ELEC-E5531
Speech and Language Processing Seminar

Aalto University
School of Electrical
Engineering

Automatic screening of mild cognitive impairment and Alzheimer's disease by means of posterior-thresholding hesitation representation

Computer Speech and Language

September 2022

Ying Liu and Zehang Li

Outline

- Background
- Motivation and Contribution
- Model
- Experiment
- Result
- Conclusion

Background

Mild cognitive impairment (MCI) : transitional phase between normal cognitive aging and dementia

- the prevalence of MCI ranges from **15%** to **20%** in individuals of 60 years and older
- the annual progression rate from MCI to dementia is between **8%** and **15%**
- MCI may be present up to **15 years** before the clinical manifestation of dementia

methods that can aid the screening of the disease are needed

Aalto University
School of Electrical
Engineering

Motivation

How to detect MCI?

- spoken language can reliably reflect cognition: **lexical-semantic abilities**, **memory,** and **executive functions**
- Compared to healthy controls:
  a. lower **speech rate**
  b. an increased number and length of **hesitations**

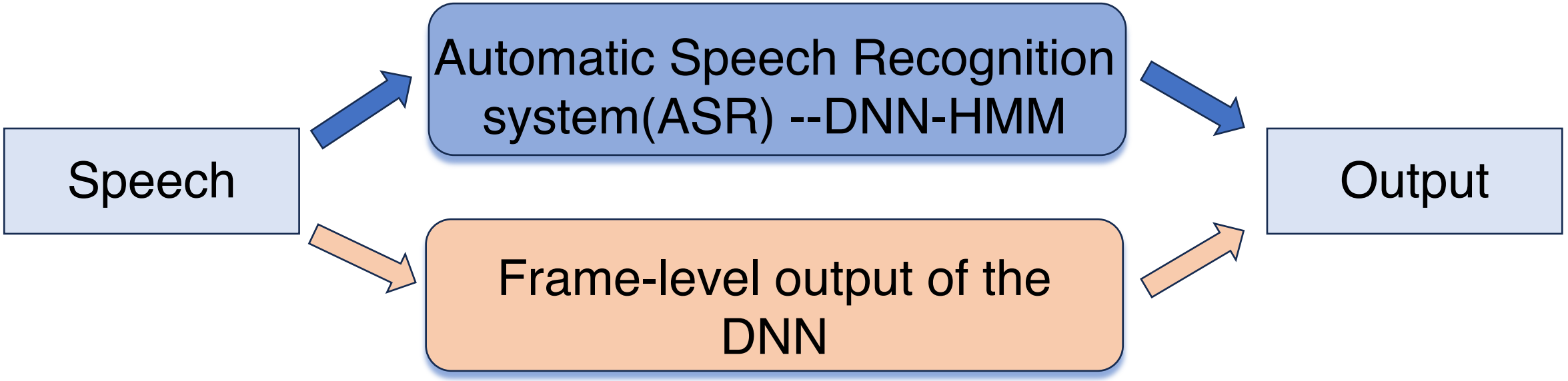
Analyzing the temporal aspects of speech allows the indirect investigation of cognition

Contribution

Hesitation is defined as an absence of speech:

silent pauses

filled pauses: vocalizations such as ‘er’, ‘umm’ etc



posterior-thresholding hesitation representation

same (or even better) classification performance, more resource-efficient

Data

Three categories of subjects:

- **mild cognitive impairment (MCI)**
- **early-stage Alzheimer's disease (mild AD or mAD)**
- **healthy controls (HC)**

	Subject groups			Test statistics	
	HC (n = 25)	**MCI** (n = 25)	**mAD** (n = 25)		
Gender (male/female)	7/18	10/15	12/13	$\chi^2 = 2.136$	$p = 0.344$
Age (mean ± SD)	69.24 ± 5.118	70.68 ± 6.725	72.56 ± 6.192	$F = 1.894$	$p = 0.158$
Years of education (mean ± SD)	11.40 ± 2.041	10.80 ± 2.102	10.92 ± 2.515	$H = 2.437$	$p = 0.296$
MMSE score (mean ± SD)	29.12 ± 0.526	27.04 ± 0.841	24.16 ± 2.135	$H = 59.596$	$p < 0.001$
CDT score (mean ± SD)	8.48 ± 2.064	7.00 ± 2.646	6.12 ± 3.032	$H = 11.828$	$p = 0.003$
ADAS-Cog score (mean ± SD)	7.75 ± 2.543	10.75 ± 2.924	18.08 ± 5.994	$F = 32.824$	$p < 0.001$
GDS score (mean ± SD)	3.32 ± 3.024	5.04 ± 3.272	4.08 ± 2.629	$F = 2.082$	$p = 0.132$

Data

Spontaneous speech recorded

- **immediate recall:** a specially designed one-minute-long animated film
- **previous day:** their previous day
- **delayed recall:** one-minute break then recall the second film

(1) *"I am going to show you a silent movie lasting about a minute. Try to remember the story, the actors, the objects and the places, paying attention to the details".*

(2) *"Please tell me about your previous day in as much detail as you can.*

(3) *"Now, I am going to show you another clip. Try to remember the story, the actors, the objects and the places, paying attention to the details. OK, I am going to start it now".*
The Patient watches the clip. If he starts talking about it, he is reminded that he is not yet allowed to talk about it. When the clip ends:
"Now we will take a one-minute break".
If the Patient starts talking during the break, he is reminded that it is still break-time, and he has to wait until the minute is over. After the one-minute break is over:
"Right, could you please tell me what you saw in the clip?"

