HEAR: HUMAN-LIKENESS EVALUATION WITH ACOUSTIC REASONING FOR GENERATIVE SPEECH

Anonymous authors
Paper under double-blind review

ABSTRACT

Assessing the perceived human-likeness of generated speech is important but challenging. Listening tests are costly to scale, quality-oriented metrics do not directly measure this target, and raw-audio AudioLLM judges offer limited explicit acoustic evidence. We propose HEAR, a training-free method for Human-likeness Evaluation with Acoustic Reasoning. HEAR uses off-the-shelf tools and explicit rules to convert acoustic measurements into semantic tags organized at global, word, and segment levels. This representation links measured behavior to its location and surrounding speech. A dedicated human-likeness prompt guides an AudioLLM to judge the original audio using the target text and evidence, producing an overall human-likeness score, eight dimension-specific scores, and evidence-grounded reasons. On Humanlike-100, HEAR raises Qwen3-Omni-Thinking's Pearson and Spearman correlations with mean human overall ratings to $2.42\times$ and $2.07\times$ those of matched audio-only judging, respectively. HEAR improves overall agreement with human ratings across all five evaluated AudioLLM backbones. Ablation studies support the contributions of structured acoustic evidence and the human-likeness evaluation rubric. Beyond human-likeness evaluation, task-adapted HEAR improves audio-only judging on speech preference and perceptual-quality benchmarks. Code is available at `https://github.com/HEAR-6/HEAR`.

1 INTRODUCTION

Clear and fluent speech can still sound mechanical, highlighting the distinction between intelligibility and perceived human-likeness. Recent text-to-speech (TTS) models have improved intelligibility, acoustic quality, and speaker similarity (Anastassiou et al., 2024; Zhang et al., 2024; 2025; Du et al., 2024; 2025; Liao et al., 2024; 2026), sharpening the need to evaluate perceived human-likeness: how closely an utterance sounds like a person speaking naturally (Wang et al., 2025b). In this work, the evaluation rubric emphasizes natural conversational delivery. This impression depends on how speech is delivered, including its timing, emphasis, and vocal details. A repeated word may convey hesitation or sound like a synthesis loop; emphasis may highlight the intended focus or interrupt speech flow.

Interpreting these cues requires context: judging a pause involves its duration, position, and surrounding prosody. Listening tests integrate this information but are costly to scale (Lameris et al., 2023; Baird et al., 2018). Automatic quality metrics summarize specific properties (Seebauer et al., 2023), leaving a need to connect local acoustic behavior with overall human-likeness.

Audio large language models (AudioLLMs) offer rubric-based judgment and natural-language explanations (Wang et al., 2025b; Shirali-Shahreza & Penn, 2026; Manakul et al., 2026; Zhao et al., 2026). They commonly align encoded acoustic features with a language model (Chu et al., 2023; Yang et al., 2025). The judge must recover local timing and vocal detail from these features and interpret them in context. Precise event localization, multidimensional speech understanding, and human-aligned assessment remain challenging (Shi et al., 2026; Li et al., 2026; Zhang et al., 2026b).

We investigate whether combining direct listening with explicit acoustic evidence organized across temporal scales helps AudioLLMs make human-likeness judgments that better align with human perception.Our contributions are threefold:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- **Human-likeness evaluation with acoustic reasoning.** We propose HEAR, a framework that combines tool-extracted acoustic evidence with direct AudioLLM listening to evaluate the perceived human-likeness of generated speech. A dedicated rubric guides the joint assessment of the original audio, target text, and evidence, producing an overall human-likeness score, eight dimension-specific scores, and evidence-grounded reasons.
- **Temporally organized acoustic evidence.** We design a rule-based representation that maps selected acoustic measurements to descriptive tags and organizes them at global, word, and segment levels. This representation links local acoustic cues to their temporal locations and surrounding speech while retaining utterance-wide summaries, providing explicit evidence for contextual judgment.
- **Empirical validation across backbones and tasks.** On Humanlike-100, HEAR achieves higher Pearson and Spearman correlations with human overall ratings than matched audio-only controls across all five evaluated AudioLLM backbones. Beyond human-likeness evaluation, HEAR improves preference accuracy on SpeechJudge-Eval, SpeakBench, and S2S-Arena, as well as system-level correlations with human MOS on TTSDS2, over matched audio-only judging.

2 Related Work

2.1 Speech Evaluation Metrics and Protocols

Automatic speech metrics quantify complementary properties: mean opinion score (MOS) predictors estimate perceptual quality (Lo et al., 2019; Reddy et al., 2021; Huang et al., 2022), automatic speech recognition (ASR)-based metrics assess intelligibility, and speaker embeddings assess identity similarity (Deja et al., 2022). TTSDS and TTSDS2 compare synthetic and real speech distributions across prosody, speaker identity, and intelligibility (Minixhofer et al., 2024; 2026). TTScore uses text-conditioned prediction of discrete speech tokens to evaluate intelligibility and prosody (Ulgen et al., 2026). Audio Turing Test uses multidimensional human judgments to assess whether generated Chinese speech sounds human (Wang et al., 2025b).

Automatic metrics summarize selected properties or system-level distributions, leaving local acoustic behaviors underlying human-likeness largely implicit. Listening protocols assess perceived human-likeness directly, but ratings remain costly to collect at scale.

2.2 AudioLLM-as-Judge for Speech Evaluation

Some methods prompt pretrained AudioLLMs for assessment. AudioJudge studies audio characteristic detection and preference simulation (Manakul et al., 2026). Audio-aware judges assess emotion, pace, emphasis, pitch, and non-verbal elements (Chiang et al., 2025); EmergentTTS-Eval evaluates expressive, prosodic, and linguistic scenarios with model-based judges (Manku et al., 2025).

Others adapt models using evaluation data. Auto-ATT fine-tunes an AudioLLM on human-likeness judgments (Wang et al., 2025b); SpeechJudge learns a generative reward model from naturalness preferences (Zhang et al., 2026b); JASTIN trains an audio adapter and fine-tunes an LLM with evaluation instructions (Zhang et al., 2026a).

