TAUKADIAL: Speech-Based Cognitive Assessment in Chinese and English

Saturnino Luz[1], *Sofia De La Fuente Garcia*[2], *Fasih Haider*[3], *Davida Fromm*[4], *Brian MacWhinney*[4], *Alyssa Lanzi*[5], *Ya-Ning Chang*[6], *Chia-Ju Chou*[7], *Yi-Chien Liu*[7]

[1]Usher Institute, Medical School, [2]School of Health in Social Science, [3]School of Engineering, University of Edinburgh, UK
[4]Department of Psychology, Carnegie Mellon University, [5]Communication Sciences & Disorders, University of Delaware, USA
[6]Miin Wu School of Computing National Cheng Kung University, [7]Department of Neurology Cardinal Tien Hospital, Taipei, Taiwan
s.luz@.ed.ac.uk

Abstract

We present a novel benchmark data set and prediction tasks for the investigation of approaches to assess cognitive function through analysis of spontaneous speech. The data set consists of speech samples and clinical information for speakers of Mandarin Chinese and English with different levels of cognitive impairment as well as individuals with normal cognition. These data have been carefully matched by age and sex by propensity score analysis to ensure balance and representativity in model training. The prediction tasks encompass classification (diagnosis) of speakers as having normal cognition or mild cognitive impairment, and prediction of cognitive test scores. This cognitive health assessment framework was designed to encourage the development of approaches to speech-based cognitive assessment which generalise across languages. We illustrate it by presenting baseline prediction models that employ language-agnostic and comparable features for diagnosis and cognitive test score prediction.

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

1. Introduction

Cognitive problems, such as memory loss, speech and language impairment, and reasoning difficulties, occur frequently among older adults and often precede the onset of dementia syndromes. Due to the high prevalence of dementia and the costs this implies to health systems worldwide [1], research into cognitive impairment for the purposes of dementia prevention and early detection has become a priority in healthcare. There is a need for cost-effective and scalable methods for assessment of cognition and detection of impairment, from its most subtle forms to severe manifestations of dementia. Speech is an easily collectable behavioural signal which reflects cognitive function, and therefore could potentially serve as a digital biomarker of cognitive function, presenting a unique opportunity for application of speech technology [2].

We aim to explore speech as a marker of cognition in a global health context by investigating its application to modelling cognitive health indicators in two major languages, namely, Chinese and English. In this paper, we focus on prediction of cognitive test scores and diagnosis of mild cognitive impairment (MCI) in older speakers of Chinese and English, using samples of connected speech. We are particularly interested in investigating approaches that are language independent or build on comparable features. To this end, we have created, and are sharing with the research community, a data set comprising recorded speech from study participants carrying out picture description tasks along with clinical and neuropsychological test data.

This data set can be used as a benchmark for speech processing and machine learning tasks that are relevant to the detection of cognitive decline through analysis of connected speech data. We hope that this new resource will stimulate research on speech biomarkers among members of the speech, signal processing, machine learning, natural language processing and biomedical research communities, enabling them to test existing methods or develop novel approaches on a new, standardised dataset which will remain available to the community for future research and replication of results.

2. Background

The field of speech-based approaches to detecting cognitive decline has grown considerably over the last two decades, with a major focus on detecting dementia or Alzheimer's dementia (AD), in comparison to a control (normal cognition, NC) group [2]. A smaller proportion of studies has focused on MCI detection. Most available studies, however, either report relatively high levels of accuracy but on imbalanced datasets (where accuracy is a biased measure), or lower levels of accuracy on more balanced datasets. Mirzaei et al. [3] report 62%, 3-way classification accuracy discriminating HC, MCI, and AD on a class-balanced dataset. Many studies report accuracy figures without class-balance, which can lead to findings that are difficult to interpret and compare. For example, Nasrolahzadeh et al. [4] report 97.71%, 4-way classification accuracy in a highly imbalanced dataset while Mirheidari et al. [5] reported 62%, 4-way classification accuracy in a more balanced dataset.

