

Article

Not peer-reviewed version

Affordable Audio Hardware and Artificial Intelligence Can Transform the Dementia Care Pipeline

[Ilyas Potamitis](#) *

Posted Date: 10 September 2025

doi: 10.20944/preprints202509.0861.v1

Keywords: dementia; elderly care; LLM; IoT; ASR; ESR; MEMS



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Affordable Audio Hardware and Artificial Intelligence Can Transform the Dementia Care Pipeline

Ilyas Potamitis

Hellenic Mediterranean University, Dept. of Music Technology and Acoustics, Rethymno, Greece;
potamitis@hmu.gr

Abstract

Population ageing is increasing dementia care demand. We present an audio-driven monitoring pipeline that operates either on mobile phones, microcontroller nodes, and smart television sets. The system combines audio signal processing with AI tools for structured interpretation. Preprocessing includes voice activity detection, speaker diarization, automatic speech recognition for dialogues, and speech-emotion recognition. An audio classifier detects home-care-relevant events (cough, cane taps, thuds, knocks, speech). A large language model integrates transcripts, acoustic features, and a consented household knowledge base to produce a daily caregiver report covering orientation/disorientation (person, place, time), delusion themes, agitation events, health proxies, and safety flags (e.g., exit seeking, possible fall). The pipeline targets continuous monitoring in homes and facilities and it is an adjunct to caregiving, not a diagnostic device. Evaluation focuses on human-in-the-loop review, various audio/speech modalities, and the ability of AI to integrate information and reason. Intended users are low-income households in remote settings where in-person caregiving cannot be secured, enabling remote monitoring support for older adults with dementia.

Keywords: dementia; elderly care; LLM; IoT; ASR; ESR; MEMS

Introduction

Population ageing is accelerating worldwide, shifting the balance between those needing care and the working-age population available to provide it. Globally, dementia imposes substantial health-system and societal costs—estimated at US\$1.3 trillion in 2019 and projected to rise sharply in coming decades—while informal carers already deliver roughly half of total care hours [1–4]. In Europe, older adults (≥65 years) accounted for 21.6% of the EU population on 1 January 2024, and projections indicate a sustained increase in the old-age dependency ratio through mid-century [2]. United Nations projections similarly forecast a rapid expansion of the ≥65 population over the next three decades [3]. Against this backdrop, enabling safe, scalable home-based care for people living with dementia is a pressing priority, both to preserve independence and to mitigate escalating institutional and family costs [1,4]. In many cases, in-home caregiving is provided by non-specialist workers with limited formal training; services can be expensive, and reliable high-quality support is not readily available on demand. In this work, we explore the integration of audio-based signal processing with new possibilities offered by artificial intelligence (AI) tools within the care pipeline for elderly patients with dementia. Our aim is to leverage recent advances in AI, and large language models (LLMs) in particular, to deliver an affordable service designed primarily for people of low-income and/or remote locations where full-time human care is not practically feasible. Our approach is also applicable to the important and pressing case of childless elderly people who live alone and develop mobility and mental health conditions.

Ambient audio sensing is an attractive modality for home care: microphones are inexpensive, unobtrusive, and easily installed; they can capture both speech and non-speech events (e.g., coughs,

thuds, alarms, cane taps, expressions of pain or distress) without the visual privacy trade-offs of cameras and without the adherence burden of wearables. Prior research has demonstrated that speech carries information relevant to cognitive status. Early work using automatic speech analysis differentiated healthy controls, mild cognitive impairment (MCI), and Alzheimer's disease (AD), establishing the feasibility of acoustic biomarkers [5]. Subsequent reviews and empirical studies have evaluated acoustic (paralinguistic) features—prosody, timing, pauses—alone and alongside linguistic variables for AD screening and monitoring, including remote collection paradigms [6–9]. Recent studies have further explored non-semantic, acoustic-only features and the feasibility and test-retest reliability of multi-day, remote assessments of speech acoustics, including associations with amyloid status and deep-learning approaches to voice recordings [10–15]. Collectively, this body of work supports speech-based digital biomarkers as a noninvasive window into cognitive health, but translation to day-to-day home-care workflows remains limited [8].

Beyond speech, audio has also been investigated for characterizing behavioral and psychological symptoms of dementia and care environments. Persistent or inappropriate vocalizations are a common and burdensome symptom in advanced dementia; integrative reviews synthesize their phenomenology and implications for care [16]. Soundscape interventions in nursing homes show that targeted monitoring and staff feedback can improve acoustic environments and staff evaluations, suggesting that environmental audio is a modifiable determinant of wellbeing [17]. Concurrently, a growing literature examines audio-based detection of safety-critical events such as falls [18,19], spanning classical machine-learning methods and transformer-based approaches and complemented by reviews of fall-detection methods and wearable sensing for older adults [20,21]; related work has demonstrated real-time acoustic detection of critical incidents on edge devices, underscoring feasibility for low-latency deployment [22]. However, most prior studies focus on single tasks (e.g., diagnostic screening or environmental monitoring), use longer or highly structured recordings, or are conducted in clinical or institutional settings rather than ordinary homes [23–28]. Integration of heterogeneous audio evidence into actionable, caregiver-facing summaries is rarely addressed.

Smart-home and telecare research for older adults has advanced through successive projects that illustrate different sensing and inference paradigms. An adaptive framework for activity recognition in domestic environments incorporated user feedback to refine models, thereby embedding personalization into resident-centered automation [29,30]. Practical issues of scaling were examined in multi-country deployments of assistive equipment for people with dementia, where installation procedures and technical support were identified as critical for sustained operation [31]. Inference methods progressed with plan-recognition models capable of interpreting incomplete observations to detect disorientation and atypical behaviors in Alzheimer's patients [32]. Multimodal capture of daily living activities was enabled through wearable audio–video devices, with indexing strategies to allow efficient clinical review of extended recordings [33]. For nocturnal monitoring, unobtrusive load-cell systems placed under beds were developed to quantify movement and posture changes, supporting analysis of sleep patterns and risk states [34]. Activity recognition in naturalistic home settings using simple state-change sensors demonstrated that ubiquitous instrumentation could yield meaningful behavior classification without requiring cameras [35]. Within clinical contexts, models were proposed for quantifying patient activity in hospital suites, producing indicators such as mobility levels and displacement distributions that could be tracked continuously [36]. Earlier, lifestyle-monitoring systems established the feasibility of using ambient data streams to identify deviations from typical behavior and support independent living [37]. Finally, integrated e-healthcare prototypes in experimental “techno houses” combined physiological and environmental sensors to enable continuous monitoring across multiple aspects of daily life, including personal hygiene and rest [38].

This work targets that translational gap by proposing and evaluating a home-based pipeline that couples short-duration audio sensing (15 s snippets) with an Audio Spectrogram Transformer (AST) [39] trained on the Audioset database [34] for non-speech event detection and a LLM (ChatGPT-5) constrained to emit auditable, structured reports. Speech and non-speech events salient for home

care—coughs (including chronic cough burden), cane-tap sequences (as a mobility proxy via cadence variability), thuds, appliance beeps, and alarms—are time-stamped by AST and exported as a JSON record. The LLM receives as text input the context of the application, dialogue transcripts, acoustic features, and event metadata together with a small, consented household knowledge base, and returns reports covering orientation/disorientation (person/place/time), delusion themes and their persistence (e.g., fixed beliefs about “another home”), agitation likelihood, comprehension difficulty, instruction follow-through, and safety flags (e.g., possible fall). Unlike diagnostic-only approaches, the outputs are designed to guide daily care (e.g., de-escalation prompts, check-ins) and to populate longitudinal dashboards for trend analysis.

Our contributions are threefold. First, we integrate speech and non-speech audio within a home pipeline, moving beyond single-task speech-only screening to multi-task monitoring that reflects real-world caregiving priorities (safety, agitation, communication efficacy). We use recently developed AI tools that were not available ten years ago (i.e. transformers, large databases and LLMs). Second, we develop a real-time, affordable, hardware solution and we open source its code. Third, we use an LLM to integrate explicitly separate acoustic observations to reach a higher level of interpretation that cannot be achieved using single audio transcription—for example, detecting delusional themes. In aggregate, the combination of audio, transformer classifiers, and LLM receiving input from real patients reframes ambient home audio from a raw signal into validated, longitudinal indicators of safety and wellbeing, complementing—not replacing—clinical assessment while addressing pressing needs in dementia care. A depiction of our approach can be seen in Figure 1.

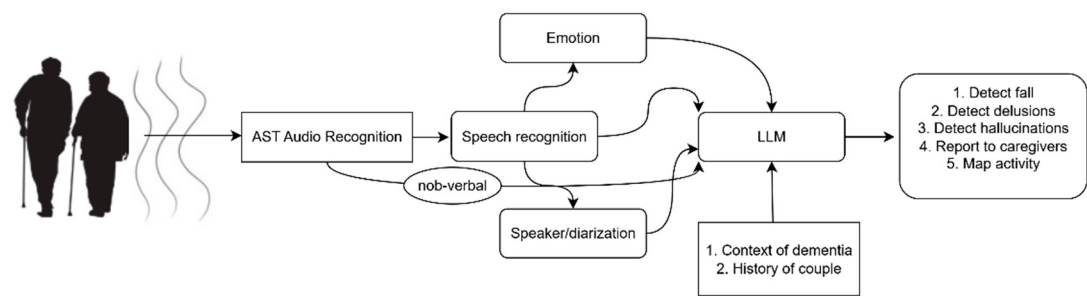


Figure 1. Audio based care of elders with dementia based on the cooperation of audio and AI. Audio events are classified as verbal/non-verbal by an AST transformer. The verbal queues are directed at an emotion recognizer, a speaker diarization module and a speech recognizer. All audio events transcribed to text are fed to an LLM that detects events of interest (falls, delusions, pain) by integrating multiple information and reasons so that it prepares reports for the caregivers.

Materials & Methods

The study participants are a couple living in their own home—a woman aged 87 and a man aged 90 (as of 2025)—both formally diagnosed with stage-2 dementia. The main symptoms are: Increasing difficulty recalling recent events; forgetting names of familiar people, word-finding problems, short sentences, confusion about time and place, repetitive questioning. The female subject has developed delusions and increased risk for falls. Both subjects have additional mobility problems and use cane and other mobility aids.

The subjects are attended by two caregivers in two 8-hour shifts per day and occasionally receive additional care from practitioners. The establishment has in-house cameras with audio in all rooms attended by professional security personnel during nights. Due to low mobility, the couple moves only to one floor and only to the bedroom, the bathroom and the living room.

This study is a proof-of-concept demonstrating the feasibility of an audio processed by AI pipeline for home dementia care. Our goal is methodological and systems-oriented—to show that ambient audio can be captured on very low-cost hardware, transformed into text labels

corresponding to the recognized audio events out of 527 audio classes, and summarized by an LLM (in this work ChatGPT 5 in thinking mode) that receives the labels and the context of the application into auditable, caregiver-facing outputs. We do not claim population-level diagnostic performance; instead, we establish engineering viability, safety guardrails, and a structured scheme for longitudinal monitoring.

