Vinod Subramanian, Namhee Kwon, Raymond Brueckner, Nate Blaylock, Henry O'Connell,
"Detecting Mild Cognitive Impairment using Vocal Biomarkers from Spontaneous Speech,"
in Proceedings of the Workshop on Speech, Music and Mind 2024 (SMM), September 2024.

Detecting Mild Cognitive Impairment using Vocal Biomarkers from Spontaneous Speech

Vinod Subramanian[1], *Namhee Kwon*[1], *Raymond Brueckner*[1], *Nate Blaylock*[1], *Henry O'Connell*[1]

[1]Canary Speech, Inc., USA

{vinod, namhee, ray, nate, henry}@canaryspeech.com

Abstract

Mild Cognitive Impairment (MCI) is a condition where an individual's ability to perform regular activities starts deteriorating. It is often an early indicator of Dementia. By identifying individuals with MCI as soon as possible, affected individuals can receive care to help alleviate their symptoms as well as slow down the progression of MCI towards Dementia. In this paper, we explore the adoption of speech-based features as well as embeddings from pre-trained models to train machine learning models for detecting MCI. Our results show that using an XGBoost tree algorithm in combination with TRILLsson features as the input can successfully predict MCI with an accuracy of 81%. We compare the answers of patients to two different questions to detect MCI, the first being a "How are you?" question and the second asking the patient to describe a particular picture. Our results show that approaches based on the answers to the picture description question typically perform better than those using the "How are you?" answers.

Index Terms: MCI, mild cognitive impairment, speech analysis, vocal biomarkers, dementia

1. Introduction

Mild Cognitive Impairment (MCI) is described by Peterson et al. [1] as "a condition in which individuals demonstrate cognitive impairment with minimal impairment of instrumental activities of daily living (IADL)." IADL refers to activities such as using the telephone, the ability to eat and bathe, etc. [2]. The definitions and diagnostic criteria for identifying MCI are constantly evolving [3, 4, 5, 6] as medical professionals realize the importance of detecting MCI in patients. Kasper et al. [5] state that the two main reasons behind identifying MCI is to help alleviate symptoms caused due to cognitive impairment and to help prevent decline into dementia.

Typical methods to detect MCI require a high level of expertise [6]. Simpler tests such as the mini-mental state examination (MMSE) [7] and the Montreal Cognitive Examination (MoCA) [8] are not as sensitive to MCI [9, 6]. Liu et al. [10] conducted a survey on the rate of diagnosis of MCI for medicare patients in the United States of America and showed that only about 8% of expected cases were diagnosed with MCI. So there is a clear need for accessible and reliable tools to detect MCI so that patients can be identified earlier if they require medical intervention.

There is evidence that features extracted from speech correlate to cognitive impairment [11, 9, 12]. Speech is a promising direction to improving the accessibility of MCI assessments because automated detection algorithms can be used to predict MCI from speech features. Broadly, the approaches towards audio-based automatic MCI prediction can be divided into two groups: those based on signal processing-derived features on the one hand and those leveraging pre-trained deep learning model embeddings on the other hand.

Signal processing-based approaches extract features from speech using signal processing algorithms, which comprise speech rate, speech duration, pause rate etc. Themistocleus et al. [13] identify mild cognitive impairment in Swedish speakers by focusing on the vowel sounds uttered by participants. They extract features such as vowel formant frequencies, fundamental frequency, and vowel duration. In addition, they use age and gender as meta features. Those features are fed into a fully-connected network architecture and the authors train 10 different versions by varying the number of hidden layers from one to ten. They perform a 5-fold cross-validation as well as a 90-10 split ensuring that recordings from the same participants are not present in the different folds. For both evaluation strategies the model with 6 hidden layers shows the best performance. In the 90-10 split it has the highest F_1-score of 72% and for the 5-fold cross-validation it achieves the higest accuracy of 83%. Konig et al. [14] propose a mobile solution to measure cognitive impairment in elderly people. There is no healthy control in their experiments, instead, their categories are subjective cognitive impairment (SCI), mild cognitive impairment (MCI), and Alzheimer's Disease (AD). Each participant had to perform six spoken tasks such as sentence repetition, counting backwards, positive storytelling, etc., resulting in a total average duration of 15 minutes. They compute 55 vocal biomarkers for each speech task and choose a subset for classification based on the p-value from the Mann-Whitney test, then they use an SVM classifier for binary classification. For SCI vs. MCI they obtain 86% classification accuracy.