In direct audio judging, the prompt specifies evaluation criteria, but the judge is not additionally supplied with acoustic measurements extracted by external tools. Supervised adaptation can improve alignment but requires task-specific data and training. Localizing within-utterance naturalness remains challenging (Zhang et al., 2026b).

2.3 Evidence-Grounded Reasoning for Audio Evaluation

Music-perception experiments report stronger performance with symbolic inputs than with audio, suggesting that explicit representations can support language-model reasoning on the evaluated tasks (Carone et al., 2026). For speech-to-speech evaluation, TRACE constructs textual blueprints from extracted audio cues, which a text-only LLM uses to assess Content, Voice Quality, and Paralinguistics (Chandra et al., 2026). TRACE combines these decisions into an overall preference using a deterministic rule.

Explicit representations make selected acoustic observations available for language-model reasoning. However, a text-only judge relies on the cues retained in the representation and on the accuracy of their extraction, without direct access to the recording to check those cues. These studies focus on music understanding and speech-to-speech preference; evidence-grounded human-likeness assessment remains less explored.

3 METHOD

3.1 OVERVIEW

For human-likeness evaluation, HEAR takes an utterance $\hat{a}_i$ and its intended text y_i and produces $O_i = (H_i, \mathbf{D}_i, R_i)$: an overall perceived human-likeness score H_i, eight diagnostic scores $\mathbf{D}_i$, and reasons R_i for each score. HEAR separates acoustic measurement from perceptual interpretation (Figure 1). Stage 1 measures acoustic properties and locates words and sound events in the recording, producing observations Z_i. Stage 2 converts selected measurements into categorical tags and groups observations into word, segment, and utterance-level records, producing evidence $E_i = f_{\text{rule}}(Z_i)$. Stage 3 supplies the original audio, intended text, and this evidence to an AudioLLM judge under the human-likeness prompt P_{HL}. The case study in Appendix A follows one utterance through the pipeline and compares its model and human ratings.

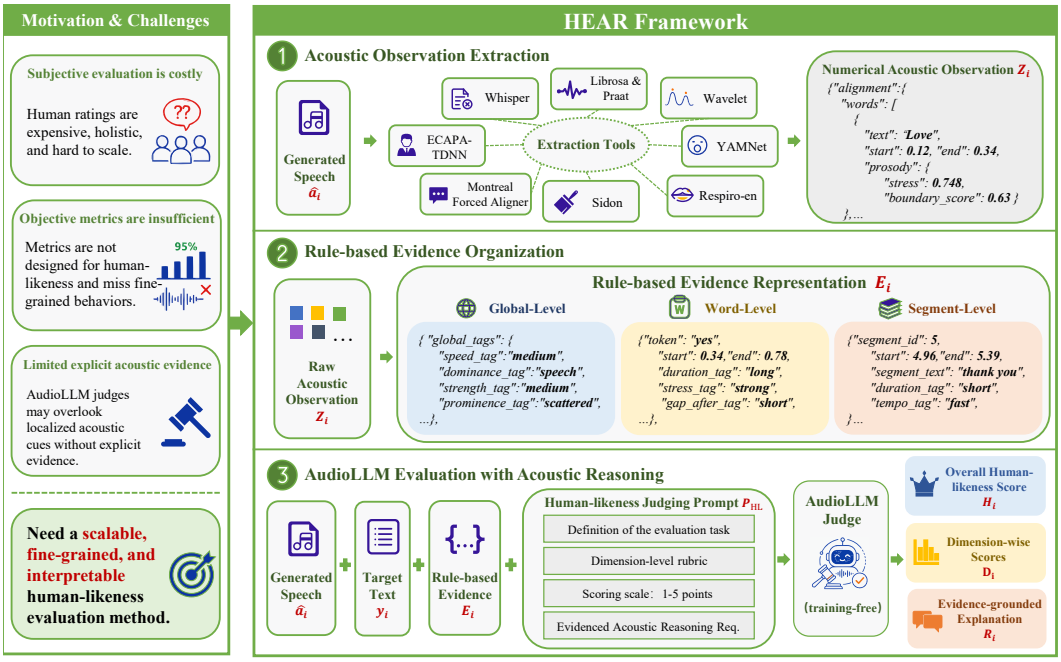


Figure 1: Overview of HEAR: acoustic observation extraction, rule-based evidence organization, and AudioLLM evaluation with acoustic reasoning.

3.2 ACOUSTIC OBSERVATION EXTRACTION

Stage 1 uses speech-analysis tools to obtain observations Z_i for evidence construction. The outputs retain acoustic measurements and timing information, from word intervals and event frames to utterance-wide summaries. Tools process mono 16-kHz copies; the judge receives the original recording.

Whisper large-v3 (Radford et al., 2022) transcribes the recording, and Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) aligns this recognized transcript with the audio to obtain word and phoneme timestamps. Each word record contains a recognized token and its start and end times. We use these records to measure word durations and inter-word gaps and to identify repetitions in the token sequence. Wavelet prosody analysis (Suni et al., 2017) assigns prominence and boundary

strengths to the aligned words, describing local acoustic emphasis and prosodic separation, respectively.

We also measure acoustic properties and detect sound events. Librosa (McFee et al., 2015) provides spectral summaries, including means and standard deviations of mel-frequency cepstral coefficients (MFCCs); Praat (Boersma & Weenink, 2026) supplies harmonic-to-noise ratio (HNR) for voice quality. Respiro-en (Yang et al., 2024) detects breath intervals and durations. YAMNet (Plakal & Ellis, 2020) analyzes 0.96-s windows at 0.48-s hops, retaining timestamps and the five highest-scoring event labels per frame. Together, these capture utterance-wide signal characteristics and localized vocal activity.

