Chloe Tu
ITAI 2373 – Natural Language Processing
Professor: Anna Devarakonda
Date: September 11, 2025

Assignment 03: Voice Tech in the Multiverse - 🎮 Gaming World

Step 1: Universe Selection (setting)

For this challenge, I chose the world of Legends of the Shattered Realms, which is a groundbreaking VR MMORPG designed to challenge every sense, especially hearing. For Mystical Meihua, a mage specializing in vocalized arcane spells, the audio landscape is a key factor in her success. The game's intricate design, while immersive, introduces a series of acoustic challenges that demand a custom-built solution.

Step 2: Complete All Four Deliverable Parts

Part 1: World Analysis

Unique Acoustic Challenges

- **Dungeon reverb:** Echoes in caves and stone corridors persist and overlap, creating blurred speech and making recognition unstable. The stone halls and deep caverns of the game world are acoustically challenging. Persistent echoes blur spoken commands. This creates overlapping sound that makes it nearly impossible for a standard system to accurately recognize and execute spells like "CAST FIREBALL."
- **Chaotic raids:** During large-scale battles, dozens of players speak simultaneously. This creates a cluttered audio environment. In these moments, pinpoint accuracy is critical. A misfired spell due to audio confusion could lead to a team wipe.
- **Fantasy spellwords:** The game's command vocabulary includes invented words like "Furra-Kai" and "Zyn-Rathos." These unique syllables do not exist in standard linguistic models. Any voice technology must be flexible enough to learn and recognize this specialized language.
- **Ultra-low Latency:** Gameplay demands immediate action. Spells and commands must execute within milliseconds. A delay of over 150 ms is a critical failure point that could lose a player the game.
- **Overlapping noise sources:** Both environmental and artificial sounds (monster roars, spell explosions, crowds) mask crucial speech signals.

Environmental Factors

- **Underwater realms:** In underwater zones, fluid resistance and bubbles muffle and distort speech. This radically changes the acoustic signature of voice commands.
- **Volcanic Regions:** Areas near lava and fire introduce hissing and crackling sounds. This variable interference makes consistent voice recognition nearly impossible.
- **Mountain Summits:** High winds bend sound waves and mask quieter commands like whispers. A system must be able to filter out this powerful natural noise.
- **Arenas:** Massive, vibrant crowds create an immense noise floor. This combines monster roars and spell impacts. The system must be able to separate a single player’s voice from this chaotic roar. This is especially true when a boss like The Stone Tyrant Rorgash lets out his signature concussive roar.

User Characteristics

- **Play styles:** A player's voice can shift dramatically. They may use quiet whispers for stealth or a strategic shout for a charge command. A good system must recognize both. The system must also accommodate the dramatic roleplay of some players.
- **Equipment diversity:** Players use a variety of headsets and microphones ranging from basic to advanced. This introduces a wide range of signal quality.
- **Accessibility:** The system must be inclusive. This means it must be able to recognize natural speech patterns, diverse accents, and overcome speech impairments without penalizing the player.

Noise Sources and Acoustic Mapping

- **In-game:** Monsters, explosions, NPC banter, and ambient weather effects combine for omnipresent audio clutter.
- **Out-of-game:** Background home noises and dogs barking, roommates shouting, mechanical keyboard clicks, merge unpredictably into gameplay.
- **Risk:** Real-world and in-game sounds blur together, making it difficult for the system to reliably separate player commands from extraneous input.

Non-Human Vocal Anatomy (Optional)

