Computer Speech & Language 69 (2021) 101215

Contents lists available at ScienceDirect

Computer Speech & Language

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

Cross-lingual detection of mild cognitive impairment based on temporal parameters of spontaneous speech

Gábor Gosztolya*[a,b], Réka Balogh[c], Nóra Imre[c], José Vicente Egas-López[b], Ildikó Hoffmann[d,e], Veronika Vincze[a,b], László Tóth[b], Davangere P. Devanand[f,g], Magdolna Pákáski[c], János Kálmán[c]

[a] MTA-SZTE Research Group on Artificial Intelligence, Eötvös Loránd Research Network, Szeged, Hungary
[b] Institute of Informatics, University of Szeged, Szeged, Hungary
[c] Department of Psychiatry, University of Szeged, Szeged, Hungary
[d] Department of Linguistics, University of Szeged, Szeged, Hungary
[e] Research Center for Linguistics, Eötvös Loránd Research Network, Budapest, Hungary
[f] Department of Psychiatry, Columbia University Medical Center, New York, NY, USA
[g] Division of Geriatric Psychiatry, New York State Psychiatric Institute, New York, NY, USA

ARTICLE INFO

Article History:
Received 30 July 2020
Revised 11 December 2020
Accepted 3 March 2021
Available online 14 March 2021

Keywords:
Mild cognitive impairment
Dementia
Temporal speech parameters
Automatic speech recognition
Cross-linguality

ABSTRACT

Mild Cognitive Impairment (MCI) is a heterogeneous clinical syndrome, often considered as the prodromal stage of dementia. It is characterized by the subtle deterioration of cognitive functions, including memory, executive functions and language. Mainly due to the tenuous nature of these impairments, a high percentage of MCI cases remain undetected. There is evidence that language changes in MCI are present even before the manifestation of other distinctive cognitive symptoms, which offers a chance for early recognition. A cheap non-invasive way of early screening could be the use of automatic speech analysis. Earlier, our research team developed a set of speech temporal parameters, and demonstrated its applicability for MCI detection. For the automatic extraction of these attributes, a Hungarian-language ASR system was employed to match the native language of the MCI and healthy control (HC) subjects. In practical applications, however, it would be convenient to use exactly the same tool, regardless of the language spoken by the subjects. In this study we show that our temporal parameter set, consisting of articulation rate, speech tempo and various other attributes describing the hesitation of the subject, can indeed be reliably extracted regardless of the language of the ASR system used. For this purpose, we performed experiments both on English-speaking and on Hungarian-speaking MCI patients and healthy control subjects, using English and Hungarian ASR systems in both cases. Our experimental results indicate that the language on which the ASR system was trained only slightly affects the MCI classification performance, because we got quite similar scores (67-92%) as we did in the monolingual cases (67-92% as well). As our last investigation, we compared the proposed attribute values for the same utterances, utilizing both the English and the Hungarian ASR models. We found that the articulation rate and speech tempo values calculated based on the two ASR models were highly correlated, and so were the attributes corresponding to silent pauses; however, noticeable differences were found regarding the filled pauses (still, these attributes remained indicative for both languages). Our further analysis revealed that this is probably due to a difference regarding the annotation of the English and the Hungarian ASR training utterances.

*Corresponding author at: Gábor Gosztolya, MTA-SZTE Research Group on Artificial Intelligence, H-6720 Szeged, Hungary, Tisza Lajos krt. 103.
E-mail address: ggabor@inf.u-szeged.hu (G. Gosztolya).

https://doi.org/10.1016/j.csl.2021.101215
0885-2308/© 2021 Published by Elsevier Ltd.

1. Introduction

Dementia is a chronic or progressive clinical syndrome, affecting mainly elderly people worldwide. It is characterized by the deterioration of memory, language and problem-solving skills, which are severe enough to adversely affect the patients' ability to carry out everyday activities (Alzheimer's Association, 2020). According to the estimates, the number of affected individuals, which at present exceeds 46.8 million, may double by 2050 (Prince et al., 2015). Since the currently available therapeutic agents are shown to be more effective in the earliest or preclinical stages of dementia (Szatlóczki et al., 2015), recognizing the disease in the earliest phase is of utmost importance.

The most widely used term to describe the preclinical stage of dementia is Mild Cognitive Impairment (MCI), which condition is often considered to be the borderline between normal aging and dementia (Petersen et al., 2014). This syndrome shows similar characteristics to dementia, although in the case of MCI the symptoms do not interfere with the patients' activity of daily living (Foster et al., 2019). However, given its high conversion rate to dementia (2−31% annually, see e.g. Bruscoli and Lovestone, 2004), MCI should be regarded as a severe condition. As the transition phase from MCI to dementia can last even 15 years (Laske et al., 2015), there is a wide time window in which the subtle signs of cognitive decline could be detected. Since the timely identification of MCI could provide more effective therapeutic interventions to delay progression, the importance of developing methods that allow early recognition has been emphasized in the recent years.

It has been shown that changes in language production are associated with subclinical declines in memory, e.g. the fluency of spontaneous speech has been proven to deteriorate in people with early MCI (Mueller et al., 2018). During the course of the disease, filled pauses (i.e. vocalizations like ‘uhm’ and ‘er’) and disfluencies become more and more frequent in the speech (de Ipiña et al., 2018), corresponding to the word-finding or word-retrieval difficulties of patients (Szatlóczki et al., 2015). Earlier studies also indicated that compared to healthy controls, MCI patients tend to have lower speech rate, and an increased number and length of hesitations (Szatlóczki et al., 2015). The above-mentioned characteristics can strongly influence the overall time course of the speech; therefore, the analysis of such temporal aspects can help us explore the relationships between language and memory.

Since collecting speech samples is a quick, non-invasive, cheap and relatively easy way of gathering data and as it minimizes the burden of the examination for the patients, the analysis of speech is a promising method for early MCI screening. There are several established methods for obtaining speech samples from participants. On the one hand, structured speech samples can be obtained by applying spoken tasks e.g. reading, counting backwards or sentence repeating (König et al., 2018; Fraser et al., 2019). On the other hand, unstructured or spontaneous speech samples can be collected, which seems to be a reliable source for language analysis as well. Spontaneous speech can be elicited by using narrative recall tasks (e.g. picture description tasks) or by asking the participants to talk about a given topic (e.g. their hobbies, or their previous day). According to some results, the latter type might be more sensitive when discriminating between HC and MCI subjects (Beltrami et al., 2018).

In the last decade, numerous attempts have been made to distinguish cognitively healthy control (HC) subjects from people with MCI or with Alzheimer's disease (AD) using different speech analysis techniques. In the earlier studies, analyzed speech features were extracted mainly from manually transcribed data, which is rather labor-intensive. In more recent studies the goal was to find out whether extraction by automated techniques could produce similar results. In the past few years, several such automatic speech analysis studies have been published (e.g. König et al., 2015; Laske et al., 2015; de Ipiña et al., 2018; König et al., 2018; Themistocleous et al., 2018; Tóth et al., 2018; Fraser et al., 2019; Gosztolya et al., 2019; Sluis et al., 2020; Themistocleous et al., 2020).

