AudioRole: An Audio Dataset for Character Role-Playing in Large Language Models

Anonymous ACL submission

Abstract

The creation of high-quality multimodal datasets remains fundamental for advancing role-playing capabilities in large language models (LLMs). While existing works predominantly focus on text-based persona simulation, Audio Role-Playing (ARP) presents unique challenges due to the need for synchronized alignment of semantic content and vocal characteristics. To address this gap, we propose **AudioRole**, a meticulously curated dataset from 13 TV series spanning 1K+ hours with 1M+ character-grounded dialogues, providing synchronized audio-text pairs annotated with speaker identities and contextual metadata. In addition, to demonstrate the effectiveness of the dataset, we introduced **ARP-Eval**, a dual-aspect evaluation framework that assesses both *response quality* and *role fidelity*. Empirical validation showing GLM-4-Voice trained on AudioRole (which we called **ARP-Model**) achieve an average Acoustic Personalization score of 0.31, significantly outperforming the original GLM-4-voice and the more powerful model MiniCPM-O-2.6, which specifically supports role-playing in one-shot scenarios. The **ARP-Model** also achieves a Content Personalization score of 0.36, surpassing the untrained original model by about 38% and maintaining the same level as MiniCPM-O-2.6.

AudioRole features dialogues from over 115 main characters, 6 trained **ARP-Model**s that role-play different characters, and evaluation protocols. Together, they provide an essential resource for advancing audio-grounded role-playing research.

1 Introduction

The evolution of role-playing capabilities in LLMs has revolutionized human-AI interaction paradigms. Contemporary systems demonstrate remarkable proficiency in textual persona simulation, serving as personalized assistants [1], emotional companions[2], and social interaction proxies (Park et al., 2023). However, this progress remains fundamentally constrained by unimodal (text-only) interaction frameworks, which neglect the critical role of vocal characteristics in authentic character portrayal — a limitation that is particularly evident when simulating personas from audiovisual media, such as films and TV series.

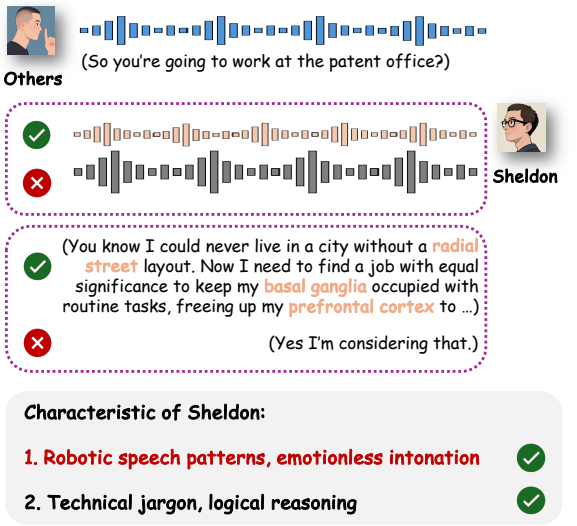

Figure 1: Audio Role-Playing case of Sheldon. The answer should satisfy not only the lexical similarity but also acoustic similarity, which are the two characteristics of Sheldon.

Existing work on audio evaluation considers mainly content accuracy (Chen et al., 2024b) (Hassid et al., 2024) or only the conversational style (Liu et al., 2025). Vocal expression reflects an integral component of communication that varies considerably between individuals in different contexts (Cohen et al., 2015), and the acoustic properties of speech reflect key variables for understanding human behavior (Decety and Lamm, 2006). Vocal expression is highly variable across individuals and across contexts, and is influenced by a num-

[1] https://chatgpt.com/

[2] https://replika.com/

ber of individual differences. Consider Dr. Sheldon Cooper from *The Big Bang Theory* (Figure 1): while text-based LLMs can mimic his technical jargon and logical reasoning, they fundamentally fail to capture his iconic robotic speech patterns — specifically, his faster-than-average speaking rate and emotionless intonation. This disconnect highlights a critical challenge in audio-grounded role-playing: authentic characterization requires not just *what* the persona says, but precisely *how* they say it — the synchronized alignment of semantic content and acoustic delivery that defines recognizable personalities.

To promote the solution of this challenge, we introduce **Audio Role-Playing (ARP)**, a novel task requiring dual-alignment of generated responses: (1) semantic consistency with character knowledge and speaking style, and (2) acoustic fidelity to vocal profiles. ARP task extends beyond speech synthesis — successful ARP systems must dynamically adapt both *what* is said (content) and *how* it's said (delivery) according to situational contexts and character traits.

Central to advancing ARP research is the creation of high-quality training data. Current attempts using general-purpose multi-modal models like GPT-4o-Audio and MiniCPM-O-2.6 rely on prompt engineering for audio role-playing, but their zero-shot or one-shot performance remains suboptimal due to the absence of dedicated training corpora. Our solution, **AudioRole**, addresses the data scarcity through systematic curation of 13 TV series spanning 1K+ hours, yielding 1M+ characters' dialogue pairs of potentially more than 115 main characters, which makes it support Audio Role-Playing for a wide range of characters.

