USING SPECTRAL SEQUENCE-TO-SEQUENCE AUTOENCODERS TO ASSESS MILD COGNITIVE IMPAIRMENT

Mercedes Vetráb[1], *José Vicente Egas-López*[1], *Réka Balogh* [2], *Nóra Imre* [2], *Ildikó Hoffmann* [3,4], *László Tóth* [1], *Magdolna Pákáski* [2], *János Kálmán* [2], *Gábor Gosztolya* [1,5]

[1] University of Szeged, Institute of Informatics, Szeged, Hungary
[2] University of Szeged, Department of Psychiatry, Szeged, Hungary
[3] Research Institute for Linguistics, Budapest, Hungary
[4] University of Szeged, Department of Linguistics, Szeged, Hungary
[5] ELRN-SZTE Research Group on Artificial Intelligence, Szeged, Hungary

ABSTRACT

Dementia is a chronic or progressive clinical syndrome, mainly characterized by the deterioration of memory, thinking, reasoning and language. In Mild cognitive impairment (MCI), often considered as the prodromal stage of dementia, there is also a subtle deterioration of these functions, but they do not affect the daily life of the patient. However, due to the slight nature of the changes, it is quite hard to diagnose MCI. In this study, we employ sequence-to-sequence deep autoencoders in order to extract compact, robust and efficient attributes from the spontaneous speech of 25 MCI subjects and 25 healthy controls. From our results, this approach gives a competitive performance, as we significantly outperformed x-vectors even though they were trained on more data. Our additional efforts to identify mild Alzheimer's (mAD) subjects as well were less successful; but since the focus is on the early detection of dementia, this is not a limitation of the methodology from a practical point of view.

Index Terms— mild cognitive impairment, dementia, sequence-to-sequence autoencoders

1. INTRODUCTION

Mild cognitive impairment (MCI) is a heterogeneous clinical syndrome characterized by the deterioration of memory, language, and problem-solving skills. It is often viewed as the transitional stage between normal aging and dementia [1]. However, in contrast to those with dementia, the cognitive impairments that occur in MCI are not severe enough to affect the patients' ability to carry out simple everyday activities [1, 2]. MCI may be present up to 15 years before the clinical manifestation of dementia [3], and this time window offers a chance for early MCI detection, which can provide an opportunity to reduce the rate of cognitive decline [4].

Changes in language performance can act as an early indicator of MCI, since language-related alterations can appear long before the manifestation of other distinctive cognitive symptoms [5]. Changes in language production are related to the subclinical decline in memory; for example, the fluency of spontaneous speech has been shown to deteriorate for subjects with early MCI [6]. Their speech contains an increasing amount of pauses and disfluencies with the progression of the disease [7], most likely attributable to the word retrieval difficulties of the patients [8]. These characteristics can have a strong effect on the patient's speech; therefore, analyzing speech permits the indirect investigation of cognition.

Taking this into account, automatic speech analysis could prove to be a cheap, easy-to-apply, remote and non-invasive tool for detecting the symptoms of MCI. Quite recently, several studies were published on detecting MCI and other forms of dementia [9, 10]. However, it is still unknown which feature types are worth extracting from the speech of the subjects. A plausible choice is to employ the ever-growing pool of general (that is, non-specific to the actual disease) feature extractors, such as i-vectors [11] and x-vectors [12]. (These two techniques were developed for speaker verification, but were later employed as feature extractors in other tasks as well [13, 14].) The main advantage of these approaches is that they do not have to be trained on limited-sized MCI and healthy control (HC) speech data, but instead general, large speech corpora (like those used to train Automatic Speech Recognition acoustic models) might be utilized for this aim.

Noting the popularity of deep learning-based techniques, in this study we apply sequence-to-sequence deep autoencoders for extracting features in order to distinguish the

This study was supported by the National Research, Development and Innovation Office of Hungary via contract NKFIH FK-124413. It was also supported by grant NKFIH-1279-2/2020 of the Ministry for Innovation and Technology, Hungary, and by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program (MILAB). G. 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 Program ÚNKP-21-5-SZTE.

