Research Article

A Novel Implementation of Large Language Model Based Turn Taking Conversational Intelligent Assistance Technology for Seniors

Anda Xie

West High School, 241 N 300 W, Salt Lake City, Utah, 84103, USA

ABSTRACT

Seniors living alone or in nursing homes are often isolated from interpersonal interaction. Existing senior care Intelligent Assistance Technology (IAT) faces challenges, including a rigid conversational structure, a lack of proactive responses, and an inability to address interruptions in a timely fashion. To address these issues, I present a framework that aims to a) develop an Automatic Speech Recognition (ASR) Natural Language Processing (NLP) IAT conversational platform that can parse user speech, analyze speech sentiment, save speech content, and respond with a situationally appropriate tone and content and b) test and implement novel interruption detecting models to simulate authentic conversation. A variety of interruption detection methods were evaluated using the ASR-NLP IAT framework, including facial sentiment analysis, head direction tracking and pupil tracking. The final iterations of this turn taking technology demonstrator involving facial sentiment analysis reached 84.6% accuracy and an F1 score of 0.6. In conclusion, it is proven that ASR-NLP IAT has matured to the phase where it can effectively simulate person-to-person conversation and fluidly exchange conversational roles.

Keywords: Senior Care; Turn Taking; Conversation; Intelligent Assistance Technology; Natural Language Processing; CNN; Computer Vision; Automatic Speech Recognition

INTRODUCTION

In 2020, every one in five people in the United States were seniors. One hundred years ago, in 1920, less than one in twenty were (1).

Despite the geometric growth of the United States' senior population, senior care professionals have reported alarming stagnation in care resources (2, 3). Seniors often face a stark reality - one characterized by isolation from meaningful interpersonal interactions, limited assistance, and a growing vulnerability that compounds with the increasing aging population and stagnation of available support resources. Seniors within assisted living environments are not immune to these negative trends either. Declining wages for senior care workers have created an underpaid, overworked industry that struggles to provide for its existing consumer base. This problematic status quo not only poses significant challenges to the well-being and quality of life of senior citizens but also to the wider community. An increasing senior dependency ratio has consistently proven to increase social security

Corresponding author: Anda Xie, E-mail: daguo2017@gmail.com.

Received April 1, 2025**; Accepted** April 24, 2025
https://doi.org/10.70251/HYJR2348.3296109

maintenance and familial senior care costs (4). With the severity of difficulties an aging society poses to seniors and the greater community, this research examines if utilizing a technological solution to reduce senior isolation and care costs is possible.

Efforts have been made in the past to address these challenges through the development of Intelligent Assistance Technology (IAT). IAT has three main advantages: economic viability, uninterrupted monitoring, and enhancing the quality of senior care. Not only as an economic alternative to nursing homes costing an average of one hundred thousand dollars annually nationwide, IAT also has the potential to assist seniors in a timelier fashion than caregivers due to its uninterrupted presence (5). The implementation of IAT in senior care can also reduce the workload of overworked caregivers without any threat to job security, a threat commonly attributed to postindustrial automation.

Existing IAT systems have a serious set of flaws stemming from their rigid response structures (mostly based on archaic logic trees), limited conversational proactivity, and an inability to respond to the natural flow of conversation (3). Pre-built IAT conversational structure does not provide seniors with the familiarity of natural conversation due to its rigidity, making improvements to social isolation unlikely to materialize. The inability to detect interruption in existing IAT also hampers its natural conversational ability. Interrupting, defined as the process of exchanging conversational roles of speaker and listener, is neglected in standard IAT technologies. This lack of functionality can cause users to lose interest in the conversation and lose valuable time in responding to emergency situations (6).

To fully tap into the potential of this emerging technology, this research project will implement an Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) IAT framework further enhanced with interruption detection capabilities, to provide seniors with natural and appropriate support. ASR-NLP word processing Artificial Intelligence (AI can mitigate the issues of rigid frameworks due to its ability to generate original conversation. With this framework at its center, this project will create methods to discover when seniors desire conversation and wish to interrupt proactively and accurately.

This paper will delve into the methodology, results, and discussions behind a comprehensive exploration of ASR-NLP IAT. This paper will also explore the psychological framework of turn taking, its physical manifestation and the results of models built on finding those signals.

LITERATURE REVIEW

The structure of natural language was researched due to its foundational role in communication, key to reducing isolation in seniors. This area of research was chosen because of the significant role that turn taking through interruption plays in natural human conversation and research gap in NLP applications. While proper turn taking facilitates understanding between listener and speaker, a lack of turn taking mechanisms may cause severe frustration between the two parties due to hampered understanding (7). With a lack of interactivity being one of the main reasons why senior care IAT did not garner significant success, proper research into this field is required to tap into its potential (2). The author classified the studies consulted regarding natural language under three categories: timing and appropriateness in turn taking, interruption cues and challenges to successful conversational turn taking.

Timing and Appropriateness in Turn taking

Turn taking itself relies heavily on timing and appropriateness when exchanging the roles of listener and speaker and is highly important to maintaining the information transfer in conversation. Within this field, Nie and Guo as well as Nguyen et al. were consulted (8, 9). Highlighting the importance of appropriate interruptions in preserving conversation quality, standard interruption time periods and the inherent rhythmic pattern of conversation, these studies proved instrumental in identifying successful timing within interruption detection technologies implemented in this project.

Nie and Guo's study raises a crucial point regarding the timing and appropriateness of turn taking in conversations (8). Studying the relay of information among groups through conversation in the format of a memory test, Nie and Guo found that traditional turn taking in group conversation, when relaying information, did not provide substantial benefits to memory compared to individual work. In fact, a significant drop in memory accuracy was detected among groups. This drop highlights the importance of considering when and how turn taking occurs, emphasizing the need to avoid disrupting the flow of thought during conversations. This finding resonates with the notion that turn taking should be strategically timed to maximize its benefits – not interrupting trains of thought and nudging minds in a mental standstill.

Nguyen et al.'s systematic review and Bayesian meta-analysis of turn taking explored the natural interruption patterns in adult-child vocal interactions, suggesting that

interruptions typically take around one second (9). Among other age groups (including seniors), it was found that this one-second gap remained consistent (except among neurodivergent populations that exhibited a one-second extension to this phenomena). With this pattern in mind, a question is raised for the design of synthetic systems replicating conversation on whether this gap should be preserved. While preserving this natural gap may create more natural-sounding conversation and increase users' familiarity with this system, this gap could also slow the conversation pace in time-sensitive use cases. Future testing after this literature review will seek to identify a balance between the two possibilities to increase an understanding of the natural rhythm and generate the most efficient form of conversation.

