INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Alzheimer Disease Recognition Using Speech-Based Embeddings From Pre-Trained Models

Lara Gauder[1,2]*, Leonardo Pepino*[1,2]*, Luciana Ferrer*[1]*, Pablo Riera*[1]

[1]Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina
[2]Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires (UBA), Argentina
{mgauder,lpepino,lferrer,priera}@dc.uba.ar

Abstract

This paper describes our submission to the ADreSSo Challenge, which focuses on the problem of automatic recognition of Alzheimer's Disease (AD) from speech. The audio samples contain speech from the subjects describing a picture with the guidance of an experimenter. Our approach to the problem is based on the use of embeddings extracted from different pre-trained models — trill, allosaurus, and wav2vec 2.0 — which were trained to solve different speech tasks. These features are modeled with a neural network that takes short segments of speech as input, generating an AD score per segment. The final score for an audio file is given by the average over all segments in the file. We include ablation results to show the performance of different feature types individually and in combination, a study of the effect of the segment size, and an analysis of statistical significance. Our results on the test data for the challenge reach an accuracy of 78.9%, outperforming both the acoustic and linguistic baselines provided by the organizers.
Index Terms: computational paralinguistics, ADreSSo challenge, Alzheimer's Disease recognition

1. Introduction

For many health problems, like speech pathologies, Parkinson's disease, Alzheimer's Disease (AD), and respiratory problems, the patient's speech is routinely used by doctors as one of the tools for diagnosis and monitoring of disease progression [1]. In particular, AD is characterized by a progressive decline of cognitive and functional abilities over time [2] often including language impairment, even at early stages [3]. As a consequence, many studies rely on the analysis of the speech signal as a source of clinical information for AD [4, 5].

In this work, we present results and analysis of our submission to the ADreSSo (Alzheimer's Dementia Recognition through Spontaneous Speech only) Challenge [6]. This challenge is focused on the automatic detection of AD using recordings of interviews with the subjects. A previous version of this challenge, called ADreSS, took place last year. In that case, manual transcriptions of the speech signals were provided to the participants along with the recordings. In this year's challenge, manual transcriptions are not provided, so systems have to rely solely on the speech signal for classification. The challenge includes three tasks: AD classification, Mini-Mental State Examination (MMSE) score regression, and cognitive decline inference. In this work, we present results on the AD classification task.

AD may affect the patient's speech production in terms of paralinguistic aspects like the prosodic patterns, pause patterns or quality of speech, and in terms of linguistic aspects, like choice of words or grammatical forms. Previous works have found that both acoustics and linguistic information can be used for automatic prediction of AD. In a paper about the ADreSS challenge [7], a comparison of acoustic and linguistic features showed that acoustic features resulted in an accuracy of 64.5% while linguistic features from manual transcriptions resulted in an accuracy of 85.42%. A similar trend is observed in the baseline results for this year's challenge [6], although with relatively poorer performance for the linguistic features due to the absence of manual transcriptions, which are replaced by automatic ones. In our work for this challenge, we focus on the use of acoustic features, without extracting automatic word transcriptions. Further, considering the sparsity of the available training data, we propose to use transfer learning approaches. To this end, we leverage recently released speech-based embedding models that aim to represent different aspects of the speech signal.

Pre-trained speech-based embeddings are currently being used in several speech recognition tasks, such as speech emotion recognition [8, 9, 10, 11] and automatic speech translation [12]. These compact representations can encode different speech attributes depending on the way the models are trained. Information about prosody, phonetic or lexical content may be emphasized in the representations, depending on the task used to train the models. The use of these representations, in combination with neural networks for the modeling stage, often provides an improvement over directly using signal processing features like mel frequency cepstral coefficients (MFCC) or Mel-spectrograms.

