Assessing Alzheimer's Disease from Speech Using the i-vector Approach

José Vicente Egas López[1](✉), László Tóth[1], Ildikó Hoffmann[2,3], János Kálmán[4], Magdolna Pákáski[4], and Gábor Gosztolya[1,5]

[1] Institute of Informatics, University of Szeged, Szeged, Hungary
egasj@inf.u-szeged.hu
[2] Department of Linguistics, University of Szeged, Szeged, Hungary
[3] Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
[4] Department of Psychiatry, University of Szeged, Szeged, Hungary
[5] MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary

Abstract. One of the world's chronic neuro-degenerative diseases, Alzheimer's Disease (AD), leads its sufferers, among other symptoms, to suffer from speech difficulties. In particular, the inability to recall vocabulary which makes patients' speech different. Furthermore, Mild Cognitive Impairment (MCI) is usually considered as a prodromal neuro-degenerative state of AD. The key to abate the progress of both disorders is their early diagnosis. However, actual ways of diagnosis are costly and quite time-consuming. In this study, we propose the extraction of features from speech through the use of the i-vector approach, by which we seek to model the speech pattern of the three mental conditions from the subjects. To the best of our knowledge, no previous studies have utilized i-vector features to assess Alzheimer's before. These i-vectors are extracted from Mel-Frequency Cepstral Coefficients (MFCCs), then they are given to a SVM classifier in order to identify the speech in one of the following manners: AD - Alzheimer Disease, MCI - Mild Cognitive Impairment, HC - Healthy Control. We tested these i-vector features by performing a 5-fold cross-validation and we achieved an F1-score of 79.2%.

Keywords: i-vectors · Alzheimer's · SVM · Speech recognition

1 Introduction

Speech difficulties among patients suffering from Alzheimer's Disease (AD) become palpable from the moderate stage of the disease and such adversities are often characterized by the incapacity to recall vocabulary, leading to constant incorrect word substitutions, also known as paraphasias [8]. The language of the AD patient is diminished to simple phrases or single words; progressively, the patient may entirely lose their speech, resulting in a substantial decrease in

A. A. Salah et al. (Eds.): SPECOM 2019, LNAI 11658, pp. 289–298, 2019.
https://doi.org/10.1007/978-3-030-26061-3_30

the quality of life [8,9]. In most cases, these factors create the structure of speech of a patient suffering from Alzheimer's, which is generally formed by syntactic complexity, insufficient speech fluency, and vocabulary limitation.

Insufficient screening techniques have made Alzheimer's too complex to diagnose. The early diagnosis of the disease could lead to a more effective confrontation of the AD in order to slow down its development; this stage of diagnosis is difficult to achieve [14,24]. Generally speaking, patients arrive at the clinic when Alzheimer's is already in an advanced state, which lowers the ratio of early AD detection cases. MCI (Mild Cognitive Impairment), as part of the process of dementia, is prone to start around the age of 40. Screening tests to detect MCI take a long time, they shortage of pre-clinical state diagnosis and require a high budget to fund them [18].

Speech recognition tools are widely used for similar tasks within this branch of medicine. Fraser et al. [10–12] utilized speech recognition to detect aphasia. Lehr et al. [22] applied speech recognition in order to diagnose MCI. Other groups [1,26] diagnosed Alzheimer's through the use of speech recognition tools. To detect and assess other neuro-degenerative diseases such as Parkinson's (PD), the i-vector approach has been successfully applied to model the speech of PD patients by extracting i-vectors from it and performing classification through the comparison with those of the test speakers by means of cosine distance scoring [16]; likewise, classifying them using of Support Vector Machines (SVM) [17]. Also, i-vectors have been used to perform classification and regression of the speaker's age. To be precise, Grzybowska et al. [19] carry out an examination of the use of i-vectors both for age regression and for age classification based on the speech of the subjects.

To the best of our knowledge, no previous studies exist that classify Alzheimer's Disease based on utterances by applying the i-vector approach. Here, we fit a (linear) Support Vector Machines (SVM) classifier which is given i-vectors features extracted from the Mel-Frequency Cepstral Coefficients (MFCCs) of the utterances. The diagnosis is predicted as one of the following three states: HC (Healthy Control), MCI (Mild Cognitive Impairment), and AD (Alzheimer's Disease).