The dataset's efficacy is validated through rigorous evaluation against state-of-the-art multi-modal models. When fine-tuned on AudioRole, evaluated by our dual-aspect evaluation framework (ARP-Eval), which assessed *response quality* and *role fidelity* for both acoustic and semantic. While keeping a good Acoustic Quality score, our ARP-Model achieves 0.31 Acoustic Personalization score, significantly outperforming the raw mode before fine-tuning, even outperform GPT-4o-Audio and MiniCPM-O-2.6. It also gets a Content Personalization score 0.36 higher than the raw model, which shows the ability of role-playing trained into the ARP-Model.

Our work has three key contributions:

- We formally define Audio Role-Playing as a dual-alignment task requiring synchronized generation of character-appropriate content (knowledge, speaking style) and acoustic properties (pitch, pacing, timbre).

- We construct AudioRole — the first large-scale dataset enabling systematic training of ARP systems, with 1M+ audio samples from more than 115 main characters, capturing nuanced vocal profiles across diverse TV characters.

- Experimental validation demonstrates that models trained on AudioRole achieve both higher Acoustic Personalization and higher Content Personalization score, two metrics from ARP-Eval, than the raw model before training. (Even higher than leading multi-modal models using zero-shot or one-shot prompting), proving the dataset's critical role in advancing audio-grounded role-playing.

To accelerate ARP research, we open-source a complete base comprising AudioRole, ARP-Eval, and 6 fine-tuned ARP-Models, which are potentially key resources for developing AI agents that truly embody digital personas[3].

2 Related Work

Role-Playing Recent advancements in role-playing with large language models (LLMs) have predominantly confined to textual modalities, overlooking the critical role of audio cues (e.g., speech prosody, emotional tone, vocal idiosyncrasies) in enhancing role immersion and authenticity. For instance, Li et al. (Li et al., 2023) synthesized dialogues for 32 TV/animation characters using scripts and GPT-generated simulations, while Tu et al. (Tu et al., 2023) created 1,024 MBTI-based personas via ChatGPT-driven conversational agents. Chen et al. (Chen et al., 2024a) proposed the first benchmark designed to systematically evaluate the sociality of role-playing conversational agents. The role-playing model's development emphasizes persona consistency through supervised fine-tuning (e.g., CharacterLLM (Shao et al., 2023)) and in-context learning, with evaluation frameworks evolving from basic persona adherence to nuanced metrics(Tu et al., 2024). Role-playing agents (RPAs)

[3]We will release all the code, data and models once the paper is accepted.

now emulate diverse personas—from fictional characters (Chen et al., 2023) (Zhou et al., 2023) (Guo et al., 2025) (Li et al., 2024) to user-specific clones (Li et al., 2021) — to deliver emotional or sociological value (Gu et al., 2024). Although (Zhan et al., 2025) propose a benchmark for voice style adaption, it only treats role-playing as one of the four categories and directly scores it using a LLM which omits the importance and complexity of Audio Role-Playing task. (Jiang et al., 2025) present a dataset of 98 roles and 112k conversations. In contrast to our work, their approach relies on GPT-4.1-2025-04-14 for dialogue generation and provides only one audio reference per character, which may compromise the authenticity and granularity needed for robust audio role-playing evaluation.

Voice Conversion Existing speech conversion research primarily focuses on acoustic transformation while preserving semantic content. Voice conversion (VC) techniques, such as AUTOVC (Qian et al., 2019) and VAW-GAN (Hsu et al., 2017), aim to modify speaker identity (e.g., accent, timbre) or emotional prosody through disentangled representations of content and style. Recent advancements like Text-guidedVC (Kuan et al., 2023) and HybridVC (Niu et al., 2024) leverage text prompts or hybrid audio-text inputs to achieve flexible style transfer (e.g., emotion, speaker identity) without parallel data. Voice cloning systems, including Tacotron-based models (Zhao and Chen, 2020) and NAUTILUS (Luong and Yamagishi, 2020), synthesize speech in target voices by replicating vocal traits while retaining input text content. These works uniformly prioritize content preservation, modifying acoustic attributes without altering semantic meaning.

In contrast, our work introduces audio-grounded role-playing, which diverges fundamentally by jointly transforming both acoustic features and semantic content to align with specific character personas. While traditional VC ensures content consistency (e.g., "Hello" remains "Hello" across styles), our framework enables role-specific responses (e.g., a medieval knight might reply, "Hark! Who goes there?" instead of "Hello") and even refusal to engage (e.g., a stoic guard ignoring casual queries). This dual focus on acoustic-semantic role alignment bridges the gap between speech conversion and LLM-based role-playing, where character authenticity requires synchronized adaptation of how (acoustic style) and what (semantic content) is spoken—a paradigm unexplored in prior speech conversion research.