Interruption Cues

Consciously or not, the desire to interrupt and exchange roles of speaker to listener in human-to-human conversation is primed by a variety of cues. To be able to identify these cues properly is key to creating a synthetic replication of human interruption detection. Kendrick et al. confirm the existence of facial cues and hand gestures in signaling the desire to interrupt and provide a comprehensive list of these cues (10). Preisig et al., as well as Degutyte and Astell, detail the specifics of anticipatory eye gazes in determining the desire to interrupt and how external factors like lexico-syntactic information, social status, and conversation format determine an eye gaze's true meaning (6, 11). Dawson and Foulsham detail an interesting phenomenon that, while previous patterns observed around eye gaze still held true for human-to-computer interaction, humans tended to have more extended periods of eye-gazing with screens compared to human-to-human conversation (12).

Kendrick et al. confirmed the presence of multimodal cues, such as hand gestures and gaze direction, in indicating the desire to interrupt during face-to-face interactions (10). For example, a raised hand would indicate the desire to question or comment on existing speech (as demonstrated in primary schools). Apart from socially weighted physical cues, motions and rapid eye gaze shifting may indicate user agitation – a common precursor to interruption. As observed with earlier studies that concluded that interruption was often mentally preceded by speech preparation and increased mental activity, this increased agitation was naturally displayed through increased hand gestures and agitated eye gaze shifting. These non-verbal cues play a pivotal role in the coordination of turn transitions, emphasizing the significance of incorporating visual and gestural elements in AI systems' interaction interfaces.

Preisig et al.'s study highlighted the role of anticipatory eye gaze shifts in conversational turn transitions between speaker and listener (6). Often, increased eye contact occurs prior to transitioning, indicating attention/affirmation to a perceived conversational transition or trying to signal to the speaker the desire to interrupt. Once again, the theme of agitation is invoked to explain this phenomenon. While eye gaze shift is a customary practice, Preisig et al. cautioned against relying solely on gaze shifts for determining turn transitions and recommended considering lexico-syntactic (verbal) information instead. However, this suggestion is considered but not implemented due to the conversational structure of the chat-bot having no way to parse added information when it is speaking. This insight underscores the need for this system to monitor the eye gaze of the subject yet not use it as a sole determinant in the desire for user interruption.

Degutyte and Astell's study examined the role of eye gaze in regulating turn taking, emphasizing that the intensity of eye contact increases as a turn approaches its end (in a question-answer conversational format) (11). Degutyte and Astell were able to summarize two opposing theories regarding the influence of eye gaze in determining the exchange of conversational roles. It was theorized in the past by two separate research schools that gazing towards or away from the listener determined the desire to interrupt. Degutyte and Astell were able to summarize that in question-and-answer format gazing away indicated the desire to interrupt, while the opposite was true in a regular conversational format. Degutyte and Astell concluded deviations in behavior observed in earlier studies were caused by differences in conversation format. Apart from conversational format, speaker and listener's social status and societal norms were also found to influence the degree of eye contact. With the cultural significance of eye gaze varying across these scales, some cultures observing eye contact as a form of respect while others observing it as a challenge to authority, Degutyte and Astell emphasized the importance of applying information in its proper context. This insight underscores the complexity of non-verbal communication and its relevance in AI systems' design, as well as the importance of implementing said system properly.

Dawson and Foulsham noted a phenomenon in modern interactions where individuals tended to look at screens rather than each other during conversations (12). While previous findings in eye gaze were confirmed, this new finding may cause slight deviations with the expected

eye gaze state of the user. Using a statistical approach to analyzing a user's gaze patterns, to remove margins of error created by this observed phenomenon, could be a way to prevent this project from running against established research. This finding underscores the importance of considering the role of eye gaze and social attention in the context of technology-mediated interactions.

Challenges to Successful Turn taking

Implementation of an interruption detection system complementing ASR-NLP requires due diligence and consideration for its success, as demonstrated by the research below. Bögels and Levinson introduced their research into the physical characteristics of natural speech preparation, not only providing a framework for identifying these characteristics but also posing the challenge of how to properly replicate these signals in computer-to-human interaction (13). Donnarumma et al. outlined the importance of minimizing speech overlap between speaker and listener for proper information transfer and posing the challenge of replicating this behavior for an ASR-NLP system (7).

Bögels and Levinson's research delved into the process of speech preparation during turn taking (13). By using ultrasound measurements to measure cerebral activity for speech preparation, their findings indicated that speech preparation began immediately upon hearing crucial information, accompanied by heightened stress, physical indicators, and in-breaths. While confirming previous research into physical indications of interruptions, this finding posed the question of how these indicators can be detected by a computer system.

Donnarumma et al. emphasized the importance of minimizing periods of overlapping speech in conversation for information transfer (7). Donnarumma found that most conversations had as little overlapping speech as possible, with more overlap leading to greater losses of information and interactivity. Action prediction emerged as a critical aspect in this study, aligning with the development of AI systems that anticipate and respond to user intentions seamlessly. Information overlap posed a challenge of creating an AI system that can properly track a user's desire to interrupt.

In summary, turn taking in conversation is crucial for the development of authentic synthetic ASR-NLP systems, particularly those aimed at aiding seniors and individuals with unique communication needs. These insights encompass the timing and appropriateness of turn taking, natural interruption patterns, multimodal cues, the role of lexico-syntactic information, minimizing overlap and silence, speech preparation challenges, and the significance of eye gaze and social status in determining interruptions. AI systems can provide more natural and effective interaction experiences for users by integrating these qualities and considerations.