Although the cohort is small, the dataset is longitudinal and dense (thousands of 10–15 s snippets across weeks), enabling robust within-subject demonstrations of trend tracking (e.g., cough burden, exit-seeking themes, cane-tap variability). For personalization-centric care, repeated measures from the same individuals are more informative than sparse cross-sectional samples.

We acknowledge that two participants do not support population-level inference. However, we believe that reported evidence in this work justifies the position that LLM and AI-bots can provide significant value to this problem by first monitoring and in the near future engaging the task with actions. Accordingly, we: (i) report per-subject results and descriptive statistics rather than pooled significance tests; (ii) release code and de-identified audio to enable replication; and (iii) outline a prospective, multi-site study with power considerations for clinical endpoints (e.g., sensitivity to agitation/exit-seeking, caregiver workload impact). This sequencing—proof-of-concept → pilot → powered trial—is standard for translational systems entering safety-critical care.

Hardware

The approach presented is based on the ESP32 and MEMS type microphone that are very low cost. We implemented a distributed architecture that supports cloud, on-premises, and edge deployments. A network of ESP32-S3-DevKitC-1-N8R8 nodes with SPH0645LM4H digital microphones transmits mp3 compressed 10–15 s audio snippets over Wi-Fi to a secure remote server for classification.

The SPH0645LM4H provides a low-power, bottom-port I²S output, eliminating the need for an external codec. Its flat 0.1–8 kHz response aligns with our 16 kHz sampling rate for AST-based analysis. The ESP32-S3, a dual-core Xtensa LX7 MCU with Wi-Fi/BLE, vector instructions, ample GPIO, and I²S, was selected to handle audio capture and MQTT transport. The total cost of the device is at the order of 20 Euros (as per 14/8/2025). However, the same service can be provided by a smartphone that streams audio snippets captured by a voice activity detector (VAD) to Wi-Fi. A more detailed description of the hardware can be found in [22].

Embedded Software in ESP32

Using Espressif IoT Development Framework version 5.3 (ESP-IDF 5.3), the embedded nodes in C language capture audio from the microphone via the Inter-IC Sound (I²S) digital audio interface in non-overlapping 1024-sample windows at 16 kHz. Then, they perform basic level checks using the Root Mean Square (RMS) value, encode the audio into MP3 format with an embedded MP3 Encoder (LAME) library, and publish the compressed data using the Message Queuing Telemetry Transport (MQTT) protocol secured with Transport Layer Security (TLS). Processing is divided between the two cores of the ESP32-S3 microcontroller (Core 0 handles capture and processing, while Core 1 handles encoding and transmission), with a ring buffer used for inter-task communication. Configuration is stored in Non-Volatile Storage (NVS) and supports remote updates via MQTT. Pseudo-Static Random Access Memory (PSRAM) is used for efficient buffering. Networking is provided over Wi-Fi by default.

MQTT was chosen for its reliability under constrained bandwidth and variable link quality. Streams are handled by a multithreaded MQTT client with Quality of Service (QoS) tuning, buffering, error detection, and fast reconnection protocols. Latency is reduced through optimized packetization, I²S Direct Memory Access (DMA) buffering, and adaptive send intervals; a User Datagram Protocol (UDP) fallback is enabled if MQTT delays increase. Local brokers operating over 2.4 GHz Wi-Fi minimize Wide Area Network (WAN) hops, while server-side asynchronous processing and

Network Time Protocol (NTP) synchronization maintain throughput and temporal alignment across nodes.

A Python 3.11 wrapper (see Appendix) manages audio acquisition, preprocessing, and model execution. At the server, MP3 clips are decoded using Fast Forward Moving Picture Experts Group (ffmpeg), converted into spectrograms via the Short-Time Fourier Transform (STFT), and rendered as time–frequency images for the Audio Spectrogram Transformer (AST). The AST, which follows a Vision Transformer (ViT) architecture, operates directly on spectrogram patches for audio event classification. Spectrograms for spoken snippets and environmental events (e.g., discussions, cough, cane taps) are processed to produce labeled, timestamped outputs suitable for downstream integration.

The Elders with Dementia Dataset

The ESP devices work 24/7 and are activated by voice to record at 16kHz sampling rate, the subsequent 15 seconds. No audio recordings are kept on the device. Depending on the activity inside home, several hundred recordings can be produced per day. The recording sessions started on 4/8/2025 and continue up to now reaching several thousands of recordings. Recordings contain mainly speech, non-verbal human sounds, cane hits and impact sound of various objects (e.g., drawers, closets, doors, switches). The database they produced is the subject of this study.

The Audio Event Recognizer

The Audio Spectrogram Transformer (AST) [39] formulates audio tagging as patch-based transformer classification on log-Mel spectrograms and is commonly pre-trained or fine-tuned on AudioSet. Concretely, a waveform is converted to 128-dimensional log-Mel filterbanks using a 25 ms window with 10 ms hop; the 2-D time–frequency map is then split into overlapping 16×16 patches that are linearly projected to tokens, augmented with positional embeddings, and processed by a ViT-style encoder (12 layers, 12 heads, 768-dim embeddings). The model is initialized from ImageNet ViT weights via patch and positional-embedding adaptation and optimized with binary cross-entropy using standard audio augmentations (mixup, SpecAugment) and weight-averaged checkpoints. Evaluated on AudioSet—the 2.1 M-clip, 5.8 k h, 527-class collection of human-labeled 10-second YouTube excerpts—AST reports mean average precision (mAP) 0.4590 for a single weight-averaged model and up to 0.4850 with model ensembling on the full training split, indicating that transformer attention over spectrogram patches is competitive with or superior to CNN/attention hybrids for weakly labeled sound event recognition. In this configuration, AST leverages AudioSet’s breadth to learn general sound representations that transfer to downstream tagging tasks without architectural changes.

AudioSet [40] is a large-scale, ontology-driven dataset of human-labeled sound events released by Google, comprising over 2 million 10-second audio clips drawn from YouTube videos. Its ontology includes 527 distinct classes spanning speech, environmental sounds, animal vocalizations, and human activities, providing a broad coverage for training and evaluating audio event recognition models. The AudioSet ontology includes the sounds that daily activity care produces, namely, impact, collisions, and knocks labels that are falls / accidents proxies. Specifically, the Knock, Tap, Bang, Thump/thud, Slam, Smash, crash, Clatter, Shatter (glass) are included in the ontology. Regarding sounds of human vocal distress, pain, and agitation that can be associated with elderly activity and accidents, the following labels are included in the ontology and are semantically related: Screaming, Yell, Shout, Wail/moan, Whimper, Groan, Gasp, Sigh. We are also interested in monitoring chronic conditions related to respiratory and health-conditions and the following are included in the available set of labels: Cough, Sneeze, Breathing, Snoring, Throat clearing and Wheeze appears under respiratory sounds in the ontology as well. Regarding mobility and movement (activity / restlessness cues), the labels: Footsteps, bouncing, wood, Stomp, stamp, Surface contact (e.g., scrape, scratch), Creak (e.g., floorboards/chairs), Door (open/close). Safety / alerting signals (hazard or attention) are included in the ontology. Smoke detector, smoke alarm / Fire alarm, Alarm

clock, Siren, Buzzer, Beep, bleep, Doorbell, Telephone ringing. Conversation & caregiver interaction (dialog state cues) are handled by the Speech, Conversation, Whispering, and loudness-related: Shout / Yell categories.

The wide representation of audio categories in Audioset ontology, practically cover all cases appearing in the context of our work.

Ordinary Automatic Speech Recognition Models Do Not Currently Work Properly for People with Dementia

Whisper large-v3 is an open, transformer-based encoder–decoder ASR model that delivers strong accuracy across many languages with competitive inference speed. Automatic speech recognition (ASR) with Whisper 3-large (see Appendix) underperformed on conversational Greek from older adults with dementia because multiple, interacting factors degrade the signal and violate the model’s assumptions. We subsequently analyze the reasons for this failure. First, speaker physiology changes with age—presbyphonia, breathy or tremulous phonation, reduced articulatory precision from dentures or dry mouth, and frequent comorbid dysarthria—alter formant structure and sibilants and reduce the effective signal-to-noise ratio, making phonetic cues less distinct. Second, cognitive–linguistic patterns characteristic of dementia (see also Figure 2)—long hesitations, false starts, repetitions, elliptical utterances, and frequent repair sequences—disrupt the fluent, clause-level structures that end-to-end ASR models implicitly expect from their training data, increasing deletions and punctuation errors. Third, the home acoustic environment adds nonstationary interference: televisions and radios, and hard-surface reverberation all produce cross-talk and spectral interference that cannot be discriminated by front-end feature extraction. Fourth, Greek-specific variability—regional accents, intonation patterns, and rich inflectional morphology with clitics and enclitics—is underrepresented in common web-scale corpora, so systems tuned on broadcast or YouTube speech from younger speakers generalize poorly to elderly, informal Greek dialogue. Fifth, pipeline constraints compound errors: tight voice-activity detection and diarization cuts fragment utterances; 10–15 s windows limit language-model context for disfluency recovery. In combination, these factors yield low confidence scores, insertion/deletion spikes, and semantic drift in transcripts, which in turn mislead downstream LLM components tasked with behavioral inference. Model adaptation to include the speech of elders is feasible but nontrivial: since systems like Whisper are trained on paired audio–text annotations, effective domain adaptation would require a sizable, consented corpus of elderly Greek conversational speech with accurate transcripts spanning dialects, acoustic conditions, and symptom severities—resources that are rarely available and costly to curate (see [41] for an adaptation of ASR models to elders using a 12 hours adaptation corpus).

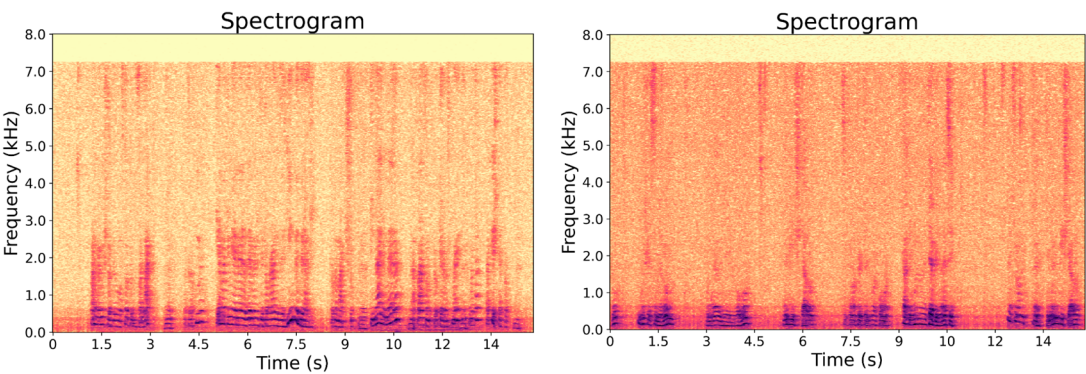


Figure 2. (Left) An example of female elder speech (Event_id: 172751). (Right) Male elder (Event_id: 172738). Both speakers were 1-1.5m away from the microphone. Notice the lack of frequencies above 2 kHz and the pauses.

