Dementia Detection from Speech Samples

Saturnino Luz
Usher Institute
Edinburgh Medical School
The University of Edinburgh

Brian MacWhinney
Language Technologies and
Modern Languages
Carnegie Mellon University

Benchmarking: Past, Present and Future, ACL 2021

Better health, better futures

THE UNIVERSITY of EDINBURGH | usher institute

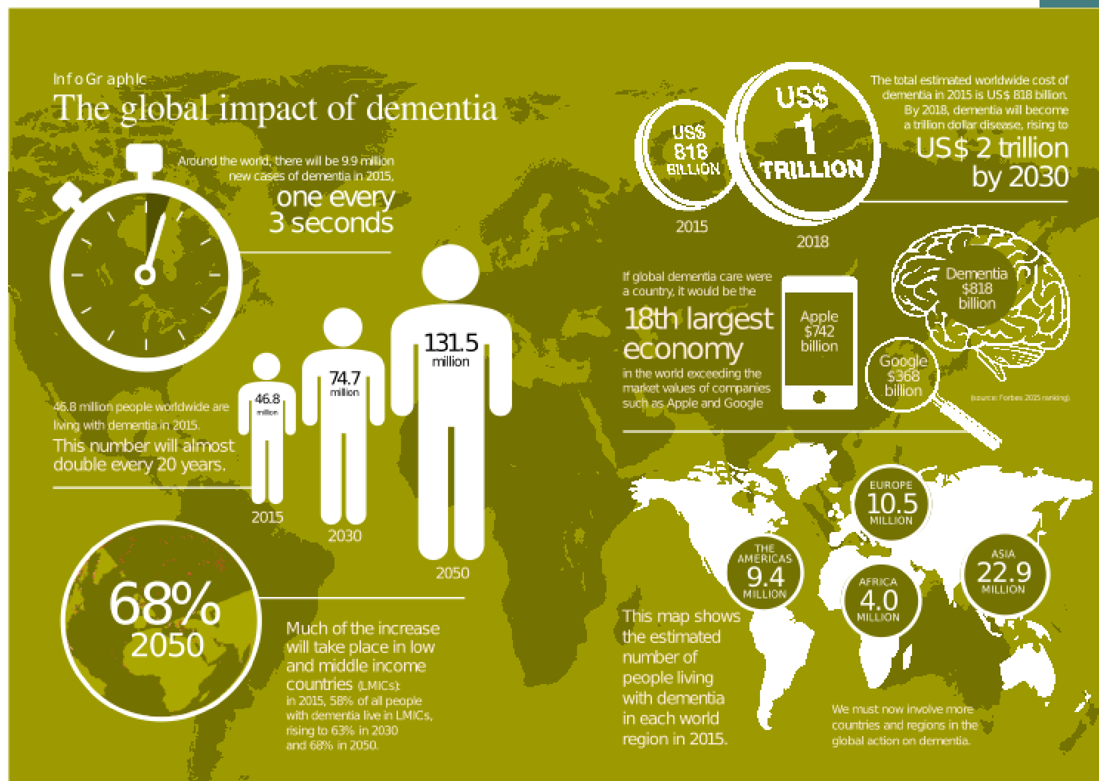
Carnegie Mellon University

Outline

- Alzheimer's detection: background
- Speech as "digital biomarkers"
- Benchmarking for Alzheimer's detection from speech:
  - The ADrESS challenges

Alzheimer's Dementia (AD)

- Alzheimer's is a **neurodegenerative disease** that entails long-term and usually gradual decline in cognitive functioning.
- Clinical manifestations include:
  - Subjective Memory Loss (**SML**)
  - Mild Cognitive Impairment (**MCI**) and
  - Alzheimer's Type Dementia (**ATD**)
- A disease of **increasing global impact**.



Detecting Alzheimer’s Type Dementia

- Reasons of testing:
  - Diagnosis
  - Scrrening for clinical trials
  - characterisation of impairment
  - monitoring of interventions/therapy
  - **Characterisation of communication difficulties** involving persons with ATD for **speech therapy** interventions, carer **coaching**, etc.
- **Cognitive Tests** to detect **MCI** and **ATD**
- More costly and/or invasive tests
  - **neuroimaging** (PET, MRI)
  - CSF, blood tests

M. Mortamais, J. A. Ash, J. Harrison, J. Kaye, J. Kramer, C. Randolph, C. Pose, B. Albala, M. Ropacki, C. W. Ritchie, and K. Ritchie. Detecting cognitive changes in preclinical Alzheimer’s disease: A review of its feasibility. *Alzheimer’s & Dementia*, 13(4):468–492, 2017

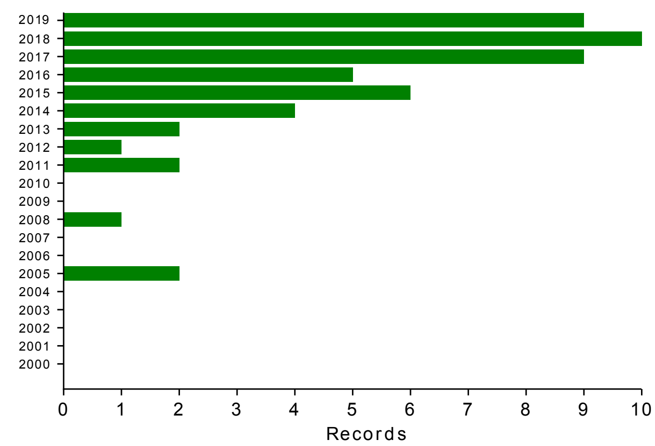
Focus on speech and language

- Much information on **cognitive status** can be gathered through speech
- Can be captured in **natural settings**, over time, and
- Might overcome **daily fluctuations** that affect cognitive test **performance**:
  - fatigue, mood, attentiveness, short-term illnesses, test anxiety, etc
- Data sources:
  - word tests,
  - **narration** (scene descriptions),
  - **interviews**,
  - spontaneous **conversations**, ...

