frontiers in Psychology

BRIEF RESEARCH REPORT
published: 12 February 2021
doi: 10.3389/fpsyg.2020.623237

Towards Computer-Based Automated Screening of Dementia Through Spontaneous Speech

Karol Chlasta[1,2]* *and Krzysztof Wołk*[1]

[1] Department of Computer Science, Polish-Japanese Academy of Information Technology, Warsaw, Poland, [2] Institute of Psychology, SWPS University of Social Sciences and Humanities, Warsaw, Poland

Dementia, a prevalent disorder of the brain, has negative effects on individuals and society. This paper concerns using Spontaneous Speech (ADReSS) Challenge of Interspeech 2020 to classify Alzheimer's dementia. We used (1) VGGish, a deep, pretrained, Tensorflow model as an audio feature extractor, and Scikit-learn classifiers to detect signs of dementia in speech. Three classifiers (LinearSVM, Perceptron, 1NN) were 59.1% accurate, which was 3% above the best-performing baseline models trained on the acoustic features used in the challenge. We also proposed (2) DemCNN, a new PyTorch raw waveform-based convolutional neural network model that was 63.6% accurate, 7% more accurate then the best-performing baseline linear discriminant analysis model. We discovered that audio transfer learning with a pretrained VGGish feature extractor performs better than the baseline approach using automatically extracted acoustic features. Our DepCNN exhibits good generalization capabilities. Both methods presented in this paper offer progress toward new, innovative, and more effective computer-based screening of dementia through spontaneous speech.

Keywords: dementia detection, prosodic analysis, affective computing, transfer learning, convolutional neural network, machine learning, speech technology, mental health monitoring

OPEN ACCESS

Edited by:
Fasih Haider,
University of Edinburgh,
United Kingdom

Reviewed by:
Fahim Salim,
University of Twente, Netherlands
Maria Koutsombogera,
Trinity College Dublin, Ireland

***Correspondence:**
Karol Chlasta
karol@chlasta.pl

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *29 October 2020*
Accepted: *30 December 2020*
Published: *12 February 2021*

Citation:
Chlasta K and Wołk K (2021) Towards Computer-Based Automated Screening of Dementia Through Spontaneous Speech. Front. Psychol. 11:623237. doi: 10.3389/fpsyg.2020.623237

1. INTRODUCTION

One of the most important social problems in developed countries is the constant rise of the percentage of the elderly population. A major health issue affecting this segment of population is the appearance Alzheimer's dementia (AD), affecting around 50 million people worldwide and expected to grow three times over the next 50 years (Baldas et al., 2010).

Dementia is estimated to be responsible for 11.2% of years lived with disability in people over 60 years of age, compared with 9.5% for stroke, 5.0% for cardiovascular disease, and 2.4% for cancer. In Europe, the prevalence of AD increases exponentially with age. The incidence also increases with age, although with a plateau in extreme old age (Todd and Passmore, 2009).

Comorbidity of several physical and mental health disorders was studied in relation to age and socioeconomic deprivation. The presence of mental health disorders increased as the number of physical morbidities increased, and was much greater in more deprived than in less deprived people. Physical-mental health comorbidity is very common, with depression and painful disorders as key comorbidities, and with dementia seen in a small reverse gradient (Barnett et al., 2012).

There is a significant relation between old-age depression and subsequent dementia in patients over the age of 50. This supports the hypothesis of old-age depression being a predictor, and possibly a causal factor of subsequent dementia (Buntinx et al., 1996).

Speech is a well-established early indicator of cognitive deficits including dementia (Bucks et al., 2000). Speech processing methods offer great potential to fully automatically screen for prototypic indicators in near real time, and they can be used as an additional information source when diagnosing Alzheimer's disease (Weiner et al., 2016).

Dementia was detected in speech with voice activity detection and speaker diarization followed by extraction of acoustic features. The unsupervised system achieved up to 0.645 unweighted average recall (UAR). Authors detected dementia using speech segments as short as 2.5 min, but achieved the best results using segments in the range between 10 and 15 min (Weiner et al., 2018).

