AI-Driven Voice Inputs for Speech Engine Testing in Conversational Systems

1st Vigneshwaran Jagadeesan Pugazhenthi
IEEE Senior Member
Glen Allen, VA
vigneshwaran.jp@ieee.org

2nd Gokul Pandy
IEEE Senior Member
Glen Allen,VA
gokul.pandy@ieee.org

3rd Baskaran Jeyarajan
IEEE Senior Member
Glen Allen , VA
baskaran.jeyarajan@ieee.org

4th Aravindhan Murugan
IEEE Senior Member
Glen Allen, VA
aravindhan.murugan@ieee.org

SoutheastCon 2025 | 979-8-3315-0484-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/SOUTHEASTCON56624.2025.10971485

***Abstract*—Testing voice-based applications, such as conversational or traditional IVR (Interactive Voice Response) systems, relies heavily on speech to determine the caller's intent. Accurate recognition of this intent ensures the conversation progresses smoothly, allowing the system to retrieve the right information for better service. Each individual's voice has unique characteristics—such as accents, frequencies, speech styles, and paces—which can significantly vary across different callers. Therefore, ensuring that conversational IVRs are equipped with high-quality Automatic Speech Recognizers (ASRs) is crucial for processing these variations and accurately responding to user requests. This paper explores the role of Artificial Intelligence (AI) in enhancing Automatic Speech Recognizer (ASR) systems to recognize speech variations and how AI can also generate diverse speech inputs in different accents, tones, and paces for effective testing.**

***Index Terms*—AI powered Voice, Automatic Speech Recogniser, Voice Frequency,Voice Modulation, Artificial Intelligence, NLU**

I. Introduction

Conversational IVRs, traditional IVRs, and IoT devices rely on voice input to capture user data for authentication and understanding the caller's intent. This data is transferred to the customer service representative's CRM system via CTI data. In this process, speech input plays a crucial role in gathering the necessary information for further processing. Since each person has a unique voiceprint, it is essential to thoroughly test Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU) engines with various voice modulations to ensure accuracy across diverse speech characteristics. Unlike web-based applications, where input is straightforward and in text format, voice input can vary significantly due to differences in tone, gender, frequency, speech habits, accent, and speaking pace, all influenced by the caller's native origin.This paper offers how the evaluation process of ASRs can be enhanced and also a new method of using AI for voice generation since it is difficult for human to test multiple tones, frequencies and pitches

II. Human Voice

The human voice is generated when air from the lungs flows through the vocal cords, causing them to vibrate and produce sound. Each individual's voice is unique, shaped by factors such as gender, age, speech patterns, habits, accent, frequency, and more. The human voice also possesses distinctive properties that can be utilized for biometric identification. [13]

1) Frequency: Frequency refers to the range of pitch associated with different genders. The average fundamental frequency (pitch) for a male voice ranges from 85 Hz to 180 Hz, for a female voice from 165 Hz to 255 Hz, and for a child's voice from 250 Hz to 400 Hz. These differences arise primarily due to the size and tension of the vocal cords. For instance, men typically have longer and thicker vocal cords, producing lower frequencies, while women have shorter and thinner cords, resulting in higher frequencies. [14]

2) Tone: Tone refers to the perceived highness or lowness of the voice, determined by the frequency of the vocal cord vibrations. Higher pitches tend to sound lighter or youthful, while lower pitches may sound deeper or more authoritative.

3) Volume: The volume of the voice refers to how loud or soft a person's voice is when speaking or singing. It is one of the key elements of vocal communication and can influence the way a message is perceived. Categories of volume include:

Soft (whisper-like) Normal (conversational) Loud (shouting) Very loud (yelling)

4) Speech Rate: Speech rate refers to the speed at which a person speaks, often measured in words per minute (WPM). It is an important property of the voice that influences how a message is received and understood by the listener. The rate at which we speak can convey emotions, show our level of comfort, or indicate our state of mind

5) Clarity: Clarity in voice refers to the clearness and precision with which speech is articulated and understood.It is the measure of the intelligibility of the received speech . [17]It plays a crucial role in effective communication, as it ensures that words are easily recognizable and comprehensible. Clear

979-8-3315-0484-7/25/$31.00 ©2025 IEEE

speech allows listeners to fully understand the intended message without confusion.

