Interspeech 2024
1-5 September 2024, Kos, Greece

Leveraging Universal Speech Representations for Detecting and Assessing the Severity of Mild Cognitive Impairment Across Languages

Anna Favaro[1], *Tianyu Cao*[1], *Najim Dehak*[1], *Laureano Moro-Velazquez*[1]

[1]Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, USA

afavaro1@jhu.edu

Abstract

This study examines the suitability of language-agnostic features for automatically detecting Mild Cognitive Impairment (MCI) and predicting Mini-Mental State Examination (MMSE) scores in a multilingual framework. We explored two methods for feature extraction: traditional feature engineering and pre-trained feature representation. We developed our models using the Interspeech 2024 Taukadial challenge data set, containing audios from subjects with MCI and controls in Chinese and English. Our top ensemble model achieved 75% accuracy in MCI detection and an RMSE of 2.44 in MMSE prediction in the testing set. Our results reveal the complementary nature of acoustic and linguistic representations and the existence of *universal* features that can be used *cross-lingually*. However, a statistical analysis of interpretable features did not show any shared speech patterns between the two languages, which can be attributed to differences in disease severity between the two cohorts of participants.

Index Terms: Speech biomarkers, neurodegenerative diseases, cognitive assessment, computational paralinguistics

1. Introduction

Mild Cognitive Impairment (MCI) is defined as cognitive decline greater than expected for an individual's age and educational attainment [1]. Researchers consider this period a window of opportunity for early detection of dementia because around 38% of patients with MCI develop a type of dementia within 5 years [2]. As speech can be affected in the early stages of cognitive decline, a promising solution for early diagnosis is language analysis [3].

Recently, there has been a growing interest in automatically analyzing linguistic and acoustic speech patterns, separately or in combination, to create more sensitive quantitative assessments of MCI. Fraser et al. [4] combined Mini-Mental State Exam (MMSE) scores with linguistic features derived from narratives to identify MCI within a Swedish cohort, achieving a classification accuracy of 87%. Mizuguchi et al. [5] aimed to distinguish MCI speakers from cognitively normal controls (CNs) using 17 acoustic features, such as pitch-related statistics and voice quality metrics (e.g., shimmer, jitter, and harmonics-to-noise ratio), yielding an accuracy of 66.7%. Various neural techniques have also been explored, including transformer-based pre-trained models like BERT, GPT, and HuBERT to address MCI detection [6, 7, 8, 9]. Combining different modalities at once, Calzà et al. [10] utilized 87 acoustic, rhythmical, morpho-syntactic, and lexical features, along with readability indexes and demographic data to differentiate CNs from subjects with MCI, all native speakers of Italian. The best classifier achieved F1 scores of around 75%. Similarly, by fusing duration-based, acoustic, and linguistic features, [11] achieved an accuracy of 81.9% from MCI and CNs subjects, all native speakers of Mandarin. Tang et al. [12] proposed a method to combine the linguistic (LIWC) and acoustic (MFCCs) features, and concluded that the detection of MCI using the combination of both features outperformed using one feature type alone.

Among the few studies on MCI conducted so far, the majority only considered language-dependent hand-crafted features. Furthermore, the exploration of multilingual MCI detection remains limited, with only a handful of studies delving into the identification of features that transcend language barriers and generalize across idioms [13, 14]. Yet, identifying such language-independent features is pivotal for establishing a universal framework for MCI detection and monitoring. This study fills a gap in the existing literature by combining multilingual speech analysis with MCI detection and prediction of cognitive scores. In particular:

1. We assess and compare the suitability of both traditional hand-crafted features and cutting-edge neural representations to the task of MCI detection and MMSE prediction in two major languages, Chinese and English. In doing so, we focus on approaches that are language-independent or build on comparable features.
2. We conduct an in-depth analysis of the acoustic and linguistic traits characterizing the Chinese and the English cohorts by conducting a statistical analysis. In doing so, we aim to identify shared speech patterns of cognitive decline (if any) between the two languages.[1]

2. The Taukadial Challenge

In this paper, we address the two tasks proposed by the Interspeech 2024 Taukadial challenge organizers:

- *Classification task*: automatic differentiation between participants with and without MCI using spontaneous speech recordings.
- *Regression task*: automatic prediction of participants' MMSE scores using the same data set used in the classification task.