Approaches based on pre-trained deep embeddings try to circumvent the constraints imposed by small dataset sizes by first training a large network on a relevant task on large amounts of data - often in a self-supervised manner - and then leveraging this network as a feature or embedding generator. Typically, the outputs of the pre-trained network are used as inputs to a simpler, task-specific classifier, i. e. the detection of MCI in this study. Vetrab et al. [15] use pre-trained autoencoders as input features for a classifier. They pre-train the autoencoder on a generic Hungarian spontaneous speech dataset and use the embeddings as the input to a linear kernel support vector machine. They use 25-fold stratified cross validation with each fold consisting of 1 healthy and 1 MCI participant. Classification is performed on 5-second chunks and the best performance for MCI vs Healthy is obtained for median aggregation with an Area-Under-the-Curve (AUC) of 0.782. Bertini et al. [16] also use an autoencoder but they pre-train it on MCI data of Italian speakers. Even though their dataset is quite small with only 288 audio files, they use data augmentation techniques such as frequency

masking, time masking, and time warping to increase the data size to train the autoencoder. The resulting embeddings are used to train a multi-layer perceptron. For the binary discrimination between the Healthy Control (HC) group vs. MCI ∪ eD (early dementia), they get an F_1-score of 90.70% and for the ternary classification of HC vs. MCI vs. early Dementia, the F_1-score amounts to 84.71%.

The contributions of this work can be summarized as follows:

1. Demonstrate that embeddings extracted from pre-trained models are effective at differentiating between MCI and healthy patients and that they can outperform manually engineered features, such as language and prosodic features.
2. Features and models derived from the picture description responses excel at detecting MCI. However, the performance difference to those leveraging the free speech responses is not large.
3. Models trained on one response type perform well on the data of the other response type, which suggests that they might not be as different as one would think. This finding is important because it is less resource intensive to gather data in response to the question "How are you?" than to have patients describe a picture, eventually lowering the cost of data collection.

2. Dataset

We collect data in collaboration with two major medical institutes. We recruit patients diagnosed with MCI along with a healthy control group. The diagnosis is conducted via comprehensive clinical interviews as well as cognitive tests, including MoCA [8] and SLUMS [17], and interpreted as the final label, i. e. MCI or no-MCI (healthy).

We create a mobile application and use it to collect audio samples. The audio prompts are intended to elicit two different types of spontaneous speech: the first is a response to the question "How are you?" (referred to as *free speech*) and the second is a response to the request to describe a picture displayed on the device (referred to as *picture description*). We use the cookie-theft picture [18] widely used for screening dementia. If the audio samples are too short ($< 20s$), we exclude them from our experiments.

Table 1: *Distribution of subjects and assessments per class*

Class	Number of Subjects	Number of Assessments
Healthy	49	59
MCI	34	68

Table 2: *Audio duration statistics [sec]*

Audio type	mean	std	min	max
Free speech	44.77	13.95	21.35	136.13
Picture description	56.06	21.83	20.71	146.78

Table 1 shows the number of subjects and assessments per class used in our study. The statistics on the audio sample length are shown in Table 2; the average audio recording times are 44.77 seconds for free speech and 56.06 seconds for picture description.

