# Mild Cognitive Impairment (MCI) Detection via Voice Analysis

*Namhee Kwon, Raymond Brueckner, Shahruk Hossain, Vinod Subramanian, Nate Blaylock, Henry O'Connell*

Canary Speech Research Report
January 2023

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

## Abstract

Dementia is a significant threat to our society, both in terms of personal life and public health. Early diagnosis of dementia is critical to provide timely treatment. As an easily accessible and non-invasive technique, we study voice analysis techniques to identify Mild Cognitive Impairment. We build a deep neural network model using acoustic features to represent a voice signal. A model using *x-vectors* per 5-second audio segments outperformed other feature models. The model prediction sensitivity is 0.58 and the specificity is 0.83.

**Index Terms**: Alzheimer, Dementia, MCI, Mild Cognitive Impairment

## 1. Introduction

Dementia is a collective term covering a wide range of neurodegenerative disorders that manifest a set of related symptoms such as progressive impairments in memory, cognitive abilities, and behavior. It often negatively affects a person's feelings and relationships and the ability to carry out everyday activities. While there are many known types of dementia, such as vascular, Lewy body, fronto-temporal, or sub-cortical (comprising Parkinson's, Huntington's, and multiple sclerosis diseases) variants, the pre-dominant cause for dementia is Alzheimer's disease (AD) - the World Health Organization (WHO) estimates that AD accounts for approximately 60-70% of the cases worldwide[1].

Although dementia is not an inevitable consequence of aging, it mainly affects the elder population. According to the WHO, dementia is currently the seventh leading cause of death among all diseases and one of the major causes of disability and dependency among older people globally. This negatively impacts their physical, psychological, social, and economic life, and also affects their carers, families, and society at large. Due to the severity of this situation, there is a strong need for scalable and cost-effective methods to detect dementia, from pre-clinical stages, over prodromal conditions, to late-stage dementia comprising AD.

Mild Cognitive Impairment (MCI) refers to a small but measurable degree of cognitive decline during the prodromal stage, i. e. between the expected decline of normal aging and the more serious decline of dementia. MCI is not usually detectable by casual conversation or observation, and it is estimated that approximately 60% of the people with MCI are unaware that they have a cognitive problem [1]. People with MCI struggle to remember recent conversations or events, keep track of schedules and appointments, or use new guidelines for a task. Distinguishing MCI from normal aging is a difficult task even for the most conscientious primary care physicians, yet doing so is important for timely intervention and optimal treatment results.

While in the past decades, a lot of progress was made developing biomarkers for AD diagnosis, such as neuropsychological tests [2], neuroimaging techniques [3], and amyloid-$\beta$ and hyperphosphorylated tau as fluid biormarkers [4]. These methods are typically invasive and come at a relatively high cost [5]. As a scalable and automated alternative, in the last decade, a number of studies investigated the use of speech and language features for the detection of AD and MCI, and proposed various signal processing and machine learning methods for this prediction task [6]. An additional motivation for adopting this approach is the fact that language deficits are present from the early and prodromal stages [7].

Speech-based studies aiming to detect AD, MCI, or dementia in general, can generally be divided into two groups: The first investigates the effects of linguistic and lexical content of speech, i. e. *what* people say. Related studies rely on a plethora of diverse speech-derived features, such as parts-of-speech (POS) rate, ratios, indices, and statistics [8]; connected speech including fluency errors, lexical and semantic content, and syntactic complexity [7] [9] [10]; information content and repetitiveness [11]; Linguistic Inquiry and Word Counts (LIWC) [12] [13]; complexity, hesitation, and intelligibility features [14]. Many of the cited research groups actually use a combination of the above-mentioned linguistic-lexical features.

The second group of speech-based research instead focuses on the acoustic and prosodic quality of speech, and hence analyzes i. e. *how* people speak and how this relates to the diverse forms of dementia. Voice and speech analysis are common diagnostic instruments used by some specialists, such as audiologists or speech pathologists and it is highly desirable to develop ways to detect certain forms of dementia in an automated, objective and scalable way. Several studies have identified acoustic correlates of pathological voice alterations [15] [16] and in recent years several research groups have investigated the effect of both suitable speech features [17] [18] [19] [20] on the one hand and their application to more advanced machine learning methods [21] [22] [23] on the other hand. Cummins et al. [24] compared different acoustic and linguistic methodologies for the recognition of dementia, while Haider et al. [25] assessed the suitability of paralinguistic acoustic features on the task of Alzheimer's detection in spontaneous speech. Mirzaei et al. [26] instead adopted a feature selection method of voice parameters for the

[1] https://www.who.int/publications/i/item/global-action-plan-on-the-public-health-response-to-dementia-2017–2025

prediction of early Alzheimer's disease. Instead of acoustic features directly derived from the audio signal, some research groups also leverage derived acoustic features, e. g. via an intermediate ASR system, as demonstrated by Tóth et al. [27].