In this paper, we present our results using different types of embeddings and traditional prosodic features for the task of AD classification. We use a simple deep neural network for modeling each individual feature and their combinations. The model takes relatively short segments of speech as input and averages the resulting scores over all segments in an audio sample to create the final score. We show different analysis, including an ablation study to find the most useful features, a study of statistical significance, a comparison of the effect of the window length, and an analysis of the effect of the presence of experimenter speech in the signals. Our results on the challenge data are significantly better than the acoustic-only baseline results implemented by the organizers and also outperform the linguistic baseline that uses automatic transcriptions [6].

2. Dataset

The development dataset provided by the ADreSSo challenge consists of 166 recordings of 87 patients with AD diagnosis and 79 cognitively normal subjects. All the subjects were asked to

http://dx.doi.org/10.21437/Interspeech.2021-753

describe the Cookie Theft picture from the Boston Diagnostic Aphasia Exam. Audio files contain both the speech from the subjects and the experimenter conducting the interview. A test set with audio files from 71 subjects was used for blind evaluation of the models. Challenge participants are not provided AD labels for this test data. The complete dataset description is available in [6]. The challenge includes three tasks: an AD classification, an MMSE score regression and a cognitive decline (disease progression) inference task. We participated on the classification task, where the goal is to determine whether a subject is a control (CN) subject or a patient with AD, based only on the speech signal from the interview.

Since recordings include both the speech from the subject and the experimenter, the dataset includes segmentation information indicating where each of the two speakers speaks. In our initial inspection of the development data, though, we found that this information was inaccurate for several of the audio files. Further, we found a case where the recording included speech from more than two speakers and was also not accurately segmented to identify the subject's speech. For these reasons, we decided to work with the full audio files, without using the provided manual segmentation, assuming that the speech from the experimenters and any other speakers represent only a relatively small portion of the speech present in the signal. In Section 5.4, we show results that indicate that including the segments from the experimenters did not degrade the performance of our system.

3. Acoustic Features

Our approach for AD classification is to use embeddings, i.e., vector representations, extracted from a set of pre-trained models. These models are deep neural networks (DNN) trained on large speech dataset to solve different tasks. The embeddings are then extracted from the output of some layer of the DNN. In general, embeddings are extracted over relatively short regions of the signal and may contain only local information or include contextual information about the rest of the signal. The details on the embeddings used for this paper are described below. Further, we also include traditional features, designed for tasks like emotion recognition. All features were normalized by subtracting the mean and dividing by the standard deviation of that feature over each recording. This normalization approach resulted in better performance than global normalization where every feature is normalized with the mean and standard deviation obtained over all the training data.

3.1. eGeMAPS features

The extended Geneva minimalistic acoustic parameter set (eGeMAPS) [13] is a set of features designed specifically for affective speech tasks and includes pitch, loudness, formants and voice quality features among others. The set includes both low-level features, extracted every 10 ms of speech over windows of 25 ms, and high-level features that correspond to different statistics extracted from the low-level features. We use only the low-level descriptors of the eGeMAPS v2.0 set which contains 25 features for every time step.

3.2. Trill features

The trill model was trained to generate a non-semantic representation of speech [14]. The model minimizes a triplet loss designed to solve the task of classifying whether a segment of audio comes from the same or from a different original audio file as another segment. The resulting embeddings were evaluated in many different non-semantic tasks including speaker identification, emotion recognition and others. Authors also tested the model on AD recognition using the Dementia Bank dataset [15], showing good results when fine-tuning the model to this task. We used the distilled version of the trill model, which generates embeddings of size 2048.[1] For this work, we resampled the trill embeddings, which are produced every 167 ms to 100 vectors per second (one every 10 ms) to match the resolution of the other features.

3.3. Allosaurus features

Allosaurus[2] is a universal phone recognition model that includes pre-trained acoustic and language models [16]. It can be used to generate phonetic transcriptions and phone logits given by $\log(p/(1-p))$, where p is the phone posterior probability, producing one set of logits every 10 ms. The model was trained with 11 languages and over 2000 utterances. For English, the output logits are 39-dimensional. The information contained in these logits could help us model specific pronunciation issues in the AD subjects, as well as some indirect information about word usage which could be found in the frequency of certain phones.