Voice Activity Detection

A VAD marks the part of the recording that is related to speech. There is a vast number of algorithms and implementations available. We have chosen a new approach from the TEN-VAD which has demonstrated better results compared with other state-of-the-art algorithms (see Appendix). This algorithm is a small hybrid digital signal processing (DSP) followed by a recurrent neural network VAD. It is lightweight, supports streaming and uses classic signal processing features and a small LSTM-style ONNX model for sequence modeling. It computes a 16 kHz STFT with a Hann window, applies a 40-band Mel filterbank, logs/normalizes features, and appends a pitch feature. These per-frame features are stacked over a short temporal context. The model emits a frame-level speech probability and a binary VAD flag after thresholding.

Speech Diarization and Speaker Recognition

Speaker diarization splits a clip into speaker-homogeneous time segments and can flag overlaps. It does not know the identities of people talking. Speaker recognition is a different task and tags “who is this speaker.” It matches a segment to an enrolled voiceprint (verification/identification) assuming the segment is single speaker. Determining *who spoke when*—is essential in monitoring elderly people with dementia because it attributes each utterance to the correct person (e.g., patient, caregiver, visitor). There are many approaches and implementations dealing with this task but in our case few-shot learning seems the appropriate solution. Few-shot learning relies on few samples to build a model for each speaker (~20 recordings of 15 seconds each), and the in-home application is expected to deal with a small number of speakers. In our setting, speaker tagging is convenient since the AST selects out speech that is subsequently tagged by a label of few speakers (six in our case: two patients, two caregivers, two relatives) or any of their combinations. By segmenting audio into speaker-specific streams before ASR, diarization reduces attribution errors (such as mistaking a caregiver prompt for a patient belief) and yields cleaner, per-speaker transcripts for the LLM. This improves LLM’s reasoning about orientation, agitation, and instruction follow-through, enables role-aware prompts and alerts, and supports longitudinal tracking of patient speech while preserving context about overlapping talk and background voices. Speech diarization has been quite successful in our experiments because, as usually the case is, there were a limited number of people in a house of elders. This allows modeling of the voices based on few examples and reliable attribution of speech queues. In this work we take the pyannote approach that implements diarization as a three-stage neural pipeline—local speaker segmentation, speaker embedding, and global clustering—that operates directly on mono 16 kHz audio and returns a time-stamped “who-spoke-when” annotation (see Appendix). In version 2.1, the default pipeline applies a neural speaker-segmentation model on short sliding windows (≈ 5 s with 0.5 s hop), producing frame-level (≈ 16 ms) posteriors for the activity of up to $K_{\max}$ speakers; a related “powerset” segmentation variant explicitly models single- and two/three-speaker overlap classes to handle simultaneous talkers. Local speaker traces are then converted to speaker embeddings and aggregated globally via hierarchical agglomerative clustering (AHC) to obtain a consistent speaker inventory for the whole recording; the number of speakers can be estimated automatically or constrained by the user when known. The released embedding model is an x-vector TDNN architecture augmented with SincNet learnable filterbanks (trainable sinc convolutions replacing fixed Mel filters), yielding compact speaker representations suited to clustering. The toolkit also exposes overlapped-speech detection that can be combined with the pipeline to better account for concurrent speech. Collectively, this design frames diarization as end-to-end neural segmentation on overlapping windows, representation learning for speakers, and global AHC on embeddings—rather than K-means on spectra or hand-engineered MFCCs—to robustly infer speaker turns.

Emotion Recognition

In elderly people with dementia, depending on the age, health problems and dementia stage, the emotional expression often differs significantly from that of younger populations and may be subtle, context-dependent, and shaped by the progression of cognitive decline. Neutral or baseline

states, which can blend into apathy, are common and characterized by long silences, flat prosody, and reduced motivation or expressivity. Moments of happiness or contentment do occur, often triggered by familiar people, music, or reminiscence, and may be marked by brief laughter or a brighter tone. Sadness and depressive states manifest as softer speech, a slower pace, sighing, or tearfulness, and are often accompanied by social withdrawal. Anger, irritability, and agitation are also possible, presenting as raised or shaky pitch, louder vocal bursts, or interruptions, particularly during care tasks, refusals, or episodes of confusion. Fear and anxiety are typically expressed through a tense or pressed voice, vocal tremor, and repetitive questioning, and are especially prevalent during sundowning. Apathy, a low-arousal negative state distinct from simple neutrality, can be pervasive, with minimal speech initiation and muted effect.

Emotion recognition algorithms extract low-level acoustic descriptors (pitch, loudness, speech rate and rhythm, etc.) from the audio and then use statistical models to classify into emotions (e.g. Neutral, Happy, Sad, Angry, Afraid). Since they rely on acoustic-prosodic features (i.e. paralinguistic cues) rather than speech content, they can be largely language-independent—meaning they *can* work on any language, but accuracy may be slightly lower if the prosody differs from the languages it was trained on (typically English). Since the algorithms are not ASR-based, unclear or slurred speech will not cause the same errors as in transcription. Note that speech melody and expression vary between languages and cultures, which can affect accuracy. For example, Greek has naturally higher pitch variation than English, which might be misread as “excited” by some models. Many speech-based emotion recognition datasets use young-to-middle-aged actors. Age-related changes in the voice (pitch lowering in women, reduced clarity, slower speech) can cause misclassification. Moreover, in noisy, reverberant environments accuracy drops.

In this work we use the HuBERT-Large backbone fine-tuned for the SUPERB Emotion Speech Recognition (ESR) task. The base encoder is hubert-large-ll60k, a self-supervised model trained on approximately 60,000 hours of 16-kHz speech (LibriLight). During pretraining, HuBERT learns robust speech representations by masking spans of audio and predicting discrete pseudo-labels obtained from clustering acoustic features. For ESR, a lightweight classification head is attached to the frozen or partially fine-tuned encoder. The resulting checkpoint accepts raw 16-kHz mono waveforms and outputs posterior probabilities over four emotions.

In the SUPERB benchmark, the ESR task is framed as four-class utterance-level classification with the categories angry, happy, sad, and neutral. The widely used setup draws on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database and follows the common practice of focusing on these four balanced classes (other minority labels are omitted). The superb/hubert-large-superb-er checkpoint is fine-tuned to predict this label set.

Old adults with dementia do not demonstrate the expressivity of the young. Therefore, we merge the initial labels to improve robustness in this task. Agitation is a merged category of (‘angry’, ‘surprised’, ‘happy’, ‘disgust’) and indicates any vivid divergence from the neutral state. The neutral state is a merged category of (‘fear’, ‘sad’, ‘neutral’). Regarding the decision procedure, to handle long, real-world recordings, we apply sliding-window inference. Each input file is resampled to 16 kHz and segmented into fixed windows (default 2.0s) with overlap (0.5s hop). For every window, the model produces a probability distribution over the four emotions. To meet the binary requirement of the application, we map categories into two states: AGITATED includes angry and happy, while NEUTRAL includes neutral and sad. For each window we compute $p(\text{agitated})$ as the sum of the angry and happy probabilities, and we mark the window as AGITATED when $p(\text{agitated}) \geq 0.5$. A clip-level decision then uses a simple fraction rule: if at least 20% of windows are AGITATED, the entire clip is labeled AGITATED (see Figure 3a); otherwise, it is NEUTRAL (see Figure 3b). Window length, hop size, and the fraction threshold can be tuned to trade off sensitivity (catching brief shouts) versus specificity (avoiding false alarms on quiet, neutral speech).

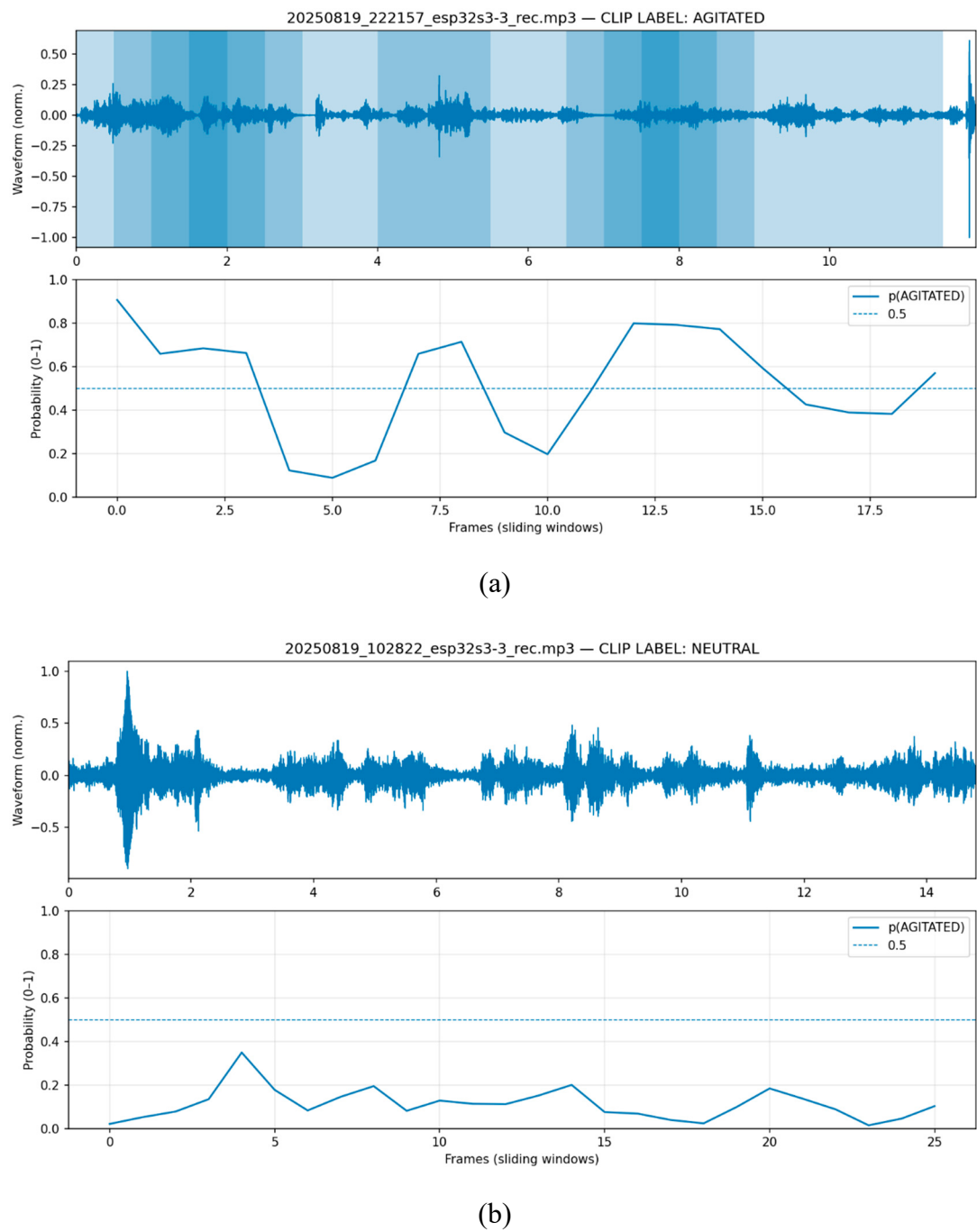


Figure 3. (a) ESR predicts an emotional state for each speech frame because agitation sometimes manifests itself in bursts, (b) A calm conversation.