2.1. The Taukadial Data Set

The Taukadial challenge data set is described in [15]. The training data set consisted of spontaneous speech recordings of picture descriptions produced by 55 CNs and 74 speakers with MCI. It contained both Chinese and English samples with three picture descriptions per participant. The test set comprised recordings from 21 speakers with MCI and 19 CNs, maintaining the same mix of languages and picture descriptions.

[1]The link to the code repository can be found here: https://github.com/Annafavaro/TAUKADIAL-2024.git.

10.21437/Interspeech.2024-2030

3. Methods

This section describes the features, models, and ensemble techniques used in the classification and regression experiments.

3.1. Acoustic Features

In this subsection, the acoustic features used to characterize the prosodic and articulatory abilities of the speakers are presented. These features encompass both interpretable hand-crafted features and non-interpretable characteristics, respectively.

3.1.1. Interpretable Features

Pauses. The behavior of pause-related features in subjects affected by MCI displays inconsistencies across studies emphasizing the need for further investigation into their reliability and task dependency. Some studies report that MCI subjects produce more pauses at a higher rate than CNs [2], although pause rate does not differ in other studies [16] and neither does the number of pauses [17]. Moreover, while some studies show that subjects with MCI produce pauses with longer duration than CNs [2], others do not report any significant difference between experimental groups [17]. In this work, we use DigiPsychProsody[2] to compute pause-related features. These include total speech time, total pause time, percentage pause time, speech pause time, mean pause duration, and pause variability, for a total of 21 features.

Intensity and formants. To model the alterations in articulation that can be related to emerging difficulties in lexical access [18], we extract descriptors (e.g., mean, std, median) related to the first four formats, and vocal tract-related measures [19]. Furthermore, pitch and intensity-related measures are extracted such as F0 (std) and intensity (std), for a total of 59 features.[3]

3.1.2. Non-interpretrable Features

Mel-Frequency Cepstral Coefficients. MFCCs were originally developed for speech recognition and have found diverse uses as voice descriptors. We extract 1400 MFCC-related descriptors using the openSMILE feature extraction library. These descriptors constitute a subset of the ComParE feature set.

VGGish. This is a feature embedding front-end for audio classification models. We use a pre-trained model trained on the AudioSet data set [20].

Wav2vec. The wav2vec features are extracted from the models pretrained using 53 and 128 languages, respectively [21]. We segment long recordings into $10\,\mathrm{s}$ segments to extract the features. The model extracts an embedding for every $20\,\mathrm{ms}$ within each $10\,\mathrm{s}$ audio segment at each layer. A 1-D feature vector of size 512 representing each audio segment is obtained by calculating the mean of the embeddings along the time axis. The final embedding for each recording is set as the average of the feature vectors extracted from its $10\,\mathrm{s}$ segments.

Whisper. The whisper features are extracted with the whisper-large-v2 model pretrained in 99 languages [22]. The fixed-dimensional encoder features obtained at the end of the encoder module have sizes of $1 \times 1500 \times 1280$ (batch, time, features). A $1-$D feature vector of size 1280 representing each audio segment is obtained by computing the mean of the embeddings along the time axis.

3.2. Linguistic Features

We explore the effectiveness of various linguistic descriptors, both interpretable and non-interpretable, to capture the speakers' semantic, lexical, and syntactic expressions.

3.2.1. Interpretable Features

All the interpretable features are extracted using the Linguistic Feature Toolkit [23], a Python package that computes various handcrafted features commonly used in computational linguistics. Only the 129 linguistic features that are universally applicable across languages are considered. To extract these features, we use the spaCy's pre-trained pipeline available for English[4] and Chinese[5]. These features can be categorized as follows.

Surface. The surface feature set contains features that do not belong to any specific linguistic branch (e.g., semantics, syntax). They are features such as the total number of characters, the total number of words, and the total number of sentences.