2 Data

The data for the experiments in this study is defined as follows: 225 speech signals recorded from 75 subjects (*dementia dataset*), and 44 recordings taken from generic speakers (*BEA dataset*). The speech utterances used are the same as those employed in [18], which were recorded at the Memory Clinic, University of Szeged, Hungary. Three categories of utterances were recorded, namely, subjects suffering from MCI, subjects affected by the early-stage of AD, and subjects having no cognitive impairment at the time of recording. Such categories were matched for age, gender and education. We worked with the utterances of 25 speakers for each speaker group, resulting in a total of 75 speakers and 225 recordings.

Table 1. The characteristics of the three groups of the study participants. Groups: MCI = mild cognitive impairment; mAD = mild Alzheimer's Disease. Tests: MMSE = Mini-Mental State Examination; CDT = Clock Drawing Test; ADAS-Cog = Alzheimer's Disease Assessment Scale. Values are given as mean ± standard deviation.

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

Mini-Mental State Examination (MMSE, [7]), Clock Drawing Test (CDT; [13]) and the Alzheimer's Disease Assessment Scale (ADAS-Cog, [25]) were the clinical tests employed in order to assess the cognitive states of the subjects. From the MMSE test, one can get a maximum of 30 points in the following manner: 29–30 points for healthy elderly, 27–28 points for mild neurocognitive impairment, 20–26 points for mild dementia, 10–19 points for moderate dementia, and 0–9 points for severe dementia [7]. The CDT test is up to a total of 10 points, where a score below 7 corresponds to a cognitive decline [13]. The ADAS-Cog test, which employs an inverse scoring (i.e. errors are counted rather than right answers), has the following scoring system: 0–8 points for normal cognitive abilities, 9–15 points for mild neurocognitive impairment, and 16–70 points for severe neurocognitive impairment [25].

The Geriatric Depression Scale (GDS) was used to assess the state of depression. The three groups (F(2;74) = 2.202; p = 0.118) were aligned with regard to gender (X2(2) = 1.389; p = 0.499), age (F(2;74) = 2.321; p = 0.105) and years of education (F(2;74) = 2.202; p = 0.118). Table 1 lists the clinical characteristics of the control, the MCI and the mAD group. The recordings reflect a spontaneous speech of the subjects and the experimental setup for them was as follows: (1) *Immediate recall*, after the presentation of a specially designed one-minute-long film, the subjects were asked to talk about details seen on the film. (2) *Previous day*, the subjects were asked to talk in detail about their previous day. (3) *Delayed recall*, in the end, a second film was played, and after having one minute pause, the subjects were asked to speak about what they saw. The structure of the data became a set of 3 spontaneous-speech recordings per speaker, where each was edited in such a manner that we cropped parts before the subject starts to speak and after the subject's last phoneme.

3 Methods

The study was achieved by performing the extraction of the i-vectors in the following manner: (1) MFCCs features were extracted separately from 225 (i.e.

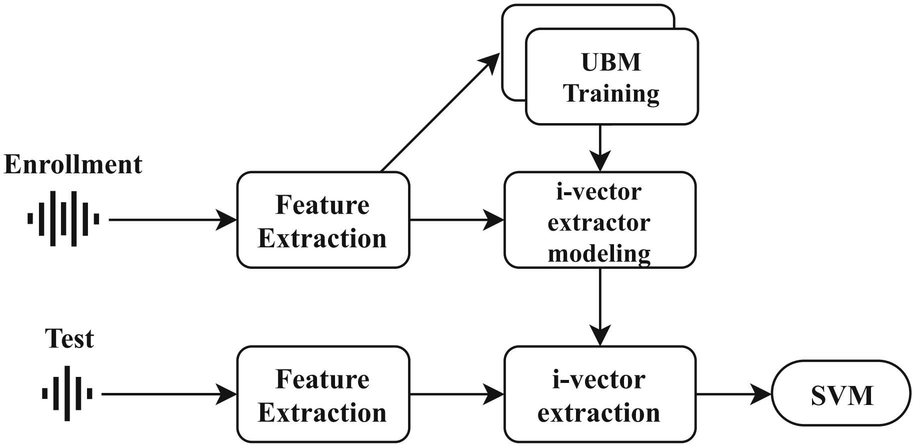

Fig. 1. The generic methodology applied in our work.

