

Article

Not peer-reviewed version

---

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

---

[Ilyas Potamitis](#) \*

Posted Date: 12 December 2025

doi: 10.20944/preprints202509.0861.v3

Keywords: AI; elderly care; dementia; LLM; ESP32



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

Department of Music Technology and Acoustics, Hellenic Mediterranean University, Heraklion 71410, Greece; potamitis@hmu.gr

## Abstract

Population aging is increasing dementia care demand. We present an audio-driven monitoring pipeline that operates either on mobile phones, microcontroller nodes, or 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 dialogs, and speech-emotion recognition. An audio classifier detects home-care-relevant events (cough, cane taps, thuds, knocks, and 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, and time), delusion themes, agitation events, health proxies, and safety flags (e.g., exit seeking and falling). The pipeline targets real-time 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:** AI; elderly care; dementia; LLM; ESP32

---

## 1. Introduction

Population aging 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 USD 1.3 trillion in 2019 and projected to rise sharply in coming decades—while informal caregivers already deliver roughly half of total care hours [1–4]. In Europe, older adults ( $\geq 65$  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  $\geq 65$  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 workers with limited or no formal training at all; 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 they 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, and 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, and 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 and 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, the 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” are suggested in [38] combined physiological and environmental sensors to enable continuous monitoring across multiple aspects of daily life, including personal hygiene and rest [39,40].

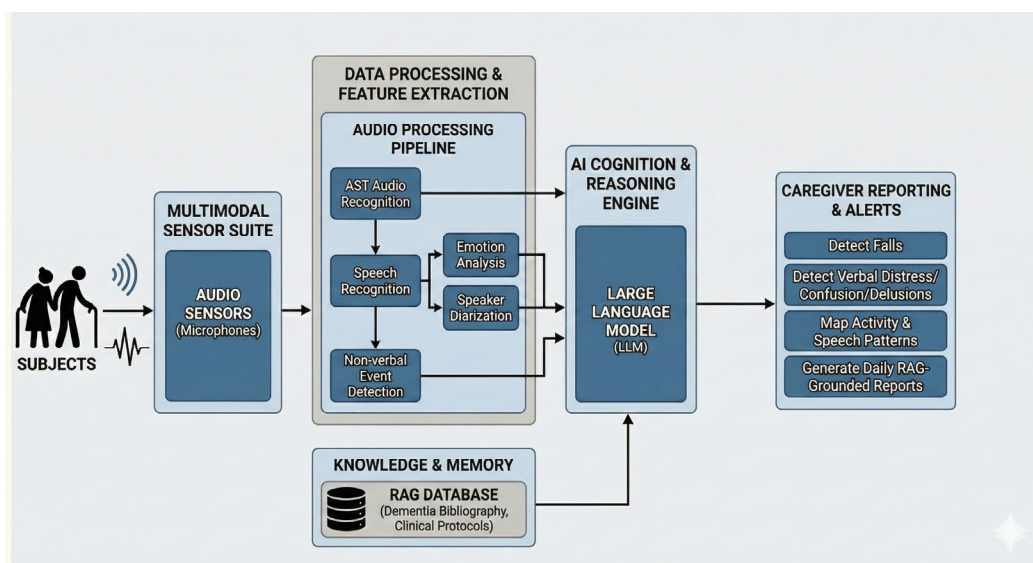
This work targets that translational gap by proposing and evaluating a home-based pipeline that couples audio signal processing with an Audio Spectrogram Transformer (AST) [41] trained on the Audioset database [42] for non-speech event detection and a LLM (ChatGPT-5.1) constrained to emit auditable, 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, dialog 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 [43–50], the outputs are designed to guide daily care (e.g., de-escalation prompts, and 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 multitask monitoring that reflects real-world caregiving priorities (safety, agitation, and communication efficacy). We use recently developed AI tools that were not available ten years ago (i.e., transformers, large databases, and LLMs) and introduce them to assisted care services. 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 falls and delusional themes. In aggregate, the combination of audio processing, 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.

This work integrates pretrained components and applies them in a zero-shot manner, with few-shot enrollment only for diarization. There is no training or finetuning on private data. The design reflects ethical and practical constraints that make large-scale end-to-end training on speech from older adults with dementia infeasible and difficult to scale (limited labeled data, privacy and consent requirements, participant burden, and annotation cost). Accordingly, we state explicitly the following: (i) all neural and transformer-based models remain frozen (i.e., AST, ASR, VAD, LLMs, and emotion recognition); (ii) emotion recognition specifically is zero-shot; and (iii) diarization uses few-shot enrollment without training from scratch or fine-tuning per se.

This study should be interpreted as a dense, longitudinal case study of two individuals rather than a generalizable clinical trial. Our conclusions are restricted to engineering feasibility and within-subject trend tracking. In Section 5.1 Future Prospects, we discuss multi-site studies with larger, diverse cohorts that will be required before any claims about diagnostic performance or broad deployment are justified.



**Figure 1.** Architecture of the audio-based dementia care system. The system utilizes an Audio Spectrogram Transformer (AST) to classify audio events. Verbal instances undergo parallel processing for emotion recognition, speaker diarization, and Speech Recognition (ASR). Both the processed verbal data and classified non-verbal events are fed into a Large Language Model (LLM). The LLM integrates this multi-modal information, applies reasoning to detect critical events (e.g., falls, delusions, pain), and generates structured reports and alerts for timely caregiver notification.

## 2. Materials and Methods

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

The subjects are attended by two caregivers in two 8 h 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.

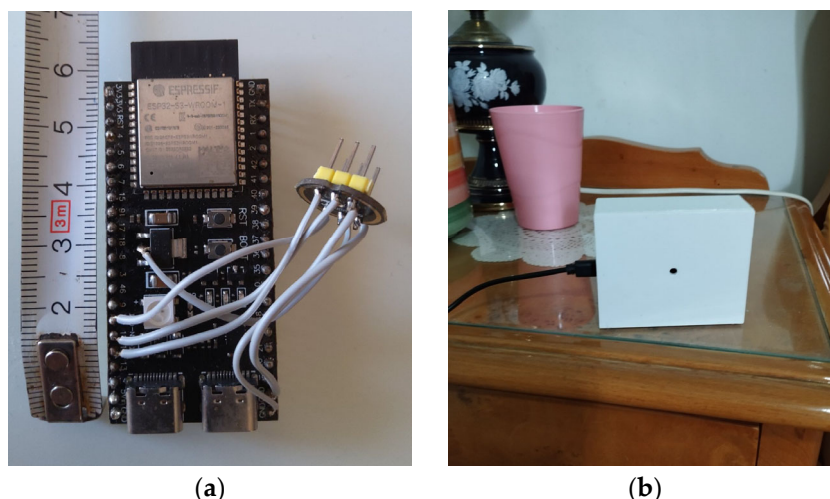
This study is a proof-of-concept demonstrating the feasibility of an audio processed by AI pipeline for home dementia care. Our goal is methodological and systems-oriented—to show that ambient audio can be captured on very low-cost hardware, transformed into text labels corresponding to the recognized audio events out of 527 audio classes, and summarized by an LLM (in this work ChatGPT 5.1 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 15 s snippets across months), enabling robust within-subject demonstrations of trend tracking (e.g., cough burden, exit-seeking themes, and 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 suitable AI-bots can provide significant value to this problem by first monitoring and, in the near future, engaging the task with actions. Accordingly, we perform the following: (i) report per-subject results and descriptive statistics rather than pooled significance tests; (ii) release code and disidentified audio to enable replication (see Appendix A); and (iii) outline a prospective, multi-site study with power considerations for clinical endpoints (e.g., sensitivity to agitation/exit-seeking and caregiver workload impact). This sequencing—proof-of-concept → pilot → powered trial—is standard for translational systems entering safety-critical care.

### 2.1. 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 (Espressif Systems, Shanghai, China) with SPH0645LM4H digital microphones (Syntiant Corp., Irvine, CA, USA) transmits MP3 compressed 15 s audio snippets over Wi-Fi to a secure remote server for classification (see Figure 2).



**Figure 2.** Hardware and deployment of the audio monitoring node. (a) ESP32-S3 low-power audio node. The node integrates an ESP32-S3 microcontroller with a microphone attachment. The device operates in a continuous, audio-triggered state. Upon activation, it captures a 15 s audio segment at a 16 kHz sampling rate, compresses the data into the MP3 format, and streams the packet over Wi-Fi to a local server for subsequent processing. (b) Bedside Deployment. A typical deployment location showing the device positioned on the bedside table. This placement ensures optimal proximity for continuously monitoring both sleep-related events (e.g., coughing and breathing), speech, paralinguistic features and general audio (cane hits, thump/thud/hits, and glass breaking).

The SPH0645LM4H provides low-power, bottom-port I<sup>2</sup>S 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 I<sup>2</sup>S, was selected to handle audio capture and MQTT. The total cost of the device is at the order of EUR 20 (as per 14 August 2025). A more detailed description of the hardware can be found in [22]. However, the same service can be provided by a spare smartphone at zero extra cost that streams audio snippets to Wi-Fi, captured by a voice activity detector (VAD).

