Impact of Environmental Noise on Alzheimer's Disease Detection from Speech: Should You Let a 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 environmental 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.

1 Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disease that affects over 40 million people worldwide (Prince et al., 2016). Current forms of diagnosis are both time consuming and expensive (Prabhakaran et al., 2018), which might explain why almost half of those living with AD do not receive a timely diagnosis (Jammeh et al., 2018). Studies have shown that ML methods can be applied to distinguish between speech from healthy and AD participants (Fraser et al., 2016; Balagopalan et al., 2018; Zhu et al., 2019; Eyre et al., 2020). 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 (Kourtis et al., 2019; Mc Carthy and Schueler, 2019; Fristed et al., 2021). However, the problem of model robustness to acoustic noise becomes increasingly important when deploying ML models in real world (Robin et al., 2020).

Current popular approaches to dealing with acoustic noise in AD detection models involve: 1) eliminating noise using various audio preprocessing techniques (Luz et al., 2021), 2) selecting features that are resilient to ASR error/noise (Zhou et al., 2016), 3) minimizing the effects of noise with multimodal fusion of features (Rohanian et al., 2021). 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 (Koziarski and Cyganek, 2017; Steffens et al., 2019), text generation (Feng et al., 2020) and relation classification (Giridhara et al., 2019), among others. The recent studies in AD classification from transcribed speech show that small levels of linguistic noise do not negatively affect performance of BERT-based models (Novikova, 2021), although there is a difference in predictive power between lexical and syntactic features, when it comes to AD detection from speech (Novikova et al., 2019).

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 environmental 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 recommendation to ML researchers and practitioners on how to utilize acoustic noise in order to achieve the best performance results.

2 Related Work

2.1 Environmental Noise and Speech Quality

Multiple previous studies attempted to investigate the influence of the environment background noise on speech quality. For example, Naderi et al. (2018) conducted a study in which participants rated the quality of speech files first in the laboratory and then in noisy speech collection settings, such as cafeteria and living room. They found that the presence of a “cafeteria" or a “crossroad" background noise would decrease the correlation to speech quality ratings.

Furthermore, multiple studies have addressed the issue of speech intelligibility under certain background noise conditions. To name some, Meyer et al. (2013) tackled the problem of speech recognition accuracy in ecologically valid natural background noise scenarios and showed the relation between the levels of noise and confusion of vowels, lexical identification and perceptual consonant confusion.

Jiménez et al. (2020) investigated the influence of environmental background noise on speech quality, where the quality of speech files was assessed under the influence of two types of background noise at different levels, i.e., street noises and tv-show. The authors found there was a certain threshold of the environment background noise level that impacted the quality of speech, and different types of noise had a different effect on the quality.

Motivated by the previous studies, in this work we analyze fifteen different types of environmental background noise in order to figure out differences in their impact. We also compare the impact of short and continuous noise to follow up on the findings of the impact threshold.

2.2 Alzheimer’s Disease Detection in Noisy Settings

Given the number of people with AD is growing and the population is aging fast in many countries (Brookmeyer et al., 2018), it becomes more and more important to have tools to help identify the presence of cognitive impairment relating to AD that can be deployed frequently, and at scale. This need will only increase as effective interventions are developed, requiring the ability to identify patients early in order to facilitate prevention or treatment of disease (Vellas and Aisen, 2021). Most of the current AD screening tools represent a significant burden, requiring invasive procedures, or intensive and costly clinical testing. However, recent shifts toward telemedicine and increased digital literacy of the aging population provide an opportunity for using digital health tools that are ideally poised to meet the needs for novel solutions. Recently, automated tools have been developed that assess speech and can be used on a smartphone or tablet, from one’s home (Robin et al., 2021). Digital assessments that can be accessed on a smartphone or tablet, completed from home and periodically repeated, would vastly improve the accessibility of AD screening compared to current clinical standards that require clinical visits, extensive neuropsychological testing or invasive procedures.

The pervasiveness of high-quality microphones in smart devices makes the recording of speech samples straightforward, not requiring additional equipment or sensors. However, there is a lack of control over the participants performing digital assessments in home environment, and often not enough information is collected about their playback system and background environment. Participants might be exposed to different environmental conditions while executing specific tasks, and as such, their recorded speech quality may be disturbed with some background noise.

