Impact of Acoustic Noise on Alzheimer’s Disease Detection from Speech: Should You Let Baby Cry?

Jekaterina Novikova

Winterlight Labs, Toronto, Canada
jekaterina@winterlightlabs.com

Abstract

Research related to automatically detecting Alzheimer’s disease (AD) is important, given the high prevalence of AD and the high cost of traditional methods. Since AD significantly affects the acoustics of spontaneous speech, speech processing and machine learning (ML) provide promising techniques for reliably detecting AD. However, speech audio may be affected by different types of background noise and it is important to understand how the noise influences the accuracy of ML models detecting AD from speech. In this paper, we study the effect of fifteen types of noise from five different categories on the performance of four ML models trained with three types of acoustic representations. We perform a thorough analysis showing how ML models and acoustic features are affected by different types of acoustic noise. We show that acoustic noise is not necessarily harmful – certain types of noise are beneficial for AD detection models and help increasing accuracy by up to 4.8%. We provide recommendations on how to utilize acoustic noise in order to achieve the best performance results with the ML models deployed in real-world applications.

Index Terms: Alzheimer’s disease, dementia detection, speech biomarker, acoustic noise, robustness to noise

1. Introduction

Alzheimer’s disease (AD) is a progressive neurodegenerative disease that affects over 40 million people worldwide [1]. Current forms of diagnosis are both time consuming and expensive [2], which might explain why almost half of those living with AD do not receive a timely diagnosis [3]. Studies have shown that ML methods can be applied to distinguish between speech from healthy and AD participants [4][5][6][7]. Currently, speech recording for AD-related research typically takes place in a quiet room with a guiding clinician. Given that smartphone technology is rapidly advancing, speech assessments using ML models trained on recordings obtained by smartphones offer a potentially simple-to-administer and inexpensive solution, scalable to the entire population, that can be performed anywhere, including the patient’s home [8][9][10]. However, the problem of model robustness to acoustic noise becomes increasingly important when deploying ML models in real-world applications [11].

Current popular approaches to dealing with acoustic noise in AD detection models involve: 1) eliminating noise using various audio pre-processing techniques [12], 2) selecting features that are resilient to ASR error/noise [13], 3) minimizing the effects of noise with multimodal fusion of features [14]. All these approaches share a common assumption of acoustic noise being definitely harmful for ML models detecting AD from speech. However, in other ML research areas, such as computer vision or NLP, adding a certain level of natural and artificial noise to data is considered a valid and advantageous practice that helps achieving better performance in tasks like image recognition [15][16], text generation [17] and relation classification [18], among others. The recent study in text-based AD classification shows that small levels of noise do not negatively affect performance of BERT-based models [19].

Motivated by the previous work, in this paper we study the effect of acoustic noise on performance of the ML models trained to detect AD from speech. The contributions of this paper are: 1) we analyze the effect of acoustic noise on the values of acoustic features extracted from speech; 2) we perform a thorough study on the effect of acoustic noise on AD classification performance across ML models, extracted acoustic features and noise categories; 3) we provide recommendations to ML researchers and practitioners on how to utilize acoustic noise in order to achieve the best performance results.

2. Methodology

2.1. Dataset

We use the ADRessO Challenge dataset [20], which consists of 166 training speech samples from non-AD (N=83) and AD (N=83) English-speaking participants. Speech is elicited from participants through the Cookie Theft picture from the Boston Diagnostic Aphasia exam. In contrast to the other datasets for AD detection such as DementiaBank’s English Pitt Corpus, the ADRessO challenge dataset is well balanced in terms of age and gender. In addition, the pre-processing step of ADRessO recordings were acoustically enhanced with stationary noise removal and audio volume normalisation applied across all speech segments to control for variation caused by recording conditions such as microphone placement. Such enhancements make this dataset a great source of the noise-clean audio, which is important for our experiments.

2.2. Feature Extraction

1) ConvFeat: We extract 182 acoustic features from the unsegmented speech audio files. Those include several statistics such as mean, std, median, etc. of mel-frequency cepstral coefficients (MFCCs), onset detection, rhythm, spectral and power features, following prior works in AD classification [20][21].