## 2.2. Embedded Software in ESP32

Using Espressif IoT Development Framework version 5.3 (ESP-IDF 5.3, Espressif Systems, Shanghai, China), the embedded nodes in C language capture audio from the microphone via the Inter-IC Sound (I<sup>2</sup>S) 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 (The LAME Project, San Diego, US), 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 to keep communications' cost low.

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, I<sup>2</sup>S 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.

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 10 s spectrogram patches for audio event classification. Spectrograms for spoken snippets and environmental events (e.g., discussions, cough, and cane taps) are processed to produce labeled, timestamped outputs suitable for downstream integration.

### 2.3. *The Elders with Dementia Dataset*

The ESP devices work 24/7, and depending on the activity inside home, several hundred recordings can be produced per day. The recording sessions started on 4 August 2025 and continue up to now, reaching several thousands of recordings. Recordings contain mainly speech, non-verbal human sounds, and impact sounds of various objects (e.g., drawers, closets, doors, switches, mobility aids). The database they produced is the subject of this study.

### 2.4. *The Audio Event Recognizer*

The Audio Spectrogram Transformer (AST) [41] 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 filter banks using a 25 ms window with 10 ms hop. The 2-D time–frequency map is then split into overlapping  $16 \times 16$  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 (mix-up, SpecAugment) and weight-averaged checkpoints. Evaluated on AudioSet—the 2.1 M-clip, 5.8 k h, 527-class collection of human-labeled 10 s YouTube excerpts—AST reports mean average precision (mAP) 0.4590 for a single weight-averaged model and up to 0.4850 with model assembling 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 [42] is a large-scale, ontology-driven dataset of human-labeled sound events released by Google, comprising over 2 million 10 s 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, and 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, and 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 which appear under respiratory sounds in the ontology as well. Regarding mobility and movement (activity/restlessness cues), the labels are as follows: Footsteps, Bouncing, Wood, Stomp, Stamp, Surface Contact (e.g., scrape and scratch), Creak (e.g., floorboards/chairs), and 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, and Telephone Ringing. Conversation and caregiver interaction (dialog state cues) are handled by the Speech, Conversation, Narration, Monolog, Whispering, and loudness-related (Shout/Yell) categories.

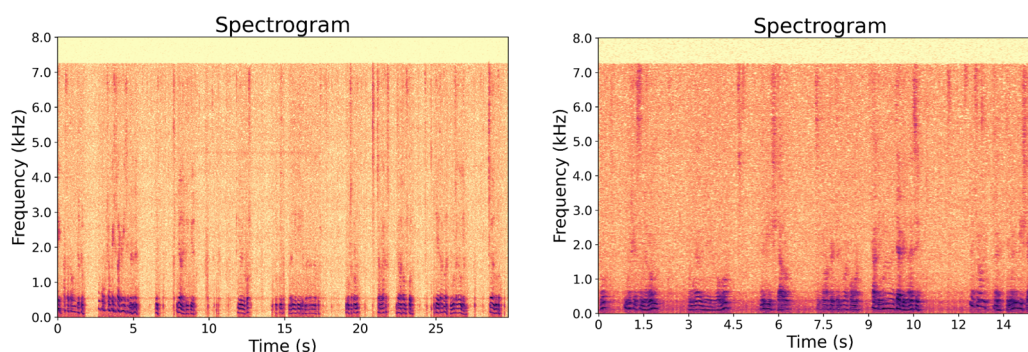
The wide representation of audio categories in Audioset ontology practically covers all cases appearing in the context of our work (see Appendix A.1).

### 2.5. 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 that has demonstrated better results compared with other state-of-the-art algorithms (see Appendix A.2). 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.

### 2.6. Ordinary Automatic Speech Recognition Models Do Not Currently Work Properly for Elders 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 A.3) underperformed on conversational Greek from older adults with dementia because multiple, interacting factors degrade the signal and violate the model’s assumptions. We need to note that transcription accuracy varies among elder individuals and the various states of dementia. While baseline testing on healthy elderly subjects (aged 60–80+) showed high transcription fidelity that match the rates reported for Whisper v3 for the Greek language (see <https://github.com/openai/whisper> accessed on 9/12/2025), the ASR accuracy for the specific couple in this study is extremely low, with the Word Error Rate (WER) consistently exceeding 90% and the Sentence Error Rate (SER) approaching saturation 99% accuracy loss. 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 3)—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 crosstalk 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 dialog. Fifth, pipeline constraints compound errors: tight voice-activity detection and diarization cuts fragment utterances and 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 of transformers 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 [51] for an adaptation of ASR models to elders using a 12 h adaptation corpus) and [52–54] for specifically adapting to dysarthric speech.



**Figure 3.** Characteristic spectrograms of speech from elders with dementia.  $y$ -axis the frequencies of speech and  $x$ -axis time in seconds. **(Left)** An example of female elder speech (Event\_id: 186692). **(Right)** Male elder (Event\_id: 172738). Both speakers were 1–1.5 m away from the microphone. Notice the fading of frequencies above 600 Hz and the prolonged pauses between vocalizations.

As shown in Figure 3, log-spectrograms are time-frequency representations of audio signals widely used in speech processing and recognition. A spectrogram is produced by first calculating the Short-Time Fourier Transform (STFT) of a signal, which yields the signal’s energy distribution across various frequencies over short, overlapping time segments. The resulting magnitude plot is then processed by applying logarithmic compression to the values. This step is critical and conforms to the way the human ear perceives loudness. Log-spectrograms effectively compress the high-energy components (such as strong speech vowels) while simultaneously amplifying the subtle, information-rich low-energy components (such as fainter fricatives or consonants). This conversion transforms the 1D audio signal into a 2D visual representation that acts as a “picture” input for machine learning models. This visibility and 2D format make log-spectrograms an ideal front-end for models like Transformers, leading to robust performance in tasks such as Automatic Speech Recognition (ASR), speaker diarization, and emotion identification.

### 2.7. 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 s 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 participants, two caregivers, and 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 A.4). In version 2.1, the default pipeline applies a neural speaker segmentation model on short sliding windows ( $\approx 5$  s with 0.5 s hop), producing frame-level ( $\approx 16$  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 since 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.

### 2.8. 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, even singing. 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 sundown. Apathy, a low-arousal negative state distinct from simple neutrality, can be pervasive, with minimal speech initiation and muted effect.

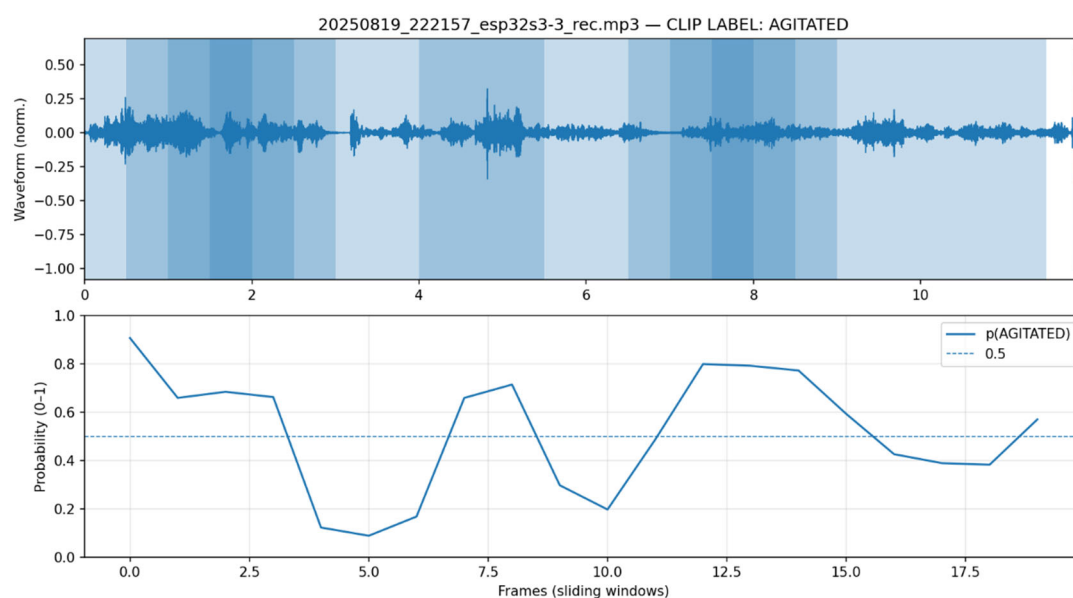
In this study, emotion recognition algorithms are based on signal processing audio only and extract low-level acoustic descriptors (pitch, loudness, speech rate, and rhythm, etc.) and then use statistical models to classify into emotions (e.g., neutral, happy, sad, angry, and afraid). Since they rely 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 (see Appendix A.5). The base encoder is `hubert-large-ll60k`, a self-supervised model trained on approximately 60,000 h 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. We did not fine-tune our data.