In the speech community, the active ongoing effort is focused on solving the problem of automated speech enhancement with the methods of noise suppression that are based on machine learning and deep learning (Zhang et al., 2022; Braun et al., 2021; Choi et al., 2018; Odelowo and Anderson, 2017, among many others). However, this problem is not considered to be solved, and the research community continues developing methods for effective noise elimination from audio recordings (Dubey et al., 2022).

These challenges motivate us asking a question whether noise suppression is absolutely necessary when it comes to the specific task of AD detection from speech. In this work, we perform a thorough study on the effect of acoustic background noise, standard for home environments, on AD classification performance across a range of ML models.

2.3 Speech Quality and Alzheimer's Disease Detection

Speech is a promising modality for digital assessments of cognitive abilities. Producing speech samples is a highly ecologically valid task that requires little instruction and at the same time is instrumental to daily functioning. Advances in signal processing and natural language processing have enabled objective analysis of speech samples for their acoustic properties, providing a window into monitoring motor and cognitive abilities. Most importantly, previous research has extensively shown that speech patterns are affected in AD, demonstrating the clinical relevance of speech for detecting cognitive impairment and dementia (Martínez-Nicolás et al., 2021; de la Fuente Garcia et al., 2020; Slegers et al., 2018).

Some of the features employed to describe acoustic characteristics of the voice applied to AD detection, include conventional acoustic features, such as fundamental frequency, jitter and shimmer, as well as pre-trained embeddings from deep neural models for audio representation, such as wav2vec (Balagopalan and Novikova, 2021). Quality of speech, which may be influenced by environmental noise, inevitably affects the values of these acoustic features extracted from speech and as a result may potentially influence the performance of ML models that use these features as internal representations of human speech.

However, in other research areas, such as computer vision or NLP, adding a certain level of natural or artificial noise to data is considered a valid and advantageous practice that helps achieving better performance in tasks like image recognition (Koziarski and Cyganek, 2017; Steffens et al., 2019), text generation (Feng et al., 2020) and relation classification (Giridhara et al., 2019), among others. Moreover, deep neural acoustic models, such as wav2vec, that are used to generate acoustic embeddings used in AD detection, are pre-trained on healthy speech. As such, it is possible that the subparts of the embeddings that are affected by environmental noise are not used for the task of AD detection directly and as a result, they do not influence the performance of such detection.

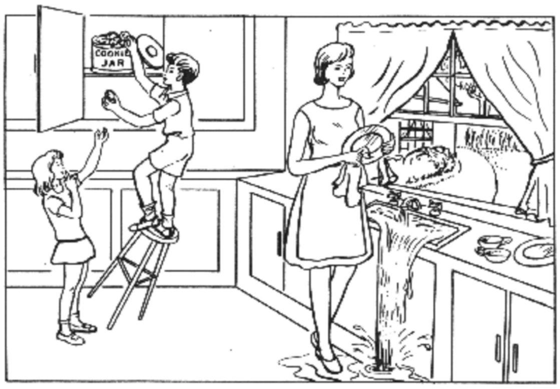

Figure 1: Cookie Theft picture used to collect speech for the ADReSS*o* dataset.

In this work, we make an attempt to understand how different types of environmental noise are impacting the values of different types of acoustic features extracted from speech, as well as how this affects performance of ML models relying on these features.

3 Methodology

3.1 Dataset

We use the ADReSS*o* Challenge dataset (Luz et al., 2021), 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 (Figure 1). In contrast to the other datasets for AD detection such as DementiaBank's English Pitt Corpus, the ADReSS*o* challenge dataset is well balanced in terms of age and gender. In addition, the pre-processing step of ADReSS*o* 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.

3.2 Feature Extraction

The following groups of features were extracted for the further use in the experiments:

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 (Fraser et al., 2016; Zhu et al., 2018; Balagopalan et al., 2020).

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 (Xue et al., 2019) and was successfully used for classifying AD from speech (Gauder et al., 2021; Pappagari et al., 2021).