In this study, we focus exclusively on this latter group, namely acoustic voice analysis, and report results on the detection of Mild Cognitive Impairment from speech recordings. While there exists a substantial body of research for Alzheimer's Detection, there the acoustic-based voice analysis of MCI is relatively underexplored. One of the reasons could be that it is considerably more difficult to differentiate between the speech of the normal population (including normal aging) and MCI.

# 2. Data Collection

## 2.1. Phone Survey

We collected speech audio samples and labels in Japan through phone surveys with senior citizens at the age of 70 years and beyond. Trained representatives called the randomly selected elderly adults to ask them a catalog of questions after they had given consent to the data collection. The representatives were given interview scripts including a list of questionnaires and example questions to elicit free speech from the participants. The free speech questions included aspects of daily life, shopping, housework, job, etc.

## 2.2. MPI Score and MCI binary label

The (binary) target labels used during training and evaluation were derived from the Memory Performance Index (MPI), which measures the pattern of recalled and non-recalled words of the *Consortium to Establish a Registry for Alzheimer's Disease Wordlist* (CWL) [28]. The MPI score ranges from 0 to 100, representing dementia severity. It was reported to have an ROC accuracy of 96% in distinguishing conditions of normal from mild cognitive impairment [29].

During the interview phone calls, the representative asked CWL to the participants and counted the correctly recalled words. The MPI score was calculated accordingly, being subsequently scaled per age, sex, education, and race (called *Millennia* in Japan). The binary MCI label is defined by the millennia-MPU score: if the millennia-MPI is less than or equal to 49.7, the person was categorized as having MCI, on the other hand, if the score was greater than or equal to 50.3, the person was categorized as normal. The gray zone, i. e. the range between 49.7 and 50.3, was excluded from the experiment to avoid confusion during training.

## 2.3. Data Pre-Processing

The audio files were collected in the uncompressed WAV format in order to eliminate any potential negative effect of compression on the final prediction of our models. Since the audio was collected via conversations over the phone - including potentially band-limiting landline, blue-tooth, or recording devices - all recordings were re-sampled to 8 kHz in order to ensure a consistent sampling rate. We obtained de-identified participant channel audio and performed an initial segmentation by checking pause and speech durations. Furthermore, we filtered out all audio segments representing answers to MPI questionnaires (recall of CWL) because our goal was to identify MCI from general speech not processing the CWL answers and scoring MPI automatically. In order to provide a fair evaluation for the case when the model is applied in general spontaneous speech, we extracted and analyzed only free speech portions by applying a set of heuristic rules and analyzing pauses and keywords.

The audio segments so selected were concatenated resulting in one audio sample per phone call. We measured the duration of voice activity by summing the duration of speech segments - determined by an ASR (automatic speech recognition) system - excluding pauses between words. We removed short audio samples based on that voice activity duration. The statistics over each sample's voice activity duration are presented in Table 1. Since very short speech samples - the minimum voice activity duration $d_{min}$ is just about 5 seconds - generally do not contain sufficient information to robustly detect MCI from speech, samples whose voice activity duration was shorter than 30 seconds were removed from the data set. Out of the 1074 phone calls (samples) that were collected, 129 short audio files were excluded, so the final number of samples we used for our study was 945.

Table 1: *Statistics of Sample-Wise Voice Activity Duration*

| Statistical Value | Voice Activity Duration [sec] |
|---|---|
| Mean | 76 |
| StdDev | 58 |
| Minimum | 5.1 |
| Maximum | 649.3 |
| Voice Activity Duration | Number of Samples |
| $\geq$ 30 seconds | 945 |
| Total collected | 1074 |

Table 2 shows the distribution of labels. Around 25% of the participants are labeled as *has-MCI* (or more severe dementia). As explained in Section 2.2, the grey zone where the MPI score is just around the cut-off score was excluded from the experiment.