3.4. Wav2vec 2.0 features

Wav2Vec 2.0 (from now on wav2vec2) is a framework for self-supervised learning of representations from raw audio [17]. The model can generate contextualized embeddings that can be later integrated in end-to-end speech-to-text models to generate transcriptions. Thus, these embeddings preserve the phone content of the signal among other information. We used the model available from the Transformer Python library which gives embeddings of size 768 and was fine-tuned with 960 hours of LibriSpeech.[3] Large audio files had to be segmented in 20 seconds long fragments to compute the features since they could not otherwise be processed by the model. This may be suboptimal since it prevents the model from extracting contextual information from the full signal. The wav2vec2 embeddings are produced every 20 ms. As for the trill features, we resampled these vectors to obtain 100 per second, matching the resolution of the other features.

4. Classification model

We use deep neural networks as models for the different individual features and their combinations. The input to the networks are 5-second segments extracted from the original audio with 1-second overlap between segments. This results in 3178 segments extracted from the development data. The final score for each audio file is obtained by computing the mean over the scores for all the segments in the file. As explained above, we use the full audio file, which means that some of the segments contain speech from the experimenter. Section 5.4 shows results on the effect of the experimenter's speech on the performance of the system.

The architecture is depicted in Figure 1 for the configuration where all features are used. Each feature set has a corresponding branch that performs a first reduction of the embedding with a 1D convolution with kernel size 1 (equivalently, a time-distributed dense layer) followed by a 1D convolution with

[1] https://tfhub.dev/google/nonsemantic-speech-benchmark/trill-distilled/3

[2] https://github.com/xinjli/allosaurus

[3] https://huggingface.co/facebook/wav2vec2-base-960h

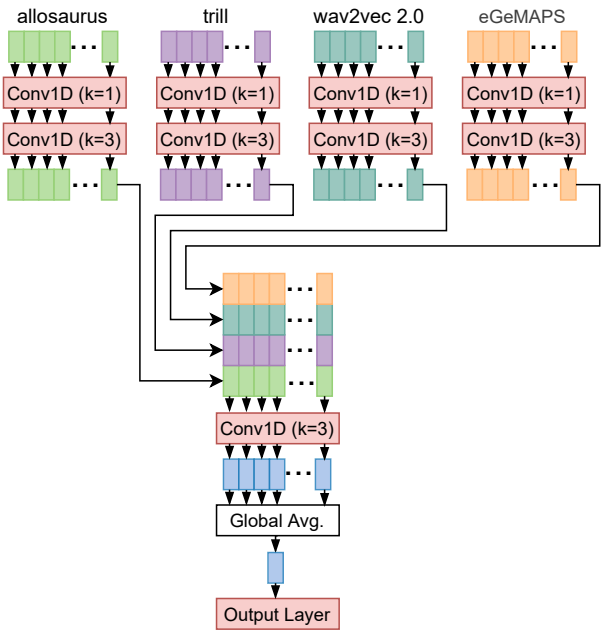

Figure 1: *Neural network architecture for the configuration in which all the features are used. The value k indicates the size of the kernel in the time dimension.*

kernel size 3. After the second convolution, the dimension of the output for each branch is 128. Then the activations from the four branches are concatenated and a second 1D convolution is performed, reducing the dimension back to 128. Finally, the output of this layer is averaged across time and a dense layer computes the prediction scores. Batch normalization and ReLu activations are used in every layer. When considering only a subset of the features, the same model is used with only the corresponding branches included.