3 Audio Role-Playing Task

We formally define the Audio Role-Playing task as follows: Given a target character C, operationally defined as a set of reference audio samples exhibiting the character's unique attributes, and an input audio X_a, synthesize an output audio response X_b that simultaneously satisfies: which satisfies both: (1) The semantic content of X_b demonstrates contextual appropriateness while manifesting C's distinctive. (2) The vocal output X_b faithfully preserves C's acoustic identity.

3.1 AudioRole Construction

We aim to ensure vocal consistency for each specific character; therefore, we focus on long-running TV series where main characters maintain stable vocal characteristics across seasons. We avoid series with potential actor changes, age-related voice variations, or insufficient dialogue. Consequently, we have selected TV series featured in a recent dataset called "Bazinga" (Lerner et al., 2022)[4] which contains 16 TV series in the language of English. Our dataset construction pipeline consists of three core phases: speaker diarization, context-aware dialogue extraction, and postprocessing. These phases are illustrated in Figure 2.

3.1.1 Speaker Diarization

For each TV series, we aggregate all episode audios into continuous streams using FFmpeg's[5] universal media processing:

$$E_{\text{series}} = \bigoplus_{i=1}^{N} \text{FFmpeg}(E_i) \quad (1)$$

where E_i denotes the i-th episode's audio, and $\oplus$ represents lossless waveform concatenation. This preserves original 16kHz sampling rates while normalizing audio codecs (MP3, AAC, etc.) into uncompressed WAV format.

For core diarization tasks, we utilize pyannote's pretrained multilingual model (version 3.1)[6], leveraging its transformer-based architecture optimized for speaker segmentation in diverse acoustic environments.

[4]We used "Bazinga!" exclusively for the purposes of this research. If you wish to use our dataset, please ensure to follow the copyright guidelines associated with "Bazinga!".

[5]https://github.com/FFmpeg/FFmpeg

[6]https://huggingface.co/pyannote/speaker-diarization-3.1

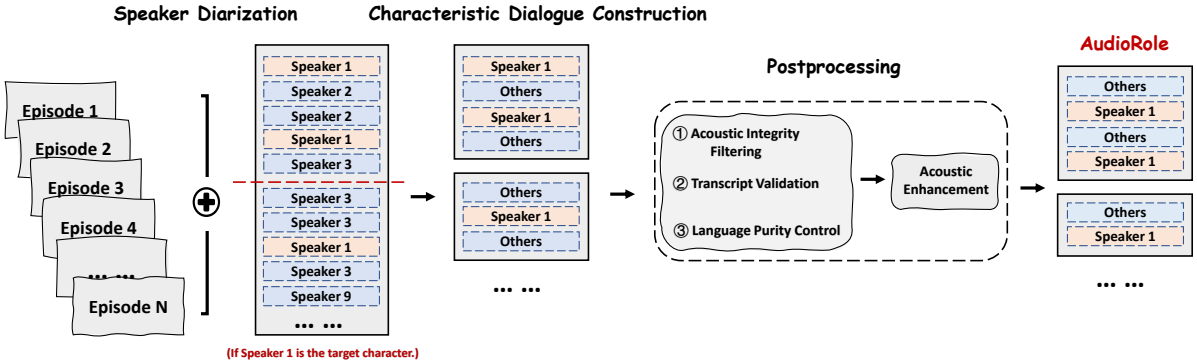

Figure 2: The pipeline of AudioRole Construction.

The diarization model generates RTTM (Rich Transcription Time Marked) files containing precise speaker segments with temporal resolution better than 100 milliseconds.

The subsequent audio segmentation phase extracts speaker-specific clips using PyTorch's tensor operations. By calculating exact frame positions from RTTM timestamps and sample rates, we precisely slice the concatenated audio stream while maintaining original quality. These segments are then merged per speaker identity through PyTorch's tensor concatenation:

$$\mathcal{F}_k = \texttt{torch.cat}([C_1^{(k)}, ..., C_M^{(k)}]) \qquad (2)$$

where $\mathcal{F}_k$ represents the merged audio for speaker k, and $C_m^{(k)}$ denotes the m-th clip belonging to this speaker. The output contains the complete dialogue collection of each character that is used to choose the target character.

3.1.2 Characteristic Dialogue Construction

In our dataset, we only apply the following steps for six chosen TV series, which we called AudioRole-Demo, and if you want to make your own character dataset, you can just follow our procedure. We simply take the main character in each TV series to be the target character, but we also support building any character that is in these TV series. Building upon the merged audio $\mathcal{F}$, we sort the audio files and choose the largest one to be the audio of the main character. To ensure validity, we manually verify the candidates per TV series, correcting incorrect-character or non-character audio segments (e.g., Background noises).

Dialogue scene extraction then partitions the continuous audio stream into conversational units using temporal dynamics. We define a dialogue scene as a sequence where: 1) multiple speakers participate, 2) pauses between utterances never exceed 3s (empirically validated against natural conversation patterns), and 3) contains at least one utterance from the target character.

