Tonal Salience in Cognitive Decline: In-Context MCI Detection with Multimodal LLMs

Christopher Song
Independent Researcher
christopher.w.song@gmail.com

Abdullah P. Rashed Ahmed
Brown University
abdullah_rashed_ahmed@brown.edu

Abstract

Speech-based screening for mild cognitive impairment offers a highly accessible way to detect early cognitive decline. While most existing work focuses on English, cross-linguistic research is emerging to examine how cognitive decline manifests across languages. Studies on the Interspeech 2024 TAUKADIAL dataset, comprising English and Chinese speech recordings, consistently report higher classification performance on Chinese, yet the cause of this cross-lingual discrepancy remains unexplored. We examine this gap using Gemini 2.5 Pro, a multimodal large language model, using zero-shot and in-context-learning (ICL) paradigms. We hypothesize that this disparity is rooted in language typology: in tonal languages like Chinese, pitch encodes lexical meaning in every syllable (tone), whereas in non-tonal languages like English, pitch carries no lexical function. To test this, we pitch-flattened audio from TAUKADIAL and compared how classification performance changed across both languages. We found that Chinese classification degraded significantly under both zero-shot and ICL conditions (-4.78 and -5.92 UAR, respectively), while English performance increased (+0.11 and +2.98 UAR), implicating tonal pitch as the cross-lingual advantage. These findings suggest language typology should inform the design of audio-based cognitive screening tools, with raw audio preferred for tonal languages and text for non-tonal languages, a distinction critical for developing equitable cross-linguistic screening.

1 Introduction

Mild cognitive impairment (MCI) occupies a critical window in the progression of neurodegenerative disease. As a frequent precursor to Alzheimer's and related dementias, MCI represents a period in which intervention can still alter outcomes, including slowing progression and preserving quality of life [2]. However, early cognitive decline is notoriously difficult to detect. Symptoms such as subtle memory lapses, hesitations in speech, and occasional word-finding failures are easily mistaken for normal aging [11, 20]. By the time these symptoms become obvious, the window for effective intervention may have already narrowed. This diagnostic gap has driven interest in non-invasive screening tools capable of detecting cognitive changes before they reach clinical thresholds.

Speech analysis has emerged as a promising candidate. Cognitive decline affects language production early, resulting in observable changes such as increased pausing, disfluencies, reduced lexical diversity, and altered prosody [7]. Unlike other diagnostics such as neuroimaging or cerebrospinal fluid analysis, speech can be collected remotely and continuously, making it well-suited for large-scale screening [8].

However, nearly all MCI speech research focuses on a single language, typically English. Prior work has identified acoustic markers, such as pause distribution and pitch variability, as well as linguistic features like decreased lexical diversity [9, 18, 32], examining how cognitive decline affects English speech but not how it manifests across different languages. Dementia is a global health crisis, with projected cases in Asia alone exceeding 70 million by 2050 [2], yet we do not know whether acoustic biomarkers of cognitive decline differ across languages, or whether classification models generalize across languages.

A growing body of work has begun extending MCI detection beyond English. The Interspeech 2024 TAUKADIAL Challenge [17] formalized this effort, providing the first bilingual benchmark for Chinese-English MCI classification. Work on this dataset includes a range of multimodal systems including CogniVoice [5] and Whisper-based ensemble models [1]. Across these, a consistent

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 913–922

pattern emerges: acoustic models achieve higher classification performance on Chinese than English, a finding corroborated by subsequent work using Qwen2-Audio [24]. CogniVoice frames one of its core contributions as reducing the cross-lingual performance gap, implicitly acknowledging the asymmetry as a known challenge. In addition, a recent systematic review of NLP methods for cognitive impairment detection independently notes that Chinese MCI studies achieve particularly high classification accuracy [25]. To our knowledge, no prior work has investigated this asymmetry in depth.

We hypothesize that this disparity is rooted in language typology. In tonal languages such as Chinese, pitch contours on every syllable (tone) are used to distinguish word meanings, requiring precise motor-cognitive control that is susceptible to cognitive decline. In non-tonal languages like English, pitch primarily serves as an expressive, prosodic function to convey emotion and sentence type without affecting lexical identity. Under this tonal salience hypothesis, Chinese audio provides a more salient acoustic signal for MCI detection.

To evaluate this, we investigate MCI detection in English and Chinese using Gemini, a multimodal large language model, with zero-shot and in-context learning on the TAUKADIAL dataset [17] across raw audio, transcribed text, and pitch-flattened audio. For the pitch-flattened experiments, we used PSOLA-based F0 manipulation, which removes tonal features while preserving other acoustic features. If the increased Chinese performance we observed is rooted in tonal features, removing pitch information should degrade Chinese classification disproportionately compared to English classification.