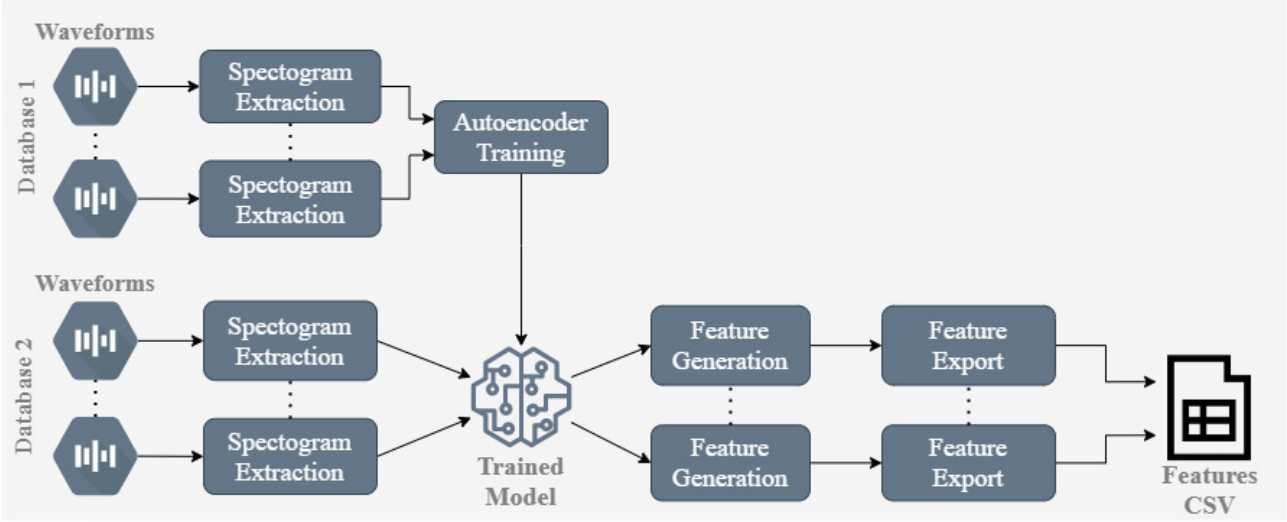

Fig. 1. The general workflow of the sequence-to-sequence autoencoder based feature extraction process, we applied.

speech of MCI and healthy control subjects. We expect benefits from the fact that they process directly the spectrum of the raw waveform, without relying on manually feature-engineered attributes like frame-level MFCCs. Furthermore, they were successfully applied to tasks like acoustic event classification [15] and categorizing the sounds of primates [16]. The novelty of our study lies in the use of sequence-to-sequence autoencoders to detect mild cognitive impairment and mild Alzheimer's disease. According to our experimental results, they are able to outperform x-vectors, even when trained only on a fraction of the data. Furthermore, to avoid peeking and to demonstrate the robustness of the extracted features, we employed a separate audio corpus for training these networks. To our knowledge, this is the first study that has employed such a cross-corpus technique in audio processing with sequence-to-sequence autoencoders.

2. SPECTRAL SEQUENCE-TO-SEQUENCE AUTOENCODERS

Autoencoders have a long history in machine learning, dating back long before deep networks [17]. The basic idea is to train a neural network to reconstruct the input (not necessarily audio), while the network structure contains a small-sized *bottleneck layer*. Evaluating the fully trained network and using the activation values of the bottleneck layer leads to a compressed representation of the input, which can be used as features in a potential classification step. For audio, due to the varying duration of the input utterances, recurrent neural networks or sequence-to-sequence autoencoders might be employed. Such techniques were successfully used in the past on various tasks like machine translation [18] and acoustic event detection [15].

Deep learning methods have shown to be more effective on raw features such as Mel-scale filter bank energies than hand-crafted attributes such as MFCCs or PLPs [19]. Therefore, the first step of the process is the extraction of Mel-scale spectograms from raw waveforms. Following the study of Amiriparian et al., the Mel-sceptra are normalized into the interval $[-1, 1]$ to match the expected input range for neural networks [15]. Next, this spectra is fed into the *encoder* part of the recurrent neural network, consisting of e.g. Long Short-Term Memory (LSTM) or Gated Recurret Unit (GRU) cells in a recurrent manner over the time axis.

The hidden states of the last cells of the encoder network (after a fully connected, compression or bottleneck layer) form the *encoded representation* of the input sequence. On the top of the encoder network, another layer of LSTM or GRU cells (i.e. the *decoder* part) is applied, which is expected to reconstruct the input frame-by-frame. Depending on the direction of this layer, the network can be unidirectional or bidirectional. The whole network is trained for input reconstruction, using the straightforward RMSE error function of the frame-level input vectors and decoder outputs. After training, the network is evaluated for each utterance, and the encoded representation (measured on the last cells of the encoder part) might be used as the compressed form (or, in practice, as a feature vector) of the utterance.