Syntactic. To represent syntactic complexity, features based on part of speech are considered. These features represent the total and unique number of various parts of speech in a text. Some examples are the total number of nouns, auxiliaries, proper nouns, conjunctions, and verbs. These descriptors are extracted considering the whole paragraph, per sentence, and per word.

Lexical-semantics. Besides the standard type-token ratio metric, to capture the richness of the vocabulary in a text, we use normalized metrics less sensitive to the length of the text (speech transcription), such as root type-token ratio, corrected type-token ratio, logarithmic type-token ratio, and uber type-token ratio.

3.2.2. Non-interpretrable Features

We employ linguistic language-agnostic embeddings extracted from pre-trained language models encompassing a wide range of languages, from ten to a hundred. [6]

Distilbert-base-multilingual-cased. This model, called DistilmBERT, is a distilled version of the BERT base multilingual model. It is trained on the concatenation of Wikipedia in 104 different languages. The model has 6 layers, 768-D, and 12 heads, totaling 134M parameters.

Text2vec-base-multilingual. This is a cosine sentence model. It is fine-tuned on 9 languages in a sentence similarity task. It maps sentences to a 384-D dense vector space.

XLM-RoBERTa-base. This is a multilingual version of RoBERTa [24]. It is pre-trained on $2.5\,\mathrm{TB}$ of filtered CommonCrawl data containing 100 languages. The pre-trained model learned an inner representation of 100 languages that can then be used to extract features useful for downstream tasks.

XLM-Roberta-Large-Vit-L-14. This is a multilingual-CLIP [25] text encoder, trained using prompts from 48 languages along with a visual encoder to maximize the similarity of (image, text) pairs via a contrastive loss.

LaBSE. The language-agnostic BERT Sentence Embedding [26] encodes text into high dimensional vectors. This model is trained and optimized to produce similar representations for bilingual sentence pairs that are translations of each other. It can be used to map 109 languages to a shared vector space.

Lealla. This is a lightweight language-agnostic sentence embedding model supporting 109 languages, distilled from LaBSE [27]. It can be used for multilingual sentence embeddings and bi-text retrieval.

Multilingual-e5-large. This model is initialized from

[2] `https://github.com/wiseman/py-webrtcvad`

[3] The following library implemented in Parselmouth is used: `https://github.com/uzaymacar/simple-speech-features`.

[4] `https://spacy.io/models/en`

[5] `https://spacy.io/models/zh`

[6] All these models are accessible via the HuggingFace library: `https://huggingface.co/`.

XLM-RoBERTa-large and continually trained on a mixture of multilingual data sets [28]. It supports 100 languages from XLM-RoBERTa, but low-resource languages may see performance degradation. This model has 24 layers and outputs embedding of 1024-D.

3.3. Classification, Regression and Statistical Analysis

We use a single linear layer with a sigmoid activation function for classification with both interpretable and non-interpretable features. For regression with interpretable features, we use Support Vector Regressor (SVR), XGB Regressor (XGBR), and Bagging Regressor (BAGR). For regression with non-interpretable features instead, we use two linear layers and one output with a linear activation function. After the different classification models are trained using their respective features, we perform score fusion as follows. Firstly, given that each speaker has three recordings, we consider individual predictions from each classifier trained with one of the features presented in section 3, and apply a *speaker-wise score average*. If this average is below 0.5, the three predictions for that speaker are labeled negative; otherwise, they are labeled positive. This decision is justified by the fact that in this cohort a speaker cannot have two recordings belonging to different classes. Then, to perform the fusion of different modalities (acoustic with linguistic models, for instance), we utilize either majority voting (MV) or the CONSEN algorithm [29]. For regression model fusion, we use either the average of the MMSE scores across models or apply the CONSEN algorithm.

For both regression and classification, two types of training-testing strategies are considered:

- *Cross-validation*: performed by training and testing with the training subset, using a 10-fold scheme where class distributions were consistent over the folds. To avoid data leakage, we ensure that training and testing subsets do not contain recordings from the same speakers.
- *Prediction on the Taukadial test set*: performed by testing the models trained in the training subset with the evaluation subset. For each separate approach, we propagate the evaluation subset through an ensemble classifier that averages the scores from the 10 cross-validation models for a certain set-up.