In the previous studies conducted by our team, we developed a set of temporal parameters that characterize hesitation in the spontaneous speech of the subjects (Tóth et al., 2015; Gosztolya et al., 2016; Tóth et al., 2018; Gosztolya et al., 2019). Hesitation is defined as an absence of speech, and it can be divided into two categories: silent pauses and filled pauses; while 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), our temporal attribute set also expressed the amount of *filled* pauses in the speech of the subject. We also experimentally demonstrated that these temporal parameters can readily be used as features in a subsequent machine learning step (e.g. classification by using Support Vector Machines SVM, Schölkopf et al., 2001). Furthermore, although in our initial studies we calculated these temporal attributes during a manual annotation process, later we demonstrated that they can efficiently be calculated by using speech processing tools, i.e. by relying on a phone-level Automatic Speech Recognition (ASR) framework.

In these studies, however, we performed our experiments in a monolingual environment. That is, both the healthy control subjects and those suffering from mild cognitive impairment were native Hungarian speakers, and the ASR tool we employed to process their speech was also a Hungarian one (i.e. it used a Hungarian phone set, a Hungarian language model and the Deep Neural Network (DNNs, Hinton et al., 2012; Tóth, 2015) used as the acoustic model was trained on Hungarian utterances). Of course, this is quite common in speech processing, and most studies also approach the MCI detection task strictly in a language-dependent manner. For example, Garcia et al. applied i-vectors for detecting Parkinson's Disease, and they both trained the Universal Background Model (UBM) of the i-vector and extracted the i-vector features on a Spanish language corpus (García et al., 2018). Similarly, other studies use spontaneous spoken utterances in English language to predict MCI by means of linguistic features (Asgari et al., 2017; Fritsch et al., 2019), while Themistocleous et al. carried out the identification of MCI from the speech of Swedish subjects by means of Deep Sequential Neural Networks trained on a Swedish corpus (Themistocleous et al., 2018).

While training the applied tools on data of the same language as the speech of the MCI and HC subjects might be well founded for some types of attributes, in other cases one might expect language-independence (at least to some extent), therefore this monolingual restriction might be obsolete. For example, i-vectors (Dehak et al., 2011) are quite general models developed for speaker segmentation and verification, and as such, they are probably practically language-independent. Similarly, x-vectors (Snyder et al., 2018) (which can be viewed as a deep learning-based improvement over i-vectors) could also be expected to work over different languages, regardless of the language of the utterances used for training the model.

From our set of temporal parameters, silent pauses can be expected to be detectable in a language-independent manner. However, the language-independence of filled pauses is not that straightforward; for example, de Leeuw found that the frequency of vocalic and vocalic-nasal hesitation markers significantly differed among Dutch, English and German speakers (de Leeuw, 2007). Nevertheless, recently several studies presented successful cross-lingual machine learning experiments in filler detection (e.g. Brueckner et al., 2017; Vetter et al., 2019), which suggests that our attribute set could also be calculated language-independently. Our attribute set contains temporal parameters describing the speech rate of the speaker as well, expressed as the uttered phones per second. Obviously, trying to fit phones of a different language to the given utterance will change the values of such measurements; still, one might expect that these can tolerate some phonetic-level misidentifications, and using an ASR system trained on a different language might only introduce slight changes in the calculated values, which makes the attributes just as indicative as they are in the monolingual case. Our hypothesis is that attribute sets like the one proposed in our previous works (e.g. in Gosztolya et al., 2019) might permit MCI detection with a similar performance when the ASR tool used for feature extraction is trained on a different language.

In this study, we experimentally investigate the language independence of the hesitation-based set of temporal speech parameters. For this, we collected the spontaneous speech of English-speaking and Hungarian-speaking MCI patients and healthy controls. Then we trained two ASR models for the automatic speech analysis step: for English, we used a subset of the TEDLium corpus (Rousseau et al., 2012), while for Hungarian we used a subset of the BEA Hungarian Spoken Language Database (Neuberger et al., 2014). We carried out classification experiments to determine the indicativeness of the different attributes. We performed both monolingual and cross-lingual experiments; furthermore, we examined the similarity of the temporal parameters calculated by the two ASR systems.

The structure of this paper is as follows. In Section 2, we describe the two MCI databases we used in our experiments. Then, in Section 3, we describe the acoustic markers we extracted from the spontaneous speech of the subjects. Next, in Section 4, we present the experimental setup of subject classification: the classification method used, the way of setting the meta-parameters, and the means of evaluation. Then we present and compare our monolingual (Section 5) and cross-lingual (Section 6) results. Lastly, in Section 7, we compare the attributes extracted by the two approaches.

2. The MCI-HC recordings

A total of 88 elderly individuals were recruited in parallel from two outpatient clinics. The two institutions were:

- Memory Disorders Center of the Department of Psychiatry, New York State Psychiatric Institute and Columbia University (New York, NY, USA), and
- Memory Clinic, Department of Psychiatry, University of Szeged (Szeged, Hungary).

Exclusion criteria included deafness, manifest speech problems (any form of aphasia), significant articulation problems (e.g. stutter), history of substance use disorder, history of stroke, previous CT or MRI showing evidence of significant abnormality that would suggest another potential etiology for MCI or dementia. In the end, a total of 66 subjects were eligible for final inclusion.

Both the English-speaking ($n = 33$, New York, NY, USA) and the Hungarian-speaking ($n = 33$, Szeged, Hungary) participants were diagnosed as MCI or HC. This decision was based on Petersen's criteria (Petersen et al., 1999) in both language groups, with the Mini-Mental State Examination (MMSE, Folstein et al., 1975) serving as a measure for objective cognitive impairment in the Hungarian-speaking sample (30-28 points: HC; 27-24 points: MCI). In all other aspects, the inclusion/exclusion criteria were the same at both sites. The inclusion criteria were a minimum age of 60 years, a minimum of 8 years of formal education, and English/Hungarian as the native language (corresponding to the country of recruitment). Bilingualism was not taken into account.

To get an overview of the participants' characteristics and to acquire eligibility data, an interview focused on demographic features and medical history was administered, as well as a brief neuropsychological test battery (including the MMSE, the Clock Drawing Test (CDT, Manos and Wu, 1994), and the Geriatric Depression Scale (GDS)). All individuals were screened for possible dementia using the MMSE, and those with a score under 24 were not involved in further participation. Following institutional protocols, the possibility of depression was also evaluated based on the 30-item (Yesavage et al., 1983) or the 15-item (Sheikh and Yesavage, 1986) version of the GDS (GDS-30 and GDS-15, for the English-speaking sample and for the Hungarian-speaking sample, respectively): patients scoring above 10 on GDS-30 or above 5 on GDS-15 were excluded from the study.

The investigation was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants at both sites. The part of the study conducted in New York was approved by the Institutional Review Board of the New York State Psychiatric Institute — Columbia University Department of Psychiatry (protocol number: 7611). The part of the study conducted in Szeged was approved by the Regional Human Biomedical Research Ethics Committee of the University of Szeged, Hungary (reference number: 231/2017-SZTE).

Table 1
Demographic characteristics (i.e. age, gender and years of education) and the results of the neurophysiological tests (MMSE, CDT and GDS) tests of the subject groups.