Clinical tests, such as the mini-mental state examination (MMSE), are often part of these studies as a mere data descriptor, but rarely used in speech-based prediction. Some studies [6, 7, 8, 9] have used MMSE results as a baseline for classification, against which to compare the speech-based classifier, but very few available studies go beyond classification and use speech-based approaches to predict MMSE scores, or similar cognitive tests.

It should be noted that MMSE has been criticised for low discrimination, especially in preclinical dementia [10]. There has been a shift of focus toward prediction of cognitive scores in recent years. For instance, [11] extracted a set of lexical-semantic features known to be affected in dementia from picture description tasks. They used these features to build a predictive model able to explain 51% of the variance of cognitive scores (3MS) at the time of speech collection, and 56% of cognitive scores in a 12-month follow-up. Another study used participants' recall of a childhood memory to predict their cognitive scores, and explained up to 16.52% variance with a covariance statistic test [12]. Artificial intelligence approaches have also been published, such as [13], which used BERT to predict

MMSE scores from denoised speech recordings from picture description tasks and report a root mean squared error (RMSE) of 3.76. Another study using speech samples from picture descriptions, but from a class, gender and age balanced dataset reported that acoustic features alone predict MMSE scores with a mean absolute error (MAE) of 5.66 and an R^2 of 0.125, with a linear regression analysis, which improved by adding age, sex and years of education to the model, yielding a MAE = 4.97 and R^2 = 0.261 [14].

Speech data are most often obtained from tasks embedded in neuropsychological batteries (e.g. verbal fluency, story recall, picture description). This is the case in our study, where for both English and Chinese, speech samples come from participants' picture description tasks conducted as part of cognitive assessment.

A multilingual study on the AZTIAHO database reported a range of accuracy between 60% and 93.79%, using only *ad hoc* acoustic features. AZTIAHO contains speech samples in English, French, Spanish, Catalan, Basque, Chinese, Arabian, and Portuguese, but it is a small dataset (40 participants) remarkably imbalanced in terms of age (25% of their control group is between 20 and 60 years old, whilst 100% of their dementia group is over 60 years old). It also presents a class imbalance, since there are 20 participants in each group, but the control group is homogeneous whereas the AD group with three different severity stages, with 4, 10 and 6 participants respectively [15].

Haider et al. [16] also used acoustic features only, and reported 78.7% accuracy generated from standardised feature sets that had been developed for computational paralinguistics. Their dataset is much larger (164 participants) and it is balanced for class, age and gender. In an imbalanced version of the same dataset, and using text-based features only, [17] obtained 85.4% accuracy. Neither of these studies addresses multilingualism.

Fraser et al. [18] used English and Swedish speech samples, also generated through picture description tasks, and word embeddings to train and test models to classify MCI and NC subjects in both languages. They obtained classification accuracy of 63% for English and 72% for Swedish. More recently, a signal processing grand challenge addressed the issue of generalising speech-based predictive models across two languages: Greek and English [19]. Differently from our experimental setting, theirs involved training of models in one language and testing on another. The top performing systems' had classification accuracy between 69% to 87% (AD vs NC), and MMSE score prediction errors RMSE between 4.79 and 3.72.

SL: Hi Sofia, here you had English and French, but the Fraser paper you referred to was English and Swedish. Perhaps you had a different paper in mind?

3. Data

The data set consists of Chinese and English speech samples collected while the speakers participated in picture description tasks conducted as part of cognitive assessment.

English-speaking participants were recruited from a community in the United States through print and online advertisements targeted to adults aged 60-90 with memory concerns. Eligible participants were at least 60 years old, spoke and understood English, had adequate hearing and vision to participate in a telehealth session, were stable on or not taking nootropic medications, and had a negative self-reported history of major psychiatric disorder or other medical disorder/illness that could cause cognitive decline (e.g., traumatic brain injury). Participants were classified as either neurotypical (NC) or MCI. To be classified as MCI, a neuropsychologist determined that participants met the following National Institute on Aging-Alzheimer's Association (NIA-AA) criteria [20]: (a) self-reported a decline in cognition, (b) documented impairment in memory (produced a score greater than or equal to -1.5 SD on an objective measure), c) preserved functional independence (obtained a global score of less than or equal to 0.5 on the Clinical Dementia Rating Scale [21] - interview with a loved one), and (d) not demented. An Institutional Review Board approved data collection.