Results

In this section we gather practical results and an evaluation of services.

Detecting Cough and Other Non-Verbal Audio Events

Cough events are counted according to the AST classification label (see Figure 4). In this dashboard we illustrate that, for 15 s snippets, the preprocessing module detects the presence of cough, enumerates cough bursts, and estimates cough intensity via the root-mean-square (RMS) amplitude of the waveform. Even in the context of a chronic cough, longitudinal variation in frequency and acoustic character is informative [42]: sudden increases may reflect intercurrent irritation or infection; changes in sound quality (e.g., wetter/drier timbre or superimposed wheeze) may indicate shifts in respiratory status; and temporal alignment with dialogue quantifies

communication impact (interruptions, post-cough cessation of speech) as well as correlations with time of day or specific activities. The AST recognizer has 527 audio sources and the frontend in Figure 4 allows filters to be applied on audio classes that queue and group the audio events of interest.

2025/08/05 → 2025/08/14

Eve...	Timestamp	Device	A ▾ ↑	Prob Cla...	Audio Clas... ≡	Prob Class 2 ≡	Audio CL...	Prob Class 3	RMS (dB)
171389	2025-08-14 15:08:32	esp32s3-3	Cough	0.65	Silence	0.19	Throat clear...	0.06	39.18
171388	2025-08-14 14:22:12	esp32s3-3	Cough	0.78	Throat clearing	0.09	Sneeze	0.08	37.84
171378	2025-08-14 10:34:52	esp32s3-3	Cough	0.6	Throat clearing	0.36	Sneeze	0.01	44.3
171371	2025-08-13 01:26:47	esp32s3-3	Cough	0.65	Throat clearing	0.31	Sneeze	0.01	35.12
171370	2025-08-13 01:25:24	esp32s3-3	Cough	0.57	Throat clearing	0.35	Sneeze	0.01	39.08
171369	2025-08-13 01:25:07	esp32s3-3	Cough	0.54	Throat clearing	0.44	Sneeze	0	40.67
171368	2025-08-13 00:18:09	esp32s3-3	Cough	0.69	Throat clearing	0.29	Sneeze	0	40.59
171270	2025-08-12 00:15:21	esp32s3-3	Cough	0.64	Throat clearing	0.34	Sneeze	0.01	38.49
171269	2025-08-12 00:14:41	esp32s3-3	Cough	0.77	Throat clearing	0.19	Sneeze	0.01	36.65
171265	2025-08-11 23:12:25	esp32s3-3	Cough	0.72	Throat clearing	0.23	Sneeze	0.02	43.35
171263	2025-08-11 23:04:14	esp32s3-3	Cough	0.68	Throat clearing	0.28	Sneeze	0.01	48.16
171262	2025-08-11 23:03:55	esp32s3-3	Cough	0.63	Throat clearing	0.35	Speech	0	47.42
171259	2025-08-11 23:00:41	esp32s3-3	Cough	0.67	Throat clearing	0.3	Speech	0	42.28
171254	2025-08-11 22:47:21	esp32s3-3	Cough	0.47	Throat clearing	0.35	Sneeze	0.05	46.19
171179	2025-08-11 20:31:00	esp32s3-3	Cough	0.57	Throat clearing	0.36	Sneeze	0.01	40.26
171172	2025-08-11 19:26:32	esp32s3-3	Cough	0.27	Speech	0.25	Throat clear...	0.09	58.65
171156	2025-08-11 00:00:31	esp32s3-3	Cough	0.71	Throat clearing	0.2	Sneeze	0.04	36.9
171083	2025-08-08 22:25:55	esp32s3-3	Cough	0.68	Throat clearing	0.25	Sneeze	0.03	42.28
171064	2025-08-08 22:04:09	esp32s3-3	Cough	0.85	Throat clearing	0.09	Sneeze	0.03	46.85
171059	2025-08-08 22:01:36	esp32s3-3	Cough	0.54	Sneeze	0.13	Throat clear...	0.11	44.03
170709	2025-08-08 13:26:01	esp32s3-3	Cough	0.03	Throat clearing	0.02	Sneeze	0.01	38.69
170697	2025-08-08 13:18:52	esp32s3-3	Cough	0.53	Throat clearing	0.27	Sneeze	0.13	34.96
170694	2025-08-08 13:12:03	esp32s3-3	Cough	0.36	Throat clearing	0.28	Sneeze	0.25	38.99
170693	2025-08-08 13:11:43	esp32s3-3	Cough	0.5	Throat clearing	0.3	Sneeze	0.13	37.62
170549	2025-08-08 12:21:18	esp32s3-3	Cough	0.07	Throat clearing	0.03	Sneeze	0.01	39.37
166843	2025-08-06 13:19:04	esp32s3-3	Cough	0.59	Throat clearing	0.3	Sneeze	0.06	34.96
166798	2025-08-06 13:04:35	esp32s3-3	Cough	0.53	Throat clearing	0.28	Sneeze	0.09	33.8

Figure 4. Asking the system to report the audio incidents of ‘cough’, ‘sneeze’, and ‘throat cleaning’ of an elderly couple in an in-house setting.

To operationalize these signals, we compute personalized baselines (e.g., coughs per hour) and multi-day deviation metrics. In Figure 5-top, we see the hours that the cough takes place over a configurable timespan. Figure 5-bottom allows us to see the rate of the cough-events per hour. The same way, mobility proxies are also captured by choosing the correct filters in the stream of captured-and-tagged audio events. In the AudioSet ontology these sounds belong to the classes ‘bouncing’, ‘wood’ and ‘tap’. Cane-tap cadence or the impact on the floor by other mobility aids variability provides a lightweight indicator of stability. Aggregating snippet-level outputs yields daily and weekly indicators (e.g., routine adherence) and supports conservative alert policies that prioritize precision for non-urgent notifications while enabling immediate escalation for high-risk events. The same framework detects and quantifies other clinically relevant sounds, including snoring, and sleep-related breathing.

Regarding accuracy of the AST transformer to classify cough (see Figure 4) and cane sounds, we have examined at least 500 positive and negative recordings without finding any error in the AST classification stage. AST’s cough classification is very accurate in in-home environments. Note the categories of ‘cough’, ‘sneeze’, ‘throat clearing’ have been merged. The VAD is used to pinpoint the event in the recording so that a precise measurement of the RMS intensity and number of incidents is logged (see Figure 6).

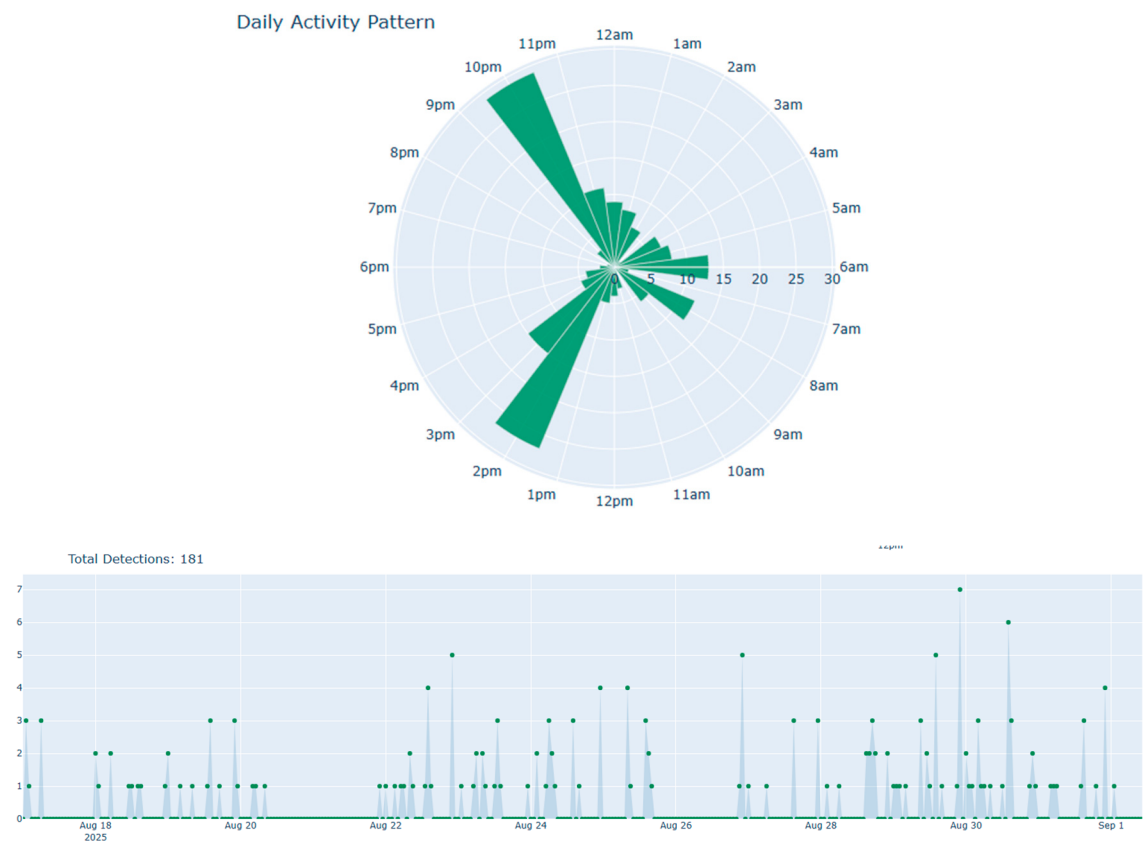


Figure 5. (Top) Hourly activity of coughing from 17/08/2025-1/9/2025 and, (Bottom) time series visualization that assesses rate and trend in cough incidents. Note that the audio category is configurable and we can visualize any of the 527 audio categories of the AST ontology.

