Detecting Dementia from Long Neuropsychological Interviews

Nauman Dawalatabad
MIT CSAIL
nauman@mit.edu

Yuan Gong
MIT CSAIL
yuangong@mit.edu

Sameer Khurana
MIT CSAIL
skhurana@mit.edu

Rhoda Au
Boston University
rhodaau@bu.edu

James Glass
MIT CSAIL
glass@mit.edu

Abstract

Neuropsychological exams are commonly used to diagnose various kinds of cognitive impairment. They typically involve a trained examiner who conducts a series of cognitive tests with a subject. In recent years, there has been growing interest in developing machine learning methods to extract speech and language biomarkers from exam recordings to provide automated input for cognitive assessment. Inspired by recent findings suggesting that the examiner's language can influence cognitive impairment classifications, in this paper, we study the influence of the examiner on automatic dementia identification decisions in real-world neuropsychological exams. To mitigate the influence of the examiner, we propose a systematic three-stage pipeline for detecting dementia from exam recordings. In the first stage, we perform audio-based speaker diarization (i.e., estimating who spoke when?) by incorporating speaker discriminative features. In the second stage, we employ text-based language models to identify the role of the speaker (i.e., *examiner* or *subject*). Finally, in the third stage, we employ text- and audio-based models to detect cognitive impairment from hypothesized subject segments. Our studies suggest that incorporating audio-based diarization followed by text-based role identification helps mitigate the influences from the examiner's segments. Further, we found that the text and audio modalities complement each other, and the performance improves when we use both modalities. We also perform several carefully designed experimental studies to assess the performance of each stage.

1 Introduction

Cognitive impairment is the condition where a human may experience cognitive decline in mental ability (e.g., Dementia and Alzheimer's disease). Alzheimer's disease is the most common form of dementia. Early detection of cognitive impairment may lead to a better lifestyle and pathways to treatment (Szekely et al., 2004; Chuang et al., 2016). Neuropsychological exams are a widely used and effective technique in assessing a subject's cognitive status (Kurlowicz and Wallace, 1999). These exams often serve as the first-stage screening tool to identify cognitive impairment before performing more expensive laboratory tests (e.g., brain imaging) (Weinstein et al., 2014) with an average estimated cost of $5,000 per brain scan. An accurate first-stage assessment is crucial as it helps avoiding additional cost and time burdens on healthcare system and the patient.

For a typical cognitive interview evaluation, a trained *examiner* interacts with a *subject* (patient) by conducting a series of tasks that are designed to assess the memory, attention, visuoperceptual, reasoning, language, and verbal skills of the subject (Alhanai et al., 2018; Alhanai, 2019). During this evaluation session, the examiner asks a series of questions related to different cognitive domains. These questions include identifying objects in an image, repeating the numbers, words, or story narrated by the examiner. Based on the answers to the questions posed by the examiner, the interview exam is assigned a set scores that reflect the subject's cognitive status. This is a long process, and each interview typically lasts for more than an hour, with interview recordings typically containing approximately equal amounts of speech from the examiner and subject.

1.1 Motivation

Recently, there is a growing interest in the speech community on learning speech-based biomarkers for cognitive impairment. Despite the success of existing automatic speech/text-based cognitive assessment models in (Alhanai et al., 2017; Pappagari et al., 2020; Balagopalan et al., 2020; Haulcy and Glass, 2021; Pérez-Toro et al., 2021), they are mostly based on either manually curated sub-

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5299 - 5312

Task	Examiner	Subject	
		Healthy	Dementia
Question	And how far did you go in school?	College.	<um> four-year college
Question	And what was that in?	I got an associates degree and a bachelor of fine arts.	let's see bachelor's degree, what did take that for, <um> this was, it was with children.
Repeating story	Anna Thompson of South Boston, employed as a scrub woman in an office building, reported at the City Hall Station that she had been held up on State Street the night before and robbed of fifteen dollars...	Anna Thompson who lived in South Boston, who worked in a state house was on her way home from work and she got robbed and she has...	Anna Thompson <um> was walking down State Street and she was <um> robbed of her purse and the policeman, she had four children with her, and the police didn't want, and <um>...
Repeating numbers	Three, nine, two, four, eight, seven.	Three, nine, two, four, eight, seven	Four, nine, three, eight, two, seven.

Table 1: Interview question and answer samples. **Key observations**: (i) Examiner and subject, in general, speak differently. (ii) Vocabulary between examiner and subject in multiple sentences may be similar, especially in the story/number "repeating" tasks. (iii) Subjects with dementia uses more filler words (e.g., "<um>") compared to healthy subject and they tend to speak broken sentences in non-fluent manner while repeating certain words.

ject speech or the entire interview exam, including the examiner's speech. Also, (Pérez-Toro et al., 2021) recently found that merely using the segments from the examiner sections can also lead to high accuracy for cognitive impairment identification. They observe this behaviour on the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge dataset (Luz et al., 2020) which consists of manually curated short speech segments, typically around one minute long, with the text transcriptions for the task of cookie theft description. The examiner's influence is potentially due to the examiner's reaction containing cues for cognitive impairment assessment. For instance, the examiner may provide some helping clues to cognitively impaired subjects to arrive at the correct answers (Egan et al., 2010). However, an automatic cognitive identification system should make decisions based on the subject, not the examiner. Hence it is desirable to process the segments of the exam belonging only to the subject.

Our goal in this work is to automatically identify dementia using text and audio from a long neuropsychological interview exam consisting recordings of both, the examiner and the subject. A related previous work by (Alhanai et al., 2018) proposed a two-stage approach using role-specific language models (LMs) to (i) diarize and identify the subject segments, and (ii) use handcrafted acoustic features with logistic regression for dementia identification. However, they did not consider the possibility of the examiner's influence on the final cognitive identification decision and ended up employing a sub-optimal diarization module. We use the work by (Alhanai et al., 2018) as our baseline.

1.2 Our Contributions

To address the aforementioned problems, we propose a three-stage pipeline where we aim to: (i) Diarize ("who spoke when"?) the neuropsychological exam, (ii) identify the regions belonging to subject and examiner (Role ID), and then (iii) detect cognitive impairment (Cognitive ID) using hypothesised subject's segments. Our work defers from the baseline system and we propose several enhancements as follows:

- **Text-speech force alignment**: The authors in (Alhanai et al., 2018) use ground truth speaker boundaries that have only speaker start timestamps. They assume the speaker's end timestamp as the start timestamp of the next speaker. This assumption is not optimal as there can be an inter-speaker silence regions. This is also insufficient for an accurate evaluation of the performance of diarization. We obtain a better speaker segmentation by force aligning text and audio.