In addition, we perform a statistical analysis using the interpretable features. To do so, the non-parametric pair-wise Mann-Whitney U tests of significance are performed using the training data[7], considering the two languages separately. As in the multilingual study of [30] for Parkinson's Disease, this analysis is conducted to determine whether there are significant differences between the MCI and the CN groups for each feature and to determine the existence of common speech patterns of cognitive decline (if any) across the two languages. We apply the Benjamini- Hochberg correction to control the False Discovery Rate (FDR).

4. Experimental Pipeline

The overall experimental pipeline can be summarized as follows:

1. *Acoustic Pre-processing*: Original recordings are resampled at 16 kHz, necessary for feature extraction algorithms. Additionally, Root Mean Square loudness normalization is applied.
2. *Language Identification*: Whisper large-v2 model is used to identify the language of the recordings.
3. *Automatic-Speech-Recognition (ASR)*: Whisper large-v2 model is utilized for ASR. English and Chinese prompts are created to provide context, incorporating punctuations, disfluencies, and hesitations to enhance transcription accuracy.
4. *Feature Extraction*: Various language-agnostic features are extracted as outlined in section 3.
5. *Training-evaluation*: Two approaches are used for training and evaluation. 1) *Monolingual*: The same acoustic and/or linguistic features are extracted for both languages, and the models are developed and evaluated separately for English and Chinese. 2) *Multi-lingual*: The same acoustic and/or linguistic features are extracted for both languages, and the models are trained using data from both languages simultaneously.

[7]The training set was exclusively employed due to the unavailability of test labels at the time of the current submission.

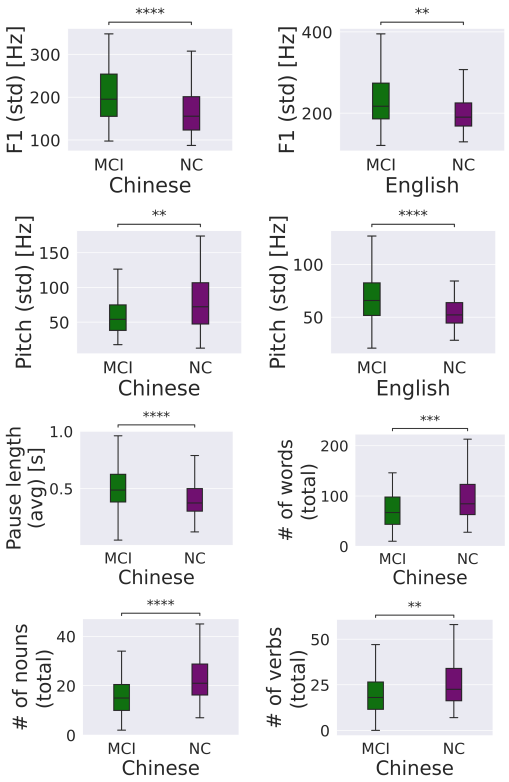

*Figure 1: Boxplots of some significant features from the statistical analysis reported on the Chinese and English training set. Asterisks highlight statistically significant differences in the median between groups, where * means $0.01 < p \leqslant 0.05$, ** means $0.001 < p \leqslant 0.01$, *** means $0.0001 < p \leqslant 0.001$, and **** means $p \leqslant 0.0001$.*

5. Results and Discussion

Experiments were performed on various combinations of extracted features and developed models. For the challenge submission, we selected the models or combinations that yielded the best results in cross-validation. Table 1 shows the results of the models submitted to the challenge for the classification and regression tasks, respectively[8]. For the classification task, our best ensemble achieved a test-data accuracy of 75% while fusing Text2vec-base-multilingual, LaBSE, Lealla, and Whisper. For the regression task, our best ensemble reached an RMSE score of 2.44 when combining interpretable linguistic features,

[8]Due to page limitations, only the performance of the submitted models is reported.

pause-related features, Multilingual-e5-large, and XLM-Roberta-Large-Vit-L-14. All the results reported are obtained adopting a monolingual training-evaluation scheme, except for the fifth ensemble, in which Whisper and XLM-RoBERTa-base models follow a multilingual training-evaluation scheme. On average, evaluation results tended to be higher than those obtained via cross-validation, possibly due to the use of ensemble models obtained from 10 cross-validation models for each submitted approach, mitigating overfitting.