	English-speaking sample		Hungarian-speaking sample	
	HC (n = 19)	**MCI** (n = 14)	**HC** (n = 20)	**MCI** (n = 13)
Age (mean ± SD)	74.47 ± 7.321	72.36 ± 6.857	69.90 ± 5.609	73.77 ± 4.969
Gender (m / f)	5 / 14	6 / 8	3 / 17	4 / 9
Years of education (mean ± SD)	17.84 ± 3.532	16.79 ± 3.118	13.15 ± 2.455	11.77 ± 2.743
MMSE score (mean ± SD)	29.16 ± 1.015	27.71 ± 1.773	28.28 ± 0.813	26.31 ± 0.751
CDT score (mean ± SD)	8.89 ± 1.197	9.21 ± 1.188	7.60 ± 3.152	7.92 ± 2.178
GDS-30 / GDS-15 score (mean ± SD)	3.16 ± 2.853	5.50 ± 2.822	1.65 ± 1.387	2.77 ± 1.013

2.1. Demographics and neuropsychological test performances

Detailed demographic characteristics and neuropsychological test scores of all groups (means and standard deviations) are presented in Table 1. Concerning demographics (age, gender, and years of education) and the CDT test, there were no statistically significant differences between the MCI and the HC groups in either of the languages. However, regarding the other neuropsychological tests, MCI patients achieved a significantly poorer performance in the MMSE than HCs (English-speaking sample: $U = 62.500$; $Z = -2.703$; $p = .009$; Hungarian-speaking sample: $U = 0.000$; $Z = -4.879$; $p < .001$), and they also had higher scores in the GDS in both languages (English-speaking sample: $U = 71.000$; $Z = -2.277$; $p = .024$; Hungarian-speaking sample: $U = 59.000$; $Z = -2.736$; $p = .008$).

2.2. The recording protocol

After the clinical evaluation, spontaneous speech samples were obtained from all the participants. During this process, one of the investigators was in the same room as the subject, while a second investigator called the subject on a mobile phone from another room. The first investigator informed the participant that a colleague would call him/her and asked the subject to pick up the phone when it rang. Then, the second investigator introduced herself over the phone and asked the participant to talk about his or her previous day (the standardized instruction was: “Hello, I’m . . . , can you hear what I’m saying? I would like you to tell me about your previous day in as much detail as you can.”). Each participant’s monologue was recorded by a call recorder application installed on the mobile phone device. The recordings obtained were then converted into an uncompressed PCM mono, 16-bit wav format with a sampling rate of 8,000 Hz. These utterances were further edited to contain only the speech of the subject (i.e. instructions and possible silent parts before/after his/her speech were removed).

3. Extracting acoustic markers from spontaneous speech

To investigate the spontaneous speech of MCI patients and HC subjects, we calculated specific temporal parameters from their spontaneous speech. We based our investigations on our previous studies (Tóth et al., 2015; Gosztolya et al., 2016; Hoffmann et al., 2017; Tóth et al., 2018; Gosztolya et al., 2019). To exploit and formalize the deterioration of the verbal fluency of the MCI subjects, we developed a set of temporal parameters which mostly focus on the amount and duration of hesitation in the speech of the subject.

This set of temporal parameters can be seen in Table 2. The articulation rate and speech tempo (i.e. parameters (1) and (2)) both describe how fast the subject speaks (although in a slightly different manner), while the duration of the utterance (parameter (3)) is related to the amount the subject could remember about his / her previous day. The remaining parameters ((4)–(7)) all describe the amount of hesitation in the spontaneous speech of the subject by focusing on the number or on the duration of pauses in some way. However, we did not clarify which hesitation types (that is, silent and/or filled pauses) we were using during this calculation. To be able to analyze both pause types, we included temporal parameters (4)–(7) in our set by calculating them

Table 2
The examined temporal speech parameters, based on our previous studies (Hoffmann et al., 2017; Tóth et al., 2018).

(1) Articulation rate was calculated as the number of phones per second during speech (excluding hesitations).
(2) Speech tempo (phones per second) was calculated as the number of phones per second divided by the total duration of the utterance.
(3) Duration of utterance, given in seconds.
(4) Pause occurrence rate was calculated by dividing the number of pause occurrences by the number of phones in the utterance.
(5) Pause duration rate was calculated by dividing the total duration of pauses by the length of the utterance.
(6) Pause frequency was calculated by dividing the number of pause occurrences by the length of the utterance.
(7) Average pause duration was calculated by dividing the total duration of pauses by the number of pauses.

for silent pauses only, for filled pauses only, and for taking all pause occurrences into account regardless of type. This led to 3 × 4 = 12 variations, hence we calculated 15 temporal parameters overall.

In our experiments we will also use specific subsets of these attributes. By ‘tempo-related attributes’ (or ‘tempo’ for short) we mean articulation rate, speech tempo and the duration of the utterance. By ‘silence-related’, we mean parameters (4) to (7) calculated for the silent pauses only; similarly, ‘filler-related’ and ‘all pause-related’ mean these attributes when calculated for the filled pauses only, and for all pause types, respectively.

3.1. *Automatic acoustic marker extraction using ASR*

While in our early studies we calculated the above-listed (or very similar) acoustic markers manually (i.e. using Praat Boersma, 2001), this process was actually quite expensive and required skilled labor. Therefore, later we sought to automate this step by using automatic speech processing techniques. Although distinguishing the silent parts and those containing speech can be done in an automated way quite easily (see e.g. Satt et al., 2014), these simple techniques cannot extract all the features of Table 2: most importantly, they cannot distinguish filled pauses from speech. Unfortunately, an off-the-shelf ASR tool (like the one employed by Fraser et al. (2013)) may also be suboptimal for several reasons. Firstly, 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 and the duration of silent and filled pauses. Secondly, 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 thirdly, 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), which in our case also makes a standard ASR tool more prone to errors.

However, notice that our speech-related markers listed in Table 2 do not require the correct identification of the phonemes: we only need to *count* them. The only phenomena we need to take special care of are the two forms of hesitation: i.e. silent and filled pauses. Because of this, we decided to use a speech recognizer that just provides a phone sequence as output (including filled pause as a special phonetic label). This allows the automatic extraction of acoustic markers, which can then be employed to perform automatic subject categorization via machine learning techniques. For our simplified workflow, see Fig. 1. Of course, recognizing the spontaneous speech of elderly people is known to be difficult (Ramabhadran et al., 2003); and when one attempts to do this without a vocabulary (i.e. only at the phonetic level), the number of errors can be expected to rise even further. 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; Tóth et al., 2018; Gosztolya et al., 2019) we found that this kind of automatic feature extraction was feasible for distinguishing speakers having MCI from healthy controls.

3.2. *ASR parameters*

Due to the bilingual nature of our study, we employed two datasets to train the DNN acoustic models of our two ASR systems (i.e. English and Hungarian). Of course, in both cases we had to employ audio datasets consisting of spontaneous speech, as only spontaneous speech is expected to contain filled pauses. For English, we used the TEDLium dataset (Rousseau et al., 2012); we made use of the utterances of 100 speakers (approximately 15 h of recordings). For Hungarian, we chose the BEA Database (Neuberger et al., 2014); we trained our DNNs on the speech of 116 subjects (44 h of recordings overall). We made sure that the annotation suited our needs for both corpora, i.e. filled pauses, breathing sounds, laughter, coughs and gasps were marked 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 modelling we used standard feed-forward Deep Neural Networks. The acoustic DNNs had an identical structure for both languages: they had 5 hidden layers, each consisting of 1024 ReLU neurons, while we used softmax neurons in the output layer. We used 40 Mel-frequency filter banks along with raw energy as frame-level features, and included the first- and second-order derivatives (i.e. Δ and $\Delta\Delta$). To improve model accuracy, we evaluated our model on a sliding window with a width of 15 frames (1845 frame-level features overall). As a language model, we employed simple phone bigrams (again, including all the above-mentioned non-verbal audio tags); and, of course, these were trained independently on the English data and on the Hungarian data.