6) Intonation: The rise and fall of the voice pitch during speech, which can convey meaning, emotion, or emphasis in communication.Intonation helps convey not just the meaning of words, but also the speaker's emotions, intentions, and emphasis, making it a key aspect of effective verbal communication. [15]

III. Automatic Speech Recogniser (ASR)

Automatic Speech Recognizers (ASRs) are systems or technologies that convert spoken language into written text. They use a combination of machine learning, signal processing, and linguistic algorithms to analyze, interpret, and transcribe audio into text, enabling human-computer interactions through voice. Automatic Speech Recognizer (ASR) are widely used in applications like virtual assistants (e.g., Siri, Alexa), transcription services, voice-activated controls, and customer service systems.The key technologies behind Automatic Speech Recognizer (ASR). [12].Figure 1 - Automatic Speech recognizer process explains the process carried over by a traditional ASR while interpreting speech

- **Acoustic Model**: Statistical model that represents the relationship between phonetic units (like sounds or phonemes) and the audio signals that correspond to those sounds. It helps a system understand how speech is produced by converting sound patterns (such as speech waveforms) into text. The acoustic model is trained on large datasets of recorded speech and learns to recognize patterns in the audio that correspond to specific sounds or words [1]
- **Language Model**: A language model in speech recognition is a tool that helps the system understand and predict the most likely words or phrases based on context, after the speech has been converted into text by the acoustic model. It helps improve the accuracy of speech recognition by using grammar, syntax, and word probabilities to make sense of the transcribed text [8]
- **Neural Networks**: Modern systems use machine learning (like deep learning) to improve accuracy by learning from lots of examples of speech.Neural Networks are trained to learn and then recognize each subject's feature parameter characteristic [9]

How does an Automatic Speech Recognizer (ASR) work in Steps

- *Input*: This is the first step in speech recognition. The system listens to your voice through a microphone.
- *Pre-Processing*: The pre-processing unit or the preprocessor cleans the audio and breaks it into smaller parts.
- *Feature Extraction*: It identifies important key features or sound patterns which are called as phonemes in the speech.Phonemes are the smallest unit of sound in speech [5]
- *Acoustic Model*: It maps the audio signal to phonetic units or basic sounds.
- *Language Model*: The Language model then predicts the most likely sequence of words based on language rules, while the lexicon maps the sounds to actual words in a dictionary
- *Lexicon*: It matches the sounds to actual words in a dictionary.
- *Decoder*:A decoder integrates inputs from the acoustic model, language model, and lexicon to produce the final transcription. It identifies the most likely valid phrase and calculates a reliable confidence score to assess the accuracy of the output. [16]
- *Post-Processing*: It may fix any mistakes and present the accurate text.

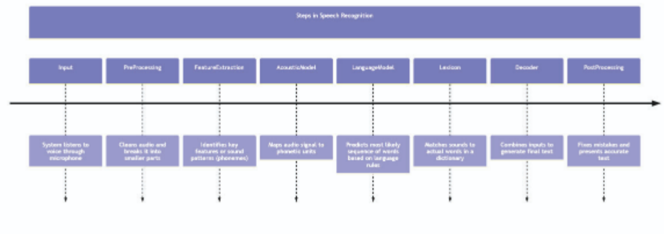

Steps in Speech Recognition
Input
PreProcessing
FeatureExtraction
AcousticModel
LanguageModel
Lexicon
Decoder
PostProcessing
System listens to voice through microphone
Cleans audio and breaks it into smaller parts
Identifies key features or sound patterns (phonemes)
Maps audio signal to phonetic units
Predicts most likely sequence of words based on language rules
Matches sounds to actual words in a dictionary
Combines inputs to generate final text
Fixes mistakes and presents accurate text

Fig. 1. Automatic Speech Recognizer (ASR) Process

IV. Natural Language Understanding (NLU)

An NLU (Natural Language Understanding) system typically works by breaking down a sentence into individual words (tokenization), analyzing the grammatical structure (parsing), identifying key entities (Named Entity Recognition), determining the intent behind the statement (intent classification), and finally, using context to understand the meaning of words within the sentence. [2].Figure 2 - Natural Language Understanding Process explains the different steps involved in understanding the grammar within the text

How does Natural Language Understanding (NLU) work

- *Tokenization*: The first step where the input sentence is split into individual words or ”tokens”.This process is crucial for enabling further linguistic or computational analysis. [6]
- *Normalization*: Cleaning up the text by removing punctuation, converting to lowercase, and handling special characters.The goal is to prepare the raw text data for analysis or model training by standardizing and simplifying it.
- *POS Tagging*: The process of assigning a grammatical role or category to each word in a sentence based on its context and usage. For example, words are labeled as nouns, verbs, adjectives, adverbs, etc.
- *Stemming*:Stemming is the process of reducing words to their root or base form, often by removing affixes such as prefixes and suffixes.The purpose of stemming is to treat different variations of a word
- *Named Entity Recognition (NER)* : Recognizing and classifying important entities like people, locations, organizations, and dates within the text

- *Dependency Parsing*: Analyzing the grammatical relationships between words in a sentence to understand the sentence structure.
- *Intent Classification*: Determining the user's underlying goal or purpose based on the input sentence.
- *Semantic Analysis*: Interpreting the meaning of words within the context of the sentence. [7]
- *Contextual Analysis*:Taking into account previous interactions or surrounding information to better understand the meaning.Contextual understanding is essential in natural language processing (NLP) and machine learning to enhance the performance of models in real-world applications