3. Methods

This study investigates machine learning techniques for automatic identification of Mild Cognitive Impairment from spontaneous speech samples. We develop a system that constructs fixed-length feature vectors from variable-length audio recordings. The high dimensionality of raw audio signals presents a challenge in defining low-dimensional features and can lead to overfitting, particularly with the limited training data commonly encountered in the healthcare domain. To address this, we incorporate a pre-trained network, trained on a substantially larger dataset, into our overall architecture. We extract features, such as x-vectors and Trillson embeddings, from the pre-trained network's output layer and integrate them into our final machine learning model. Additionally, we explore the use of low-dimensional speech pattern features, including language and prosody information, for comparative analysis.

3.1. Features

3.1.1. X-vector features

X-vectors encode several aspects of a speaker's voice [19] and have been used in many applications, although they were initially proposed for the task of speaker identification [20]. Recently, x-vectors were adopted in the bio-informatics field, such as for the detection of Parkinson's and Alzheimer's disease [21][22][23] and the detection of depression [24] from speech.

In this study, we use the publicly available 16 kHz English VoxCeleb x-vector model [25], pre-trained on over 1 million audio recordings from more than 7000 celebrities contained in the VoxCeleb 1 and 2 datasets. The x-vector model is a deep neural network consisting of five 1-dimensional time-delay neural network (TDNN) layers [26], a statistics pooling layer, and a fully connected (FC) layer. The inputs to the model are 30-dimensional Mel Frequency Cepstral Coefficients (MFCC), computed for a frame length of 25 ms and a hop size of 10 ms. The MFCC frames are normalized using Cepstral Mean Variance Normalization (CMVN) before being fed to the network. The output of the last layer of the pre-trained model is used as a 512-dimensional x-vector feature. We generate a single x-vector for each audio input.

3.1.2. TRILLsson features

The TRILLsson model is a universal paralinguistic speech representation, derived via distillation from the pre-trained, much bigger TRILL models, as proposed by Shor et al. [27]. The TRILLsson models are available in five different flavors, differing in their underlying network architectures and network sizes. We adopt the largest TRILLsson model (v1), which is based on the Audio Spectrogram Transformer (AST) [28], a transformer-based model for audio classification. This model has 88.6 million parameters and is publicly available[1]. The generation process of the features via the TRILLsson model is configurable with respect to the frame size and frame shift, aggregating the internal frame-level latent representation over frames of the specified configuration. In our experiments, a single (global) 1024-dimensional feature vector is generated for each audio input.

[1] https://www.kaggle.com/models/google/TRILLsson

3.1.3. Language features

The language features are generated based on automatic speech recognition (ASR) results. Using the start and end time of the spoken words, we compute syllable duration, the number of spoken words per second, pause duration, the number of long pauses (longer than 150 ms), the maximum length of a pause, etc. We also compute filler (*"ah", "hmm", "eh", "uh"*, etc.) and word repetition ratios from the transcribed words. Finally, we add semantic features such as lexical complexity (n-gram probability based on a language model generated using the Librispeech dataset [29]), word usage (SMOG reliability) score [30], word ambiguity, and familiarity score based on linguistic research [31]. However, we do not create content-based semantic features to avoid overfitting to a specific domain or topic of the training data. The resulting feature vector dimension is 53.

3.1.4. Prosody features

The prosody features capture aspects of the modulations caused by the movement of the human articulatory organs during speech. These are mainly represented by the statistics of the fundamental frequency (F_0), energy, and speech rate, derived via supra-frame spectral analysis. The values are normalized following the approach in [32] and the minimum and maximum F_0 values and energy over the full audio signal are added. Further, as a measure related to speech rate, the rhythm of the acoustic energy pattern is analyzed, i. e. the duration and the number of syllables, pauses, speech segments, phonation, articulation, and average speaking duration. The pauses and durations are computed via energy-based voice activity detection instead of relying on automatic speech recognition (ASR) results.

3.2. Models

Given the small size of the dataset and to avoid overfitting, we opt for models that are generally known to be better suited to small datasets. For that reason, we compare three types of models, namely *Support Vector Machines* (SVM), *Random Forests* (RF), and *XGBoost*.