Once events are recognized and logged, visualizations can support caregiver interpretation. Figure 7 illustrates one such summary: the couple briefly woke around 05:00 (speech detected) and then slept again, waking for the day after ~09:00. Speech dominates the day’s activity (~68% of 177 detections). Notably, there are extended conversations after the 22:30 bedtime.

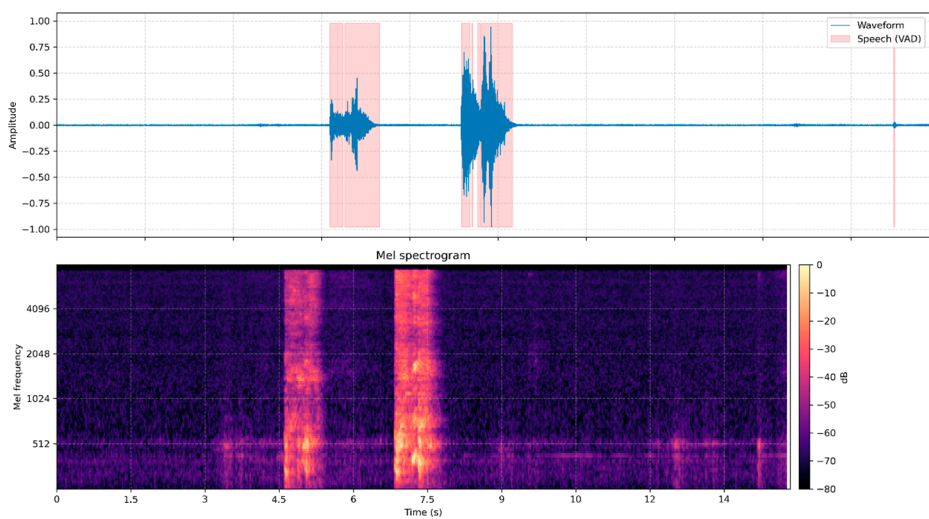


Figure 6. Voice activity detection on a cough recording. (Top) The vocal segments are identified and shaded, and the RMS value of the event and the number of events can be measured precisely. (Bottom) Spectrogram of a typical cough event.

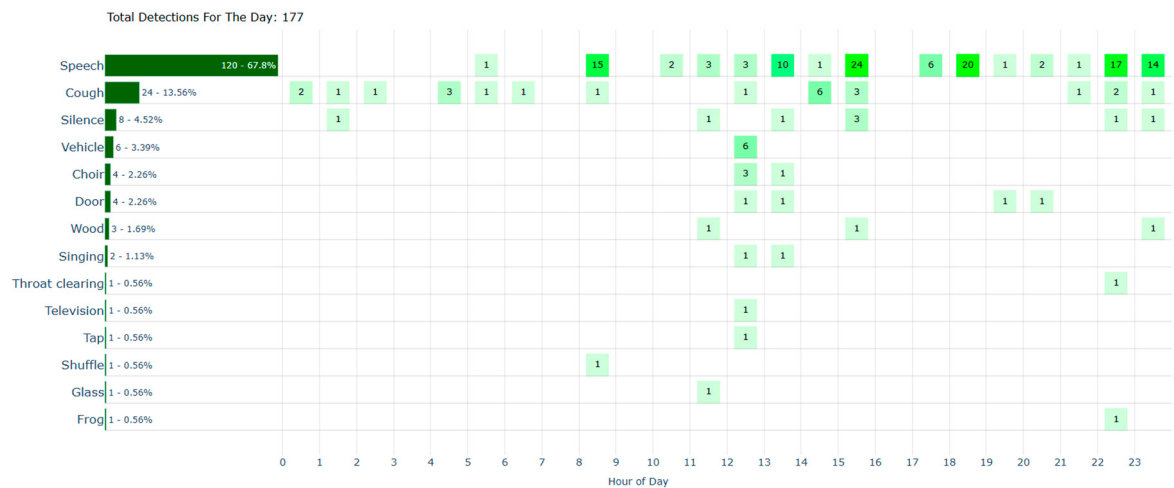


Figure 7. Audio event classification and hourly distribution (n=177) of an elderly couple with dementia in home. The Y-axis shows the detected classes and green boxes their relative rates, while the X-axis localizes their occurrence over time.

In Figure 7, aside from intermittent coughs during the night, sleep appears relatively uninterrupted. Cough events constitute ~14% of detections and occur intermittently across the day—few in the early morning, a larger cluster in the early–mid afternoon, and several in the late evening. Cane-related impacts (“bounce,” “tap,” “wood”) are rare and isolated, suggesting minimal cane striking. No sounds indicative of a fall, pain, or distress were detected on this day.

Diarization

Diarization has been notably successful even in a few-shot learning scheme. We extracted 20 recordings 15 sec each for every person in the home environment to train the diarization application (see Appendix). The people are A: Relative 1, B: Male patient, C: Female patient, D: Caregiver 1, E: Caregiver 2, F: Relative 2. For finetuning, we included recordings including conversations such as B+C, A+B+C etc. 20 recordings per case. The test set of 100 recordings was taken within one month from the training set. The results are very promising and have been achieved with few recordings per speaker entailing that more thorough training can achieve near perfect results.

In Tables 1-3 we present various accuracy metrics for this multilabel, multiclass problem. In multilabel evaluation, subset accuracy is the proportion of instances for which the predicted label set exactly matches the true set (all labels correct, non missing or extra). Hamming loss measures the average fraction of misclassified label decisions over all instance–label pairs; it penalizes both missing positives and spurious positives and is lower when performance is better. The Jaccard index (intersection-over-union) compares predicted and true; Jaccard (micro) first pools true positives, false positives, and false negatives over all labels and instances, then applies the Jaccard formula; Jaccard (macro) computes Jaccard per label and averages labels equally, regardless of frequency. The F1 score is the harmonic mean of precision and recall: F1 (samples) computes per-instance F1 and averages across instances; F1 (micro) pools counts globally before computing precision, recall, and F1 (favoring frequent labels); and F1 (macro) averages per-label F1 scores (treating each label equally and highlighting performance on rare labels). In Table 2 we gather the results with emphasis on the multilabel case

Table 1. Multilabel Multiclass Metrics.

Metric	Value
Subset accuracy	0.66
Hamming loss	0.07

Jaccard (samples)	0.80
Jaccard (micro)	0.73
Jaccard (macro)	0.80
F1 (samples)	0.86
F1 (micro)	0.84
F1 (macro)	0.88

In classification tasks, precision quantifies the proportion of predicted positives that are truly positive, i.e., the ability of the model to avoid false positives. Recall (also called sensitivity) measures the proportion of true positives that are correctly identified among all actual positives, reflecting the ability to avoid false negatives. The F1-score is the harmonic mean of precision and recall, balancing the trade-off between them and providing a single metric that is especially informative when class distributions are imbalanced). In Table 3 we gather the results on a per speaker basis.

Table 2. Single class Metrics.

Speaker	Precision	Recall	F1-score	Support
A	1.00	0.92	0.96	12
B	0.98	0.77	0.86	57
C	0.97	0.60	0.74	53
D	1.00	0.89	0.94	9
E	0.78	0.95	0.86	19
F	0.89	1.00	0.94	8

In multi-class or multi-label evaluation, micro average computes a metric (e.g., precision, recall, F1, Jaccard) after aggregating true positives, false positives, and false negatives across all classes and instances, emphasizing performance on frequent classes. Macro average first computes the metric separately for each class and then takes the unweighted mean, giving equal weight to rare and common classes. Weighted average is like macro but weighs each class’s metric by its support (number of true instances), mitigating the influence of very rare or very common classes on the overall score. Samples average is specific to multi-label settings: it computes the chosen metric per instance by comparing its predicted label set to its true set, then averages these per-instance values over all samples, reflecting how well complete label sets are recovered for each example. In Table 4 we gather the results on micro and macro averaging.

Table 3. Micro and macro-averaging Metrics.

Average type	Precision	Recall	F1-score	Support
Micro avg	0.94	0.77	0.84	158
Macro avg	0.94	0.85	0.88	158
Weighted avg	0.95	0.77	0.84	158
Samples avg	0.95	0.82	0.86	158

Table 4. Metrics of the emotion recognition task. There are two classes: ‘agitated’ and ‘neutral’.

Class	Precision	Recall	F1
Neutral	0.91	0.91	0.91
Agitated	0.91	0.91	0.91
Accuracy			0.91
Macro avg	0.91	0.91	0.91
Weighted avg	0.91	0.91	0.91

Emotion Recognition

We have pulled together a corpus of 22 recordings in agitated and 22 in neutral state. The agitated state is not common for this couple while in neutral there are thousands of recordings. The algorithm has been tuned so that it picks up agitation even in small phrases (abrupt and short-time raise of voice). In Table 4, we gather the metrics and in Figure 8 we show the confusion matrix of this task. The metrics are the same as in the diarization case.

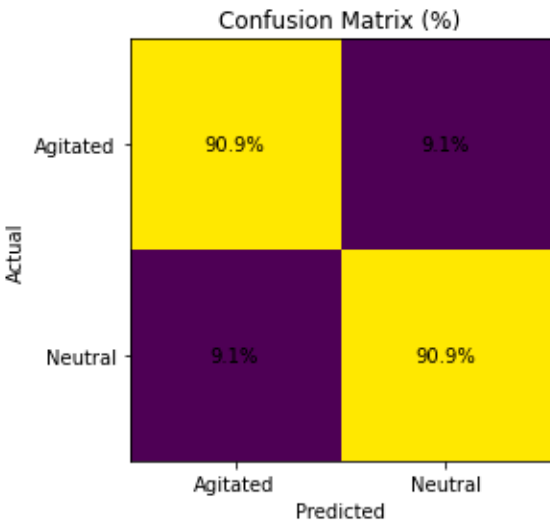


Figure 8. Confusion Matrix of emotion recognition between the ‘agitated’ and ‘neutral’ class.

Safety & Emergency Detection

Pharmacologic sleep aids do not reliably ensure uninterrupted nocturnal rest in dementia; patients may still engage in nighttime wandering or purposeless activity, such as searching for food, watching television, or repeated trips to the toilet. These unattended nocturnal trips are dangerous for their safety as they may result in a fall. This has happened repeatedly for this couple resulting in bone fractures, agony incidents and hospitalization. Elders falling at this age are much more likely to experience bone fractures due to osteoporosis. Caregiving during nights is exceptionally problematic because it raises abruptly the cost of this service and reliable service is hard to find. Our audio-based monitoring device provided unobtrusive nocturnal activity surveillance to detect out-of-bed movement and wandering. The benefits of visualizing activity proxies during night is presented in detail in the Discussion section. The device can record events that indicate pain and discomfort are sometimes identifiable through groans, moans, or sharp exclamations, and may overlap acoustically with anger or anxiety. Finally, prolonged silence detection during active hours prompts a check-in. The LLM can integrate the information before it reaches a decision on probable falls from characteristic impact transients (thuds/crashes) followed by silence or distress vocalizations and recognize shouting or calls for help trigger immediate alerts to relatives and caregivers. The program at the server can subsequently send push notification via e-mail and short messages to relatives and caregivers for a probable emergency.