2) eGeMAPSv02: The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features are a selected standardized set of statistical features that characterize affective physiological changes in voice production. We extracted these features for the entire recording, as this feature set was shown to be usable for atypical speech [22] and was successfully used for classifying AD from speech [23][24].

3) wav2vec: in order to create audio representations using this approach, we make use of the huggingface[1] implementation of the wav2vec 2.0 [25] base model wav2vec2-base-960h. This base model is pretrained and fine-tuned on 960 hours of LibriSpeech on 16kHz sampled speech audio. Closely following [26] that used these representations for AD classification, we...
we extracted the last hidden state of the wav2vec2 model and used it as an embedded representation of audio.

2.3. Adding Noise

We used the audiomentations library to add two types of audio noise that are common when recording audio with smart devices - 1) background noise, and 2) short noise. We used a reduced version of the ESC-50 dataset to generate noisy audio, where we select three classes of noise from all the five presented major categories:

1. Animal sounds: dog, cat, crow
2. Natural sounds: rain, wind, chirping birds
3. Human sounds: crying baby, sneezing, coughing
4. Domestic / interior sounds: clock ticking, washing machine, vacuum cleaner
5. Urban / exterior noises: train, car horn, siren

2.4. Experiments

We first analyze how significantly addition of noise changes the values of acoustic features ConvFEAT and eGE MAPSv02 . We calculate the ratio of features that are impacted significantly by noise, with the Mann–Whitney U test used to estimate significance of difference.

Next, we experiment with the effect of noise addition to the performance of AD classification models. Following multiple previous studies on AD classification from speech, we use a set of linear and non-linear ML models: Logistic regression (LR), Support Vector Machines (SVM), Neural Network (NN), and Decision Tree (DT).

We use 10-fold cross-validation approach to evaluate the performance of classifiers, with the F1 score being the main classification performance evaluation metric.

3. Results and Discussion

3.1. Effect of Noise on the Values of Acoustic Features

The results in Table show that different types of noise have very different impact on the acoustic features, where sneezing sound introduced several times within recordings for short periods only affects 15% of ConvFEAT , while continuous background sound of rain significantly changes more than 90% of these features. Unsurprisingly, background noise affects recordings much stronger than short noise. Notably, conventional acoustic features are on average more vulnerable than eGE MAPSv02 to both short noise (12.5% higher ratio of significantly affected features) and background noise (19.8% higher ratio), with the categories of natural sounds, domestic/interior and urban/exterior bringing the strongest difference between the ConvFEAT and eGE MAPSv02.

Both ConvFEAT and eGE MAPSv02 are quite robust to the human non-speech noise, especially the sound of sneezing. Out of all the noise types analyzed in this work, sneezing is the only one that only affects up to 50% of acoustic features, both in a format of short and background noise. Natural sounds, such as rain, wind or chirping birds, affect the acoustic features the strongest.

The above results suggest that noise strongly disturbs the quality of audio samples, as represented by both ConvFEAT and eGE MAPSv02. Next, we analyze whether such a disturbance is beneficial or harmful when it comes to AD detection.
from disturbed speech.

3.2. Effect of Noise on Performance of AD Classification

Four types of ML models (SVM, NN, LR and DT) were trained on noisy and original audio recordings represented using CONVFEAT, eGEMAPSv02 and WAV2VEC. Each set of features was extracted from both original audio recordings and the recordings with added 20 subcategories of noise. Each ML model was evaluated with the F1 score on three different random seeds. As such, it is possible to analyse the mean classification performance level per feature type, where performance is averaged across all the seeds, for each model, noise subcategory and feature type.

**Analysis Per Feature Type:** The best mean F1 score represents the model that performs the best on average (across three random seeds) for some specific noise subcategory. Based on the best mean F1 score, the WAV2VEC-based model outperforms substantially the eGEMAPSv02-based model, while the CONVFEAT-based model achieves the lowest best mean level of performance (see Figure 1). Interestingly in all three cases, the best mean level of performance is achieved by the models trained on the original audio without noise addition.