Classification			
Feature	**Ensemble**	**CV (Acc)**	**Test (Acc)**
Baseline [15]	–	–	0.59
LaBSE	MV	0.66	0.67
DistilmBERT	MV	0.66	0.70
Lealla, Wav2vec53, LaBse	MV	0.67	0.65
Text2vec-base-multilingual, LaBSE, Lealla, Whisper	MV	0.70	**0.75**
Ling., MFCCs, Whisper, LaBSE, Lealla, VGGish, Whisper*, XLM-RoBERTa-base*	CONSEN	**0.72**	0.68
Regression			
Feature	**Ensemble**	**CV (RMSE)**	**Test (RMSE)**
Baseline [15]	–	–	2.89
Ling. (SVR), Pause (XGR, SVR, BAGR)	Avg.	3.06	2.62
Ling. (XGR), Pause (XGR, SVR, BAGR), Multilingual-e5-large, XLM-Roberta-Large-Vit-L-14	Avg.	3.06	**2.44**
Pause (XGR), Intensity (XGR), Ling. (SVR)	Avg.	2.99	2.62
MFCCs (SVR), Ling. (XGR, RFR), XLM-Roberta-Large-Vit-L-14, Multilingual-e5-large, Wav2vec128	Avg.	2.87	2.93
The same ensemble used for prediction	CONSEN	**2.77**	2.47

*Table 1: The upper section of the table reports the results obtained in the MCI detection task during Cross-Validation (CV) and on the Taukadial test set in terms of accuracy (Acc). The lower section of the table reports the results obtained in the MMSE regression task during CV and on the test set in terms of Root Mean Squared Error (RMSE). When not stated otherwise, model performance was obtained following a monolingual training-evaluation scheme. If a multilingual evaluation was adopted, * is reported next to the model name. Abbreviations: MV: majority voting; Ling., linguistic (interpretable) features.*

From these experiments, several key observations emerged.

1. In classification, models leveraging semantic embeddings, such as DistilmBERT, surpassed most acoustic feature-based models and some of the ensembles on both validation and testing sets. This is consistent with studies suggesting that lexical-semantic processing is early affected in AD, making it a crucial target for MCI detection and prognosis [31]. Conversely, articulatory, phonological, and syntactic aspects of language production typically remain relatively intact until the later stages of AD.
2. Fusion with majority voting of linguistic and acoustic representations yielded the highest classification accuracy on the test set, underscoring the generalizability and complementarity of these features. Similarly, in the regression task, the lowest RMSE was achieved using a combination of interpretable and non-interpretable linguistic and acoustic representations. Notably, our observations indicate that while neural embeddings excel in classification, relying solely on interpretable features (linguistic in particular) can yield comparable, and in certain instances, superior performance to neural representations or ensemble methods in regression, as reported by [15].
3. Interestingly, our experiments revealed that a monolingual training-evaluation scheme tended to yield higher classification accuracies compared to training models with data from both languages simultaneously. A difference in the severity of cognitive impairment could explain the observed differences between the Chinese and English cohorts in the data set, as shown by the higher mean of the MMSE scores reported for the Chinese MCI group (23.45 ± 4.34) compared to their English counterpart (27.77 ± 1.24). Furthermore, when considering both subjects with MCI and CNs, the range of values of the MMSE scores in the English partition is very narrow ($\min$: 25, $\max$: 30), making the regression task more challenging than in the Chinese partition.

To identify the most salient features in differentiating between subjects with MCI and CNs, we performed a statistical analysis using the interpretable features extracted from the Chinese and English training splits, respectively. Among the 80 acoustic features, 27 were significant on the Chinese split, while only 7 reached significance on the English split. Notably, only two features—F1 and F0 standard deviation—were significant in both languages simultaneously (see Figure 1). The former was significantly higher for the MCI group on both languages and was the only interpretable feature displaying a robust cross-lingual behavior. The latter instead displayed an opposite behavior on the English split and the Chinese split for the MCI group compared to the CN group (see first two boxplots in Figure 1), which indicates a lack of robust cross-lingual behavior. Furthermore, while 30 linguistic features, such as the number of words, nouns, and verbs, were significant in the Chinese split (see Figure 1), none of the linguistic features considered reached significance in the English split after FDR correction. Overall, the lack of shared patterns between the two languages and the inconsistencies reported in the severity distributions motivate why better results can be achieved when adopting a monolingual training-evaluation scheme and when considering the Chinese partition of the data set by itself.