For speaker stability, Sidon (Nakata et al., 2025) enhances audio in the speaker-analysis branch to reduce non-speech interference in speaker embeddings. ECAPA-TDNN (Desplanques et al., 2020) extracts embeddings from this audio using 6-s windows with a 3-s hop. Cosine similarity between adjacent-window embeddings characterizes local speaker stability.

3.3 Rule-based Evidence Organization

Stage 2 organizes observations Z_i into evidence E_i at three temporal levels: global tags G_i, word records W_i, and segment records S_i. Word records store each recognized token together with its timestamps and tags for duration, following gap, prominence, and boundary strength. Segment records group consecutive words using pauses and prosodic boundaries, summarizing local patterns together with associated sound-event and speaker-stability cues. Global tags summarize timing, prosody, acoustic properties, and speaker stability across the recording. Together, these levels allow the judge to examine individual cues within their local context and the utterance as a whole.

From measurements to word tags. For word j with interval $[s_j, e_j]$, we compute duration $d_j = e_j - s_j$ and following gap $g_j = \max(0, s_{j+1} - e_j)$. Field-specific thresholds map duration, gap, prominence, and boundary strength to semantic categories. Records retain tokens, timestamps, and tags. Adjacent tokens identical ignoring case receive a repeat flag when $g_j < 0.1\,\text{s}$. The judge interprets repetition using the audio and intended text. Detailed conversion rules are provided in Appendix C.1.

Segment-level evidence. We group consecutive words into segments. A segment ends when the following gap is at least $0.3\,\text{s}$ or boundary strength exceeds 0.8; the final word closes the last segment. Each segment spans its first word's start to its last word's end. We summarize its word count, duration, speaking rate, prominence, repetition, final boundary, and following pause, and retain words with strong emphasis or boundary cues.

We associate event and speaker observations with each segment. For YAMNet, we sum each event label's scores across frames within the segment and select the highest-scoring label after excluding `Speech`. Each speaker comparison is assigned to adjacent windows' shared 3-s interval. When a segment overlaps several intervals, it receives the lowest-stability tag.

Global summaries. Global tags summarize utterance duration and speaking rate, mean prominence, maximum boundary strength, the proportion of prominent tokens, detected breath duration as a fraction of utterance duration, and overall speaker stability, together with acoustic observations.

3.4 AudioLLM Evaluation with Acoustic Reasoning

The AudioLLM evaluates the original recording alongside the intended text and organized acoustic evidence under the human-likeness prompt P_{HL}. The intended text provides a reference for interpreting omissions, disfluencies, and corrections, while the evidence supplies acoustic measurements and the locations of relevant cues. The prompt instructs the judge to base its assessment primarily on the audio and to prioritize what it hears when the auxiliary evidence conflicts with the recording.

We select eight dimensions to capture how human-likeness is conveyed through prosody, vocal behavior, and consistency across the utterance (Suni et al., 2017; Schuller et al., 2013; Cong et al., 2021; Perrotin et al., 2025). Rhythm & Pacing and Stress & Intonation address how timing and emphasis organize spoken content. Paralinguistic Cues and Non-lexical Vocal Sounds capture ex-

pressive events, such as fillers and laughter, and physiological vocal details, such as breathiness and tension. Identity Consistency and Acoustic Realism assess whether the voice and acoustic setting remain coherent across the utterance. Difficult-content Handling assesses sustained delivery through complex passages, while Error Correction addresses local self-monitoring and repair. This design connects local vocal details with utterance-wide coherence, providing complementary perceptual perspectives for the holistic judgment.

The judge assigns an overall human-likeness rating alongside eight dimension scores, all integers on a 1–5 scale, with an evidence-grounded reason for each score. The overall rating integrates the listening impression across these perspectives. The full prompt, dimension criteria, and output schema are provided in Appendix B.2.

4 Experimental Setup

4.1 Evaluation Tasks and Datasets

Humanlike-100. To evaluate perceived human-likeness in speech, we design Humanlike-100, a test set containing 50 human recordings and 50 generated utterances, with five generated utterances from each of ten TTS systems. The represented systems are Qwen (Alibaba Cloud, 2026), ElevenLabs (ElevenLabs, 2026), Gemini (Google, 2026), OpenAI TTS (OpenAI, 2026), MiniMax (MiniMax, 2026), IndexTTS (Deng et al., 2025), CosyVoice3 (Du et al., 2025), F5-TTS (Chen et al., 2024), HiggsAudio (Boson AI, 2026), and ZipVoice (Zhu et al., 2025). Target texts cover Chinese, English, and mixed-script utterances. Recordings last 1.48–31.95 seconds (mean 10.02 seconds).

Ten annotators rate overall human-likeness and eight dimensions on a 1–5 scale, producing 9,000 ratings. The rubric emphasizes natural conversational delivery. Detailed annotation instructions are provided in Appendix B.1.

HEAR-H data resource. To study how speech generators realize the behaviors assessed by HEAR, we construct HEAR-H: 2,702 prompt–target pairs covering expressive speech, paralinguistic behavior, and difficult linguistic content. Shared audio prompts and target texts support comparisons across TTS systems under common generation conditions. Appendix C.4 describes its design and data format.

General speech evaluation. Humanlike-100 evaluates the perceived human-likeness of individual utterances under a dedicated rubric. We use three external benchmark groups to test whether HEAR's acoustic evidence also helps AudioLLMs align with human judgments under different evaluation targets and decision formats.

SpeechJudge-Eval (Zhang et al., 2026b) tests pairwise naturalness judgment between two synthesized readings of the same target text. We evaluate 1,000 pairs by comparing the judge's selected candidate with the human preference.

The *TTSDS2 benchmark* (Minixhofer et al., 2026) evaluates speech naturalness using human mean opinion scores (MOS) on a five-point scale. It covers four domains: clean read speech (Clean), read speech without signal-to-noise filtering (Noisy), speech with varied recording conditions and speaking styles (Wild), and child speech (Kids). We evaluate 460 utterances from 20 TTS systems and measure agreement between predicted and human scores after averaging within each system. This tests whether evidence supports graded naturalness judgments and system ranking across recording conditions and speaker populations.