Speech and Language AD research

- In recent years, **several research groups** have investigated **ATD detection** based on **speech and language**.
- A recent systematic review (de la Fuente Garcia et al., 2020) identified 51 articles on speech/language approaches to monitoring AD
- Data sources:
  - word tests,
  - **narration** (scene descriptions),
  - **interviews**,
  - spontaneous **conversations**, ...

A growing field



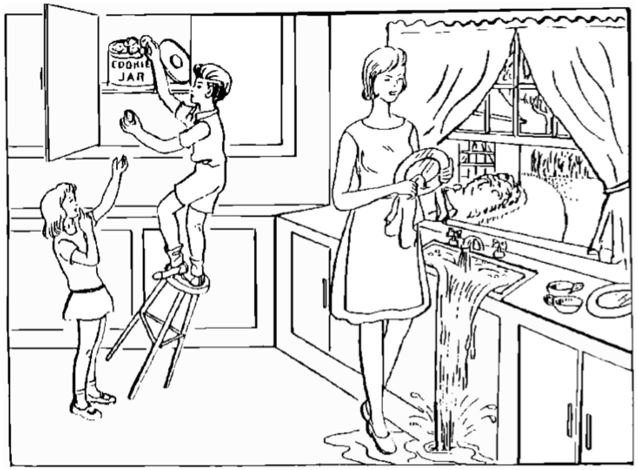
- Datasets:
  - 8 (in 51) used DementiaBank
  - But the majority of studies (36 out of 51) did not report data availability

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*, 78(4):1547–1574, 2020

The Pitt Dataset from DementiaBank

Recorded speech data for a number of neuropsychological tests:

- Fluency
- Word recall
- Sentence production
- Cambridge **Cookie Theft test**:
  - **Probable AD** speech
  - **Normal control** speech



Control	**242**
MCI	43
Memory	3
PossibleAD	21
ProbableAD	**236**
Vascular	5

J. Becker, F. Boller, O. Lopez, J. Saxton, and K. McGonigle. The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594, 1994

An example: testing AD detection algorithms on DementiaBank's Pitt dataset

Balanced and **acoustically enhanced speech** dataset:

	AD		non-AD	
Age Interval	Male	Female	Male	Female
$[50, 55)$	2	1	2	1
$[55, 60)$	7	8	7	8
$[60, 65)$	4	9	4	9
$[65, 70)$	10	14	10	14
$[70, 75)$	9	11	9	11
$[75, 80)$	4	3	4	3
Total	36	46	36	46

- Assessed several voice **feature sets**:
  - emobase **(Eyben et al., 2010)**
  - ComParE **(Schuller et al., 2014)**,
  - eGeMAPS **(Eyben et al., 2016)**
  - Multiresolution cochleagram **(Haider and Luz, 2019)**

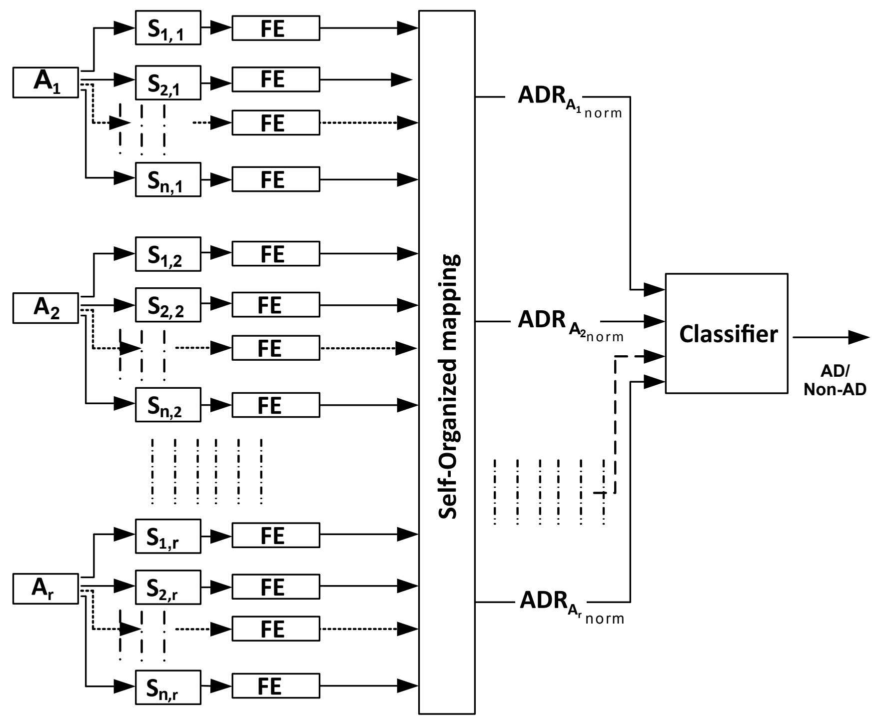
F. Haider, S. de la Fuente, and S. Luz. An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):272–281, 2020