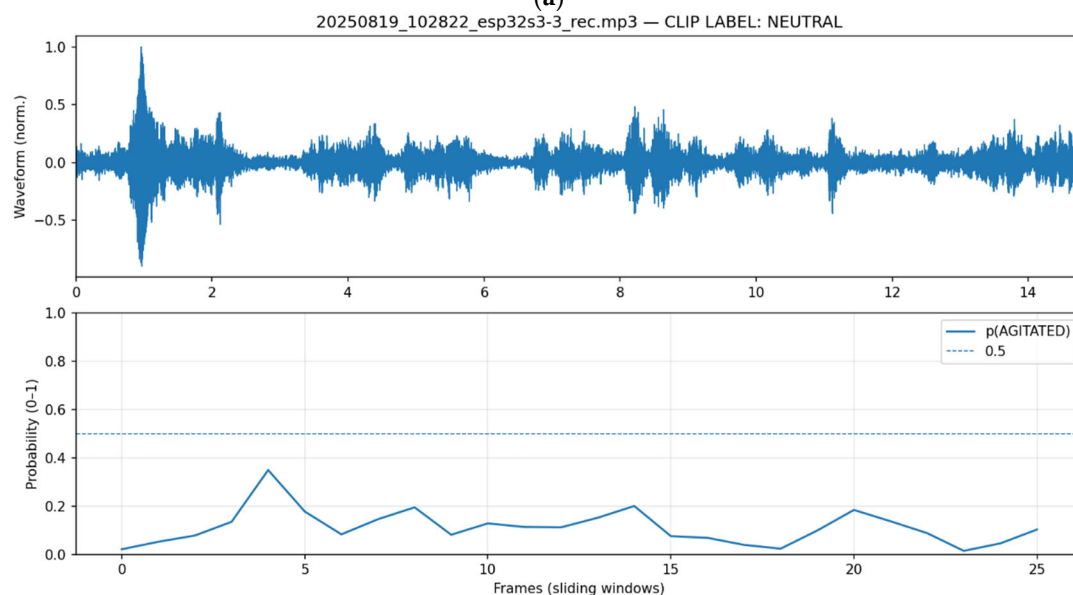
In the SUPERB benchmark, the ESR task is framed as a four-class utterance-level classification with the categories angry, happy, sad, and neutral. The widely used setup draws on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database and follows the common practice of

focusing on these four balanced classes (other minority labels are omitted). The superb/hubert-large-superb-er checkpoint is fine-tuned to predict the label set—angry, happy, sad, and neutral.

Old adults with dementia do not demonstrate the expressivity of the young. Therefore, we merge the initial labels to improve robustness in this task. 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. Regarding the decision procedure, to handle long, real-world recordings, we apply sliding-window inference. Each input file is segmented into fixed windows (default 2.0 s) with overlap (0.5 s hop). For every window, the model produces a probability distribution over the four emotions. 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 4a), otherwise, it is NEUTRAL (see Figure 4b). 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). The 0.5 frame-level probability threshold and the “ $\geq 20\%$  agitated windows” clip criterion were chosen to favor sensitivity based on pilot inspection, not tuned on a held-out set.



(a)



(b)

**Figure 4.** Figures (a,b) depict two recordings that are analyzed to assess ‘agitation’ vs. ‘neutral’ emotional state. The top subfigure is the oscillogram and the bottom subfigure an agitation metric. (a) ESR predicts an emotional state for each speech frame because agitation sometimes manifests itself in bursts. (b) A calm conversation.

### 2.9. Privacy, Data Handling, and Ethics

The ESP32/S3 nodes are triggering autonomously and produce short snippets (fixed 15 s windows). They are designed not to include an SD shield, and their memory suffices for a single recording, therefore audio is buffered only for encoding and transmission and is not written to persistent storage. Continuous raw audio is not streamed.

The server either in the home or facility scenarios can be held on premises so nothing is transmitted outside, including audio visual content. On the server, raw MP3 is decoded to a spectrogram and discarded after feature extraction. We retain only timestamped, de-identified event metadata (AST labels, RMS, diarization tags, agitation flags) and, for the small manually transcribed subset, text without names or identifiers.

Data transport is encrypted, and all MQTT traffic is protected by TLS. Moreover, the experiments in this paper used an on-premises broker and server on the same local network. Participants are referred to only by anonymous IDs in all stored data. Access to the server and logs is restricted to the research team. Caregivers can see audiovisual data but can acquire only reports. We clarify that event logs used in this study are retained for 12 months for research analysis and then will be deleted or irreversibly anonymized. There will be no permanent archive of identifiable audio that can be uploaded to third parties.

## 3. Results

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

### 3.1. Detecting Cough and Other Non-Verbal Audio Events

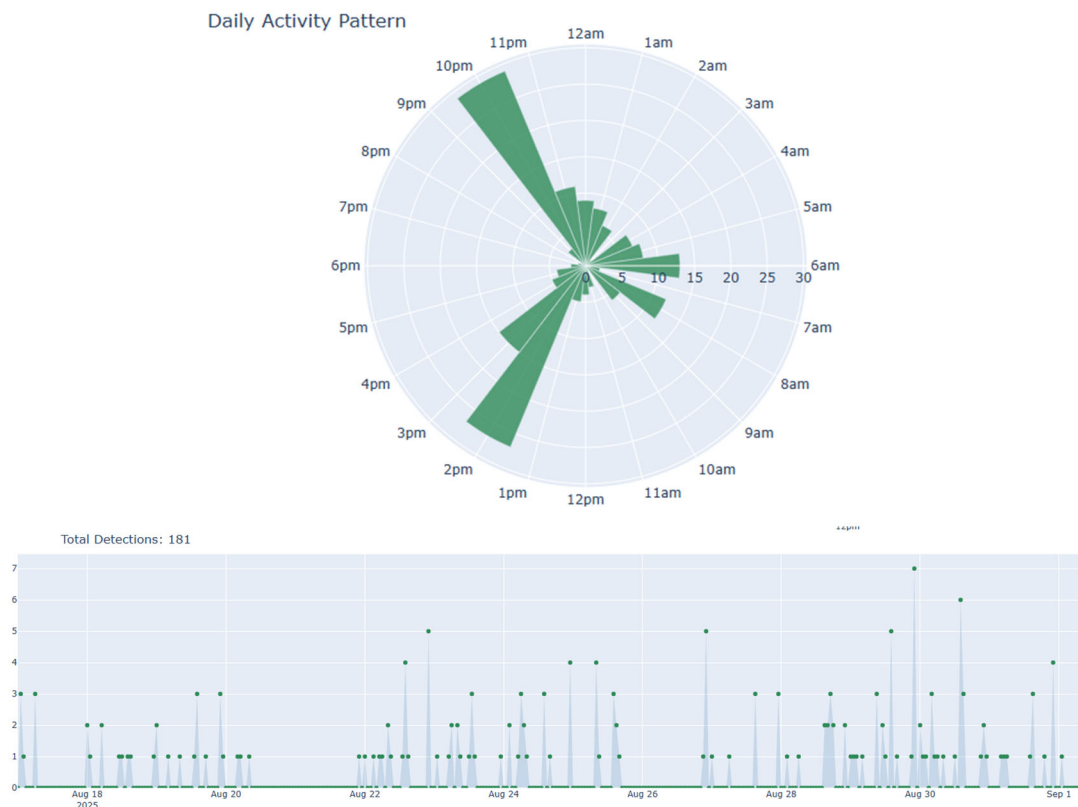
Cough events are counted according to the AST classification label (see Figure 5). In this dashboard we illustrate that, for 15 s snippets, the preprocessing module detects the presence of cough, enumerates cough bursts, and estimates cough intensity via the root-mean-square (RMS) amplitude of the waveform. Even in the context of a chronic cough, longitudinal variation in frequency and acoustic character is informative [55]: sudden increases may reflect intercurrent irritation or infection, and changes in sound quality (e.g., wetter/drier timbre or superimposed wheeze) may indicate shifts in respiratory status, and temporal alignment with dialog quantifies communication impact (interruptions and 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 5 allows filters to be applied to audio classes that queue and group the audio events of interest.

To operationalize these signals, we compute personalized baselines (e.g., coughs per hour) and multi-day deviation metrics. In Figure 6-top, we see the hours that the cough takes place over a configurable time span. Figure 6-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.

2025/08/05 → 2025/08/14

Event ID	Timestamp	Device	Class 1	Prob Class 1	Class 2	Prob Class 2	Class 3	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 5.** Audio events are classified using the AST, and they are gathered in a dashboard that has the role of a situation assessment board. The dashboard depicts from left to right columns: the ID of the audio event, its timestamp, the device node that originated from, the three higher-ranked class labels of the event with their corresponding probabilities, and the RMS (dB) of the event. Once events are cataloged, one can ask the system to report the audio incidents of ‘cough’, ‘sneeze’, and ‘throat cleaning’ of an elderly couple in an in-house setting and perform various statistics and time-series analysis spanning days to weeks.



**Figure 6. (Top)** Pie chart of hourly activity of coughing from 17 August 2025–1 September 2025 and, **(Bottom)** time series visualization that assesses rate and trend in cough incidents and relates it to pre- and after treatment.

$x$ -axis denotes time and  $y$ -axis number of incidents for the target label. Note that the audio category is configurable, and we can visualize any of the 527 audio categories of the AST ontology.

The Audio Spectrogram Transformer (AST) demonstrated exceptional reliability in differentiating cough events from domestic background noise. To evaluate the model's performance in a real-world setting, we implemented a reliable sampling protocol based on continuous "in-the-wild" recordings from the deployed ESP32 nodes.