2 Background

2.1 Speech-Based CI Detection

Speech-based MCI detection pipelines typically extract two categories of features from recordings. Acoustic features capture paralinguistic properties of the speech signal, including pitch variability, speech rate, voice quality, and pause distributions. Standard implementations use feature sets such as eGeMAPS (88 descriptors) and ComParE (6,373 descriptors), which aggregate frame-level measurements into utterance-level statistics [9, 18]. Linguistic features operate on transcribed speech (transcripts typically generated manually or via ASR), extracting lexical and syntactic information either through hand-crafted metrics (type-token ratio, syntactic complexity) or dense embeddings from pretrained language models like BERT [32]. These features are fed into supervised classifiers – commonly SVMs, random forests, or shallow neural networks – trained on labeled datasets.

More recent work on the TAUKADIAL dataset extends this to multilingual settings, including English and Chinese [17]. While effective, these are supervised approaches, still requiring task-specific feature engineering and labeled training data. This motivates our use of large language models.

2.2 Transformers and Large Language Models

Gemini, the model we use for MCI classification, is a large language model built on the transformer architecture. The transformer architecture [29] introduced a self-attention mechanism that enables models to weigh the relationships between all elements in a sequence simultaneously, replacing the sequential processing of recurrent neural networks (RNNs). This parallel processing enables transformers to capture long-range dependencies efficiently and scale to very large datasets and parameter counts. Scaling transformer-based language models to billions of parameters yields emergent capabilities not explicitly trained for [4]. Among these is in-context learning, the ability to adapt to new tasks from a few examples provided in the prompt without any parameter updates (further discussed in 2.4). The combination of flexible sequence modeling, scalability, and emergent few-shot capabilities has made the transformer the dominant architecture for both language and multimodal systems. Because self-attention operates over arbitrary token sequences, the transformer architecture can be used across multiple input modalities, including audio.

2.3 Audio LLMs

Audio Large Language Models (AudioLLMs) typically consist of an audio encoder, often a pretrained model like Whisper [22], that converts audio into a sequence of embeddings then processed by an LLM. This architecture enables AudioLLMs to perform diverse audio understanding tasks through natural language prompting, without task-specific output heads.

Qwen2-Audio [6] uses a Whisper-large-v3 encoder paired with the Qwen-7B LLM, trained on

approximately 30 audio tasks including speech recognition, emotion recognition, and speaker analysis.

Gemini [12] is a family of multimodal models developed by Google that natively process text, images, audio, and video. Unlike Qwen2-Audio's encoder-decoder architecture, Gemini is trained end-to-end as a multimodal model, directly processing raw audio alongside other modalities. This difference, combined with Gemini's larger scale and broader training data, motivates our investigation of its capabilities for MCI detection.

While prior work demonstrated that AudioLLMs can perform zero-shot MCI classification [24], the potential of few-shot learning remains underexplored.

2.4 In-Context Learning

In-context learning (ICL) refers to the ability of large language models to perform tasks given a few input-output examples provided in the prompt [3]. Rather than tuning on a training set, ICL presents the model with labeled demonstrations of the target task followed by a new query, and the model generates a prediction conditioned on those demonstrations. Because it uses only inference-time computation, it is attractive for domains with limited labeled data.

For MCI detection, prior AudioLLM work on cognitive assessment Shahin et al. [24] used only zero-shot prompting, leaving open the question of whether few-shot examples can improve classification performance for this task. By providing the model with examples of speech from cognitively healthy and impaired individuals, we can supply implicit information about the acoustic and linguistic patterns associated with cognitive decline, without requiring gradient-based training.

2.5 Tone in Language Typology

Tonal languages, which account for approximately one-third of the world's languages, use pitch contour at the syllable level to distinguish words with completely different meanings [16]. For example, Mandarin Chinese has 4 tones, where a syllable like "ma" can mean "mother" (first tone, level), "hemp" (second tone, rising), "horse" (third tone, dipping), or "scold" (fourth tone, falling). Crucially, production of tone is obligatory (a speaker must produce tones when speaking) and perceptually salient (tone errors are noticeable to listeners). In non-tonal languages like English, pitch variation instead primarily serves as a prosodic role, marking emphasis, conveying emotion, or signaling sentence type (questions, exclamations). It does not change the identity of individual words. This typology difference has implications for the cognitive and motor demands of speech production. Producing lexical tone requires real-time control of laryngeal muscles, as intracranial recording work has shown that local populations in the bilateral laryngeal motor cortex (LMC) encode articulatory kinematic information to generate the pitch dynamics of lexical tone [16]. This contrasts with coarser prosodic modulation of non-tonal languages, where pitch variation tolerates greater variability without impacting meaning. Evidence suggests that fine-grained pitch control is vulnerable to neurological impairment, motivating investigation of whether tonal features are similarly affected in MCI.