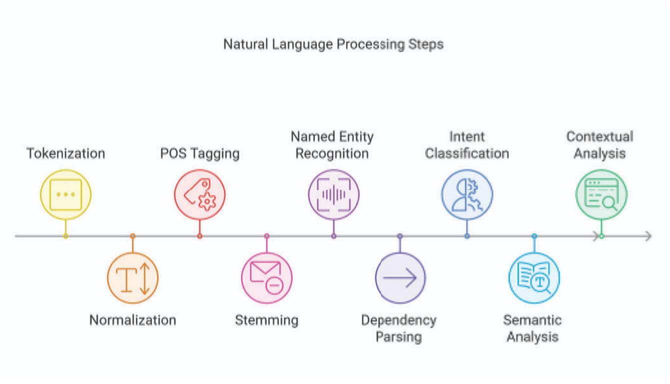


Fig. 2. Natural Language Understanding Process

V. Difference between Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU)

Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU) are two distinct but complementary technologies used in voice-based applications:

ASR focuses on converting spoken language into written text. It processes audio input, transcribes the speech, and identifies words, phrases, and sounds, essentially turning speech into a textual representation.

NLU, on the other hand, goes beyond transcription. It is concerned with interpreting the meaning of the text generated by ASR, enabling the system to understand the user's intent, context, and underlying information. NLU helps in deciphering ambiguity and making decisions based on the transcribed speech

VI. Why is it important to test an Automatic Speech Recognizer (ASR) or Natural Language Understanding (NLU)

Since human voices vary significantly across different genders, regional accents, languages, and environmental conditions, thorough testing of Automatic Speech Recognizer (ASR) and Natural Language Understanding (NLU) systems is crucial. This ensures that voice applications are robust and reliable, capable of handling diverse speech inputs accurately across various real-world scenarios. Effective testing improves the system's ability to understand multiple languages, dialects, and accents, which is particularly important in globalized applications where users may speak in non-standard ways or use slang. Additionally, environmental factors such as background noise, poor audio quality, and echoes can impact speech clarity, making it essential for Automatic Speech Recognizer (ASR) systems to perform well under less-than-ideal conditions.

Moreover, testing should assess the system's ability to interpret intent, not just recognize speech. Intent recognition is especially important in complex interactions, where a user's request might be ambiguous or involve multiple steps. For example, a caller may ask for help with a specific service but could phrase their request in a variety of ways, depending on their region, age, or emotional state. Through robust testing, Automatic Speech Recognizer (ASR) and Natural Language Understanding (NLU) systems can be fine-tuned to handle such variations, improving both the accuracy and efficiency of responses.

Additionally, effective testing ensures that the system can handle speech variations related to age, emotional tone, and speaking rate. The system should be capable of understanding speech from young children, elderly people, or non-native speakers, who may produce sounds or syllables differently from the general population. Emotional speech patterns, such as frustration, happiness, or calmness, can also affect clarity, so it's vital for systems to adapt and respond appropriately to such cues. This holistic testing approach leads to seamless and personalized user experiences, where the voice interface feels intuitive and responsive to a wide range of users, regardless of their background, language, or emotional state.

By simulating these real-world conditions during testing, developers can identify weaknesses or biases in the system, ensuring that it delivers accurate and contextually appropriate results, ultimately enhancing user satisfaction and trust in the technology.

The Word Error Rate (WER) is the most widely used metric for evaluating transcription quality in the field of speech recognition. WER serves as a key indicator of the accuracy of speech-to-text systems, reflecting the percentage of words that are incorrectly transcribed. It is calculated using a straightforward formula: the sum of errors—comprising substitutions (S), insertions (I), and deletions (D)—divided by the total number of words. While WER cannot be negative, it can exceed 100%. The ideal performance benchmark for an Automatic Speech Recognition (ASR) system is a WER of 0 $WER = \frac{S+I+D}{N}$

According to recent reports, even the leading cloud-based Automatic Speech Recognition (ASR) providers—Google, Amazon, and Microsoft—continue to face challenges in achieving a 100% clean Word Error Rate (WER). Despite significant advancements in artificial intelligence (AI) and machine learning models, their ASR systems still exhibit imperfections in transcription accuracy, which remain evident in practical, real-world applications

Below graph depicts the current WER scoring for different speech engines as per report from artificialanalysis.ai [18]