SpeakBench and S2S-Arena evaluate instruction-conditioned spoken responses. SpeakBench includes instructions specifying content, vocal style, or paralinguistic behavior, while S2S-Arena covers instruction following, speech perception, and conversational scenarios. We use the TRACE annotations (Chandra et al., 2026) for 497 SpeakBench pairs and 314 S2S-Arena pairs. Each example contains a spoken instruction and two responses. The Content, Voice Quality, and Paralinguistics judgments yield a four-way preference, testing whether evidence helps evaluate speech delivery in the context of the requested response.

4.2 AUDIOLLM BACKBONES AND INFERENCE

On Humanlike-100, we evaluate HEAR with five AudioLLM backbones: Qwen2.5-Omni, Qwen3-Omni-Instruct, Qwen3-Omni-Thinking, Gemini-3-Flash, and Gemini-3.5-Flash (Xu et al., 2025a;b; Google DeepMind, 2025). Each matched **audio-only** control uses the same backbone, audio, target text, rubric, and response format as HEAR, with acoustic evidence omitted. We average each score over the first three schema-valid outputs per utterance. Appendix C.2 details response validation, tool versions, and inference settings.

For general speech evaluation, HEAR and the corresponding audio-only controls use Qwen3-Omni-Thinking on SpeechJudge-Eval and TTSDS2, and Gemini-3.5-Flash on SpeakBench and S2S-Arena. Each comparison uses the same samples and task-specific scoring criteria. Detailed descriptions of task inputs, prompts, evidence use, and score aggregation are provided in Appendix B.3.

4.3 MATCHED CONTROLS AND ABLATION DESIGN

All diagnostic ablations use Qwen3-Omni-Thinking on Humanlike-100. To examine the contribution of selected evidence, we remove records or fields from the saved input supplied to the judge while keeping acoustic extraction fixed. Record removals omit entire word, segment, or tool-observation records; field removals omit selected attributes within the saved evidence.

We also compare HEAR's structured evidence with alternative representations. Raw JSON supplies the raw extraction output in place of the evidence organized by Stage 2. Natural-Instruct and Natural-Thinking use the corresponding Qwen3-Omni models to convert evidence into English descriptions, with Qwen3-Omni-Thinking retained as the judge. These comparisons change both how the observations are presented and which information is retained. Prompt variants omit individual rubric dimensions. Output variants retain the overall score (O) while varying whether dimension scores (D), reasons (R), or both are also requested.

4.4 METRICS AND BASELINES

Metrics. Utterance-level correlations measure agreement between predicted scores and mean human ratings for individual recordings. We compute them for human-likeness ratings on Humanlike-100 and naturalness MOS on TTSDS2. System-level correlations compare one predicted score per TTS system with that system's mean human rating. We use them on the TTSDS2 benchmark and for the distribution-based TTSDS2 baseline on Humanlike-100. Both levels use Pearson r for linear association and Spearman ρ for rank agreement.

Preference accuracy is the fraction of audio pairs for which the predicted decision matches the human label. On SpeechJudge-Eval, there are two possible decisions: speech A is more natural or speech B is more natural. On SpeakBench and S2S-Arena, there are four: prefer response A, prefer response B, both responses are good, or both responses are bad under the task's evaluation criteria.

Human reference. For Humanlike-100, the human reference averages correlations between each annotator's scores and the mean scores of the remaining annotators:

$$\mathrm{HC} = \frac{1}{N}\sum_{j=1}^{N} \mathrm{corr}\left(\{s_{i,j}\}_{i=1}^{M}, \left\{ \frac{1}{N-1} \sum_{k \neq j} s_{i,k} \right\}_{i=1}^{M} \right), \tag{1}$$

where $M = 100$, $N = 10$, and $s_{i,j}$ is annotator j's rating of utterance i. For overall ratings, ordinal Krippendorff's α is 0.614 and absolute-agreement ICC(A,10) is 0.950.

Baselines. We compare HEAR with four groups of baselines: recognition-error and perceptual-quality measures, including WER, SQUIM, DNSMOS, UTMOS, UTMOSv2, NISQA, and Audiobox Aesthetics (Kumar et al., 2023; Reddy et al., 2022; Saeki et al., 2022; Baba et al., 2024; Mittag et al., 2021; Tjandra et al., 2025); the TTSDS2 metric, which compares generated and reference speech distributions (Minixhofer et al., 2026); specialized speech evaluators Auto-ATT, SpeechEval, EmergentTTS-Eval, SpeechJudge-BTRM, and TRACE (Wang et al., 2025b;a; Manku et al., 2025; Zhang et al., 2026b; Chandra et al., 2026); and matched audio-only AudioLLM judges. All baselines are reproduced locally. Table 1 and Figure 2 show the comparisons for each task.

5 Results and Analysis

5.1 Main Results

Adding HEAR improves both overall Pearson and Spearman correlations across all five AudioLLM backbones (Table 1). With Qwen3-Omni-Thinking, correlations rise from $0.197/0.253$ for audio-only judging to $0.477/0.524$. The gains extend to most diagnostic dimensions for each backbone; Qwen3-Omni-Thinking improves in all eight, most notably Stress & Intonation ($0.240 \rightarrow 0.457$).

Table 1: Correlation with human ratings on Humanlike-100. Bold/underline: best/second-best utterance-level results, excluding human reference.