Other AD detection approaches combined extraction of acoustic and linguistic features (Speech to Text and Human Transcriptionist), and applied a one-way ANOVA for feature selection. The reported binary classification accuracy on brief (less than 10 min) spontaneous speech samples reached 88%, with recall of 0.920 (Jarrold et al., 2014).

We target the classification task of AD Recognition through Spontaneous Speech (ADReSS Challenge 2020). The AD classification task consists of creating binary classification models to distinguish between AD and non-AD patient speech on the ADReSS dataset. The authors of that challenge prepared the dataset and provided five baseline, machine learning classification models, that used both acoustic and linguistic features for the detection of AD in spontaneous speech. Their acoustic approaches were based on emobase (Eyben et al., 2010), ComParE 2013 (Eyben et al., 2013), Multi-resolution Cochleagram features (MRCG) proposed by Chen et al. (2014), the Geneva minimalistic acoustic parameter set (eGeMAPS) by Eyben et al. (2015), and minimal feature set (Luz, 2017). The best baseline accuracy was achieved by linear discriminant analysis (LDA) model using ComParE features.

In this paper, we propose two methods for speech-based screening of AD. Our models perform significantly better than the ADReSS challenge baseline for classification task, as evaluated on the same, official ADReSS challenge dataset.

2. METHODS

2.1. Dataset

The dataset for the 2020 ADReSS challenge consists of speech recordings elicited for the Cookie Theft picture description task from the Boston Diagnostic Aphasia Exam (Goodglass et al., 2001). These data were balanced by the organizers in terms of age, gender, and the distribution of labels between the training and test partitions in order to minimize the risk of bias in the prediction tasks. The dataset from 78 non-AD subjects, and 78 AD subjects, was labeled for binary classification and regression tasks. The labels for the binary classification include Alzheimer's dementia and healthy control, whereas the labels for the regression task are Mini-Mental State Examination (MMSE) scores (Folstein et al., 1975), which provide a means for dementia diagnosis based on linguistic tests. For more details regarding the dataset, including the segmentation and voice activity detection algorithm, we refer the reader to the ADReSS challenge baseline paper (Luz et al., 2020).

2.2. VGGish Model and Scikit-Learn Classifiers

We extended the method of Pons Puig et al. (2018) and conducted two-step classification experiments to detect cognitive impairment due to AD (as shown in **Figure 1**). This consisted of a two-stage classification process, where a classifier was trained with features to predict whether a speech segment was uttered by a non-AD or AD patient, and majority vote (MV) classification, which assigned each subject an AD or non-AD label based on the majority labels classification.

2.2.1. Feature Extraction

We used VGGish (Hershey et al., 2017), a deep, pretrained Tensorflow (Abadi et al., 2016) model as a feature extractor. VGGish is an audio embedding produced by training a modified VGGNet model (Simonyan and Zisserman, 2014) to predict video tags from the Youtube-8M dataset (Abu-El-Haija et al., 2016). Principal component analysis (PCA) (Cao et al., 2003) was used for dimensionality reduction, with PCA set to 128. VGGish model converted audio input features into high-level 128-D embedding, which was fed as an input to a downstream classification model. The features were extracted from non-overlapping audio patches of 0.96 s, where each audio patch covered 64 mel bands and 96 frames of 10 ms each.

2.2.2. Classification Methods

We performed classification experiments using five different methods, namely support vector machines (SVM, with a radial basis function kernel and scaling gamm), linear support vector machines (LSVM), perceptron, multi-layer perceptron classifier (MLP, with 20 hidden layers, using a stochastic gradient descent solver, 600 iterations, learning rate of 0.001), and nearest neighbor (1NN, for KNN with $K = 1$ and cosine metric).

2.3. DemCNN—Custom Convolutional Neural Network

Current deep convolutional neural network (CNN) performs considerably better than the previous state-of-the-art (Krizhevsky et al., 2012). Transfer learning was often used in medical image analysis (Cheplygina et al., 2019). Applying transfer learning on a wide range of tasks nearly always gave better results (Kornblith et al., 2019). CNN-based methods have been successfully employed to medical imaging tasks and achieved human-level performance in classification tasks (Esteva et al., 2019). CNNs have proven very effective in image classification and show promise for audio (Hershey et al., 2017). We extend the audio classification work presented in Wołk, K., and Wołk (2019) and Chlasta et al. (2019).