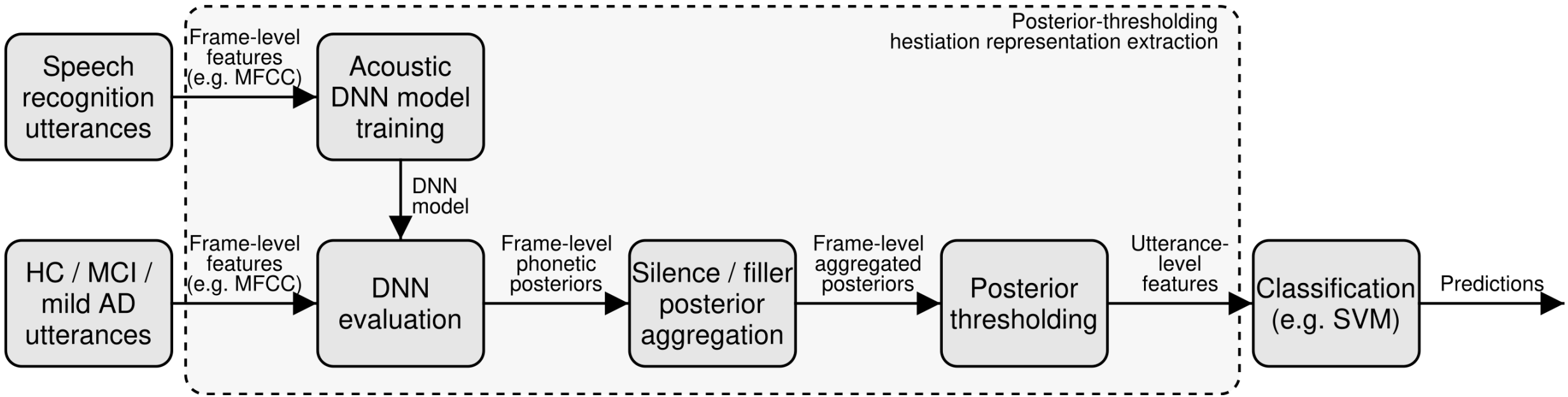
Model

Posterior-thresholding hesitation representation

(1) Frame-level DNN evaluation

(2) Hesitation posterior estimation

(3) Posterior-based utterance-level feature extraction

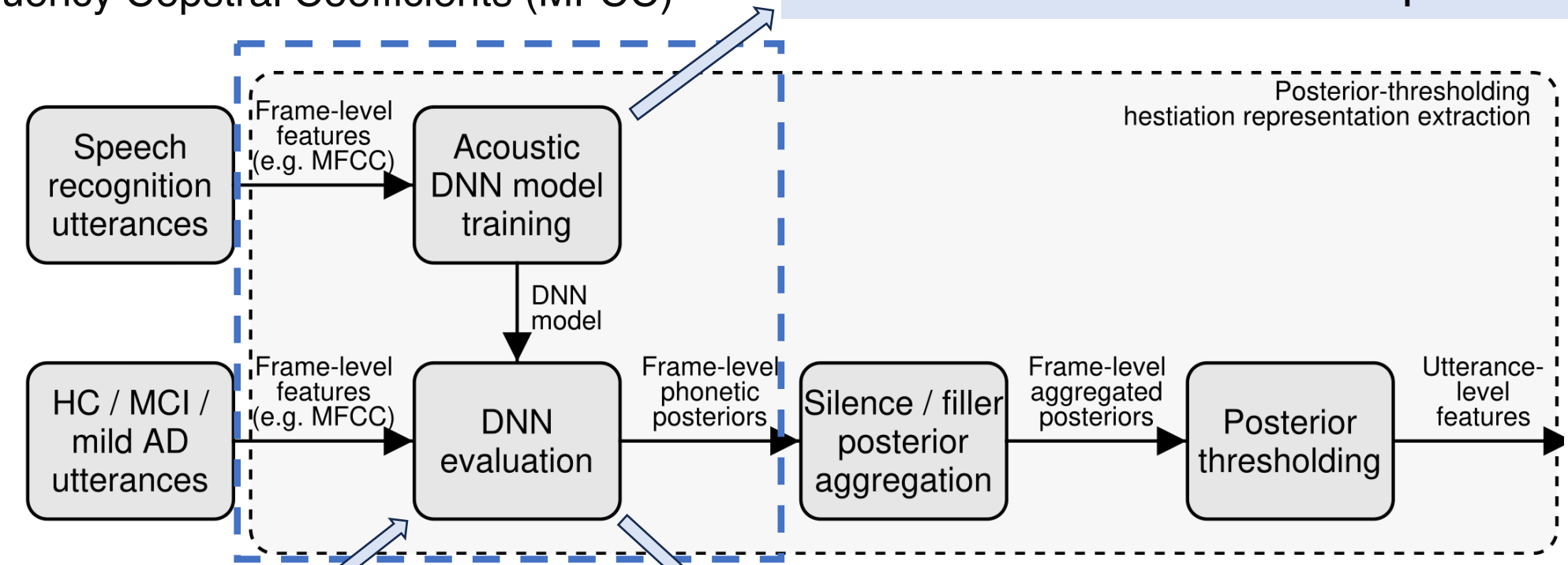


Model

(1) Frame-level DNN evaluation

trained on an audio corpus that contains occurrences of filled pauses

Mel-Frequency Cepstral Coefficients (MFCC)