- **Study**: We study examiner's influence on cognitive ID in full long interview exams and propose a system to mitigate this influence.

- **Diarization**: (Alhanai et al., 2018) used a single-stage role-specific language model for performing both diarization and role ID. This is sub-optimal because, (i) As shown in Table 1, although there are differences in examiner's and subject's vocabulary, there are multiple segments in the interview exam where the vocabulary of examiner and subject might be similar (e.g., a task of repeating a story or numbers). The language models may fail to properly distinguish the roles while doing segment-wise diarization. (ii) Speaker discriminative information present in audio is not used in diarization.

  Hence, unlike previous work by (Alhanai et al., 2018), we propose to decouple this sin-

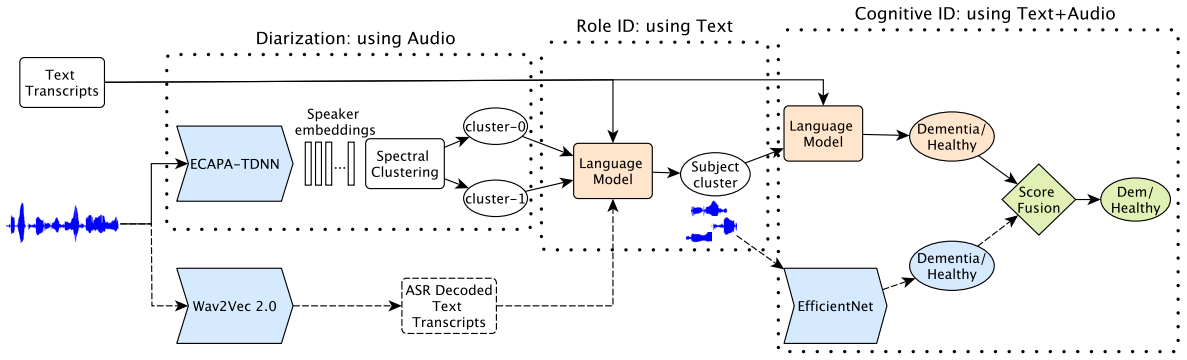


Figure 1: The proposed 3-stage cognitive ID system. (i) Audio-based diarization using ECAPA-TDNN embeddings and spectral clustering. (ii) Text-based role ID (examiner/subject) using LMs trained on manual or ASR decoded transcripts. (iii) Text- and audio-based cognitive ID with LMs and EfficientNet model, respectively. The scores from text and audio are fused in the late score fusion stage for making the final cognitive decision.

gle stage into two separate stages. We propose the diarization stage using audio, and the role identification stage using text-based role-specific language models.

- **Models**: For the diarization stage, we employ Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network (ECAPA-TDNN) to obtain robust speaker discriminative embeddings from audio (Desplanques et al., 2020). Apart from text-based role-specific LMs for role identification, we also employ wav2vec 2.0 self-supervised model to convert audio to text for training role-specific language models (Baevski et al., 2020).

- **Cognitive ID**: Motivated by differences in vocabulary of healthy subjects and the subjects with dementia (Table 1, observation (iii)), we propose text-based cognitive ID using LMs trained along with the filler words. Further, we propose cognitive ID from audio using EfficientNet model (Gong et al., 2022). Finally we merge cluster-level scores from the text and audio modalities in using late score fusion to further improve the performance.

- **Dataset curation**: Our system automatically curates the conversation dataset at all stages, i.e., diarization (speaker segmentation), role ID (speaker identity), ASR (speech to text), and cognitive ID (subject's cognitive condition) which is useful for research in these domains.

2 Text and speech forced alignment

We use the gold standard Framingham Heart Study (FHS) dataset as used by (Alhanai et al., 2018). FHS dataset comprises a total of 92 interview recordings with orthographically transcribed text. In the original (*orig*) annotations for the FHS dataset, only the start timestamps of the speaker's segments are marked by the annotators. The end timestamp for the current speaker segment is assumed to be the same as the start timestamp of the next speaker segment. However, this is insufficient for an accurate evaluation of diarization performance as silence regions are considered as speech. Hence, we force align text and speech for each of the *orig* segment to obtain the word level alignments *FAligned*.

To force align text and speech, we employ Gaussian Mixture Model/ Hidden Markov Model (GMM/HMM) Automatic Speech Recognition (ASR) pipeline using Montreal Forced Aligner (MFA) toolkit (McAuliffe et al., 2017). During training, mono-phone GMM models are iteratively trained to generate primary alignments. Then triphone GMMs are trained to account for the context phones. We obtain the end timestamps of each of the *orig* segments as the end timestamp of the last word in corresponding *FAligned* word sequence. This removes the inter-speaker silences. The final boundaries are used for evaluation purpose. An illustration is given in the Appendix (Figure 8).

3 Methodology

Figure 1 shows the proposed three stages for detection of dementia from the long neuropsychological

interview examinations, including (i) Diarization using audio, (ii) Role ID using text, and (iii) Cognitive ID using both text and audio.

The diarization stage takes input long interview recording and divides it into smaller chunks and processes using the ECAPA-TDNN model to extract speaker discriminative embeddings. These embeddings are clustered using spectral clustering to obtain two clusters. Both clusters are fed to the role ID module to identify the cluster belonging to the subject. The hypothesized subject cluster is then used to identify dementia using text and audio. The scores from text and audio are fused to make the final cognitive ID decision.

3.1 Speaker diarization using audio

In order to partition different speakers into different clusters, the audio segment representations must capture speaker discriminative information. The ECAPA-TDNN model (Desplanques et al., 2020) has shown impressive performance on speaker diarization of meeting recordings (Dawalatabad et al., 2021). We extend the ECAPA-TDNN model embeddings to diarize the FHS neuropsychological exams. The long recording is segmented into smaller chunks such that each segment can be of length $maxSegLen$ or smaller. The extracted ECAPA-TDNN embeddings from these short chunks are clustered using spectral clustering as given in (Dawalatabad et al., 2021).