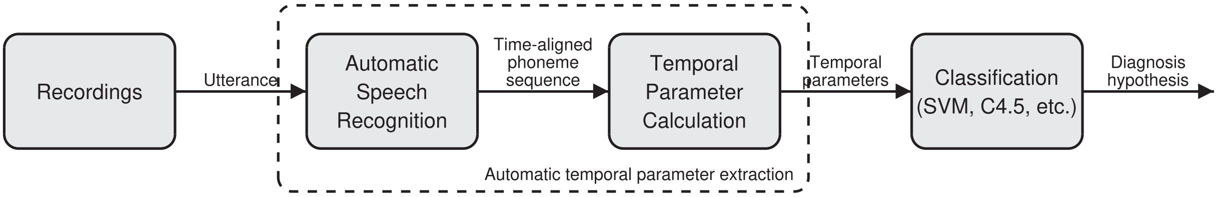
Recordings
Utterance
Automatic Speech Recognition
Time-aligned phoneme sequence
Temporal Parameter Calculation
Temporal parameters
Classification (SVM, C4.5, etc.)
Diagnosis hypothesis
Automatic temporal parameter extraction

Fig. 1. A simplified diagram of the applied subject classification workflow.

4. Subject classification by machine learning

A Support Vector Machine classifier was employed to discriminate the utterances of the subjects. We used the libSVM implementation (Chang and Lin, 2011); we tested both linear and Radial basis function (RBF) kernels. The C complexity parameter was set in the range $10^{-5}, 10^{-4}, \ldots, 10^{2}$, while in the case of the RBF kernel, γ was set in the range $2^{-10}, 2^{-9}, \ldots, 2^{5}$. The extracted temporal parameter vectors were standardized before using them in the machine learning step. From a machine learning perspective, having fewer than 100 examples (i.e. subjects) is an extremely small dataset. However, the number of diagnosed MCI patients is limited, collecting recordings of their speech and obtaining a medical diagnosis is time-consuming; therefore, in studies similar to ours it is quite common to have fewer than 100 patients (e.g. Satt et al., 2014; Asgari et al., 2017; Themistocleous et al., 2018; 2020; Sluis et al., 2020).

Having so few (i.e. $n = 33$ for each language) examples, we did not create separate training and test sets, but opted for speaker-wise cross-validation (CV); that is, we always trained our SVM model on the features extracted from the speech of 32 speakers. In the next step, this machine learning model was evaluated on the remaining speaker. We repeated this for all speakers, and aggregated the results into one final score. To balance class distribution during model training, we employed upsampling by repeating examples (i.e. subjects) belonging to the class with fewer examples; in our case, this was the MCI subject category for both languages.

The meta-parameters of SVM (i.e. C and γ) were determined by a technique called *nested cross-validation* (Cawley and Talbot, 2010). That is, each time we trained on the data of 32 subjects, we performed *another* (speaker-wise, i.e. 32-fold) cross-validation session, looking for the C and γ meta-parameter values that led to the highest AUC score. After this, we trained an SVM model with the selected meta-parameters on the data of all 32 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.

4.1. Evaluation

The choice of evaluation metric is not a clear-cut issue for this task. Perhaps the most straightforward choice is to use the traditional classification accuracy metric. However, since we have somewhat fewer subjects belonging to the MCI category than healthy controls, it might be beneficial to measure classification performance by other evaluation metrics as well. Therefore we decided to also report 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 of the above-listed metrics, sensitivity being equivalent to recall, while specificity is practically the recall of the negative class (in this case, healthy controls). We calculated these values at Equal Error Rate (EER); that is, we chose the decision threshold between the posterior estimates of the two classes (provided by the SVM classifier) in a way which minimizes the absolute difference between the scores sensitivity and specificity (i.e. recall). This technique is quite common in the medical speech processing area (see e.g. Satt et al., 2013; König et al., 2015; Vaiciukynas et al., 2017; Moro-Velázquez et al., 2018; Fritsch et al., 2019).

Furthermore, we also give the area under the ROC curve (AUC) as a metric frequently applied in medical speech processing studies; note that, since we have only two subject categories (i.e. classes), their AUC scores happen to be identical. Lastly, we also employ the evaluation metrics of log-likelihood ratio cost (C_{llr}) and minimum log-likelihood ratio cost ($\min C_{llr}$), frequently used in forensic studies and in the area of speaker recognition (see e.g. Morrison et al., 2010; Frost and Ishihara, 2015; Nautsch et al., 2019). These values were calculated using the Bosaris toolkit (Brümmer and de Villiers, 0000). Note that, as C_{llr} and $\min C_{llr}$ are cost values, a lower value indicates a more accurate system.

From the list of our applied evaluation metrics, we find AUC and $\min C_{llr}$ to be the most important one for two reasons. Firstly, since we have a limited number of subjects, the other metrics can take their values only from a quite limited domain. For example, recall can take only one of 15 and 14 values (since we had 14 and 13 MCI subjects for English and Hungarian, respectively), therefore classifying one further MCI subject correctly leads to an increase of about 7%. In contrast, Area-Under-Curve and min C_{llr} allow us to express classification performance in much finer detail. The second reason is that, during our meta-parameter selection for the Support Vector Machine classifier (i.e. at the inner loop of the nested cross-validation procedure), we maximized the AUC metric, selecting the SVM meta-parameters C and γ accordingly.

5. Monolingual results

First, we present our monolingual results: that is, the classification scores we obtained by extracting the temporal parameters by using English-language ASR for our English-speaking subjects, and by using Hungarian-language ASR for our Hungarian-speaking speakers. The measured evaluation metric values are listed in Table 3.

5.1. English-speaking subjects

Regarding the English-speaking MCI and HC subjects (see the upper half of Table 3), when we calculated our temporal parameters by relying on the English TEDLium corpus, we noticed that the achieved scores were quite high. Besides an accuracy score of 84.8%, the recall/sensitivity and specificity values were also around 85%. With a precision score of 80%, we achieved an

Table 3
Accuracy scores obtained in the monolingual cases.