Caregiver Support & Reporting

We extract the audio events classified by the AST audio recognizer into a JSON file format. The structure of each event is as follows: {"eventid": "evt171390", "timestamp": "2025-08-14 15:43:20", "audio_events": [{"class": "Cough", "probability": 0.37}, {"class": "Silence", "probability": 0.21}, {"class": "Throat clearing", "probability": 0.15}], "rms": 38.69}. The Python program time-stamps the event and the AST registers the first higher in rank classes and their corresponding probabilities and the RMS intensity of the audio event. The LLM (i.e. ChatGPT 5 in our case) was given the JSON file for the current week, the context of the application, the context of the subjects and was asked to prepare a

report summary of the previous day for the caregiver for the following day. The structure of the report would be a short log of their activity, emotional state, and detected issues as well as flagging important audio snippets for medical review. Full reports are included in the Appendix. In Table 5 we give a short extract.

Table 5. ChatGPT-5 prepares a report summarizing key points of health-relevant notes for supporting the caregiver based on a daily file of recognized audio events, the context of the application and a short history of the couple. At this stage of analysis, speech and conversations are not transcribed.

Health-relevant signals (audio only) • Cough burden (afternoon spike). Six high-confidence cough events occurred 13:45–15:48, aligning with the known chronic cough. Consider hydration, inhaler timing, or irritant exposure review. • Breath/vocal strain (AM). One gasp/breathing snippet at 08:47 without follow-up distress markers; continue to watch for clustering. • Mobility / impact sounds. Short bouncing/wood-tap series around 16:55–16:57 may reflect cane contacts or small object taps; no escalations or falls detected. Engagement & orientation • Conversation density. Peak conversational activity occurred 13:35–13:48 and 17:47–17:51 near the sensor, suggesting caregiver interaction or couple dialogue proximal to bedroom/bath corridor. • Night-time/bedtime not confirmed: last record 18:14; no data later in the evening to verify final bathroom/bedroom transition.
--

Delusion Detection

Detecting delusions and/or hallucinations or even suicidal ideation [43] is valuable information for care giving as they can track the development of dementia and detect a possibly harmful action. A delusion is a fixed false belief held despite clear evidence to the contrary. *Example:* the person insists “this isn’t my home—I need to go to my other house,” starts packing, and rejects reassurance. A hallucination is a sensory perception without an external stimulus (seeing, hearing, or feeling something that isn’t there). *Example:* the person says “the kids are sitting in the sofa” when no one is present. What we suggest in this work is that these events, if manifested verbally, can be recognized by the LLM and, therefore, counted. A more elaborate report, considering these issues, can be asked if we feed the LLM with the transcribed speech of the subjects and features of their emotional state and dialogs. The aim is to identify events of distress, pain, increasing confusion, agitation, and delusions. In our case, the female has developed two recurrent delusional patterns: a) she needs to abandon the current home and move back to the first house of her youth. Although she has been living in her last home for over 30 years, she does not recognize it as her own house, b) a form of delusion that maybe develops to a hallucination where another ‘lady’, is inside the house. Besides the caregivers, sporadic visits by relatives, there are no other people living in the house. The validity of

such events has been verified by the caregivers themselves and the IP cameras installed in the rooms. In Appendix, we have such a report an extract of which we give in Table 6 for clarity.

The AST classifier tags the speech events as ‘Speech’, ‘Conversation’, ‘Male speech, man speaking’, ‘Female speech, woman speaking’. These events can be automatically directed to an ASR application, but this faces the limitations of current ASR we presented above. To investigate if ChatGPT could actually discern such events given reliable transcriptions from ASR, we transcribed manually the dialogues from 50 recordings and asked the LLM to pinpoint the events if present and prepare a report. Regarding the other speech processing modules, diarization ESR, and VAD work quite efficiently and achieve high accuracy rates. We have instructed ChatGPT-5 to have a look at the transcribed conversations and based on a small history of the couple, to detect delusional incidents (see Appendix for history and prompts). In this corpus 5 cases have delusional speech and 45 cases of normal speech recordings. Note that the delusional patterns of the patients have not been revealed explicitly and *a-priori* to GPT and the detection is based on the trained knowledge and reasoning ability of GPT-5 to decompose and integrate information and decide based on the logical contradiction of facts between the short personal history of the couple and their dialogues. Remarkably, ChatGPT-5 discovered easily all five delusional dialogues in the corpus of 50 dialogues. The system has not misinterpreted non-delusional conversations that were consistent with confusion, memory error, or everyday difficulty. In Appendix, the prompts, dialogues and the tasks are given so that the results can be examined.

Table 6. Prompting ChatGPT-5 with a) small history of the couple, b) the context of the application, c) discussions, audio labels, diarization (Male, Female), emotion tagging (Neutral, Agitated). Diarization and emotion tags are in brackets.

PROMPT From now on, act as my expert assistant with access to all your knowledge in the domain of elderly care, elderly psychology, medicine, dementia and best treatment practices of patients with Alzheimer’s. Always provide: A clear, direct answer to my request. A step-by-step explanation of how you reached a conclusion. A practical summary or action plan I can apply immediately. Never give vague answers and act like a professional. Push your reasoning to 100% of your capacity. <Small history of the couple>

Transcribed dialogues with automated diarization and emotional tagging

1. [F][N] Maki did she leave? [M][N] Who? [F][N] Gogo, did she leave? [M][A] I don't know leave me alone!
2. [F][N] I am going to pack our things? [M][N] Why? [F][N] To go home. [M][N] What are you talking about? [F][N] The owners may come back
3. [M][N] Did we eat today? [F][N] I don't remember

RESULTS

Dialogue 2 shows a delusional theme (place misidentification / “this isn’t our home”) with exit-seeking intent. The lines “I am going to pack our things... To go home... The owners may come back” indicate a fixed false belief that they are in someone else’s property and must leave. This contradicts the history (they have lived in this home for ~30 years) and fits the female’s known occasional delusions. **Conclusion: Delusion detected (high likelihood).**

Dialogue 1 suggests anxiety/checking behavior, not a delusion. Asking if the caregiver “Gogo” left is a reality-check; no false belief is asserted. The male’s irritable reply does not introduce delusional content. **Conclusion: No delusion detected (low likelihood).**

Dialogue 3 reflects memory impairment, not a delusion. “Did we eat today?” / “I don’t remember” is short-term memory loss. No false belief is stated. **Conclusion: No delusion detected (very low likelihood).**

Discussion

Practical Achievements so Far

Our suggestion is affordable and easily deployable intervention. Caregiving is typically an arrangement under budget constraints, and the configuration studied here—without loudspeakers or actuators—targets practical gains from passive audio monitoring. We gather practical gains after using the system for over two months continuously: a) Our approach for overnight surveillance lowers staffing needs for night shifts, which are costly and often inefficient. Continuous fall monitoring at night, reducing family anxiety given the difficulty of calling for help after a fall. b) The recognition of audio and its hourly distribution revealed that after caregivers left the subjects frequently rose, walked, and conversed, leading to daytime exhaustion that had been misattributed; inspection of the system’s “speech” and “conversation” labels between 00:00 and 07:00 prompted a medication adjustment by a doctor that resolved this pattern. c) The suggested pipeline quantifies delusional episodes over time, enabling trend tracking, and measures cough burden before and after treatment to evaluate treatment’s efficacy. d) Systematic analysis of conversational content provided actionable insight into psychological status and stressors, enabling individualized care plans that target minor yet recurring practical stressors (e.g., arrangement of objects, room temperature, lighting conditions) that degrade well-being of dementia patients and are commonly not reported the following day due to memory impairment. Finally, e) We valued the system’s uninterrupted (24/7) reliability, which ensured steady monitoring and augmented care during staffing constraints or caregiver absence, easing the burden on adult children with legal guardianship responsibilities.

Scaling up the Service

Although our primary goal is to deliver affordable, automated in-home services, the architecture scales directly to facility-wide deployments. Field visits to elderly care facilities in the city of Patras, Greece revealed multi-floor layouts with consecutive rooms housing 2–4 residents in heterogeneous clinical states, often with only a subset of institutions admitting people with dementia or Alzheimer’s disease and with wide variation in service level and cost. These characteristics—repeatable room topology, diverse care needs, and operational heterogeneity—make a centralized, multi-room deployment both feasible and valuable, enabling consistent monitoring policies across wards while preserving room-level personalization. The proposed combination of audio signal processing and AST classifiers combined with LLMs stack yields a live “situational awareness board” that could monitor all rooms simultaneously, surface deviations (e.g., sudden spikes in cough burden, impact-like thuds, delusional dialogues), enabling oversight across an entire ward. The infrastructure needed can be deployed unobtrusively, it is low-cost, and all nodes can be processed by a single session software on a laptop. Bedside evolve into conversational endpoints where the LLM engages residents and staff through brief, goal-directed dialogues and concise screen summaries—probing orientation, interpreting confusion or fixed false beliefs, and continuously “listening” for accidents, pain, distress and alarms—while emitting auditable, structured assessments and suggested de-escalation or check-in scripts. Cameras with audio can also be used but our audio modality is more discrete, less invasive, more affordable and installed in less than 3 seconds. Affordability and easy deployment are the key advantages of our approach.

In the envisioned deployment, large language models (LLMs) coordinate with assistive robots and screens—both mobile platforms and bedside devices—within a closed-loop control framework. The LLM translates detected needs into verifiable action primitives, monitors execution via on-board sensor feedback, and hands off to human staff when uncertainty or risk thresholds are exceeded. Continuous monitoring and self-adaptation target conservative escalation policies that prioritize precision for non-urgent events and immediate alerts for hazards. Over time, federated learning enables room- and resident-specific personalization while preserving central governance and auditability. This approach is intended to augment—not replace—human care by converting ambient sound into continuously updated, clinically relevant signals and delegating low-risk, routine assistance to restless machines, thereby reserving clinician and caregiver effort for difficult judgment and empathy that are far from what a machine can provide.

Retrieval-Augmented Generation

LLMs can hallucinate as well and to reduce this chance future work will rely on retrieval-augmented generation (RAG) to keep the LLM factual, personalized, and auditable. Retrieval-augmented generation (RAG) pairs a generator (e.g., an LLM) with a search/retrieval step that pulls relevant, up-to-date documents or records at query time; the model then grounds its answer onto those sources. Its use can reduce hallucinations, inject patient/home context (logs, protocols, prior episodes), and enable auditable, citation-backed outputs for care summaries, alerts, and decision support.