3. wav2vec : In order to create audio representations using this approach, we make use of the huggingface[1] implementation of the wav2vec 2.0 (Baevski et al., 2020) 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 (Balagopalan and Novikova, 2021) that used these representations for AD classification, we extracted the last hidden state of the wav2vec2 model and used it as an embedded representation of audio.

3.3 Adding Noise

We used the audiomentations[2] library to add two types of audio noise that are common when recording audio with smart devices - 1) background noise, and 2) short noise (Vhaduri et al., 2019; Dibbo et al., 2021). We use a reduced version of the ESC-50 dataset (Piczak, 2015) 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** soundscapes: 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

3.4 Experiments

We first analyze how significantly addition of noise changes the values of acoustic features ConvFeat and eGeMAPSv02 . 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 (Balagopalan et al., 2020, 2021; Balagopalan and Novikova, 2021), 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.

4 Results and Discussion

4.1 Effect of Noise on the Values of Acoustic Features

The results in Table 1 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 10% 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 eGeMAPSv02 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 eGeMAPSv02 .

Both ConvFeat and eGeMAPSv02 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

[1] https://huggingface.co/models

[2] https://github.com/iver56/audiomentations

Noise category	Subcategory	Features	Ratio of sign diff features Short noise	 Background noise
Animals	cat	ConvFeat	32.42%	68.68%
		eGeMAPSv02	32.95%	50.00%
	crow	ConvFeat	55.49%	80.22%
		eGeMAPSv02	45.45%	59.09%
	dog	ConvFeat	23.08%	63.19%
		eGeMAPSv02	23.86%	50.00%
Natural	chirping birds	ConvFeat	69.23%	71.43%
		eGeMAPSv02	44.32%	54.55%
	rain	ConvFeat	67.58%	90.11%
		eGeMAPSv02	32.95%	69.32%
	wind	ConvFeat	48.35%	78.02%
		eGeMAPSv02	42.05%	60.23%
Human	coughing	ConvFeat	37.36%	52.20%
		eGeMAPSv02	27.27%	32.95%
	crying baby	ConvFeat	53.30%	68.68%
		eGeMAPSv02	40.91%	67.05%
	sneezing	ConvFeat	10.44%	41.21%
		eGeMAPSv02	27.27%	25.00%
Domestic/ interior	clock ticking	ConvFeat	48.35%	63.74%
		eGeMAPSv02	23.86%	30.68%
	vacuum cleaner	ConvFeat	63.19%	87.36%
		eGeMAPSv02	42.05%	60.23%
	washing machine	ConvFeat	51.10%	82.97%
		eGeMAPSv02	28.41%	65.91%
Urban/ exterior	car horn	ConvFeat	39.01%	81.32%
		eGeMAPSv02	27.27%	45.45%
	siren	ConvFeat	53.30%	74.73%
		eGeMAPSv02	42.05%	62.50%
	train	ConvFeat	56.04%	83.52%
		eGeMAPSv02	39.77%	57.95%

Table 1: Impact of noise addition on the value of **ConvFeat** and **eGeMAPSv02** . Ratio of sign. diff. features shows the percentage of all the features that is significantly ($p < 0.05$) different from the original values as a result of adding short noise and background noise to original audio samples. Lighter cell color indicates higher than 50% ratio, darker - higher than 80% ratio.

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 **eGeMAPSv02** . Next, we analyze whether such a disturbance is beneficial or harmful when it comes to AD detection from disturbed speech.

4.2 Effect of Noise on Performance of AD Classification

Four types of ML models (SVM, neural network / NN, logistic regression / LR and decision tree / 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

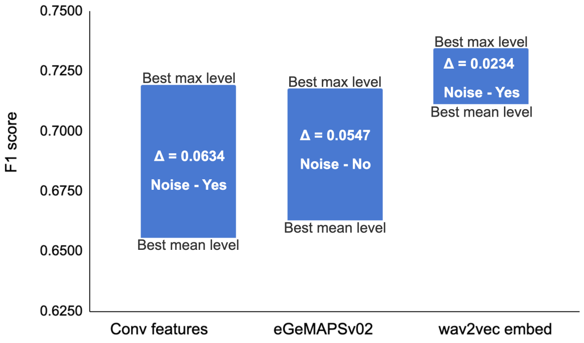

Figure 2: AD classification performance by feature type.