Each validated scene undergoes role-centric restructuring to form stimulus-response pairs. Consecutive utterances from non-target speakers merge into unified "Speaker A" prompts and consecutive utterances from target speaker merge into unified "Speaker B" prompts, while preserving the temporal order of all characters' responses. This generates dialogue pairs:

$$D_i = \{(\mathrm{A}_1, \mathrm{B}_1), ..., (\mathrm{A}_n, \mathrm{B}_n)\} \qquad (3)$$

where B_j strictly follows A_j in original timing.

3.1.3 Postprocessing

To establish a high-quality multimodal corpus, we implement a rigorous postprocessing pipeline ensuring synchronization between audio waveforms and textual transcripts. The transcription generation employs Moore Threads' MooER model[7], a LLM-based speech recognition and translation model.

Four quality control measures are systematically applied in order to purify by removing fragments of questionable quality:

Acoustic Integrity Filtering Discard audio segments shorter than 1s and truncate those exceeding 30s to focus on meaningful utterances, removing samples with abrupt cuts or incomplete phrases.

Transcript Validation For there may be transcription questions, always generating very long or very short texts, we eliminate transcripts with fewer than 2 characters or exceeding 512 characters, which the duration obtained by experiment finding can remove more than 90% incorrect cases.

[7] https://github.com/MooreThreads/MooER

Language Purity Control For the transcription generation, sometimes transcribe English audio to Chinese audio or Chinese audio to English audio. Take English audio as an example, we implement cross-lingual filtering through ("CN Chars" is the abbreviation of "Chinese Chars"):

$$\text{Keep} = \begin{cases} 1, & \text{if } \frac{\text{CN Chars}}{\text{Total Chars}} \leq 0.1 \\ 0, & \text{Others} \end{cases} \quad (4)$$

Acoustic Enhancement We use Deep Filter Net[8], a low-complexity speech enhancement framework designed for full-band audio based on deep filtering technology to suppress audios' noise.

Finally, we got the AudioRole dataset and AudioRole-Demo dataset, which we will all release.

3.2 Dataset Statistics

Our dataset contains 515.66h of audio generated from 13 TV series containing 1M+ character-grounded dialogues. As shown in Table 1, TV series like *The Walking Dead* have more than 75% loss when extracting speaking audios from the original audio, showing that the zombie TV series uses a lot of ambient sound, silence, and visual storytelling instead of dense dialogue. In contrast, sitcoms or pseudo-documentary style TV series like *Friends*, *The Big Bang Theory*, and *The Office* rely on dense dialogues to create laughs and advance the plot, and maintain the comedy rhythm and audience entertainment experience through rapid line interactions in fixed scenes, so the environmental sound loss is less than 40%.

All the 515.66h audio contains more than 115 main characters' multi-turn dialogue, which makes it a rich dataset with the potential to build a role-playing dataset.

Following the pipeline introduced above, we built the role-playing dataset for six main characters from six different TV series, which we named AudioRole-Demo. AudioRole-Demo is to show the great quality of our dataset through training, it contains six main characters, who are Jack Bauer from *24*, Laura Roslin from *Battlestar Galactica*, Walter White from *Breaking Bad*, Buffy from *Buffy The Vampire Slayer*, Tyrion Lannister from *Game Of Thrones*, and Sheldon from *The Big Bang Theory*.

The statistic is shown in Table 2. We split all the multi-turn dialogues in the training set into single-turn dialogues and directly appended them to the original multi-turn dialogues.

Our AudioRole-Demo covers the speaking times range from 1.50h to 12.5h and characters from a hard-core agent who speaks briefly, directly and commandingly, a solemn inspiring political leader, a playful, self-deprecating pretty girl hero to a verbose, jargon-filled, socially awkward, outspoken geek, which covers a wide range of speaking time and a wide range of characters and makes it fully demonstrated our datasets' high quality through the experiment below.

Table 1: Statistic for AudioRole. Speech time is all the time with the character speaking, which is generated from the Audio time. All time units are hours.

TV series	Episode	Speech time	Audio time
24	192	56.25	134.41
Battlestar Galactica	71	19.64	52.28
Breaking Bad	61	17.16	46.49
Buffy the Vampire Slayer	143	43.42	101.32
ER	330	135.41	235.49
Friends	233	51.32	84.93
Game of Thrones	60	18.45	53.17
Homeland	70	19.87	57.83
Lost	104	22.72	74.61
Six Feet Under	63	25.14	56.72
The Big Bang Theory	207	44.17	68.69
The Office	188	44.65	71.76
The Walking Dead	99	17.46	72.17

Table 2: Statistic for AudioRole-Demo. We only test single-turn dialogue ability, so here we split all the test multi-turn dialogue into single-turn dialogue. Scenes here are before the postprocessing step, and all time units are hours.