F. Haider and S. Luz. Attitude recognition using multi-resolution cochleagram features. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3737–3741, 2019

Active data representation (ADR)

ADR **feature extraction** process:

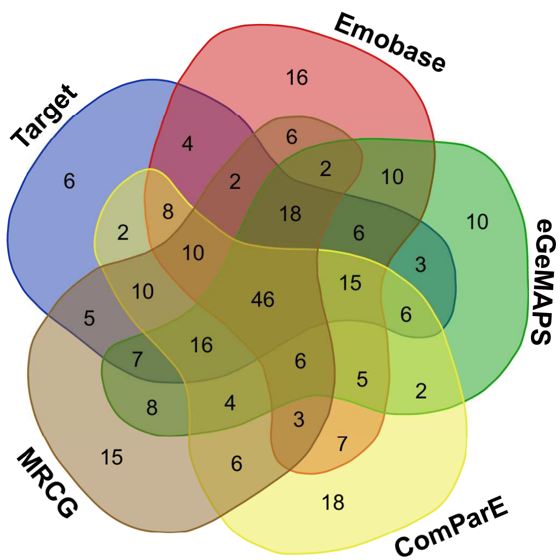
1. Audio **segmentation**
2. SOM **clustering**
3. Generation of **histograms** for segment duration and number
4. Computation of **rates of change** in cluster membership
5. **Normalisation** (L1 norm)



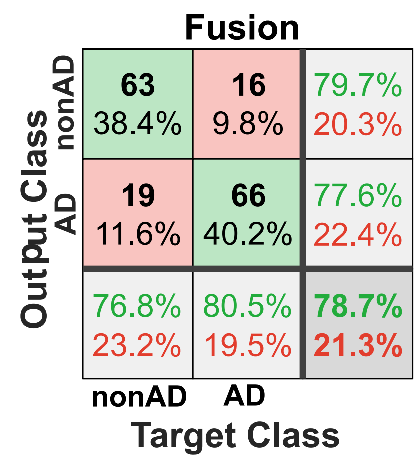
F. Haider, S. de la Fuente, and S. Luz. An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):272–281, 2020

Combining feature sets

Analysing the performance of the ADR for different feature sets with a DT classifier:



The ADRs capture different aspects of classification, and **ADR fusion** (using simple DT classifiers) produces relatively **good results**:



Study	accuracy	modality	fully automatic	privacy
Our approach	**78.7%**	acoustic	yes	yes
Hernández et AL., 2018	62.0%	acoustic	yes	no(?)
Luz, 2017	68.0%	acoustic	yes	yes
Mirheidari et Al., 2018	62.3%	text	yes (ASR)	no
Fraser et Al., 2016	**81.9%**	text/acoustic	no	no (text)
Yancheva & Rudzicz, 2016	80.0%	text/acoustic	no (text)	no
Hernández et AL., 2018	68.0%	text	no	no
Mirheidari et Al., 2018	75.6%	text	no	no

Challenges and shortcomings

- Lack of **balanced and standardised data** sets on which different approaches can be compared;
- Reproducibility, **methodological inconsistencies** across studies;
- Scarcity of spontaneous speech/interaction **data**, particularly **longitudinal data**;
- Challenges in data **pre-processing**
  - segmentation,
  - diarisation,
  - feature extraction (including ASR).
- **Disconnect** between studies' aims and **clinical research and/or practice**;

Challenges and shortcomings

- **Lack of balanced and standardised data sets on which different approaches can be compared**;
- **Reproducibility, methodological inconsistencies across studies**;
- Scarcity of spontaneous speech/interaction **data**, particularly **longitudinal data**;
- Challenges in data **pre-processing**
  - segmentation,
  - diarisation,
  - feature extraction (including ASR).
- **Disconnect** between studies' aims and **clinical research and/or practice**;

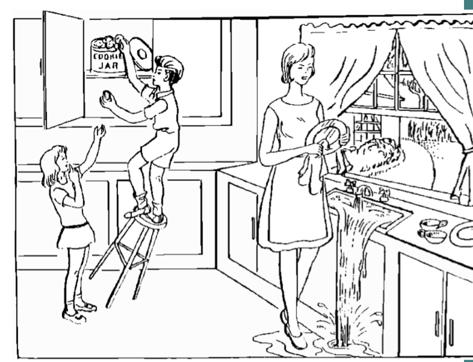
A Benchmark to ”ADReSS” some of these challenges

- **ADReSS**: **A**lzheimer’s **D**ementia **Re**cognition through **S**pontaneous **S**peech.
- Special session at **INTERSPEECH’20**
- Two automatic **prediction tasks**:
  - Alzheimer’s Dementia **classification** task
  - Cognitive test **(MMSE) score regression** task
- The ADReSS Challenge **dataset** is
  - acoustically **pre-processed**
  - **balanced** in terms of age and gender
  - Available through DementiaBank: **https://dementia.talkbank.org/**

S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney. Alzheimer’s dementia recognition through spontaneous speech: The ADReSS Challenge. In *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020. URL `https://arxiv.org/abs/2004.06833`

The ADReSS data set

- **Cookie Theft** picture description task, from
  - the **Boston Diagnostic Aphasia** Exam
- Part of DementiaBank's **Pitt Corpus**
- Transcripts **annotated** using the **CHAT** coding system (MacWhinney, 2019)
- Recordings were **acoustically enhanced** with stationary noise removal
- Audio **volume was normalised** across all speech segments
  - control for variation caused by **recording conditions**, such as microphone placement.