Usually, the network is trained on the same corpus that is used during the classification experiments. However, in the medical speech processing area the amount of data is extremely limited due to the availability of subjects with the given disease, and the fact that trained personnel (e.g. doctors) are required to diagnose the patients. To resolve this, we trained our neural network models on a different, general audio dataset, in the hope that we could demonstrate the robustness of the feature extraction technique. For the general workflow of the approach employed, see Figure 1.

3. DATA

Our utterances were recorded at the Memory Clinic at the Department of Psychiatry of the University of Szeged, Hungary. The data sets were recorded with a digital voice recorder and a tie clip microphone. Recordings had a sampling rate of 44.1 kHz in stereo. Later the recordings were converted to 16 kHz mono with a 16 bit resolution. A total of 50 subjects, selected from a larger pool of test participants, were used in the current study: 25 MCI patients and 25 healthy controls. We selected these subjects in order to ensure that the two study groups did not significantly differ from each other with regard to gender ($p = 0.734$), age ($p = 0.150$) and years of education ($p = 0.214$). All the subjects were right-handed and native speakers of Hungarian. The exclusion criteria were drug or alcohol consumption; being under pharmacological treatment affecting cognitive functions; depression; a medical history of head injuries or psychosis; and visual or auditory deficits. MCI patients were selected after a medical diagnosis supported by neuropsychological tests and CT or MRI. We focused on spontaneous speech: in our protocol, the subjects were asked to talk about their previous day. The duration of the responses lay in the range of 25...325 seconds with a mean duration of 89.8s.

4. EXPERIMENTAL SETUP

4.1. Sequence-to-Sequence Autoencoders

The autoencoder feature extractor models were trained on a subset of the BEA corpus [20], containing Hungarian spontaneous speech. We employed a small subset, consisting of the speech of 16 subjects with a total duration of 3 hours and 59 minutes. We used the `AuDeep` toolkit [21], which was written in Python. Following the results of preliminary tests, we applied 128 log-scale Mel-spectogram filters with 0.08ms wide windows and a 0.04ms overlap. We used the Adam optimiser with a learning rate of 0.001, and applied dropout with a 0.2 probability. We used 2 recurrent layers, each one consisting of 128 GRU cells, and a bidirectional decoder. We trained our models with a mini-batch size of 64 for 32 epochs. `AuDeep` normalizes all the computed spectrograms to 0 dB. As suggested by the literature (see e.g. [15, 16]), we experimented with removing background noise by clipping power levels below a given dB value. When we set a threshold it was applied after the spectogram normalization. For this, we applied thresholds of -30, -45, -60 and -75 dB, and we tried concatenating the feature vectors of these four variations ("Merged" approach), and without clipping as well.

4.2. Data Preprocessing

Although the sequence-to-sequence autoencoders in theory can handle utterances with any duration, due to implementation constraints of `Tensorflow`-based toolkits, in practice only objects with a limited size (in our case, duration) colud be processed. Therefore, we split all recordings of the BEA corpus into 5-second-long chunks before training our models. Similarly, we repeated this split with the responses of the subjects; from the 50 utterances, we got 1371 chunks overall. Since the length of the utterances (containing the responses of the subjects) varied from subject to subject, the number of such chunks ranged from 5 to 60, the mean being 27.42.

Table 1. The accuracy (Acc.) and AUC scores obtained with the different approaches tested.

Feature extraction approach		Acc.	AUC
Sequence-to-sequence autoencoders	-30 dB	64%	0.694
	-45 dB	60%	0.706
	-60 dB	68%	0.734
	-75 dB	**72%**	**0.763**
	Merged	68%	0.643
	Unclipped	68%	0.715
Duration only		60%	0.615
x-vectors		60%	0.680

4.3. Classification

A linear SVM was employed for classification, using the libSVM implementation [22]; the C complexity parameter was set in the range $10^{-5}, 10^{-4}, \ldots, 10^{1}$. We used 25-fold stratified cross-validation (CV): each fold consisted of the data of one healthy and one MCI subject. Performance was measured by the classification accuracy and equal error rate (EER), and by the AUC value. All features were standardized before utilizing them in the classification step. Besides the embeddings extracted from the autoencoders, we used one further attribute: the number of chunks associated with the given speaker (roughly estimating the utterance length).

We performed classification at the level of the 5-second chunks. To aggregate the predictions obtained for the chunks of the same speaker into one prediction for the given subject, we simply took the (unweighted) arithmetical mean of the posterior scores for the two classes (i.e. MCI and HC).