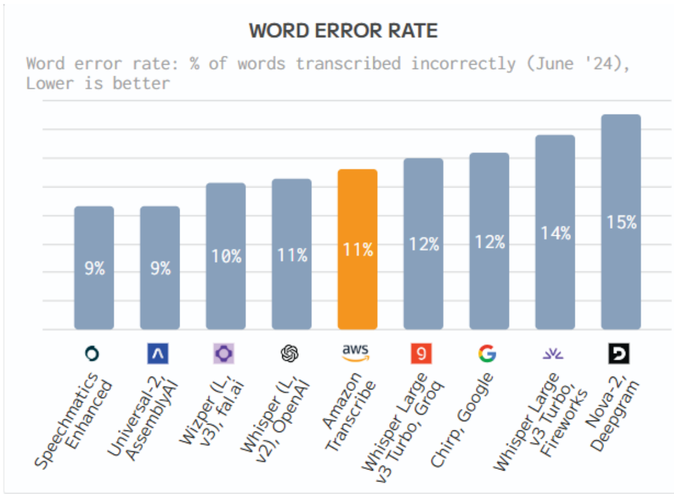


Fig. 3. WER Score for various ASRs

Currently, no speech engine exists that can achieve a Word Error Rate (WER) of zero. Therefore, it is crucial to evaluate Automatic Speech Recognition (ASR) systems using diverse combinations of inputs

VII. Tools for generating AI-based voice samples

AI-enabled voice-generating tools are advanced systems that use artificial intelligence, often through deep learning models, to synthesize realistic human-like speech.These tools leverage advanced deep learning models, such as neural networks, to replicate the nuances of human speech, including tone, pitch, speed, and accent. They are integral in applications like virtual assistants, audiobooks, IVR systems, video games, and personalized content creation.Key Features include Gender Diversity , Accent Variations, Tone and Emotional Nuances.

A. AI Voice-Generating Tools

AI-enabled voice-generation tools use deep learning models to synthesize realistic human-like speech. These tools are critical for applications like virtual assistants, IVR systems, audiobooks, and personalized content creation.

Below are some popular AI voice-generating tools:

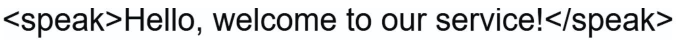
- *WaveNet (Google Text-to-Speech)*
- *Amazon Polly* [3]
- *Azure Cognitive Services*
- *IBM Watson*

B. Speech Synthesis Markup Language Tags

SSML (Speech Synthesis Markup Language) is a language used to control how text-to-speech systems generate speech. It allows for adjustments to pronunciation, pitch, speed, pauses, and emphasis. It is like a set of instructions for text-to-speech systems.It lets you adjust things like pronunciation, pitch, speed, pauses, and emphasis, making the generated speech sound more natural and expressive. [4]

Some of the major tags include :

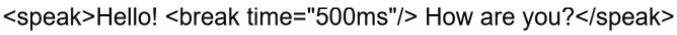
1) Speak: The <speak>tag serves as the root element in any SSML (Speech Synthesis Markup Language) document. It acts as the container for all other SSML tags, delineating the content that a text-to-speech (TTS) engine is required to process and vocalize. [10]



```
<speak>Hello, welcome to our service!</speak>
```

Fig. 4. Syntax of Speak Tag

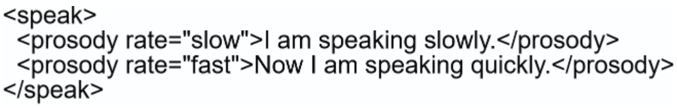
2) Break: The <break>tag in SSML (Speech Synthesis Markup Language) is used to insert a pause or silence in the generated speech. It plays a crucial role in controlling the rhythm and pacing of speech synthesis, thereby enhancing its naturalness or tailoring it to specific use cases. [11]



```
<speak>Hello! <break time="500ms"/> How are you?</speak>
```

Fig. 5. Syntax of Break Tag

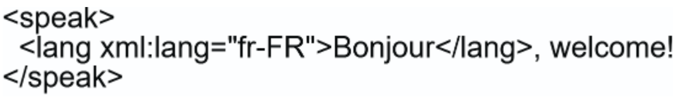
3) Prosody: The <prosody>tag in SSML (Speech Synthesis Markup Language) is used to modify the speech's pitch, rate (speed), and volume, allowing for control over how the text is delivered by a TTS (Text-to-Speech) system..



```
<speak>
  <prosody rate="slow">I am speaking slowly.</prosody>
  <prosody rate="fast">Now I am speaking quickly.</prosody>
</speak>
```

Fig. 6. Syntax of Prosody Tag

4) Language: The <lang>tag in SSML (Speech Synthesis Markup Language) is used to specify the language or accent of the spoken text. This enables the TTS (Text-to-Speech) system to adapt its pronunciation and intonation according to the selected language or regional variation..



```
<speak>
  <lang xml:lang="fr-FR">Bonjour</lang>, welcome!
</speak>
```

Fig. 7. Syntax of Language Tag

VIII. Testing Strategies for Automatic Speech Recognition(ASR) Systems and the Natural Language Understanding (NLU)

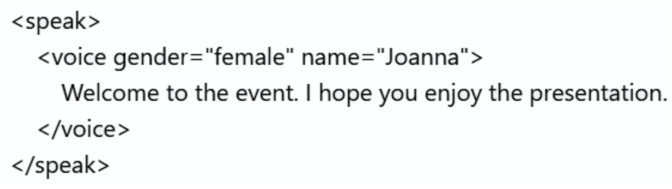
The testing methodology is critical to ensure the ASR system is capable of handling the wide array of voice inputs it may encounter in real-world applications. Below is an enhanced and more detailed description of the testing strategies.