MATERIALS AND METHODS

The creation of a system based on Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) technology to respond to a senior's request naturally and appropriately for conversation, emergencies, entertainment, and lifelong learning is proposed in this solution (Figure 1). To begin, an Automatic Speech Recognition (ASR) system parses speech from the user to determine if they wish to request a service. Unlike existing services, this system operates in the background of a user's environment in a non-disruptive way. It determines if it should intervene based on the sentiment detected from musings. Simultaneously, a Face-mesh Computer Vision (CV) system captures images of the user to identify sentiment and potential needs (like the wish to interrupt) using a convolutional neural network (CNN). This data is sent to a cloud storage-base and automatically processed by a state machine to identify the state of the senior and react accordingly. The state of the user, as well as their speech, is piped to a large language model (LLM) that returns a text response and keeps a transcript to create personal profiles of users. This response is then output as speech via a Text-To-Speech (TTS) system to the senior in a conversational way, which allows them to respond. To maintain a timely response for this system, this code is

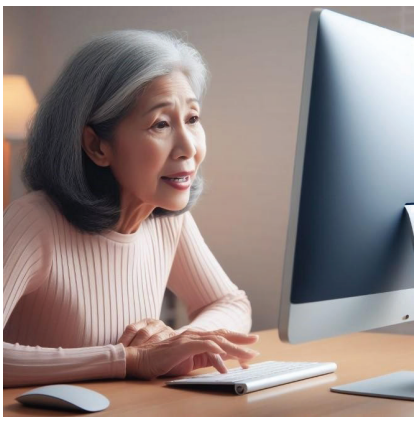

Figure 1. This graphic demonstrates the desired use case of this service – involving a senior in front of a computer screen. Credits to Bing Image Generator.

offloaded among multiple devices in the spirit of parallel computation (Figure 2).

Concerns and Considerations

While ASR-NLP was chosen as the backbone of the IAT framework used for this project, other technologies were considered and rejected for determining senior sentiment. Voice Stress Analysis (VSA) was explored as a potential way to determine user sentiment. This method was rejected due to established VSA frameworks being inherently unstable and pseudoscientific and novel deep learning-based methods being too data-heavy to respond in a practical timeframe. Deep learning models were not used in the Computer Vision (CV) portion of this project (14, 15). Haar Cascade to CNN formats were adopted instead for the sake of data processing speed. Apart from its current sentiment analysis role, CV systems could also be trained to recognize specific actions signifying emergency situations, like a fall, in future iterations of this project involving more computationally robust hardware.

Non-technical concerns have also been properly addressed in this proposal. Firstly, senior autonomy – a repeated theme among ethics specialists in IAT, has been ensured in this program by hard-coding the ASR-NLP service to respect senior boundaries (16). Coupled with the fact that this service only provides advice and has no physical manifestation, there is a clear boundary between senior and service. Secondly, ASR-NLP's open-source framework also maintains the economically affordable qualities of IAT. This affordability ensures that innovation does not come at economic inaccessibility. Thirdly, job security for senior care workers is improved by ASR-NLP due to its purely virtual qualities requiring real-world professionals to manage physical maneuvers. Time saved in this way gives professionals time to engage in meaningful activity with seniors. Lastly, the ASR-NLP framework can build a user profile while maintaining privacy using state-of-the-art secure databases.

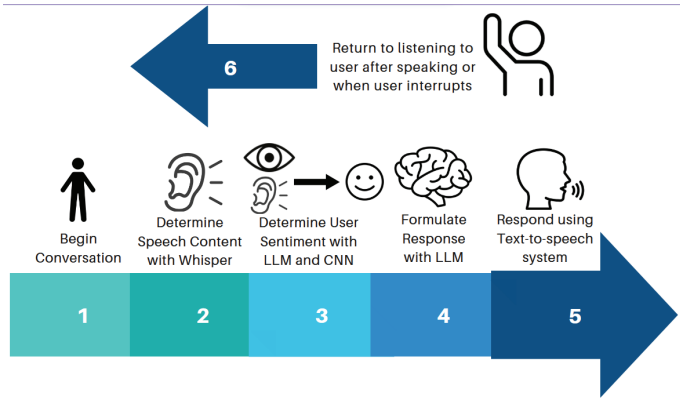

Figure 2. A flowchart of the working process of the ASR-NLP system. The system listens to the user using Whisper API, and then determines the emotional state of the user using visual and speech clues to generate an appropriate response. This response is formulated with an LLM and output with a text-to-speech system. While this main line functionality occurs, a background program monitors the user to see if they display any desire to interrupt the current conversation. If so, all output speech is paused, and the program proceeds to listen to the user.

Most of the research within the literature review relied on human input through manual labeling to detect cues for interruption. This form of labeling, although able to create a ground truth, cannot be used in a timely manner for conversational purposes. Interruption cues would be identified using a CNN to balance speed and accuracy to derive these signals in real-time. Without adding more hardware to the ASR-NLP system, the most logical next step would be to investigate the use of a device's built-in camera to detect facial manifestations of the desire to interrupt. With a variety of physical interruption cues present on the face, and the face being the natural source of sentiment detection among human-to-human conversation, it seems the ideal platform to begin with and expand upon. Monitoring the face can simultaneously facilitate the change of the user's mouth movements and state, open or closed. Mouth movement as a prerequisite for speech ensures that a facial monitoring position can identify when the user is speaking. Lastly, thorough research has been done in facial interruption detection, ensuring the scientific backing of detection methods.

Tested Interruption Detection Methods

After a review of theoretical and practical research of the use of technology in the detection of conversational turn taking, the following portion of this project focused on implementing technology using an ASR-NLP framework. Specifically, three methods of facial visual sentiment detection were tested: Haar Cascade CNN facial expression classification to classify facial sentiment through a collection of labeled data of "normal" and "agitated" faces, pupil tracking to determine user agitation and distraction (due to either a lack of interest in the conversation or a desire to add more information to the conversation) and head direction tracking to determine user distraction. These models were also chosen due to their sole reliance on facial signals to complete their classification. Existing classification support models, such

as the Haar-cascade eye tracker for pupil tracking, could also be implemented in parallel after slight tweaks and without requiring boilerplate coding. This improves the speed and efficacy of testing.

Haar Cascade CNN. The Haar Cascade CNN facial expression classification method uses labeled data to find hidden and existing patterns between normal faces and interrupting faces. To begin, a dataset of two thousand images consisting of normal faces (any face without a desire to interrupt) and interrupting faces (faces expressing agitation or boredom by showing a (large) open mouth, heavily furrowed eyebrows, a raised hand (near the face) or raised eyebrows). This dataset was constructed by taking a one-minute video of the researcher under both non-interrupting and interrupting facial conditions, spliced into individual frames at a rate of 20 Frames Per Second (FPS). This data is used as the training dataset for a CNN model, with its algorithmic architecture shown below, and is saved as an .h5 model file and imported to a Colab Notebook once finished. Prior to using this model, the Colab notebook utilizes a JavaScript extension to take an image of the user which is sent to a Haar Cascade based Frontal Face classifier to identify the user's facial area. Afterwards, the preloaded CNN model classifies the facial sentiment and saves this sentiment within the shared cloud database.