Most relevant to the present study, Feng et al. [10] investigated categorical perception of Mandarin tones in seniors with MCI and found that while tone identification (categorizing a sound) remained intact, tone discrimination (comparing two sounds to determine if they're the same or different) degraded significantly. Considering that discrimination places a higher demand on working memory compared to identification, they attribute the impairment to impaired working memory and long-term phonological memory. For non-tonal languages, dementia-related prosodic changes have been documented, including reduced F0 variability and flattened emotional intonation [13, 19], but these changes do not alter lexical meaning.

3 Methods

To investigate the effects of tone on MCI classification, we evaluate Gemini on the TAUKADIAL dataset under three conditions: (1) zero-shot classification on audio, (2) zero-shot classification on text transcripts, and (3) in-context learning on audio, to assess baseline performance and the cross-lingual asymmetry across modalities and learning paradigms. We then repeat conditions (1) and (3) with pitch-flattened audio to isolate whether tonal information drives the cross-lingual gap.

3.1 Dataset

We evaluate on TAUKADIAL-24 [17], a multilingual corpus from the INTERSPEECH 2024

	English	Chinese	Total
NC	93 (38%)	129 (49%)	222
MCI	153 (62%)	132 (51%)	285
Total	246	261	507

Table 1: TAUKADIAL combined (train + test) dataset distribution by language and cognitive status.

challenge on speech-based cognitive assessment. The dataset, hosted by DementiaBank [14], contains recordings of elderly speakers performing a clinician-administered picture description task, with each participant describing three images. The dataset contains 507 speech samples (261 Chinese and 246 English). English-speaking participants described the "Cookie Theft", "Cat Rescue", and "Coming and Going" pictures. Chinese-speaking participants described three pictures depicting Taiwanese culture; individual picture names are not documented in the dataset. Recordings are labeled as either Normal Cognition (NC) or Mild Cognitive Impairment (MCI) based on Mini Mental State Examination (MMSE) or Montreal Cognitive Assessment (MoCA) screening, with MCI diagnosis following National Institute on Aging-Alzheimer's Association (NIA-AA) criteria.

To control for demographic confounds, the dataset was balanced for age and sex using propensity score matching [23], using logistic regression to estimate each participant's probability of MCI classification. All standardized mean differences for covariates fell below 0.1, indicating adequate balance [17].

Following Shahin et al. [24], we evaluated on a combined training and test split, originally a ratio of approximately 3:1 (Table 1).

3.2 Model

For all experiments, we use Gemini 2.5 Pro [12], a multimodal LLM developed by Google DeepMind featuring a sparse mixture-of-experts architecture with native support for text, image, and audio. It features a context window of up to 1M tokens and an audio encoding rate of approximately 32 tokens per second, which enables inclusion of multiple full-length recordings in a single prompt. Unlike models that rely on an external encoder like Whisper, Gemini processes audio end-to-end as a native modality.

All experiments were conducted via the Google Cloud Vertex AI SDK using the model identifier gemini-2.5-pro, accessed July-December 2025. For cross-model comparison, we also evaluated gemini-3.0-pro under the ICL audio condition. For all runs, the temperature was set to the default value of 1 and no other parameters were modified.

3.3 Prompt Design

Following Shahin et al. [24], we designed a direct classification prompt instructing the model to assess cognitive status for our zero-shot and ICL experiments:

> *Assess the cognitive condition based on the input audio, where an elderly speaker describes one of three images as part of a clinician-guided task. Indicate the diagnosis using one of these labels: NC (Normal Cognitive) or MCI (Mild Cognitive Impairment). Output only "NC" or "MCI" as your response. Do not include any explanation, reasoning, or additional text.*

For text evaluation, we substitute "input audio" with "transcript". This prompt design mirrors the "Contextual" prompt type from Shahin et al. [24]. We do not use chain-of-thought (CoT) prompting, as they found it did not improve performance for this task.