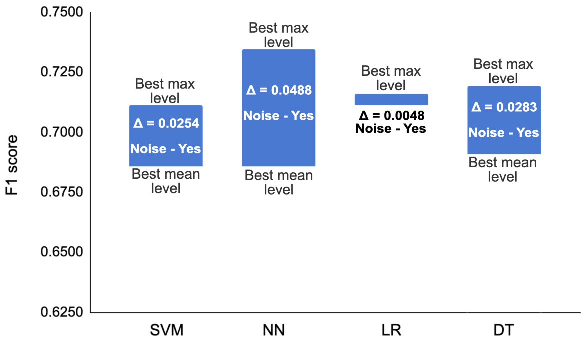

Figure 3: AD classification performance by model.

per feature type, where performance is averaged across all the seeds, for each model, noise subcategory and feature type.

4.2.1 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 2). 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 2 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.

4.2.2 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 3). 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 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.

4.2.3 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/ noise* columns in Table 2). However, the maximum performance is consistently higher for the models trained on the noisy audio (columns *Max F1 w/ 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/exterior 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

Noise category	Subcategory	Features	Count	Mean F1 w/ noise	Max F1 w/ noise	Best mean F1 w/o noise	Best max F1 w/o noise
Animals	cat	ConvFeat	24	0.6232	**0.6842***	0.6557	0.6557
		eGeMAPSv02	24	0.6284	**0.6907***	0.6557	0.6557
		wav2vec	24	0.6222	0.7006	0.7111	**0.7200**
	crow	ConvFeat	24	0.6217	**0.6591**	0.6557	0.6557
		eGeMAPSv02	24	0.6345	**0.6800***	0.6557	0.6557
		wav2vec	24	0.6400	0.6878	0.7111	**0.7200***
	dog	ConvFeat	24	0.6255	**0.6704**	0.6557	0.6557
		eGeMAPSv02	24	0.6318	**0.7014***	0.6557	0.6557
Natural	chirping birds	ConvFeat	24	0.6273	**0.7018***	0.6557	0.6557
		eGeMAPSv02	24	0.6443	**0.6995***	0.6557	0.6557
		wav2vec	24	0.6275	0.6966	0.7111	**0.7200***
	rain	ConvFeat	24	0.6229	**0.6882***	0.6557	0.6557
		eGeMAPSv02	24	0.6506	**0.7135***	0.6557	0.6557
	wind	ConvFeat	24	0.6156	0.6480	**0.6557**	**0.6557**
		eGeMAPSv02	24	0.6138	**0.7019***	0.6557	0.6557
Human	coughing	ConvFeat	24	0.6182	**0.6923***	0.6557	0.6557
		eGeMAPSv02	24	0.6387	**0.7120***	0.6557	0.6557
	crying baby	ConvFeat	24	0.6182	**0.6816***	0.6557	0.6557
		eGeMAPSv02	24	0.6472	**0.7079***	0.6557	0.6557
		wav2vec	24	0.6344	***0.7345****	0.7111	0.7200
	sneezing	ConvFeat	24	0.6071	0.6444	**0.6557**	**0.6557**
		eGeMAPSv02	24	0.6406	**0.6800***	0.6557	0.6557
Domestic/ interior	clock ticking	ConvFeat	24	0.6013	**0.6557**	**0.6557**	**0.6557**
		eGeMAPSv02	24	0.6284	**0.6990***	0.6557	0.6557
	vacuum cleaner	ConvFeat	24	0.5775	0.6292	**0.6557**	**0.6557**
		eGeMAPSv02	24	0.5937	**0.6561**	0.6557	0.6557
	washing machine	ConvFeat	24	0.6254	**0.6919***	0.6557	0.6557
		eGeMAPSv02	24	0.6391	**0.6990***	0.6557	0.6557
		wav2vec	24	0.6194	0.6816	0.7111	**0.7200***
Urban/ exterior	car horn	ConvFeat	24	0.6324	**0.7111***	0.6557	0.6557
		eGeMAPSv02	24	0.5868	**0.6832**	0.6557	0.6557
		wav2vec	24	0.6069	0.6631	0.7111	**0.7200***
	siren	ConvFeat	24	0.6274	**0.7191***	0.6557	0.6557
		eGeMAPSv02	24	0.6098	**0.6919***	0.6557	0.6557
		wav2vec	24	0.5961	0.6667	0.7111	**0.7200***
	train	ConvFeat	24	0.6112	**0.6818***	0.6557	0.6557
		eGeMAPSv02	24	0.6382	**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.