B. MacWhinney. Understanding spoken language through talkbank. *Behavior research methods*, 51(4):1919–1927, 2019

More about the data set

- Carefully selected so as to mitigate **common biases**:
  - **repeated occurrences** of speech from the same participant,
  - variations in **audio quality**, and
  - imbalances of **gender** and **age** distribution.
- Segmented for voice activity based on a signal energy threshold.
  - 65dB, maximum of 10 seconds per segment.
  - 1,955 speech segments from 78 non-AD participants and
  - 2,122 speech segments from 78 AD participant.
  - The average number of speech segments per participant 24.86 (sd=12.84)

Baseline features

- Acoustic features:
  - emobase **(Eyben et al., 2010)**
  - ComParE **(Schuller et al., 2014)**,
  - eGeMAPS **(Eyben et al., 2016)**
  - Multiresolution cochleagram **(Haider and Luz, 2019)**
  - Minimal: statistics (mean, standard deviation, median, minimum and maximum) of the duration of vocalisations and pauses, speech rate, and a vocalisation count (20 features).
- Linguistic features:
  - basic set of 34 language outcome measures (e.g., duration, total utterances, MLU, type-token ratio, open-closed class word ratio, percentages of 9 parts of speech) on the CHAT transcripts.

F. Eyben, M. Wöllmer, and B. Schuller. openSMILE: the Munich versatile and fast open-source audio feature extractor. In *Procs. of ACM-MM*, pages 1459–1462. ACM, 2010

B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang. The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load. *Proc. Interspeech, Singapore, Singapore*, 2014

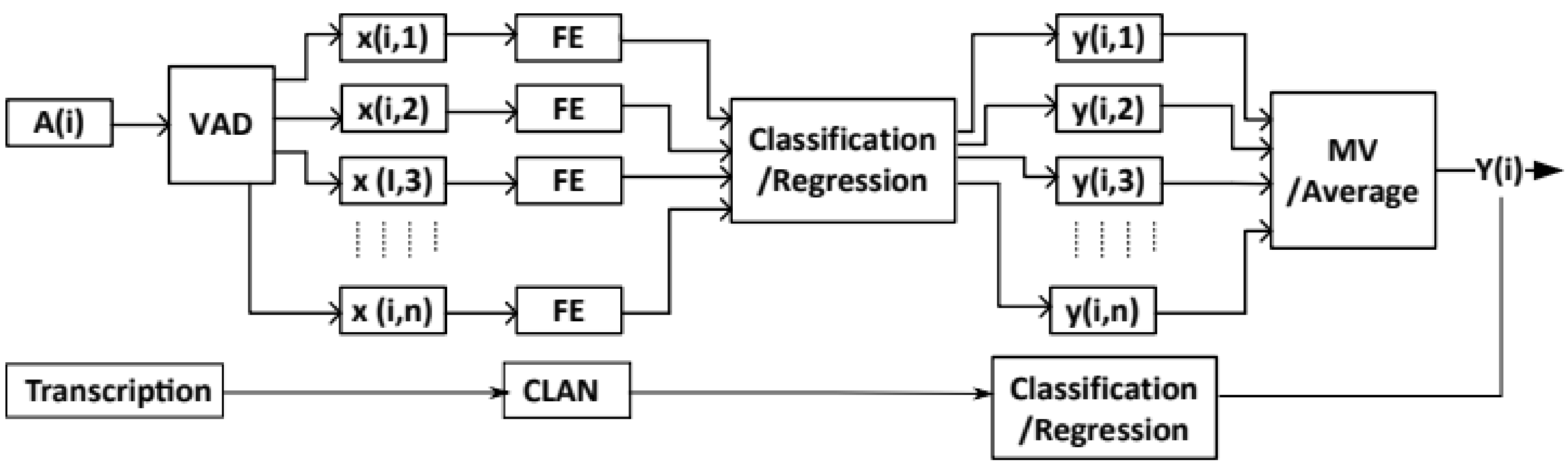
F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al. The Geneva minimalistic acoustic parameter set GeMAPS for voice research and affective computing. 7(2):190–202, 2016

F. Haider and S. Luz. Attitude recognition using multi-resolution cochleagram features. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3737–3741, 2019

Rules

- Participants could use acoustic features and linguistic features, separately or combined:
- They could attempt one of the tasks or both,
- were provided with access to a training set,
- and were given access to a separate set on which models were tested two weeks prior to the paper submission deadline.
- They could send results to us for scoring up to 5 times
- but were required to submit all attempts (up to 5 per task) together, in separate files.
- Evaluation metrics for AD classification: accuracy, precision, recall, F1
- Metric for MMSE score prediction: root mean squared error (RMSE)

Baseline system



Baseline results

- AD classification results on test set (LDA classifier)