5. RESULTS

Examining the results (see Table 1), we can see that clipping the power levels below a certain dB threshold clearly affects the MCI classification performance. In this case, the highest threshold (-75 dB) led to the best accuracy and AUC scores (72% and 0.763, respectively), although the values corresponding to the -60 dB case were also quite similar. Surprisingly, concatenating the four variations led to a clear fall in the values: although the 68% classification accuracy is only slightly lower than the best 72% value, the AUC score of 0.643 is the lowest one for all six cases. We also notice

Table 2. The AUC scores obtained for the approaches tested in the 3-class case.

Feature extraction approach		AUC		
		HC	MCI	mAD
Autoencoders	-30 dB	0.706	0.618	0.503
	-45 dB	0.714	0.633	0.569
	-60 dB	0.732	**0.706**	0.606
	-75 dB	**0.771**	**0.710**	0.589
	Merged	0.701	0.622	0.598
	Unclipped	0.682	0.703	**0.629**
Duration only		0.637	0.641	0.417
x-vectors		0.753	0.546	0.606

Table 3. The AUC scores obtained for the different aggregation formulas applied.

Speaker Category	Aggregation (Mean)	AUC		
		HC	MCI	mAD
MCI	Arithmetic	0.763	0.763	—
	Median	**0.782**	**0.782**	—
	Geometric	0.760	0.760	—
	Harmonic	0.749	0.749	—
MCI + mAD	Arithmetic	0.771	0.710	0.589
	Median	0.755	**0.712**	0.586
	Geometric	**0.789**	**0.716**	**0.606**
	Harmonic	**0.801**	**0.733**	**0.611**

that the scores clearly exceed those obtained by using only the utterance length (i.e. number of 5-second chunks).

For reference, we also trained an x-vector feature extractor network on the same BEA dataset, but on 60 hours and 14 seconds of data with 165 speakers, using 40 Mel-frequency filter bank energies ("FBANK"). Surprisingly, we see that both the accuracy and the AUC value significantly lag behind the scores obtained via sequence-to-sequence autoencoders. This finding shows that the methodology employed in our study indeed provides competitive scores.

5.1. Experiments With Mild Alzheimer's Subjects

In the next experiments we investigated how these features could be used to discriminate three speaker categories. That is, besides the 25 MCI and the 25 control subjects, we utilized the speech recordings of 25 mild Alzheimer's (mAD) patients as well; of course, they were also matched to the other groups in terms of age, gender and level of education. We retrained our SVM models on a 3-class task, again in a stratified cross-validation fashion. Table 2 shows our results; now we focused only on the AUC values of the individual speaker categories. Apart from, noting that, just like in the MCI and HC case, using the -75 dB cut-off threshold led to the best results, we also observe that the mAD patients could be distinguished from the other speakers with the lowest efficiency. (Or, in the -30 dB case, they could not be identified at all.) This is surprising as distinguishing the mAD subjects from healthy controls is usually regarded an easier task than detecting MCI due to the more prominent symptoms.

5.2. Experiments With Further Aggregation Approaches

Recall that, after evaluating the classifier SVM models on the 5-second long speech chunks, we had to aggregate the corresponding predictions in some way to obtain a prediction for each speaker. For this, we simply took the mean of the posterior estimates. Next, we took a look at the efficiency of other forms of aggregation; that is, besides arithmetic mean, we evaluated the median, geometric mean and harmonic mean of the posterior scores obtained for the chunks.

In the HC vs. MCI case (see the upper half of Table 3), we can see that employing the median of the chunk-level posterior estimates was slightly better than using the standard arithmetic mean, while geometric and harmonic means gave almost identical or even slightly worse values. In the HC vs. MCI vs. mAD three-class setup, however, we note the opposite trend (see the lower half of Table 3). That is, compared to the straightforward approach of using the arithmetic mean, relying on the median value made the AUC score of the HC speaker category slightly worse (although the AUC values corresponding to the MCI and mAD patients were practically unaltered). Utilizing the geometric and the harmonic means, however, improved all three AUC values, the latter increasing it to 0.801 for the healthy control subjects. These opposing trends, however, seem to indicate the lack of robustness of these aggregation strategies.

6. CONCLUSIONS

In this study we focused on the detection of Mild cognitive impairment (MCI), often considered as a prodromal stage of dementia, from the spontaneous speech of the subjects. To employ deep learning-based feature extraction, we trained spectral sequence-to-sequence autoencoder networks on an external speech corpus containing spontaneous discussions, and utilized the compressed hidden states of the cells of the last time frame as features. Our experimental results indicate that this procedure might be an efficient way for MCI detection, as we significantly outperformed standard x-vector encoders, even though we used a fraction of its training data. Our findings might contribute to the development of an automatic, efficient, non-invasive and cost-effective MCI screening method, which does not even require a personal meeting with the subject, and would be a useful tool in the current Covid-19 pandemic situation.