	Feature Categories	Acc.	Prec.	Recall	Spec.	F_1	AUC	C_{llr}	**min**C_{llr}
English-speaking *(English ASR)	All features	84.8%	80.0%	85.7%	84.2%	82.8	0.932	0.685	0.305
	Silence-related	63.6%	56.3%	64.3%	63.2%	60.0	0.756	0.847	0.570
	Filler-related	78.8%	73.3%	78.6%	78.9%	75.9	0.816	1.037	0.597
	All pause-related	63.6%	56.3%	64.3%	63.2%	60.0	0.748	0.889	0.610
	Tempo + silence-related	78.8%	73.3%	78.6%	78.9%	75.9	0.880	0.852	0.458
	Tempo + filler-related	84.8%	80.0%	85.7%	84.2%	82.8	0.883	0.756	0.481
	Tempo + all pause-related	69.7%	62.5%	71.4%	68.4%	66.7	0.797	0.851	0.583
Hungarian-speaking (Hungarian ASR)	All features	75.8%	66.7%	76.9%	75.0%	71.4	0.727	0.957	0.690
	Silence-related	39.4%	29.4%	38.5%	40.0%	33.3	0.519	0.990	0.887
	Filler-related	90.9%	85.7%	92.3%	90.0%	88.9	0.912	0.947	0.306
	All pause-related	75.8%	66.7%	76.9%	75.0%	71.4	0.642	1.065	0.793
	Tempo + silence-related	51.5%	41.2%	53.9%	50.0%	46.7	0.531	1.018	0.944
	Tempo + filler-related	84.8%	78.6%	84.6%	85.0%	81.5	0.812	0.938	0.534
	Tempo + all pause-related	78.8%	71.4%	76.9%	80.0%	74.1	0.792	1.015	0.600

F-measure value of 82.8. Furthermore, the AUC and minC_{llr} values are also quite good (0.932 and 0.305, respectively), indicating that the temporal parameters described in Table 2 allowed accurate MCI detection in this case.

Examining the scores achieved with only a subset of these parameters (i.e. calculating the indicators (4)–(7) for the silent pauses only, for the filler events only, for both hesitation types, and extending these sets with parameters (1)–(3)), we see that there are several identical values. We think that this is probably due to the fact that we calculated these metric scores at the Equal Error Rate, and that we had a relatively low number of subjects (i.e. $n = 33$). Despite this, we observe significant differences among the different feature subsets in terms of AUC; this is expected, though, because we optimized for the AUC score in the nested cross-validation steps.

Based on the AUC values, using just the filled pause-related attributes led to a slightly better performance than just using the silent pause-related ones or when we did not distinguish between the two hesitation types; on the other hand, the minC_{llr} values were quite similar in the three cases ($0.570 - 0.610$). Notice, however, that we got the worst values with the last case, which is quite reasonable, considering that the two hesitation types might have different temporal characteristics (such as average duration): treating them as the same phenomenon might lead to less descriptive temporal parameters, and lead to a lower classification performance.

Interestingly, when we combined these attribute subsets with articulation tempo, speech tempo and utterance duration, silence-related and filler-related attributes produced similar (and similarly high) AUC scores (values 0.880 and 0.883, silent and filled pauses, respectively), and the minC_{llr} values were also quite close (0.458 and 0.481). However, the other evaluation metrics differed significantly, fillers producing better scores; while not distinguishing the two hesitation types led to the lowest (although still relatively high) metric values. As expected, we achieved the highest scores when we used all 15 temporal parameters.

5.2. *Hungarian-speaking subjects*

As for the classification scores of the Hungarian-speaking subjects based on attributes calculated by the Hungarian ASR model (see the lower half of Table 3), we got slightly different results (although, of course, the two parts of Table 3 cannot be compared, as the values were measured on different subjects). Using all 15 temporal parameters led to a fine performance, but exploiting the silent pause-related attributes was only scarcely better than simple guessing (AUC scores of 0.519 and 0.531, without and with the speech tempo-derived attributes, respectively). However, filler-related temporal parameters turned out to be surprisingly useful: besides an AUC value of 0.912 and a minC_{llr} value of 0.200, we achieved classification accuracy, recall (sensitivity) and specificity scores above 90%; with a precision score of 85.7%, it led to an F-measure value of 88.9. Not distinguishing the two hesitation types led to an average performance when judging by AUC (0.642) or minC_{llr} (0.793), but the other evaluation metric values were the same as those obtained by using all the attributes.

We can also notice the difference between the corresponding C_{llr} and minC_{llr} values. Indeed, the actual C_{llr} scores are usually significantly higher (i.e. worse), sometimes even going above 1.0. In our opinion this indicates that the decision threshold between the two speaker categories was usually not set reliably. The other metric values did not reflect this, since they are insensitive to the actual threshold value: accuracy, precision, recall, specificity and F_1 were calculated at Equal Error Rate (as common in the literature), and the AUC metric takes all possible decision levels into consideration. On the other hand, the AUC and minC_{llr} values behaved quite similarly; the 14 value pairs presented in Table 3 have a correlation coefficient of 0.969.

6. Cross-lingual results

Next, we turn to the results of our cross-lingual experiments. That is, we will now calculate our set of temporal parameters (the features used for classification) with a Hungarian ASR system (trained on the BEA corpus) for the English-speaking MCI and

Table 4
Accuracy scores obtained in the cross-lingual cases.

	Feature Categories	Acc.	Prec.	Recall	Spec.	F_1	AUC	C_{llr}	**min**C_{llr}
English-speaking (Hungarian ASR)	All features	78.8%	73.3%	78.6%	78.9%	75.9	0.835	0.804	0.469
	Silence-related	72.7%	66.7%	71.4%	73.7%	69.0	0.808	0.826	0.590
	Filler-related	72.7%	66.7%	71.4%	73.7%	69.0	0.744	0.882	0.656
	All pause-related	78.8%	73.3%	78.6%	78.9%	75.9	0.838	0.810	0.466
	Tempo + silence-related	78.8%	73.3%	78.6%	78.9%	75.9	0.850	0.845	0.571
	Tempo + filler-related	78.8%	73.3%	78.6%	78.9%	75.9	0.820	0.852	0.526
	Tempo + all pause-related	78.8%	73.3%	78.6%	78.9%	75.9	0.820	0.839	0.454
Hungarian-speaking (English ASR)	All features	75.8%	66.7%	76.9%	75.0%	71.4	0.819	0.895	0.538
	Silence-related	75.8%	66.7%	76.9%	75.0%	71.4	0.665	0.999	0.769
	Filler-related	90.9%	85.7%	92.3%	90.0%	88.9	0.923	0.941	0.200
	All pause-related	69.7%	60.0%	69.2%	70.0%	64.3	0.685	0.981	0.815
	Tempo + silence-related	75.8%	66.7%	76.9%	75.0%	71.4	0.788	0.976	0.663
	Tempo + filler-related	84.8%	78.6%	84.6%	85.0%	81.5	0.785	0.971	0.595
	Tempo + all pause-related	69.7%	60.0%	69.2%	70.0%	64.3	0.677	0.963	0.763

HC subjects, while for the Hungarian-speaking subjects these were calculated by an English-language speech recognizer (trained on the TEDLium corpus). Note that, after this step, we still train and evaluate our SVM classifiers on subjects belonging only to one language (with the use of the nested cross-validation technique). Our results can be seen in Table 4.

6.1. English-speaking subjects

The accuracy scores achieved when classifying the English-speaking MCI and HC subjects are listed in the upper half of Table 4. When using all 15 attributes, we obtained slightly lower scores than in the corresponding monolingual case (shown in the upper half of Table 3), but these values are still quite high: accuracy, recall/sensitivity, specificity and UAR all appeared to be around 79%, while precision and the F_1-score were 73.3% and 75.9, respectively. This, in our opinion, indicates that the attribute set proposed in our previous studies, and outlined in Section 3, is quite robust, as it could be reliably extracted by an ASR system trained on a different language.