Pupil Tracking. Pupil tracking has similarities to facial expression classification, but with some fundamental differences. The model begins by using a pre-built Haar Cascade face identifier to create a smaller frame for detecting user eyes. Afterwards, another Haar Cascade eye tracker processes the image to identify user's eye position. The program isolates a single eye (left by default) and crops out the rest of the photo. This finalized image of the user's eye is manipulated by cutting out user eyebrows and using the "blob" function to amplify the pupil of the user. Based on the proportions shown in the graphic below, the user is identified as looking right, center, or left. This data is then analyzed to determine if the user is distracted by looking away. Among humans, apt attention is expressed through looking directly at the subject being spoken too (the listener). Therefore, the conclusion can be made that the opposite of staring at the listener – looking away, can be interpreted as a lack of focus/interest in the conversation engaged. This phenomenon of looking away signaling distraction is further amplified by the human tendency to stare at screens for extended periods of time (Figure 3).

Head Direction Tracking. Head direction tracking is the most straightforward and simple model of the three tested – its main feature and detractor. While the simplicity of this model allowed for the quickest and least data-heavy classification, it could only analyze rare edge cases of facial direction changes in determining interruption. The preliminary model extended the previous pupil tracking program. When the device could not detect pupils, it was assumed that the user was facing away (enough to only show one eye) and distracted. However, another method was considered for this classifier – the use of the Haar Cascade side face identifier. This classifier method was not implemented in this preliminary model of testing due to a lack of awareness toward its properties but was considered for use in later iterations.

Rejected Detection Methods

Other detection methods considered were rejected during this stage of testing due to their reliance on external hardware. However, research has shown promising results with their implementation. In this category, the two main methods considered were heart rate tracking and ultrasound activity measurement. Heart rate monitoring was demonstrated as successful in determining a user's increased agitation and internal speech preparation, as demonstrated by the above studies in turn taking. Future use of a heart rate tracker or developing models of optical facial heart rate tracking (a technology that relies on slight shifts in the user's facial coloration in detecting their pulse) can be pursued to determine this factor. Ultrasound activity measurement was also determined as successful when discovering internal speech preparation. However, no reliable implementation of this technology was able to be implemented and is left as an opportunity for future research.

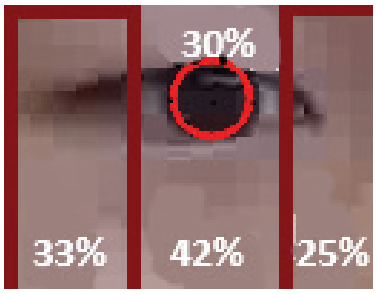

Figure 3. This image shows the proportions of eyes considered in the pupil tracking. If the pupil is present in the leftmost one-third and one-fourth of the image, the user is considered as looking away. Pupil tracking is completed by using a CNN to identify the position of the pupil within one of these three areas.

Continuous/Discrete Models

All models tested accepted image input as a basis for classification; however, "continuous" classification-based models were implemented and tested in this phase of testing to compare their accuracy with their discrete counterparts. A continuous model is defined as a model that captures a series of images of the user in a short period of time to analyze rather than analyze a single image only (as done with discrete models). This route was pursued due to concerns of various physical fluctuations of the user facial expressions compromising the accuracy of single-frame analysis models. Continuous classification was accomplished, therefore, by feeding multiple frames of the user's face into the continuous model in a single instance. Each of these frames would be analyzed and have its sentiment classified. The final sentiment would be output as the one most common among these classifications.

This ASR-NLP framework utilizes the Python programming language in Google Colab due to its easy-to-start structure, providing the most flexible codebase, its ample collection of AI/ML libraries (e.g. TensorFlow), and its online availability in the form of Intelligent Python Notebook (.ipynb) services (Google Colab) (Figure 4).

Specifically, this framework will follow the modular structure detailed below:

1. Speech Sentiment Determination. To begin a conversation, this system assumes the role of "listener" with an Automatic Speech Recognition (ASR) system – OpenAI's Whisper AI. This ASR system parses the audio of a user's background as input before isolating any speech detected. Afterwards, the speech content is transcribed and saved in a cloud database before being analyzed by a built-in NLP sentiment analysis platform. The NLP platform tested was the GPT-3.5-Turbo API model due to ease of access. Transcripts of the conversation and the inferred sentiment are classified by purpose and stored in a cloud database (Figure 5).

2. Sentiment Recognition through Computer Vision (CV). In conjunction with the ASR-NLP pipeline, this system employs a Computer Vision (CV) based sentiment identification system. This CV system assists in identifying the sentiment and non-verbal cues of seniors. It works concurrently with the ASR system to provide a comprehensive understanding of the user's state. Considering the nuance of language, a map of the user's facial expression helps give this machine a greater understanding of states like "sarcasm" or double-

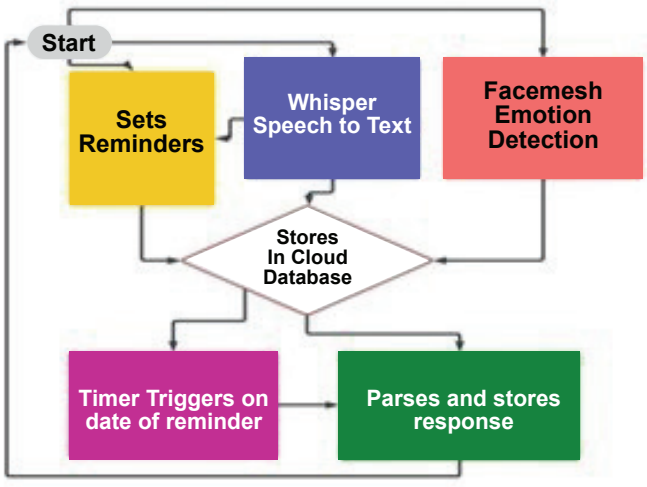

Figure 4. This flowchart shows the logical flow of this ASR-NLP system. It begins with ASR speech recognition and visual sentiment reaction, followed by a reaction to previous data.