5. Results

In this section we show results for the proposed systems, including an ablation study, statistical significance results, an analysis of the influence of the segment size and of the effect of the experimenter speech on the performance of the system. We run experiments using 6-fold cross-validation (CV) on the development data provided to challenge participants. The folds are determined by subject to prevent segments from the same subject being in the training and the test set for a certain fold, which would make the CV results overly optimistic. The models used to obtain scores for the challenge's evaluation data were trained on the full training set.

5.1. Ablation Results

The middle column in Table 1 shows the results on the development set obtained with cross-validation, for several systems including single feature sets, 2-way combinations, and the 4-way combination. The best individual features are trill and wav2vec2. We hypothesize that this may be partly because these two features sets have large dimensionality, 2048 for trill and 768 for wav2vec2 (compared to the other two which have 25 features for eGeMAPS and 39 for allosaurus), allowing these features to contain a richer representation of the audio. Larger dimensions could also result in the model overfitting the training data, which would lead to poor results, but this effect is discouraged by the small architectures we use.

Feature set	Dev	Test
eGeMAPS	63.9%	-
trill	72.9%	69%
allosaurus	66.3%	-
wav2vec2	**75.3%**	**78.9%**
eGeMAPS, allosaurus	63.9%	-
eGeMAPS, wav2vec2	72.3%	-
eGeMAPS, trill	71.1%	-
trill, allosaurus	72.9%	70.4%
trill, wav2vec2	75.2%	69%
allosaurus, wav2vec2	70.5%	-
all	74.7%	70.4%

Table 1: *Accuracy values for different combinations of features for development and test data. The five best performing models on the development set were submitted to the challenge. Results on the test set are shown for those cases.*

Fusion results show no gains with respect to the best individual system, wav2vec2. This could imply that the other three sets of features are redundant given the wav2vec2 features. That is, that wav2vec2 features contain all the information in eGeMAPs, trill and allosaurus features that is important for AD classification. In fact, we would expect allosaurus and wav2vec2 features to be somewhat redundant since they are trained to solve similar tasks: phone recognition and speech recognition, respectively. On the other hand, we would also expect trill or eGeMAPs to provide some complementary information to those two set, since they are designed to contain information beyond the phonetic content. Hence, a more likely explanation for the lack of gain from fusion is that our downstream model is not able to effectively combine the information from all these sets. In the future, we will continue exploring different architectures for the combination of these features.

Finally, the right column in Table 1 shows the results for the 5 best systems based on the development results, which were the ones selected for submission to the challenge. In the test results, as in the development results, wav2vec2 alone was the best performing model, with an accuracy of 78.9%. This result is superior to the acoustic baseline results presented in [6], which have an accuracy of 64.79%. Further, they are also superior to the linguistic baseline results in that paper, which has an accuracy of 77.46%. This is not too surprising since wav2vec2 features are designed to contain phonetic information and, hence, are probably able to implicitly represent some information about word-usage, as well as pronunciation patterns. Further, wav2vec2 embeddings have a very distinct pattern over non-speech regions. Hence, our downstream model could potentially be learning patterns of usage of pauses, which are likely to be useful for differentiating AD from control subjects.

5.2. Statistical Significance Study

Given the relatively small number of samples available both in the development and the evaluation sets, we conducted a bootstrapping analysis on the development scores to determine confidence intervals for each of the systems submitted to the challenge. We sampled with replacement the 166 development scores obtained with CV to get 5000 new sets of scores, each with 166 samples. For each of these bootstrap sets, we computed the accuracy. The purple bars in Figure 2 show the 5% and 95% percentiles of the resulting set of accuracy values. We can see that the intervals are wide: all systems overlap with the others making it impossible to conclude whether there is, in-

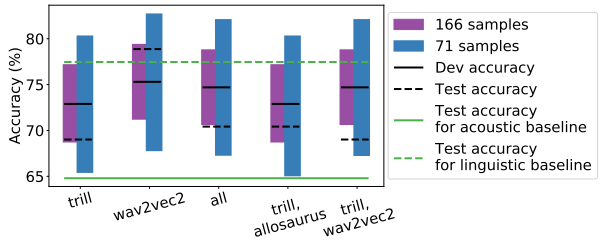

Figure 2: *Confidence intervals from bootstrapping experiment for the 5 best-performing models on the development set. The purple bars show the confidence intervals with bootstrap sets of size 166, the same as the original set, while the blue bars show the intervals with sets of size 71, the size of the test set. Black lines show the accuracy for each model on the development and test sets. Further, the green lines correspond to the test accuracy for the two baseline systems in [6].*

deed, a significant difference between them.