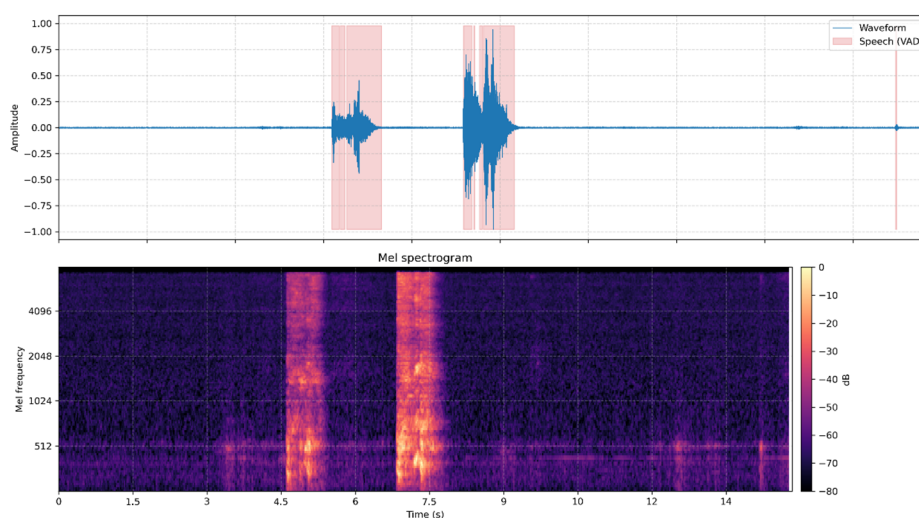
**Sampling Methodology:** Ground truth was established through manual annotation of the audio snippets. We curated a balanced validation set comprising 100 confirmed positive instances of respiratory events (merging 'cough', 'sneeze', and 'throat clearing' due to their similar acoustic signatures and clinical relevance) and 400 confirmed negative instances. The negative set included a diverse array of "hard negatives" typical of a home environment, such as speech, television audio, door slams, and silence, to rigorously test the model's false-positive rejection rate.

To accommodate the polyphonic nature of real-world home environments, where target sounds often overlap with background noise, we adopted a "Top-3" classification criterion. An audio segment was classified as a True Positive (TP) if any of the target labels—'cough', 'sneeze', or 'throat clearing'—appeared within the AST's top three predicted classes. Conversely, a false positive (FP) was recorded if any of these labels appeared within the AST's top three predicted classes for an audio segment containing only hard negative events. This threshold was selected to ensure robust detection even in multi-source scenarios where the target event might be identified as a secondary prediction rather than the primary output.

Under these criteria, the AST classifier achieved perfect separation between classes across the 500 sample validation set, effectively filtering out all non-respiratory acoustic interference. The system demonstrated perfect accuracy with the following performance metrics: accuracy, 1.0; precision, 1.0; recall (sensitivity), 1.0; and F1-score, 1.0.

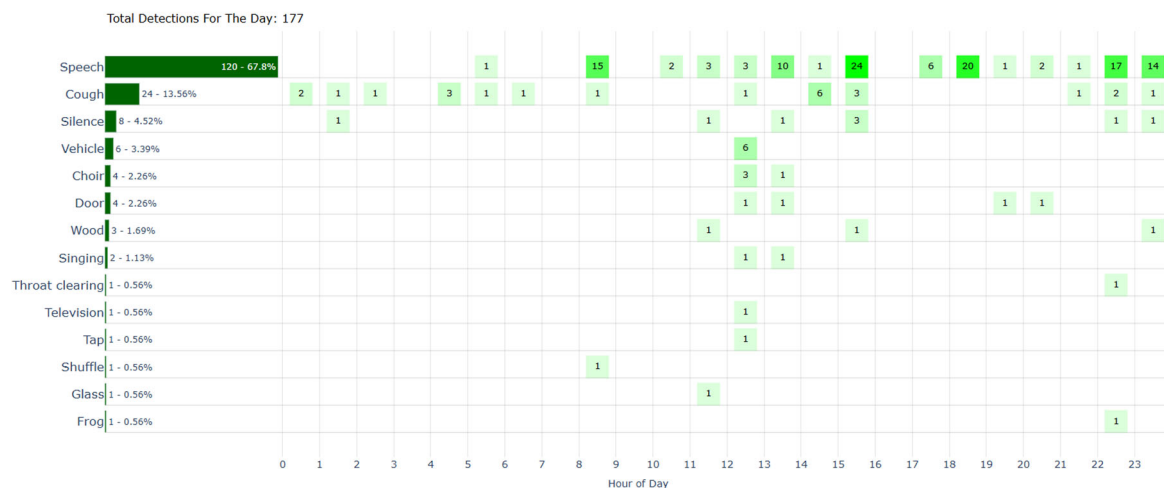
These results confirm the AST's capability to serve as a highly reliable pre-filtering trigger, ensuring that the computationally intensive downstream LLM analysis is only initiated by verified health-related acoustic events.

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 7). Note that the audio category is configurable, and we can visualize any of the 527 audio categories of the AST ontology.



**Figure 7.** 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.

Once events are recognized and logged, visualizations can support caregiver interpretation. Figure 8 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 ( $\approx 68\%$  of 177 detections). Notably, there are extended conversations after the 22:30 bedtime.



**Figure 8.** Automated audio event classification and hourly distribution of 177 audio events 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 hourly occurrence in a single day.

In Figure 8, aside from intermittent coughs during the night, sleep appears relatively uninterrupted. Cough events constitute  $\sim 14\%$  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,” and “wood”) are rare and isolated, suggesting minimal cane striking. No sounds indicative of a fall, pain, or distress were detected on this day.

### 3.2. Diarization

Diarization has been notably successful, even in a few-shot learning scheme. We extracted 20 recordings, 15 s each, for every person in the home environment to enroll speakers in the diarization application. 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, indicating 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, with no missing or extra labels). Hamming loss measures the average fraction of misclassified label decisions over all instance–label pairs and 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) pool counts globally before computing precision, recall, and F1 (favoring frequent labels). F1 (macro) averages per-label F1 scores (treating each label equally and highlighting performance on rare labels). In Table 1, 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

**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

**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

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 2, we gather the results on a per-speaker basis.

In multi-class or multi-label evaluation, microaverage 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 4.** Metrics of the emotion recognition task. There are two classes: 'agitated' and 'neutral'.

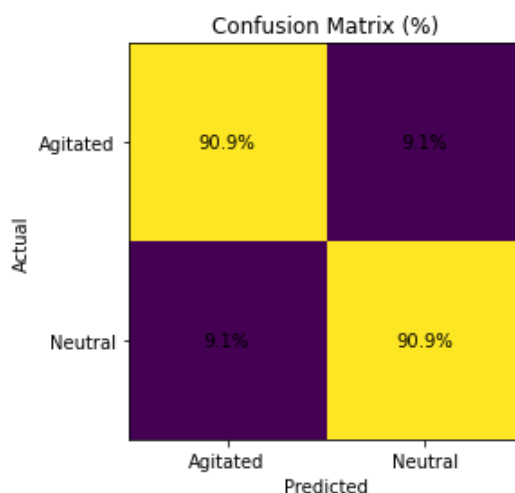
Class	Precision	Recall	F1
Neutral	0.91	0.91	0.91
Agitated	0.91	0.91	0.91
Accuracy			0.91
Macro avg	0.91	0.91	0.91

Weighted avg	0.91	0.91	0.91
--------------	------	------	------

### 3.3. Emotion Recognition Results

Caregivers and the first author have pulled together a corpus of 22 recordings in agitated and 22 in neutral state. They are based on observed behaviors cross-checked by the cameras and voice characteristics. The 22 Agitated and 22 Neutral clips were selected by querying the longitudinal data for episodes and confirming the label by listening. These class labels are not on formal scales such as Cohen–Mansfield Agitation Inventory (CMAI) or Neuropsychiatric Inventory (NPI).

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 9 we show the confusion matrix of this task. The metrics are the same as in the diarization case. The reported 91% accuracy should be interpreted as agreement with caregiver-level labels in this specific household, not as a validated clinical biomarker. In Supplementary Materials, we offer a sample of agitated/neutral recordings for the reader.



**Figure 9.** Confusion matrix of emotion recognition between the ‘agitated’ and ‘neutral’ class for the emotion recognition module.

### 3.4. Safety and 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 are presented in detail in Section 4. 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 to trigger immediate alerts to relatives and caregivers. The

program at the server can subsequently send push notifications via e-mail and short messages to relatives and caregivers for a probable emergency.

### 3.5. Caregiver Support and 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 3.11 program time-stamps the event and the AST registers the first three higher in rank classes and their corresponding probabilities and the RMS intensity of the audio event. The LLM 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 A.6. In Table 5 we give a short extract.

**Table 5.** The LLM prepares a report summarizing key points of health-relevant notes for supporting the caregiver based on a daily file of recognized and timestamped 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.