6. Conclusions and Future Work

This study presents a general processing framework for multilingual cognitive assessment, encompassing language-agnostic features alongside comparable linguistic features tailored to Chinese and English. The study's primary objective was to identify universal representations that can be used to detect MCI and predict MMSE scores across languages. The acoustic features adopted consist of speaker and speech recognition embeddings, prosodic, and articulatory descriptors, whereas the linguistic ones are lexical and syntactic features as well as semantic embeddings extracted from multilingual language models previously trained on various downstream tasks. Our findings suggest that acoustic and linguistic approaches contain complementary information for automatic detection and assessment of MCI. Namely, in both classification and regression, the best performance was achieved by ensembles of acoustic and linguistic models. Notably, interpretable linguistic features demonstrated superior performance in regression tasks as reported by [15], while non-interpretable features (and their combinations) outperformed interpretable ones in classification. Lastly, the statistical analysis reveals almost the absence of common speech patterns of cognitive decline between the two languages. This discrepancy likely stems from the advanced disease stages of the Chinese MCI group, contributing to distinct linguistic and acoustic manifestations in the two cohorts. In the future, we aim to explore multi-modal approaches that leverage aligned linguistic and acoustic information to extract more nuanced cues and harness the complementary nature of both modalities.

7. References

[1] R. C. Petersen, B. Caracciolo, C. Brayne, S. Gauthier, V. Jelic, and L. Fratiglioni, "Mild cognitive impairment: a concept in evolution," *Journal of internal medicine*, vol. 275, no. 3, pp. 214–228, 2014.

[2] D. Beltrami, G. Gagliardi, R. Rossini Favretti, E. Ghidoni, F. Tamburini, and L. Calzà, "Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline?" *Frontiers in aging neuroscience*, vol. 10, p. 369, 2018.

[3] J. E. Karr, R. B. Graham, S. M. Hofer, and G. Muniz-Terrera, "When does cognitive decline begin? a systematic review of change point studies on accelerated decline in cognitive and neurological outcomes preceding mild cognitive impairment, dementia, and death." *Psychology and aging*, vol. 33, no. 2, p. 195, 2018.

[4] K. C. Fraser, K. Lundholm Fors, M. Eckerström, C. Themistocleous, and D. Kokkinakis, "Improving the sensitivity and specificity of mci screening with linguistic information," in *Proceedings of the LREC 2018 Workshop "Resources and processing of linguistic, para-linguistic and extra-linguistic data from people with various forms of cognitive/psychiatric impairments (RaPID-2)"(2015)*, 2018, pp. 19–26.

[5] D. Mizuguchi, T. Yamamoto, Y. Omiya, K. Endo, K. Tano, M. Oya, and S. Takano, "Novel screening tool using non-linguistic voice features derived from simple phrases to detect mild cognitive impairment and dementia," *JAR life*, vol. 12, p. 72, 2023.

[6] B. Mirheidari, Y. Pan, D. Blackburn, R. O'Malley, and H. Christensen, "Identifying cognitive impairment using sentence representation vectors." in *Interspeech*, 2021, pp. 2941–2945.

[7] A. P. Fard, M. H. Mahoor, M. Alsuhaibani, and H. H. Dodgec, "Linguistic-based mild cognitive impairment detection using informative loss," *arXiv preprint arXiv:2402.01690*, 2024.

[8] B. S. Runde, A. Alapati, and N. G. Bazan, "The optimization of a natural language processing approach for the automatic detection of alzheimer's disease using gpt embeddings," *Brain Sciences*, vol. 14, no. 3, p. 211, 2024.