The best maximum F1 score represents the best possibly achievable performance across all the seeds, i.e. the model that performs the best on a single best seed. The difference between the best mean level and the best maximum level shows the potential of the models to achieve higher level of performance. Figure 1 shows that such a potential is the strongest for the CONVFEAT-based models (+6.3%), and there is not that much room for improvement for the WAV2VEC-based models (+2.3%). However, given the strong starting point, i.e. the strong best mean level, the absolute best maximum level of performance is achieved by the WAV2VEC-based model. Interestingly, this best maximum level is achieved by the model trained on the noisy data, not the original audio. The same is true for the second-best maximum performance, i.e. of the CONVFEAT-based model.

**Analysis Per Model Type:** The best mean F1 score is achieved by the LR model, while SVM and NN both share the lowest level of the best mean performance (Figure 2). The growth potential of both linear models (LR and SVM) is weaker than that of the non-linear models (DT and NN), with the NN model being the potential of the models to achieve higher level of performance.

| Noise category | Subcategory | Features | Count | Mean F1 w/ noise | Max F1 w/ noise | Best mean F1 w/ noise | Best max F1 w/ noise |
|----------------|-------------|----------|-------|-----------------|----------------|----------------------|---------------------|
| **Animals**    | crow        | CONVFEAT | 24    | 0.6273          | 0.7018*        | 0.6557               | 0.6557              |
|                | dog         | eGEMAPSv02 | 24   | 0.6284          | 0.6907*        | 0.6557               | 0.7200              |
|                |             | WAV2VEC  | 24    | 0.6222          | 0.7006         | 0.7111               | 0.7200              |
| **Natural**    | chirping    | CONVFEAT | 24    | 0.6275          | 0.6704*        | 0.6557               | 0.6557              |
|                | birds       | eGEMAPSv02 | 24   | 0.6443          | 0.6995*        | 0.6557               | 0.6557              |
|                | rain        | WAV2VEC  | 24    | 0.6257          | 0.7114         | 0.6557               | 0.6557              |
|                | wind        | CONVFEAT | 24    | 0.6506          | 0.7135*        | 0.6557               | 0.6557              |
|                |             | eGEMAPSv02 | 24  | 0.6506          | 0.7135*        | 0.6557               | 0.6557              |
| **Human**      | coughing    | CONVFEAT | 24    | 0.6182          | 0.6293*        | 0.6557               | 0.6557              |
|                | crying baby | eGEMAPSv02 | 24 | 0.6387          | 0.7120*        | 0.6557               | 0.6557              |
|                |             | WAV2VEC  | 24    | 0.6275          | 0.6966         | 0.7111               | 0.7200              |
|                | sneezing    | CONVFEAT | 24    | 0.6506          | 0.6410         | 0.6557               | 0.6557              |
|                |             | eGEMAPSv02 | 24  | 0.6506          | 0.6410         | 0.6557               | 0.6557              |
| **Domestic/   | clock       | CONVFEAT | 24    | 0.6013          | 0.6557         | 0.6557               | 0.6557              |
| interior**     | vacuum      | eGEMAPSv02 | 24  | 0.6284          | 0.6990*        | 0.6557               | 0.6557              |
|                | cleaner     | WAV2VEC  | 24    | 0.5725          | 0.7100*        | 0.6557               | 0.7200              |
|                | washing     | CONVFEAT | 24    | 0.5937          | 0.6561         | 0.6557               | 0.6557              |
|                | machine     | eGEMAPSv02 | 24 | 0.6254          | 0.6919*        | 0.6557               | 0.6557              |
|                |             | WAV2VEC  | 24    | 0.6391          | 0.6900*        | 0.6557               | 0.6557              |
|                | car horn    | CONVFEAT | 24    | 0.6406          | 0.6800*        | 0.6557               | 0.6557              |
|                |             | eGEMAPSv02 | 24  | 0.6406          | 0.6800*        | 0.6557               | 0.6557              |
|                |             | WAV2VEC  | 24    | 0.6194          | 0.6816         | 0.7111               | 0.7200              |
| **Urban/      | siren       | CONVFEAT | 24    | 0.6324          | 0.7111*        | 0.6557               | 0.6557              |
| exterior**     |             | eGEMAPSv02 | 24 | 0.5868          | 0.6832         | 0.6557               | 0.6557              |
|                |             | WAV2VEC  | 24    | 0.6069          | 0.6631         | 0.7111               | 0.7200              |
|                | train       | CONVFEAT | 24    | 0.6282          | 0.6818*        | 0.6557               | 0.6557              |
|                |             | eGEMAPSv02 | 24  | 0.6328          | 0.6866*        | 0.6557               | 0.6557              |