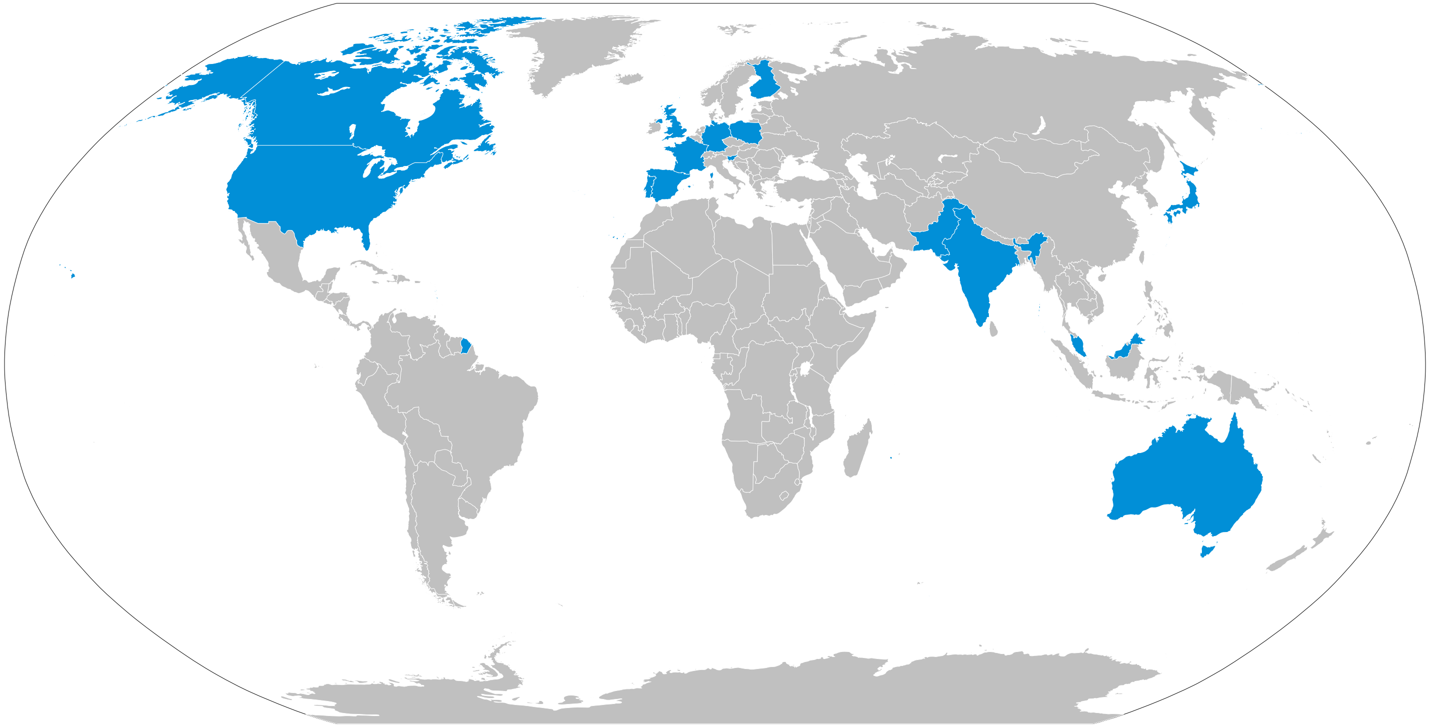
	class	Precision	Recall	F1 Score	Accuracy
$LOSO_{Acous}$	non-AD	0.56	0.61	0.58	0.56
	AD	0.57	0.52	0.54	
$TEST_{Acous}$	non-AD	0.67	0.50	0.57	0.62
	AD	0.60	0.75	0.67	
$LOSO_{ling}$	non-AD	0.76	0.78	0.77	0.77
	AD	0.77	0.76	0.77	
$TEST_{ling}$	non-AD	0.70	0.87	0.78	0.75
	AD	0.83	0.62	0.71	

- MMSE prediction results on test set

Features	Linear	DT	GP	SVM	LSBoost	mean
emobase	6.80	6.78	6.36	6.18	6.73	6.57
ComParE	6.47	6.52	6.33	6.19	6.72	6.45
eGeMAPS	6.90	5.99	6.28	6.12	6.41	6.34
MRCG	6.70	**6.14**, $r = 0.22$	6.33	6.20	6.31	6.33
Minimal	6.29	6.84	6.58	6.19	7.71	6.72
Linguistic	4.78	**5.20**, $r = 0.57$	5.54	6.24	6.62	5.68
mean	6.32	**6.25**	6.24	6.19	6.75	–