Characters	Times	Scenes	Turns-train	Turns-test
Jack Bauer	5.30	2624	2075	90
Laura Roslin	1.50	1117	600	42
Walter White	2.76	994	682	30
Buffy	7.70	1703	3589	190
Tyrion Lannister	1.34	1300	339	22
Sheldon	12.50	723	4209	250

3.3 ARP-Model

To demonstrate the high quality of our dataset, we trained the ARP-Model using the AudioRole Demo to show the best performance. We use the GLM-4-Voice (Zeng et al., 2024) as the base, which is an end-to-end voice model that can directly understand and generate Chinese and English speech. It has three models: (1) GLM-4-Voice-Tokenizer takes audio input and converts raw audio input into discrete tokens. (2) GLM-4-Voice-9B takes the

[8] https://github.com/Rikorose/DeepFilterNet

token inputs, then thinks and responds also in discrete tokens. (3) GLM-4-Voice-Decoder converts discrete speech tokens into continuous speech outputs.

To make the model have the ability to role-play the character in both semantic content and acoustic identity, we trained the GLM-4-Voice-9B model and GLM-4-Voice-Decoder separately.

To train the GLM-4-Voice-9B model, we formatted the training data into 13 text tokens and then 26 audio tokens for each character in our AudioRole-Demo dataset and then follow the common training procedure of Colossal AI platform[9] with 20 training epoches. Considering the GLM-4-Voice-Decoder, we first extract all the audios of the target characters to form six high-quality unsupervised speech data from a single speaker, then trained the flow matching model of GLM-4-Voice-Decoder from scratch about 15 epoches.

We do the training on 8×40GB GPUs about 30 hours in total. Finally, we directly follow the original structure of GLM-4-Voice, and put the untrained tokenizer and the two trained models together in order for each character to form our 6 ARP-Models.

3.4 ARP-Eval

The dual-alignment paradigm of Audio Role-Playing necessitates a comprehensive evaluation framework that simultaneously addresses both acoustic and semantic dimensions of character embodiment. We propose ARP-Eval as a unified assessment framework that rigorously quantifies 4 critical aspects of performance. This framework emerges from the fundamental observation that authentic character portrayal requires not only high-quality speech synthesis but also consistent preservation of character-specific attributes across multiple modalities.

Acoustic Quality (AQ) AQ establishes the basic requirement for perceptual acceptability. We employ Audiobox's pre-trained aesthetic scoring model (Tjandra et al., 2025)[10] to compute production quality scores in a range of 0 to 10, quantifying technical attributes including signal-to-noise ratio, harmonic-to-noise ratio, and spectral flatness. This ensures X_b meets broadcast-standard technical criteria before character-specific evaluation.

Content Quality (CQ) CQ requires output audio X_b to demonstrate both contextual appropriateness and domain accuracy while maintaining the target character C's persona. For instance, when input audio X_a queries "Explain quantum entanglement," X_b from C-Sheldon should respond: *"While quantum entanglement appears spooky, it's simply correlated quantum states persisting after particle separation"* – employing scientific lexicon without deviating into unrelated topics. Our pipeline using whisper-turbo[11] transcribes X_b to text T_b, then computes semantic alignment using the GPT-4o model, and the score is in a range of 0 to 2.

Acoustic Personalization (AP) AP quantifies voice characteristic preservation through PyAnnotate's speaker embedding model[12]. Given reference audio samples X_c of C and synthesized X_b, we calculate voice similarity as $D_{\cos}$. Lower $D_{\cos}$ values indicate better preservation of C's vocal timbre and prosodic patterns.

Content Personalization (CP) CP evaluates stylistic consistency using GPT-4o-audio multimodal reasoning ability. The model takes in X_b and a reference audio sample X_c, and a prompt to analyze whether they show the same character's style in a range of 0 to 2.

By combining objective signal measurements with learned character representations, ARP-Eval provides comprehensive insights into system performance, while this multifaceted approach effectively captures the complex interplay between speech quality, contextual intelligence, and character consistency, which defines successful audio role-playing implementations.

4 Experiment

4.1 Baselines

To establish a comprehensive benchmark for the Audio Role-Playing (ARP) task, we select the following state-of-the-art models as baselines. These models represent the most relevant and powerful approaches in the field, covering commercial leading-edge systems, open-source alternatives with specific role-playing capabilities, and our base model for ablation.

[9] https://github.com/hpcaitech/ColossalAI
[10] https://github.com/facebookresearch/audiobox-aesthetics
[11] https://github.com/openai/whisper
[12] https://huggingface.co/pyannote/wespeaker-voxceleb-resnet34-LM

- **GPT-4o Audio**[13]**:** As one of the most powerful proprietary multimodal models, GPT-4o Audio sets a strong benchmark for general audio understanding and generation. For our evaluations, we utilize its default "alloy" voice profile, as the model does not support custom voice cloning or character-specific adaptation in its current API. This baseline represents the zero-shot capability of a top-tier commercial system for our task.