3.2 Role ID using text

The output of the diarization system consist of relative cluster labels (*cluster-0* and *cluster-1*). Hence, we need a role ID module to identify which cluster belongs to the subject. Due to differences in vocabulary of the examiner and subject, we train an n-gram word-based role-specific LM for each of the roles (i.e., examiner and subject) using the text sentences. The LMs are trained using the KenLM toolkit (Heafield, 2011). Classification of the diarized clusters into examiner and subject is done using Log-Likelihood Ratio (LLR) test at the cluster-level. A score for a cluster is estimated by averaging all the scores of segments in that cluster. The decision threshold on the development set is tuned based on the standard Youden's J statistics as:

$$thr = \arg\max_{t}(Sensitivity_t + Specificity_t - 1) \quad (1)$$

where, thr is the best score threshold, $t \in \{t_1, \ldots, t_n\}$, and t_i is the i-th score threshold.

We employ two strategies for role identification: (i) *Text Transcripts*: The given manually transcribed text is used to train LMs and role identification. (ii) *ASR Decoded Text Transcripts*: Here, we use a popular wav2vec 2.0 model (Baevski et al., 2020) to convert audio into text. The decoded text is used to train LMs and for role identification.

3.3 Cognitive ID using text and audio

Text: We observe that subjects with cognitive impairments often use some filled pause words (for example "<um>", "<uh>"). Apart from these differences, the subjects with dementia tend to speak sentences in a broken form and a non-fluent manner as they slowly try to recall the story/words/numbers, as shown in Table 1 (observation (iii)). These sentences are different from fluent sentences spoken by healthy subjects. Hence, we propose to use n-gram LMs to capture these differences between dementia and healthy subjects using the text transcriptions along with the filled pauses. A hypothesized subject cluster is used to calculate a classification score with respect to the dementia class.

Audio: Apart from the text cues as used above, audio contains biomarkers that distinguish dementia subjects from healthy subjects (Alhanai et al., 2017). Hence we also use audio for cognitive ID. Unlike previous works that use handcrafted features, we propose to use Convolutional Neural Network (CNN) based model (LeCun and Bengio, 1995; Trigeorgis et al., 2016) to learn these features automatically from the audio. Specifically, we use an EfficientNet model (Tan and Le, 2019) that has shown impressive performance for speech and vocal sound classification tasks (Gong et al., 2021, 2022).

Text+Audio: A score for hypothesised subject cluster (text- and audio-based) is obtained by averaging scores over segments in the cluster. We merge the scores obtained from the text- and audio-based classifier using a late score fusion approach:

$$f_{score} = w.t_{score} + (1 - w)a_{score} \quad (2)$$

where t_{score}, and a_{score} are scores from text and audio respectively. f_{score} denotes fused score and w is weight assigned to the text score such that $w \in [0,1]$. The final cognitive impairment decision is taken on the fused scores.

4 Experimental Setup

4.1 Data split

There are 92 interview recordings with orthographically transcribed text in the FHS dataset (Alhanai et al., 2018). We use a development set (dev set) with 42 recordings (with nine dementia cases) for tuning hyperparameters of the diarization system, training role-specific LMs, and finetuning the wav2vec 2.0 model. The evaluation set (eval set) with 50 recordings (with 12 dementia cases) is kept blind for evaluation purposes. Interview recordings were randomly selected to be part of the Dev or Eval sets. Due to very few recordings for the dementia class (overall 21/92 cases), similar to (Alhanai et al., 2018) we employ leave-one-out cross-validation (Loocv) for evaluating the cognitive ID module.

4.2 Evaluation setup

Diarization setup

Diarization performance is evaluated under two conditions: (i) *Oracle VAD (OrcVAD)*: speech/non-speech details are obtained from the ground truth, and (ii) *Estimated VAD (EstVAD)*: A Voice Activity Detection (VAD) system is used to obtain speech/non-speech details from audio. Diarization Error Rate (DER) is a sum of Missed Speech (MS), False Alarms (FA), and Speaker Error Rate (SER) (Xavier Anguera, 2008). MS and FA denote errors made by the VAD system. Since our goal is to improve clustering in diarization, following a standard procedure similar to (Alhanai et al., 2018; Pal et al., 2019; Dawalatabad et al., 2021), the speech/non-speech labels are taken from the ground truth for the OrcVAD case. For the EstVAD case, we tune robust VAD (rVAD) system (Tan et al., 2020) such that MS is very small and we allowed false alarm silence. This is important in our context as for the role ID and cognitive ID stage, we are primarily concerned with the spoken content, and hence we do not miss any speech regions. We use SER for evaluating the diarization module using an open-source National Institute of Standards and Technology (NIST) evaluation toolkit (Neville Ryant, 2018). The configuration used for the evaluation toolkit is the same as (Alhanai et al., 2018).

Role ID and cognitive ID setup

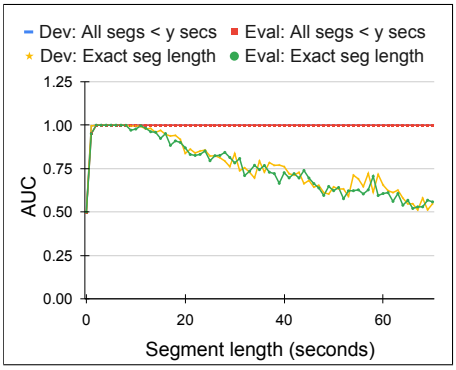
We conducted experiments for role ID using given text transcripts (denoted as *Text*) and also using text generated from the ASR wav2vec 2.0 model (denoted as *ASR*). For cognitive ID, we use *Text* as the ASR system may not generate filled pause words accurately. For both role ID and cognitive ID, we use Area Under the Receiver Operating Characteristic Curve (AUC) as evaluation metric (Huang and Ling, 2005) which is also used by (Alhanai et al., 2018). It gives a complete picture of the system's performance under different thresholds unlike other metrics (accuracy or F1-score) that operate only on specific thresholds. As diarized cluster may also contain segments from the other speaker, the target cluster labels for AUC calculation (for role ID) are obtained by mapping the diarized clusters and the actual clusters using a duration-based cluster matching algorithm provided in the NIST evaluation toolkit (Neville Ryant, 2018). Since the diarized cluster may have segments from the other speaker, for an accurate estimate of the subject cluster, we also report the cluster purity of the hypothesized subject's clusters. The percentage of examiner's segments present in the cluster is considered as the impurity for that cluster. Notice that the cluster purity is a better measure in this context compared to segment-wise accuracy, as the latter does not account for the varying duration nature of the segments.