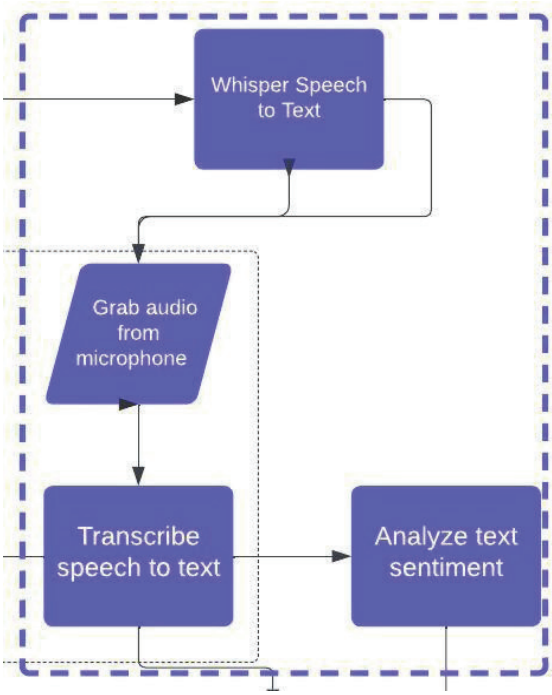

Figure 5. This flowchart demonstrates the logical flow of this project's speech sentiment determination phase. In this portion, the device first parses audio from the device's microphone, transcribes this text using OpenAI Whisper speech-to-text, and then uses a large language model as a sentiment analyzer to determine the sentiment of this speech.

meanings that would be missed with solitary ASR-NLP.

The identification of the desire to interrupt is also present in this portion. Code has purposefully been made malleable to allow for testing of a variety of models identifying varying nonverbal cues that could point to the user interrupting. A major portion of this project will be concentrated in this section to determine what specific nonverbal cues impact conversation. Both the user's overall sentiment and the desire to interrupt are also detected and stored in a cloud storage-base (Figure 6).

3. Cloud-Based State Machine. Classifications of the user's state are then stored within a cloud-based storage-base housing a state machine. This state machine analyzes the user's sentiment, desire to interrupt, and situational context to make informed decisions. It continuously processes data piped over from ASR and CV systems and reacts accordingly to provide assistance tailored to the user's emotional state. Apart from constantly processed data, this state machine also houses permanent personal data and serves as an anchor for model improvement and quality control. Permanent personal data, such as the user's name, is stored and easily accessed by the ASR-NLP service to create more personalized conversations. Providing a transcript of user conversations serves to prevent model hallucination and identify repeated questions for model improvement. Model hallucination is one of the most significant sources of reluctance to apply ASR-NLP in healthcare. To prevent hallucinations like this, the cloud database is connected to another ASR-NLP service that monitors the main GPT pipelines for problematic responses. Transcripts of the GPT-user conversation are also stored in this database to identify repeated questions to create more advanced iterations of this service.

4. NLP Output Generation. User speech and situational information are piped from the state machine to the ASR-NLP framework. ASR-NLP uses contextual understanding to generate a text response that is not only contextually relevant but also emotionally attuned to the user's sentiment and desire to interrupt. For example, a user looking sad and musing about "wishing to chat with someone" would receive a response of, "Sure, I am here to help! What would you like to know or talk about regarding [interest]?" from the GPT service. Reminders for time-based events like appointments set in the previous sections are also output in this section (Figure 7).

5. Conversational Delivery. To ensure a natural and conversational interaction, a Text-To-Speech (TTS) system delivers the GPT-generated response to the senior user. After receiving the ASR-NLP generated response, this TTS system will split the passage with dividers as significant conversational pauses. These speech dividers (small pauses based on the total length of the ASR-

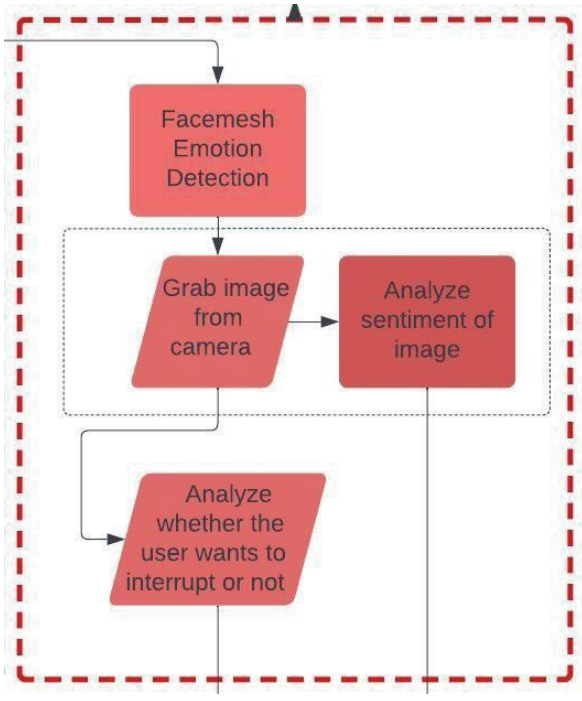

Figure 6. This flowchart demonstrates the logical flow of the Computer Vision portion of this project. In this portion, the device first takes a picture of the user using the device's built-in camera. This image is then analyzed for signs of interruption as well as general emotion.

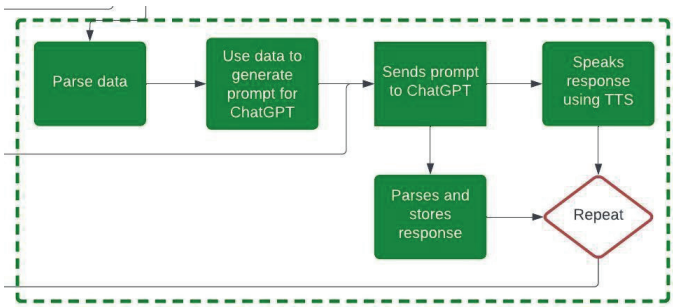

Figure 7. This flowchart describes the process of reacting to previously stored data. User speech sentiment, speech, general emotion, as well as desire to interrupt, are first parsed from a cloud database before being combined into a text prompt. This prompt is then fed to a GPT-3.5 service via API, which returns a generated response. This response is output via TTS and stored in a cloud database.