After providing informed consent, participants completed an assessment session via Zoom that lasted approximately 90 minutes. During the tele-session, participants completed the discourse protocol and cognitive-linguistic battery with an assessor. The discourse protocol tasks relevant to this project are: 1) the "Cookie Theft" picture description task [22] elicited with the prompt, "Please tell me everything you see going on in this picture"; 2) the "Cat Rescue" picture [23] elicited with the prompt, "Tell me a story with a beginning, a middle, and an end"; and 3) the Norman Rockwell print "Coming and Going" [24] elicited with the same prompt as the Cat Rescue task. The cognitive-linguistic battery included the MoCA [25]. The assessor used a standardized script and materials to deliver the discourse protocol and audio-recorded the administration using high-quality audio recording guidelines. The study data collection was managed using Research Electronic Data Capture [26, 27] tools.

In the Taiwanese corpus, inclusion criteria were participants between 60 and 90 years old, with at least six years of education, and no history of neurological or psychiatric disorders. The neurologist evaluated participants with MCI according to the NIA-AA criteria. The evaluation was based on their CDR scores, which had a global score of 0.5, and brain magnetic resonance imaging (MRI) conducted within two years before recruitment, which showed atrophy in regions related to Alzheimer's disease.

Picture description tasks were employed to elicit connected speech and recorded the responses using a digital recorder. Participants described a set of three pictures depicting Taiwanese culture, with the instruction to report everything they observed in each one. The evaluators refrained from providing feedback but encouraged participants to elaborate if their responses were insufficient. The speech data was transcribed manually, subsequently. The transcribers were unaware of the clinical diagnosis and only transcribed the words spoken by the participants. The remaining words were segmented into utterances and annotated as pauses, filled pauses (such as "uh," "um," "er," and "ah"), repetitions, and revisions. Punctuation is limited to periods, exclamation marks and question marks at the end of a sentence, and slash, centered dot, commas within a sentence. Correct mispronounced or orally used words to their written form. Filled pauses were not considered words, and multiple attempts to say the same word were only recorded once (e.g., "They brew-brew a pot of tea" was recorded as "They brew a pot of tea"). Words were grouped by part of speech and tagged using the Chinese Knowledge and Information Processing Lab[1].

Ethical approval was obtained from the Institutional Review Board of Cardinal Tien Hospital in Taipei, Taiwan (CTH-110-3-8-041), and all participants signed a written informed consent document.

SL: Anonymise this for submission

The full data set (English and Chinese) was age and gender balanced to avoid bias in modelling. We ensured that the speech recordings met suitable audio quality standards for processing. Propensity score matching [28] was employed to generate an unbiased training set. The data set was matched to scores de-

[1] https://github.com/ckiplab

fined in terms of the probability of an instance being treated as AD given covariates age and sex estimated through logistic regression, and matching instances were selected. All standardised mean differences for the covariates, standardised mean differences for squares, and two-way interactions between covariates were well below 0.1, indicating that the resulting set was adequately balanced.

The training set contained both Chinese and English samples with three picture descriptions per participant. The test set comprised recordings from different participants, with the same mix of languages and picture descriptions. Basic descriptive statistics of training and test set are shown in Tables 1 and 2. Overall, there are 507 speech samples (picture description recordings) with total duration of 528 minutes, ratio of training to test samples is just over 3:1. The data set has been made available to the wider research community via DementiaBank (`https://dementia.talkbank.org/`).