<p>Health-relevant signals (audio only)</p> <ul style="list-style-type: none"> <li>• 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.</li> <li>• Breath/vocal strain (AM). One gasp/breathing snippet at 08:47 without follow-up distress markers; continue to watch for clustering.</li> <li>• 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.</li> </ul> <p>Engagement &amp; orientation</p> <ul style="list-style-type: none"> <li>• 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.</li> <li>• Night-time/bedtime not confirmed: last record 18:14; no data later in the evening to verify final bathroom/bedroom transition.</li> </ul>
--

The deployment of LLMs in interactions with vulnerable populations necessitates robust mitigation strategies against hallucinations and departures from predefined safety guardrails. In the following section, we develop a technique where LLM interactions are augmented with domain-certified knowledge extracts (e.g., from validated professional texts or medical guidelines). This process effectively grounds the LLM's responses within established safety and care protocols.

### 3.6. Retrieval-Augmented Generation

Large language models (LLMs) generate continuations from task context and prompts. In extended dialog interactions, initial constraints (e.g., 'don't suggest medicine or outdoor activities') may drift in time and lead to suboptimal or impermissible suggestions. To reduce this risk, we employ retrieval-augmented generation (RAG) to promote factuality, personalization, and auditability. RAG pairs a generator with a retrieval step that, at query time, selects relevant documents or records, and the model is instructed to ground its output in these sources. This design reduces unsupported inferences, incorporates patient and home context (e.g., logs, protocols and prior episodes), and enables citation-backed outputs for care summaries, alerts, and decision support.

In this proof of concept, we indexed a source document on dementia care [56]. In operational deployments the corpus can be augmented to include entire books, guides, and manuals. The system ingests daily Audio Spectrogram Transformer event tags in JSON form, caregiver guidance PDF files, and a small household history, and then executes an indexing–retrieval–generation pipeline to produce shift-level reports. During indexing, text is segmented, embedded, and stored in a vector database for semantic search and retrieval supplies the most relevant passages and generation constrains outputs to the retrieved evidence. The code inspects the JSON file for what audio events happened in a day (cough, thuds, speech, and audio related to cane, etc.). It then builds a dynamic retrieval query from the books on treating patients with dementia concerning these events (on top of your base query). Finally, it builds the final prompt that is directed to the LLMs. The LLM is instructed to create a caregiver report with the following sections: 1. summary of the day; 2. behavioral interpretation; 3. caregiver action items (specific; practical); 4. environment/routine adjustments; 5. safety flags; 6. what to monitor tomorrow; and 7. evidence used (cite pages from corpora). In Appendix A, the code link describes an implementation using ChromaDB. Table 6 presents a report with page-level references, and the reader is encouraged to contrasted it with Table 5. Since the JSON is file is different every day the daily report to caregivers remains up to date. Hereinafter, we give a more analytic view of what happens during RAG which is a query targeting communication strategies, safety, and routine management is embedded and used to select the most relevant passages from the indexed corpus. During generation, the system integrates three inputs—retrieved guidance text, a structured summary of daily audio event annotations with timestamps and confidence scores, and a fixed household profile describing history and constraints of the care recipients—to produce a shift-level report for caregivers. The report synthesizes observed behaviors with guideline-concordant responses, highlights potential risks, proposes environment or routine adjustments, and specifies near-term monitoring items. The design emphasizes transparent evidence use by passing the retrieved excerpts and their page references to the language model, enabling traceable outputs without modifying model parameters.

**Table 6.** A RAG based system prepares a report after vectorizing books, manuals and directives for best treatment of patients with dementia. Note that this report changes daily based on the JSON file that includes the audio recognition events of the previous day.

Daily updated caregiver report (RAG version)

(1) SUMMARY OF THE DAY:

The couple had a relatively active day with 82 speech events, 37 coughs, and 17 sneezes. Distress-like events were detected around 15 h, 05 h, and 04 h.

(2) BEHAVIORAL INTERPRETATION:

The high number of speech events suggests they are communicating, which is positive. However, the distress-like events, especially during the early hours, could indicate sleep disturbances, a common issue in dementia patients. The frequent coughing and sneezing may suggest a health issue, possibly a cold or respiratory infection.

(3) CAREGIVER ACTION ITEMS:

- Monitor their health closely, especially for signs of a cold or respiratory infection.
- Ensure they are keeping hydrated, particularly if they are unwell.
- Encourage them to communicate and engage in conversations.
- Try to maintain a calm and quiet environment, especially during the night to promote better sleep.

(4) ENVIRONMENT/ROUTINE ADJUSTMENTS:

- Introduce relaxing activities in the late afternoon or early evening to help reduce sundowning.

- Ensure their bedroom is quiet, dark, and cool, with a clock showing whether it's day or night next to the bed.
  - Limit their screen time towards the end of the day and ensure they get plenty of daylight.
- (5) SAFETY FLAGS:
- The distress-like events during the early hours could indicate they are getting up during the night, which could increase the risk of falls.
  - The frequent coughing and sneezing could indicate a health issue that needs medical attention.
- (6) WHAT TO MONITOR TOMORROW:
- Monitor their sleep patterns and any signs of distress during the night.
  - Keep an eye on their health, particularly for any signs of a cold or respiratory infection.
  - Observe their communication and social engagement.
- (7) EVIDENCE USED:
- Guidance from "Caring for a person with dementia: A practical guide" (pages: 43, 63, 75, 79, 139, 140) was used to interpret the data and suggest practical actions.

### 3.7. Delusion Detection

Detecting delusions and/or hallucinations or even suicidal ideation [57] is valuable information for caregiving, 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 is not there). *Example:* the person says "the kids are sitting on 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 than that in Table 6, considering these issues, can be asked if we feed the LLM with the transcribed speech of the subjects and features of their emotional state. 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 may develop into a hallucination is the one where another 'lady' is inside the house or addressing her husband as her father sporadically. 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 the Appendix A.6, we have such a report, an extract of which we give in Table 7 for clarity. Note that the system currently does *not* perform automatic delusion detection in real time due to restrictions on ASR reliability. However, if patients do not have dysarthria and a small, but concise history of the patient is given, then it is feasible.

We set the filters to sort speech incidents on the dashboard described in Figure 5, and we listened to discussions, sampling 50 dialogs to include 10 episodes that caregivers and the author had flagged as containing delusional content and 40 routine conversations. Manual Greek transcripts were produced and used as ground truth for the LLM experiment. The 10 "delusional" cases were selected from periods where caregivers or family members documented clear delusional content (place misidentification, phantom outdoors activities, etc.), and the 40 "non-delusional" dialogs were randomly sampled from routine conversations without such flags.

Regarding the annotation protocol, for each dialog, manual Greek transcripts were produced with diarization tags ([M]/[F]) and agitation labels ([A]/[N]). The author then assigned a binary label "Delusional" vs. "Non-delusional" based on the definition of delusion (fixed false belief held despite contrary evidence). We describe the decision rules (e.g., confusion or memory gaps without a fixed

false belief are not labeled as delusion). The system discovers delusional schemes by cross-referencing the transcribed narrative against the short family's background to locate factual contradictions. The corpus is small, household-specific, and enriched for clear-cut cases. The detection rate should be interpreted as an existence proof of reasoning ability under idealized conditions as regards ASR, not as an estimate of sensitivity/specificity.

**Table 7.** Prompting the LLM with (a) small history of the couple (Appendix A.6), (b) the context of the application (prompt), (c) transcribed speech, diarization (male or female), emotion tagging (neutral or agitated). Diarization and emotion tags are in brackets.

<p>PROMPT</p> <p>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 nursing 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. &lt;Small history of the couple&gt;</p> <p>TRANSCRIBED DIALOGUES WITH AUTOMATED DIARIZATION AND EMOTIONAL TAGGING</p> <ol style="list-style-type: none"> <li>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!</li> <li>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</li> <li>3. [M][N] Did we eat today? [F][N] I don't remember</li> </ol> <p>RESULTS</p> <p>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).</p> <p>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).</p> <p>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).</p>
---

## 4. Discussion

### 4.1. Practical Achievements So Far

Our proposal is an affordable plug-and-play device whose data is classified at multiple levels, with results jointly interpreted by an LLM. Caregiving is typically an arrangement under budget constraints, and the configuration studied here—without loudspeakers or actuators—targets practical gains from passive audio monitoring. Hereinafter, we gather practical gains after using the

system for almost five months continuously: (a) Our approach for overnight surveillance lowers staffing needs for night shifts, which are costly and often inefficient. The system can report the accident (usually a fall) in real time to guardians, reducing the response time and relieving the pain and agony of the elderly that are typically unable to rise. Continuous fall monitoring at night also reduces family anxiety, given the difficulty of elders calling for help after a fall. The system successfully detected a real fall at night and recognized it as such using only audio (see Supplementary Materials). (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., cold bedsheets, arrangement of furniture, room temperature, and lighting conditions) that degrade wellbeing of dementia patients and are commonly not reported the following day due to memory impairment.

We placed particular value on the system's uninterrupted (24/7) reliability, which enables continuous monitoring and provides added support during periods of staffing constraints or caregiver absence, thereby easing the burden on adult children with legal guardianship responsibilities. Our approach is designed to offer a remote yet direct connection between guardians and patients without requiring the patients to use a phone or computer application to send or receive messages. Guardians can quickly review the situation (see Figures 5–7), listen to specific events of interest by clicking on them on the dashboard, and—depending on the individual's cognitive state and acceptance—will soon be able to respond via the device's Wi-Fi loudspeaker to reassure, remind, or gently redirect (duplex communication). This design acknowledges that many older adults do not respond well to handling gadgets: even reaching for a mobile phone, pressing the correct button, or navigating a laptop or tablet application can be difficult or unreliable for them depending on age, mobility, and mental condition.