	Overall		Dimension-wise Spearman ρ							
Evaluator	r	ρ	**R&P**	**Para.**	**Non.**	**Iden.**	**Stress**	**Acous.**	**Diff.**	**Error**
Human reference†	.796	.759	.716	.696	.715	.589	.707	.661	.621	.653
Objective metrics										
WER	.159	-.054	-.106	-.077	-.116	-.007	-.107	-.060	-.033	.035
SQUIM	.066	.076	.100	.116	.022	.169	.029	.024	.095	.058
DNSMOS	-.247	-.328	-.257	-.258	-.285	-.289	-.252	-.248	-.162	-.276
UTMOS	-.279	-.300	-.291	-.252	-.252	-.226	-.216	-.169	-.175	-.231
UTMOSv2	-.290	-.342	-.279	-.237	-.251	-.294	-.198	-.186	-.210	-.278
NISQA	-.214	-.203	-.268	-.176	-.207	-.255	-.188	-.126	-.154	-.218
AE-CE	-.239	-.223	-.161	-.147	-.204	-.160	-.106	-.111	-.192	-.231
AE-CU	-.281	-.262	-.207	-.198	-.241	-.207	-.153	-.163	-.248	-.289
AE-PQ	-.331	-.338	-.237	-.284	-.300	-.241	-.200	-.227	-.269	-.312
Speech evaluators										
SpeechEval	-.235	-.301	-.241	-.198	-.287	-.141	-.265	-.191	-.244	-.272
EmergentTTS-Eval	.050	.013	-.030	.021	.033	.049	.005	.064	-.011	.008
Auto-ATT	.351	.332	.225	.258	.250	<u>.320</u>	.290	.271	.215	.286
TRACE	<u>.356</u>	.333	.212	.327	.273	.313	.245	.293	.119	.193
AudioLLM judges: audio-only and +HEAR										
Qwen2.5-Omni	-.014	-.017	.156	.144	.242	.091	.178	.101	.072	.150
+ HEAR	.119	.143	.181	.164	.234	.119	.147	.099	.125	.220
Qwen3-Instruct	.260	.298	.226	.257	.268	.033	.155	-.104	.114	.269
+ HEAR	.341	.357	.321	.354	.335	.253	.314	.221	.250	.212
Qwen3-Thinking	.197	.253	.348	.327	.307	.302	.240	.270	.158	.211
+ HEAR	**.477**	**.524**	.443	**.493**	**.466**	**.345**	**.457**	.293	.224	<u>.362</u>
Gemini-3-Flash	.147	.233	.268	.335	.260	.007	.189	.112	.096	.178
+ HEAR	.333	.394	**.473**	.447	.310	-.018	.348	**.327**	**.317**	.258
Gemini-3.5-Flash	.317	.373	<u>.455</u>	.422	.360	.168	.359	.273	<u>.313</u>	.297
+ HEAR	.354	<u>.409</u>	.410	<u>.482</u>	<u>.407</u>	.216	<u>.402</u>	<u>.311</u>	.244	**.366**
System-level distributional evaluation ($n = 10$; separate scope)										
TTSDS2	.071	-.212	-.036	-.249	-.418	.176	-.115	-.219	-.321	-.091

† Human reference (blue): mean leave-one-annotator-out correlation with the remaining annotators; an empirical measure of inter-annotator agreement. In the AudioLLM group, each model row is an audio-only control; the following green row adds HEAR to the same backbone. Auto-ATT and TRACE each supply one scalar per utterance; TRACE uses Gemini-2.5-Flash. R&P: rhythm and pacing; Para.: paralinguistic cues; Non.: non-lexical sounds; Iden.: identity consistency; Acous.: acoustic realism; Diff.: difficult content. AE: Audiobox Aesthetics; CE: Content Enjoyment; CU: Content Usefulness; PQ: Production Quality. Appendix C.3 specifies baseline outputs, score direction, and scalar-score reuse. Dimension-wise Pearson results are in Appendix D.1.

Objective metrics show weak or negative agreement with human-likeness ratings. WER and SQUIM obtain Spearman correlations of -0.054 and 0.076, while DNSMOS, UTMOSv2, and NISQA correlate negatively. These quality predictors favor generated speech in mean scores, whereas human-likeness ratings favor real recordings. The opposing group ordering contributes to their negative pooled Pearson correlations. Thus, intelligibility and predicted quality do not reliably recover perceived human-likeness on this dataset.

Auto-ATT and TRACE provide stronger agreement than the generic objective metrics, reaching $0.351/0.332$ and $0.356/0.333$, respectively. Their positive correlations show that specialized speech evaluation and explicit audio descriptions capture useful aspects of human-likeness. HEAR further improves on both, achieving the strongest overall correlations without task-specific training.

HEAR reaches 59.9% and 69.0% of the empirical human-reference Pearson/Spearman correlations ($0.796/0.759$), respectively, placing it closest to the human reference among the automatic utterance-level evaluators.

Across AudioLLM configurations, higher mean dimension-wise Spearman correlations generally accompany higher overall agreement; HEAR improves both for every backbone. This pattern supports multidimensional assessment as a useful complement to holistic judgment, connecting overall alignment with human ratings to the assessment of specific perceptual behaviors.

5.2 Ablation Studies

We examine how audio and acoustic observations, evidence organization, and scoring design contribute to agreement with human ratings. Table 2 reports the ablations on Humanlike-100.

Table 2: Humanlike-100 configuration comparisons (Pearson r and Spearman ρ). O: overall score; D: dimension scores; R: reasons.

Full HEAR		**Audio-only (no evidence)**		**Without audio**	
$r = \mathbf{.477}$ $\rho = \mathbf{.524}$		$r = .197$ $\rho = .253$		$r = .210$ $\rho = .205$	
Removed tool observations	r ρ	**Removed rubric dimension**	r ρ		
Whisper (Radford et al., 2022)	.368 .394	Rhythm and pacing	.353 .365		
MFA (McAuliffe et al., 2017)	.341 .378	Paralinguistic cues	.337 .339		
Wavelet prosody (Suni et al., 2017)	.463 .465	Non-lexical sounds	.367 .379		
Librosa and Praat (McFee et al., 2015; Boersma & Weenink, 2026)	.494 .439	Identity consistency	.351 .339		
YAMNet (Plakal & Ellis, 2020)	.370 .404	Stress and intonation	.332 .344		
Sidon (Nakata et al., 2025)	.470 .402	Acoustic realism	.334 .349		
Respiro-en (Yang et al., 2024)	.418 .463	Difficult-content handling	.305 .319		
ECAPA-TDNN (Desplanques et al., 2020)	.356 .405	Error correction	.378 .373		
Evidence organization	r ρ	**Removed input**	r ρ		
Raw JSON	.288 .355	Tool-observation records	.228 .313		
Natural-Instruct	.333 .386	Word records	.313 .343		
Natural-Thinking	.293 .337	Segment records	.175 .293		
Requested outputs (r/ρ)					
O only .376/.388		O + R .392/.418		O + D .393/.406	