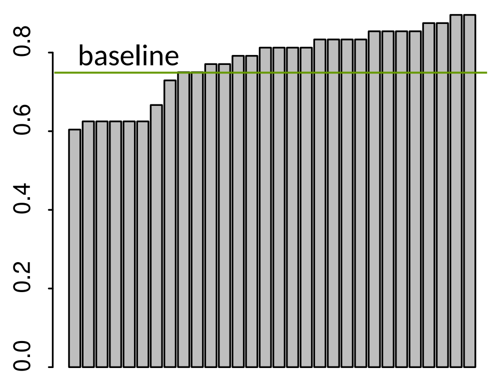
Participation in the ADReSS Challenge

- 33 teams from around the world entered the challenge



Results of the classification task

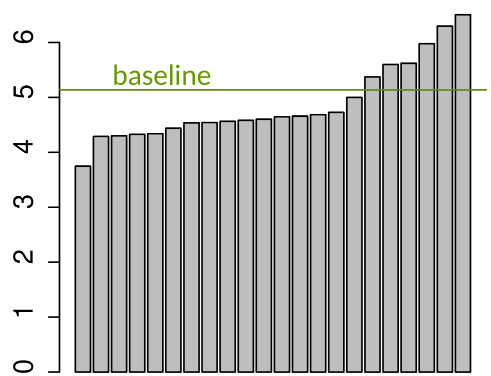
Part cipant (top 12 scores among accepted papers)	***Accuracy***	F1-Score (nonAD)	F1-Score (AD)	F1-Score (mean)
Baidu USA	***0.8958***	0.902	0.8889	0.8955
RMIT, Australia and Mehran University, Pakistan	**0.8542**	0.8627	0.8444	0.8536
Winterlight Labs, Toronto, Canada	0.8333	0.8261	0.84	0.8331
MIT Media Lab, Massachuset s Inst tute of Technology	0.8333	0.8333	0.8333	0.8333
INESC-ID's, Portugal	0.8125	0.8364	0.7805	0.8085
Music & Audio Res. Group, Seoul Nat onal University	0.8125	0.8085	0.8163	0.8124
Augsburg, Sheif eld, Nijmegen & Philips Res.	0.8125	0.800	0.8235	0.8118
Kings College London	0.8125	0.8085	0.8163	0.8124
Verisk & Aalto Univ	0.7917	0.8	0.7826	0.7913
Queen Mary University London	0.7917	0.7917	0.7917	0.7917
JSI	0.7708	0.7843	0.7556	0.7700
John Hopkins University	0.7500	0.7143	0.7778	0.7461



All accuracy scores

Results of the regression task

Part cipants (top 10 scores of accepted papers)	*RMSE*
Music and Audio Research Group at Seoul Nat onal University	***3.747***
RMIT University, Australia & Mehran Univ, Pakistan	**4.301**
University of Illinois Chicago	4.340
JSI	4.439
QMUL	4.537
Winterlight Labs, Toronto	4.563
Kings College London	4.583
MIT Media Lab, Massachuset s Inst tute of Technology	4.602
Universit es of Augsburg, Sheif eld and Nijmegen & Philips Reseasch	4.659
Johns Hopkins University	0.530



All RMSE scores

Extended results: Journal Special Issue

- Joint Frontiers in **Aging Neuroscience**/**Frontiers in Computer Science** special issue:
  - `https://www.frontiersin.org/research-topics/13702/alzheimers-dementia-recognition-through-spontaneous-speech`

Research Topic

Alzheimer's Dementia Recognition through Spontaneous Speech

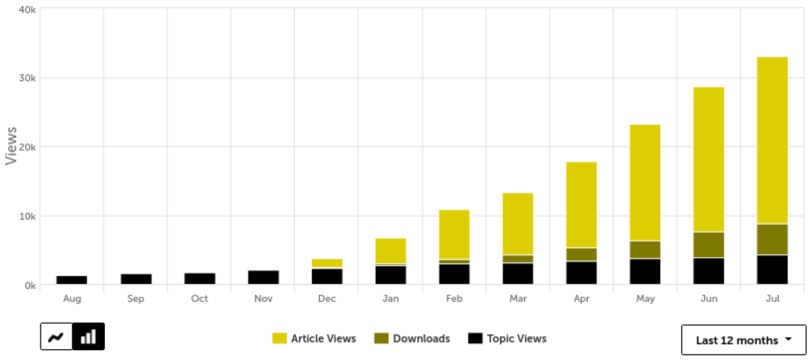
Submission closed.

Overview | 20 Articles | Authors | Impact

33,039 Views | Demographics

Last 12 months

33,039 total views **24,290** article views **4,547** article downloads **4,202** topic views



ADReSSo 2021

Alzheimer's Dementia Recognition through Spontaneous Speech The ADReSSo Challenge

News:

- NEW! ADReSSo Challenge announced! (18-1-21)

- More information at:
  - `https://edin.ac/3p1cyaI`

    and

  - `https://://www.interspeech2021.org/special-sessions-challenges`

Changes this year

- Expanded data set
  - AD vs CN
  - Longitudinal MMSE scores
- No transcripts provided
- Three tasks:
  - detection of Alzheimer's Dementia,
  - inference of cognitive testing scores, and
  - prediction of cognitive decline (disease progression).

The data sets

- speech recordings of Alzheimer's patients performing a category (semantic) fluency task at their baseline visit, for prediction of cognitive decline over a two year period,
- picture descriptions produced by cognitively normal subjects and patients with an AD diagnosis, as in ADReSS'20
- data also includes speech from different experimenters who gave instructions to the patients and occasionally interacted with them in short dialogues.
- segmentation of the recordings into vocalisation sequences with speaker identifiers made available, but no transcripts.
- Data marched using a proprensity score approach, to minimise risk of bias (Rosenbaum and Rubin, 1983, Rubin, 1973)

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects.
Biometrika, 70(1):41–55, 04 1983.
ISSN 0006-3444.
doi: 10.1093/biomet/70.1.41