Table 2: *Label Distribution*

| Category | Count | Percentage |
|---|---|---|
| Normal | 704 | 74.5% |
| has-MCI | 233 | 24.7% |
| N/A (Grey Zone) | 8 | 0.8% |
| Total | 945 | 100% |

# 3. Method

## 3.1. Feature Extraction

The variable-length audio input signal is converted to fixed-sized representations, called *feature vectors*. We either represent the statistics of per-frame values over the full audio input (acoustic feature) or we generate a representative vector compared to other audio signals (i-vector and x-vector). Although text-based features analyzing the speech contents were reported more successful in identifying Alzheimer's disease in previous

work [30], we focus on the acoustic and signal-based representation that is less dependent on language or speech topic. The details about the used features are described in the following:

- **Acoustic** features are calculated on a per-frame basis, where a frame is defined as a 25 ms sliding window over the audio signal being computed every 10 ms. For each window, a 36-dimensional vector is calculated consisting of Mel-Frequency Cepstral Coefficients (MFCC) [31], Perceptual Linear Prediction (PLP) [32], and fundamental frequency ($F_0$) features. From the sequence of these vectors, their first-order and second-order time derivatives are computed and added to the base features. Based on a word-boundary Voice Activity Detector (VAD) of an ASR system, all non-speech feature segments are subsequently removed. Finally, 18 statistical values such as mean, percentile, slope, skewness, kurtosis, quartile, etc., are computed per individual feature across all (speech) frames resulting in a fixed-length representation of the variable-length audio input.

- **I-vectors** are data-driven low-dimensional speech representations initially invented for the purpose of robust speaker recognition [33], which encodes the speaker's voice characteristics. A Universal Background Model (UBM) is trained on a large dataset of high-dimensional acoustic features such as MFCCs per frame, to model the variability across different speakers. An *i-vector* is extracted using factor analysis on the Gaussian Mixture Models (GMM) derived from the UBM [34]. For a robust *i-vector* representation of audio inputs a pre-trained model is generated based on a large audio corpus - independent of the current task of MCI prediction - that covers diverse speech styles. Since no pre-trained 8 kHz i-vector models are publicly available, we trained a model on the *Corpus of Spontaneous Japanese (CSJ)*[2] [35] and the *Callhome Japanese* corpus [36].

- **X-vectors** were proposed as an improvement over i-vectors and are built using a deep neural network (DNN) [37]. They, too, are trained in a supervised, data-driven fashion and return a fixed-sized representation of a variable-length input. The x-vector model is trained on the feature frames containing 23 MFCCs extracted using a frame length of 25 milliseconds and shift by 10 milliseconds. Additionally, the MFCC frames are normalized using Cepstral Mean Variance Normalization (CMVN) before they are fed to the network. The x-vector model itself is a neural network composed of 5 stacked one-dimensional convolutional layers, also called *time-delay neural network (TDNN)* layers [38], a statistics pooling layer, and a final fully connected layer. In our study, we use the statistics layer output instead of the final fully-connected layer output, producing a feature vector of 3000 dimensions per input sequence. As in the *i-vector* approach, we need an *x-vector* model trained on a large corpus. For our purposes we leverage a publicly available English x-vector model[3] trained on 8 kHz audio data from several data sets including *Switchboard* [39], *Mixer 6* [40] and *NIST SREs* [41]. In addition, we trained our own x-vector model for Japanese on the *CSJ* and *Callhome Japanese* corpora.

[2]https://clrd.ninjal.ac.jp/csj/en/index.html

[3]https://kaldi-asr.org/models/3/0003_sre16_v2_1a.tar.gz

### 3.2. Model Architecture

Using the audio representations described in Section 3.1, we build models at two different levels: first, *per-speaker* models, and second, *per-segment* models. For the per-speaker models, we generate a single feature vector given a segment of per-speaker (concatenated) audio, pre-processed as described in Section 2.3. We apply a fully-connected DNN architecture composed of 256 nodes with ReLU activation followed by a final softmax layer. This network is trained using the Adam optimizer [42] and the learning rate, initialized to a value of $10^{-3}$, follows an *inverse time decay* schedule. Further, to avoid overfitting, a 20% dropout rate is applied.