NLP response) will be placed where the passage has punctuation present. Between each pause, this code piece will check the state machine to determine if the CV-based interruption detection portion of the code has detected if the user wants to interrupt – like in human-to-human conversation (7). By doing so, speech overlap between the user and the computer is minimized, preserving conversational quality.

This conversational framework seeks to replicate human conversation as accurately as possible. Starting as the listener, this system exemplifies human conversational qualities of emotional detection, response generation, and response output (13). Beyond their significance in interruption detection, conversational pauses in this framework also improve understanding among seniors with hearing impairments by giving them more time to react to information presented. This feature is key to serving the one in four seniors having hearing impairments (17). With a proper conversational framework created, interruption detection tests can proceed with the created framework.

Preliminary testing was completed to remove any unusable models and find gaps in model accuracy. For conversation data, a two-minute, person-to-person conversation was recorded and split into frames. Every five seconds, a collection of three frames was stored and sent to the trained model for a total of twenty-six instances. Five seconds was chosen as the time between each collection to prevent overloading the model with data. Among these twenty-six instances, six showed the user interrupting with a variety of interrupting expressions – misaligned pupils [2], opened mouth [2], and head facing away [2].

RESULTS

The following data tables represent the results of the models tested. The top right corner represents a confusion matrix. The bottom statistics are measurements of model traits, accuracy being the number of correct predictions over total predictions. False positives are defined as when the user was classified as interrupting when the user was in a normal state, while false negatives were defined as the opposite case. This data is expressed as a percentage with the total number of false negatives/positives compared to the total number of tests conducted. Model precision was defined as the number of interruptions correctly classified over the total number of interruptions detected. Model recall was defined as the total number of interruptions predicted over the number of correct predictions. These two metrics were combined in an F1 score, two times the inverse of the sum of the recall and precision, to better indicate the predictive performance of the tested model (Table 1, 2, 3).

DISCUSSION

Of all the models presented, the model based on facial expressions appeared to be the most successful based on average accuracy and F1 score. However, this model was consistently unable to classify the user with an open mouth, potentially due to a lack of substantial data of a user with an open mouth in its training dataset. This model also had a tendency toward identifying false positives over false negatives. This would lead to conversations where interruption cues were unlikely to be left ignored, yet

Table 1. Facial Expression

Facial Motion - Discrete		**Actual**		**Facial Motion**	**Discrete**	**Continuous**
		Normal	Interrupt	**Accuracy**	84.60%	84.60%
Predicted	Normal	19	1	**False Positives**	11.50%	11.50%
	Interrupt	3	3	**False Negatives**	11.50%	11.50%
				F1	0.6	0.6
Facial Expression - Continuous		**Actual**				
		Normal	Interrupt			
Predicted	Normal	19	1			
	Interrupt	3	3			

This data represents the confusion matrix and analysis of the facial expression-based model when detecting interruption cues. The discrete and continuous model produced a 84.6% accuracy, potentially due to human facial expressions not having large fluctuations in short time periods normally.

overall conversational quality would have more frequent interruptions. On a positive note, the similarity of results between continuous and discrete classification shows that human emotion does not widely fluctuate during brief time periods and that continuous classification is not necessary for this model. Although reaching an acceptable threshold of accuracy – greater than 80% of classifications – this model can still be improved with more training data.

Pupil tracking has produced positive initial results also reveals potential issues. In faces where the eye was visible to the camera, pupil tracking was consistently able ($x > 90\%$) to identify the eye and its corresponding pupil position. However, this model was hampered by blinking. Whenever the eye was closed, this model misclassified the user as looking away and therefore considered the user distracted. This may be the major factor in why continuous models (which were more likely to detect blinking) have a suboptimal accuracy compared to their discrete counterpart. Although not an issue in this test, another potential problem is the rapid fluctuation of human eye motion, leading to inconsistent data. To mitigate this, it is proposed that all frames without a face or eye detected to be not analyzed in future iterations of continuous models.

Head direction tracking, dependent upon pupil tracking, had its results heavily influenced by the accuracy of the above model. Therefore, the single-shot head direction model had the same accuracy as its pupil tracking counterpart while the continuous-classification model exhibited high rates of misclassification. To prevent further errors caused by mutual dependency, it is proposed head direction models rely on the Haar Cascade side-face classifier, which was specifically built for tracking

Table 2. Head Direction

Heading - Discrete		**Actual**		**Head Direction**	**Discrete**	**Continuous**
		Normal	Interrupt	**Accuracy**	80.80%	53.80%
Predicted	Normal	18	2	**False Positives**	11.50%	42.30%
	Interrupt	3	3	**False Negatives**	7.70%	3.80%
Heading - Continuous		**Actual**		**F1**	0.55	0.455
		Normal	Interrupt			
Predicted	Normal	9	1			
	Interrupt	11	5			

This data represents the confusion matrix and analysis of the head direction-based model when detecting interruption cues. The discrete model produced a statistically significant increase in accuracy over the continuous model.

Table 3. Pupil Tracking

Pupil Tracking - Discrete		**Actual**		**Pupil Tracking**	**Discrete**	**Continuous**
		Normal	Interrupt	**Accuracy**	80.80%	73.10%
Predicted	Normal	18	2	**False Positives**	11.50%	23.10%
	Interrupt	3	3	**False Negatives**	7.70%	3.80%
Pupil Tracking - Continuous		**Actual**		**F1**	0.55	N/A
		Normal	Interrupt			
Predicted	Normal	19	1			
	Interrupt	6	0			

This data represents the confusion matrix and analysis of the pupil tracking-based model when detecting interruption cues. The discrete model produced a statistically insignificant increase in accuracy over the continuous one, and warrants future research.

head direction, instead of tracking the number of pupils detected.

To further improve these models, specific testing was completed on revised versions of the single-shot models. The facial emotion recognition model received more training data. The head direction tracking model tested the Haar Cascade side-face classifier. The continuous pupil tracking model only treated frames with a face and pupil identified as significant. These tests will focus solely on one aspect of interruption – open mouth, facing away, and consistently shifting pupils.