Support Vector Machines are large-margin. binary classifiers, which means that their predictions lack clear probabilistic interpretation. In our experiments we adopt SVMs with a radial basis function kernel and a complexity of 1.0 for all experiments. On the other hand, Random Forests and XGBoost are tree based classifiers that do feature probabilistic outputs allowing for the possibility to tune their decision threshold required for generating a prediction. A Random Forest is a randomized ensembling approach, in which the constituent trees are independent from one another, while XGBoost trains each tree in an ensemble sequentially to compensate for the weaknesses of the previous tree. SVM and Random Forest models were trained using the SKLearn toolkit [33], while XGBoost was trained utilizing the Python xgboost package [34].

3.3. Metrics

To assess model performance we report *Sensitivity* and *Specificity*, which are commonly used metrics in the medical field to evaluate diagnostic efficacy. They are defined as the proportion of correctly diagnosed positive and negative labels, respectively. Hence, sensitivity is the recall of positive labels (MCI) and specificity is the recall of the negative labels (healthy control). *Unweighted Average Recall* (UAR), also known as *Unweighted Accuracy* (UA), is an average of Sensitivity and Specificity without giving a penalty to the minority class recall in situations where the label distribution is imbalanced. We finally report the standard (weighted) accuracy (Acc) and the *Area Under the Curve* (AUC).

4. Experiments and Results

As described in Section 3.2 the three models we investigated are Support Vector Machines, Random Forests and XGBoost, which we combined with the four feature sets x-vectors, TRILLsson, language, and prosody features. These model-feature combinations are assessed on the two types of speech responses, namely picture description and free speech.

The dataset is split following a stratified 10-fold cross-validation (CV) strategy, ensuring that the speakers are disjunct between the folds and that each fold has a label distribution that reflects the distribution of the whole dataset. While the data is split according to the subject names, the features are extracted per assessment and the prediction results are reported on the assessment level. Consequently, each fold of data can have a different number of assessments, since the number of assessments varies across subjects.

For the XGBoost and Random Forest models, the (default) threshold of 0.5 is selected as the binary decision threshold for prediction in order to calculate the Accuracy, UA, Sensitivity and Specificity.

4.1. Experiment 1: Cross fold validation

The aim of this experiment is to identify the most suitable features and models for predicting MCI and to observe whether the choice of question has any impact on performance. The performance metrics are computed using cross fold validation.

The results for this experiment are shown in Table 3. The XGBoost model with TRILLsson features on the picture description task is unquestionably the best model. It shows the highest UA, Accuracy, Specificity, and AUC across all model-feature combinations. For the XGBoost model with x-vector and TRILLsson on Free Speech, there is a large discrepancy between the UA and AUC, which suggests that a tuned decision threshold could lead to higher performance on the data. As a consequence for the rest of this experiment the AUC will be used as the main metric to compare the performance of different models.

Looking at just the AUC, the x-vector and TRILLsson features perform much better than language or prosody features across all models and speech types. This indicates that embeddings of pre-trained models capture audio characteristics more effectively than explicitly designed features, such as language and prosody features. The one exception in our experiment is the AUC for language features and the picture description, which is similar to the x-vector and TRILLsson results with the Random Forest and XGBoost models. This exception could be explained by the additional structure present in the picture description task, which language features might potentially exploit more successfully compared to the open ended nature of free speech.

4.2. Experiment 2: Cross speech evaluation

This experiment aims to assess the performance of a model trained on one type of one question type's speech, but tested on the other. Since both types of speech use the same folds for cross fold validation, the cross speech predictions can be