#### 4.2. Limitations and Generalizability

In interpreting our findings, it is important to recognize the limits of the cohort and analysis. The study focuses on a single couple with stage 2 dementia, both Greek-speaking, living in a specific home environment, and is therefore not representative of the broader dementia population in terms of language, cultural background, disease severity, or living arrangements. All quantitative results are reported at the individual level and are intended as descriptive characterizations of system behavior and within subject patterns rather than as estimates of general performance. That is, the emphasis is on the engineering part. Consequently, we do not make claims about sensitivity, specificity, or other diagnostic metrics at population scale, and we view the present work as a feasibility oriented case study that motivates, rather than replaces, larger and more diverse evaluations.

## 5. Future Prospects

### 5.1. 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 dialogs), enabling oversight across an entire ward. Bedsides evolve into conversational endpoints where the LLM engages residents and staff through brief, goal-directed dialogs 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 equipped with audio and loudspeaker functionalities are viable, and in the Supplementary Materials, we include as an example of what is currently feasible with an analysis of a video depicting real incidents of falls using an audiovisual LLM. However, the audio-only modality alone is more discreet, less invasive, lower cost, and requires no infrastructure changes. In-house deployments and large facilities, processing and inference can run on an on-premises server for security and to adhere to ethical normative.

In the envisioned deployment, 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, physical exertion, and empathy that are far from what a machine can yet provide. The next step is a planned multi-site pilot with a larger, more heterogeneous cohort (different dementia stages, living arrangements, and languages) and pre-registered endpoints (e.g., sensitivity to agitation/exit-seeking and caregiver workload).

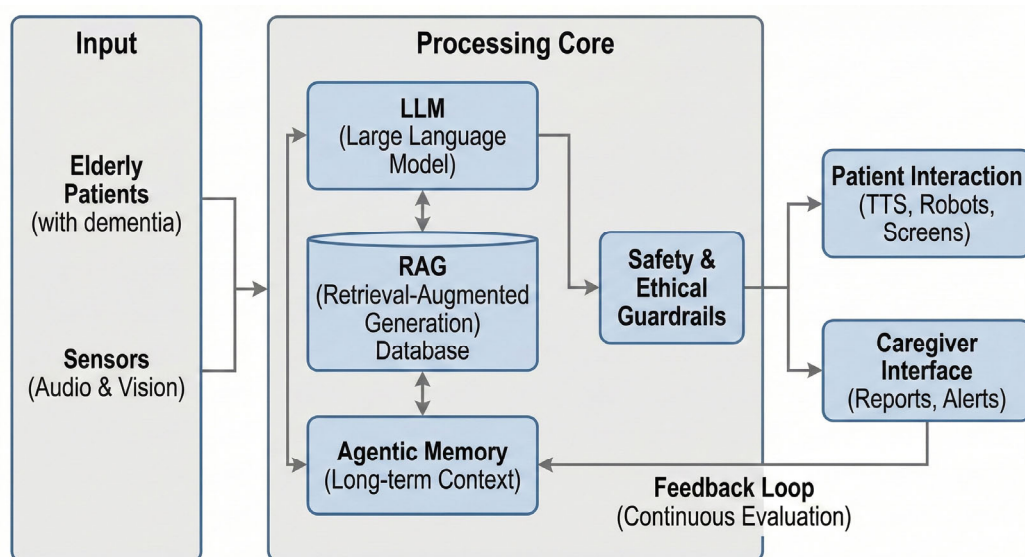
### 5.2. Conversation and Cognitive Support Using Actuators

In the present state, we keep a human in the loop as an indispensable part (see Section 4.1). However, we envision that an agentic service will finally autonomously engage in conversations (AI-bots/agents) that stimulate participants’ memory, monitor hazardous situations, and mitigate the sense of loneliness. This task is currently technically feasible, especially with the advent of large audiovisual models (e.g., of the class Qwen3-VL) and voice agents in online decoding mode with emotional speech, barge-in capability, and duplex communication. Our aim is to provide this service under the unconditional restriction of financial affordability using autonomous Wi-Fi audio recorders, spare smartphones, and affordable Wi-Fi cameras with audio.

In dementia care, an effective agent must integrate six capabilities within explicit safety and ethics guardrails. First, reasoning must be bounded and evidence-driven: the agent fuses multimodal inputs (e.g., short audio snippets, camera feed, ASR key phrases, affect tags, time of day, and a consented personal profile) to draw low-risk inferences and select pre-approved utterances rather than generate open-ended advice to vulnerable people. Second, acting is constrained and assistive, not clinical: the agent performs digital actions—brief TTS prompts via Wi-Fi loudspeakers, on-screen cues, and caregiver/guardian notifications. This would be useful, especially in the case of the important category of an alone elder that will otherwise not speak a lot. It will operate in supervised autonomy and will generate speech only from curated templates. Third, observing relies on privacy-preserving sensing (on-device preprocessing where feasible, minimal metadata upstream) to maintain situational awareness while protecting dignity. Fourth, planning converts near-term goals (hydration, toileting, calm bedtime, and orientation) into simple, time-appropriate steps with escalation paths to humans when uncertainty or risk is detected. Upon detecting agitation or

perseverative questioning—either through upstream classifiers or conversational cues—the agent employs gentle redirection strategies to achieve de-escalation. Fifth, it treats caregivers as partners, exposing transparent logs, rationales, and easy overrides so human judgment remains central. Finally, self-refining allows personalization via supervised updates based on psychological profiling of the users. The system self-adapts speaking rate, prosody, and linguistic complexity to the user’s current cognitive state that is unobtrusively assessed continuously and maintains a consented personal knowledge database to tailor references and improve engagement. This implies that the speech recognizer is already adapted for elderly speech [52–54]. The broad availability of LLMs, ASR, and TTS across most major languages inherently supports the global scalability. The agent can manage telecommunications by autonomously handling incoming calls and initiating contact with guardians or emergency services when necessary. Modern TTS technologies enable the synthesis of expressive, affect-aware speech, enhancing the naturalness of these interactions. While challenges to fully autonomous operation remain, we maintain that these are technically tractable through robust architectural safeguards. Key risks to be mitigated include hallucinatory guidance (highly reduced with RAG), ASR bias against elderly speech patterns (reduced through adaptation), the propagation of false positives from upstream classifiers, and the necessity of enforcing strict topic boundaries (e.g., preventing the LLM from offering medical diagnoses or medication advice).

The system we propose (see Figure 10) differs fundamentally from existing voice assistants such as Alexa/Siri/Cortana, which rely solely on speech-based interaction and reactive responses. In contrast, our system will be proactive, i.e., it will take initiative in starting and maintaining a conversation based on audiovisual feed, continuously integrating audio and visual sensing to monitor older adults for safety-critical events such as falls, choking, or potentially hazardous events such as unattended cooking and handling of objects, etc. This multimodal, context-aware, and memory-preserving framework shifts the paradigm from passive communication to active, personalized care monitoring. In this work we focus only on audio, but in the Supplementary Materials, we offer some direction for vision as well. To safeguard vulnerable users, the design process must be led by a joint medical-technical team that continuously monitors how patients at different stages of disease react to and accept the system.



**Figure 10.** Proposed architecture for future AI-driven dementia care. A Retrieval-Augmented Generation (RAG) enhanced Large Language Model (LLM) engages patients in conversation via a text-to-speech (TTS) module and loudspeakers, continuously re-evaluating the situation in real-time. The system detects emotional speech cues, adapting its voice and responses accordingly. A “long-term memory” database stores patient history, personal data, reports, and audiovisual logs. The system interfaces with external devices (screens, robots, smartphones) and responds to caregiver queries with evidence-backed insights to ensure accountability.

### 5.3. 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. Clinical deployment cannot rely on free-field ASR for safety-critical decisions until such adaptation has been completed and evaluated.

## 6. Conclusions

Affordable, commodity sensors and AI can supplement—and in underserved cases partially substitute—in-home caregiving for older adults with dementia when dependable services are unaffordable or unavailable. We integrated novel, affordable devices and 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. Our goals are feasibility, systems integration, and safety/guardrail design rather than diagnostic performance. This manuscript is a proof-of-concept case study, 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 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, robots and screens.

**Supplementary Materials:** The following supporting information can be downloaded at the website of this paper posted on Preprints.org, [Picture Device.jpg](#). [Dialogues.pdf](#): Transcribed human dialogues. [sound\\_events\\_single\\_day.json](#): A JSON log of audio events over a single day from the automated system. Videos: [male\\_fall\\_subs\\_anonymized.mp4](#), [male\\_8\\_11\\_captioned.mp4](#), [female\\_fall\\_subs.mp4](#), [male\\_fall2\\_captioned.mp4](#), [female\\_fall2\\_short.mp4](#): Qwen3-VL scans videos for accidents. Video: [male\\_audio\\_only\\_fall\\_subs\\_anonymized.mp4](#). Audio only scanning. Emotions folder: Anonymized mp3 audio sample recordings of agitated/neutral state and figures of agitation probability vs time for each recording.