dementia dataset) and 44 speech recordings (i.e. BEA dataset) (2) the UBM was trained using the MFCCs obtained from the BEA dataset (3) the i-vector extractor model was trained using the UBM of the previous step, and MFCCs from the dementia dataset (4) MFCCs from the dementia dataset were processed to extract a set of 225 i-vectors, and lastly, (5) a Support Vector Machines (SVM) performed the classification process. These stages are outlined in Fig. 1.

3.1 Feature Extraction

Among the most popular short-term acoustic features are the MFCCs (Mel-Frequency Cepstral Coefficients), which are obtained by implementing the following operations on the utterances: power spectrum, logarithm, and Discrete Cosine Transform (DCT), these deliver the first coefficients plus one more coefficient associated with the energy of the frame. Velocity and acceleration (first and second derivatives) are affixed to the MFCCs together with their energy's coefficients. In this study, we will use MFCCs because this technique has proved to be one of the most effective when it comes to creating a speaker model [15,20].

3.2 The i-vector Approach

GMM (Gaussian Mixture Model) supervectors [2] and JFA (Joint Factor Analysis) [21] are successful approaches that were once the state-of-the-art systems for robust speaker recognition. In an attempt to combine of both techniques, JFA speaker factors were used as features for SVM classifiers [5]. It found that the channel factors estimated with JFA not only contain channel effects but speaker-dependent information as well; hence, speaker and channel factors were combined into a single space. Factor Analysis (FA), which is used as a feature extractor, defines a new low-dimensional *total variability space* in which a speech utterance is defined by a new vector called *i-vector* [6] that contains the estimates of the *total factors*:

$$M = m + Tw, \tag{1}$$

where M is the Gaussian Mixture Model (GMM) speaker supervector for a given signal; m is the speaker/channel-independent component, namely, the UBM supervector; T is the Total Variability matrix (TV); and w is a standard normal distributed hidden variable, i.e. the i-vector. This vector can be thought of as a representation of a given recording in a lower-dimension space.

In contrast to JFA, i-vectors do not make any distinction between speaker and channel; here, each utterance is assumed to be acquired from a different speaker. The i-vector approach is, in plain words, a dimensionality reduction technique of the GMM supervector.

To the best of our knowledge, no previous studies described in the literature used i-vector features specifically to predict AD from speech. We think that, owing to the nature of factor analysis, which is used to obtain information about speaker and channel variabilities, i-vector features are able to capture efficiently the information needed in order to model an AD subject's speech in a proper way.

4 Experiments and Results

Here, we describe the experiments carried out using the i-vectors as features obtained from the speech of 225 bio-signals (i.e. utterances). Moreover, we will analyze the classification results given by the Support Vector Machines algorithm which utilized the k-fold cross-validation technique.

4.1 i-vectors Extraction

Bob Kaldi [3] was used to perform the i-vectors extraction process, it being a python wrapper for the Kaldi speech recognition toolkit [23]. In our work, 20 MFCCs features are extracted from the audio signals, which were 25 ms in duration and had a 10 ms time-shift.

Our UBM was trained relying on the BEA Hungarian Spoken Language Database that consists of spontaneous speech similar to the recordings collected from the patients. We worked with a 120 min-long set of recordings from the BEA corpus, mostly utilizing utterances from elderly subjects so as to match the age group targeted audience. The UBM was supplied with the MFCCs related to the BEA dataset in order to get a universal model of the speakers. The values of the following parameters were adjusted in order to train the UBM: the number of Gaussian components, C, from 2 to 256; and the number of Gaussians to keep per frame, C_f, was given by $log_2(C)$.

MFCC features extracted from the utterances of the MCI, HC, and AD subjects (i.e dementia dataset), were used both to model the i-vector extractor, for which we used training utterances only, and to extract i-vectors from each

Table 2. Scores obtained when SVM classifies with i-vectors.