Table 3: *Model performance over assessments*

Model	Features	Free Speech					Picture Description				
		UA	Acc	Specificity	Sensitivity	AUC	UA	Acc	Specificity	Sensitivity	AUC
SVM	x-vector	0.71	0.72	0.59	0.84	0.71	0.72	0.72	0.64	0.79	0.72
	TRILLsson	0.71	0.72	0.49	**0.93**	0.72	0.67	0.68	0.52	0.82	0.67
	language	0.64	0.63	0.80	0.48	0.64	0.63	0.62	0.80	0.47	0.63
	prosody	0.56	0.56	0.51	0.60	0.56	0.57	0.57	0.49	0.65	0.57
Random Forest	x-vector	0.72	0.72	0.71	0.72	0.78	0.72	0.72	0.69	0.75	0.78
	TRILLsson	0.73	0.73	0.69	0.76	0.82	0.76	0.76	0.75	0.76	0.83
	language	0.66	0.67	0.59	0.74	0.64	0.70	0.70	0.63	0.76	0.76
	prosody	0.50	0.51	0.41	0.60	0.57	0.63	0.64	0.59	0.68	0.68
XGBoost	x-vector	0.69	0.69	0.71	0.66	0.78	0.72	0.72	0.71	0.74	0.73
	TRILLsson	0.66	0.66	0.66	0.66	0.77	**0.81**	**0.81**	**0.83**	0.79	**0.87**
	language	0.66	0.66	0.66	0.66	0.67	0.73	0.73	0.73	0.74	0.79
	prosody	0.54	0.54	0.44	0.63	0.54	0.61	0.61	0.54	0.68	0.66

computed without any concerns for speaker overlap between the train and test sets. The AUC is again used as the metric to compare the performance.

Table 4: *Comparison of AUC values between testing on picture description and free speech (FS: Free Speech, PD: Picture Description)*

Model	Features	Train on FS		Train on PD	
		FS	PD	FS	PD
SVM	x-vector	0.72	0.72	0.77	0.72
	TRILLsson	0.71	0.67	0.71	0.67
	language	0.64	0.63	0.64	0.63
	prosody	0.56	0.55	0.55	0.57
Random Forest	x-vector	0.78	0.79	0.78	0.78
	TRILLsson	0.82	0.84	0.82	0.84
	language	0.74	0.65	0.72	0.76
	prosody	0.57	0.63	0.61	0.68
XGBoost	x-vector	0.78	0.77	0.75	0.73
	TRILLsson	0.77	0.83	0.82	0.87
	language	0.67	0.70	0.68	0.79
	prosody	0.54	0.60	0.60	0.66

The results for this experiment are shown in Table 4. For most of the models, the performance is similar and in some cases better when testing on the other type of speech as opposed to its own type of speech. For the x-vector and TRILLsson features which were identified as the best features in Experiment 4.1, the AUC is similar irrespective of the type of data used during testing. In fact, in some cases, e. g. TRILLsson features with SVM trained on Picture Description, the AUC is better on the free speech than on the picture description task itself.

These results indicate that the used features and models trained on one type of speech prompt transfer well to the other. Note that all feature types in this work do not analyze the semantic content, but instead only capture the inherent speech patterns. This is important because in practical settings it might not always be possible to assure the use of a specific picture across assessments picture, whereas it is considerably easier to collect data for simple and generic question like “How are you?”.

5. Conclusion and Future work

In this study we explore a cross section of different models, features,q and types of speech in order to gain a deeper insight into effective techniques to detect mild cognitive impairment from speech. The results show that the best-performing model is XGBoost operating on TRILLsson features utilizing the picture description speech resulting in an AUC of 0.87. The second highest AUC value of 0.83 belongs to the Random Forest utilizing TRILLsson features on the picture description task. In general, the picture description task performs slightly better than free speech, but the difference is relatively limited. This is an important result because obtaining free speech data requires less effort than the picture description objective, so in a practical setting the former approach might be advantageous.

Comparing the performance of the free speech models on the picture description data and vice versa shows a surprising trend: there is none to little degradation and in some cases the performance is even better on the cross-task than on the same type of speech. This suggests that there is consistency in the features being used to identify MCI between the two types of speech.

In the future, we will continue to explore improved solutions for deploying these models in practical scenarios, which are characterized by varying environmental situations, background noise, different acquisition devices, etc. These aspects often adversely impact model performance and any model deployed ”in the wild” needs to be robust to these variations.