2.3.1. Classification Method

We introduce DemCNN, a custom PyTorch (Paszke et al., 2019) CNN. We designed and implemented a custom sequential architecture consisting of six Conv1D layers using ReLU activation function, batch normalization and dropout, with

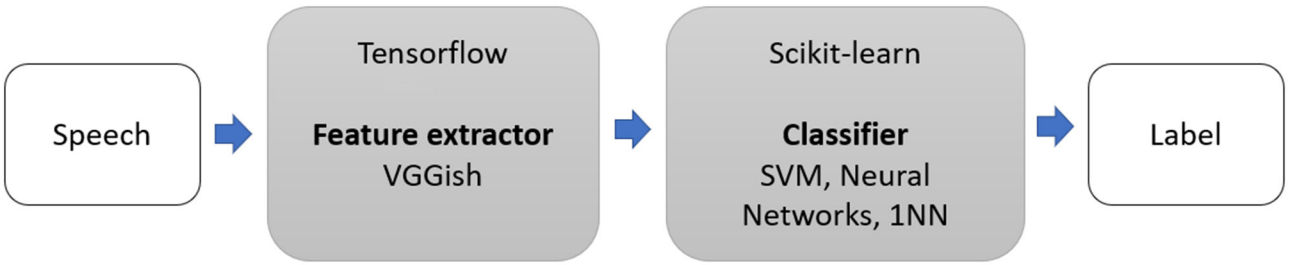

FIGURE 1 | Two-stage architecture: VGGish model and Scikit-learn classifiers.

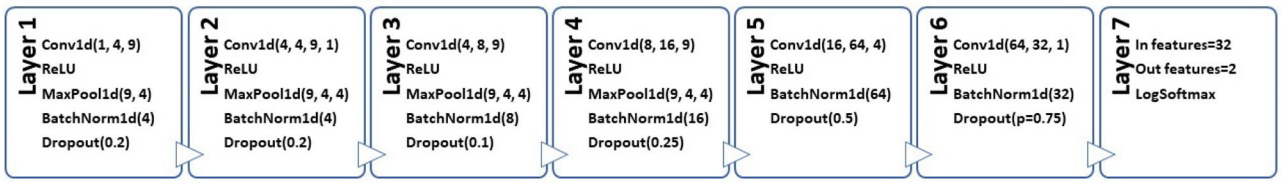

FIGURE 2 | Architecture diagram of DemCNN, a custom PyTorch convolutional neural network for speech classification.

the final (seventh) output layer being a dense layer. The output layer had 2 nodes (num_labels), which matched the number of possible classifications outputs. **Figure 2** presents a more detailed architecture diagram of our custom CNN for speech classification.

We unpacked a byte-string for each file into a 1D numpy array of numbers that could be analyzed by the CNN. Subsequently, the dataset was downsampled with a low-pass filter (with downsampling factors of 4, 4, 2).

We performed a two-step training of our CNN model using a cross-entropy loss function. We fine-tuned learning rate, the number of training cycles, and the number of training iterations per cycle. We set the first (training) batch size to 32, and the second (deployment) batch size to 2. The selection of the second learning rate for each step of our method was automated using a custom function operating on standard lr_finder. We trained the classifier for 2 or 4 epochs.

3. EXPERIMENTS AND RESULTS

All experiments were implemented in Python using Scikit-learn (Pedregosa et al., 2011), Tensorflow (Abadi et al., 2016), and PyTorch (Paszke et al., 2019) on the Google Colaboratory Platform (Bisong, 2019). The platform uses Jupyter Notebook standard that facilitates exchange of source code and reproducibility of results. The source code and accompanying results are available on GitHub.[1]

The ADReSS development data were split into train and test sets by randomly assigning 80% of the speakers to the train set and 20% to the test set. Results obtained for different classifier setups are summarized in **Table 1**.