Examining the classification performances corresponding to the various attribute subsets, we see that when we relied only on the silent pause-related attributes, we measured slightly higher scores than in the mono-lingual case; most importantly, the AUC score rose from 0.756 to 0.808 (but the other metric values improved as well with the exception of minC_{llr}, which appeared to be slightly worse). This supports our expectation that silent pauses can be located in a language-independent way. Still, in the case of using only the filled pause-related temporal parameters we observe lower performance scores, as the AUC value fell from 0.816 to 0.744 and the minC_{llr} score rose from 0.579 to 0.656. Calculating the pause-related attributes for all pause types, however, again led to an increase in the classification performance: we measured practically identical scores as we did for the full feature set (in the cross-lingual case), while for the monolingual case, AUC was 0.797 and the other metric values were around 60%.

6.2. Hungarian-speaking subjects

The classification scores for the Hungarian-speaking subjects, when relying on the English ASR (see the lower part of Table 4) are usually slightly better than those obtained with the corresponding (i.e. Hungarian) ASR system (presented in the lower part of Table 3). More importantly, though, the tendencies of the evaluation metric values measured mirror those of the monolingual case: the silence-related attributes were not remarkably useful (AUC value of 0.665 and minC_{llr} score of 0.769), while filled pauses led to a very high (0.923) AUC and a very low (0.200) minC_{llr} score, while the other metrics also lay between 85% and 92%. When we included the articulation rate, speech tempo and utterance duration in our feature set, the AUC and minC_{llr} scores improved slightly for the silent pause case, but for the filled pauses we found a slight drop in the scores (i.e. from 85.7−92.3% to 78.6−85.0%).

6.3. Summary

The classification results seem to support our hypothesis that the temporal speech parameters developed by our team can be calculated language-independently. For the English-speaking subjects, using all the attributes was, without a doubt, the most effective approach: although we obtained the same classification scores (not taking AUC, C_{llr} and minC_{llr} into account) once for the English ASR case, and for most of the Hungarian ASR case, we never managed to outperform it significantly. Looking at the feature subsets, attributes related to either hesitation types (i.e. silent or filler pauses) were of similar importance, although the filled pauses happened to be slightly more important in the monolingual case. Furthermore, the articulation tempo-related attributes proved to be beneficial in 5 out of the total of 6 cases.

From the classification results involving the Hungarian-speaking MCI and HC subjects, we also observe similarities among the two ASR systems. First of all, we obtained the same metric values (again, with the exception of AUC) when we used all the temporal parameters, regardless of the language of the ASR model employed. For these speakers, filled pauses turned out to be significantly more useful than silent ones, while treating both pause types indifferently led to mediocre classification results regardless of the ASR language used. Lastly, the speech tempo-derived features were of little or no use (silent and filled pause-related attributes, respectively); and they only slightly improved the classification performance for the monolingual case when we treated both hesitation types as identical in the monolingual case, while in the cross-lingual case there was no improvement at all.

Based on these results we can say that the language, for which the automatic speech recognition system was trained on, does not really affect the performance of the subsequent MCI-HC classification step. Therefore it seems that we can actually calculate the speech temporal parameters (described in Section 3) by using an ASR system of a language different from that used by the MCI and HC subjects.

7. Comparing the temporal parameter values

Up until now we have examined the temporal parameters extracted by the English and the Hungarian ASR systems by performing classification; however, the attributes themselves are interpretable, and (at least, by our hypothesis) for the two ASR systems they are expected to be quite similar for the same input (i.e. utterance). This allows us to make a more direct comparison of the calculated values. Therefore, next we took the temporal parameter values calculated for the English speech recordings got via both ASR systems, and calculated the (Pearson's) correlation scores. The resulting values can be seen in Fig. 2.

While it was not surprising that the duration of the utterances matched perfectly, we also measured high correlation scores for the articulation rate and for the speech tempo (0.959 and 0.962, respectively). These also support our hypothesis that these attributes can be reliably calculated with the phone set of a different language.

For the silent pause-related attributes, we also notice quite high correlation values (i.e. in the range 0.801 ... 0.945); this supports our initial assumption that the silent pauses (and, of course, their starting and ending positions) can be located in a language-independent manner; in this case, using an English and a Hungarian ASR system for the English utterances. From the four values, the average duration was the most correlated attribute: it had a correlation coefficient of 0.945, while we got the lowest correlation coefficient for pause frequency (0.801). This probably means that not all (but still, most) pause occurrences were found by the two ASR models, but they (roughly) agreed on the starting and ending points of these occurrences.

Examining the correlation coefficients measured for the filled pause-related temporal parameters, we see much lower values. As in the silent pause case, the average duration was the most correlated attribute (Pearson's correlation coefficient value of 0.731), while the pause frequency again had the lowest value. This suggests that one of the ASR systems (actually, the English one) found fewer filled pause occurrences than the other (i.e. the Hungarian); however, we found no huge difference in the duration of the filled pause events found by both ASR models. (Note, however, that the temporal speech parameters calculated from these pause occurrences remained similarly indicative for the subsequent classification step.)

To investigate whether this difference came from the languages used by the two ASR models, we conducted a final experiment. For this, we calculated our temporal speech parameters for the *training sets* of the two ASR systems (following the annotation we used for ASR training), i.e. the 15 h subset of the TEDLium corpus and the 44 h subset of the BEA database. Next, we calculated the ratio of these values; i.e. we took each attribute for the English training utterances, and divided it by the corresponding attribute value of the Hungarian training set (case "annotated"). For example, a ratio value of 0.8 for the "average duration of filled pauses" attribute would mean that in the TEDLium corpus the annotated occurrences of filled pauses are, on average, 20% shorter than they are in the Hungarian BEA database. We repeated this process for the utterances of the English

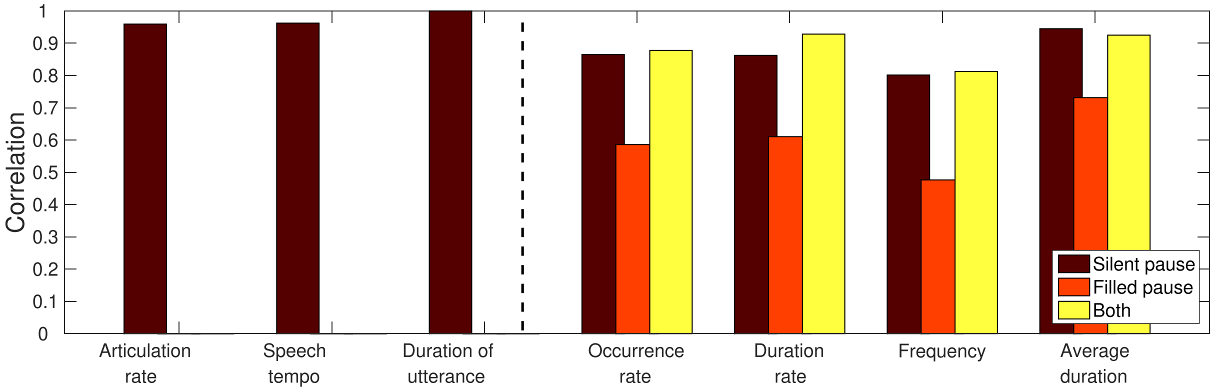

Fig. 2. Correlation coefficients of the temporal speech parameters based on the output of the English and the Hungarian ASR models, measured on the utterances of the English MCI and HC subjects.

