. San Francisco, CA, USA, pp. 1938–1942.

Weiner, M.F., Neubecker, K.E., Bret, M.E., Hynan, L.S., 2008. Language in Alzheimer's Disease. J. Clin. Psychiatry 69 (8), 1223–1227.

Yin, C., Li, S., Zhao, W., Feng, J., 2013. Brain imaging of mild cognitive impairment and Alzheimer's disease. Neural Regeneration Res. 8 (5), 435–444.

Young, S., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2006. The HTK Book. Cambridge University Engineering Department Cambridge, UK.

Zimny, A., Szewczyk, P., Trypka, E., Wojtynska, R., Noga, L., Leszek, J., Sasiadek, M., 2011. Multimodal imaging in diagnosis of Alzheimer's disease and amnestic mild cognitive impairment: value of magnetic resonance spectroscopy, perfusion, and diffusion tensor imaging of the posterior cingulate region. J. Alzheimer's Dis. 27 (3), 435–444.

Zsibrita, J., Vincze, V., Farkas, R., 2013. magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP, pp. 763–771.