Three models we developed using the first approach (VGGish + 128 PCA + linearSVM/perceptron/1NN) achieved 59% accuracy in our test set. Employing the same setup with SVN model, we achieved 55% accuracy. The best-performing baseline SVM models using MRCG features proposed by (Chen et al., 2014) and the ComParE 2013 features (Eyben et al., 2013) achieved lower accuracy of 53%. Interestingly, our 1NN model achieved better results than the best-performing baseline 1NN model using ComParE features (59% against 57%).

Our custom raw waveform DemCNN system achieved the best classification accuracy of 63.6%. The model classified 14 speakers correctly, eight incorrectly, and proved the most effective in distinguishing between AD and non-AD speech samples on the full wave enhanced ADReSS audio dataset. This result was 7% better then the best baseline classification accuracy on the ADReSS training set (Luz et al., 2020).

The final results for our custom audio DemCNN model were submitted to the 2020 ADReSS Challenge organizers after retraining the classifier on the full ADReSS training set, and predicting on the full ADReSS test set (see **Table 2** for results and the accompanying hyperparameters). Our model performed slightly better (1%) on the test partition than the best baseline LDA model trained on automatically extracted ComParE feature set (Eyben et al., 2013).

4. DISCUSSION

The main limitations of the AD field are poor standardization, limited comparability of results, and a degree of disconnect between study aims and clinical applications (de la Fuente Garcia

[1]Code: https://github.com/KarolChlasta/ADReSS-Challenge2020

TABLE 1 | Summary of classification results on AD Recognition through Spontaneous Speech (ADReSS) training set.

Model type	Precision	Recall	F1 score	Accuracy	Baseline accuracy
SVM	0.556	0.454	0.500	0.545	0.565 (SVM + Minimal)
LinearSVM	0.600	0.545	0.571	**0.591**	0.565 (SVM + Minimal)
Perceptron	0.600	0.545	0.571	**0.591**	0.565 (LDA + ComParE)
MLP	0.429	0.273	0.333	0.454	0.565 (LDA + ComParE)
1NN	0.600	0.545	0.571	**0.591**	0.574 (1NN + ComParE)
DemCNN	0.692	0.692	0.692	**0.636**	0.565 (LDA + ComParE)

Our approaches (VGGish + 128 Principal component analysis [PCA] and custom audio convolutional neural network [DemCNN]) vs. the best baseline accuracy on acoustic features. The bold values indicate best results achieved on ADReSS training dataset.

TABLE 2 | The results of Alzheimer's dementia (AD) classification task on AD Recognition through Spontaneous Speech (ADReSS) test set.

Approach	Class	Precision	Recall	F1 Score	Accuracy
DemCNN (Learning rate = 0.2;	Non-AD	0.528	0.792	0.633	0.542
Cycles = 4.4; Lengths = 8.8)	AD	0.528	0.792	0.389	0.542
DemCNN (Learning rate = 0.1;	Non-AD	0.625	0.625	0.625	**0.625**
Cycles = 2.2; Lengths = 8.8)	AD	0.625	0.625	0.625	**0.625**
Baseline acoustic features	Non-AD	0.670	0.500	0.570	0.620
(LDA + ComParE)	AD	0.600	0.750	0.670	0.620

Our approach (custom audio convolutional neural network [DemCNN]) vs. the best baseline on acoustic features (linear discriminant analysis [LDA] + CompParE). The bold values indicate best results achieved on ADReSS test dataset.

et al., 2020). Our two methods are attempting to close some of these gaps.

Data scarcity has hindered research into the relationship between speech and dementia. Recently, the community has turned to transfer learning (Yosinski et al., 2014), as a solution for a wide range of machine learning tasks for which labeled data are scarce. Selecting the right pretrained model as audio feature extractor allows to rapidly prototype competent speech classifiers.

In our first approach, we used a standard VGGish (Hershey et al., 2017), that is a popular deep audio embedding model trained on Youtube-8M video dataset (Abu-El-Haija et al., 2016). In our experiments to detect subtle changes in pathological speech, we confirmed that automatic extraction of acoustic features (Eyben et al., 2010) performs similarly to using a pretrained deep audio embedding model for feature extraction.