3.4 Transcription

The TAUKADIAL dataset does not provide text transcripts. For the zero-shot text condition, we generated transcripts from the raw audio using Whisper base [22], OpenAI's open-source automatic speech recognition model. Each audio file was transcribed independently, with no manual correction applied. We selected Whisper for its strong multilingual performance across English and Chinese, and because it serves as the audio encoder backbone of Qwen2-Audio [6].

3.5 In-Context Learning Protocol

We devised the following ICL protocol, which we applied to all 507 samples in the combined dataset. For each test sample, we inserted two exemplars into the prompt, one NC and one MCI, drawn from the same language as the test sample, excluding the test speaker themselves. To reduce variance from exemplar selection, we repeated this classification 3 times per sample, each time sampling a fresh exemplar pair, and then determined the final prediction by majority vote across the 3 predictions. We did not set a fixed seed.

[Classification instruction]

`Here are 2 exemplars:`

Label: NC
[Audio: exemplar_1.wav]

Label: MCI
[Audio: exemplar_2.wav]

`Here is the sample to be classified:`

[Audio: target.wav]

Figure 1: Structure of the ICL prompt for audio classification. The classification instruction is the prompt described in subsection 3.3.

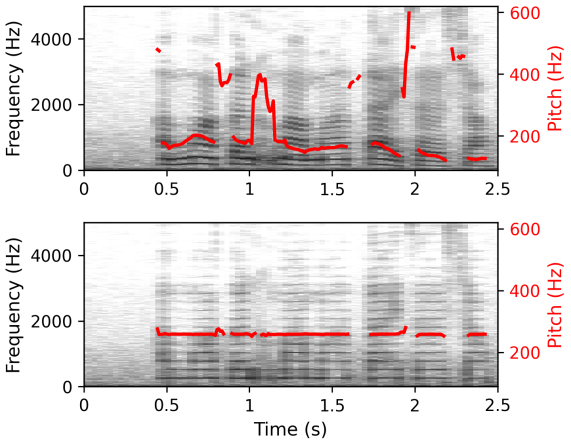

Figure 2: Example of pitch flattening applied to a 2.5-second clip of a Chinese audio file. **Top:** Original spectrogram overlaid with the extracted pitch contour. The contour exhibits the rising and falling movements that encode lexical tone. Gaps indicate unvoiced segments (e.g. stops). **Bottom:** Flattened version of the same audio, produced by replacing the original F0 contour with the mean F0 value via PSOLA. The pitch contour is now flat, while the spectral envelope, formant structure, and duration are all intact, only the pitch dynamics have been removed (see subsection 3.6).

To our knowledge, in-context learning has not been previously applied to AudioLLM-based cognitive assessment.

3.6 Pitch Flattening Protocol

To test whether the cross-lingual asymmetry is driven by tonal information, we removed pitch contours from all audio recordings prior to classification. We applied pitch flattening using Praat's Pitch-Synchronous Overlap and Add (PSOLA) algorithm [15, 28], which manipulates F0 while preserving other speech qualities including duration, intensity, and spectral envelope. We accessed Praat through the parselmouth Python library. F0 was extracted using Praat's autocorrelation method with a pitch floor of 75 Hz and ceiling of 600 Hz, and a time step of 0.01 s. The mean F0 was computed over voiced frames only, and unvoiced segments were left unmodified. Resynthesis used Praat's PSOLA implementation, with all other parameters left at default.

For each recording, the original F0 contour was replaced with the mean F0 value, removing the pitch dynamics that encode lexical tone in Chinese.

Our approach follows Liang and Levow [15], who used PSOLA to investigate pitch's role in ASR performance between tonal and non-tonal languages. They found that tonal languages experienced significantly larger ASR degradation than non-tonal languages, confirming that pitch carries crucial information in tonal systems. We adapt this paradigm from ASR to MCI classification. The pitch-flattened audio samples were classified under both zero-shot and ICL conditions, described in detail above.

We note that PSOLA flattening doesn't eliminate all traces of pitch information, as it leaves microperiodicity cues in the harmonic spectrum [15]. The flattened data therefore represents a conservative estimate of tonal removal and contribution of tonal information to performance.

3.7 Metrics

To evaluate performance on the TAUKADIAL dataset, we report Unweighted Average Recall (UAR) and macro-F1 (mF1), following the TAUKADIAL evaluation protocol. UAR is the arithmetic mean of per-class recall, giving equal weight to each class regardless of its dominance in the dataset. For binary classification:

$$\text{UAR} = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2} \quad (1)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives respectively. Unlike standard accuracy, UAR is robust to class imbalance. For example, a classifier that predicts only the majority class achieves 50% UAR regardless of class distribution.