5 Results, Analysis and Discussion

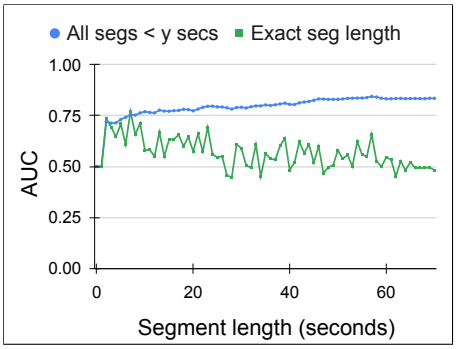
5.1 Baseline system

The evaluations for the baseline system by (Alhanai et al., 2018) are performed under the following three cases: [*B0*]: *Oracle (also topline)*: The role-specific language models are trained on the whole dataset (92 exams), and utterances were segmented according to the ground truth speaker turns. Testing is performed on the same dataset. [*B1*]: *Leave-one-out (Loocv+oracleseg)*: A language model is trained on all transcripts except the one being processed (i.e., 91 exams), and utterance speaker turns are again taken from the ground truth. [*B2*]: *Leave-one-out+autoseg (Loocv+autoseg)*: The language model is trained on all transcripts except the one being processed. The segmentation is estimated using the ASR system.

Note that a strong assumption is used by all the above mentioned baseline systems that there are no silence regions during the whole interview conversation. This may finally lead to over-optimistic cognitive ID decisions. Although it is incorrect to use the test data during training by the baseline



(a) Role ID



(b) Cognitive ID

Figure 2: Analysis on subject segments taken from ground truth for role ID and cognitive ID. The AUC when decision using LMs is taken based on (i) All the segments smaller than certain length (say, *y* sec). (ii) Only using segment with a specific length. The former condition (i.e. (i)) show better AUC.

work (i.e., *B0*), it gives an estimate of the hypothetical best performance that can be achieved by the baseline system (i.e., a topline performance).

5.2 Study on role ID and cognitive ID

The primary motivation for our work is (i) Examiner's segments influence the final cognitive ID decision, so we need to mitigate this influence. (ii) The baseline work by (Alhanai et al., 2018) does not consider the examiner's influence and ends-up employing a poor diarization system. We designed the following two experiments to study this.

Analysis on role ID

This experimental study assumes text transcripts, segmentation, and speaker IDs from the ground truth. Experiments are performed under two conditions: (i) Role ID score is estimated using all segments less than a certain length. (ii) Role ID score is calculated on segment of specific length (say, *y* sec). Segment lengths are quantized to 1 sec. For example, the segment of 5 seconds means segments in the range of 4-5 seconds. This study is designed to understand if role ID can reliably be performed only using some segments.

Segments from	Subject	Examiner	All
AUC	0.816	0.865	0.871

Table 2: A study on cognitive ID. Examiner's segments contain cues for cognitive ID.

We consider examiner/subject clusters from ground truth and use n-gram word-based LMs for the role ID of a given cluster. The AUC versus segment length is shown in Figure 2a. There are two key observations for condition when role ID score is calculated on segment of specific length: (i) It can be seen that the segments of different lengths behave differently. The AUC is high for smaller segments. This is expected as most short-length segments from the examiner are question-type sentences like *"what is your highest degree?"* which are quite different from the subject's vocabulary. (ii) The longer segments (though very few in number), on the other hand, show low AUC. This is due to the fact that the vocabulary of the examiner and subject can be very similar in longer segments, e.g., the task of repeating the story narrated by the examiner (see Table 1, point (ii)). Therefore, text-based role-specific LMs used in baseline system might not be an optimal approach to diarization. This is also evident from the baseline's poor diarization performance (as discussed later in Table 3). Hence, we propose diarization using speaker discriminative characteristics in audio.

It can be seen that the AUC stays high (1.0) for the condition when the decision is taken over multiple segments (or even all segments) of different lengths. Hence, we propose a text-based role-specific LMs for role ID using all segments in a cluster obtained from diarization module.

Analysis on cognitive ID

We repeat a similar experimental study for cognitive ID (Figure 2b) in leave-one-out (Loocv) setup. Cognitive ID is relatively difficult to estimate compared to role ID. The highly peaky nature of the curve when using exact segment length (green curve) indicates that the subject does not express dementia features all the time. Making a decision on multiple segments shows improved AUC. Hence we make collective decision using a score averaged over all segments.

Table 2 shows performance of cognitive ID using

System/Condition	Dev-set	Eval-set	Combined-set
Baseline			
B0: Oracle (topline)	-	-	31.7
B1: Loocv+oracleseg	-	-	32.9
B2: Loocv+autoseg	-	-	36.7
Ours			
Oracle VAD	19.25	18.45	**18.82**
Estimated VAD	28.5	26.1	**27.2**

Table 3: Performance of diarization system in SER. The lower SER value is better.

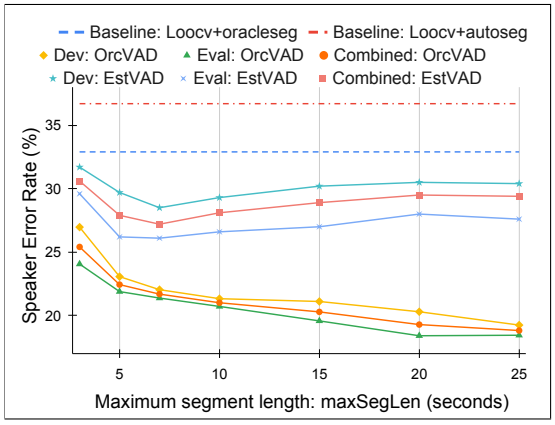


Figure 3: Diarization performance: SERs for different maximum segment lengths under oracle and estimated VAD segmentation. Increasing segment length improves SER for OrcVAD while best SER for EstVAD is observed at 7 secs. Lower SER value is better.