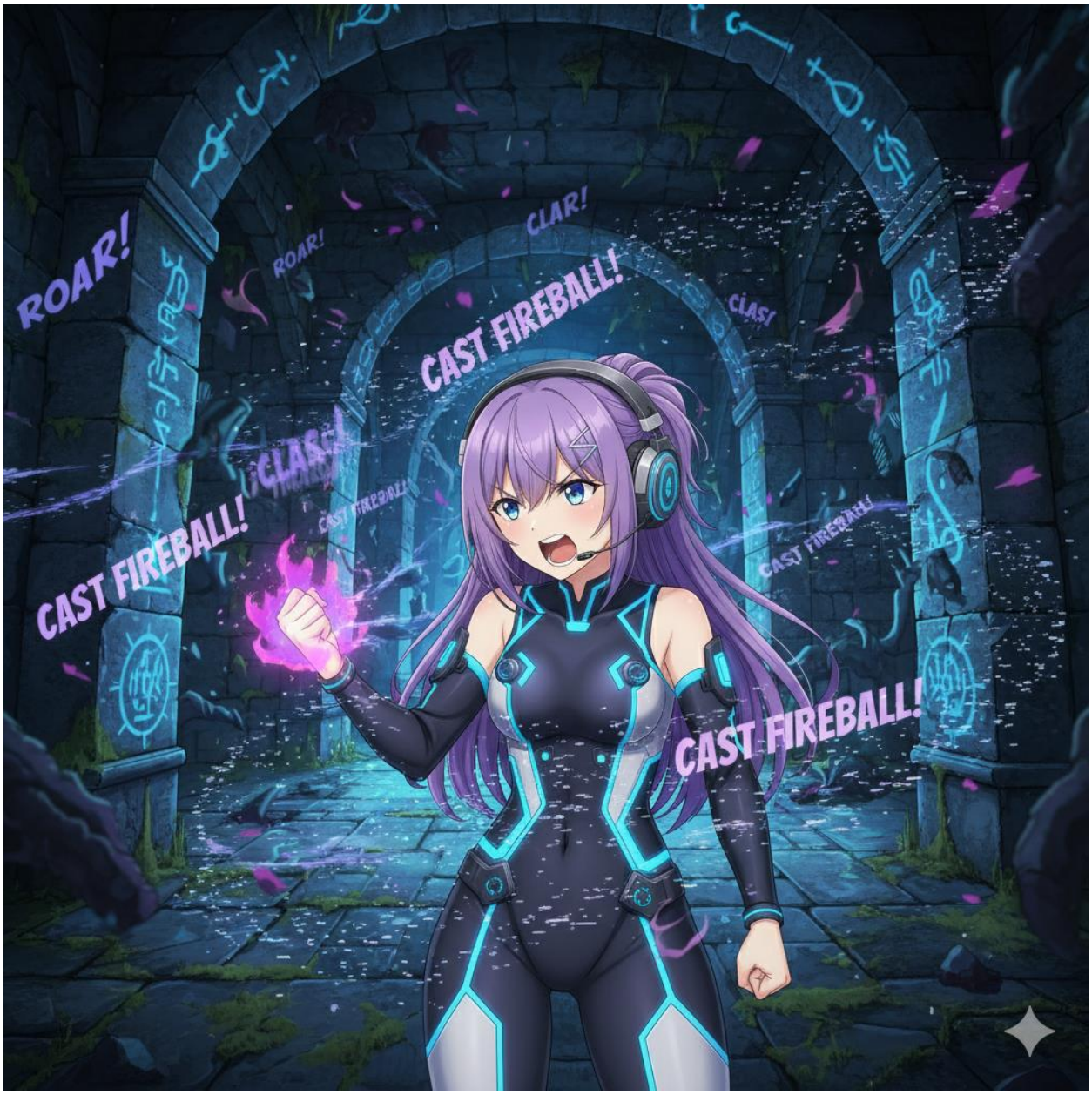
Avatars become magical creatures or machines orc growls, elf chants, robotic modulations. EchoBlade transforms spoken commands into appropriate avatar voice styles, maintaining both expressiveness and recognition accuracy.

Summary:

The game world is engineered to push voice control to its limits, demanding a flexible system that deals with noise, overlapping speech, invented languages, and varied user scenarios without lag or failure.

For Part 1:

World Analysis, envisioning the challenging environments:
Here's an image of a gamer (Mystical Meihua) in a dungeon, struggling with echoes as they try to cast a spell.



Part 2: Technical Solutions Design

Our solution, EchoBlade, is a multi-layer speech intelligence system. It is built for the unpredictable fantasy conditions of the game world. It is designed to empower every player, especially those with diverse play styles like Mystical Meihua. EchoBlade is engineered as a multi-layer speech intelligence system, custom-built for unpredictable, fantasy VR conditions.

Custom Preprocessing Pipeline (Flow)

1. Input Capture:

- Audio is first captured using a player's microphone. The system supports a wide range of sampling rates from 16 to 48 kHz. Optional throat microphones are used for whispers, ensuring stealth and accessibility.

2. Noise Filtering:

- This module removes both in-game and real-world background noise, such as dog barks and fire crackle. It uses adaptive filters that adjust dynamically to changing environments to prevent spell miscasts.

3. Dereverb Module:

- Adaptive algorithms suppress and cancel lingering echoes, especially vital in echo-heavy dungeons and cave zones.