Table 2: Change in AD classification performance when models are trained on the noisy audio recordings, by noise category, subcategory and feature type. **Bold** denotes best performance per noise subcategory+features, **bold italic** denotes best overall performance, **green background** denotes noise subcategory that has consistently highest performance when models are trained on the noisy recordings. * indicates significant difference of \( p < 0.05 \) on McNemar’s test.
showing the strongest potential across all model types. Once again, the best mean level of all the models is achieved when training the models on the original noise-free recordings, while the best maximum level is always achieved by training the models on the noisy audio recordings.

To overview, the results strongly suggest that noise has a beneficial effect on performance of AD classifiers, both linear and non-linear and utilizing different sets of features. However, all these performance results are aggregated across different categories and subcategories of noise. Next, we investigate in more detail how each specific noise category affects AD classification model performance.

### Analysis Per Noise Type

The results of classification experiments with models trained on the noise-free and noisy audio show that best average classification performance is achieved when models are trained on clean noise-free audio recording (Best mean F1 w/o noise and Mean F1 w/o noise columns in Table 2). However, the maximum performance is consistently higher for the models trained on the noisy audio (columns Max F1 w/o noise vs Best max F1 w/o noise in Table 2).

Out of all the noise categories, domestic/interior sounds seem to be the least beneficial for the AD classification models - none of the noise subcategories helps consistently improving classification performance. In the other categories, such as animal sounds, natural sounds, and urban/interior noise, at least one noise subcategory consistently achieves substantially higher performance with the models trained on the noisy recordings, with all the tested audio features. The human noise is the most beneficial noise category for getting high AD classification results: 1) the overall best classification performance is achieved by the model trained on the noisy recording of this category (model trained on wav2vec embeddings of the audio with the crying baby noise), 2) two out of three noise subcategories (coughing and crying baby) consistently achieve higher performance level across all the audio features. The best overall performance motivates us to investigate in more detail the classification performance of the models trained on the audio with the crying baby noise.

### Analysis of the Crying Baby Noise

All the ConvFeats - based models trained on the audio recordings with the sounds of crying baby present as short noise, perform better than those same models trained on the original noise-free audio recordings. Same is true for the majority of the WAV2Vec -based models, with WAV2Vec -based NN achieving the overall best performance.

When it comes to the sound of crying baby to be introduced as a continuous background noise, the overall performance level of WAV2Vec and ConvFeats -based models decreases substantially. WAV2Vec -based models are not able anymore to outperform any of noise-free models, and only linear ConvFeats -based models are still able to outperform their noise-free analogues. The eGEMAPSv02 -based SVM model is able to achieve its best performance with this type of noise.

### 3.3. Recommendations

Based on the results of our analysis, we outline a set of recommendations for the ML researchers and practitioners interested in deploying AD classification models in real world.

First, if acoustic features are extracted using conventional and not deep learning-based features, such as ConvFeats or eGEMAPSv02, it is important to use the noise removal speech pre-processing techniques to normalize the audio dataset that is used for training ML models. As explained in Section 5, even short segments of unwanted noise, such as accidental siren, craw caw or a short vacuum cleaner sound, may significantly change more than 50% of acoustic features. Having the training dataset where otherwise similar datapoints are represented by significantly different acoustic features, introduces many unnecessary challenges in model development.