For the per-segment model, we split each speaker's audio input into a set of fixed-length segments. Based on findings of preliminary experiments, the segment length is fixed to 5 seconds with a and is created every 2.5 seconds. We treat each segment as an independent sample, and the model is trained to return the corresponding label. Each segment is input into a classifier architecture which is composed of two stacked 1-D convolutional layers with 256 and 512 filters, respectively, both using a kernel size of 3, and a ReLU activation function. In addition, L2 regularization is applied on these layers during training to further reduce the risk of overfitting. After each convolutional layer, batch normalization is applied. The output of the second 1-D convolutional layer is followed by a *generalized mean pooling*, a fully connected layer with 128 nodes, and an output layer with a softmax activation function. The network is trained adopting the *Adamax* optimizer [42] with an initial learning rate of $10^{-3}$ and a cosine decay schedule (with restarts).

As shown in Table 2 the data set is imabalance with the *has-MCI* class being underrepresented. Since this has a negative effect on the final classifier performance, class weights are applied during training to adjust the cost of mis-classifying the minority class (*has-MCI*). In addition, *binary focal cross-entropy loss* was used because it was reported to be more effective for imbalanced data sets [43].

Since in each model type the final layer is a *Softmax* layer, it returns two probability values, on for *Normal* and on for *has-MCI* class (both summing up to 1). To determine the final prediction given a speaker in the per-segment model, we compute the average of the output prediction probabilities for the *has-MCI* class for all segments whose probability is greater than than the 5 percentile and lower than the 95 percentile per speaker. If the average is greater than or equal to 0.51, the final prediction is determined as *has-MCI*. In the per-sepaker model the final prediction label for a given speaker is determined by averaging each segment's prediction output.

## 4. Experiment

We split the data into three groups: training, development, and test set. We train the model on the training set and monitor the model performance on the development set, e. g. for early stopping or tuning the hyper-parameters. The test

Table 3: *Per-Speaker Model Performance (Fully-connected)*

| Features | Acc | UA | Sensitivity | Specificity | AUC |
|---|---|---|---|---|---|
| xvec(EN) | 0.60 | 0.54 | 0.42 | 0.65 | 0.56 |
| xvec(JP) | 0.57 | 0.54 | 0.47 | 0.60 | 0.49 |
| Acoustic | 0.60 | 0.48 | 0.26 | 0.69 | 0.54 |
| ivec(JP) | 0.80 | 0.60 | 0.26 | 0.94 | 0.60 |

Table 4: *Per-Segment Model Performance (CNN)*

| Model | Acc | UA | Sensitivity | Specificity | AUC |
|---|---|---|---|---|---|
| xvec (EN) | 0.78 | 0.71 | 0.58 | 0.83 | 0.66 |
| xvec (JP) | 0.45 | 0.58 | 0.36 | 0.79 | 0.53 |
| Acoustic | 0.41 | 0.52 | 0.68 | 0.36 | 0.50 |

set is completely blinded during training, and the prediction performance on the test set is compared and reported for evaluation. The data set split is done in a random manner but we approximately maintain the label distribution of full data set. The training, development, and test set size is 751, 95, and 91, respectively. For the per-segment model, 34 segments per speaker are defined on average.

We measure the sensitivity and specificity along with the accuracy (Acc). The sensitivity is the recall rate of the positive label (*has-MCI*) and the specificity is the recall rate of the negative label (*Normal*). We also report the unweighted accuracy (UA), which is the average of sensitivity and specificity, i. e. the respective class recalls. One benefit of reporting the unweighted accuracy is that we can perform a fair comparison when the label distribution is imbalanced because the model's chance UA is always 0.5 regardless of the label distribution. Similarly, we use the Area-Under-The-(ROC)-Curve (AUC) to measure the models' performance considering the trade-off between sensitivity and specificity based on a threshold on the prediction probability.

## 5. Result

We evaluate the model performance on the test set using various metrics as explained above. Each combination of features and model architectures shows different results, and the performance of the per-speaker and the per-segment models is depicted in Table 3 and Table 4, respectively. For the per-segment model, we report the final speaker-level prediction performance. For the per-speaker model, both x-vector models pre-trained on English or Japanese show the same UA with a trade-off between sensitivity and specificity, while the English *x-vector* model yields a higher AUC. The i-vector model performs best in terms of UA and AUC.