Following the analysis above, there is a good reason to cut out all continuous models due to their lack of accuracy coupled by greater data processing costs. An average drop of around 10% accuracy for continuous compared to discrete models was observed. Given a sample size of 26, a two-sample z-test for population proportion difference for the head-direction tracking model gives a p-value of 0.019. This p-value is created from the null condition that the models perform equally well, and an alternate condition the discrete model has a higher average accuracy than continuous model. At a significance level of 0.05, this provides convincing evidence to overturn the null condition and convincing evidence discrete models perform more accurately than continuous models. Continuous models also took around one second longer than their single-shot counterparts – a noticeable gap in natural conversation. This difference may be due to inherent model instability coupled with users always blinking during this frame interval which lowered model accuracy. However, a two-sample z-test for proportion difference for the facial expression and pupil tracking models give p-values of 0.5 and 0.26, failing to provide statistically convincing evidence to overturn the null condition. Future research could be done on evaluating the usefulness of continuous models in regard to facial emotion tracking and pupil tracking models.

Further testing was completed following the same methodology of the previous experiment by analyzing a video of a conversation at 20 FPS with a sampling rate of 5 seconds. Instances of when the user desired to interrupt were manually selected and used for the evaluation of the older and newer iterations of the tested models. This method of testing for the detection of the desired characteristic solely expedited testing speed.

After testing, an average increase of up to 20% in overall model precision can be observed. All newer models reached a threshold of 80% precision. It is believed further improvements could still be achieved through the implementation of more complex classification methods beyond the Haar Cascade and an increase in CNN training data. However, this method may come at the expense of processing speeds due to the advanced neural networks presented by state-of-the-art models far exceeding that of the binary operation of the Haar Cascade, hampering conversational speed. This proposal is left as an avenue for future research (Figure 8).

While all improved models have reached a higher accuracy than their previous iterations, the current sampling rate at which they were tested on (one sample every five seconds) warrants further testing due to the presence of changes in expression observed between each sample. The initial sampling rate was put in place due to Google's Cloud Database services having built-in DDoS protection, limiting the number of API calls to 60 GET/POST (read/write) calls per minute per user. The 5-second sample rate was selected to maintain a buffer between API calls to prevent the model from crashing.

An internal buffer within model code-pieces was considered to improve the sampling rate of these models, and therefore the system's overall accuracy. This buffer would sample frames of the user at one frame per 0.5 seconds, a 1000% increase from previous iterations and append these results to a local database for five seconds. Once five seconds have passed, the data would be filtered to remove any instances without eyes present. Afterwards, the most common sentiment would be analyzed from the dataset and stored within the cloud database, maintaining a buffer period while creating a more complete picture of the user's sentiment.

This model was first tested for the total number of additional interruption instances detected under ideal

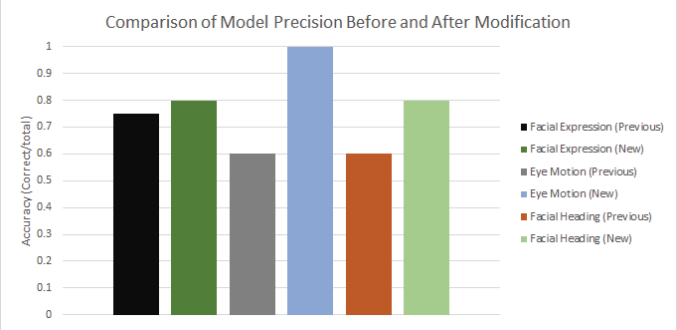

Figure 8. This bar graph displays the change in accuracy among AI/ML models in their preliminary to current iteration. The facial expression model had a 5% accuracy increase, the eye motion model has 40% and the head direction model has experienced an increase of 20%.

conditions with an improved sample rate. This was tested through analysis of the same two-minute conversation clip used during the preliminary testing of AI/ML models used in this research. Shown above in Figure 9, an additional two cases of interruption missed by a five-second sampling rate have been observed using a higher sampling rate and identical reporting rate proposed in the above experiment. With this data, it can be observed that an updated sampling rate has a significant increase in a model's ability to report interruption, up 30% in this instance (Figure 9).

Future improvements in terms of updating the sampling and reporting rate were also considered following this round of experimentation. Between each five-second POST request to the cloud database, it has been observed that the desire to interrupt had already existed for around 0.5-3 seconds prior to each reporting instance. Compounding this additional interruption time with the five seconds it took to report this observation to the state machine, a 5.5-8 second gap would be created as a significant delay within the conversation. A five-second reporting rate would also eliminate any intent of interruption from the user lasting less than 2.5 seconds. This phenomenon was observed at least two times within the two-minute sample conversation. To prevent these interruptions from being ignored, it is proposed to separate the state machine responsible for GET/POST requests to be split into two entities - one responsible for the ASR-NLP IAT and its transcript, and the other responsible for storing the user's desire to interrupt or not. This would reduce the strain on a single project's API load and allow for the increase to a reporting rate to 2.5 seconds.

Throughout this project, prototype creation and testing took place on a Personal Computer (PC). While this allowed for a simple interface for programmers and a strong computational base for testing novel features, the PC would be unsuitable for the project's eventual implementation beyond the testing phase. Due to the large configuration of the PC model (Lenovo ThinkPad) and a 5-10 minute startup process for all code pieces on online servers, the current configuration can only confirm the validity of ASR-NLP in creating a turn-conscious IAT system. Mitigation of these problems should prioritize portability, computational speed, and a user-friendly interface for creating more personable environments for senior users compared to the research-oriented PC design. This project's embedded system (ES form was created with these needs in mind, with computations (collecting user visual and audio data) completed on a microcontroller. The lightweight ES configuration can perform all startup activities in a single session, with greater packageability due to its smaller size.

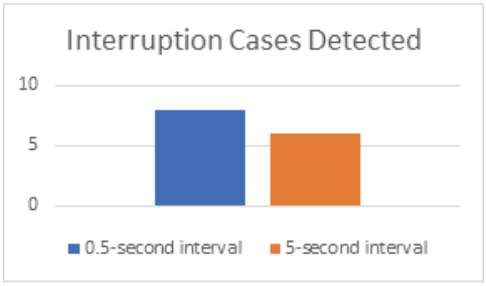

Figure 9. This graph compares interruption cases detected among sampling rate intervals. Within a 2-minute conversation, two additional interruption cases were detected using a 0.5-second interval.