Used recording(s)	UBM size	Performance (%)			
		Acc.	Prec.	Rec.	F_1
Immediate recall	32	42.7%	82.6%	76.0%	79.2%
Previous day	32	41.3%	72.2%	78.0%	75.0%
Delayed recall	4	46.7%	78.7%	74.0%	76.3%
All utterances	16	56.0%	80.9%	76.0%	78.4%

MFCC feature vector (i.e. using train, development and test utterances, respectively). The i-vector extractor model was fitted using the UBM as well as the MFCC features extracted from the *dementia* dataset. Then, the i-vector extractor makes use of the i-vector extractor model together with the UBM to extract the i-vector features from each utterance.

4.2 Evaluation

We performed our classification with the use of Support-Vector Machines [27] and we relied on the libSVM implementation [4]. To avoid overfitting due to having a large number of meta-parameters, we applied a linear kernel; the value of complexity (C) was set in the range $10^{\{-5,-4,\ldots,0,1\}}$. The subjects were classified using 5-fold cross-validation. Each fold contained the utterances of 5 healthy controls, 5 speakers having AD, and 5 speakers suffering from MCI. Each SVM model was trained on the utterances of 60 subjects.

The evaluation was carried out in 4 ways, where we measured the performance of the recordings: immediate recall, previous day, delayed recall, and all utterances together, respectively. Table 2 lists the results got in terms of F1-scoring and accuracy. The best F1-score outcome belongs to the immediate recall measurement. However, the best accuracy score was obtained when using all the utterances. It can be seen that Immediate Recall and Previous Day recordings performed the best with 32 Gaussian components in the UBM; but this is not true for Delayed recall, and All utterances evaluations, they performed the best when the size of the UBM was 4 and 16, respectively.

Figure 2 shows a big difference between the values of accuracy related to the set 'All tasks' and the accuracy scores from the other set of tasks (i.e. Immediate recall, Previous day, and Delayed recall). This happens because the accuracy score was measured as a 3-wise set, that is, it was obtained in terms of the AD, MCI, and HC classifications. This means that SVM had a 3-class classification with an accuracy score of 56%. In contrast, a 2-wise set used in the rest of the scores, that is, AD and MCI were treated as one class, while HC was the other class, which allowed the classifier to perform better. Thus here the evaluation was basically whether the subject has dementia (AD or MCI) or the subject is healthy (HC). The same figure describes the number of Gaussian components required to get the best results in terms of accuracy, it turns out that the best configurations

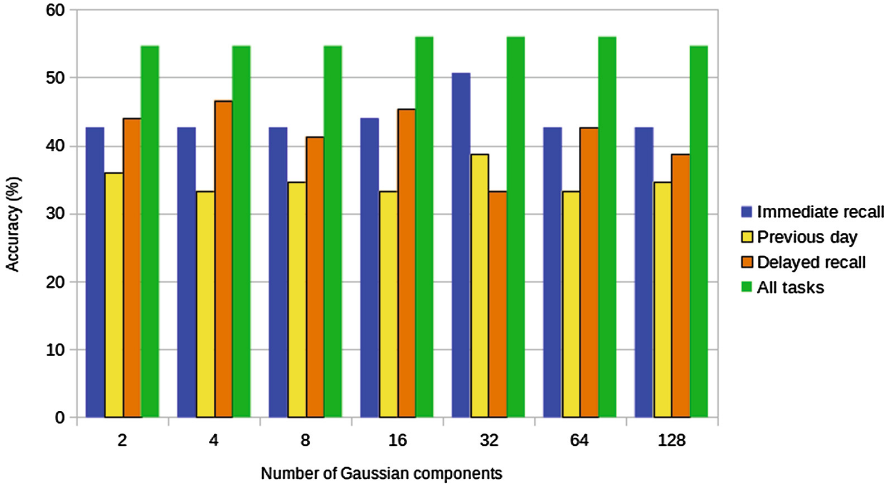

Fig. 2. Achieved accuracy scores in terms of the number of Gaussian components.

were obtained when using the number of Gaussian components was less than 32 in the case of Immediate Recall and Previous Day tasks. For Delayed Recall just 4 components were needed. When all the utterances were combined, it was enough to use 16 Gaussian components so as to achieve the best accuracy scores with less computation time. Thus, i-vector features in these experiments performed better when using smaller number of Gaussian components.