Similarly to us, Syed et al. (2020) also used VGGish deep acoustic embeddings in the ADReSS Challenge. They used other types of feature aggregation methods: (a) Fisher Vector encodings (FVs) and (b) Bag-of-Audio-Words (BoAW). Both achieved satisfactory results. Their VGGish and FVs model overperformed ours (59.1%) with 62.96% accuracy on the train partition, whereas their VGGish and BoAW model achieved even higher accuracy of 75%.

Our second method, the DemCNN model, for which we only performed a basic hyperparameter tuning, improved the classification results further. Moreover, the results achieved by DemCNN were similar in training and testing (63.6 vs. 62.5%), which is a good indicator of the lack of overfitting during the training process. This can be explained by a larger dropout defined in layers 5 and 6 of the network. An expected consequence of that is a good generalization capacity of our DemCNN model, which would positively impact the overall performance in clinical practice, when working with new data.

A similar approach to our DemCNN in the ADReSS Challenge was proposed by Cummins et al. (2020). Their raw segment based End-to-End CNN had four convolution layers, with the first convolution layer used to model voice source-related information or vocal tract information, such as formants. This approach achieved 71.3% accuracy on the training partition, but the reported result on the test partition was only 66.7%. Although this result is 4% better than our DemCNN, an expected consequence of a large difference between the results in training and test partitions is possibly a worse generalization capability of the network when working with new data.

An interesting opportunity for future research would be to use a combination of acoustic and linguistic features in detecting dementia. The latter approach, derived from automatic speech recognition (ASR) output, or from manual transcripts, had already been proven to detect dementia (Weiner et al., 2017), but relatively small gains were found when fusing acoustics and linguistics approaches (Cummins et al., 2020; Rohanian et al., 2020).

ADReSS Challenge 2020 helped to establish that although the linguistic systems outperforms the acoustic systems in AD (Cummins et al., 2020; Yuan et al., 2020), this result is unsurprising given that a human observer generated the transcripts manually, and they contain considerably fewer

sources of noise than the audio recordings. As a result, such systems would be difficult to implement in clinical practice.

An option to overcome that would be to combine acoustic information with linguistics systems based on transcripts generated from ASR systems. This idea would introduce automation, but also increase the complexity, and dependency on errors rate for ASR in a given language.

It may also be useful for future work to gather a large dataset combining spontaneous speech samples for several pathologies (starting with depression and dementia, especially for old-age patients) to train an improved DepCNN to distinguish different types of disorders in pathological speech.

Finally, the DementiaBank's Pitt corpus (Jost and Grossberg, 1995) is large enough for considering experiments with other, custom, or off-the-shelf deep neural network architectures.

5. CONCLUSION

In this paper, we proposed and compared two acoustic-based systems: VGGish, a pretrained Tensorflow model as audio feature extractor and Scikit-learn classifiers with DemCNN, a custom raw waveform based CNN.

In the first approach, we selected the VGGish model as feature extractor and PCA for dimensionality reduction. This approach achieved the accuracy of 59.1%, 3% better than the best baseline accuracy achieved on the train partition with acoustic feature extraction for the respective classification algorithms.

In the second approach, we presented DemCNN, our custom PyTorch audio CNN to detect signs of dementia in spoken language. According to the experiments, the proposed architecture achieved promising performance and demonstrated the effectiveness of our method, as well as good generalization capabilities. DemCNN overperformed the best baseline accuracy of LDA model (ComParE feature set) by 7% on the ADReSS training set (accuracy of 63.6%), and 1% on the test ADReSS test set (accuracy of 62.5%). Our DemCNN and End-to-End Convolutional Neural Network (Cummins et al., 2020) produced the strongest performance of the acoustic systems on the ADReSS 2020 classification task, highlighting the benefits of self-learning features.

To conclude, we demonstrated a proof-of-concept, and applicability of (1) audio transfer learning for feature extraction, (2) DemCNN, a custom raw waveform based CNN in detecting dementia through spontaneous speech. We demonstrated that (1) audio transfer learning with a pretrained VGGish feature extractor performs better then the baseline approach (Luz et al., 2020) using automatically extracted acoustic features, and that these are relatively minor improvements. Our DemCNN method (2) overperforms our VGGish method (1) by 4% and the baseline on the test partition (Luz et al., 2020) by roughly 1%.