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.

4.2.4 Analysis of the *Crying Baby* Noise

All the ConvFeat -based models trained on the audio recordings with the sounds of *crying baby* present as short noise, perform better than those

F1	Model	Original noise-free audio			Short noise			Background noise		
		ConvFeat	eGeMAPSv02	wav2vec	ConvFeat	eGeMAPSv02	wav2vec	ConvFeat	eGeMAPSv02	wav2vec
Mean	SVM	0.6557	0.6480	0.6857	🟩 0.6816	🟩 0.6484	🟩 0.6885	0.6096	🟩 **0.7079**	0.6067
	LR	0.5698	0.6630	0.7111	🟩 0.6441	0.6413	🟩 **0.7159**	🟩 0.6243	0.6369	0.5914
	NN	0.6289	0.6541	0.6857	🟩 0.6355	🟩 0.6603	🟩 **0.7061**	0.6206	🟩 0.6595	0.5901
	DT	0.5882	0.5567	**0.6908**	🟩 0.6142	🟩 0.6004	0.6113	0.5154	🟩 0.6234	0.5653
Max	SVM	0.6557	0.6480	0.6857	🟩 0.6816	🟩 0.6484	🟩 0.6885	0.6096	🟩 **0.7079**	0.6067
	LR	0.5698	0.6630	0.7111	🟩 0.6441	0.6413	🟩 **0.7159**	🟩 0.6243	0.6369	0.5914
	NN	0.6519	0.7177	0.7200	🟩 0.6705	0.6832	🟩 **0.7345**	🟩 0.6484	0.6634	0.6034
	DT	0.6250	0.5795	**0.7045**	🟩 0.6292	🟩 0.6077	0.6328	0.5263	🟩 0.6292	0.5843

Table 3: Classification performance of the models trained on the noisy audio recordings with the sounds of crying baby. Mean F1 is averaged across three random seeds. **Bold** denotes the best performance per noise type (*short* and *background*), 🟩 green background denotes performance that is higher than the analogous one for the model+feature set trained on the original noise-free audio.

same models trained on the original noise-free audio recordings (see Table 3 for details). Same is true for the majority of 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 ConvFeat -based models decreases substantially. wav2vec -based models are not able anymore to outperform any of noise-free models, and only linear ConvFeat -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.

4.3 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 ConvFeat 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 4.1, 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 4.2, 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 ConvFeat 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.

5 Conclusions

In this paper, we study the effect of fifteen types of acoustic noise, standard in home environments, on AD classification from speech. We perform a thorough analysis showing how four ML models and three types of acoustic features are affected by different types of acoustic noise. We show that natural environmental noise is not necessarily harmful, with certain types of noise even being beneficial for AD classification performance and helping increase 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 in order to facilitate the use of scalable digital health tools for AD detection from speech. Further research is necessary to investigate the effect of more types of acoustic noise common in real world scenarios.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, 33.

Aparna Balagopalan, Benjamin Eyre, Jessica Robin, Frank Rudzicz, and Jekaterina Novikova. 2021. Comparing pre-trained and feature-based models for prediction of Alzheimer's disease based on speech. *Frontiers in aging neuroscience*, 13:189.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection. *INTERSPEECH 2020*.

Aparna Balagopalan and Jekaterina Novikova. 2021. Comparing Acoustic-based Approaches for Alzheimer's Disease Detection. *INTERSPEECH 2021*.