frame-level features (e.g. MFCCs)

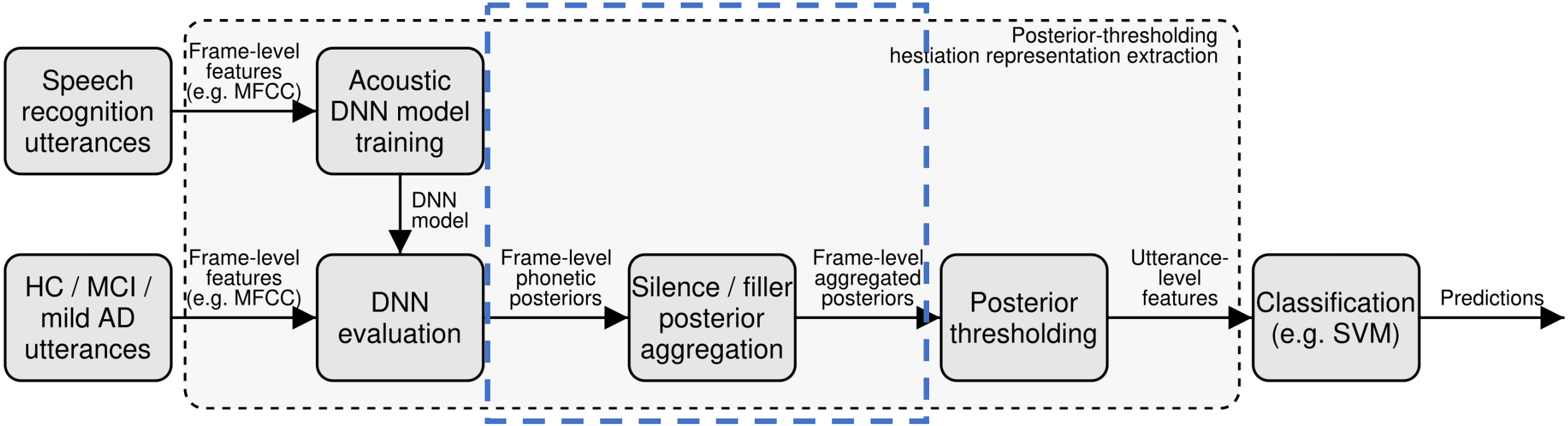
a sequence of frame-level posterior estimate vectors of all phonetic states

Aalto University
School of Electrica
Engineering

frame(i) = […, phonemes, silence, filled pause, laughter, coughs…]

Model

(2) Hesitation posterior estimation



frame(i) = […, phonemes, **silence pause**, **filled pause**, laughter, coughs…]