Macro-F1 is the unweighted mean of per-class F1 scores, where each class's F1 is the harmonic

mean of its precision and recall:

$$\mathrm{F1}_c = \frac{2 \cdot \mathrm{Precision}_c \cdot \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c} \quad (2)$$

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{F1}_c \quad (3)$$

where C is the number of classes. Like UAR, macro-F1 treats each class equally, but additionally penalizes low precision — a classifier with high recall but many false positives will receive a lower macro-F1 than UAR.

3.8 Code and Data Release

All experiment code, including the zero-shot classification, ICL classification, and PSOLA flattening pipeline is available at https://github.com/ChispTheLegend/mci-tonality.

The TAUKADIAL audio data is distributed by DementiaBank under a data use agreement and cannot be redistributed by us. Researchers can request access from them directly. We provide our generated Whisper transcripts and model predictions in the repo to support partial reproduction without dataset access.

4 Results

4.1 Classification Performance

To assess Gemini's ability to detect MCI from speech and whether it also produces the cross-lingual performance gap, we evaluated Gemini on the TAUKADIAL dataset under three conditions: zero-shot classification on audio, zero-shot classification on transcripts (text), and in-context learning with audio using the two-exemplar, three-run majority vote protocol described in subsection 3.5. Table 2 presents classification performance across all conditions, alongside prior supervised baselines and AudioLLM performance.

The cross-lingual gap emerges across our results and prior work: Chinese outperforms English in audio-based classification conditions, persisting across previous, zero-shot, and ICL paradigms. Under zero-shot audio classification, Chinese already outperforms English (60.41 vs. 58.70 UAR). This gap widens under ICL (67.81 vs. 61.43 UAR), suggesting that exemplars sharpen the model's attention to the features underlying the Chinese advantage. Critically, the pattern reverses for our text-only classification, where English outperforms Chinese (50.02 vs. 60.44 UAR), suggesting the Chinese advantage is specific to audio rather than textual content. We also evaluated Gemini 3.0 Pro to assess whether the observed pattern was specific to Gemini 2.5 Pro, but found that performance is nearly identical across both models (63.35 vs. 63.37 overall UAR). The cross-lingual asymmetry is preserved (67.5 vs. 61.41 UAR for Chinese and English, respectively, under ICL). Given the consistency, we used Gemini 2.5 Pro for all subsequent pitch-flattening experiments due to its lower token cost.

4.2 Pitch Flattening Performance

To assess whether the cross-lingual gap for audio classification is rooted in tonal features, we repeated both zero-shot and ICL experiments on pitch-flattened audio. See subsection 3.6 for more detail on the ablation protocol. Table 3 presents the change in performance in both languages after pitch flattening.

Chinese classification degrades substantially under flattening, while English remains essentially unchanged or improves under flattening. For Chinese, UAR drops 4.78 points under zero-shot (60.41 to 55.63) and 5.92 points under ICL (67.81 to 61.89). F1 drops 7.92 points under zero-shot (60.23 to 52.31) and 6.75 points under ICL (67.81 to 61.06). For English, UAR under zero-shot is largely unchanged (58.70 to 58.81), and increases 2.98 points under ICL (61.43 to 64.41). F1 increases by 3.71 under zero-shot (48.06 to 51.77) and 5.6 points under ICL (55.37 to 60.97).

5 Discussion

5.1 ICL Effectiveness

We utilized ICL rather than supervised training to investigate the cross-lingual gap. Providing two labeled exemplars yielded a 6.2 UAR point improvement over zero-shot, exceeding the supervised baselines from the challenge organizers and prior AudioLLM approaches without any gradient updates or feature engineering. Beyond the raw performance gain, our ICL approach also replicates the cross-lingual performance pattern observed in supervised models. In zero-shot classification, the model relies entirely on its pretraining to distinguish MCI from normal cognition, while ICL, which anchors the model in this specific task, increases the chance that the model is attending to task-relevant features. It enables us to test the tonal hypothesis without model fine-tuning or feature ex-