Aparna Balagopalan, Jekaterina Novikova, Frank Rudzicz, and Marzyeh Ghassemi. 2018. The Effect of Heterogeneous Data for Alzheimer's Disease Detection from Speech. In *NeurIPS 2018 Workshop Machine Learning for Health (ML4H)*.

Sebastian Braun, Hannes Gamper, Chandan KA Reddy, and Ivan Tashev. 2021. Towards efficient models for real-time deep noise suppression. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 656–660. IEEE.

Ron Brookmeyer, Nada Abdalla, Claudia H Kawas, and María M Corrada. 2018. Forecasting the prevalence of preclinical and clinical alzheimer's disease in the united states. *Alzheimer's & Dementia*, 14(2):121–129.

Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee. 2018. Phase-aware speech enhancement with deep complex u-net. In *International Conference on Learning Representations*.

Sofia de la Fuente Garcia, Craig W Ritchie, and Saturnino Luz. 2020. Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: a systematic review. *Journal of Alzheimer's Disease*, 78(4):1547–1574.

Sayanton V Dibbo, Yugyeong Kim, and Sudip Vhaduri. 2021. Effect of Noise on Generic Cough Models. In *2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–4. IEEE.

Harishchandra Dubey, Vishak Gopal, Ross Cutler, Ashkan Aazami, Sergiy Matusevych, Sebastian Braun, Sefik Emre Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, et al. 2022.

Icassp 2022 deep noise suppression challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9271–9275. IEEE.

Benjamin Eyre, Aparna Balagopalan, and Jekaterina Novikova. 2020. Fantastic features and where to find them: detecting cognitive impairment with a subsequence classification guided approach. *W-NUT at EMNLP 2020*.

Steven Y Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. 2020. 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*, pages 29–42.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2016. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Emil Fristed, Caroline Skirrow, Marton Meszaros, Raphael Lenain, Udeepa Meepegama, Stefano Cappa, Dag Aarsland, and Jack Weston. 2021. Evaluation of a speech-based AI system for early detection of Alzheimer's disease remotely via smartphones. *medRxiv*.

Lara Gauder, Leonardo Pepino, Luciana Ferrer, and Pablo Riera. 2021. Alzheimer Disease Recognition Using Speech-Based Embeddings From Pre-Trained Models. In *Proc. INTERSPEECH 2021*, pages 3795–3799.

Praveen Kumar Badimala Giridhara, Chinmaya Mishra, Reddy Kumar Modam Venkataramana, Syed Saqib Bukhari, and Andreas Dengel. 2019. A Study of Various Text Augmentation Techniques for Relation Classification in Free Text. *ICPRAM*, 3:5.

Emmanuel A Jammeh, B Carroll Camille, W Pearson Stephen, Javier Escudero, Athanasios Anastasiou, Peng Zhao, Todd Chenore, John Zajicek, and Emmanuel Ifeachor. 2018. Machine-learning based identification of undiagnosed dementia in primary care: a feasibility study. *BJGP Open*, page bjgpopen18X101589.

Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller. 2020. Effect of environmental noise in speech quality assessment studies using crowdsourcing. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE.

Lampros C Kourtis, Oliver B Regele, Justin M Wright, and Graham B Jones. 2019. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *NPJ digital medicine*, 2(1):1–9.

Michał Koziarski and Bogusław Cyganek. 2017. Image recognition with deep neural networks in presence of noise–dealing with and taking advantage of distortions. *Integrated Computer-Aided Engineering*, 24(4):337–349.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021. Detecting cognitive decline using speech only: The ADReSSo Challenge. *arXiv preprint arXiv:2104.09356*.

Israel Martínez-Nicolás, Thide E Llorente, Francisco Martínez-Sánchez, and Juan José G Meilán. 2021. Ten years of research on automatic voice and speech analysis of people with alzheimer's disease and mild cognitive impairment: a systematic review article. *Frontiers in Psychology*, 12:620251.

Marie Mc Carthy and P Schueler. 2019. Can Digital Technology Advance the Development of Treatments for Alzheimer's Disease?

Julien Meyer, Laure Dentel, and Fanny Meunier. 2013. Speech recognition in natural background noise. *PloS one*, 8(11):e79279.