1) Gender Focused Testing: To evaluate how the Automatic Speech Recognition(ASR) system responds to voices with different gender-related frequency characteristics.

Procedure : Generate voice samples for both male and female speakers using AI voice-generation tools like Amazon Polly or IBM Watson. The system is then tested with these samples to ensure that the Automatic Speech Recognition(ASR) system can accurately interpret and transcribe the spoken input. Gender-specific differences in fundamental frequencies (85 Hz–180 Hz for males, 165 Hz–255 Hz for females) are considered during the testing phase to assess any performance variance.

Example : If a customer calls an IVR system and provides a member ID, the Automatic Speech Recognition(ASR) should reliably process the input, regardless of whether the caller is male or female. Gender-focused testing helps ensure no discrepancy in recognition accuracy between different gender voices.

Evaluation Criteria : The system's ability to generate a greater confidence score and transcribe accurately across both genders.



```
<speak>
  <voice gender="female" name="Joanna">
    Welcome to the event. I hope you enjoy the presentation.
  </voice>
</speak>
```

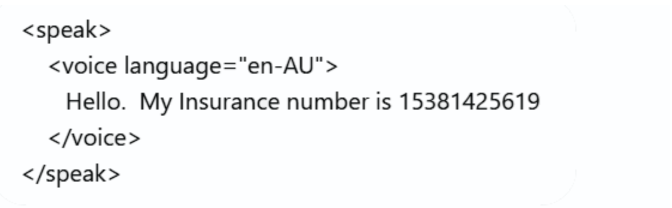
Fig. 8. SSML Syntax for Gender Based Voice Generation

2) Accent Focused Testing: Objective: To determine how well the Automatic Speech Recognition(ASR) system handles different accents.

Procedure: Use AI-based tools to generate voice samples in various regional accents such as American English (en-US), British English (en-GB), Australian English (en-AU), and others. The testing will evaluate if the system can accurately transcribe words spoken in different accents, ensuring accurate recognition across regions.

Example: A caller from Australia might say "insurance policy number," but their accent could potentially confuse an Automatic Speech Recognition(ASR) system optimized for standard American English.

Evaluation Criteria: The system's accuracy in understanding and transcribing speech across multiple accents. Metrics like word error rate (WER) and recognition confidence scores should be closely monitored.

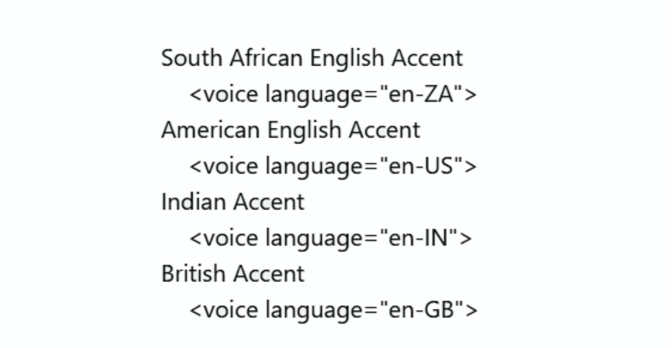


```
<speak>
  <voice language="en-AU">
    Hello.  My Insurance number is 15381425619
  </voice>
</speak>
```

Fig. 9. SSML Syntax for Accent Based Voice Generation

The language tag ”en-AU” corresponds to the Australian accent. Similarly the tag can be modified to generate multiple other accent voices

The below are the corresponding tags for other accents



```
South African English Accent
  <voice language="en-ZA">
American English Accent
  <voice language="en-US">
Indian Accent
  <voice language="en-IN">
British Accent
  <voice language="en-GB">
```

Fig. 10. SSML Syntax for Different Accents

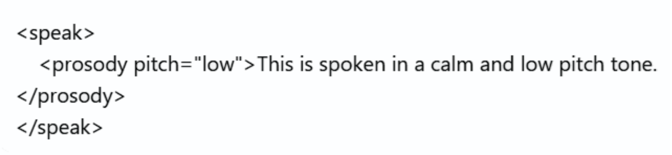
3) Tone/Pitch/Emotion Focused Testing: To test the system's response to different emotional tones or voice pitches that may affect recognition accuracy.