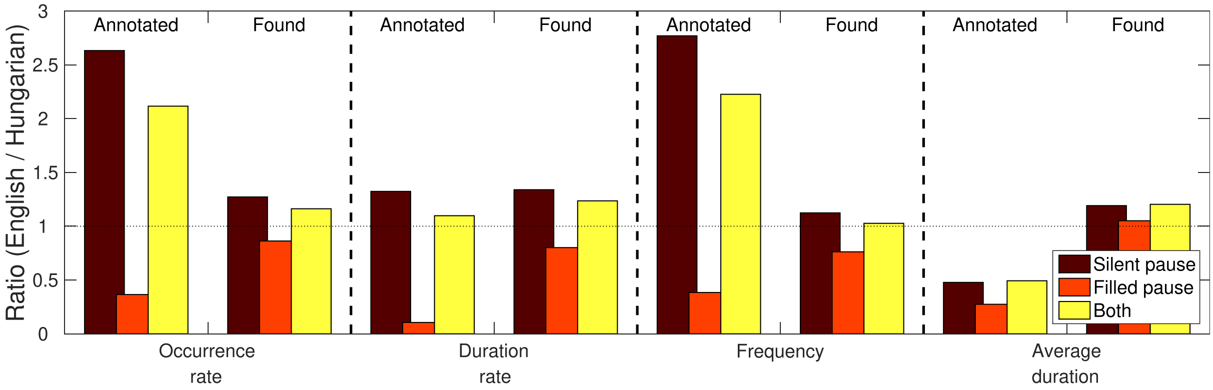

Fig. 3. Ratio of the temporal speech parameters calculated for the training set of the English and the Hungarian ASR systems (case "Annotated"), and for the output of the English and the Hungarian ASR models on the utterances of the English MCI and HC subjects (case "Found").

MCI and HC subjects; here we used the phonetic segmentation produced by the English and the Hungarian ASR models, and then we took the ratio of these two scores (case "found"). The resulting ratio values of our attributes are shown in Fig. 3.

Looking at this graph, we see that the trends of the calculated temporal speech parameter values using the two ASR systems, generally follow the distribution of the samples in the two training sets. Regarding silent pauses, the TEDLium corpus contains significantly more (i.e. 2.5 times) silent pause occurrences than BEA does; and although their average duration is much smaller, their combined duration is still higher by 32%. (Surprisingly, 29% of the used subset of the TEDLium dataset corresponds to silence, while it is only 22% for the BEA corpus.) Examining the detected silent pause occurrences in the speech of the (English) MCI and HC subjects, we observe similar trends: the English one found more pause occurrences than the Hungarian one, expressed by any of our temporal parameters (i.e. occurrence rate, duration rate or frequency).

Regarding filled pauses, we see just the opposite. This subset of the TEDLium corpus contained fewer and shorter filled pauses than the 44 h long BEA subset did: there, only 0.6% of the phones corresponded to filled pauses (BEA: 1.7%), and only 0.5% of the total duration of the utterances were filled pauses (BEA: 5.0%). This was also reflected by the recognition results: in the very same utterances of the 14 MCI and 19 HC subjects, fewer filled pauses were hypothesized by the English ASR system than by the Hungarian one. Although the difference in this case is not as large as it was for the training corpora (for example, using the English ASR system 5.5% of the duration was marked as filled pause, while for the Hungarian ASR model it was 6.8%), the temporal speech parameters calculated from the spontaneous speech utterances show the same tendencies as we found for the training sets. The only parameters which seem to be insensitive to this training data difference are the average durations: they appeared to be quite similar for all three cases (i.e. for silent pauses only, for filled pauses only, and for treating the two as the same event).

These values, in our opinion, suggest that the actual language of the ASR model is of secondary importance when calculating the proposed set of temporal speech parameters. However, the *type of utterances* and the *annotation* of the corpus used to train the speech recognition system might lead to noticeable differences in these attributes. In our case, the TEDLium corpus consists of planned speech; and although such talks do contain filled pauses, they occur less frequently than in narrated spontaneous conversations, which comprise the BEA dataset. In contrast, lots of short pauses were marked in the annotation of TEDLium, which (at least, according to Fig. 3) were probably unannotated in the BEA Hungarian database. The ASR models trained on these two corpora led to phonetic transcripts which followed this tendency: the English system, evaluated on the same utterances, found more silent and fewer filled pause occurrences than its Hungarian counterpart.

8. Conclusions

In this study, we investigated the language-dependence of our speech processing workflow, developed for distinguishing between patients suffering from Mild Cognitive Impairment (MCI) and healthy controls (HCs) based on the analysis of their spontaneous speech. We decided to focus on a set of temporal speech parameters, consisting of the articulation rate, speech tempo, utterance duration, and attributes describing various characteristics of hesitation present in the speech of the subject. We distinguished two different types of hesitations: silent pauses, being the absence of any sound uttered by the speaker, and filled pauses, referring to vocalizations like 'uhm', 'er' and 'ah'. We used a phone-level ASR system to obtain a phonetic-level, time-aligned transcription of each utterance, serving as the basis of temporal speech parameter calculation.

To test the language-independence of this attribute set, we performed experiments using two different languages. That is, we collected speech samples both from English and from Hungarian MCI and HC subjects, and analyzed their speech using English-language and Hungarian-language ASR models. The calculated attribute vectors were then used as features for subject classification by Support Vector Machines in a nested cross-validation process into two categories. Our hypothesis was that in the cross-lingual setup we could extract similarly indicative attributes for articulation rate, speech tempo, for the silent pause-related attributes and even for the filled pause-related ones.

The monolingual cases, i.e. when we applied the English ASR system to analyze the speech of the English-speaking subjects, and when we used the Hungarian ASR model for the Hungarian speakers, served as our baseline. For the English subjects, we achieved high classification scores (in the range 80.0−85.7%, an Area-Under-Curve score of 0.932 and $\min C_{llr} = 0.305$), while for Hungarian, the classification performance was acceptable (with classification metrics falling into the range 66.7−76.9% and with an AUC value of 0.727). By only using specific subsets of our temporal parameters, we noticed that filled pauses were more useful for both speaker groups (i.e. English and Hungarian) than silent pauses; surprisingly, silent pauses were not useful at all for distinguishing the Hungarian subjects. Calculating the temporal attributes and treating the two pause types as the same phenomenon proved to be less indicative than focusing on filled pauses only, for both languages.

In the next part of our study, we turned to the cross-lingual experiments: we used our Hungarian ASR model to analyze the English-speaking subjects, while the English ASR system was employed for the Hungarian subjects. The classification results confirmed our hypothesis: although we got slightly lower scores for the English-speaking subjects than we did in the monolingual case, for the Hungarian-speaking sample our scores remained the same or even improved. Furthermore, silent and filled pauses led to a similar classification performance for the English-speaking subjects. For Hungarian, filled pauses remained the most indicative despite using an ASR model trained for a different language (i.e. English), while the metric values corresponding to the silent pause-related attributes improved relative to the monolingual case. These results, in our opinion, support our hypothesis that both silent and filled pauses can be detected robustly, even when the ASR system used is trained on a different language.