4. Voice Separation:

- Beamforming isolates key speakers (raid leaders, main spellcasters) and ignores off-axis chatter, improving command clarity in group scenarios. This allows the system to ignore off-axis chatter and other players' voices, which is critical for clear communication in raids

5. Feature Extraction:

- Log-mel spectrograms for core speech recognition.
- **Prosody tracking:** Captures pitch, loudness, and rhythm to distinguish between whispers and shouts.
- **Phoneme embeddings:** Flexible model recognizes invented syllables or spellwords unique to Legends of the Shattered Realms.

Recognition and Synthesis

- **Command ASR:** Specialized to reliably identify structured actions ("shield ally," "attack," "summon") even in noise-heavy contexts.
- **Spellword Recognizer:** A sequence-to-sequence (Seq2Seq) neural model interprets unusual or fantasy-driven syllables for spellcasting.
- Synthesis: NPC and player "voice skins" enable real-time voice transformation, players' commands sound like their chosen avatars.

Why MFCCs Won't Work Here

Standard MFCC (Mel-frequency cepstral coefficient) algorithms are not suitable for this environment. They lose crucial information that the EchoBlade system needs.

- Standard MFCC algorithms drop pitch/energy cues needed to separate whispers from shouts.
- MFCC features collapse under intense dungeon echoes.
- The approach ignores the wide range of fantasy syllables, resulting in failed spell recognition.
- MFCC models expect steady speech,unrealistic during active gaming, where commands are brief and abrupt.

Feature Extraction Strategy

- Log-mel for standard speech clarity.
- Prosody and formant tracking to separate whispers/yells and model fantasy spellwords.
- Game-aware Voice Activity Detection isolates short bursts for timely command response.

Acoustic Modeling Considerations

EchoBlade uses a Streaming Conformer model for ultra-low latency ASR. This is essential for the <100 ms response time needed for competitive gameplay. The system also creates personal

profiles for each player. These profiles tune the model to a player's specific accent and microphone. This allows the system to learn and adapt to a player's unique voice.

Game-based data augmentation trains the system to handle new environments, invented words, and large-scale battles. This makes the system more robust and reliable.

- Streaming Conformer delivers ultra-low-latency ASR essential for competitive gameplay.
- Personal profiles tune models to each player's accent, microphone characteristics, and roleplay preferences.
- Game-based data augmentation teaches the system to handle novel environments, invented words, and epic scale battles.

ASR / TTS Adaptations

- The system's grammar is inclusive. It recognizes a core set of game commands and a flexible set of spellword tokens.
- For NPCs, **TTS (Text-to-Speech**) voices are stylized for the lore of the game. This means that a wizard's voice will have an echo, a fairy's voice will be a whisper, and an orc's voice will be a growl.
- The system also includes **voice skins**. This feature transforms a player's voice to sound like their chosen avatar, whether it is a mythical warrior, a demon, or an android.

EchoBlade is more than a speech recognizer; it is an immersive, adaptive pipeline for VR fantasy, giving every player's voice instant power and control.

Deeper Technical Detail: The Spellword Recognizer

The Spellword Recognizer is a specialized, sequence-to-sequence (Seq2Seq) neural model. This model is trained on a custom, game-specific dictionary of invented words and syllables created by the game developers. It is not limited to phonetic sounds found in Earth-based languages. The model is also designed to be adaptable. When a new expansion or content update is released with new spells, the model can be quickly fine-tuned with the new vocabulary, ensuring that Mystical Meihua's new arcane chants are recognized instantly without the need for a full re-training. This ensures the system remains current and flexible as the game world evolves.

Part 3: Demo Scenario

Storyboard Flow (8 Panels)

Panel 1 – Failed Fireball: The gamer raises her fist and shouts, "CAST FIREBALL!" but cavern echoes confuse the system. Instead of firing, the spell fizzles out mid-air.

Panel 2 – Missed Heal: She whispers softly, "heal me," but in the middle of raid chaos her voice is buried under overlapping shouts and roaring monsters. The system fails to recognize it.

Panel 3 – Boss Interference: A giant stone boss lets out a thunderous laugh, "RUA HA HA!" The roar overwhelms the recognition system, which mistakes it for a player command. Another misfire follows.

Panel 4 – EchoBlade Activates: The gamer steadies herself as EchoBlade powers up. Smart filters and dereverb modules glow in her UI, separating echoes from real speech and untangling the noisy battlefield.

Panel 5 – Fireball Success: She tries again, shouting, “CAST FIREBALL!” This time, EchoBlade processes the voice correctly. A blazing fireball shoots forward, striking the enemy with precision.

Panel 6 – Whisper Detected: She leans closer to her mic and whispers, “heal me.” EchoBlade’s throat mic input and prosody tracking pick it up instantly, triggering a glowing wave of healing magic.