Table 1: *Training set description*

	MCI
Age	73.36 (SD 6.14, range 61-87)
Men	39.2% (n = 87)
Women	60.8% (n = 135)
MMSE	25.84 (SD 3.73, range 13-30)
Duration	58.92 (SD 36.61, range 12.7-240.9)
	NC
Age	71.85 (SD 6.65, range 61-87)
Men	38.2% (n = 63)
Women	61.8% (n = 102)
MMSE	29.07 (SD 1.08, range 25-30)
Duration	63.07 (SD 33.85, range 10.2-209.6)

Table 2: *Test set description*

	MCI
Age	77.90 (SD 9.15, range 59-91)
Male	52.4% (n = 33)
Female	47.6% (n = 30)
MMSE	25.86 (SD 3.27, range 18-30)
Duration	72.33 (SD 46.94, range 20.46-257.65)
	MCI
Age	67.68 (SD 4.71, range 62-82)
Male	36.8% (n = 21)
Female	63.2% (n = 36)
MMSE	29.05 (SD 1.06, range 26-30)
Duration	63.47 (SD 48.55, range 10.7-253.471)

4. Cognitive assessment tasks

The benchmark presented in this paper encompasses the following tasks: (a) a classification task, where we aim to create models to distinguish NC speech from MCI speech, and (b) a cognitive test score prediction (regression) task, where we create models to infer the subject's MMSE scores based on connected (spontaneous) speech data.

The MCI classification task will be evaluated through specificity (σ), sensitivity (ρ) and F_1 scores for the MCI category. These metrics will be computed as follows: $\sigma = \frac{T_N}{T_N+F_P}$, $F_1 = \frac{2\pi\rho}{\pi+\rho}$, where $\pi = \frac{T_P}{T_P+F_P}$, $\rho = \frac{T_P}{T_P+F_N}$, N is the number of patients, T_P is the number of true positives, T_N is the number of true negatives, F_P is the number of false positives and F_N the number of false negatives. The balanced accuracy metric (unweighted average recall, UAR) will be used for the overall ranking of this task's results. It is defined as follows: $\mathrm{UAR} = \frac{\sigma+\rho}{2}$.

The MMSE regression task will be assessed using the RMSE, defined as $\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}}$, where $\hat{y}$ is the predicted MMSE score, y is the patient's actual MMSE score, and $\bar{y}$ is the mean score.

5. Modelling approach

As our goal is to explore models that generalise across languages, we aimed to create a single predictive model for each task which encompassed features extracted from both languages. Thus, the general architectures of our classification and regression systems is shown in Figure 1, where *comparable* features extracted from both languages are combined into a single predictive model.

5.1. Acoustic Feature extraction

The acoustic feature extraction procedure aimed to identify speech features that could generalise well across the two languages. We tested two different approaches: a traditional feature engineering approach, with a feature set that has been found useful in emotion recognition and other computational paralinguistics tasks (eGeMAPs), and a self-supervised feature learning approach (wav2vec). These are described below:

eGeMAPs: this feature set comprises the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, Hammarberg index, and slope V0 features, along with numerous statistical functions applied to these features. This results in a total of 88 features for every audio recording [29].

wav2vec: we used the pre-trained model wav2vec[2] and extracted features directly from raw audio [30]. To balance the duration of all audio recordings, we zero-pad the audio recordings for feature extraction. Next, we applied a dropout layer, followed by a feature aggregation layer and another dropout layer. For dimensionality reduction, we used MaxPool1d layer (with a size of 42000, and a stride of 10,000). The result was used as input features for the multilayer perceptron (MLP) models. This results in 512 features per audio recording.

5.2. Multi-layer of Perceptron

MLP models, with the Adam solver, were employed for both classification and regression. We set $\alpha = 10^{-4}$, hidden layers of sizes 55, 160,160 and 55, and a maximum of 10,000 iterations. In both cases, we used 20-fold cross-validation for model assessment.

SL: We should also try adding comparable language features to these models. Perhaps this will improve the classification baselines. The regression baseline seems quite strong with wav2vec alone though.

SL: We need to add (as per Interspeech instructions: "A description of the computing infrastructure used and the average runtime for each model or algorithm (e.g. training, inference etc), and the number of parameters in each model.