Both approaches presented are active attempts to close the gaps in standarization of automatic AD detection, and to improve the overall comparability of results to better embed computational speech technology into clinical practice. They offer simplicity, easy deployment, and they are language independent, which could result in a wide adoption and improved accessibility in a short space of time.

This contribution is especially important now, in the time of current COVID-19 pandemic, when the need for a remote digital health assessment tool is greater than ever for the elderly and other vulnerable populations.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://dementia.talkbank.org/.

AUTHOR CONTRIBUTIONS

KC was responsible for conceptualization, algorithmic development, data analysis, investigation, validation, and writing of original draft. KW supervises the entire work and contributes to idea conceptualization, algorithmic development, manuscript revision, and approval of the submission.

ACKNOWLEDGMENTS

We thank Michael Connolly for assistance with editorial reviews that improved the manuscript.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "Tensorflow: a system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (Savannah, GA), 265–283.

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., et al. (2016). Youtube-8m: a large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*. Available online at: https://research.google/pubs/pub45619/

Baldas, V., Lampiris, C., Capsalis, C., and Koutsouris, D. (2010). "Early diagnosis of Alzheimer's type dementia using continuous speech recognition," in *International Conference on Wireless Mobile Communication and Healthcare* (Berlin; Heidelberg: Springer), 105–110. doi: 10.1007/978-3-642-20865-2_14

Barnett, K., Mercer, S. W., Norbury, M., Watt, G., Wyke, S., and Guthrie, B. (2012). Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet* 380, 37–43. doi: 10.1016/S0140-6736(12)60240-2

Bisong, E. (2019). "Google colaboratory," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform* (Berkeley, CA: Springer), 59–64. doi: 10.1007/978-1-4842-4470-8_7

Bucks, R. S., Singh, S., Cuerden, J. M., and Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of alzheimer type: evaluation of an objective technique for analysing lexical performance. *Aphasiology* 14, 71–91. doi: 10.1080/026870300401603

Buntinx, F., Kester, A., Bergers, J., and Knottnerus, J. A. (1996). Is depression in elderly people followed by dementia? A retrospective cohort study based in general practice. *Age Ageing* 25, 231–233. doi: 10.1093/ageing/25.3.231

Cao, L., Chua, K. S., Chong, W., Lee, H., and Gu, Q. (2003). A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. *Neurocomputing* 55, 321–336. doi: 10.1016/S0925-2312(03)00433-8

Chen, J., Wang, Y., and Wang, D. (2014). A feature study for classification-based speech separation at low signal-to-noise ratios. *IEEE/ACM Trans. Audio Speech Lang. Process.* 22, 1993–2002. doi: 10.1109/TASLP.2014.2359159

Cheplygina, V., de Bruijne, M., and Pluim, J. P. (2019). Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.* 54, 280–296. doi: 10.1016/j.media.2019.03.009

Chlasta, K., Wołk, K., and Krejtz, I. (2019). Automated speech-based screening of depression using deep convolutional neural networks. *Proc. Comput. Sci.* 164, 618–628. doi: 10.1016/j.procs.2019.12.228

Cummins, N., Pan, Y., Ren, Z., Fritsch, J., Nallanthighal, V. S., Christensen, H., et al. (2020). A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition. *Proc. Interspeech* 2020, 2182–2186. doi: 10.21437/Interspeech.2020-2635

de la Fuente Garcia, S., Ritchie, C., and Luz, S. (2020). Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. *J. Alzheimers Dis.* 78, 1547–1574. doi: 10.3233/JAD-200888

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al. (2019). A guide to deep learning in healthcare. *Nat. Med.* 25, 24–29. doi: 10.1038/s41591-018-0316-z

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., et al. (2015). The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* 7, 190–202. doi: 10.1109/TAFFC.2015.2457417

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia* (Barcelona), 835–838. doi: 10.1145/2502081.2502224

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia* (Firenze), 1459–1462. doi: 10.1145/1873951.1874246