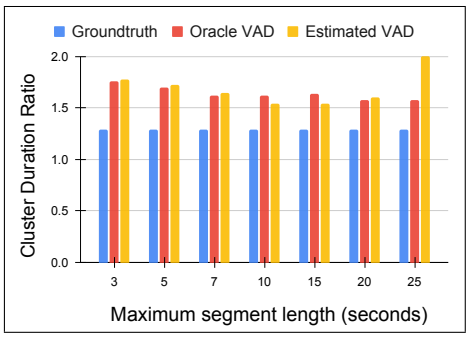


Figure 4: The ratio of the average duration of the large cluster and the smaller cluster on dev set. Both oracle and estimated VAD cases show comparable duration ratios and closer to the ground truth duration ratio.

segments from different roles (Subject/ Examiner/ All). This experiment analyzes the influence of the examiner's segments on cognitive ID decision. It can be seen that the cognitive ID can be estimated from the examiner's segments with a reasonably high AUC. When the examiner's segments are mixed with the subject (*All*), it still shows a high AUC. Hence, the final cognitive ID performance can be sub-optimal (or over-optimistic) if examiner segments get mixed with subject segments. To mitigate the examiner's influence, we propose a better diarization system that helps to partition the examiner's and subject's segment with high accuracy.

5.3 Diarization performance

Table 3 shows the results obtained by our audio-based diarization module for different VAD conditions. It can be seen that our systems outperform baseline systems by a significant margin. Figure 3 shows performance of the diarization system under different maximum segment lengths ($maxSegLen$ as defined in Section 3.1). Diarization with oracle VAD (OrcVAD) segments shows improved performance with increasing $maxSegLen$. This is because of the TDNN layer in ECAPA-TDNN, the longer-range dependencies help retain speaker's information. For the estimated VAD (EstVAD) case, we obtain the best performance with 7-second segments as too long segments tend to contain silences and other sounds, possibly corrupting the speaker embeddings. The proposed systems outperform baseline systems irrespective of segment length.

It is important to check if there are enough segments in each of the diarized clusters to make sure diarization does not yield an empty cluster or a cluster with too few segments. In Figure 4 we show the ratio of the duration of the large cluster to the duration of the smaller cluster. It can be seen that both proposed systems show cluster duration ratios similar to that in the ground truth irrespective of $maxSegLen$. For the estimated VAD case, a very high ratio for the segment size of more than 25 sec is observed. This is due to more silence/noise getting included in the segment, potentially corrupting the speaker embedding. Hence, for the estimated VAD case, we use shorter segments of 7 secs that give the best performance on the dev set.

5.4 Role ID performance

Figure 5 shows AUC for role ID using the segments from the diarized clusters. Due to the differences in the segmentation across audio and text transcripts, the following are the valid cases: Case *Ro-1*: Diarization using oracle VAD and role ID using given text transcripts ([Di]:OrcVAD, [Ro]: Text). Case *Ro-2*: Diarization using estimated VAD and role ID using ASR-decoded text transcripts ([Di]:EstVAD, [Ro]: ASR-text). It can be seen that the trend in the AUC for both the conditions is similar to the

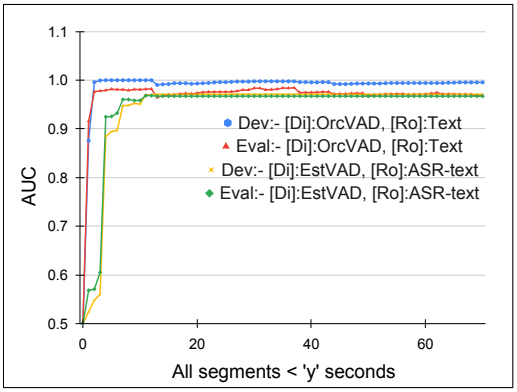
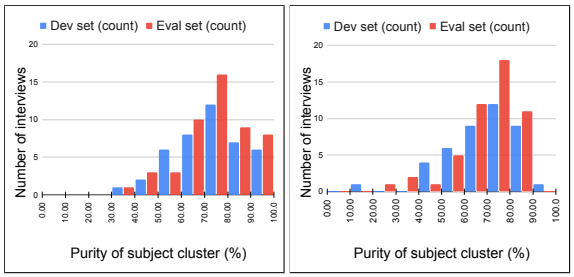


Figure 5: Role ID performance: AUC for role ID on different systems using all segments under y seconds length. All systems show improved performance when using more segments for role ID. [Di] denotes Diarization, and [Ro] denotes role ID.



(a) Oracle VAD diarized and Text-based RoleID

(b) Estimated VAD diarized and ASR text-based RoleID

Figure 6: Role ID performance: Purity of hypothesized subject cluster vs. number of interviews. Both the distributions are left-skewed, i.e., most of the hypothesized subject clusters have high average purity of 70% and above. Hence, the cognitive decisions taken in the next stage are majorly based on the subject's segments.

ground truth (as earlier discussed in Figure 2a). For the case *Ro-1*, we obtain the best AUC of 1.0 on dev set with cluster-level accuracy of 98% on the eval set using all segments in a cluster. For the case *Ro-2*, we obtain an AUC of 0.978 on the dev set with an accuracy of 93.5% on the eval set.

The average subject cluster purity for case *Ro-1* on dev set and eval set are 72.73% and 74.03% respectively. The average subject cluster purity for the case *Ro-2* for dev set and eval set are 68.41% and 69.8% respectively. The distributions of the subject cluster purities shown in Figure 6 are left-skewed for both the cases. It can be seen that most diarized interviews have highly pure subject clusters with the purity ranging between 70% and 100%. This ensures that the cognitive decision taken in the next stage is majorly based on the subject's segments.

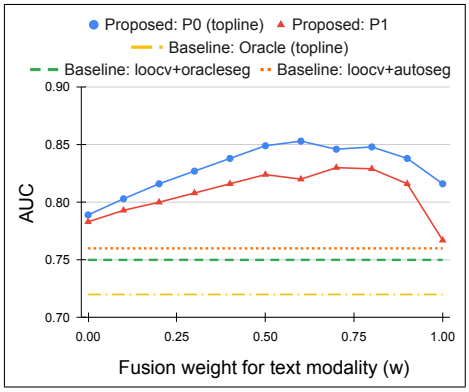


Figure 7: Cognitive ID performance: Text- and audio-based cognitive ID in Loocv setup. The sum of weights assigned to text and speech is 1. Fusing text and audio scores help to improve the performance.

Sys.	VAD	Role ID	Cognitive ID AUC (Eval / Combined)		
			Text	Audio	Text+Audio
P0	Topline	Text	0.875 / 0.816	0.811 / 0.789	0.875 / 0.853
P1	Oracle	Text	0.721 / 0.767	0.825 / 0.783	**0.825 / 0.830**
P2	Estim.	ASR	-	**0.781 / 0.763**	-

Table 4: Performance on cognitive ID. Hypothesized subject's segments are used for cognitive ID.