AST event tags (e.g., cough, cane taps), diarization spans, N-best ASR hypotheses with word confidences, keyword-spotting hits, and per-snippet acoustics—can be serialized into time-stamped “episode” documents and stored locally in a vector+BM25 index. A second index can hold a consented household knowledge base (names, address, routines, safety rules), caregiver playbooks (de-escalation scripts, orientation prompts), and facility protocols (alert thresholds, fall procedures). For each new snippet, a compact “query document” (current features + intent, e.g., “assess exit-seeking/place disorientation”) retrieves the top-k episodes and relevant playbook passages with time decay and speaker filters; retrieval scoring weights high-confidence words, repeated themes, and corroborating non-speech events. The LLM is then constrained to emit a structured JSON record (observations vs. inferences, confidence, and source references to retrieved snippets) and never reason beyond retrieved evidence.

Conversation & Cognitive Support Using a Loudspeaker

We envision that an LLM autonomously engages in conversations (AI-bots) that stimulate patients’ memory, mitigate the sense of loneliness but at the same time the LLM self-adapts its form of interaction to the special case of each patient’s history and behavioral patterns but also performs primary diagnostical tests by gathering evidence from the dialogue and ESR (see Figure 9). This would be useful, especially in the case of the important category of an alone elder that will otherwise not speak a lot. Therefore, though mentioned briefly, the present work can be further expanded to an AI-bot, that is a conversational machine that cooperates with actuators such as loudspeakers and screens and engages in conversations with the elderly. This implies that the speech recognizer is already adapted for elderly speech. A loudspeaker-based conversational interface employs a large-language-model agent (e.g., ChatGPT) as a real-time dialogue partner to mitigate loneliness and support memory and cognition. Messages are created with a text-to-speech (TTS) interface and are communicated via loudspeakers located in mobile phones, smart TV’s or standalone devices. The agent conducts brief, socially engaging exchanges and delivers multimodal memory prompts (spoken and on-screen) that queue names, relationships, and upcoming events. When agitation or persevering questioning is detected by upstream classifiers or conversational signals, the agent resorts to gentle redirection to de-escalate. It also elicits autobiographical reminiscence to stimulate memory. Personalization is integral: the system adapts speaking rate, prosody, and linguistic complexity to the user’s current cognitive state; maintains a consented personal knowledge store to tailor references and improve recognition; and learns daily routines so that reminders and prompts are context- and time-appropriate.

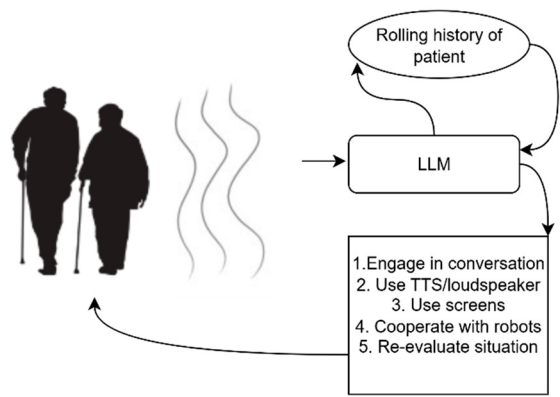


Figure 9. Future audio based elderly care with dementia based on Audio and AI. The LLM engages in conversations with the subjects through a TTS and loudspeakers and re-evaluates the situation as it evolves in time. It cooperates with actuators (screens, robots, telephone) and answers queries of a caregiver based on historical audiovisual data.

Adapting Speech Processing Modules

As the older-to-younger demographic imbalance is likely to persist in the Western world—driven by cultural factors and advances in medicine—older adults present a significant opportunity for advanced AI applications. Existing ASR, emotion-recognition, and diarization systems should be adapted to this population.

Conclusions

Affordable, commodity sensors and AI can supplement—and in underserved settings partially substitute—in-home caregiving for older adults with dementia when dependable services are unaffordable or unavailable. We integrated recent advances in audio processing with LLM-based reasoning and highlighted the limitations of ASR in age-related voice changes, motivating a pipeline

that leverages non-speech acoustic events alongside transcripts. This manuscript is a proof-of-concept, not a definitive clinical study. We prototyped an ESP32-S3 node with an onboard MEMS microphone and note that commodity smartphones and smart TVs are viable, low-cost recorders. Short audio snippets from daily activity are streamed almost in real time to a recognizer with 527 acoustic classes, with speech transcription and emotion inference producing structured cues that are passed to an LLM for situation assessment. We demonstrate end-to-end feasibility in a private home with real patients, introduce reusable tooling (open code/data schemas, structured outputs, safety policies), and provide early longitudinal evidence that clinically relevant signals can be tracked. The range of services under the same concept can be significantly expanded with the future incorporation of loudspeakers, house-robots and screens.

Informed Consent Statement: The study subjects are the author’s parents, who due to advanced dementia are no longer able to provide informed consent. Legal consent for participation, recording, and the open sharing of de-identified audio data was granted by the author and the author’s sister, acting as their guardians by law. The caregivers have been notified that all rooms are under surveillance using cameras and audio for the prevention of accidents or crime. Physical mute control was available at all times.

Data Availability Statement: The dataset of audio events is available from the corresponding author upon reasonable request.

Acknowledgement: The study was conducted under guardian-provided, explicit consent in accordance with GDPR Art. 6(1)(a) and Art. 9(2)(a), with capacity loss and the guardianship order documented. A concise DPIA was completed covering home audio capture, EU cloud storage with client-side encryption, retention limits (raw ≤30 days), room scoping, assent/dissent procedures, and a no-open-raw-data policy (controlled access under data-use agreement only)

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DSP	Digital signal processing
RMS	Root, mean square value
VAD	Voice activity detection
ASR	Automatic speech recognition
ESR	(Automatic) Emotion Speech Recognition
TTS	Text to speech
AST	Audio Spectrogram Transformer
IoT	Internet of things
MQTT	Message Queuing Telemetry Transport
RAG	retrieval-augmented generation
VIT	Vision Transformer
JSON	JavaScript Object Notation
MP3	Moving Picture Experts Group Audio Layer III

LLM	Large language model
MEMS	Micro Electro Mechanical Systems

Appendix

Code used in this work: <https://github.com/potamitis123>

The links of all toolboxes used can be found below.

General audio recognition

Audio Spectrogram Transformer (AST) applies a Vision Transformer to audio by converting waveforms to log-Mel spectrograms, tokenizing them into patches, and feeding them to a transformer encoder for classification. The library provides ASTFeatureExtractor to compute and normalize Mel features (AudioSet mean/std by default) and ASTForAudioClassification/ASTModel for inference or fine-tuning, with configuration parameters like patch size, time/frequency strides, and number of Mel bins exposed via ASTConfig. The docs highlight practical tips and show ready-to-run examples with the pretrained MIT/ast-finetuned-audioset-10-10-0.4593 checkpoint.

https://huggingface.co/docs/transformers/model_doc/audio-spectrogram-transformer

VAD

TEN is an open-source ecosystem for creating, customizing, and deploying real-time conversational AI agents with multimodal capabilities including voice, vision, and avatar interactions.

<https://github.com/TEN-framework/ten-vad/>

ASR

Whisper is a state-of-the-art model for automatic speech recognition (ASR) and speech translation. The Whisper large-v3 model was trained on 1 million hours of weakly labeled audio and 4 million hours of pseudo-labeled audio collected using Whisper large-v2. The model was trained for 2.0 epochs over this mixture dataset. The large-v3 model shows improved performance over a wide variety of languages, showing 10% to 20% reduction in errors compared to Whisper large-v2

<https://huggingface.co/openai/whisper-large-v3>

ESR

superb/hubert-large-superb-er is a HuBERT-Large (hubert-large-ll60k) model fine-tuned for the SUPERB Emotion Recognition task. It takes 16-kHz mono audio and predicts one of four utterance-level emotions—angry, happy, sad, or neutral—following the standard SUPERB/IEMOCAP setup that drops minority classes to balance the dataset. The checkpoint is a port of the S3PRL implementation, providing a simple sequence-classification head on top of HuBERT’s self-supervised speech representations. Usage is straightforward via the Transformers audio-classification pipeline or the underlying model/feature extractor; the model card reports accuracy on the SUPERB demo split and includes example code. In practice, it’s a compact, well-documented baseline for categorical speech emotion recognition that you can adapt or fold into application-specific labels (e.g., “agitated” vs “neutral”).

https://huggingface.co/superb/hubert-large-superb-er?utm_source=chatgpt.com

Diarization

pyannote.audio is an open-source toolkit written in Python for speaker diarization. Based on PyTorch machine learning framework, it comes with state-of-the-art pretrained models and pipelines, that can be further finetuned to our data for even better performance.

<https://github.com/pyannote/pyannote-audio>

Delusion detection

GPT-5 set at ‘thinking mode’. The prompts, the history and the dialogues should recreate the results on the delusional patterns detection.

PROMPT: From now on, act as my expert assistant with access to all your knowledge in the domain of elderly care, elderly psychology, medicine, dementia and best treatment practices of

patients with Alzheimer’s. Always provide: A clear, direct answer to my request. A step-by-step explanation of how you reached a conclusion. A practical summary or action plan I can apply immediately. Never give vague answers and act like a professional. Push your reasoning to 100% of your capacity.

TASK: We have in-house recordings of the activity of a couple with dementia stage 2. I will provide their discussions, and you need to answer queries based on these discussions and the small history of the couple.

Query #1: Detect delusional expressions in dialogues.

SMALL HISTORY OF THE COUPLE

- a) Female: 87; name variants: Vagelia / Vagelio / Vago / Vangelitsa. Male: 90+; name variants: Maki / Makis / Mak.
- b) Diagnosis: both have dementia (stage 2); only the female has occasional delusions.
- c) Mobility: both have mobility impairments; use canes; no capability to leave the home; they can’t access the garden due to a steep staircase; they can’t cook or make hot beverages.
- d) Memory: acute short-term memory loss/forgetfulness.
- e) Care context: they live in their own home; two daytime caregivers (Eve and Gogo) cover 8-hour shifts; caregivers have been working with the couple for two years; one adult daughter and one other relative visit occasionally; no other people are in the home; Their parents died fifty years ago.
- f) Residency & property: they have lived in the current home for ~30 years; before that, only their youth home (>30 years ago); they do not own any other house.

Query #2: At night, the couple is unattended by caregivers and there is no one other than the couple in home. There are no pets, no music, no tv set on. Pay attention to transcribed audio patterns in the attached JSON file that may involve a hit, fall, followed by vocal expressions of distress, pain and/or anxiety.

Query #3: Prepare a report summarizing the main events of the previous day by parsing the attached JSON file containing tagged audio events of the previous day. Each event receives the 3 higher rank labels and their corresponding probabilities and a timestamp.

DIALOGUES

Dialogues and results on Q1 are appended to the manuscript as supporting material.