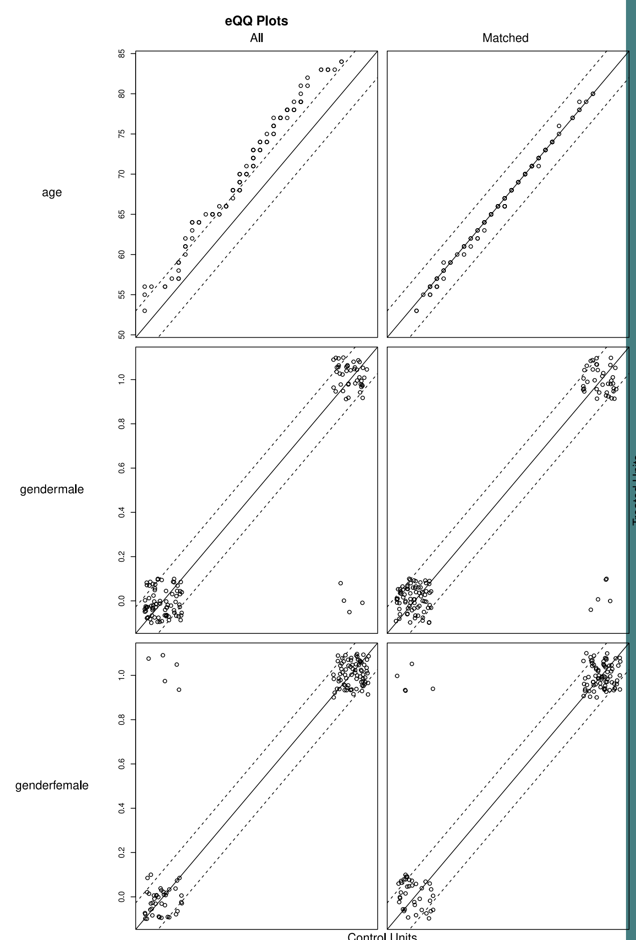
D. B. Rubin. Matching to remove bias in observational studies.
Biometrics, 29(1):159–183, 1973

Data matching

Table: Composition of the datasets.

	Tasks 1 and 2		Task 3	
	AD	CN	Decline	No decline
Age	69.38 ($sd = 6.9$)	66.06 (6.3)	69.84 (9.3)	70.26 (8.5)
Men	35.2% ($n = 43$)	34.8% (40)	24.0% (6)	47.5% (38)
Women	64.8% (79)	65.2% (75)	76.0% (19)	52.5% (42)
MMSE	17.8 (5.5)	28.9 (1.2)	17.9 (4.6)	20.7 (5.2)
Duration	65.7s (38.6)	61.6s (26.9)	58.2s (16.0)	48.9s (19.5)

- Quantile-quantile plots for data before (left) and after matching (right) by age and gender



Baseline system

- Acoustic features
  - eGeMAPS features extracted from 100ms time windows
  - ADR method used for generation of final feature set (Haider et al., 2020)
- Linguistic features generated from ASR transcripts encoded in CHAT format and processed with the CLAN software (MacWhinney, 2017):
  - EVAL to create a composite profile of 34 measures, and
  - FREQ to compute the Moving Average Type Token Ratio

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 Journal of Selected Topics in Signal Processing*, 14(2):272–281, 2020

B. MacWhinney. Tools for analyzing talk part 2: The CLAN program, 2017. URL http://talkbank.org/manuals/CLAN.pdf. Pittsburgh, PA: Carnegie Mellon University

S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney. Detecting cognitive decline using speech only: The ADReSS$_o$ challenge. *medRxiv*, 2021. doi: 10.1101/2021.03.24.21254263

Baseline results: AD detection

Table: Task1: AD classification accuracy on CV and test data

		LDA	DT	SVM	TB	KNN	mean (sd)
CV	Acoustic	62.65	**78.92**	69.28	65.06	65.06	68.19 (6.4)
	ASR	72.29	**72.89**	72.89	75.90	65.06	71.81 (4.0)
	Transcript	*80.12*	*77.71*	*80.72*	*76.51*	*69.28*	*76.87 (4.6)*
Test	Acoustic	50.70	60.56	**64.79**	63.38	53.52	58.59 (6.2)
	ASR	76.06	74.65	**77.46**	73.24	59.15	72.11 (7.4)
	Transcript	*76.06*	*67.61*	*78.87*	*66.20*	*60.56*	*69.86 (7.5)*

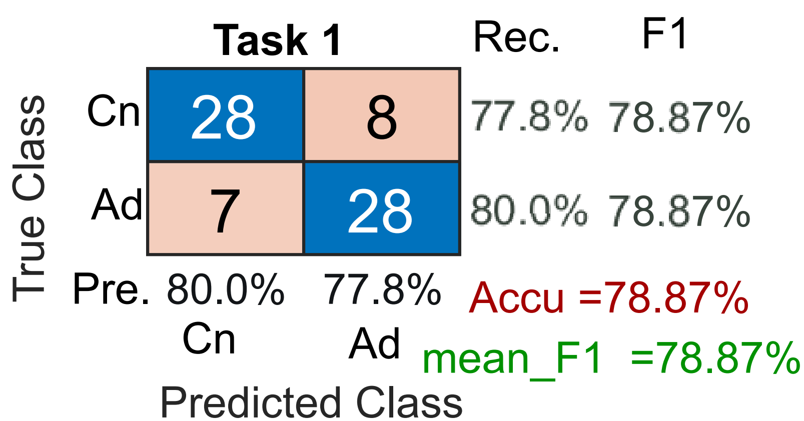


Figure: Late (decision) fusion of the best results of acoustic and linguistic models for Task 1.