6. References

[1] R. C. Petersen, O. Lopez, M. J. Armstrong, T. S. Getchius, M. Ganguli, D. Gloss, G. S. Gronseth, D. Marson, T. Pringsheim, G. S. Day *et al.*, “Practice guideline update summary: Mild cognitive impairment: Report of the guideline development, dissemination, and implementation subcommittee of the american academy of neurology,” *Neurology*, vol. 90, no. 3, pp. 126–135, 2018.

[2] M. P. Lawton and E. M. Brody, “Assessment of Older People: Self-Maintaining and Instrumental Activities of Daily Living1,” *The Gerontologist*, vol. 9, no. 3_Part_1, pp. 179–186, Oct. 1969, _eprint: https://academic.oup.com/gerontologist/article-pdf/9/3_Part_1/179/1466322/9-3_Part_1-179.pdf. [Online]. Available: https://doi.org/10.1093/geront/9.3_Part_1.179

[3] B. Winblad, K. Palmer, M. Kivipelto, V. Jelic, L. Fratiglioni, L.-O. Wahlund, A. Nordberg, L. Bäckman, M. Albert, O. Almkvist *et al.*, “Mild cognitive impairment–beyond controversies, towards a consensus: report of the international working group on mild cognitive impairment,” *Journal of internal medicine*, vol. 256, no. 3, pp. 240–246, 2004.

[4] R. C. Petersen, B. Caracciolo, C. Brayne, S. Gauthier, V. Jelic, and L. Fratiglioni, “Mild cognitive impairment: a concept in evo-

lution," *Journal of internal medicine*, vol. 275, no. 3, pp. 214–228, 2014.

[5] S. Kasper, C. Bancher, A. Eckert, H. Förstl, L. Frölich, J. Hort, A. D. Korczyn, R. W. Kressig, O. Levin, and M. S. M. Palomo, "Management of mild cognitive impairment (mci): the need for national and international guidelines," *The World Journal of Biological Psychiatry*, vol. 21, no. 8, pp. 579–594, 2020.

[6] R. A. Dunne, D. Aarsland, J. T. O'Brien, C. Ballard, S. Banerjee, N. C. Fox, J. D. Isaacs, B. R. Underwood, R. J. Perry, D. Chan *et al.*, "Mild cognitive impairment: the manchester consensus," *Age and ageing*, vol. 50, no. 1, pp. 72–80, 2021.

[7] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state": a practical method for grading the cognitive state of patients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[8] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[9] L. Tóth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatloczki, Z. Bánréti, M. Pakaski, and J. Kalman, "A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech," *Current Alzheimer Research*, vol. 15, no. 2, pp. 130–138, Nov. 2017.

[10] Y. Liu, H. Jun, A. Becker, C. Wallick, and S. Mattke, "Detection Rates of Mild Cognitive Impairment in Primary Care for the United States Medicare Population," *The Journal of Prevention of Alzheimer's Disease*, Oct. 2023. [Online]. Available: https://doi.org/10.14283/jpad.2023.131

[11] K. W. Kim, S.-H. Na, Y.-C. Chung, and B.-S. Shin, "A Comparison of Speech Features between Mild Cognitive Impairment and Healthy Aging Groups," *Dementia and Neurocognitive Disorders*, vol. 20, pp. 52 – 61, 2021.

[12] K. E. Forbes-McKay and A. Venneri, "Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task," *Neurological Sciences*, vol. 26, no. 4, pp. 243–254, Oct. 2005. [Online]. Available: https://doi.org/10.1007/s10072-005-0467-9

[13] C. Themistocleous, M. Eckerström, and D. Kokkinakis, "Identification of Mild Cognitive Impairment From Speech in Swedish Using Deep Sequential Neural Networks," *Frontiers in Neurology*, vol. 9, 2018. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fneur.2018.00975

[14] A. Konig, A. Satt, A. Sorin, R. Hoory, A. Derreumaux, R. David, and P. H. Robert, "Use of Speech Analyses within a Mobile Application for the Assessment of Cognitive Impairment in Elderly People," *Current Alzheimer Research*, vol. 15, no. 2, pp. 120–129, 2018.