The difference between models increases in the per-segment model. The model using x-vectors pre-trained on English outperforms the model using the Japanese x-vectors. The reason is un-clear - however, the number of speakers in the English pre-trained model is higher than in the Japanese pre-trained model, so we conjecture that the English x-vector model generalizes better. It is very encouraging to see that the speech representation based on English is transferable to Japanese, because this allows for language-independent models and in general, there exist more publicly available English resources to pre-train robust models.

The best model performance is achieved by the per-segment CNN model using the x-vector pre-trained on English. The accuracy is 0.78 and the unweighted accuracy 0.71, with a sensitivity and specificity of 0.58 and 0.81, respectively.

## 6. Conclusion

In this study, we built a DNN model to predict MCI labels from spontaneous speech collected via phone calls with elderly adults in Japan. We generated two types of models: a per-speaker model and a per-segment model. The per-segment model using x-vector features pre-trained on English outperformed other models. The model was trained on 5-second segments using a TDNN/CNN architecture and the final prediction per speaker was determined by averaging each segment's class prediction probability. The best model's accuracy was 0.78 and the unweighted accuracy 0.71. When considering that the task of identifying MCI is comparably more difficult than identifying other forms of severe dementia, our acoustic voice analysis approach looks very promising. We will extend our research looking into different features and model architectures to identify MCI and will explore more sophisticated language-adaptive approaches in the future.

## 7. References


[1] J. L. Purser, G. G. Fillenbaum, and R. B. Wallace, "Memory complaint is not necessary for diagnosis of mild cognitive impairment and does not predict 10-year trajectories of functional disability, word recall, or short portable mental status questionnaire limitations," *Journal of the American Geriatrics Society*, vol. 54, no. 2, pp. 335–338, Feb. 2006.

[2] R. M. Chapman, M. Mapstone, A. P. Porsteinsson, M. N. Gardner, J. W. McCrary, E. DeGrush, L. A. Reilly, T. C. Sandoval, and M. D. Guillily, "Diagnosis of Alzheimer's disease using neuropsychological testing improved by multivariate analyses." *Journal of clinical and experimental neuropsychology*, vol. 32, no. 8, pp. 793–808, Oct. 2010.

[3] T. Varghese, R. Sheelakumari, J. S. James, and P. Mathuranath, "A review of neuroimaging biomarkers of Alzheimer's disease," *Neurology Asia*, vol. 18, no. 3, pp. 239–248, 2013.

[4] J. C. Lee, S. J. Kim, S. Hong, and Y. Kim, "Diagnosis of Alzheimer's disease utilizing amyloid and tau as fluid biomarkers," *Experimental & Molecular Medicine*, vol. 51, no. 5, pp. 1–10, May 2019.

[5] I. Martínez-Nicolás, T. E. Llorente, F. Martínez-Sánchez, and J. J. G. Meilán, "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*, vol. 12, 2021.

[6] M. L. B. Pulido, J. B. A. Hernández, M. F. Ballester, C. M. T. González, J. Mekyska, and Z. Smékal, "Alzheimer's disease and automatic speech analysis: A review," *Expert Systems with Applications*, vol. 150, p. 113213, Jul. 2020.

[7] F. Cuetos, J. C. Arango-Lasprilla, C. Uribe, C. Valencia, and F. Lopera, "Linguistic changes in verbal expression: a preclinical marker of Alzheimer's disease," *Journal of the International Neuropsychological Society*, vol. 13, no. 3, pp. 433–439, May 2007.

[8] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, Jan. 2000.

[9] S. Ahmed, A.-M. F. Haigh, C. A. de Jager, and P. Garrard, "Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease," *Brain*, vol. 136, no. 12, pp. 3727–3737, Dec. 2013.

[10] V. Boschi, E. Catricalà, M. Consonni, C. Chesi, A. Moro, and S. Cappa, "Connected Speech in Neurodegenerative Language Disorders: A Review," *Frontiers in Psycholigy*, 2017.

[11] 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, Jan. 2016.

[12] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 27–37.

[13] B. Peintner, W. Jarrold, D. Vergyri, C. Richey, M. L. G. Tempini, and J. Ogar, "Learning diagnostic models using speech and language measures," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2008, pp. 4648–4651.