frame-level posterior estimate vectors of all phonetic states

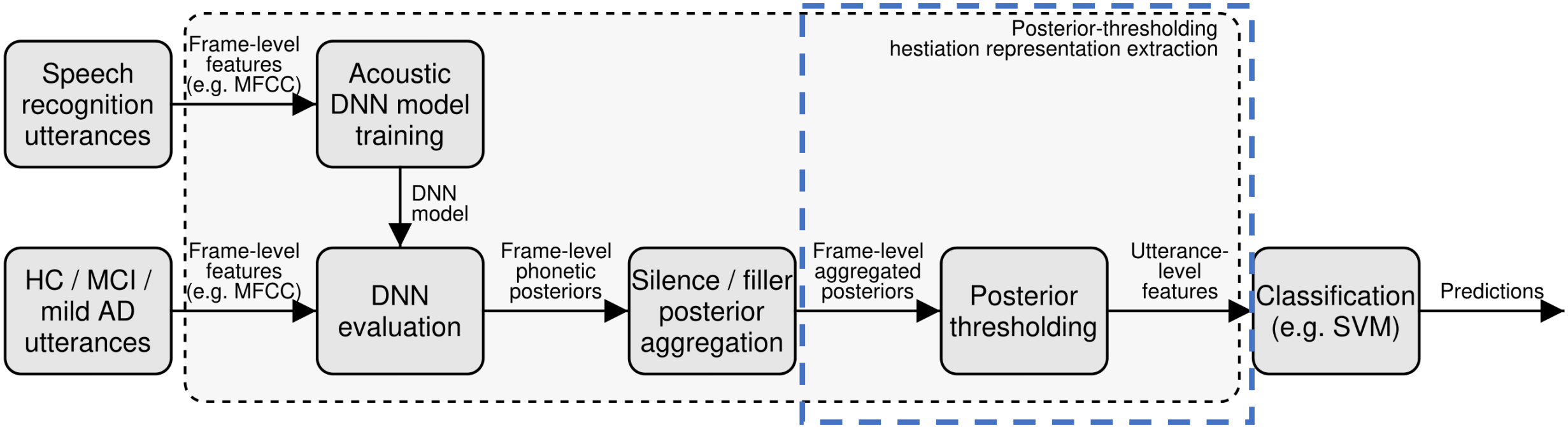
⇒

frame-level posterior estimates of silence and filled pause

Aalto University
School of Electrical Engineering

Model

(3) Posterior-based utterance-level feature extraction



The size of their vector is related to length of the given utterance

frame-level posterior estimates of silence and filler events ⟹ Utterance-level features with a fixed-size vector

Aalto University
School of Electrical Engineering

Model

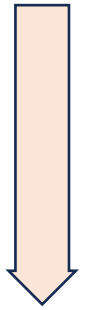
(3) Posterior-based utterance-level feature extraction

A. given threshold value $0 \leq \boldsymbol{th} \leq 1$

B. count the number of frames where the posterior estimate is greater than $\boldsymbol{th}$, then divide this sum by the total number of frames (i.e. we normalize them)

C. repeat $\boldsymbol{m}$ times with different $\boldsymbol{th}$. step size is $\boldsymbol{s}$: $\boldsymbol{s} = \boldsymbol{1/m}$

$$\boldsymbol{frame_{level} = [fram_1, frame_2, \ldots, frame_i, \ldots frame_n]}$$



$\boldsymbol{fram_i}$ represent the posterior estimate of the $\boldsymbol{i}$ frame

$$\boldsymbol{th = \quad (1 * s), (2 * s), \ldots, (i * s), \ldots, 1}$$

$$\boldsymbol{utterence_{level} = [num_1, num_2, \ldots, num_i, \ldots, num_n]}$$

Experiment

(1) The DNN acoustic model

- Trained on a subset of the BEA Hungarian corpus, involving 116 subjects which equated to 44 hours of recordings across 9.7k instances
- Were meticulously annotated to identify various non-verbal vocal sounds like:
  - filled pauses, breathing sounds, laughter, coughs, gasps
  - needs to accurately distinguish hesitations in speech (= indicative of cognitive impairments) from regular speech patterns
- Mel-frequency filter banks, energy features, and their derivatives (1st, 2nd order)
- Context-Independent (CI) VS Context-Dependent (CD) phonetic mappings
  - to explore if simple CI mappings could achieve comparable performance to the more complex CD ones in identifying silent and filled pauses.
  - comprehensive strategy to balance complexity with performance

Experiment

Context-Independent (CI) VS Context-Dependent (CD)

- CI models
  - could potentially lower computational costs
  - simplify the model without significantly compromising accuracy
- CD models
  - more accurate due to their nuanced understanding of phonetic context
  - more complex and computationally demanding

Experiment

(2) Posterior-thresholding hesitation (PTHR) representation

- Focusing on hesitation in speech, specifically silent and filled pauses
- Setting a step size of 0.02
- Yielded 50 features for each hesitation type
- Considering silent pauses (including gasps, breath intakes, and sighs) and filled pauses (treated as special phonetic instances) separately or combined under the term "all hesitation“
- Excluding the segments where the likelihood of silent pauses exceeded a threshold of 0.9
- The PTHR method serves as the foundation for the experiment's feature extraction phase, directly influencing the classification process.

Experiment

(3) Utterance-level Classification

- Performed using the SVM algorithm, suited for datasets with limited data.
  - provides a flexible decision boundary, which is essential when dealing with complex, high-dimensional data derived from speech.
- 25-fold cross-validation method was employed, ensuring each model was trained on the speech of 72 subjects and evaluated on three subjects (one from each category: HC, MCI, and mAD).
- SVM's complexity parameter C was optimized using a nested cross-validation technique to find the value that yielded the highest AUC score.
  - the cross-validation approach ensures robust model training and validation, reducing bias and overfitting.