Learning	Model	Modality	UAR (%)			Macro-F1 (%)		
			Overall	**ZH**	**EN**	**Overall**	**ZH**	**EN**
Prior work								
eGeMAPS + MLP [17]		Audio	54.89	—	—	—	—	—
Wav2Vec + eGeMAPS [17]		Audio	59.18	60.04	60.00	60.20	—	—
CogniVoice† [5]		Audio+Text	75.10	77.50	58.40	84.10	81.70	81.30
Qi et al.† [21]		Audio	68.20	70.30	55.00	78.90	71.80	79.20
Qwen2-Audio (best prompt) [24]		Audio	57.50	60.94	54.60	56.79	60.91	52.34
Qwen2-Audio (majority vote) [24]		Audio	59.00	—	—	—	—	—
Our results								
Zero-shot	Gemini 2.5 Pro	Text	51.58	50.02	**60.44**	50.40	34.92	**58.01**
Zero-shot	Gemini 2.5 Pro	Audio	57.14	**60.41**	58.70	55.50	**60.23**	48.06
ICL	Gemini 2.5 Pro	Audio	63.35	**67.81**	61.43	61.88	**67.81**	55.37
ICL	Gemini 3.0 Pro	Audio	63.37	**67.50**	61.41	61.06	**67.34**	54.24

Table 2: Classification performance across methods. ZH and EN columns show per-language UAR and Macro-F1 for Chinese and English, respectively. Dashes indicate metrics not reported in the original work. Bolded results represent the higher performance between Chinese and English. †Results are from k-fold cross-validation on the training set, not the official test split.

Condition	UAR (%)			Macro-F1 (%)		
	Orig.	**Flat.**	**Δ**	**Orig.**	**Flat.**	**Δ**
English						
Zero-shot	58.70	58.81	+0.11	48.06	51.77	+3.71
ICL	61.43	64.41	+2.98	55.37	60.97	+5.60
Chinese						
Zero-shot	60.41	55.63	−4.78	60.23	52.31	−7.92
ICL	67.81	61.89	−5.92	67.81	61.06	−6.75

Table 3: Effect of pitch flattening on classification performance for Gemini 2.5 Pro (Flattened – original, in percentage points). English improves under flattening; Chinese performance degrades substantially, consistent with the tonal hypothesis.

traction.

5.2 Cross-Lingual Asymmetry

Our experiments demonstrate a consistent performance advantage for Chinese over English in audio classification. Chinese outperforms English in zero-shot (60.41 vs. 58.70 UAR) and widens under ICL conditions (67.81 vs. 61.43), as the model is more attuned to the task. These results suggest this gap reflects how cognitive impairment manifests in tonal vs. non-tonal languages, rather than being an artifact of model training, architecture, or dataset.

One might attribute the performance discrepancy to model features, such as training-data composition. However, the cross-lingual gap replicates across Gemini, developed by Google DeepMind (US), and Qwen, developed by Alibaba (China). English accounts for approximately 50% of web content [30]. Google's open source model family Gemma, which shares the same research base and tokenizer as Gemini, is explicitly trained on "primarily-English data" [26, 27]. Gemini's training distribution is not publicly disclosed, but its shared lineage with Gemma suggests a similar composition. Qwen2-Audio is explicitly pretrained with a focus on Chinese and English [6, 31]. Rather than each model performing better on the language associated with its origin, both achieve higher accuracy on Chinese. Consistency across separate multimodal architectures, Gemini's native audio processing and Qwen's Whisper-based encoder-decoder pipeline, further suggests the gap is rooted in the acoustic properties of the languages rather than model properties.

When Gemini received only text transcripts for classification, which discards acoustic features in transcription, the pattern reversed. English outperformed Chinese (50.0 vs. 60.4 UAR), suggesting that the Chinese advantage is rooted in acoustic features present in audio but lost in transcription.

5.3 Pitch Flattening Supports the Tonal Hypothesis

To test whether tonal information drives the cross-lingual gap, we replaced each recording's pitch contour with its mean F0, effectively removing lexical tone while preserving duration, intensity, and

spectral envelope.

We found that Chinese classification degraded substantially under flattening: UAR dropped 4.8 points under zero-shot and 5.9 points under ICL. English classification is largely unaffected, being essentially unchanged under zero-shot and actually increasing under ICL. The cross-lingual gap entirely reverses under both zero-shot and ICL flattening conditions (58.81 vs 55.63 UAR under zero-shot, and 64.41 vs. 61.89 UAR under ICL).

This asymmetric degradation isolates pitch contour as the acoustic feature which drives the higher performance in Chinese over English. When tone information is provided, the Chinese classification performs better than English; when it is removed, the Chinese performs comparably to or worse than English. Our results corroborate Liang and Levow [15], who applied the same PSOLA protocol to flatten audio across 6 languages and found that tonal languages experienced disproportionately greater ASR degradation than non-tonal languages. Their parallel finding supports the conclusion that pitch carries crucial information in tonal languages, which, in our case, can carry markers of cognitive decline.