5.5 Cognitive ID performance

The best performing cognitive ID baseline systems are based on *B0 (oracle)*, *B1 (loocv+oracleseg)* and *B2 (loocv+autoseg)* with an AUC of 0.72, 0.75, and 0.76 as described in Section 5.1, respectively. As the baseline has a strong assumption that there are no silences in the whole conversation, we compare these results with our oracle VAD case. Table 4 shows AUC obtained using the proposed cognitive ID for different systems on eval and combined sets. It can be seen that our system performs better than the baseline system.

Further, we merge the text and audio scores at the cluster level using late score fusion. Figure 7 shows the AUC on the combined set under different score fusion weights. It can be seen that merging text and audio scores helps to improve the performance. This shows that the text and the audio modalities contain complementary information. As shown in Table 4, the best performance for the proposed system P1 obtained using text and audio score fusion is 0.825 and 0.830 on eval and combined set, respectively. This is close to performance on the best possible P0 system (topline). Finally, on fully audio-based proposed system P2, we obtain a competitive AUC of 0.781 and 0.763 on eval and combined sets, respectively. It is worth mentioning that the proposed system P2 does not make any

assumptions about silences in the recording. Also, note that cognitive ID decisions are less influenced by examiner's segments due to a better partitioning of subject and examiner by audio-based diarization module.

6 Conclusions

In this study, we found that the examiner's segment influence the cognitive ID performance in long exams. To mitigate the examiner's influence, we propose a three-stage cognitive identification system for long neuropsychological interviews. Our system achieves state-of-the-art performance on the FHS dataset. More importantly, the following are the main conclusions from our study:

- **Force alignment:** Proper segmentation is essential for accurate evaluation of performance.
- **Diarization:** Audio-based diarization focuses on speaker discriminative features and is more suitable compared to text-based role-specific LMs as multiple text sentences can be similar between examiner and subject.
- **Role ID and cognitive ID:** decision using *multiple segments* gives better performance as opposed to using some segments of particular lengths. Merging text and audio scores for cognitive ID improves performance.

Finally, our system automatically curates the conversation dataset at all stages, i.e., diarization (who spoke when?), ASR (speech to text), role ID (speaker identity), and cognitive ID (cognitive condition of a subject) which is useful for research.

7 Limitations

Since the proposed (and also the baseline) system have multiple stages (pipeline/cascaded system), there is a possibility of error propagation through these stages. However, using the best possible modules in the pipeline may reduce the possibility of error propagation.

8 Open directions for future research

- Our pipeline-based cascaded system has interpretable intermediate outputs. This is an important factor in healthcare applications. However, as mentioned in Section 7, it may suffer from error propagation when improper modules are used. End-to-End (E2E) deep neural network based systems on the other hand, are difficult to interpret but can reduce the possibility of error propagation. Considering the scarcity of clinical datasets, it is challenging to build such an E2E model that can handle multiple tasks and simultaneously mitigate the examiner's influence on cognitive ID decisions. We believe this can be a potential direction for the future research to have such an E2E system without losing the interpretability of intermediate outputs.
- We focus on mitigating the examiner's influence by eliminating examiner's sections from the final cognitive ID decision. It will also be an interesting study to analyse how much examiner can influence the decision and find the sections that are responsible for the influence.

Acknowledgements

We thank Tuka Alhanai for helping with the initial data setup. This work was supported by the Framingham Heart Study's National Heart, Lung, and Blood Institute contract N01-HC-25195; National Institutes of Health grants U19-AG068753, R01-AG016495, R01-AG008122, R01-AG033040. The authors would also like to thank Cody Karjadi and the staff and participants of the Framingham Heart Study.

References

Tuka Alhanai. 2019. Detecting cognitive impairment from spoken language. [Doctoral dissertation, Massachusetts Institute of Technology].

Tuka Alhanai, Rhoda Au, and James Glass. 2017. Spoken language biomarkers for detecting cognitive impairment. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 409–416.

Tuka Alhanai, Rhoda Au, and James Glass. 2018. Role-specific language models for processing recorded neuropsychological exams. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 746–752. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. To BERT or not to BERT: Comparing Speech and Language-Based Approaches for Alzheimer's Disease Detection. In *Proc. Interspeech*, pages 2167–2171.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

YF. Chuang, Y. An, M. Bilgel, DF. Wong, JC. Troncoso, RJ. O'Brien, JC. Breitner, L. Ferruci, SM. Resnick, and Thambisetty M. 2016. Midlife adiposity predicts earlier onset of alzheimer's dementia, neuropathology and presymptomatic cerebral amyloid accumulation. *Mol Psychiatry*, 21:910–915.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. In *Proc. Interspeech*, pages 1086–1090.

Nauman Dawalatabad, Mirco Ravanelli, François Grondin, Jenthe Thienpondt, Brecht Desplanques, and Hwidong Na. 2021. ECAPA-TDNN embeddings for speaker diarization. In *Proc. Interspeech*, pages 3560–3564.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Proc. Interspeech*, pages 3830–3834.

Mary Egan, Daniel Bérubé, Geneviève Racine, Carol Leonard, and Elizabeth Rochon. 2010. Methods to enhance verbal communication between individuals with alzheimer's disease and their formal and informal caregivers: A systematic review. *Int J Alzheimers Dis.*

Yuan Gong, Yu-An Chung, and James Glass. 2021. Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3292–3306.

Yuan Gong, Jin Yu, and James Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155.

R'mani Haulcy and James Glass. 2021. Classifying alzheimer's disease using audio and text-based representations of speech. *Frontiers in Psychology*, 11.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Jin Huang and C.X. Ling. 2005. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310.

Lenore Kurlowicz and Meredith Wallace. 1999. The mini-mental state examination (MMSE). *Journal of gerontological nursing*, 25(5).

Yann LeCun and Yoshua Bengio. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge. In *Proc. Interspeech*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using kaldi. In *Proc. Interspeech*, pages 498–502.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. In *Proc. Interspeech*, pages 2616–2620.

Neville Ryant. 2018. Dscore: Rich transcription time marked evaluation tool. `https://github.com/nryant/dscore`. [Online; accessed 21-Oct-2022].

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Monisankha Pal, Manoj Kumar, Raghuveer Peri, Tae Jin Park, So Hyun Kim, Catherine Lord, Somer Bishop, and Shrikanth Narayanan. 2019. Speaker diarization using latent space clustering in generative adversarial network. ArXiv:1910.11398.