Experiment

(4) Prediction combination

- Combining predictions from multiple attribute sets (silence-related and filler-related features) to enhance classification performance.
- Weighted mean of the posterior probability estimates from individual classifiers was used to combine predictions.
- Prediction combination is a technique to increase the robustness of classification results by leveraging the strengths of individual feature sets.
  - based on the premise that different features might capture different aspects of the speech patterns indicative of cognitive impairment.
  - may capitalize on the complementary information provided by different hesitation types, leading to a more accurate overall prediction than any single feature set could achieve.
  - system effectively creates a consensus mechanism that could potentially reduce false positives and negatives, leading to more reliable diagnostics.

Experiment

(5) Temporal Speech Parameters (S-GAP)

- Revisited attributes defined in previous research:
  - utterance duration, speech rate, articulation rate, the total length of pauses in relation to the duration, pause rate, the average length of pauses
- Derived from the phonetic decoding output of a phone-level Automatic Speech Recognition system
- Differentiated between using all S-GAP attributes and focusing solely on silence-related attributes
  - utterance duration, silence occurrence rates, average silence length, silence frequency.
- Crucial for providing a baseline comparison for the more focused posterior-thresholding hesitation representation explored elsewhere in the study

Experiment

(6) Evaluation

- Area Under the Receiver Operating Characteristics Curve (AUC) score
  - healthy controls (HC), Mild Cognitive Impairment (MCI), mild Alzheimer's disease (mAD), with a report on the mean of these three scores.
- Information Retrieval metrics - balanced class distribution of the dataset
  - precision, recall, and the F-measure (harmonic mean of precision and recall)
- MCI and mAD categories were combined to form a single positive class against the HC category as the negative one
- These metrics were calculated by setting the decision threshold along with the Equal Error Rate (EER).

Result

(1) Results using the temporal speech parameters (S-GAP)

Features	Speaker task	Classification metrics					Area-Under-Curve			
		Acc.	Prec.	Recall	Spec.	F_1	HC	MCI	mAD	Mean
Silence	Immediate recall	38.7%	75.0%	60.0%	60.0%	66.7	0.637	0.474	0.705	0.605
	Previous day	41.3%	68.4%	52.0%	52.0%	59.1	0.502	0.646	0.562	0.570
	Delayed recall	41.3%	68.4%	52.0%	52.0%	59.1	0.558	0.536	0.774	0.623
	All three tasks	42.7%	77.5%	62.0%	64.0%	68.9	0.619	0.374	0.689	0.561
All	Immediate recall	40.0%	71.8%	56.0%	56.0%	62.9	0.580	0.566	0.743	0.630
	Previous day	42.7%	72.5%	58.0%	56.0%	64.4	0.622	0.611	0.569	0.601
	Delayed recall	50.7%	77.5%	62.0%	64.0%	68.9	0.673	0.592	0.805	0.690
	All three tasks	60.0%	83.7%	72.0%	72.0%	77.4	0.728	0.600	0.780	0.705

- The accuracy metrics obtained with S-GAP temporal speech parameters developed in prior research showed modest classification accuracy overall.
- There was no significant difference in performance between the three speaker tasks (immediate recall, previous day, delayed recall), with accuracy ranging from 40.0% to 50.7%, precision from 71.8% to 77.5%, and recall and specificity between 56% and 64%.
- The immediate recall and delayed recall tasks were more effective in detecting mild AD, with the highest AUC values.

Result

Features	Speaker task	Classification metrics					Area-Under-Curve			
		Acc.	Prec.	Recall	Spec.	F_1	HC	MCI	mAD	Mean
Silence	Immediate recall	38.7%	75.0%	60.0%	60.0%	66.7	0.637	0.474	0.705	0.605
	Previous day	41.3%	68.4%	52.0%	52.0%	59.1	0.502	0.646	0.562	0.570
	Delayed recall	41.3%	68.4%	52.0%	52.0%	59.1	0.558	0.536	0.774	0.623
	All three tasks	42.7%	77.5%	62.0%	64.0%	68.9	0.619	0.374	0.689	0.561
All	Immediate recall	40.0%	71.8%	56.0%	56.0%	62.9	0.580	0.566	0.743	0.630
	Previous day	42.7%	72.5%	58.0%	56.0%	64.4	0.622	0.611	0.569	0.601
	Delayed recall	50.7%	77.5%	62.0%	64.0%	68.9	0.673	0.592	0.805	0.690
	All three tasks	60.0%	83.7%	72.0%	72.0%	77.4	0.728	0.600	0.780	0.705

- The delayed recall task was the most efficient overall, while the previous day task was more useful for the MCI category.
- When predictions from all three tasks were combined, the classification performance improved significantly, with accuracy increasing to 60%, precision to 83.7%, and recall and specificity to 72%, resulting in an F1-score of 77.4.
- The data suggests that while individual tasks provide useful insights, a combined approach yields a more accurate and robust assessment. This aligns with the complexity of cognitive impairments, which may affect various aspects of speech differently. The improvement seen in combined predictions suggests that the interplay between different speech elements could be key in developing effective diagnostic tools for cognitive decline.