References

1. World Health Organization: <https://www.who.int/news-room/fact-sheets/detail/dementia?> (Accessed on 14/9/2025)
2. EUROSTAT: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_structure_and_ageing (Accessed on 14/9/2025)
3. United Nations: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/undesd_pd_2024_wpp_2024_advance_unedited_0.pdf?utm_source=chatgpt.com (Accessed on 14/9/2025)
4. Alzheimer’s Association: <https://www.alz.org/alzheimers-dementia/facts-figures?> (Accessed on 14/9/2025)
5. König, A., Satt, A., Sorin, A., Hoory, R., Derreumaux, A., David, R., Robert, P.-H. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer’s disease. *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1 (1), 112–124. <https://doi.org/10.1016/j.dadm.2014.11.012>
6. Haider, F, De La Fuente Garcia, S & Luz, S 2020, 'An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer’s Dementia in Spontaneous Speech', *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272-281. <https://doi.org/10.1109/JSTSP.2019.2955022>
7. Qi X, Zhou Q, Dong J, Bao W. Noninvasive automatic detection of Alzheimer’s disease from spontaneous speech: a review. *Front Aging Neurosci*. 2023 Aug 24;15:1224723. doi: 10.3389/fnagi.2023.1224723.
8. Saeedi S, Hetjens S, Grimm MOW, Latoszek BB. Acoustic speech analysis in Alzheimer’s disease: a systematic review and meta-analysis. *J Prev Alzheimers Dis*. 2024. PMID: 39559890.

9. Martínez-Nicolás I, Llorente TE, Martínez-Sánchez F, Meilán JJG. Ten years of research on automatic voice and speech analysis of people with Alzheimer's disease and mild cognitive impairment: a systematic review article. *Front Psychol.* 2021;12:620251. doi:10.3389/fpsyg.2021.620251.
10. van den Berg, R.L., de Boer, C., Zwan, M.D. *et al.* Digital remote assessment of speech acoustics in cognitively unimpaired adults: feasibility, reliability and associations with amyloid pathology. *Alz Res Therapy* **16**, 176 (2024). <https://doi.org/10.1186/s13195-024-01543-3>
11. Liu, J.; Fu, F.; Li, L.; Yu, J.; Zhong, D.; Zhu, S.; Zhou, Y.; Liu, B.; Li, J. Efficient Pause Extraction and Encode Strategy for Alzheimer's Disease Detection Using Only Acoustic Features from Spontaneous Speech. *Brain Sci.* **2023**, *13*, 477. <https://doi.org/10.3390/brainsci13030477>
12. Meilán JJ, Martínez-Sánchez F, Carro J, López DE, Millian-Morell L, Arana JM. Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia? *Dement Geriatr Cogn Disord.* 2014;37(5-6):327-34. doi: 10.1159/000356726.
13. Xue C, Karjadi C, Paschalidis ICh, Au R, Kolachalama VB. Detection of dementia on voice recordings using deep learning: a Framingham Heart Study. *Alzheimers Res Ther.* 2021;13:146. doi:10.1186/s13195-021-00888-3. BioMed Central
14. Ding K, He C, Chen Z, Xu X, Li W, Hu B. Using acoustic voice features to detect mild cognitive impairment: machine learning model development study. *JMIR Aging.* 2024;7:e57873. doi:10.2196/57873.
15. Vincze V, Sztalóczy G, Tóth L, Gosztolya G, Pákási M, Hoffmann I, Kálmán J. Telltale silence: temporal speech parameters discriminate between prodromal dementia and mild Alzheimer's disease. *Clin Linguist Phon.* 2021. doi:10.1080/02699206.2020.1827043.
16. Sefcik JS, Ersek M, Hartnett SC, Cacchione PZ. Integrative review: Persistent vocalizations among nursing home residents with dementia. *Int Psychogeriatr.* 2019 May;31(5):667-683. doi: 10.1017/S1041610218001205.
17. Kusters J, Janus SIM, Van Den Bosch KA, Zuidema S, Luijendijk HJ, Andringa TC. Soundscape Optimization in Nursing Homes Through Raising Awareness in Nursing Staff With MoSART. *Front Psychol.* 2022 Jun 1;13:871647. doi: 10.3389/fpsyg.2022.871647.
18. Prabhjot Kaur, Qifan Wang, Weisong Shi, Fall detection from audios with Audio Transformers, *Smart Health*, Volume 26, 2022, 100340, ISSN 2352-6483, <https://doi.org/10.1016/j.smhl.2022.100340>.
19. Newaz NT, Hanada E. The Methods of Fall Detection: A Literature Review. *Sensors (Basel).* 2023 May 30;23(11):5212. doi: 10.3390/s23115212
20. Li, Y.; Liu, P.; Fang, Y.; Wu, X.; Xie, Y.; Xu, Z.; Ren, H.; Jing, F. A Decade of Progress in Wearable Sensors for Fall Detection (2015–2024): A Network-Based Visualization Review. *Sensors* **2025**, *25*, 2205. <https://doi.org/10.3390/s25072205>
21. Rocha, I.C.; Arantes, M.; Moreira, A.; Vilaça, J.L.; Morais, P.; Matos, D.; Carvalho, V. Monitoring Wearable Devices for Elderly People with Dementia: A Review. *Designs* **2024**, *8*, 75. <https://doi.org/10.3390/designs8040075>
22. Saradopoulos, I.; Potamitis, I.; Ntalampiras, S.; Rigakis, I.; Maniavas, C.; Konstantaras, A. Real-Time Acoustic Detection of Critical Incidents in Smart Cities Using Artificial Intelligence and Edge Networks. *Sensors* **2025**, *25*, 2597. <https://doi.org/10.3390/s25082597>
23. Parlak S, Feldstein Ewing SW, Mecoli M, et al. Alzheimer's disease-related differences in acoustic measures of brief, unstructured speech. [Journal info to verify]. PMID: [to verify]. (Listed on PMC; please confirm journal/volume/pages for final formatting.)
24. Yusupov I, Douglas H, Lane R, Ferman T. Vocalizations in dementia: a case report and review of the literature. *Case Rep Neurol.* 2014;6(1):126-133. doi:10.1159/000362159.
25. Fillit H, Aigbogun MS, Gagnon-Sanschagrin P, Cloutier M, Davidson M, Serra E, et al. Impact of agitation in long-term care residents with dementia in the United States. *Int J Geriatr Psychiatry.* 2021;36(12):1959-1969. doi:10.1002/gps.5604.
26. Pedersen SKA, Andersen PN, Lugo RG, Andreassen M, Sütterlin S. Effects of Music on Agitation in Dementia: A Meta-Analysis. *Front Psychol.* 2017 May 16;8:742. doi: 10.3389/fpsyg.2017.00742
27. Miller S, Vermeersch PE, Bohan K, Renbarger K, Kruep A, Sacre S. Audio presence intervention for decreasing agitation in people with dementia. *Geriatr Nurs.* 2001 Mar-Apr;22(2):66-70. doi: 10.1067/mgn.2001.115200.

28. Shah R, Basapur S, Hendrickson K, Anderson J, Plenge J, Troutman A, Ranjit E, Banker J. Does an Audio Wearable Lead to Agitation Reduction in Dementia: The Memesto AWARD Proof-of-Principle Clinical Research Study. *Res Sq [Preprint]*. 2025 Feb 17;rs.3.rs-6008628. doi: 10.21203/rs.3.rs-6008628/v1
29. Alsina-Pagès, R.M.; Navarro, J.; Alías, F.; Hervás, M. HomeSound: Real-Time Audio Event Detection Based on High Performance Computing for Behaviour and Surveillance Remote Monitoring. *Sensors* **2017**, *17*, 854.
30. Rashidi, P.; Cook, D.J. Keeping the Resident in the Loop: Adapting the Smart Home to the User. *IEEE Trans. Syst. Man Cibern. J.* **2009**, *39*, 949–959.
31. Adlam, T.; Faulkner, R.; Orpwood, R.; Jones, K.; Macijauskiene, J.; Budraitiene, A. The installation and support of internationally distributed equipment for people with dementia. *IEEE Trans. Inf. Technol. Biomed.* **2004**, *8*, 253–257.
32. Bouchard, B.; Giroux, S.; Bouzouane, A. A Keyhole Plan Recognition Model for Alzheimer Patients: First results. *Appl. Artif. Intell.* **2007**, *21*, 623–658.
33. Mégret, R.; Dovgalecs, V.; Wannous, H.; Karaman, S.; Benois-Pineau, J.; El Khoury, E.; Pinquier, J.; Joly, P.; André-Obrecht, R.; Gäestel, R.; et al. The IMMED Project: Wearable Video Monitoring of People with Age Dementia. In Proceedings of the 18th ACM international conference on Multimedia, New York, NY, USA, 2010; pp. 1299–1302.
34. Adami, A.; Pavel, M.; Hayes, T.; Singer, C. Detection of movement in bed using unobtrusive load cell sensors. *IEEE Trans. Inf. Technol. Biomed.* **2010**, *14*, 481–490.
35. Tapia, E.; Intille, S.; Larson, K. Activity Recognition in the Home Using Simple and Ubiquitous Sensors. In Proceedings of the International Conference on Pervasive Computing, Linz and Vienna, Austria, 21–23 April 2004; Volume 3001, pp. 158–175.
36. LeBellego, G.; Noury, N.; Virone, G.; Mousseau, M.; Demongeot, J. A Model for the Measurement of Patient Activity in a Hospital Suite. *IEEE Trans. Inf. Technol. Biomed.* **2006**, *10*, 1.
37. Barnes, N.; Edwards, N.; Rose, D.; Garner, P. Lifestyle monitoring-technology for supported independence. *Comput. Control Eng.* **1998**, *9*, 169–174.
38. Tamura, T.; Kawarada, A.; Nambu, M.; Tsukada, A.; Sasaki, K.; Yamakoshi, K.I. E-healthcare at an experimental welfare techno house in Japan. *Open Med. Inform.* **2007**, *1*, 1–7.
39. Gong, Y.; Chung, Y.-A.; Glass, J. AST: Audio Spectrogram Transformer. arXiv 2021, arXiv:2104.01778.
40. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An ontology and human-labeled dataset for audio events. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 776–780.
41. L. Chen and M. Asgari, "Refining Automatic Speech Recognition System for Older Adults," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 7003-7007, doi: 10.1109/ICASSP39728.2021.9414207.
42. Barry, S.J., Dane, A.D., Morice, A.H. et al. The automatic recognition and counting of cough. *Cough* **2**, 8 (2006). <https://doi.org/10.1186/1745-9974-2-8>
43. Henry Brodaty, The practice and ethics of dementia care, *International Psychogeriatrics*, Volume 27, Issue 10, 2015,1579-1581, ISSN 1041-6102, <https://doi.org/10.1017/S1041610215001234>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.