Baseline results: MMSE Prediction

Table: Task2: MMSE score prediction error scores (RMSE).

		LR	DT	SVR	RF	GP	mean (sd)
	Acoustic	6.88	6.88	6.96	7.89	6.71	7.06 (0.47)
CV	ASR	6.65	5.92	6.42	7.02	6.50	6.50 (0.40)
	Transcript	*5.77*	*6.20*	*5.75*	*6.94*	*5.52*	*6.04 (0.56)*
	Acoustic	6.23	6.47	**6.09**	8.18	6.81	6.75 (0.84)
Test	ASR	5.87	6.24	**5.28**	6.94	5.43	5.95 (0.67)
	Transcript	*4.49*	*6.06*	*4.65*	*6.07*	*4.35*	*5.12 (0.87)*

Baseline results: Prognosis

Table: Task3: cognitive decline progression results (mean F_1) for leave-one-subject-out CV and test data.

		LDA	DT	SVM	TB	KNN	mean (sd)
Val	Acoustic	59.89	84.94	55.64	63.85	65.92	66.05 (11.27)
	ASR	55.19	76.52	45.24	63.10	55.25	59.06 (11.64)
test	Acoustic	**61.02**	53.62	40.74	40.74	38.46	46.91 (9.89)
	ASR	54.29	**66.67**	40.74	56.56	39.62	51.58 (11.41)

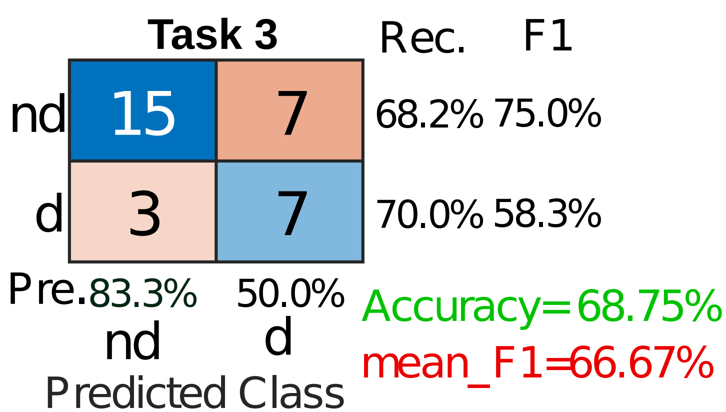


Figure: Late fusion of the best results of acoustic and linguistic models for cognitive decline prediction (Prognosis) Task.

Participation in ADReSSo’21

- More than 30 systems submissions
- 12 papers accepted for presentation at INTERSPEECH’21
- For those not attending INTERSPEECH’21, papers will be made available at the ISCA website...

Challenges and shortcomings

- Lack of balanced and standardised data sets on which different approaches can be compared; (**partly covered by ADReSS**)
- Reproducibility, methodological inconsistencies across studies; (**partly covered by ADReSS**)
- **Scarcity of spontaneous speech/interaction data**, particularly **longitudinal data**;
- Challenges in data **pre-processing**
  - segmentation,
  - diarisation,
  - feature extraction (including ASR).
- **Disconnect** between studies' aims and **clinical research and/or practice**;

References
I

J. Becker, F. Boller, O. Lopez, J. Saxton, and K. McGonigle. The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594, 1994.

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*, 78(4):1547–1574, 2020.

F. Eyben, M. Wöllmer, and B. Schuller. openSMILE: the Munich versatile and fast open-source audio feature extractor. In *Procs. of ACM-MM*, pages 1459–1462. ACM, 2010.

F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al. The Geneva minimalistic acoustic parameter set GeMAPS for voice research and affective computing. 7(2):190–202, 2016.

F. Haider and S. Luz. Attitude recognition using multi-resolution cochleagram features. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3737–3741, 2019.

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 Journal of Selected Topics in Signal Processing*, 14(2):272–281, 2020.

S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney. Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge. In *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020. URL `https://arxiv.org/abs/2004.06833`.

S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney. Detecting cognitive decline using speech only: The ADReSS$_o$ challenge. *medRxiv*, 2021. doi: 10.1101/2021.03.24.21254263.

B. MacWhinney. Tools for analyzing talk part 2: The CLAN program, 2017. URL `http://talkbank.org/manuals/CLAN.pdf`. Pittsburgh, PA: Carnegie Mellon University.

B. MacWhinney. Understanding spoken language through talkbank. *Behavior research methods*, 51(4):1919–1927, 2019.

M. Mortamais, J. A. Ash, J. Harrison, J. Kaye, J. Kramer, C. Randolph, C. Pose, B. Albala, M. Ropacki, C. W. Ritchie, and K. Ritchie. Detecting cognitive changes in preclinical Alzheimer's disease: A review of its feasibility. *Alzheimer's & Dementia*, 13(4):468–492, 2017.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70 (1):41–55, 04 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41.

D. B. Rubin. Matching to remove bias in observational studies. *Biometrics*, 29(1):159–183, 1973.

B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang. The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load. *Proc. Interspeech, Singapore, Singapore*, 2014.