In the last part of our study, we compared the calculated attributes more directly. To do this, we calculated Pearson's correlation coefficient for all 15 temporal parameters obtained for the English subjects by both (i.e. English and Hungarian) ASR models. We found that the articulation rate and speech tempo were well correlated (coefficients around 0.96); we also got high correlation coefficients for the silent pause-related attributes (0.801 to 0.945), indicating that silence was detected by the two ASR systems to a similar extent. For the filled pauses, however, the attributes were significantly less correlated ($0.476 \ldots 0.731$), reflecting a difference in the outputs of the two ASR models in this regard. By examining the training sets of the two ASR systems, however, we found that this is most likely not due to the language difference, but it can be accounted for the distribution of silent and filled pauses in the training utterances. Indeed, the TEDLium corpus, used to train the English ASR system, contained significantly more silent pause occurrences than its Hungarian counterpart (the BEA Hungarian Database), while it had fewer filled pauses present (or at least annotated). The outputs of the two ASR systems mirrored these trends: there were more silent and fewer filled pauses in the phonetic-level output of the English ASR model than in the output of the Hungarian one.

Overall, in our opinion, the differences we found in the temporal parameters appear to be the effect of a difference in the *training databases*, and they have little to do with the difference in the languages. However, speech recognition datasets which do contain filled pause occurrences, and, more importantly, where these occurrences are properly annotated, are not so common. Finding datasets for different languages with the same (or at least similar) recording and annotation protocols is quite difficult. Fortunately, according to our experimental results, it is not even necessary: our temporal speech parameters proved to be indicative both for the English-speaking and for the Hungarian-speaking subjects, regardless of the speech recognition dataset used for acoustic DNN model training.

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.

Acknowledgments

This study was supported by the National Research, Development and Innovation Office of Hungary via contract NKFIH FK-124413, by the grant NKFIH-1279-2/2020 of the Hungarian Ministry of Innovation and Technology, and by the Hungarian Artificial Intelligence National Laboratory. Gábor Gosztolya was also funded by the János Bolyai Scholarship of the Hungarian Academy of Sciences and by the Hungarian Ministry of Innovation and Technology New National Excellence Programme ÚNKP-20-5.

References

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

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

Asgari, M., Kaye, J., Dodge, H., 2017. Predicting mild cognitive impairment from spontaneous spoken utterances. Alzheimer's Dement. Transl. Res. Clin. Interv. 3 (2), 219–228.

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.

Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot Int. 5 (9/10), 341–345.

Brueckner, R., Schmitt, M., Pantic, M., Schuller, B., 2017. Spotting social signals in conversational speech over IP: a deep learning perspective. In: Proceedings of the Interspeech, pp. 2371–2375.

Brümmer, N., de Villiers, E.,. The BOSARIS toolkit: theory, algorithms and code for surviving the new DCF. arXiv: 1304.2865.

Bruscoli, M., Lovestone, S., 2004. Is MCI really just early dementia? A systematic review of conversion studies. Int. Psychogeriatr. 16 (2), 129–140.

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. 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.

Foster, N.L., Bondi, M.W., Das, R., Foss, M., Hershey, L.A., Koh, S., Logan, R., Poole, C., Shega, J.W., Sood, A., Thothala, N., Wicklund, M., Yu, M., Bennett, A., Wang, D., 2019. Quality improvement in neurology. Neurology 93 (16), 705–719.

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

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.

Fritsch, J., Wankerl, S., Nöth, E., 2019. Automatic diagnosis of Alzheimer's disease using neural network language models. In: Proceedings of the ICASSP, pp. 5841–5845.

Frost, D., Ishihara, S., 2015. Likelihood ratio-based forensic voice comparison on L2 speakers: a case of Hong Kong native male production of English vowels. In: Proceedings of the ALTA. Parramatta, Australia, pp. 39–47.

García, N., Vásquez-Correa, J.C., Orozco-Arroyave, J.R., Nöth, E., 2018. Multimodal i-vectors to detect and evaluate Parkinson's disease. In: Proceedings of the Interspeech. Hyderabad, India, pp. 2349–2353.

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 the 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.

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.

Hoffmann, I., Tóth, L., Gosztolya, G., Szatlóczki, G., Vincze, V., Kárpáti, E., Pákáski, M., Kálmán, J., 2017. Beszédfelismerés alapú eljárás az enyhe kognitív zavar automatikus felismerésére spontán beszéd alapján. Általános nyelvészeti tanulmányok 29 (1), 385–405.

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.

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. Diagn. Assess. Dis. Monit. 1 (1), 112–124.

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.

de Leeuw, E., 2007. Hesitation markers in English, German, and Dutch. J. Ger. Linguist. 19 (2), 85–114.

Manos, P.J., Wu, R., 1994. The ten-point clock test: a quick screen and grading method for cognitive impairment in medical and surgical patients. Int. J. Psychiatry Med. 24 (3), 229–244.

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.

Moro-Velázquez, L., Gómez-García, J.A., Godino-Llorente, J.I., Villalba, J., Orozco-Arroyave, J.R., Dehak, N., 2018. Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's disease. Appl. Soft Comput. 62 (10), 649–666.

Morrison, G.S., Thiruvaran, T., Epps, J., 2010. Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system. In: Proceedings of the Odyssey. Brno, Czech Republic, pp. 63–70.

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.

Nautsch, A., Jiménez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., Delgado, H., Todisco, M., Hmani, M.A., Mtibaa, A., Abdelraheem, M.A., Abad, A., Teixeira, F., Matrouf, D., Gomez-Barrero, M., Petrovska-Delacrétaz, D., Cholleth, G., Evans, N., Schneider, T., Bonastre, J.-F., Raj, B., Trancoso, I., Busch, C., 2019. Preserving privacy in speaker and speech characterisation. Comput. Speech Lang. 58 (Nov), 441–480.

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 the TSD. Brno, Czech Republic, pp. 424–431.

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.

Prince, M., Wimo, A., Guerchet, M., Ali, G.-C., Wu, Y.-T., Prina, M., 2015. World Alzheimer Report 2015. The Global Impact of Dementia. Alzheimer's Disease International, London, UK.

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

Rousseau, A., Deléglise, P., Estève, Y., 2012. TED-LIUM: an automatic speech recognition dedicated corpus. In: Proceedings of the LREC, pp. 125–129.

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 the Interspeech Singapore, pp. 2538–2542.

Satt, A., Sorin, A., Toledo-Ronen, O., Barkan, O., Kompatsiaris, I., Kokonozi, A., Tsolaki, M., 2013. Evaluation of speech-based protocol for detection of early-stage dementia. In: Proceedings of the Interspeech, pp. 1692–1696.

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.

Sheikh, J.I., Yesavage, J.A., 1986. Geriatric depression scale (GDS) – recent evidence and development of a shorter version. Clin. Gerontol. 5 (1–2), 165–173.

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. Am. 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 the 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.

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

Themistocleous, C., Eckerström, M., Kokkinakis, D., 2020. Voice quality and speech fluency distinguish individuals with mild cognitive impairment from healthy controls. PLoS ONE 15 (7).

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 the 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.

Vaiciukynas, E., Verikas, A., Gelzinis, A., Bacauskiene, M., 2017. Detecting Parkinson’s disease from sustained phonation and speech signals. PLoS ONE 12 (10).

Vetter, M., Sakti, S., Nakamura, S., 2019. Cross-lingual speech-based Tobi label generation using bidirectional LSTM. In: Proceedings of the ICASSP. IEEE, pp. 6620–6624.

Yesavage, J.A., Brink, T.L., Rose, T.L., Lum, O., Huang, V., Adey, M., Leirer, V.O., 1983. Development and validation of a geriatric depression screening scale: a preliminary report. J. Psychiatr. Res. 17 (1), 37–49.

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.