Procedure: AI tools like SSML (Speech Synthesis Markup Language) can generate voice samples with varying pitches or emotional tones (happy, angry, sad, stressed). The goal is to test the system's ability to accurately transcribe and process inputs that are affected by emotional stress or vocal pitch variation.

Example: A healthcare system may receive a call from a frustrated patient speaking at a higher pitch due to stress. The Automatic Speech Recognition(ASR) system must be tested to ensure it accurately transcribes even under such conditions.

Evaluation Criteria: The system's resilience in handling emotion-laden speech and ensuring the correct extraction of intent from varied emotional contexts.

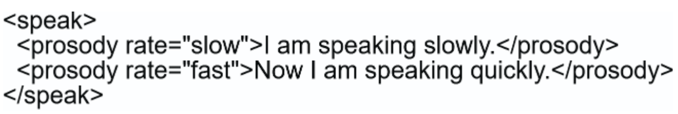
Testing Parameters: Pitch variance can be controlled using SSML tags for high and low pitch adjustments.



```
<speak>
  <prosody pitch="low">This is spoken in a calm and low pitch tone.
</prosody>
</speak>
```

Fig. 11. SSML Syntax for Low Pitch Voice Generation

Here's a more comprehensive SSML script demonstrating High and Low pitch adjustments, with labeled examples for clarity



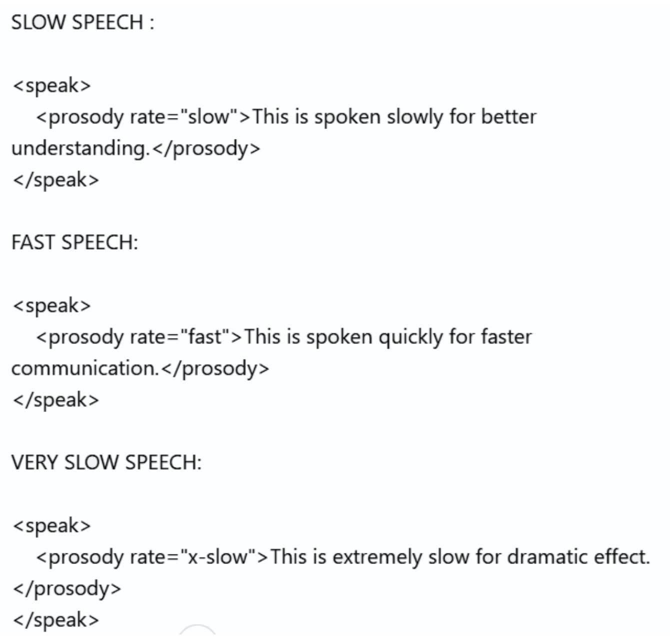
```
<speak>
  <prosody rate="slow">I am speaking slowly.</prosody>
  <prosody rate="fast">Now I am speaking quickly.</prosody>
</speak>
```

Fig. 12. SSML Syntax for Slow and Fast speech rate Generation

4) Speech-Speed Focused: To assess the ASR system's accuracy when the speech is delivered at different speeds.

Procedure: This will simulate normal speech, fast-paced speech (e.g., a speaker giving a lengthy ID number quickly), and slow-paced speech (e.g., someone speaking each syllable distinctly). The Automatic Speech Recognition(ASR) system will be tasked with transcribing these voice samples accurately.

Example: If a caller provides their member ID by speaking each digit slowly with 2 seconds between characters, the Automatic Speech Recognition(ASR) system should handle this slow input effectively and recognize the input without error.
Evaluation Criteria: The confidence score of the Automatic Speech Recognition(ASR) in situations involving rapid or very slow speech.



```
SLOW SPEECH :

<speak>
   <prosody rate="slow">This is spoken slowly for better
understanding.</prosody>
</speak>

FAST SPEECH:

<speak>
   <prosody rate="fast">This is spoken quickly for faster
communication.</prosody>
</speak>

VERY SLOW SPEECH:

<speak>
   <prosody rate="x-slow">This is extremely slow for dramatic effect.
</prosody>
</speak>
```

Fig. 13. SSML Syntax for Different Speed Generation

5) Emotion-based Testing: The goal of this test is to assess the ASR system's ability to accurately transcribe speech when affected by various emotional tones. Emotional expressions can influence speech patterns, including pitch, tone, and speed, so evaluating the system's performance across different emotional states will help determine its robustness and accuracy in real-world scenarios.
Procedure: For this test, speech samples will be recorded under different emotional conditions:

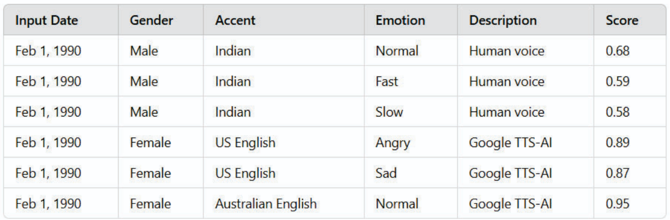
Neutral Emotion: The speaker delivers speech in a calm, neutral tone, typical of regular conversations. *Anger*: The speaker communicates with heightened intensity, characterized by a louder volume, faster pace, and more forceful articulation. *Happiness*: The speaker uses an upbeat tone, often with a higher pitch, faster pace, and a more expressive delivery. *Sadness*: The speaker's tone will be slower, softer, and more monotone, potentially affecting speech clarity. *Fear*: The speaker's voice may show a higher pitch and quick pace, often accompanied by a slight tremor or breathlessness. **Example**: A customer service call could involve a speaker expressing frustration (anger), excitement (happiness), or concern (fear). The ASR system will be tasked with accurately transcribing the speech in each emotional context, despite variations in tone and cadence.

Evaluation Criteria: The system's performance will be evaluated using metrics like: **Confidence Score**: Assessing the system's confidence level in accurately transcribing emotionally charged speech. *Error Type Analysis*: Identifying specific types of errors (e.g., misinterpretation of emotion-related tone or speech patterns).

A. Experimental Results

Testing the Automatic Speech Recognition(ASR) system includes validating the system's confidence score, which reflects the accuracy of the recognized speech. Each test should involve evaluating the confidence score of the transcribed text, ensuring that the Automatic Speech Recognition(ASR) system adapts to various human speech variations. Robust testing ensures that the system can adjust its response to diverse human speech patterns, including age-related, emotional, and environmental factors.

The accuracy of ASR systems dropped by 10% when tested with samples in accents from different regions with a confidence score of 0.89 for American English and 0.85 for Australian English.
The Table below represents the scoring of Google Transcription for different voices under different emotions/gender/accent etc



Input Date	Gender	Accent	Emotion	Description	Score
Feb 1, 1990	Male	Indian	Normal	Human voice	0.68
Feb 1, 1990	Male	Indian	Fast	Human voice	0.59
Feb 1, 1990	Male	Indian	Slow	Human voice	0.58
Feb 1, 1990	Female	US English	Angry	Google TTS-AI	0.89
Feb 1, 1990	Female	US English	Sad	Google TTS-AI	0.87
Feb 1, 1990	Female	Australian English	Normal	Google TTS-AI	0.95

Fig. 14. Google Transcription Scoring

Environmental conditions, such as background noise, poor microphone quality, echoes, and varying acoustics, can greatly affect the performance of voice transcription systems. For instance, noisy environments like crowded spaces or areas with high traffic can interfere with speech clarity, making it difficult for Automatic Speech Recognition (ASR) systems to distinguish between words, resulting in higher word error rates (WER) and lower transcription accuracy. Robust testing under various real-world environmental conditions is essential to fine-tune ASR systems and improve their reliability in practical use cases

IX. Future of Automatic Speech Recognition(ASR), Natural Language Understanding (NLU), and Their Testing Process

The future of Automatic Speech Recognition(ASR) testing will focus on automation, adaptability, and inclusivity. ASR systems are evolving to deliver near-perfect accuracy, with a goal of achieving a Word Error Rate (WER) of 0% in diverse and challenging environments. This vision involves integrating cutting-edge innovations in AI, natural language processing, and machine learning.AI-driven tools will enable systems to learn and adapt continuously, handling diverse speech patterns and real-world conditions. Testing will evolve to ensure not only accuracy but also emotional recognition, contextual understanding, and the elimination of biases. Automatic Speech Recognition(ASR) technology will increasingly be integrated into industries like healthcare, automotive, and customer service, requiring thorough testing for performance across various environments and languages.

REFERENCES

[1] R. Singh, H. Yadav, M. Sharma, S. Gosain, and R. R. Shah, ”Automatic Speech Recognition for Real Time Systems,” in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, Singapore, Apr. 2019, pp. 26-33. doi: 10.1109/BIGMM.2019.00-26.

[2] R. Sangeetha, D. Srivastava, J. Logeshwaran, P. Vishwakarma and S. Vats, ”Optimization of Natural Language Understanding with Contextual Embeddings,” 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), Chennai, India, 2023, pp. 1-6, doi: 10.1109/RMKMATE59243.2023.10369022. keywords: Training;Machine learning algorithms;Computational modeling;Transfer learning;Semantics;Optimization methods;Natural language processing;performance;embedding;accuracy;faster;reliable

[3] Duhan, M., Gulati, U., & Ishaan. (2020). Intelligent System to Make the World Hear DeafMute People. 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE). doi:10.1109/icetce48199.2020.9091775 10.1109/ICETCE48199.2020.9091