6. Results

For the classification (diagnostic) task , our model achieved a test-data UAR of 59.18% while fusing the wav2vec and eGeMAPs features. The full set of results is reported in Table 3. The baseline result for this task is 59.18% UAR obtained on test data with sensitivity of 0.5873, Specificity of 0.5965, precision of 0.6167, Negative Predictive Value of 0.5667, False Positive Rate of 0.4035, False Discovery Rate of 0.3833, False Negative Rate of 0.4127, Accuracy of 0.5917, F1 Score of 0.6016

[2]https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_large.pt

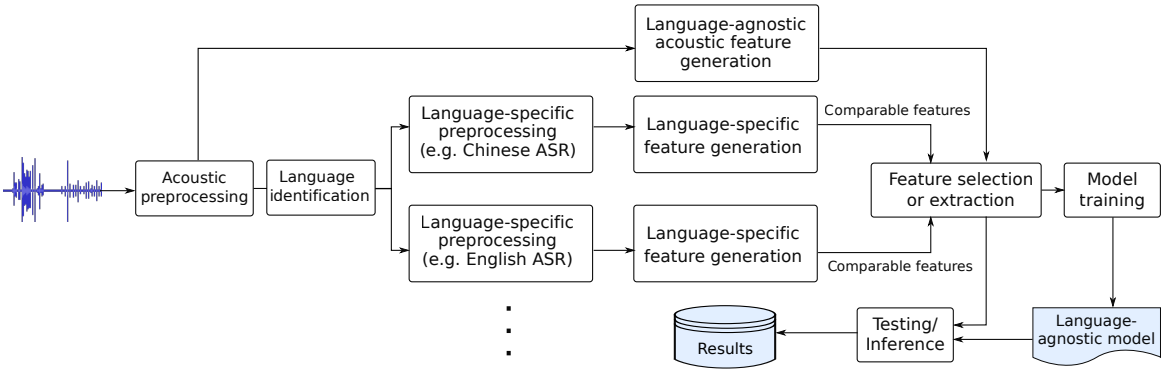

Figure 1: *General architecture for multilingual cognitive assessment based on recorded speech.*

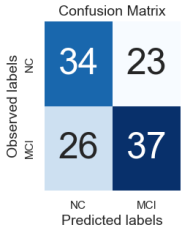

Figure 2: *fusion test results*

and Matthews Correlation Coefficient of 0.1836. The confusion matrix is shown in Figure 2.

SL: The problem with these results is that if we were to choose a model purely based on CV we would have chosen the wav2vec, which clearly overfits, giving a test set UAR of just over 46%. We need to address this.

Table 3: *Summary of results for the classification tasks (in % UAR) and the MMSE regression task (in RMSE). The features were validated in 20 fold Cross-Validation (CV) using MLP.*

	Feature	eGeMAPs	wa2vec	Fusion
Task 1	CV	49.56	61.6	50.94
	Test	44.95	46.05	**59.18**
Task 2	CV	13.18	3.70	13.14
	Test	17.00	**4.48**	13.28

For the regression task, wa2vec features on their own proved to be the most effective features, with RMSE scores of 3.70 (r=0.280) and 4.48 (r=-0.136), for validation and test sets respectively.

FASIH: add language features. Run bootstrap to get confidence intervals.

7. Discussion

The present data set is considerably less heterogeneous in terms of diagnoses and cognitive test scores than most public data sets used to data in research on predictive models for cognitive function assessment, including the few existing cross- and multi-lingual speech data sets used in this area [19, 18]. This makes the learning tasks defined in this paper harder, as they need to discriminate over a narrower range of values. However, the performance of our baseline models is comparable to those models.

A distinctive characteristic of our approach is the use of languages-agnostic and comparable languages-specific features.

SL: This is not quite true yet, as we haven't added the comparable linguistic features yet.

ALL: here we will discuss our results, the limitations of the data and approaches, etc.
Fasih: can you add some comments about the feature sets, and in particular about the notably superior performance of wav2vec over the alternatives in MMSE prediction?

8. Conclusion

This paper presented a novel benchmark data set for the development and testing of models for cognitive assessment through automatic analysis of connected speech. In particular, it defined learning tasks for diagnosis of MCI and prediction of MMSE. A general processing architecture for cross-lingual cognitive assessment was proposed which encompassed language-agnostic acoustic features and comparable linguistic features in a single predictive model for English and Chinese speech. Baseline models illustrated these predictive tasks and approach to feature extraction. The data and metadata have been made available to the research community. With the increasing interest by the medical community in speech biomarkers as a convenient and cost-effective approach to early detection and monitoring of cognitive problems, we expect this new resource will stimulate further research in the little explored field of cross-lingual modelling of cognitive function.