Monisankha Pal, Manoj Kumar, Raghuveer Peri, Tae Jin Park, So Hyun Kim, Catherine Lord, Somer L. Bishop, and Shrikanth S. Narayanan. 2021. Meta-learning with latent space clustering in generative adversarial network for speaker diarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1204–1219.

Raghavendra Pappagari, Jaejin Cho, Laureano Moro-Velázquez, and Najim Dehak. 2020. Using State of the Art Speaker Recognition and Natural Language Processing Technologies to Detect Alzheimer's Disease and Assess its Severity. In *Proc. Interspeech*, pages 2177–2181.

Tae Jin Park, Kyu J. Han, Manoj Kumar, and Shrikanth Narayanan. 2020. Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap. *IEEE Signal Processing Letters*, 27:381–385.

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech Language*, 72:101317.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

P.A. Pérez-Toro, S.P. Bayerl, T. Arias-Vergara, J.C. Vásquez-Correa, P. Klumpp, M. Schuster, Elmar Nöth, J.R. Orozco-Arroyave, and K. Riedhammer. 2021. Influence of the Interviewer on the Automatic Assessment of Alzheimer's Disease in the Context of the ADReSSo Challenge. In *Proc. Interspeech*, pages 3785–3789.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. ArXiv:2106.04624.

Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, and Sanjeev Khudanpur. 2018. Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge. In *Proc. Interspeech*, 2808–2812.

Christine A Szekely, Jennifer E Thorne, Peter P Zandi, Mats Ek, Erick Messias, John C S Breitner, and Steven N Goodman. 2004. Nonsteroidal anti-inflammatory drugs for the prevention of alzheimer's disease: a systematic review. *Neuroepidemiology*, 23(4):159–169.

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Zheng-Hua Tan, Achintya kr. Sarkar, and Najim Dehak. 2020. rvad: An unsupervised segment-based robust voice activity detection method. *Computer Speech Language*, 59:1–21.

George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE.

Ulrike von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and Computing*, 17(4).

G Weinstein, AS Beiser, SH Choi, SR Preis, TC Chen, D Vorgas, Au Rhoda, A Pikula, Wolf PA, AL DeStefano, Vasan RS, and S. Seshadri. 2014. Serum brain-derived neurotrophic factor and the risk for dementia: the framingham heart study. *JAMA Neurol*, 71(1).

Xavier Anguera. 2008. Diarization Error Rate. `http://www.xavieranguera.com/phdthesis/node108.html`. [Online; accessed 21-Oct-2022].

A General challenges with clinical interview data

Information such as "who is speaking when", "what is spoken?", and "cognitive condition" are of interest to the clinical research community. Even though recording an audio of the exam is an easy task, automatically obtaining such information of interest, on the other hand, is not trivial (especially for clinical data). Automatically obtaining such information from clinical data involves several challenges that are different from other conversational interviews (Sell et al., 2018; Alhanai et al., 2018; Pal et al., 2021). For example, neuropsychological interview exams have significant variations in terms of speaker turn-taking, have significant background noise, and some times improper text transcriptions. Further, the amount of clinical data (in general) is an order of magnitude smaller compared to other datasets. All these conditions make it highly challenging to process such real-world conversational clinical data.

A.1 FHS dataset, recording conditions and annotations

We use the gold standard Framingham Heart Study (FHS) dataset (Alhanai et al., 2018) in this paper. It includes a total of 92 real-world neuropsychological interview exam recordings in English with 100 hours of total duration. There are 20 examiners, and all exams have a unique subject. In total, there are 21 cognitively impaired subjects, accounting for 22.8% of the total dataset size. The average length of each recording is around 65 mins. The mean age of subjects under evaluation is 68 years, and the gender ratio is 49:51 (female:male). Exams were recorded at different sampling rates of 8kHz, 16kHz, and 44kHz. We re-sampled all recordings to 16kHz for consistency. The exam recordings have various background noise, e.g., some recordings are recorded in open settings, and sound from background speakers also gets recorded. In addition, the quality and placement of microphones used are inconsistent across the exams. All these variations of real clinical interviews are challenging compared to other general conversation datasets. These variations make our task more challenging.

All the recordings have been transcribed orthographically by human annotators. Annotators were also asked to mark non-speech filled pauses (e.g., <um>, <uh>).

B Text and speech forced alignment

Figure 8 shows an illustration of the text and speech forced alignment. The *orig* represents speaker boundaries in the original annotation, *FAligned* are the word boundaries obtained using MFA (McAuliffe et al., 2017). The *impr* are the final output boundaries. Each *orig* segment is processed separately to obtain the end timestamp of that segment as described in Section 2.

C Spectral clustering for diarization

Spectral clustering is a graph-based approach to clustering (von Luxburg, 2007) and widely used backend clustering approach to cluster speaker embeddings (Dawalatabad et al., 2021; Park et al., 2022). The affinity matrix representing the similarities between the pairs of the embeddings is calculated. Smaller values of similarities are pruned using pruning threshold. The Laplacian matrix is estimated using the affinity matrix as described in (von Luxburg, 2007; Park et al., 2020). We then perform Eigen decomposition of the Laplacian, and top-k eigenvectors are estimated. The rows of the eigenvectors become the spectral embeddings. These spectral embeddings are clustered using the standard k-means algorithm.

D Additional evaluation and model configuration details

All the models trained in this paper are from popular publicly available toolkits. We share the complete configuration files for all the models used in this paper with all the hyperparameters used in our experiments. We also share the code for the models that differ from the publicly available toolkits. Since the FHS dataset is in the English language, all the models are trained using English language datasets. Table 5 shows the statistics for all the models used for the experiments in this work.

D.1 Text and speech forced alignment

We force align the word sequences with the input audio (McAuliffe et al., 2017). We use the pretrained Librispeech model and adapt it with the FHS dataset. We perform adaptation with 10 hours of FHS data from dev set. The alignments are generated as described in Section 2. Note that MFA is only used to get the speech segment endpoints. No speaker turn information is used in the diarization system.