Second, it is important to make sure the deployed models are not be used in certain types of real world environments where certain noises are common. As explained in Section 5.2, domestic noise, such as washing machine or vacuum cleaner, may decrease classification performance. As such, it is important to recommend the real world users of the AD classification model to avoid this type of noise when recording audio in order to expect better accuracy of the model. Other noises, such as baby cry, cough or dog bark, are not harmful and there is no need to avoid them. This is also important to know because these types of noise are much more difficult to securely avoid in real world scenarios than sounds of a vacuum cleaner or washer.

Third, model developers should expect different effects of noise on the AD classification performance depending on the type of audio representation and model used. Deep features, such as WAV2Vec, are affected less strongly by the presence of noise comparing to more conventional acoustic features, such as ConvFeats and eGEMAPSv02, although models utilizing all three types of features may benefit from certain noises in audio. More simplistic linear models, such as SVM and LR, may be impacted positively but not very strongly (up to 2.5%) by the presence of appropriate noise in the recordings. The more complex non-linear models, such as DT and NN, may experience twice stronger positive effect (+4.8%) due to appropriate noise.

### 4. Conclusions

In this paper, we study the effect of acoustic noise on AD classification from speech. We show that certain types of noise are beneficial for AD classification performance. Further research is necessary to investigate the effect of more types of acoustic noise common in real world scenarios.
5. References

[1] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou, “World Alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future,” 2016.

[2] G. Prabhakaran, R. Bakschi et al., “Analysis of Structure and Cost in a Longitudinal Study of Alzheimer’s Disease,” Journal of Health Care Finance, 2018.

[3] E. A. Jammeh, B. C. Camille, W. P. Stephen, J. Escudero, A. Anastasiou, P. Zhao, T. Chenore, J. zajicek, and E. Ifeachor, “Machine-learning based identification of undiagnosed dementia in primary care: a feasibility study,” BJGP Open, p. bigpopen18X101589, 2018.

[4] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, “Linguistic features identify Alzheimer’s disease in narrative speech,” Journal of Alzheimer’s Disease, vol. 49, no. 2, pp. 407–422, 2016.

[5] A. Balagopalan, J. Novikova, F. Rudzicz, and M. Ghassemi, “The Effect of Heterogeneous Data for Alzheimer’s Disease Detection from Speech,” in NeurIPS 2018 Workshop Machine Learning for Health (ML4H), 2018.

[6] Z. Zhu, J. Novikova, and F. Rudzicz, “Detecting cognitive impairments by agreeing on interpretations of linguistic features,” in NAACL 2019, 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics, no. arXiv preprint arXiv:1808.06570, 2019.

[7] B. Eyre, A. Balagopalan, and J. Novikova, “Fantastic features and where to find them: detecting cognitive impairment with a subsequence classification guided approach,” W-NUT at EMNLP 2020, 2020.

[8] L. C. Kourtis, O. B. Regele, J. M. Wright, and G. B. Jones, “Digital biomarkers for Alzheimer’s disease: the mobile/ wearable devices opportunity,” NPJ digital medicine, vol. 2, no. 1, pp. 1–9, 2019.

[9] M. Mc Carthy and P. Schueler, “Can Digital Technology Advance the Development of Treatments for Alzheimer’s Disease?” pp. 217–220, 2019.

[10] E. Fristed, C. Skirrow, M. Moszaros, R. Lenain, U. Meepegama, S. Cappa, D. Aarsland, and J. Weston, “Evaluation of a speech-based AI system for early detection of Alzheimer’s disease remotely via smartphones,” medRxiv, 2021.

[11] J. Robin, J. E. Harrison, L. D. Kaufman, F. Rudzicz, W. Simpson, and M. Yancheva, “Evaluation of speech-based digital biomarkers: review and recommendations,” Digital Biomarkers, vol. 4, no. 3, pp. 99–108, 2020.

[12] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, “Detecting cognitive decline using speech only: The ADReSsIo Challenge,” arXiv preprint arXiv:2104.06156, 2021.