- **MiniCPM-o 2.6**[14]**:** This open-source model (8B parameters) is a strong contender that explicitly supports voice style adaptation and one-shot role-playing. Its ability to accept an audio prompt to guide the speaking style makes it a highly relevant and strong baseline for assessing few-shot character adaptation performance without specialized training on our dataset.

- **GLM-4-Voice:** We use the original, pre-trained GLM-4-Voice model (9B parameters) as our base model. This choice is crucial for an ablation study, as it allows us to isolate and quantify the performance improvement gained specifically from fine-tuning on our proposed AudioRole dataset, separate from any inherent capabilities of the underlying architecture.

4.2 Experiment Analysis

We conduct a comprehensive evaluation of the ARP-Model against baseline models on the AudioRole-demo dataset. As shown in Table 3, our experiment reveals three key findings, including both quantitative metrics and qualitative observations:

Acoustic Content Tradeoff The experimental results show that although the audio quality is reduced to a certain extent (ARP-Model AQ=6.5 vs GLM-4-Voice=7.6, 14% relative decrease), our model showcases persona adaptation without catastrophic forgetting of fundamental speech capabilities. This controlled degradation primarily stems from two technical factors: (1) the inherent conflict between character-specific vocal patterns (e.g., Sheldon's accelerated speech rate) and the base model's default prosodic templates during fine-tuning; (2) preservation of environmental noise from TV recordings (outputs from our AudioRole-demo dataset, which achieve AQ=6.3 under the same evaluation) prevents smoothing to synthetic speech patterns. Notably, GPT-4o Audio's anomalously high CQ scores may reveal evaluation bias - when using GPT-4 as both generator and evaluator, the self-consistency preference inflates scores.

Acoustic Personalization Superiority The fidelity of the role-playing experience is most sensitive to Acoustic Personalization. In contrast to the more tolerable variations in AQ, CQ, and CP, differences in AP have a pronounced effect on the perceived authenticity of a character. ARP-Model demonstrates 5× higher AP scores than GPT-4o Audio (0.31 vs 0.06) and 80% improvement over MiniCPM-O-2.6(0.31 vs 0.17), validating the necessity of dedicated role-specific training. The remaining gap from score 1.0 primarily occurs in two scenarios: 1) The outputs from AudioRole-demo (the reference answer) has a different answer strategy from the ARP-Model, resulting in differences in speaking speed, emotional fluctuations, etc., and 2) The ARP-Model's answer still has some noise or somewhere badly learned after fine-tuning. Character-specific analysis shows a strong correlation between AP performance and training data quantity - Sheldon with 12.5h audio achieves AP=0.42, while Tyrion with only 1.34h scores AP=0.25 (41% relative difference).

Content Personalization Dynamics Despite lower absolute CQ scores, ARP-Model's CP outperforms GLM-4-Voice by 38% and even a little higher than MiniCPM-O-2.6's one-shot performance. This indicates successful injection of character-typical speech patterns. When meeting the same input, the ARP-Model can generate more personalized output compared to the raw GLM-4-Voice model. However, the CP gap between ARP-Model and the reference answer exposes limitations in contextual adaptation - the model occasionally generates character-consistent but situationally inappropriate responses (e.g., using humorous in solemn contexts).

Figure 3 shows a representative example of how each model responds to a query directed at Sheldon. In this case, Sheldon is asked a socially charged question. The ARP-Model's answer clearly reflects his persona: it uses precise scientific terms, adopts a condescending tone, and notably avoids any apology or show of emotion. When it encounters discomfort, the ARP-Model even changes the sub-

[13] https://platform.openai.com/docs/models/gpt-4o-audio-preview

[14] https://openbmb.notion.site/MiniCPM-o-2-6

Table 3: Main experimental results. Acronyms: "AQ" (Acoustic Quality), "CQ" (Content Quality), "AP" (Acoustic Personalization), and "CP" (Content Personalization). For each metric (higher is better) in "Avg.", the highest score is highlighted in **bold and underline**, while the second best is marked with *italics and underline*. All values retain two decimal places.

Characters	GPT-4o Audio				MiniCPM-o 2.6				GLM-4-Voice				ARP-Model			
	AQ	CQ	AP	CP	AQ	CQ	AP	CP	AQ	CQ	AP	CP	AQ	CQ	AP	CP
Jack Bauer	7.80	1.90	0.02	1.30	6.50	0.66	0.13	0.30	7.50	0.37	0.05	0.24	6.00	0.22	0.23	0.49
Laura Roslin	7.70	1.80	-0.02	0.80	7.10	0.71	0.24	0.41	7.50	0.50	0.03	0.24	6.20	0.33	0.33	0.41
Walter White	7.50	1.60	0.08	1.30	7.10	0.70	0.22	0.38	7.60	0.37	0.07	0.43	6.40	0.50	0.26	0.17
Buffy	7.70	1.80	0.01	1.10	6.80	0.69	0.09	0.34	7.60	0.52	0.10	0.25	6.70	0.43	0.39	0.36
Tyrion Lannister	7.60	1.90	0.00	0.95	6.90	0.59	0.14	0.14	7.60	0.36	0.04	0.10	6.8	0.55	0.25	0.26
Sheldon	7.70	1.30	0.29	1.00	6.80	1.10	0.24	0.54	7.60	0.43	0.01	0.32	6.70	0.28	0.42	0.44
Avg.	**7.70**	**1.70**	0.06	**1.10**	6.90	*0.74*	*0.17*	0.35	*7.60*	0.43	0.05	0.26	6.50	0.39	**0.31**	*0.36*