Result

(2) Results using the posterior-thresholding hesitation representation with context-dependent states

Speaker task	Features	Classification metrics					Area-Under-Curve			
		Acc.	Prec.	Recall	Spec.	F_1	HC	MCI	mAD	Mean
Immediate recall	Silence	52.0%	77.5%	62.0%	64.0%	68.9	0.587	0.614	0.759	0.653
	Filler	33.3%	68.4%	52.0%	52.0%	59.1	0.533	0.561	0.498	0.530
	All hesit.	46.7%	75.0%	60.0%	60.0%	66.7	0.570	0.687	0.746	0.668
	All	50.7%	78.0%	64.0%	64.0%	70.3	0.580	0.643	0.769	0.664
Previous day	Silence	50.7%	77.5%	62.0%	64.0%	68.9	0.613	0.684	0.594	0.630
	Filler	38.7%	78.0%	64.0%	64.0%	70.3	0.734	0.522	0.510	0.589
	All hesit.	40.0%	64.9%	48.0%	48.0%	55.2	0.415	0.657	0.574	0.549
	All	50.7%	78.0%	64.0%	64.0%	70.3	0.665	0.680	0.610	0.652
Delayed recall	Silence	60.0%	86.4%	76.0%	76.0%	80.9	0.755	0.670	0.746	0.724
	Filler	33.3%	68.4%	52.0%	52.0%	59.1	0.455	0.497	0.455	0.469
	All hesit.	68.0%	88.9%	80.0%	80.0%	84.2	0.857	0.706	0.802	0.788
	All	68.0%	89.1%	82.0%	80.0%	85.4	0.842	0.700	0.775	0.773
All three tasks	Silence	61.3%	86.4%	76.0%	76.0%	80.9	0.773	0.683	0.734	0.730
	Filler	41.3%	81.0%	68.0%	68.0%	73.9	0.758	0.566	0.526	0.617
	All hesit.	68.0%	88.9%	80.0%	80.0%	84.2	0.854	0.712	0.806	0.791
	All	70.7%	93.5%	86.0%	88.0%	89.6	0.911	0.709	0.794	0.804

Interpretation

- Silent pause-related attributes in the immediate recall speaker task exhibited an acceptable performance with a 52% accuracy and an F1 score of 68.9, which are above random guessing, and a mean AUC of 0.653.
- Filler events were not as useful for detecting MCI and mAD in the immediate recall task, suggesting they do not present a reliable pattern for these conditions in this context.
- Combining silent and filled pauses ('All hesitation' case) showed similar values to the silent pause case alone. However, using all three attribute types together resulted in a slight improvement across all metrics.

Interpretation - continued

- For the previous day task, filled pauses yielded higher scores than in the immediate recall, indicating different hesitation patterns among the subject types for this task.
- In the delayed recall task, silent pauses were more useful than filled pauses, with the 'All hesitation' case showing a slightly more successful outcome.
- Combining predictions across all three speaker tasks led to a general improvement, although slight in most cases.

Result

(3) Results using the posterior-thresholding hesitation representation with context-independent states

Speaker task	Features	Classification metrics					Area-Under-Curve			
		Acc.	Prec.	Recall	Spec.	F_1	HC	MCI	mAD	Mean
Immediate Recall	Silence	50.7%	77.5%	62.0%	64.0%	68.9	0.577	0.590	0.758	0.642
	Filler	30.7%	64.9%	48.0%	48.0%	55.2	0.467	0.509	0.553	0.510
	All hesit.	48.0%	75.0%	60.0%	60.0%	66.7	0.574	0.693	0.769	0.678
	All	50.7%	77.5%	62.0%	64.0%	68.9	0.580	0.597	0.759	0.645
Previous day	Silence	46.7%	77.5%	62.0%	64.0%	68.9	0.618	0.648	0.550	0.605
	Filler	40.0%	81.0%	68.0%	68.0%	73.9	0.749	0.516	0.454	0.573
	All hesit.	33.3%	58.3%	42.0%	40.0%	48.8	0.373	0.640	0.557	0.523
	All	49.3%	80.5%	66.0%	68.0%	72.5	0.659	0.645	0.561	0.622
Delayed recall	Silence	58.7%	86.4%	76.0%	76.0%	80.9	0.772	0.690	0.750	0.738
	Filler	33.3%	67.6%	50.0%	52.0%	57.5	0.478	0.517	0.488	0.494
	All hesit.	62.7%	86.0%	74.0%	76.0%	79.6	0.778	0.684	0.798	0.753
	All	62.7%	86.4%	76.0%	76.0%	80.9	0.786	0.685	0.794	0.755
All three tasks	Silence	62.7%	88.9%	80.0%	80.0%	84.2	0.786	0.693	0.746	0.742
	Filler	40.0%	81.4%	70.0%	68.0%	75.3	0.680	0.519	0.540	0.580
	All hesit.	64.0%	86.4%	76.0%	76.0%	80.9	0.772	0.696	0.802	0.757
	All	69.3%	91.3%	84.0%	84.0%	87.5	0.866	0.703	0.770	0.780

Interpretation

- The results obtained using the context-independent (CI) DNN acoustic model show very similar trends to those with the context-dependent (CD) model. Silent pauses remained a stronger feature than filled pauses for identifying mAD subjects.
- For the immediate recall task, silent pauses were more useful than filled pauses, with mAD subjects identified more precisely than HC or MCI subjects.
- In the previous day task, silent pauses and filled pauses performed similarly, with filled pauses resulting in a high AUC for HC subjects.
- The delayed recall task was the most informative when silent pauses and all hesitations were considered.
- The use of CI DNN models led to only a slight decrease in performance scores, or none at all, compared to the CD models.