It should be mentioned that the best configuration of the number of components C in the SVM classifier differed in relation to the type of recordings used, i.e. for the best F1-score (Immediate recall) $C = 10^{-2}$, while for the best accuracy (All utterances) $C = 10^{-3}$. A complexity constant value that is too large may lead to overfit the model; on the other hand, a value that is too small may result in over-generalization. Here, the best SVM complexity constant values, which set the tolerance for misclassification, were low in the two best cases, which means that C just needed 'hard' boundaries of tolerance to perform the best, and over-fitting was controlled by the cross-validation.

5 Conclusions and Future Work

Alzheimer's Disease is currently very difficult to diagnose accurately, and the methods of diagnosis generally comprise several costly and time-consuming tasks that the patient may be asked to repeat more than once. A successful and precise diagnosis might be relative due to the fact that it is strongly dependent of the expertise of the physician. Mild Cognitive Impairment is commonly viewed as a prodromal stage of Alzheimer's, it causes a gentle-yet-noticeable decline in cognitive abilities (i.e. memory and thinking). Generally speaking, a person with MCI has a relatively high risk of developing AD or another type of dementia disease. Unfortunately, the successful diagnosis of MCI greatly depends on the doctor's experience and judgement which may not be the most accurate. MCI diagnosis is also based on the costly biomaker tests (e.g. brain imaging and cerebrospinal fluid tests).

In this paper, we showed how speech analysis offers a non-intrusive, non-expensive and faster way to perform the diagnosis of Alzheimer's by means of the utterances (i.e. speech recordings) of subjects. Here, we presented the advantage of i-vectors as features to model the particular speech of an Alzheimer's sufferer. Two groups of speech signals were represented via MFFCs features, one for the BEA Hungarian Spoken Language Database and the other got from the *dementia* dataset. Next, i-vector modeling was performed over these features with the goal of extracting their total factors (i.e. i-vector features). SVM utilized these i-vectors and classified them using a linear kernel. It achieved an F1 score of 79.2% for the three groups, namely, Alzheimer Disease (AD), Mild Cognitive Impairment (MCI), and Healthy Control (HC).

We tested the i-vector features by means of 5-fold cross-validation to avoid overfitting. Evaluation took place over three types of recordings (Immediate recall, Previous day, Delayed recall) from each of the 75 speakers, plus one more evaluation over all these together.

In a future study, we intend to perform a standard i-vector preprocessing before classifying them with the SVM. LDA (Linear Discriminant Analysis) and WCCN (Within-class Covariance Normalization) are commonly used on i-vector features in order to achieve the compensation for the intersession problem. We expect that, with the use of LDA, undesired information may be removed from the total factors (i.e. i-vectors) and that the variance between speakers can be maximized (discrimination of multiple classes); on the other hand, WCCN can be utilized to compensate the intersession variability. Such processes on i-vectors may lead to a dimension reduction in the features which should cut CPU time and make it easier to classify them.

Acknowledgments. This study was partially funded by the National Research, Development and Innovation Office of Hungary via contract NKFIH FK-124413 and by the Ministry of Human Capacities, Hungary (grant 20391-3/2018/FEKUSTRAT). László Tóth was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

References

1. Baldas, V., Lampiris, C., Capsalis, C., Koutsouris, D.: Early diagnosis of Alzheimer's type dementia using continuous speech recognition. In: Lin, J.C., Nikita, K.S. (eds.) MobiHealth 2010. LNICST, vol. 55, pp. 105–110. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20865-2_14
2. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: Support vector machines using GMM supervectors for speaker verification. IEEE Signal Process. Lett. **13**(5), 308–311 (2006)
3. Cernak, M., Komaty, A., Mohammadi, A., Anjos, A., Marcel, S.: Bob speaks Kaldi. In: Proceedings of Interspeech, August 2017

4. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 1–27 (2011)
5. Dehak, N., et al.: Support vector machines and joint factor analysis for speaker verification. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4237–4240. IEEE (2009)
6. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2011)
7. Folstein, M., Folstein, S., McHugh, P.: Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. **12**(3), 189–198 (1975)
8. Förstl, H., Kurz, A.: Clinical features of Alzheimer's disease. Eur. Arch. Psychiatry Clin. Neurosci. **249**(6), 288–290 (1999). https://doi.org/10.1007/s004060050101
9. Frank, E.: Effect of Alzheimer's disease on communication function. J. S. C. Med. Assoc. **9**(90), 417–23 (1994)
10. Fraser, K., Rudzicz, F., Graham, N., Rochon, E.: Automatic speech recognition in the diagnosis of primary progressive aphasia. In: Proceedings of SLPAT, Grenoble, France, pp. 47–54 (2013)
11. Fraser, K.C., et al.: Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. Cortex **55**, 43–60 (2014)
12. Fraser, K.C., Rudzicz, F., Rochon, E.: Using text and acoustic features to diagnose progressive aphasia and its subtypes. In: Proceedings of Interspeech, Lyon, France, pp. 25–29 (2013)
13. Freedman, M., Leach, L., Kaplan, E., Winocur, G., Shulman, K., Delis, D.: Clock Drawing: A Neuropsychological Analysis. Oxford University Press, New York (1994)
14. Galvin, J.E., Sadowsky, C.H.: Practical guidelines for the recognition and diagnosis of dementia. J. Am. Board Fam. Med. **25**(3), 367–382 (2012)
15. Ganchev, T., Fakotakis, N., Kokkinakis, G.: Comparative evaluation of various MFCC implementations on the speaker verification task. In: Proceedings of the SPECOM, vol. 1, pp. 191–194 (2005)
16. García, N., Orozco-Arroyave, J.R., D'Haro, L.F., Dehak, N., Nöth, E.: Evaluation of the neurological state of people with Parkinson's Disease using i-vectors. In: INTERSPEECH (2017)
17. García, N., Vásquez-Correa, J., Orozco-Arroyave, J.R., Nöth, E.: Multimodal i-vectors to detect and evaluate Parkinson's disease. In: Proceedings of Interspeech 2018, pp. 2349–2353 (2018)
18. Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., Hoffmann, I.: Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. Comput. Speech Lang. **53**, 181–197 (2019). http://www.sciencedirect.com/science/article/pii/S088523081730342X
19. Grzybowska, J., Kacprzak, S.: Speaker age classification and regression using i-vectors. In: INTERSPEECH, pp. 1402–1406 (2016)
20. Hansen, J.H.L., Hasan, T.: Speaker recognition by machines and humans: a tutorial review. IEEE Signal Process. Mag. **32**(6), 74–99 (2015). https://doi.org/10.1109/MSP.2015.2462851
21. Kenny, P.: Joint factor analysis of speaker and session variability: theory and algorithms. CRIM, Montreal, (Report) CRIM-06/08-13, vol. 14, pp. 28–29 (2005)
22. Lehr, M., Prud'hommeaux, E., Shafran, I., Roark, B.: Fully automated neuropsychological assessment for detecting mild cognitive impairment. In: Proceedings of Interspeech, Portland, OR, USA, pp. 1039–1042 (2012)

23. Madikeri, S., Dey, S., Motlicek, P., Ferras, M.: Implementation of the standard i-vector system for the Kaldi speech recognition toolkit. Idiap-RR Idiap-RR-26-2016, Idiap, October 2016
24. Nelson, L., Tabet, N.: Slowing the progression of Alzheimer's disease; what works? Ageing Res. Rev. **23**(B), 193–209 (2015)
25. Rosen, W., Mohs, R., Davis, K.: A new rating scale for Alzheimer's disease. J. Psychiatry Res. **141**(11), 1356–1364 (1984)
26. Satt, A., Hoory, R., König, A., Aalten, P., Robert, P.H.: Speech-based automatic and robust detection of very early dementia. In: Proceedings of Interspeech, Singapore, pp. 2538–2542 (2014)
27. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. **13**(7), 1443–1471 (2001)