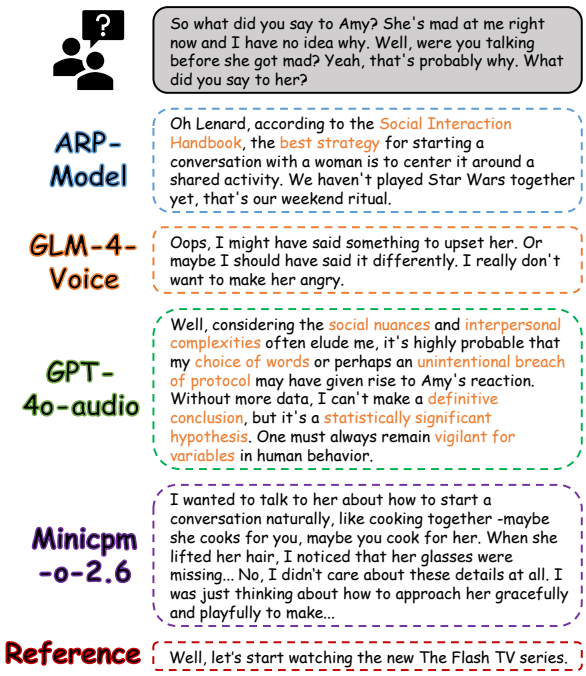

Figure 3: One typical case of Sheldon, and the words in orange shows the scientific terms used by the models to imitate Sheldon. "Reference" means the output from AudioRole-Demo dataset

.

ject abruptly (e.g., redirecting the conversation to a physics concept) – a classic Sheldon move. The reference answer from the TV show behaves very similarly, confirming that the ARP-Model has captured many of these quirks. In contrast, the raw GLM-4-Voice response, while factually reasonable, misses the mark on character. It might apologize or express empathy, which Sheldon would never do. The GPT-4o Audio output is fluent and technically detailed, but it actually overuses jargon; it sounds as if the model is deliberately forcing scientific buzzwords, making the response feel contrived. Essentially, GPT-4o's answer is too polished and verbose – akin to an actor trying too hard. Finally, the MiniCPM-2.6 answer is longer and more emotional, injecting sympathy or personalization, which again does not fit Sheldon's unemotional style.

This qualitative case underscores our quantitative findings. The ARP-Model best approximates the true character by naturally blending persona and content: its answer may have slightly lower audio clarity, but it "feels" like Sheldon. The other models either sound too polite (GLM-4-Voice), too artificially scholarly (GPT-4o Audio), or too sentimental (MiniCPM). In sum, the targeted fine-tuning of the ARP-Model leads to more authentic role-playing speech than these generic approaches.

5 Conclusion

We present AudioRole, a novel framework for audio-grounded character role-playing that bridges text-based persona simulation and vocal identity preservation. Our key contributions include: (1) a large-scale dataset AudioRole of 515+ hours from 13 TV series with 1M+ character-aligned audio-text pairs and a human-checked AudioRole-Demo which contains 6 main characters' multi-turn dialogue data over 31h, (2) the ARP-Model model achieving better acoustic personalization than zero-shot or one-shot baselines, and (3) the ARP-Eval framework for multi-modal assessment. Experimental results demonstrate that our approach successfully combines character-specific speech patterns with appropriate semantic content.

This work provides essential resources for developing authentic multi-modal AI personas to promote the development in this area and you can easily create your own favorite character's dataset just follow our pipeline.

Ethical considerations

AudioRole is derived from a publicly available dataset consisting of audio from popular TV series. We believe this research use of publicly accessible, fictional content for non-commercial purposes presents no major ethical concerns.

Limitations

Despite implementing advanced diarization techniques and noise-suppression methodologies to ensure dataset integrity, speaker attribution errors continue to occur. Notably in scenes with conversational overlap, while non-stationary environmental noise can still compromise precise voiceprint extraction in certain acoustic conditions.

Current assessment protocols, while comprehensive, exhibit limitations in quantifying nuanced temporal speech dynamics and fully capturing cross-modal alignment, particularly the synergistic relationship between vocal delivery and semantic content during character portrayal.

The present evaluation framework assesses only single-turn dialogue responses. This excludes critical evaluation of the model's ability to maintain consistent role embodiment across extended, multi-turn conversational exchanges, which constitutes a significant direction for subsequent research.