Removing audio or evidence records lowers both correlations; segment-record removal gives the lowest Pearson value (0.175). Removing each tool's observations lowers Spearman; Pearson also falls except without Librosa/Praat. All ten additional observation groups lower both correlations under fixed conversion and judging (Appendix D.2). Structured records outperform raw JSON and both natural-language formats, with differences in both organization and retained information. Every rubric-dimension removal lowers both correlations, most strongly for Difficult-content Handling ($0.305/0.319$). O+D+R performs best among the tested output configurations; both O+R and O+D improve on O alone. Dimension averaging ($0.475/0.521$) nearly matches the direct overall score ($0.477/0.524$), retained as HEAR's primary prediction.

Together, these results support the effectiveness of combining audio with acoustic evidence, structuring the evidence, and using multidimensional scoring to improve agreement with human judgments.

5.3 General Speech Evaluation

As shown in Figure 2, HEAR yields higher preference accuracy on SpeechJudge-Eval, SpeakBench, and S2S-Arena, and higher system-level MOS correlations on TTSDS2, than the corresponding audio-only controls.

SpeechJudge-Eval. HEAR raises single-run preference accuracy from 62.1% for the matched audio-only control to 68.2% (paired gain 6.1 percentage points, 95% CI $[2.7, 9.5]$ points). This gain supports using acoustic evidence for relative naturalness judgments between alternative readings of the same text.

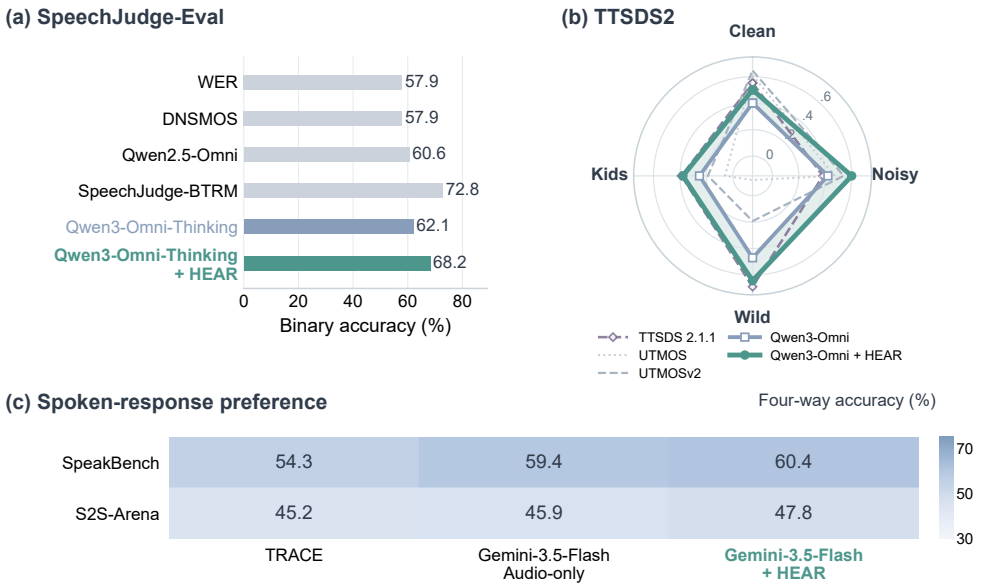


Figure 2: General speech evaluation: (a) binary preference accuracy; (b) system MOS Spearman ρ; (c) four-way accuracy.

TTSDS2. We average utterance-level predictions and human MOS across the four domains within each of the 20 systems, then correlate the resulting system means. Under this pooled evaluation, HEAR raises Pearson/Spearman correlations from $0.660/0.671$ for audio-only to $0.806/0.711$. System-level Spearman improves in all four domains: Clean ($0.401 \rightarrow 0.501$), Noisy ($0.420 \rightarrow 0.600$), Wild ($0.470 \rightarrow 0.643$), and Kids ($0.252 \rightarrow 0.377$). The largest gains occur in Noisy and Wild. The gains within each domain complement the pooled result, showing closer agreement with human system rankings across the evaluated conditions.

SpeakBench and S2S-Arena. With Gemini-3.5-Flash, HEAR reaches 60.4% versus 59.4% for audio-only on SpeakBench (+1.0 percentage point), and 47.8% versus 45.9% on S2S-Arena (+1.9 percentage points). These results show higher overall preference accuracy when HEAR evidence is supplied to the Voice Quality and Paralinguistics judgments; the Content judgment remains based on audio and the instruction transcript.

6 Conclusion

Across all five evaluated AudioLLM backbones, HEAR improves Pearson and Spearman correlations with human overall ratings on Humanlike-100 over matched audio-only judging. The improvements in preference accuracy and system-level MOS correlations support HEAR's generalization to other general speech evaluation tasks beyond human-likeness evaluation.

Ablations highlight the roles of evidence organization and scoring design. Audio and acoustic evidence work better together than either alone, and structured records outperform the tested raw-JSON and natural-language formats. Removing segment records causes a larger decline than removing word records, suggesting the value of local context. Jointly producing overall ratings, dimension-specific scores, and reasons yields the highest correlations among tested output configurations; removing any rubric dimension lowers overall agreement. These results support multidimensional assessment as part of holistic judgment.

Together, these results show that HEAR helps AudioLLMs make more accurate speech judgments by supplementing direct listening with explicit, localized acoustic evidence, bringing their evaluations closer to human assessments.