Aalto University
School of Electrical
Engineering

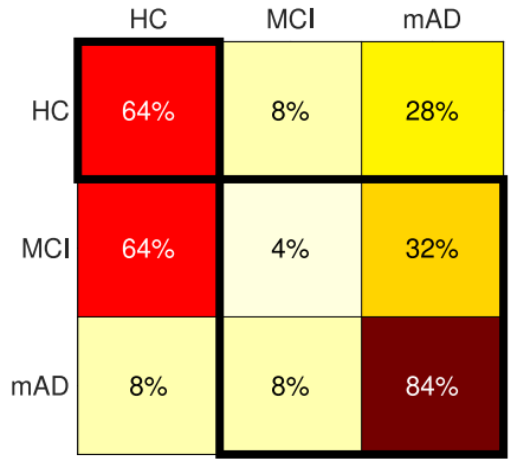
Evidences

- Similar accuracy rates for silent pauses across both CD and CI models, with 52.0% for CD and 50.7% for CI in immediate recall.
- High AUC scores for filled pauses in the previous day task, notably 0.749 for HC in the CI model.
- The delayed recall task showed strong performance for silent pauses and all hesitations in both models, with an F1 score of 85.4 for CD and 80.9 for CI.
- Generally, a slight decline or maintenance of metric scores when using CI models compared to CD models.

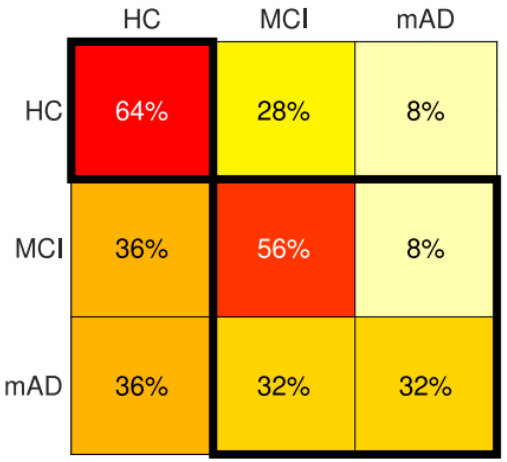
Aalto University
School of Electrical
Engineering

Result

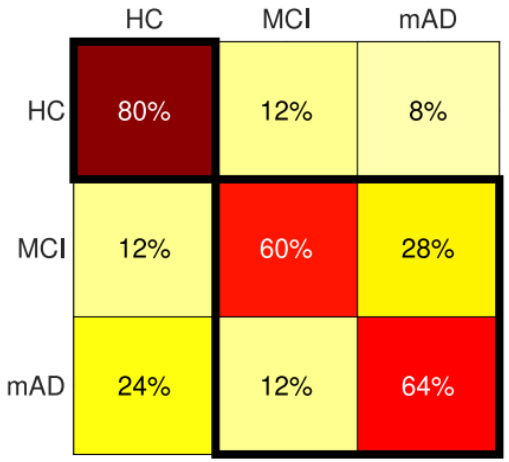
(4) The performance of speaker tasks and feature subsets