[13] L. Zhou, K. C. Fraser, and F. Rudzicz, “Speech Recognition in Alzheimer’s Disease and in its Assessment,” in INTERSPEECH 2016, 2016, pp. 1948–1952.

[14] M. Rohanian, J. Hough, and M. Purver, “Alzheimer’s Dementia Recognition Using Acoustic, Lexical, Disfluency and Speech Pause Features Robust to Noisy Inputs,” arXiv preprint arXiv:2106.15684, 2021.

[15] M. Kozierski and B. Cyganek, “Image recognition with deep neural networks in presence of noise–dealing with and taking advantage of distortions,” Integrated Computer-Aided Engineering, vol. 24, no. 4, pp. 337–349, 2017.

[16] C. R. Steffens, L. V. Messias, P. L. J. Drews, and S. S. da Costa Botelho, “Can exposure, noise and compression affect image recognition? an assessment of the impacts on state-of-the-art convnets,” in 2019 Latin American Robotics Symposium (LARS), 2019 Brazilian Symposium on Robotics (SBR) and 2019 Workshop on Robotics in Education (WRE). IEEE, 2019, pp. 61–66.

[17] S. Y. Feng, V. Gangul, D. Kang, T. Mitamura, and E. Hovy, “GenAug: Data Augmentation for Finetuning Text Generators,” in Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, 2020, pp. 29–42.

[18] P. K. B. Giridhara, C. Mishra, R. K. M. Venkataramana, S. S. Bukhari, and A. Denga, “A Study of Various Text Augmentation Techniques for Relation Classification in Free Text,” ICPRAM, vol. 3, p. 5, 2019.

[19] J. Novikova, “Robustness and Sensitivity of BERT Models Predicting Alzheimer’s Disease from Text,” in Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), 2021, pp. 334–339.

[20] Z. Zhu, J. Novikova, and F. Rudzicz, “Semi-supervised classification by reaching consensus among modalities,” in NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language IARAL, 2018.

[21] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, “To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer’s Disease Detection,” INTERSPEECH 2020, 2020.

[22] W. Xue, C. Cucchiarini, R. van Hout, and H. Strik, “Acoustic correlates of speech intelligibility: the usability of the egemaps feature set for atypical speech,” in Workshop on Speech and Language Technology in Education, 2019.

[23] L. Gauler, L. Pepino, L. Ferrer, and P. Riera, “Alzheimer Disease Recognition Using Speech-Based Embeddings From Pre- Trained Models,” in Proc. INTERSPEECH 2021, 2021, pp. 3795–3799.

[24] R. Pappagari, J. Cho, S. Joshi, L. Moro-Velázquez, P. Zelasko, J. Villalba, and N. Dehak, “Automatic Detection and Assessment of Alzheimer Disease Using Speech and Language Technologies in Low-Resource Scenarios,” in Proc. INTERSPEECH 2021, 2021, pp. 3825–3829.

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “w2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” Advances in Neural Information Processing Systems, vol. 33, 2020.

[26] A. Balagopalan and J. Novikova, “Comparing Acoustic-based Approaches for Alzheimer’s Disease Detection,” INTERSPEECH 2021, 2021.

[27] S. Vhaduri, T. Van Kessel, B. Ko, D. Wood, S. Wang, and T. Brunswijler, “Nocturnal cough and snore detection in noisy environments using smartphone-microphones,” in 2019 IEEE international conference on healthcare informatics (ICHI). IEEE, 2019, pp. 1–7.

[28] S. V. Dibbo, Y. Kim, and S. Vhaduri, “Effect of Noise on Generic Cough Models,” in 2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN). IEEE, 2021, pp. 1–4.

[29] K. J. Piczak, “ESC: Dataset for environmental sound classification,” in Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 1015–1018.

[30] A. Balagopalan, B. Eyre, J. Robin, F. Rudzicz, and J. Novikova, “Comparing pre-trained and feature-based models for prediction of Alzheimer’s disease based on speech,” Frontiers in aging neuroscience, vol. 13, p. 189, 2021.