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

Goodglass, H., Kaplan, E., and Barresi, B. (2001). *BDAE-3: Boston Diagnostic Aphasia Examination, 3rd Edn.* Philadelphia, PA: Lippincott Williams and Wilkins.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, C., et al. (2017). "CNN architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (New Orleans, LA). doi: 10.1109/ICASSP.2017.7952132

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., et al. (2014). "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (Baltimore, MD), 27–37. doi: 10.3115/v1/W14-3204

Jost, B. C., and Grossberg, G. T. (1995). The natural history of Alzheimer's disease: a brain bank study. *J. Am. Geriatr. Soc.* 43, 1248–1255. doi: 10.1111/j.1532-5415.1995.tb07401.x

Kornblith, S., Shlens, J., and Le, Q. V. (2019). "Do better imagenet models transfer better?" in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2661–2671. (Long Beach, CA: IEEE). 2661–2671. doi: 10.1109/CVPR.2019.00277

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Lake Tahoe, NV: Curran Associates, Inc.), 1097–1105.

Luz, S. (2017). "Longitudinal monitoring and detection of Alzheimer's type dementia from spontaneous speech data," in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)* (Thessaloniki: IEEE), 45–46. doi: 10.1109/CBMS.2017.41

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). "Alzheimer's dementia recognition through spontaneous speech: the ADReSS Challenge," in *Proceedings of INTERSPEECH 2020* (Shanghai). doi: 10.21437/Interspeech.2020-2571

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems* 32: Annual Conference on Neural Information Processing Systems 2019, eds H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Vancouver, BC), 8024–8035. Available online at: https://proceedings.neurips.cc/paper/2019

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://jmlr.org/papers/v12/pedregosa11a.html

Pons Puig, J., Nieto Caballero, O., Prockup, M., Schmidt, E. M., Ehmann, A. F., and Serra, X. (2018). "End-to-end learning for music audio tagging at scale," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018* (Paris: International Society for Music Information Retrieval), 637–644.

Rohanian, M., Hough, J., and Purver, M. (2020). "Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech," in *Proc. Interspeech* (Shanghai), 2187–2191. doi: 10.21437/Interspeech.2020-2721

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. Available online at: https://www.robots.ox.ac.uk/~vgg/publications/2015/Simonyan15/simonyan15.pdf

Syed, M. S. S., Syed, Z. S., Lech, M., and Pirogova, E. (2020). "Automated screening for Alzheimer's dementia through spontaneous speech," in *Interspeech*, eds W. Hess and M. Cooke (Shanghai: International Speech Communication Association). 1–5. doi: 10.21437/Interspeech.2020-3158

Todd, S., and Passmore, P. (2009). Alzheimers disease, the importance of early detection. *Eur. Neurol. Rev.* 110, 18–21. doi: 10.17925/ENR.2008.03.02.18

Weiner, J., Angrick, M., Umesh, S., and Schultz, T. (2018). "Investigating the effect of audio duration on dementia detection using acoustic features," in *Interspeech*, eds W. Hess and M. Cooke (Hyderabad: International Speech Communication Association), 2324–2328. doi: 10.21437/Interspeech.2018-57

Weiner, J., Engelbart, M., and Schultz, T. (2017). "Manual and automatic transcriptions in dementia detection from speech," in *Interspeech*, eds W. Hess and M. Cooke (Stockholm: International Speech Communication Association), 3117–3121. doi: 10.21437/Interspeech.2017-112

Weiner, J., Herff, C., and Schultz, T. (2016). "Speech-based detection of Alzheimer's disease in conversational German," in *Interspeech*, eds W. Hess and M. Cooke (San Francisco, CA: International Speech Communication Association), 1938–1942. doi: 10.21437/Interspeech.2016-100

Wołk, K., and Wołk, A. (2019). Early and remote detection of possible heartbeat problems with convolutional neural networks and multipart interactive training. *IEEE Access* 7, 145921–145927. doi: 10.1109/ACCESS.2019.2919485

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, NeurIPS 2014*, eds Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Q. Weinberger (Montréal, QC), 3320–3328. Available online at: https://proceedings.neurips.cc/paper/2014

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., and Church, K. (2020). Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease. *Proc. Interspeech* 2020, 2162–2166. doi: 10.21437/Interspeech.2020-2516

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.