Panel 7 – Victory Cheers: The stone boss crumbles into rubble as the raid team cheers. Even with everyone yelling in triumph, EchoBlade isolates the gamer’s voice, keeping recognition clean.

Panel 8 – Triumphant Close: The gamer lifts her glowing headset proudly. Across the final frame, bold words appear: “EchoBlade: Your Voice, Your Power.”

Scenario Technical Notes

- Smart dereverb removes dungeon echoes, boosting spell accuracy.
- Beamforming and prosody detection separate the player's distinct commands from ambient game sound.
- Seq2Seq recognizer ensures even fantasy syllables trigger gameplay as intended.

Summary:

The scenario demonstrates EchoBlade's resilience: it overcomes chaos, adapts to multiple failure points, and empowers every command.

Part 4: Executive Pitch

Branding

- **System Name:** EchoBlade™
- **Tagline:** "Your Voice is Your Ultimate Weapon."

Key Technical Features

- **Dynamic dereverb:** It eliminates echoes in cavernous dungeons and epic battles.
- **Prosody analysis:** It accurately recognizes whispers and shouts, giving players precise control over their volume.
- **Unique Spellword Detection:** It can recognize and interpret fantasy syllables.
- **NPC TTS and Voice Skins:** This feature creates a fully immersive audio experience.
- **Ultra-fast Voice Processing:** The system processes voice in less than 100 ms.

Business Impact and ROI

EchoBlade is more than just a technological feature; it is a core business driver. By providing a truly seamless and reliable voice interface, our system will increase player engagement and satisfaction. This leads to higher player retention rates, reducing churn. A superior gameplay experience will also generate positive word-of-mouth marketing, attract new users and building a loyal community around Legends of the Shattered Realms. In a competitive market where players demand next-level immersion, EchoBlade's unique capabilities provide a distinct competitive advantage that can be leveraged in marketing campaigns and justify a premium price point for the game or its expansions. Our solution not only solves a critical technical problem but also creates a significant revenue stream by ensuring players feel empowered and committed to the game world.

Competitive Advantages

- Recognizes and interprets fantasy syllables standard ASR ignores.
- Filters and separates overlapping raid or ambient chatter effortlessly.
- Adapts to dungeons, arenas, and environments where other systems fail.
- Lets all players craft their own in-game voice identity.
- Sets new speed and reliability standards for voice tech in gaming.

Legends of the Shattered Realms. In a competitive market where players demand next-level immersion, EchoBlade's unique capabilities provide a distinct competitive advantage that can be leveraged in marketing campaigns and justify a premium price point for the game or its expansions. Our solution not only solves a critical technical problem but also creates a significant revenue stream by ensuring players feel empowered and committed to the game world.

Character Name: Mystical Meihua
Summary Description:
Mystical Meihua is a spirited and highly skilled mage, renowned across the Shattered Realms for her mastery of arcane vocalizations. With her vibrant, amethyst hair often tied back in a practical yet stylish braid, she sports custom-designed gaming gear in sleek black and neon blue, perfectly integrated with her EchoBlade headset. Her sharp, intelligent blue eyes are always scanning the battlefield, but it's her voice—ranging from precise whispers for stealth spells to powerful shouts for devastating incantations—that truly defines her. Meihua thrives on strategic challenge and is fiercely loyal to her raid team. Her journey with EchoBlade isn't just about winning; it's about pushing the boundaries of what a mage can achieve when her every command is perfectly understood, turning her voice into a formidable weapon against the chaotic forces threatening her world.

Big Bad Monster Name: The Stone Tyrant, Rorgash
Summary Description:
Rorgash, the Stone Tyrant, is a colossal, ancient elemental forged from the very bedrock of the deepest dungeons. His form is a jagged, imposing mass of granite and obsidian, with glowing crimson cracks pulsing like veins of molten fury beneath his stony hide. His face is a craggy, snarling mask, and his eyes burn with malevolent, primordial energy. Rorgash's most terrifying weapon, beyond his immense physical strength, is his voice. His "RUA HA HA!" laugh is not just a sound; it's a concussive force, a cacophony designed to utterly overwhelm and shatter communication, specifically targeting and jamming player voice commands with its raw, earth-shaking resonance. He embodies the ultimate acoustic challenge, a living embodiment of the "Boss Interference" that EchoBlade was designed to overcome, seeking to silence all resistance with his thunderous roars and chaotic presence.
Reference: (AI was used to create all the images and comic strip)