Together, these findings suggest that the cross-lingual asymmetry in MCI classification is not a result of the dataset or model bias, but language typology: tonal languages encode cognitive decline saliently in their acoustic features, because the motor and cognitive demands of tone production are vulnerable to cognitive impairment.

The English improvement under flattening and ICL is unexpected. One possibility is that pitch variation in non-tonal languages, which encodes emotional and prosodic information, functions as noise for the purpose of MCI classification, so removing it may allow the model to attend more cleanly to more relevant acoustic features. This is consistent with our tonal hypothesis more broadly: pitch carries distinct functional roles across language types, with different implications for cognitive decline markers. Validating this interpretation would require further experimentation isolating acoustic feature contributions, which we leave to future work.

5.4 Limitations

Differing stimuli. The English and Chinese subsets of TAUKADIAL use different picture stimuli for the description task, which may elicit different speech patterns or cognitive demands. This confound cannot be fully ruled out; future work should evaluate performance using matched tasks across languages.

Model coverage. Our pitch-flattening experiments were conducted on a single model, Gemini 2.5 Pro. Validating the tonal salience hypothesis across additional LLMs – such as GPT-4o, Phi-4-multimodal, or Qwen2.5-Omni – would strengthen claims of generalizability. However, the underlying cross-lingual asymmetry has already been observed across independent architectures, including Whisper-based supervised systems [5] and Qwen2-Audio [24], suggesting the pattern persists across models. The asymmetric degradation under pitch flattening, which is the core contribution of this work, remains to be replicated in models other than Gemini, and we leave this to future work. Replicating the zero-shot, ICL, and pitch-flattening experiments across additional LLMs was bottlenecked by inference cost. We hope future work with broader access can extend this analysis.

Dataset scope. Generalization to other language populations, recording conditions, and cognitive assessment tasks remains unexplored. Validation of the tonal language hypothesis would require testing across (a) other tonal languages within the Sinitic family (Cantonese), (b) tonal languages from other families (Vietnamese, Thai), (c) pitch-accent languages as an intermediate between tonal and non-tonal languages (Japanese, Swedish), and (d) additional non-tonal controls (Spanish, German). Such validation is currently bottlenecked, as TAUKADIAL remains the only multilingual MCI speech corpus we are aware of.

Exemplar selection. Our ICL approach uses randomly selected exemplars. Performance might improve with more sophisticated selection strategies, such as selecting maximally informative exemplars.

Binary classification. Following the TAUKADIAL dataset classification task, we frame MCI detection as binary classification (NC vs. MCI). Eventual clinical deployment could benefit from finer-grained severity estimation – the TAUKADIAL dataset also includes the MMSE scores of each sample as a separate regression task.

Interpretability. LLMs operate as black boxes; we cannot directly inspect which acoustic features drive classification decisions. While our text vs. audio comparison provides indirect evidence that prosodic features matter, we cannot ”open up” the model to identify the specific markers that the

model relies on. This limits clinical interpretability and may hinder trust in deployment contexts.

6 Conclusion

We investigated MCI detection from speech samples using Gemini, a multimodal large language model, with in-context learning on the TAUKADIAL dataset. First, our ICL approach improves classification over zero-shot prompting, outperforming prior zero-shot approaches. Second, audio-based classification consistently performs better in Chinese than English. Third, pitch-flattening disproportionately degrades Chinese classification while improving English, confirming that tonal information is the acoustic feature behind the cross-lingual gap. Our findings suggest that language typology should inform the design of cognitive screening tools. For tonal languages, audio is necessary to capture tonal features of cognitive decline, whereas for non-tonal languages, text-based classification may offer comparable or improved accuracy at a lower inference cost. Should speech-based screening scale globally, this distinction will matter for how resources are allocated across modalities. Future work should validate the tonal hypothesis across a broader range of languages, including other tonal languages like Cantonese, Vietnamese, and Thai, as well as other typologies, such as pitch-accent languages like Japanese. Additionally, investigating which specific pitch features are most disrupted by cognitive impairment would contribute to the understanding of the mechanism and clinical applicability of these findings.

7 Acknowledgments

We would like to thank Google for providing the free Vertex AI credits we used to experiment with Gemini on Google Cloud.

References

[1] Felix Agbavor and Hualou Liang. 2024. Multilingual prediction of cognitive impairment with large language models and speech analysis. *Brain sciences*, 14(12):1292.

[2] Alzheimer's Association. 2023. 2023 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 19(4).

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, and 1 others. 2020. Language models are few-shot learners. In *NeurIPS*.