[9] E. Kurtz, Y. Zhu, T. Driesse, B. Tran, J. A. Batsis, R. M. Roth, and X. Liang, "Early detection of cognitive decline using voice assistant commands," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] L. Calzà, G. Gagliardi, R. Rossini Favretti, and F. Tamburini, "Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia," *Computer Speech & Language*, vol. 65, p. 101113, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230820300462

[11] Z. Liu, Z. Guo, Z. Ling, S. Wang, L. Jin, and Y. Li, "Dementia detection by analyzing spontaneous mandarin speech," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 289–296.

[12] F. Tang, J. Chen, H. H. Dodge, and J. Zhou, "The joint effects of acoustic and linguistic markers for early identification of mild cognitive impairment," *Frontiers in digital health*, vol. 3, p. 702772, 2022.

[13] K. C. Fraser, K. L. Fors, and D. Kokkinakis, "Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment," *Computer Speech & Language*, vol. 53, pp. 121–139, 2019.

[14] H. Lindsay, P. Müller, I. Kröger, J. Tröger, N. Linz, A. Konig, R. Zeghari, F. R. Verhey, and I. H. Ramakers, "Multilingual learning for mild cognitive impairment screening from a clinical speech task," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, R. Mitkov and G. Angelova, Eds. Held Online: INCOMA Ltd., Sep. 2021, pp. 830–838. [Online]. Available: https://aclanthology.org/2021.ranlp-1.95

[15] S. Luz, S. d. l. F. Garcia, F. Haider, D. Fromm, B. MacWhinney, A. Lanzi, Y.-N. Chang, C.-J. Chou, and Y.-C. Liu, "Connected speech-based cognitive assessment in chinese and english." arXiv, 2024, dOI: 10.48550/ARXIV.2404.nnnnn (TBA).

[16] L. Tóth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatlóczki, Z. Bánréti, M. Pákáski, and J. Kálmán, "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, 2018.

[17] R. A. Sluis, D. Angus, J. Wiles, A. Back, T. Gibson, J. Liddle, P. Worthy, D. Copland, and A. J. Angwin, "An automated approach to examining pausing in the speech of people with dementia," *American Journal of Alzheimer's Disease & Other Dementias®*, vol. 35, p. 1533317520939773, 2020.

[18] J. J. Meilán, F. Martínez-Sánchez, I. Martínez-Nicolás, T. E. Llorente, J. Carro *et al.*, "Changes in the rhythm of speech difference between people with nondegenerative mild cognitive impairment and with preclinical dementia," *Behavioural neurology*, vol. 2020, 2020.

[19] W. T. Fitch, "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques," *The Journal of the Acoustical Society of America*, vol. 102, no. 2, pp. 1213–1222, 1997.

[20] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[21] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[23] B. W. Lee and J. Lee, "LFTK: Handcrafted features in computational linguistics," in *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1–19. [Online]. Available: https://aclanthology.org/2023.bea-1.1

[24] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *CoRR*, vol. abs/1911.02116, 2019. [Online]. Available: http://arxiv.org/abs/1911.02116

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[26] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," *arXiv preprint arXiv:2007.01852*, 2020.

[27] Z. Mao and T. Nakagawa, "Lealla: Learning lightweight language-agnostic sentence embeddings with knowledge distillation," 2023.

[28] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Improving text embeddings with large language models," 2024.

[29] L. Jin, Y. Oh, H. Kim, H. Jung, H. J. Jon, J. E. Shin, and E. Y. Kim, "Consen: Complementary and simultaneous ensemble for alzheimer's disease detection and mmse score prediction," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–2.

[30] A. Favaro, L. Moro-Velázquez, A. Butala, C. Motley, T. Cao, R. D. Stevens, J. Villalba, and N. Dehak, "Multilingual evaluation of interpretable biomarkers to represent language and speech patterns in parkinson's disease," *Frontiers in Neurology*, vol. 14, p. 1142642, 2023.

[31] V. Taler and N. A. Phillips, "Language performance in alzheimer's disease and mild cognitive impairment: a comparative review," *Journal of clinical and experimental neuropsychology*, vol. 30, no. 5, pp. 501–556, 2008.