The microcontroller configuration, tested on a Raspberry Pi (RPi model 4B 8Gb board, is more efficient and occupies only one ninth of the volume (e.g. space) compared to the previous Lenovo ThinkPad used for research. This board, chosen for its user-friendly configuration and compatibility with Colab's existing software on an Ubuntu Linux Operating System (OS, features a Google Coral TPU (Tensor Processing Unit hardware accelerator for offloading internal calculations needed for CV-based emotion recognition. This hardware accelerator system was added to address the computational power loss when transitioning from a PC to ES system to preserve processing speed requirements. The board is equipped with a camera and microphone system to parse user speech and sentiment. Basic sentiment calculations are offloaded onto the attached TPU, and the information is stored in the project's common cloud storage. This data is then sent to another processing unit (a standard PC sufficesor directly to the RPi for synthesis into a prompt for a language model (LLM) service. While the prompt synthesis occurs offline on the hosting servers of the LLM service, the RPi microcontroller handles sending and receiving input/output to and from the online LLM service. This configuration significantly reduces the lengthy startup process of the audio and video collection and analysis portion of the project, thanks to the connection of an external TPU.

CONCLUSION

The pressing issue of senior isolation, whether in solitary living or within nursing homes, necessitates

effective solutions to bridge the gap in interpersonal interaction. The growing aging population and limited support resources accentuate the vulnerability faced by seniors. Recognizing the potential of senior care Intelligent Assistance Technology (IAT), this research sought to address the shortcomings of existing technologies characterized by rigid conversational structures, a lack of proactive responses, and inefficiencies in handling interruptions during conversations. This project accomplished two key objectives: firstly, the creation of a proactive Natural Language Processing (NLP) Generative Pre-Trained Transformer (GPT) IAT capable of accommodating diverse user needs; and secondly, the development and implementation of innovative interruption detection technology to emulate natural conversational dynamics. Through experimentation and testing, an ASR-NLP IAT was successfully integrated with an interruption detection framework, paving the way for true-to-life conversation simulations. As demonstrated by the final prototype, ASR-NLP IAT has evolved to a stage where it can effectively replicate person-to-person interactions, offering a promising solution to address the challenges of senior isolation and vulnerability. Research into turn taking has resulted in the creation of a model with an average of 85% precision in detecting the desire to interrupt – creating a testbed as proof for the potential application of ASR-NLP IAT in senior care. Such a comprehensive system can serve as a base for future research in senior care using large language models.

ACKNOWLEDGEMENTS

The author would like to thank Professor Clark Hochgraf of the Rochester Institute of Technology in his mentorship of this project.

DECLARATION OF CONFLICT OF INTERESTS

The author declares that there are no conflicts of interest regarding the publication of this article.

REFERENCES

1. Caplan Z. The Older Population: 2020. Available from: https://www.census.gov/library/publications/2023/decennial/c2020br-07.html (accessed on 2023-10-6).
2. Interview with Mendenhall, J., Utah Senior Advisors, Licensed and Certified Assisted Living Administrator. Phone Call, July 2023. https://utahseniorcareadvisors.com/.
3. Interview with Nausheen, Serenity Elder Care, LLC. Phone Call, August 2023 https://serenityec.com/.
4. Lee R, Mason A. Cost of Aging. Available from: https://www.imf.org/external/pubs/ft/fandd/2017/03/lee.htm (accessed on 2023-10-6).
5. National Academies of Sciences, Engineering, and Medicine. The National Imperative to Improve Nursing Home Quality: Honoring Our Commitment to Residents, Families, and Staff. Available from: https://www.ncbi.nlm.nih.gov/books/NBK584657/ (accessed on 2024-7-14).
6. Preisig BC, *et al*. Eye Gaze Behavior at Turn Transition: How Aphasic Patients Process Speakers' Turns during Video Observation. *Journal of Cognitive Neuroscience.* 2016;28(10):1613–1624. EBSCOhost. https://doi.org/10.1162/jocn_a_00983
7. Donnarumma F, *et al*. You Cannot Speak and Listen at the Same Time: A Probabilistic Model of Turn taking. *Biological Cybernetics*. 2017;111(2):165–183. EBSCOhost. https://doi.org/10.1007/s00422-017-0714-1
8. Nie A & Guo B. Benefits and Detriments of Social Collaborative Memory in Turn taking and Directed Forgetting. *Perceptual & Motor Skills*. 2023;130(3):1040–1076. EBSCOhost. https://doi.org/10.1177/00315125231163626
9. Nguyen V, *et al.* A Systematic Review and Bayesian Meta-analysis of the Development of Turn Taking in Adult–child Vocal Interactions. *Child Development.* 2022;93(4):1181–1200. EBSCOhost. https://doi.org/10.1111/cdev.13754
10. Kendrick KH, *et al*. Turn taking in Human Face-to-Face Interaction Is Multimodal: Gaze Direction and Manual Gestures Aid the Coordination of Turn Transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2023;378(1875):1–12. EBSCOhost. https://doi.org/10.1098/rstb.2021.0473
11. Degutyte Z & Astell A. The Role of Eye Gaze in Regulating Turn Taking in Conversations: A Systematized Review of Methods and Findings. *Frontiers in Psychology.* 2021;11, N.PAG. EBSCOhost. https://doi.org/10.3389/fpsyg.2021.616471
12. Dawson J & Foulsham T. Your Turn to Speak? Audiovisual Social Attention in the Lab and in the Wild. *Visual Cognition*. 2022;30(1/2):116–134. EBSCOhost. https://doi.org/10.1080/13506285.2021.1958038
13. Bögels S & Levinson SC. Ultrasound Measurements of Interactive Turn taking in Question-Answer Sequences: Articulatory Preparation Is Delayed but Not Tied to the Response. *PloS One*. 2023;18(7):e0276470. EBSCOhost. https://doi.org/10.1371/journal.pone.0276470
14. Damphousse KR. Voice Stress Analysis: Only 15 Percent of Lies about Drug Use Detected in Field Test. Available from: https://nij.ojp.gov/topics/articles/voice-stress-analysis-only-15-percent-lies-about-drug-use-detected-field-test (accessed on 2024-7-15).
15. Bromuri S, Henkel AP, Iren D & Urovi V. Using AI to

predict service agent stress from emotion patterns in service interactions. *Journal of Service Management*. 2021;32(4): 581–611. https://doi.org/10.1108/JOSM-06-2019-0163.

16. Zhu J, *et al*. Ethics of Smart Home-Based Elderly Care. Available from: https://onlinelibrary.wiley.com/doi/10.1111/jonm.13521. (accessed on 2024-7-15).
17. Quick Statistics about Hearing, Balance, & Dizziness. Available from: https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing (accessed on 2025-1-10).