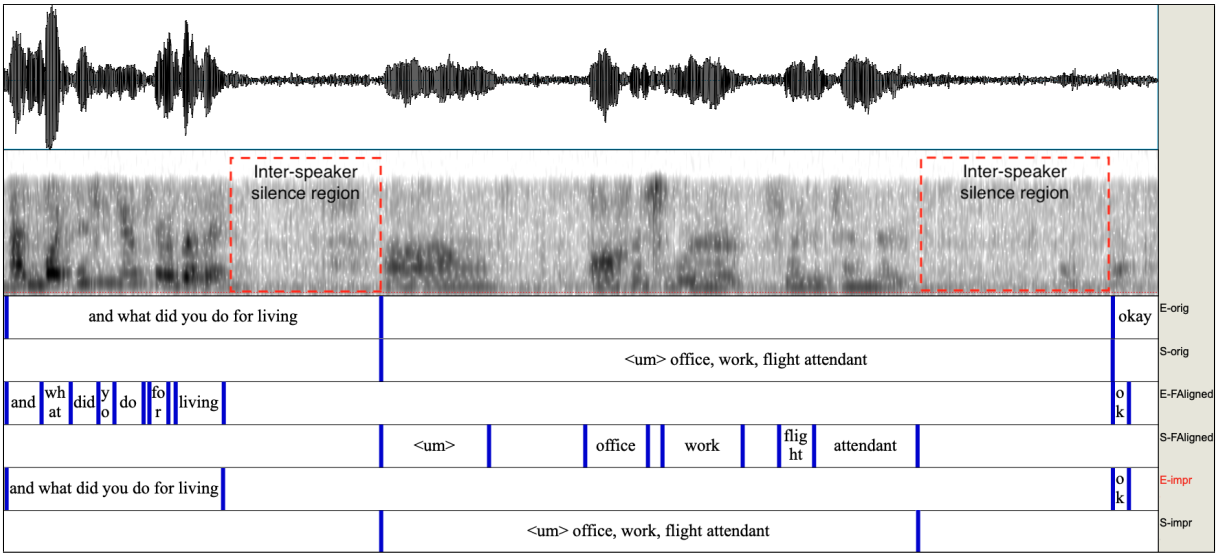


Figure 8: Forced alignment of text and speech: The first and second tiers shows audio and spectrogram respectively. The prefixes E and S are used to represent examiner and subject respectively. The suffix *orig* represents speaker boundaries in the original annotation. The *FAligned* are the word boundaries obtained after forced aligning text and speech. The *impr* denotes the improved speaker boundaries with speaker end timestamps.

Categories	KenLM (Training)	ECAPA-TDNN	Wav2Vec 2.0 (Finetuning)	EfficientNet (Training)
GPU / CPU Name	INTEL CPU 2.7 GHz	NVIDIA V100	NVIDIA RTX A5000	NVIDIA GTX 1080
GPU / CPU Memory	3 GB	32 GB	24 GB	8 GB
Number of such GPUs/CPUs	1	1	4	2
Run time for 1 Epoch	-	23 hours	20mins	5 min x 92 recordings
Tot. num of Epochs	-	12	20	5
Avg. runtime for Training	1 hour	11 days (pretrained model used)	8 hours	40 hours
Avg. runtime for Inference	5 mins	1 hour	30 mins	1 min x 92 recordings
Num of model parameters	100,000	21 Million	300 Million	4 Million

Table 5: Experiment and resource details for different models used in this work. All the values are approximated to the closest integer.

D.2 Diarization

For diarization with estimated VAD, we tune the rVAD system on dev set such that we do not miss any speech. This is important as the role ID and the cognitive ID modules are dependent on the words spoken by the subject. The hyperparameter of $ftThres$ and $vadThres$ are set to 0.91 and 0.001 respectively (Tan et al., 2020). With this configuration, we obtain the MS and FA are 0.3% and 21.3%, respectively.

We use the pre-trained ECAPA-TDNN model (Desplanques et al., 2020) for our experiments (Ravanelli et al., 2021). The model was originally trained using Voxceleb1 and Voxceleb2 (Nagrani et al., 2017; Chung et al., 2018) datasets. The best pruning threshold value ($pval$) for spectral clustering obtained on the dev set is 0.02 and 0.5 for oracle VAD and estimated VAD, respectively.

We use a standard forgiveness collar of 250ms (Alhanai et al., 2018) for diarization evaluation. It does not penalize the error within the collar from the speaker's boundaries. It is helpful to ignore these small errors as they might have originated due to improper human annotations. We use the NIST evaluation tool to calculate the SER and cluster purity (Neville Ryant, 2018).

D.3 Role ID

The ASR decoded text used for training role-specific LMs is obtained from a popular wav2vec 2.0 model (Baevski et al., 2020). The pre-trained model from Fairseq (Ott et al., 2019) is finetuned on the FHS dev set to obtain the ASR decoded text. All the AUC calculations are performed using sklearn toolkit (Pedregosa et al., 2011; Buitinck et al., 2013).

D.4 Cognitive ID

Since there is a class imbalance between dementia and healthy, models for cognitive ID are trained using class balancing strategies. For both LM-based and EfficientNet cognitive ID, we over-sample the sentences from the dementia class. For cognitive ID evaluation, we found that considering the top N highest scoring segments from the hypothesized subject cluster for audio modality gives better performance. Hence we consider top N scoring segments for audio modality, while for text modality, we use all the segments in the hypothesized subject cluster.

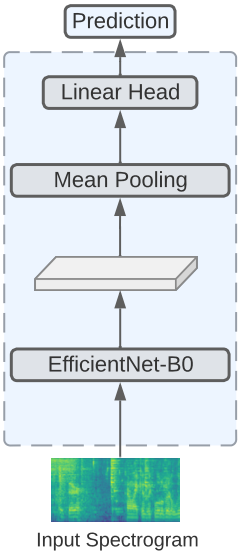


Figure 9: The EfficientNet model architecture for cognitive ID used in this paper.

We slightly modified the EfficientNet model in (Gong et al., 2022) for cognitive ID (illustrated in Figure 9). Each speech waveform is first converted to a sequence of 128-dimensional log Mel filterbank (fbank) features computed with a 25ms Hanning window every 10ms. The $t \times 128$ fbank feature vector is input to an EfficientNet-B0 model (Tan and Le, 2019). The EfficientNet-B0 model effectively downsamples the time and frequency dimensions by a factor of 32, and the feature dimension d is 1280. Thus, the penultimate output of the model is a $\lceil t/32 \rceil \times 4 \times 1280$ tensor. We apply mean pooling over the four frequency dimensions and all time dimensions to produce a 1280-dimensional segment-level representation that is fed to a binary linear classifier for cognitive ID. We train the EfficientNet model with a batch size of 48, an initial learning rate of 1e-3 for up to 5 epochs.