7. REFERENCES

[1] R.C. Petersen, B. Caracciolo, C. Brayne, S. Gauthier, V. Jelic, and L. Fratiglioni, "Mild Cognitive Impairment: A concept in evolution," *Journal of Internal Medicine*, vol. 275, no. 3, pp. 214–228, 2014.

[2] Alzheimer's Association, "2020 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 16, no. 3, pp. 391–460, 2020.

[3] C. Laske, H.R. Sohrabi, S.M. Frost, K. López-de Ipiña, P. Garrard, M. Buscema, J. Dauwels, S.R. Soekadar, S. Mueller, and at. al, "Innovative diagnostic tools for early detection of Alzheimer's disease," *Alzheimer's & Dementia*, vol. 11, no. 5, pp. 561–578, 2015.

[4] E.A. Hahn and R.Andel, "Nonpharmacological therapies for behavioral and cognitive symptoms of mild cognitive impairment," *Journal of Aging and Health*, vol. 23, no. 8, pp. 1223–1245, 2011.

[5] K.C. McCullough, K.A. Bayles, and E.D. Bouldin, "Language performance of individuals at risk for Mild Cognitive Impairment," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 3, pp. 706–722, 2019.

[6] K.D. Mueller, R.L. Koscik, B.P. Hermann, S.C. Johnson, and L.S. Turkstra, "Declines in connected language are associated with very early mild cognitive impairment: Results from the Wisconsin registry for Alzheimer's prevention," *Frontiers in Aging Neuroscience*, vol. 9, 2018.

[7] K. López-de Ipiña, U. Martinez-de Lizarduy, P.M. Calvo, B. Beitia, J. García-Melero, E. Fernández, M. Ecay-Torres, M. Faundez-Zanuy, and P. Sanz, "On the analysis of speech and disfluencies for automatic detection of Mild Cognitive Impairment," *Neural Computing and Applications*, pp. 1–9, 2018.

[8] G. Szatlóczki, I. Hoffmann, V. Vincze, J. Kálmán, and M. Pákáski, "Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease," *Frontiers in Aging Neuroscience*, vol. 7, pp. 195, 2015.

[9] R. Haulcy and J. Glass, "Classifying alzheimer's disease using audio and text-based representations of speech," *Frontiers in Psychology*, vol. 11, 2020.

[10] P.A. Pérez-Toro, S.P. Bayerl, T. Arias-Vergara, J.C. Vásquez-Correa, P. Klumpp, M. Schuster, Elmar Nöth, and at. al, "Influence of the interviewer on the automatic assessment of Alzheimer's disease in the context of the ADReSSo challenge," in *Proceedings of Interspeech*, 2021, pp. 3785–3789.

[11] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, V. Hubeika, and F. Castaldo, "Support Vector Machines and Joint Factor Analysis for speaker verification," in *Proceedings of ICASSP*, 2009, pp. 4237–4240.

[12] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker verification," in *Proceedings of ICASSP*, 2018, pp. 5329–5333.

[13] J. Grzybowska and S. Kacprzak, "Speaker age classification and regression using i-vectors.," in *Proceedings of Interspeech*, 2016, pp. 1402–1406.

[14] M. Huckvale, A. Beke, and M. Ikushima, "Prediction of sleepiness ratings from voice by man and machine," in *Proceedings of Interspeech*, 2020, pp. 4571–4575.

[15] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence-to-sequence autoencoders for unsupervised representation learning from audio," in *Proceedings of DCASE*, 2017, pp. 17–21.

[16] B.W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, and at. al, "The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates," in *Proceedings of Interspeech*, 2021, pp. 431–435.

[17] R. Hecht-Nielsen, "Replicator Neural Networks for universal optimal source coding," *Science*, vol. 269, pp. 1860–1863, 1995.

[18] M.-T. Luong, Q.V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," in *Proceedings of ICLR*, 2016.

[19] A-R. Mohamed, G.E. Dahl, and G. Hinton, "Acoustic modeling using Deep Belief Networks," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 14–22, 2011.

[20] T. Neuberger, D. Gyarmathy, T.E. Gráczi, V. Horváth, M. Gósy, and A. Beke, "Development of a large spontaneous speech database of agglutinative Hungarian language," in *Proceedings of TSD*, 2014, pp. 424–431.

[21] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and Björn Schuller, "auDeep: Unsupervised learning of representations from audio with Deep Recurrent Neural Networks," *Journal of Machine Learning Research*, vol. 18, no. 173, pp. 1–5, 2018.

[22] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.