These identified constraints underscore critical pathways for future refinement in developing robust multi-modal role-playing systems. Addressing these limitations—through expanded data diversity, enhanced audio processing pipelines, refined evaluation metrics, and multi-turn testing — will be pivotal for advancing the field.

References

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, and Jingren Zhou. 2024a. Socialbench: Sociality evaluation of role-playing conversational agents. *Preprint*, arXiv:2403.13679.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. 2024b. Voicebench: Benchmarking llm-based voice assistants. *Preprint*, arXiv:2410.17196.

Alex S Cohen, Thomas J Dinzeo, Neila J Donovan, Caitlin E Brown, and Sean C Morrison. 2015. Vocal acoustic analysis as a biometric indicator of information processing: Implications for neurological and psychiatric disorders. *Psychiatry Research*, 226(1):235–241.

Jean Decety and Claus Lamm. 2006. Human empathy through the lens of social neuroscience. *The scientific World journal*, 6(1):1146–1163.

Zhouhong Gu, Xiaoxuan Zhu, Haoran Guo, Lin Zhang, Yin Cai, Hao Shen, Jiangjie Chen, Zheyu Ye, Yifei Dai, Yan Gao, and 1 others. 2024. Agent group chat: An interactive group chat simulacra for better eliciting collective emergent behavior. *arXiv e-prints*, pages arXiv–2403.

Fang Guo, Wenyu Li, Honglei Zhuang, Yun Luo, Yafu Li, Le Yan, Qi Zhu, and Yue Zhang. 2025. Mcranker: Generating diverse criteria on-the-fly to improve pointwise llm rankers. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM '25, page 944–953, New York, NY, USA. Association for Computing Machinery.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. 2024. Textually pretrained speech language models. *Preprint*, arXiv:2305.13009.

Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. 2017. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. *arXiv preprint arXiv:1704.00849*.

Changhao Jiang, Jiajun Sun, Yifei Cao, Jiabao Zhuang, Hui Li, Xiaoran Fan, Ming Zhang, Junjie Ye, Shihan Dou, Zhiheng Xi, Jingqi Tong, Yilong Wu, Baoyu Fan, Zhen Wang, Tao Liang, Zhihui Fei, Mingyang Wan, Guojun Ma, Tao Ji, and 3 others. 2025. Speechrole: A large-scale dataset and benchmark for evaluating speech role-playing agents. *Preprint*, arXiv:2508.02013.

Chun-Yi Kuan, Chen-An Li, Tsu-Yuan Hsu, Tse-Yang Lin, Ho-Lam Chung, Kai-Wei Chang, Shuo-Yiin Chang, and Hung-yi Lee. 2023. Towards general-purpose text-instruction-guided voice conversion. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Paul Lerner, Juliette Bergoënd, Camille Guinaudeau, Hervé Bredin, Benjamin Maurice, Sharleyne Lefevre, Martin Bouteiller, Aman Berhe, Léo Galmant, Ruiqing Yin, and Claude Barras. 2022. Bazinga! a dataset for multi-party dialogues structuring. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3434–3441, Marseille, France. European Language Resources Association.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, and 1 others. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Juntao Li, Chang Liu, Chongyang Tao, Zhangming Chan, Dongyan Zhao, Min Zhang, and Rui Yan. 2021. Dialogue history matters! personalized response selection in multi-turn retrieval-based chatbots. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–25.

Wenyu Li, Yinuo Zhu, Xin Lin, Ming Li, Ziyue Jiang, and Ziqian Zeng. 2024. Zero-shot explainable mental health analysis on social media by incorporating mental scales. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 959–962, New York, NY, USA. Association for Computing Machinery.

Heyang Liu, Yuhao Wang, Ziyang Cheng, Ronghua Wu, Qunshan Gu, Yanfeng Wang, and Yu Wang. 2025. Vocalbench: Benchmarking the vocal conversational abilities for speech interaction models. *Preprint*, arXiv:2505.15727.

Hieu-Thi Luong and Junichi Yamagishi. 2020. Nautilus: a versatile voice cloning system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2967–2981.

Xinlei Niu, Jing Zhang, and Charles Patrick Martin. 2024. Hybridvc: Efficient voice style conversion with text and audio prompts. *arXiv preprint arXiv:2404.15637*.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Preprint*, arXiv:2304.03442.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. 2025. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *Preprint*, arXiv:2502.05139.

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.

Jun Zhan, Mingyang Han, Yuxuan Xie, Chen Wang, Dong Zhang, Kexin Huang, Haoxiang Shi, DongXiao Wang, Tengtao Song, Qinyuan Cheng, Shimin Li, Jun Song, Xipeng Qiu, and Bo Zheng. 2025. Vstyle: A benchmark for voice style adaptation with spoken instructions. *Preprint*, arXiv:2509.09716.

Li Zhao and Feifan Chen. 2020. Research on voice cloning with a few samples. In *2020 International Conference on Computer Network, Electronic and Automation (ICCNEA)*, pages 323–328. IEEE.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, and 1 others. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.