**Funding:** This research received no external funding.

**Ethical Issues:** In Greece, the general research–ethics framework is governed by Law 4521/2018, which requires ethical review for research involving human subjects in biomedical and social sciences. Institutional codes (e.g., the Code of Ethics and Conduct of Research of the University of West Attica, the Code of Conduct of ITE, and other universities) specify that research in the social sciences must protect anonymity and personal data and may proceed with consent if minimal-risk and non-interventional and do not fall under clinical/biomedical intervention. According to the Guide of the National Bioethics Commission, while ethics-committee review is recommended for research on human behavior, it is not mandated in every case. In this study the data collection consisted of audio-recorded conversations with older adults in a non-interventional, minimal-risk format; participants gave informed consent, data were pseudonymized, and no clinical intervention was involved. Therefore, as per the above legislative and institutional context, we determined that a full ethics-committee formal approval was not required while appropriate ethical safeguards (consent, anonymity, data protection) were adopted in accordance with local regulations and the ethical principles that have their origin in the principles of the Declaration of Helsinki.

**Institutional Review Board Statement:** Not applicable. The researchers analyzed only anonymized event metadata generated on-device; no identifiable audio or video were accessed or transferred.

**Informed Consent Statement:** Not applicable; no identifiable data or images are included.

**Data Availability Statement:** Raw recordings remain on participants' devices and are not available. Anonymized event metadata used in this study is not publicly available. A small sample of the dataset of audio events is available at Supplementary Materials.

**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 A

Code used in this work: <https://github.com/potamitis123/dementia-care> (accessed on 9/12/2025). In Supplementary Materials we analyzed two videos of falls with Qwen3-VL-235B-A22B (<https://chat.qwen.ai/> accessed on 9/12/2025) and passed the LLM's output as captions. The system recognized correctly the situation and can be set to notify guardians automatically. This is a concrete example of what audiovisual models can bring to elderly care. The first video has been tagged by our audio-only model and can be found on Supplementary Materials as well. The second video has no audio and in such cases audio only approaches are underpowered. The interested reader needs to know that no injuries happened to the elderly and that there was no way to intervene prior to the fall as the data have been analyzed offline. The videos have been defaced (i.e., blurred the face) when needed to anonymize them using <https://github.com/ORB-HD/deface/> (accessed on 9/12/2025). The links of all audio toolboxes used in this work can be found below.

### *Appendix A.1. 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 documents 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](https://huggingface.co/docs/transformers/model_doc/audio-spectrogram-transformer) (accessed on 9/12/2025)

### Appendix A.2. 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/> (accessed on 9/12/2025)

### Appendix A.3. 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> (accessed on 9/12/2025)

### Appendix A.4. Diarization

pyannote.audio is an open-source toolkit written in Python for speaker diarization. Based on PyTorch 2.7.1+cu118 machine learning framework, it comes with state-of-the-art pretrained models and pipelines that can be further fine-tuned to our data for even better performance.

<https://github.com/pyannote/pyannote-audio> (accessed on 9/12/2025)

### Appendix A.5. 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 Transformer’s 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 is 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> (accessed on 9/12/2025)

### Appendix A.6. Delusion Detection

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

“PROMPT: From now on, function as my expert assistant with access to all your knowledge in the domain of elderly care, elderly psychology, medicine, dementia and best nursing 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 speech/conversations 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.

SMALL HISTORY OF THE COUPLE

- (a) Female: 87; Male: 90+
- (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-h 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 #1: Detect delusional expressions in dialogues. [M]; Male, [F] Female, [A] Agitated emotional state. In absence of [A], emotional state Neutral"

#### DIALOGUES

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

Caregivers' daily summary

"Query #2: Prepare a text-only report summarizing the main events of the previous day by parsing the attached JSON file containing tagged audio events of the previous day <day>. Each event receives the three higher rank labels and their corresponding probabilities and a timestamp. Consider all three classes as possible valid classes. Be brief, prioritize events that can be related to health, nocturnal activity and movement patterns."

ChatGPT-5.1 response: "Here's your concise daily audio report for Wednesday, 10 September 2025 (Europe/Athens)—focused on health, nocturnal activity, and movement patterns.

- (1) Direct answer—What mattered yesterday
  - Nocturnal cough episode from 04:32–05:34 with multiple high-confidence cough events (peaks up to 0.81). Likely disturbed sleep and airway irritation.
  - Evening respiratory activity around 21:26–22:36, including a strong cough at 22:35:48 (0.82).
  - Door/movement events at typical shift/transition times: 09:17, 10:25, 12:20, 18:12, 18:47, 20:45, 21:30–21:32. No clear fall pattern.
  - Single low-confidence impact ("thump/thud") at 11:25:53—not suggestive of a fall given low probability and surrounding context.
  - Overall counts: 133 events total. Health-related: 47 (of which 18 nocturnal). Door/movement: 9. No high-confidence alarms/falls.

- (2) Practical action plan (apply today)

Health (coughs & sleep):

- Log this as a significant nocturnal cough episode (04:32–05:34). Ask caregivers to note any triggers (dry air, reflux after late meal, recent cold).
- Interventions today: ensure hydration, warm fluids (prepared by caregiver), avoid late heavy meals, keep head elevated during rest, check bedroom humidity (~40–50%).
- If cough clusters repeat ≥3 nights/week or intensify, consider GP review (rule out GERD, post-nasal drip, infection).

Night routine:

- Given repeated late-evening/early-morning events, consider a calmer wind-down from 20:30 (reduced TV volume, warm drink, bathroom check, brief reassurance).

Movement & safety:

- Door activity matches shift changes; no fall signature detected. Keep current fall-prevention setup (clear pathways, canes nearby, night lights). Caregivers to quickly check after any future "thud/door-slam" timestamps"

"Query #3: 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."

## References

1. World Health Organization. Available online: <https://www.who.int/news-room/fact-sheets/detail/dementia?> (accessed on 6 November 2025).
2. EUROSTAT. Available online: [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population\\_structure\\_and\\_ageing](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_structure_and_ageing) (accessed on 6 November 2025).
3. United Nations. Available online: [https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/undesa\\_pd\\_2024\\_wpp\\_2024\\_advance\\_unedited\\_0.pdf](https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/undesa_pd_2024_wpp_2024_advance_unedited_0.pdf) (accessed on 6 November 2025).
4. Alzheimer's Association. Available online: <https://www.alz.org/alzheimers-dementia/facts-figures?> (accessed on 6 November 2025).
5. König, A.; Satt, A.; Sorin, A.; Hoory, R.; Derreumaux, A.; David, R.; Robert, P.-H. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* **2015**, *1*, 112–124. <https://doi.org/10.1016/j.dadm.2014.11.012>.
6. Haider, F.; De La Fuente Garcia, S.; Luz, S. An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 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**, *15*, 1224723. <https://doi.org/10.3389/fnagi.2023.1224723>.
8. Saeedi, S.; Hetjens, S.; Grimm, M.O.W.; Barsties, V.; Latoszek, B. Acoustic Speech Analysis in Alzheimer's Disease: A Systematic Review and Meta-Analysis. *J. Prev. Alzheimers Dis.* **2024**, *11*, 1789–1797. <https://doi.org/10.14283/jpad.2024.132>.
9. Martínez-Nicolás, I.; Llorente, T.E.; Martínez-Sánchez, F.; Meilán, J.J.G. 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. <https://doi.org/10.3389/fpsyg.2021.620251>.
10. van den Berg, R.L.; de Boer, C.; Zwan, M.D.; Jutten, R.J.; van Liere, M.; van de Glind, M.-C.A.; Dubbelman, M.A.; Schlüter, L.M.; van Harten, A.C.; Teunissen, C.E.; et al. Digital remote assessment of speech acoustics in cognitively unimpaired adults: Feasibility, reliability and associations with amyloid pathology. *Alzheimers Res. Ther.* **2024**, *16*, 176. <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, J.J.; Martínez-Sánchez, F.; Carro, J.; López, D.E.; Millian-Morell, L.; Arana, J.M. Speech in Alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? *Dement. Geriatr. Cogn. Disord.* **2014**, *37*, 327–34. <https://doi.org/10.1159/000356726>.
13. Xue, C.; Karjadi, C.; Paschalidis, I.C.; Au, R.; Kolachalama, V.B. Detection of dementia on voice recordings using deep learning: A Framingham Heart Study. *Alzheimers Res. Ther.* **2021**, *13*, 146. <https://doi.org/10.1186/s13195-021-00888-3>.
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. <https://doi.org/10.2196/57873>.
15. Vincze, V.; Szatlóczki, G.; Tóth, L.; Gosztolya, G.; Pákáski, M.; Hoffmann, I.; Kálmán, J. Telltale silence: Temporal speech parameters discriminate between prodromal dementia and mild Alzheimer's disease. *Clin. Linguist. Phon.* **2021**, *35*, 727–742. <https://doi.org/10.1080/02699206.2020.1827043>.
16. Sefcik, J.S.; Ersek, M.; Hartnett, S.C.; Cacchione, P.Z. Integrative review: Persistent vocalizations among nursing home residents with dementia. *Int. Psychogeriatr.* **2019**, *31*, 667–683. <https://doi.org/10.1017/S1041610218001205>.
17. Kusters, J.; Janus, S.I.M.; Van Den Bosch, K.A.; Zuidema, S.; Luijendijk, H.J.; Andringa, T.C. Soundscape Optimization in Nursing Homes Through Raising Awareness in Nursing Staff With MoSART. *Front. Psychol.* **2022**, *13*, 871647. <https://doi.org/10.3389/fpsyg.2022.871647>.