SL: Please feel free add to and revise this section.

9. Acknowledgements

Acknowledgement should only be included in the camera-ready version, not in the version submitted for review. The 5th page is reserved exclusively for acknowledgements and references. No other content must appear on the 5th page.

The Interspeech 2024 organisers would like to thank ISCA and the organising committees of past Interspeech conferences for their help and for kindly providing the previous version of this template.

10. References

[1] GBD 2019 Dementia Forecasting Collaborators, "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019," *Lancet Public Health*, vol. 7, no. 2, pp. e105–e125, Feb. 2022.

[2] S. de la Fuente Garcia, C. Ritchie, and S. Luz, "Artificial intelligence, speech and language processing approaches to monitoring Alzheimer's disease: a systematic review," *Journal of Alzheimer's Disease*, vol. 78, no. 4, 2020.

[3] S. Mirzaei, M. El Yacoubi, S. Garcia-Salicetti, J. Boudy, C. Kahindo, V. Cristancho-Lacroix, H. Kerherve, and A. S. Rigaud, "Two-stage feature selection of voice parameters for early Alzheimer's disease prediction," *Irbm*, vol. 39, no. 6, pp. 430–435, 2018.

[4] M. Nasrolahzadeh, Z. Mohammadpoory, and J. Haddadnia, "Higher-order spectral analysis of spontaneous speech signals in Alzheimer's disease," *Cognitive Neurodynamics*, vol. 12, no. 6, pp. 583–596, 2018.

[5] B. Mirheidari, D. Blackburn, R. O'Malley, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Computational cognitive assessment: Investigating the use of an intelligent virtual agent for the detection of early signs of dementia," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2732–2736.

[6] E. T. Prud'hommeaux and B. Roark, "Graph-based word alignment for clinical language evaluation," *Comput. Linguist.*, vol. 41, no. 4, pp. 549–578, 2015.

[7] R. Sadeghian, J. D. Schaffer, and S. A. Zahorian, "Speech processing approach for diagnosing dementia in an early stage," pp. 0–4, 2017.

[8] K. Shinkawa, A. Kosugi, M. Nishimura, M. Nemoto, K. Nemoto, T. Takeuchi, Y. Numata, R. Watanabe, E. Tsukada, M. Ota, S. Higashi, T. Arai, and Y. Yamada, "Multimodal behavior analysis towards detecting mild cognitive impairment: Preliminary results on gait and speech," *Stud Health Technol Inform*, vol. 264, pp. 343–347, 2019.

[9] 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.

[10] C. Carnero-Pardo, "Should the mini-mental state examination be retired?" *Neurología (English Edition)*, vol. 29, no. 8, pp. 473–481, 2014.

[11] R. Ostrand and J. Gunstad, "Using automatic assessment of speech production to predict current and future cognitive function in older adults," *Journal of Geriatric Psychiatry and Neurology*, vol. 34, no. 5, pp. 357–369, 2021.

[12] A. A. Wisler, A. R. Fletcher, and M. J. McAuliffe, "Predicting Montreal cognitive assessment scores from measures of speech and language," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 6, pp. 1752–1761, 2020.

[13] Z. Liu, L. Proctor, P. N. Collier, and X. Zhao, "Automatic diagnosis and prediction of cognitive decline associated with alzheimer's dementia through spontaneous speech," in *2021 IEEE International Conference on Signal and Image Processing Applications*. IEEE, 2021, pp. 39–43.

[14] Z. Fu, F. Haider, and S. Luz, "Predicting mini-mental status examination scores through paralinguistic acoustic features of spontaneous speech," in *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 5548–5552.

[15] K. Lopez-de Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C. M. Travieso, M. Ecay-Torres, P. Martinez-Lage, and H. Egiraun, "On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature," *Cognitive Computation*, vol. 7, no. 1, pp. 44–55, 2015.

[16] F. Haider, S. de la Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech," *IEEE J Sel Top Signal Process*, vol. 14, no. 2, pp. 272–281, 2020.