Further, since the test set is smaller than the development set, containing 71 subjects instead of 166, we repeated the bootstrap analysis on the development scores, but this time selecting only 71 samples per bootstrap set. The resulting confidence intervals are, of course, wider, and reflect the variability we could expect when testing these systems on a dataset of that size. Notably, the actual test results (shown in dashed black lines in Figure 2) fall within the estimated blue intervals suggesting that the test data is well represented by the training data.

Finally, the green lines in Figure 2 show the baseline results provided by the organizers in [6]. We can see that results for all our systems are significantly better than the acoustic baseline results. The wav2vec2 results are also better than the linguistic baseline results, though not by a significant margin.

5.3. Effect of the segment size

Our downstream model takes relatively short segments as input, extracted from the original audio with some overlap. The final score for each audio file is then given by an average of the segment-level scores. In this section, we study the effect of the segment size. Figure 3 shows the accuracy results at audio level (i.e., one sample per subject, as in all other results in this paper) and at segment level, using varying segment sizes. For this figure, segments are shifted by 2 seconds instead of 4, as in previous results, so that no speech is lost when using 2-second segments. Note that, since the shift is fixed, the number of segments for a certain audio file is approximately the same for all segment sizes. Further, to improve the stability of the results, we run each model with 3 different seeds to determine the cross-validation folds. The results shown in the plot correspond to the average accuracy over those 3 runs.

Figure 3 shows an interesting trend. As intuition would suggest, segment-level results improve as the segment size increases, since more information is available to make the classification decision. On the other hand, when averaging the scores from all segments in an audio file, the optimal segment size is around 5 seconds; longer segments degrade performance. We hypothesize that this is because, in longer segments, the effect of some short-term phenomena that might be a strong indicator for classification may be *washed out*. On the other hand, when using shorter segments, the model may be able to focus on these local phenomena and produce more discriminative scores for the segments that contain them. Further analysis is necessary to prove or disprove this hypothesis. If proven true, this may

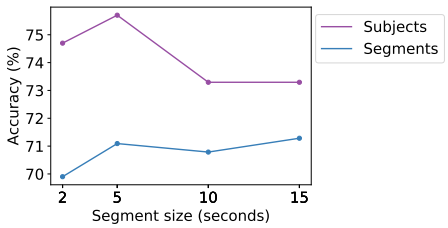

Figure 3: *Average accuracy over three seeds for the wav2vec2-only model, varying the segment size, for segment- and subject-level scores.*

suggest an interesting research direction: the development of hierarchical models that would take the output of our segment-level scores and effectively combine their outputs to emphasize the more informative scores, for example, using attention mechanisms.

5.4. Analysis of the influence of the experimenter speech

As mentioned above, our systems obtain the final scores for each subject as an average over all segments in an audio file. A portion of these segments contain at least some speech from the experimenter. To explore whether these segments had a negative impact on our results, we performed the following experiment. We computed the average accuracy over three seeds using only the audio files for which the manual segmentation had no obvious issues (139 out of the 166 files). Using the wav2vec2-only model from the previous section, with segment size of 5-seconds and shifts of 2-seconds, this gave an accuracy of 76.97%. We then discarded the scores from all the segments with any speech from the experimenter (33% of them) and recomputed the average score for each audio file. The accuracy for these new average scores did not significantly change. Given this result, we can conclude that the effect of the experimenter's speech is not harmful once the model is fixed. On the other hand, it is possible that a model trained without segments including speech from the experimenter would work better. This analysis is left for future work.