[14] A. Khodabakhsh, F. Yesil, E. Guner, and C. Demiroglu, "Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 9, Mar. 2015.

[15] M. Markaki and Y. Stylianou, "Voice Pathology Detection and Discrimination Based on Modulation Spectral Features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1938–1948, 2011.

[16] C. Poellabauer, N. Yadav, L. Daudet, S. L. Schneider, C. Busso, and P. J. Flynn, "Challenges in Concussion Detection Using Vocal Acoustic Biomarkers," *IEEE Access*, vol. 3, pp. 1143–1160, 2015.

[17] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert, and R. David, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 1, pp. 112–124, 2015.

[18] F. Martínez-Sánchez, J. J. G. Meilán, J. Carro, and O. Ivanova, "A Prototype for the Voice Analysis Diagnosis of Alzheimer's Disease," *Journal of Alzheimer's Disease*, vol. 64, no. 2, pp. 473–481, Jan. 2018.

[19] 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.

[20] A. S. Gulapalli and V. K. Mittal, "Detection of Alzheimer's Disease Through Speech Features and Machine Learning Classifiers," vol. 1, pp. 825–840, 2022.

[21] A. Shimoda, Y. Li, H. Hayashi, and N. Kondo, "Dementia risks identified by vocal features via telephone conversations: A novel machine learning prediction model," *PLOS ONE*, vol. 16, no. 7, Jul. 2021.

[22] A. Ablimit, C. Botelho, A. Abad, T. Schultz, and I. Trancoso, "Exploring Dementia Detection from Speech: Cross Corpus Analysis," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Singapore, Singapore: IEEE, May 2022, pp. 6472–6476.

[23] 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 *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Singapore, Singapore: IEEE, May 2022, pp. 6467–6471.

[24] N. Cummins, Y. Pan, Z. Ren, J. Fritsch, V. S. Nallanthighal, H. Christensen, D. Blackburn, B. Schuller, M. M. Doss, H. Strik, and A. Harma, "A Comparison of Acoustic and Linguistics Methodologies for Alzheimer's Dementia Recognition," *Proceedings INTERSPEECH*, pp. 2182–2186, Oct. 2020.

[25] F. Haider, S. de la Fuente Garcia, and S. Luz, "An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, pp. 272–281, Feb. 2020.

[26] S. Mirzaei, M. A. E. Yacoubi, S. Garcia-Salicetti, J. Boudy, C. Kahindo, V. Cristancho-Lacroix, H. Kerhervé, and A.-S. Rigaud, "Two-Stage Feature Selection of Voice Parameters for Early Alzheimer's Disease Prediction," *Innovation and Research in BioMedical engineering*, vol. 39, no. 6, pp. 430–435, 2018.

[27] 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.

[28] M. S. Rafii, C. Taylor, A. Coutinho, K. Kim, and D. Galasko, "Comparison of the Memory Performance Index With Standard Neuropsychological Measures of Cognition," *American Journal of Alzheimer's Disease  Other Dementias*, vol. 26, no. 3, pp. 235–239, May 2011.

[29] W. R. Shankle, T. Mangrola, T. Chan, and J. Hara, "Development and validation of the Memory Performance Index: reducing measurement error in recall tests," *Alzheimer's & Dementia*, vol. 5, no. 4, pp. 295–306.

[30] S. Luz, F. Haider, S. d. l. Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge," *Proceedings INTERSPEECH*, 2020.

[31] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[32] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, 1990.

[33] A. Khosravani, C. Glackin, N. Dugan, G. Chollet, and N. Cannings, "The Intelligent Voice 2016 Speaker Recognition System," *CoRR*, vol. abs/1611.00514, 2016. [Online]. Available: http://arxiv.org/abs/1611.00514

[34] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[35] K. Maekawa, "Corpus of spontaneous Japanese: its design and evaluation," in *Proceedings ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*. Tokyo, Japan: ISCA & IEEE, 2003.

[36] A. Canavan and G. Zipperlen, "Callhome japanese speech," *LDC96S37. Web Download*, 1996.

[37] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[38] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme Recognition Using Time-Delay Neural Networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[39] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1992, pp. 517–520 vol.1.

[40] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "The Mixer 6 corpus: Resources for cross-channel and text independent speaker recognition," in *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, 2010.

[41] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 2-3, pp. 225–254, 2000.

[42] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.