Babak Naderi, Sebastian Möller, and Gabriel Mittag. 2018. Speech quality assessment in crowdsourcing: Influence of environmental noise. *44. Deutsche Jahrestagung für Akustik (DAGA)*, pages 229–302.

Jekaterina Novikova. 2021. 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)*, pages 334–339.

Jekaterina Novikova, Aparna Balagopalan, Ksenia Shkaruta, and Frank Rudzicz. 2019. Lexical features are more vulnerable, syntactic features have more predictive power. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 431–443.

Babafemi O Odelowo and David V Anderson. 2017. A noise prediction and time-domain subtraction approach to deep neural network based speech enhancement. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 372–377. IEEE.

Raghavendra Pappagari, Jaejin Cho, Sonal Joshi, Laureano Moro-Velázquez, Piotr Żelasko, Jesús Villalba, and Najim Dehak. 2021. Automatic Detection and Assessment of Alzheimer Disease Using Speech and Language Technologies in Low-Resource Scenarios. In *Proc. INTERSPEECH 2021*, pages 3825–3829.

Karol J Piczak. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018.

Gokul Prabhakaran, Rajbir Bakshi, et al. 2018. Analysis of Structure and Cost in a Longitudinal Study of Alzheimer's Disease. *Journal of Health Care Finance*.

Martin Prince, Adelina Comas-Herrera, Martin Knapp, Maëlenn Guerchet, and Maria Karagiannidou. 2016. World Alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future. *Alzheimer's disease International (ADI).*

Jessica Robin, John E Harrison, Liam D Kaufman, Frank Rudzicz, William Simpson, and Maria Yancheva. 2020. Evaluation of speech-based digital biomarkers: review and recommendations. *Digital Biomarkers*, 4(3):99–108.

Jessica Robin, Mengdan Xu, Liam D Kaufman, and William Simpson. 2021. Using digital speech assessments to detect early signs of cognitive impairment. *Frontiers in digital health*, 3:749758.

Morteza Rohanian, Julian Hough, and Matthew Purver. 2021. Alzheimer's Dementia Recognition Using Acoustic, Lexical, Disfluency and Speech Pause Features Robust to Noisy Inputs. *arXiv preprint arXiv:2106.15684.*

Antoine Slegers, Renee-Pier Filiou, Maxime Montembeault, and Simona Maria Brambati. 2018. Connected speech features from picture description in alzheimer's disease: a systematic review. *Journal of Alzheimer's Disease*, 65(2):519–542.

Cristiano Rafael Steffens, Lucas Ricardo Vieira Messias, Paulo Lilles Jorge Drews, and Silvia Silva da Costa Botelho. 2019. 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)*, pages 61–66. IEEE.

B Vellas and P Aisen. 2021. New hope for alzheimer's disease.

Sudip Vhaduri, Theodore Van Kessel, Bongjun Ko, David Wood, Shiqiang Wang, and Thomas Brunschwiler. 2019. Nocturnal cough and snore detection in noisy environments using smartphone-microphones. In *2019 IEEE international conference on healthcare informatics (ICHI)*, pages 1–7. IEEE.

Wei Xue, Catia Cucchiarini, Roeland van Hout, and Helmer Strik. 2019. Acoustic correlates of speech intelligibility: the usability of the egemaps feature set for atypical speech. In *Workshop on Speech and Language Technology in Education.*

Guochang Zhang, Libiao Yu, Chunliang Wang, and Jianqiang Wei. 2022. Multi-scale temporal frequency convolutional network with axial attention for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9122–9126. IEEE.

Luke Zhou, Kathleen C Fraser, and Frank Rudzicz. 2016. Speech Recognition in Alzheimer's Disease and in its Assessment. In *INTERSPEECH 2016*, pages 1948–1952.

Zining Zhu, Jekaterina Novikova, and Frank Rudzicz. 2018. Semi-supervised classification by reaching consensus among modalities. In *NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language IRASL.*

Zining Zhu, Jekaterina Novikova, and Frank Rudzicz. 2019. 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*, arXiv preprint arXiv:1808.06570.