[17] Z. Guo, Z. Ling, and Y. Li, "Detecting Alzheimer's disease from continuous speech using language models," *Journal of Alzheimers Disease*, vol. 70, no. 4, pp. 1163–1174, 2019.

[18] K. C. Fraser, K. Lundholm 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.

[19] S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. MacWhinney, "Multilingual Alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge," in *Proccedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2023)*. IEEE Press, Jun. 2023.

[20] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen *et al.*, "The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 270–279, 2011.

[21] J. C. Morris, "The clinical dementia rating (CDR) current version and scoring rules," *Neurology*, vol. 43, no. 11, pp. 2412–2412, 1993.

[22] E. Kaplan, *Boston diagnostic aphasia examination booklet*. Lea & Febiger Philadelphia, PA, 1983.

[23] L. E. Nicholas and R. H. Brookshire, "A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 338–350, 1993.

[24] N. Rockwell, "Going and coming [oil on canvas]," *Norman Rockwell Art Collection Trust, Indianapolis, IN, United States*, 1947.

[25] 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.

[26] P. A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J. G. Conde, "Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support," *Journal of biomedical informatics*, vol. 42, no. 2, pp. 377–381, 2009.

[27] P. A. Harris, R. Taylor, B. L. Minor, V. Elliott, M. Fernandez, L. O'Neal, L. McLeod, G. Delacqua, F. Delacqua, J. Kirby *et al.*, "The REDCap consortium: building an international community of software platform partners," *Journal of biomedical informatics*, vol. 95, p. 103208, 2019.

[28] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 04 1983.

[29] F. Eyben *et al.*, "The Geneva minimalistic acoustic parameter set for voice research and affective computing," *IEEE Trans Affect Computing*, vol. 7, no. 2, 2016.

[30] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

IS24 Author check list The following checklist will be part of the submission form and includes a list of guidelines we expect Interspeech 2024 papers to follow. Not every point in the check list will be applicable to every paper. Ideally, all guidelines that do apply to a certain paper should be satisfied. Nevertheless, not satisfying a guideline that applies to your paper is not necessarily grounds for rejection. When your paper does not satisfy one or more of the guidelines in a section, the submission form will ask that you please explain why.

1. Claims and limitations - for all papers
   - The paper clearly states what claims are being investigated.
   - The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.
   - The limitations of your work are described.
   - All assumptions made in your work are stated in the paper.
2. If data sets are used, the paper includes information about the following:
   - Relevant details such as languages, audio duration distribution, number of examples, and label distributions, or a reference where that information can be found.
   - Details of train/validation (development)/test splits. Ideally, the lists should be released with the supplementary material if not already publicly available.
   - Explanation of all pre-processing steps, including any criteria used to exclude samples, if applicable.
   - Reference(s) to all data set(s) drawn from the existing literature.
   - For newly collected data, a complete description of the data collection process, such as subjects, instructions to annotators and methods for quality control and a statement on whether ethics approval was necessary.
3. If using non-public data sets:
   - The paper includes a discussion on the reason/s for using non-public data sets.
   - Full details of the dataset are included in the paper to enable comparison to similar data sets and tasks.
4. If reporting experimental results, the paper includes:
   - An explanation of evaluation metrics used.
   - An explanation of how models are initialized, if applicable.
   - Some measure of statistical significance of the reported gains or confidence intervals (a python toolkit and brief tutorial for computing confidence intervals with the bootstrapping approach can be found in https://github.com/luferrer/ConfidenceIntervals).
   - A description of the computing infrastructure used and the average runtime for each model or algorithm (e.g. training, inference etc).
   - The number of parameters in each model.
5. If hyperparameter search (including choice of architecture or features and any other development decision) was done, the paper includes:
   - Final results on a held-out evaluation set not used for hyperparameter tuning.
   - Hyperparameter configurations for best-performing models.
   - The method for choosing hyperparameter values to explore, and the criterion used to select among them.
6. If source code is used:
   - The code is or will be made publicly available and/or sufficient details are included in the paper for the work to be reproduced.
   - For publicly available software, the corresponding version numbers and links and/or references to the software.