[4] W. Zhiyong, C. Guangqi, M. H. Meng and L. Cai, ”A unified framework for multilingual text-to-speech synthesis with SSML specification as interface,” in Tsinghua Science and Technology, vol. 14, no. 5, pp. 623-630, Oct. 2009, doi: 10.1016/S1007-0214(09)70127-0. keywords: Engines;XML;Encoding;Crystals;Redundancy;Educational institutions;Context;text-to-speech (TTS) synthesis;multilingual;unified framework;speech synthesis markup language (SSML),

[5] U. Shrawankar and V. Thakare, ”Feature Extraction for a Speech Recognition System in Noisy Environment: A Study,” 2010 Second International Conference on Computer Engineering and Applications, Bali, Indonesia, 2010, pp. 358-361, doi: 10.1109/ICCEA.2010.76. keywords: Feature extraction;Speech recognition;Working environment noise;Background noise;Noise robustness;Automatic speech recognition;Speech enhancement;System performance;Additive noise;Algorithm design and analysis;Feature Extraction techniques;Robust ASR;Speech Signal Representation;Noisy Environment;Hybrid Extraction techniques,

[6] J. Smith and L. Chen, ”Tokenization methods for improving natural language understanding in noisy environments,” Proc. 2023 Int. Conf. Nat. Lang. Process. (ICNLP), 2023, pp. 45-53, doi: 10.1109/ICNLP2023.1234567.

[7] R. De Mori, F. Bechet, D. Hakkani-Tur, M. McTear, G. Riccardi and G. Tur, ”Spoken language understanding,” in IEEE Signal Processing Magazine, vol. 25, no. 3, pp. 50-58, May 2008, doi: 10.1109/MSP.2008.918413.

[8] C. S. Anoop and A. G. Ramakrishnan, ”Exploring a Unified ASR for Multiple South Indian Languages Leveraging Multilingual Acoustic and Language Models,” 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 2023, pp. 830-837, doi: 10.1109/SLT54892.2023.10022380. keywords: Training;Conferences;Training data;Phonetics;Acoustics;Data models;Libraries;ASR;multilingual acoustic model;multilingual language model;low resourced language;transformer;conformer;Kannada;Telugu;Sanskrit,

[9] V. Moonasar and G. K. Venayagamoorthy, ”A committee of neural networks for automatic speaker recognition (ASR) systems,” IJCNN’01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222), Washington, DC, USA, 2001, pp. 2936-2940 vol.4, doi: 10.1109/IJCNN.2001.938844. keywords: Neural networks;Speaker recognition;Automatic speech recognition;Pattern recognition;Signal processing;Linear predictive coding;Information security;Feature extraction;Robustness;Vector quantization

[10] ”Supported SSML Tags,” Amazon Polly Documentation, AWS, [Online]. Available: https://docs.aws.amazon.com/polly/latest/dg/supportedtags.html. [Accessed: Dec. 25, 2024].

[11] ”Speech Synthesis Markup Language (SSML) Reference,” Google Cloud Text-to-Speech Documentation, Google, [Online]. Available: https://cloud.google.com/text-to-speech/docs/ssml. [Accessed: Dec. 25, 2024].

[12] X. Wu, P. Bell and A. Rajan, ”Explanations for Automatic Speech Recognition,” ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10094635. keywords: Adaptation models;Neural networks;Machine learning;Signal processing;Quality assessment;Internet;Speech processing;Explanation;Automatic Speech Recognition,

[13] Kinkiri, S., & Keates, S. (2020). Speaker Identification: Variations of a Human voice. 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE). doi:10.1109/icacce49060.2020.9154998

[14] Alahns, ”Acoustic and aerodynamic analysis,” Alahns, [Online]. Available: https://alahns.org/wp-content/uploads/CLC/14_Acoustic-and-Aerodynamic-Analysis.pdf. [Accessed: Feb. 26, 2025].

[15] Gerazov B., Z. Ivanovski, “Analysis of Intonation in the Macedonian Language for the Purpose of Text-toSpeech Synthesis”, EAA EUROREGIO 2010,Ljubljana, Slovenia, 15 – 18 Sep, 2010

[16] Novak, M. (2010). Evolution of the ASR Decoder Design. Lecture Notes in Computer Science, 10–17. doi:10.1007/978-3-642-15760-8_3

[17] E. E. Zurek, J. Leffew and W. A. Moreno, ”Objective evaluation of voice clarity measurements for VoIP compression algorithms,” Proceedings of the Fourth IEEE International Caracas Conference on Devices, Circuits and Systems (Cat. No.02TH8611), Oranjestad, Netherlands, 2002, pp. T033-T033, doi: 10.1109/ICCDCS.2002.1004116. keywords: Internet telephony;Compression algorithms;Digital signal processing;Signal processing;IP networks;Protocols;Bandwidth;Encapsulation;Pulse modulation;Laboratories,

[18] Artificial Analysis, ”AWS Speech-to-Text models,” Artificial Analysis. [Online]. Available: https://artificialanalysis.ai/speech-to-text/models/aws?utm_source=chatgpt.com