[15] M. Vetrab, J. V. Egas-Lopez, R. Balogh, N. Imre, I. Hoffmann, L. Toth, M. Pakaski, J. Kalman, and G. Gosztolya, "Using Spectral Sequence-to-Sequence Autoencoders to Assess Mild Cognitive Impairment," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Singapore, Singapore: IEEE, May 2022, pp. 6467–6471. [Online]. Available: https://ieeexplore.ieee.org/document/9746148/

[16] F. Bertini, D. Allevi, G. Lutero, D. Montesi, and L. Calzà, "Automatic Speech Classifier for Mild Cognitive Impairment and Early Dementia," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–11, Jan. 2022. [Online]. Available: https://dl.acm.org/doi/10.1145/3469089

[17] J. E. Morley and N. Tumosa, "Saint louis university mental status examination," institution: American Psychological Association. [Online]. Available: https://doi.apa.org/doi/10.1037/t27282-000

[18] L. Cummings, "Describing the cookie theft picture: Sources of breakdown in alzheimer's dementia," *Pragmatics and Society*, vol. 10, pp. 151–174, 03 2019.

[19] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 726–733.

[20] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[21] L. Jeancolas, D. Petrovska-Delacrétaz, G. Mangone, B.-E. Benkelfat, J.-C. Corvol, M. Vidailhet, S. Lehéricy, and H. Benali, "X-vectors: New quantitative biomarkers for early parkinson's disease detection from speech," *Frontiers in Neuroinformatics*, vol. 15, p. 578369, 2021.

[22] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect parkinson's disease from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1155–1159.

[23] R. Pappagari, J. Cho, L. Moro-Velazquez, and N. Dehak, "Using state of the art speaker recognition and natural language processing technologies to detect alzheimer's disease and assess its severity." in *INTERSPEECH*, 2020, pp. 2177–2181.

[24] N. Kwon, S. Hossain, N. Blaylock, H. O'Connell, N. Hachen, and J. Gwin, "Detecting Anxiety and Depression from Phone Conversations using x-vectors," in *Proc. Workshop on Speech, Music and Mind*, 2022, pp. 1–5.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[26] A. H. Waibel, T. Hanazawa, G. E. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 328–339, 1989.

[27] J. Shor and S. Venugopalan, "TRILLsson: Distilled Universal Paralinguistic Speech Representations," in *Proc. Interspeech 2022*. Incheon, Korea: ISCA, September 2022, pp. 356–360.

[28] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*. Brno, Czech Republic: ISCA, 2021, pp. 571–575.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[30] K. L. Grabeel, J. Russomanno, S. Oelschlegel, E. Tester, and R. E. Heidel, "Computerized versus hand-scored health literacy tools: a comparison of simple measure of gobbledygook (smog) and flesch-kincaid in printed patient education materials," *Journal of the Medical Library Association: JMLA*, vol. 106, no. 1, p. 38, 2018.

[31] S. Cho, N. Nevler, S. Ash, S. Shellikeri, D. J. Irwin, L. Massimo, K. Rascovsky, C. Olm, M. Grossman, and M. Liberman, "Automated analysis of lexical features in frontotemporal degeneration," *Cortex*, vol. 137, pp. 215–231, 2021.

[32] R. Brueckner, M. Takegami, N. Kwon, N. Blaylock, V. Subramanian, E. Kiyoshige, S. Ogata, Y. Nakaoku, H. O'Connell, and K. Nishimura, "Voice Technology to Identify Fatigue from Japanese Speech," in *Proc. SMM23, Workshop on Speech, Music and Mind 2023*, 2023, pp. 31–35.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794.