6. Conclusions

We presented our work on Alzheimer's Disease recognition, using the data from this year's ADreSSo challenge. Our approach uses speech-based embeddings from three different pre-trained models recently released to the public: trill, allosaurus and Wav2vec 2.0. We also include eGeMAPS, a set of features traditionally used for emotion recognition and related tasks. The features are modeled with a simple neural network that takes short segments of audio and generates scores which are then averaged to obtain the final score for each audio file. Word transcriptions are not used by our system. We show that the best results are obtained using Wav2vec 2.0 features, though all features perform similarly, considering the wide confidence intervals. Our results significantly outperform the acoustic baseline provided by the organizers, reaching an accuracy of 78.87% on the challenge's test set.

7. Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation for the donation of a Titan Xp GPU.

8. References

[1] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, 2018.

[2] A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[3] G. W. Ross, J. L. Cummings, and D. F. Benson, "Speech and language alterations in dementia syndromes: Characteristics and treatment," *Aphasiology*, vol. 4, no. 4, pp. 339–352, 1990.

[4] K. Lopez-de Ipina, U. Martinez-de Lizarduy, P. M. Calvo, J. Mekyska, B. Beitia, N. Barroso, A. Estanga, M. Tainta, and M. Ecay-Torres, "Advances on automatic speech analysis for early detection of alzheimer disease: a non-linear multi-task approach," *Current Alzheimer Research*, vol. 15, no. 2, pp. 139–148, 2018.

[5] K. López-de Ipiña, J.-B. Alonso, C. M. Travieso, J. Solé-Casals, H. Egiraun, M. Faundez-Zanuy, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage, and U. M. d. Lizardui, "On the selection of non-invasive methods based on speech analysis oriented to automatic alzheimer disease diagnosis," *Sensors*, vol. 13, no. 5, pp. 6730–6745, 2013. [Online]. Available: https://www.mdpi.com/1424-8220/13/5/6730

[6] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo Challenge," in *Submitted to INTERSPEECH 2021*, 2021. [Online]. Available: https://edin.ac/31eWsjp

[7] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for alzheimer's dementia through spontaneous speech," *INTERSPEECH (to appear)*, pp. 1–5, 2020.

[8] M. Macary, M. Tahon, Y. Estève, and A. Rousseau, "On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 373–380.

[9] Y. Zhao, D. Yin, C. Luo, Z. Zhao, C. Tang, W. Zeng, and Z.-J. Zha, "General-Purpose Speech Representation Learning through a Self-Supervised Multi-Granularity Framework," *arXiv:2102.01930*, 2021.

[10] M. Li, B. Yang, J. Levy, A. Stolcke, V. Rozgic, S. Matsoukas, C. Papayiannis, D. Bone, and C. Wang, "Contrastive Unsupervised Learning for Speech Emotion Recognition," *arXiv:2102.06357*, 2021.

[11] R. Zhang, H. Wu, W. Li, D. Jiang, W. Zou, and X. Li, "Transformer based unsupervised pre-training for acoustic representation learning," *arXiv:2007.14602*, 2021.

[12] H. Nguyen, F. Bougares, N. Tomashenko, Y. Estève, and L. Besacier, "Investigating Self-Supervised Pre-Training for End-to-End Speech Translation," in *Proc. Interspeech 2020*, 2020, pp. 1466–1470. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1835

[13] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[14] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. d. C. Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards learning a universal non-semantic representation of speech," *arXiv preprint arXiv:2002.12764*, 2020.

[15] F. Boller and J. Becker, "Dementiabank database guide," *University of Pittsburgh*, 2005.

[16] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black *et al.*, "Universal phone recognition with a multilingual allophone system," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8249–8253.

[17] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.