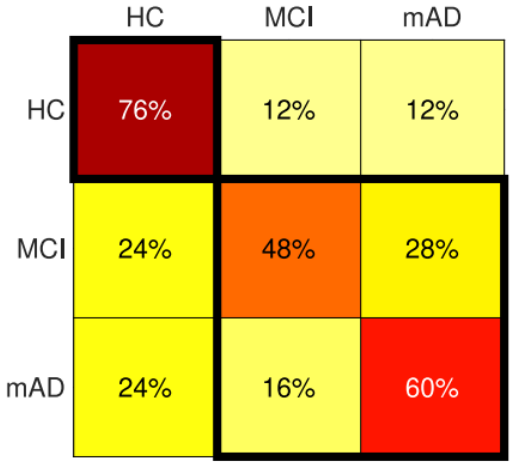
(a) Immediate Recall



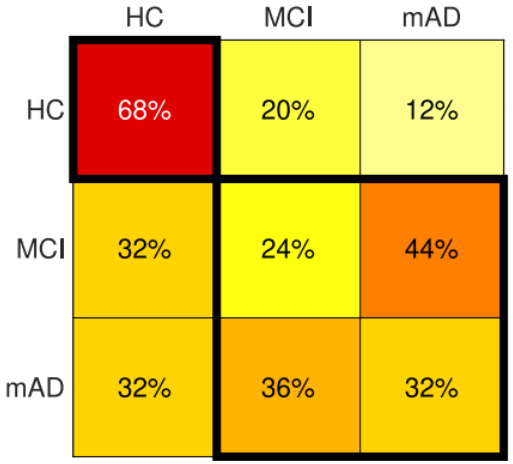
(b) Previous Day



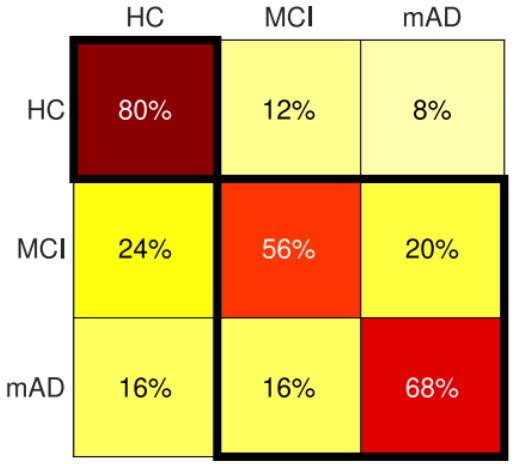
(c) Delayed Recall



(a) Silent Pause



(b) Filled Pause



(c) All Hesitation

Aalto University
School of Electrical
Engineering

What can we observe?

- For immediate recall tasks, mAD is well-identified, but MCI is often misclassified as HC, suggesting that this task is insufficient for detecting MCI.
- Previous day tasks show improved identification of MCI but poor detection of mAD.
- Delayed recall tasks are more balanced in identifying HC, MCI, and mAD.
- Silent pauses distinguish HC well but have a lower rate of correctly identifying MCI and mAD.
- Filled pauses are less effective, particularly for MCI and mAD.
- Combining silent and filled pauses improves the identification of all categories.

Further interpretation

- The ability of the immediate recall task to identify mAD but not MCI may reflect the more pronounced speech disruption in mAD.
- The previous day task may engage memory retrieval differently, hence the improved MCI detection.
- Delayed recall's effectiveness might be due to it challenging both recent and working memory, revealing deficits across HC, MCI, and mAD.
- The analysis suggests that the complexity of the recall task and the type of hesitation examined critically influence the classifiers' ability to distinguish between HC, MCI, and mAD.
- Silent pauses may be indicative of cognitive processing delays more prevalent across all impairment levels, while filled pauses may not be as diagnostically significant.
- The varying effectiveness of each task and feature subset underscores the need for multi-faceted approaches in cognitive impairment screening tools.

Conclusion

- The paper proposed a feature extraction approach that detects MCI and mild AD using DNN outputs for silent and/or filled pauses.
- Achieved a best accuracy score of 69.3%, an F1 value of 87.5, and a mean AUC score of 0.780.
- Competitive results compared to other studies using different methods and datasets.
- Showed that context-independent DNN models could achieve comparable performance to context-dependent models.
- Delayed recall was the most effective task for identifying all speaker groups.
- Silent pauses were the most indicative of mild Alzheimer's, while filled pauses were less effective.
- Combining different hesitation features resulted in better classification performance.
- Practical application favored the use of previous day task due to ease of recording and effectiveness in early MCI detection

Shortness

- Comparison with traditional diagnostic methods not extensively explored.
- The complexity of the method's implementation in real-world settings is not fully addressed

Possible future works

- Future research could explore the method's applicability to other neurodegenerative diseases.
- Conducting the study with a larger, more diverse dataset could help in validating the robustness of the model and its applicability across different demographics.
- Tracking subjects over time to observe how speech patterns evolve with the progression of cognitive decline would provide deeper insights and improve predictive modeling.
- Explore if combining speech analysis with other biomarkers (e.g., imaging, cognitive tests) could improve diagnostic accuracy and provide a more holistic view of the patient's condition.

Questions

1. How is the Posterior-based utterance-level feature extraction process applied to transform frame-level posterior estimates into a fixed-size feature vector for utterance-level classification, and what role does the step size parameter (s) play in this process?

2. Discuss the role of silent and filled pauses in speech as potential biomarkers for detecting Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD). How does the posterior-thresholding hesitation representation technique aid in distinguishing between healthy controls, MCI, and AD patients?

3. Explorative question: The study presented classification accuracies for three distinct groups: HC, MCI, and mAD. However, cognitive decline is a continuum. How might the model perform on individuals who are at the borderline between these defined categories? Would the model's performance degrade gradually or show a sharp threshold effect? Try to derive your assumption without actually implementing any models and provide your justification or other relevant published researches as references.

ELEC-E5531
Speech and Language Processing Seminar

Aalto University
School of Electrical
Engineering

Thank you!

Ying Liu and Zehang Li