18. Kaur, P.; Wang, Q.; Shi, W. Fall detection from audios with Audio Transformers. *Smart Health* **2022**, *26*, 100340. <https://doi.org/10.1016/j.smhl.2022.100340>.
19. Newaz, N.T.; Hanada, E. The Methods of Fall Detection: A Literature Review. *Sensors* **2023**, *23*, 5212. <https://doi.org/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.; Manifavas, 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. Casu, F.; Lagorio, A.; Ruiu, P.; Trunfio, G.A.; Grosso, E. Integrating Fine-Tuned LLM with Acoustic Features for Enhanced Detection of Alzheimer’s Disease. *IEEE J. Biomed. Health Inform.* **2025**. <https://doi.org/10.1109/JBHI.2025.3566615>.
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*, 126–133. <https://doi.org/10.1159/000362159>.
25. Fillit, H.; Aigbogun, M.S.; Gagnon-Sanschagrín, P.; Cloutier, M.; Davidson, M.; Serra, E.; Guérin, A.; Baker, R.A.; Houle, C.R.; Grossberg, G. Impact of agitation in long-term care residents with dementia in the United States. *Int. J. Geriatr. Psychiatry* **2021**, *36*, 1959–1969. <https://doi.org/10.1002/gps.5604>.
26. Pedersen, S.K.A.; Andersen, P.N.; Lugo, R.G.; Andreassen, M.; Sütterlin, S. Effects of Music on Agitation in Dementia: A Meta-Analysis. *Front. Psychol.* **2017**, *8*, 742. <https://doi.org/10.3389/fpsyg.2017.00742>.
27. Miller, S.; Vermeersch, P.E.; Bohan, K.; Renbarger, K.; Kruep, A.; Sacre, S. Audio presence intervention for decreasing agitation in people with dementia. *Geriatr. Nurs.* **2001**, *22*, 66–70. <https://doi.org/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.* **2025**, rs.3.rs-6008628. <https://doi.org/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. <https://doi.org/10.3390/s17040854>.
30. Rashidi, P.; Cook, D.J. Keeping the Resident in the Loop: Adapting the Smart Home to the User. *IEEE Trans. Syst. Man Cybern. – Part A Syst. Hum.* **2009**, *39*, 949–959. <https://doi.org/10.1109/TSMCA.2009.2025137>.
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. <https://doi.org/10.1109/titb.2004.834393>.
32. Bouchard, B.; Giroux, S.; Bouzouane, A. A Keyhole Plan Recognition Model for Alzheimer’s Patients: First Results. *Appl. Artif. Intell.* **2007**, *21*, 623–658. <https://doi.org/10.1080/08839510701492579>.
33. Mégret, R.; Dovgalecs, V.; Wannous, H.; Karaman, S.; Benois-Pineau, J.; El Khoury, E.; Pinquier, J.; Joly, P.; André-Obrecht, R.; Gaëstel, R.; et al. The IMMED Project: Wearable Video Monitoring of People with Age Dementia. In Proceedings of the 18th ACM international conference on Multimedia, Firenze, Italy, 25–29 October 2010; Association for Computing Machinery: 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*, 92–99. <https://doi.org/10.1109/titb.2005.856855>.

37. Barnes, N.; Edwards, N.; Rose, D.; Garner, P. Lifestyle monitoring technology for supported independence. *Comput. Control. Eng.* **1998**, *9*, 169–174. <https://doi.org/10.1049/cce:19980404>.
38. Tamura, T.; Kawarada, A.; Nambu, M.; Tsukada, A.; Sasaki, K.; Yamakoshi, K. E-healthcare at an experimental welfare techno house in Japan. *Open Med. Inform. J.* **2007**, *1*, 1–7. <https://doi.org/10.2174/1874431100701010001>.
39. Boumpa, E.; Gkogkidis, A.; Charalampou, I.; Ntaliani, A.; Kakarountas, A.; Kokkinos, V. An Acoustic-Based Smart Home System for People Suffering from Dementia. *Technologies* **2019**, *7*, 29. <https://doi.org/10.3390/technologies7010029>.
40. Periša, M.; Teskera, P.; Cvitić, I.; Grgurević, I. Empowering People with Disabilities in Smart Homes Using Predictive Informing. *Sensors* **2025**, *25*, 284. <https://doi.org/10.3390/s25010284>.
41. Gong, Y.; Chung, Y.-A.; Glass, J. AST: Audio Spectrogram Transformer. *arXiv* **2021**, arXiv:2104.01778.
42. 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.
43. Ilias, L.; Askounis, D. Multimodal Deep Learning Models for Detecting Dementia From Speech and Transcripts. *Front. Aging Neurosci.* **2022**, *14*, 830943. <https://doi.org/10.3389/fnagi.2022.830943>.
44. Llaca-Sánchez, B.A.; García-Noguez, L.R.; Aceves-Fernández, M.A.; Takacs, A.; Tovar-Arriaga, S. Exploring LLM Embedding Potential for Dementia Detection Using Audio Transcripts. *Eng* **2025**, *6*, 163. <https://doi.org/10.3390/eng6070163>.
45. Zhang, M.; Pan, Y.; Cui, Q.; Lü, Y.; Yu, W. Multimodal LLM for enhanced Alzheimer’s Disease diagnosis: Interpretable feature extraction from Mini-Mental State Examination data. *Exp. Gerontol.* **2025**, *208*, 112812. <https://doi.org/10.1016/j.exger.2025.112812>.
46. Bang, J.; Han, S.; Kang, B. Alzheimer’s disease recognition from spontaneous speech using large language models. *ETRI J.* **2024**, *46*, 96–105. <https://doi.org/10.4218/etrij.2023-0356>.
47. Li, R.; Wang, X.; Berlowitz, D.; Mez, J.; Lin, H.; Yu, H. CARE-AD: A multi-agent large language model framework for Alzheimer’s disease prediction using longitudinal clinical notes. *npj Digit. Med.* **2025**, *8*, 541. <https://doi.org/10.1038/s41746-025-01940-4>.
48. Du, X.; Novoa-Laurentiev, J.; Plasek, J.M.; Chuang, Y.W.; Wang, L.; Marshall, G.A.; Mueller, S.K.; Chang, F.; Datta, S.; Paek, H.; et al. Enhancing early detection of cognitive decline in the elderly: A comparative study utilizing large language models in clinical notes. *EBioMedicine* **2024**, *109*, 105401. <https://doi.org/10.1016/j.ebiom.2024.105401>.
49. Amini, S.; Hao, B.; Yang, J.; Karjadi, C.; Kolachalama, V.B.; Au, R.; Paschalidis, I.C. Prediction of Alzheimer’s disease progression within 6 years using speech: A novel approach leveraging language models. *Alzheimers Dement.* **2024**, *20*, 5262–5270. <https://doi.org/10.1002/alz.13886>.
50. Agbavor, F.; Liang, H. Predicting dementia from spontaneous speech using large language models. *PLOS Digit. Health* **2022**, *1*, e0000168. <https://doi.org/10.1371/journal.pdig.0000168>.
51. Chen, L.; Asgari, M. Refining Automatic Speech Recognition System for Older Adults. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7003–7007. <https://doi.org/10.1109/ICASSP39728.2021.9414207>.
52. Hu, S.; Xie, X.; Geng, M.; Jin, Z.; Deng, J.; Li, G.; Wang, Y.; Cui, M.; Wang, T.; Meng, H.; et al. Self-Supervised ASR Models and Features for Dysarthric and Elderly Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 3561.
53. Geng, M.; Xie, X.; Ye, Z.; Wang, T.; Li, G.; Hu, S.; Liu, X.; Meng, H. Speaker Adaptation Using Spectro-Temporal Deep Features for Dysarthric and Elderly Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2597–2611. <https://doi.org/10.1109/TASLP.2022.3195113>.
54. El Hajal, K.; Hermann, E.; Hovsepyan, S.; Doss, M.M. Unsupervised Rhythm and Voice Conversion to Improve ASR on Dysarthric Speech. *Proc. Interspeech* **2025**, *2025*, 2760–2764. <https://doi.org/10.21437/Interspeech.2025-2069>.

55. Barry, S.J.; Dane, A.D.; Morice, A.H.; Walmsley, A.D. The automatic recognition and counting of cough. *Cough* **2006**, *2*, 8. <https://doi.org/10.1186/1745-9974-2-8>.
56. A PRACTICAL GUIDE FOR PEOPLE WITH DEMENTIA Available online: <https://www.alzheimers.org.uk/sites/default/files/2022-07/Caring-for-a-person-with-dementia-a-practical-guide.pdf> (accessed on 20 November 2025).
57. Brodaty, H. The practice and ethics of dementia care. *Int. Psychogeriatr.* **2015**, *27*, 1579–1581. <https://doi.org/10.1017/S1041610215001234>. (accessed on 6 November 2025).

**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.