[5] Jiali Cheng, Mohamed Elgaar, Nidhi Vakil, and Hadi Amiri. 2024. Cognivoice: Multimodal and multilingual fusion networks for mild cognitive impairment assessment from spontaneous speech. *arXiv preprint arXiv:2407.13660*.

[6] Yunfei Chu, Jin Xu, Qian Yang, and 1 others. 2024. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*.

[7] S. De La Fuente Garcia, C. W. Ritchie, and S. Luz. 2020. Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease*, 78(4):1547–1574.

[8] K. Ding, M. Chetty, and 1 others. 2024. Speech based detection of Alzheimer's disease: a survey of AI techniques, datasets and challenges. *Artificial Intelligence Review*, 57(12).

[9] E. Edwards, C. Dognin, B. Bollepalli, and M. Singh. 2020. Multiscale system for Alzheimer's dementia recognition through spontaneous speech. In *Interspeech*.

[10] Yan Feng, Gang Peng, and William Shi-Yuan Wang. 2022. Categorical perception of lexical tones in Mandarin-speaking seniors. *Journal of Speech, Language, and Hearing Research*, 65(8):2833–2852.

[11] L. D. Gamble, F. E. Matthews, I. R. Jones, and 1 others. 2022. Characteristics of people living with undiagnosed dementia: findings from the CFAS Wales study. *BMC Geriatrics*, 22(1):1–12.

[12] Gemini Team, Google. 2024. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

[13] Kate Horley, Amanda Reid, and Denis Burnham. 2010. Emotional prosody perception and production in dementia of the Alzheimer's type. *Journal of Speech, Language, and Hearing Research*, 53(5):1132–1146.

[14] Alyssa M. Lanzi, Anna K. Saylor, Davida Fromm, Hui Liu, Brian MacWhinney, and Matthew Cohen. 2023. DementiaBank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*.

[15] Siyu Liang and Gina-Anne Levow. 2025. Tone in perspective: A computational typological analysis of tone function in ASR. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 82–92, Vienna, Austria. Association for Computational Linguistics.

[16] Junfeng Lu, Yuanning Li, Zehao Zhao, Yan Liu, Yanming Zhu, Ying Mao, Jinsong Wu, and Edward F Chang. 2023. Neural control of lexical tone production in human laryngeal motor cortex. *Nature Communications*, 14(1):6150.

[17] S. Luz, S. De La Fuente Garcia, F. Haider, and 1 others. 2024. Connected speech-based cognitive assessment in Chinese and English. In *Interspeech*.

[18] M. Martinc and S. Pollak. 2020. Tackling the ADReSS challenge: a multimodal approach to the automated recognition of Alzheimer's dementia. In *Interspeech*.

[19] Francisco Martínez-Sánchez, Juan José García Meilán, Julia García-Sevilla, Juan Carro, and José M Arana. 2012. Temporal parameters of spontaneous speech in Alzheimer's disease. *International Journal of Speech-Language Pathology*, 14(5):428–436.

[20] Milap A. Nowrangi, Paul B. Rosenberg, and Jeannie-Marie S. Leoutsakos. 2016. Subtle changes in daily functioning predict conversion from normal to mild cognitive impairment or dementia. *International Psychogeriatrics*, 28(12):2009–2018.

[21] Kristin Qi, Jiali Cheng, Youxiang Zhu, Hadi Amiri, and Xiaohui Liang. 2025. Unveil multi-picture descriptions for multilingual mild cognitive impairment detection via contrastive learning. *arXiv preprint arXiv:2505.17067*.

[22] Alec Radford, Jong Wook Kim, Tao Xu, and 1 others. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*.

[23] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[24] Mostafa Shahin, Beena Ahmed, and Julien Epps. 2025. Zero-shot cognitive impairment detection from speech using AudioLLM. In *Interspeech*.

[25] Ravi Shankar, Anjali Bundele, and Amartya Mukhopadhyay. 2025. A systematic review of natural language processing techniques for early detection of cognitive impairment. *Mayo Clinic Proceedings: Digital Health*, 3(2):100205.

[26] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

[27] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

[28] H. Valbret, E. Moulines, and J. P. Tubach. 1992. Voice transformation using PSOLA technique. *Speech Communication*, 11(2):175–187.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

[30] W3Techs. 2026. Usage statistics of content languages for websites. `https://w3techs.com/technologies/overview/content_language`. Accessed: 8 May 2026.

[31] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

[32] J. Yuan, Y. Bian, X. Cai, and 1 others. 2020. Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease. In *Interspeech*.