ACKNOWLEDGMENTS

The authors used AI editing tools to improve grammar and language. All research content and conclusions were developed independently by the authors.

REFERENCES

Alibaba Cloud. Non-real-time speech synthesis, 2026. URL https://www.alibabacloud.com/help/en/model-studio/non-realtime-tts-user-guide. Model Studio documentation, updated 11 September 2026; accessed 25 September 2026.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-TTS: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024. doi: 10.48550/arXiv.2406.02430. URL https://arxiv.org/abs/2406.02430.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The T05 system for the VoiceMOS challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. *arXiv preprint arXiv:2409.09305*, 2024.

Alice Baird, Emilia Parada-Cabaleiro, Simone Hantke, Felix Burkhardt, Nicholas Cummins, and Björn Schuller. The perception and analysis of the likeability and human likeness of synthesized speech. In *Proceedings of Interspeech 2018*, pp. 2863–2867, 2018. doi: 10.21437/Interspeech.2018-1093. URL https://www.isca-archive.org/interspeech_2018/baird18_interspeech.html.

Paul Boersma and David Weenink. Praat: Doing phonetics by computer. Computer program, 2026. URL https://www.praat.org/. Accessed: 2026-06-22.

Boson AI. Higgs Audio: Text-audio foundation model from Boson AI, 2026. URL https://github.com/boson-ai/higgs-audio. Official repository, accessed 25 September 2026.

Brandon James Carone, Iran R. Roman, and Pablo Ripollés. LLMs can read music, but struggle to hear it. an evaluation of core music perception tasks. In *Proceedings of the 1st International Workshop on Emerging AI Technologies for Music*, volume 303 of *Proceedings of Machine Learning Research*, pp. 1–27. PMLR, 2026. URL https://proceedings.mlr.press/v303/carone26a.html.

Arjun Chandra, Kevin Miller, Venkatesh Ravichandran, Constantinos Papayiannis, and Venkatesh Saligrama. Hearing Between the Lines: Unlocking the Reasoning Power of LLMs for Speech Evaluation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 2895–2916, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-eacl.151.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching, 2024. URL https://arxiv.org/abs/2410.06885.

Cheng-Han Chiang, Xiaofei Wang, Chung-Ching Lin, Kevin Lin, Linjie Li, Radu Kopetz, Yao Qian, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, and Lijuan Wang. Audio-Aware Large Language Models as Judges for Speaking Styles. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 467–480, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.25.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models, 2023.

Jian Cong, Shan Yang, Na Hu, Guangzhi Li, Lei Xie, and Dan Su. Controllable context-aware conversational speech synthesis. In *Proceedings of Interspeech 2021*, pp. 4658–4662, 2021. doi: 10.21437/Interspeech.2021-412.

Kamil Deja, Ariadna Sanchez, Julian Roth, and Marius Cotescu. Automatic evaluation of speaker similarity. *arXiv preprint arXiv:2207.00344*, 2022. doi: 10.48550/arXiv.2207.00344.

Wei Deng, Siyi Zhou, Jingchen Shu, Jinchao Wang, and Lu Wang. IndexTTS: An industrial-level controllable and efficient zero-shot text-to-speech system. *arXiv preprint arXiv:2502.05512*, 2025. URL https://arxiv.org/abs/2502.05512.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech 2020*, pp. 3830–3834, 2020. doi: 10.21437/Interspeech.2020-2650.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. CosyVoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024. doi: 10.48550/arXiv.2412.10117. URL https://arxiv.org/abs/2412.10117.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, and Jieping Ye. CosyVoice 3: Towards in-the-wild speech generation via scaling-up and post-training, 2025. URL https://arxiv.org/abs/2505.17589.

ElevenLabs. Text to speech, 2026. URL https://elevenlabs.io/docs/overview/capabilities/text-to-speech. Online documentation, accessed 25 September 2026.

Google. Text-to-speech generation (TTS): Gemini API, 2026. URL https://ai.google.dev/gemini-api/docs/speech-generation. Online documentation, accessed 25 September 2026.

Google DeepMind. Gemini 3 Flash Model Card. https://deepmind.google/models/model-cards/, December 2025. Updated: 2025-12-17; Accessed: 2026-06-22.

Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. The VoiceMOS challenge 2022. In *Proceedings of Interspeech 2022*, pp. 4536–4540, 2022. doi: 10.21437/Interspeech.2022-970.

Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio. In *ICASSP*, 2023. URL https://arxiv.org/abs/2304.01448.

Harm Lameris, Ambika Kirkland, Joakim Gustafson, and Éva Székely. Situating speech synthesis: Investigating contextual factors in the evaluation of conversational TTS. In *Proceedings of the 12th ISCA Speech Synthesis Workshop (SSW 2023)*, pp. 69–74, 2023. doi: 10.21437/SSW.2023-11. URL https://www.isca-archive.org/ssw_2023/lameris23_ssw.html.

Guojian Li, Zhixian Zhao, Zhennan Lin, Jingbin Hu, Qirui Zhan, Yuang Cao, Pengyuan Xie, Chuan Xie, Jie Liu, Qiang Zhang, Zhonghua Fu, and Lei Xie. Towards fine-grained multi-dimensional speech understanding: Data pipeline, benchmark, and model. *arXiv preprint arXiv:2605.12036*, 2026. doi: 10.48550/arXiv.2605.12036.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. Fish-Speech: Leveraging large language models for advanced multilingual text-to-speech synthesis, 2024.

Shijia Liao, Yuxuan Wang, Songting Liu, Yifan Cheng, Ruoyi Zhang, Tianyu Li, Shidong Li, Yisheng Zheng, Xingwei Liu, Qingzheng Wang, Zhizhuo Zhou, Jiahua Liu, Xin Chen, and Dawei Han. Fish Audio S2 technical report, 2026.

Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. MOSNet: Deep learning-based objective assessment for voice conversion. In *Proceedings of Interspeech 2019*, pp. 1541–1545, 2019. doi: 10.21437/Interspeech.2019-2003.

Potsawee Manakul, Woody Haosheng Gan, Michael J Ryan, Ali Sartaz Khan, Warit Sirichotedumrong, Kunat Pipatanakul, William Barr Held, and Diyi Yang. Audiojudge: Understanding what works in large audio model based speech evaluation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3644–3663, 2026.

Ruskin Raj Manku, Yuzhi Tang, Xingjian Shi, Mu Li, and Alex Smola. EmergentTTS-Eval: Evaluating TTS Models on Complex Prosodic, Expressiveness, and Linguistic Challenges Using Model-as-a-Judge, 2025.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017*, pp. 498–502, 2017. doi: 10.21437/Interspeech.2017-1386.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, pp. 18–25, 2015. doi: 10.25080/Majora-7b98e3ed-003.

MiniMax. Text to speech (T2A) WebSocket, 2026. URL https://platform.minimax.io/docs/api-reference/speech-t2a-websocket. Online documentation, accessed 25 September 2026.

Christoph Minixhofer, Ondřej Klejch, and Peter Bell. TTSDS – Text-to-Speech Distribution Score, 2024.

Christoph Minixhofer, Ondrej Klejch, and Peter Bell. TTSDS2: Resources and benchmark for evaluating human-quality text to speech systems. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/31888563b194f9bb33ce1aebc7e1551c-Abstract-Conference.html.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech*, pp. 2127–2131, 2021. doi: 10.21437/Interspeech.2021-299.

Wataru Nakata, Yuki Saito, Yota Ueda, and Hiroshi Saruwatari. Sidon: Fast and robust open-source multilingual speech restoration for large-scale dataset cleansing. *arXiv preprint arXiv:2509.17052*, 2025.

OpenAI. Text to speech, 2026. URL https://developers.openai.com/api/docs/guides/text-to-speech. Online documentation, accessed 25 September 2026.

Olivier Perrotin, Brooke Stephenson, Silvain Gerber, Gérard Bailly, and Simon King. Refining the evaluation of speech synthesis: A summary of the blizzard challenge 2023. *Computer Speech & Language*, 90:101747, 2025. doi: 10.1016/j.csl.2024.101747.

Manoj Plakal and Daniel P. W. Ellis. Yamnet. https://github.com/tensorflow/models/tree/master/research/audioset/yamnet, 2020. Pretrained audio event classifier.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6493–6497, 2021. doi: 10.1109/ICASSP39728.2021.9414878.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 886–890, 2022.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Interspeech*, 2022. URL https://arxiv.org/abs/2204.02152.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth S. Narayanan. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39, 2013. doi: 10.1016/j.csl.2012.02.005. URL https://doi.org/10.1016/j.csl.2012.02.005.

Fritz Seebauer, Michael Kuhlmann, Reinhold Haeb-Umbach, and Petra Wagner. Re-examining the quality dimensions of synthetic speech. In *Proceedings of the 12th ISCA Speech Synthesis Workshop (SSW 2023)*, pp. 34–40, 2023. doi: 10.21437/SSW.2023-6. URL https://www.isca-archive.org/ssw_2023/seebauer23_ssw.html.

Yanfeng Shi, Pengfei Cai, Jun Liu, Qing Gu, Nan Jiang, Lirong Dai, Ian McLoughlin, and Yan Song. Towards fine-grained temporal perception: Post-training large audio-language models with audio-side time prompt. *arXiv preprint arXiv:2604.13715*, 2026. doi: 10.48550/arXiv.2604.13715.

Sajad Shirali-Shahreza and Gerald Penn. Better naturalness evaluation of tts systems. *IEEE Open Journal of Signal Processing*, 2026.

Antti Suni, Juraj Šimko, Daniel Aalto, and Martti Vainio. Hierarchical Representation and Estimation of Prosody Using Continuous Wavelet Transform. *Computer Speech & Language*, 45: 123–136, 2017. doi: 10.1016/j.csl.2016.11.001.

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. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025. URL https://arxiv.org/abs/2502.05139.

Ismail Rasim Ulgen, Zongyang Du, Junchen Lu, Philipp Koehn, and Berrak Sisman. Objective Evaluation of Prosody and Intelligibility in Speech Synthesis via Conditional Prediction of Discrete Tokens. *IEEE Open Journal of Signal Processing*, 2026. doi: 10.1109/OJSP.2026.3653666.

Hui Wang, Jinghua Zhao, Yifan Yang, Shujie Liu, Junyang Chen, Yanzhe Zhang, Shiwan Zhao, Jinyu Li, Jiaming Zhou, Haoqin Sun, Yan Lu, and Yong Qin. SpeechLLM-as-Judges: Towards general and interpretable speech quality evaluation. *arXiv preprint arXiv:2510.14664*, 2025a.

Xihuai Wang, Ziyi Zhao, Siyu Ren, Shao Zhang, Song Li, Xiaoyu Li, Ziwen Wang, Lin Qiu, Guanglu Wan, Xuezhi Cao, Xunliang Cai, and Weinan Zhang. Audio Turing Test: Benchmarking the Human-likeness of Large Language Model-based Text-to-Speech Systems in Chinese, 2025b.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Dong Yang, Tomoki Koriyama, and Yuki Saito. Frame-Wise Breath Detection with Self-Training: An Exploration of Enhancing Breath Naturalness in Text-to-Speech. In *Interspeech 2024*, pp. 4928–4932, 2024. doi: 10.21437/Interspeech.2024-168.

Zhengdong Yang, Shuichiro Shimizu, Yahan Yu, and Chenhui Chu. When Large Language Models Meet Speech: A Survey on Integration Approaches. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20298–20315, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.1041.

Leying Zhang, Yao Qian, Long Zhou, Shujie Liu, Dongmei Wang, Xiaofei Wang, Midia Yousefi, Yanmin Qian, Jinyu Li, Lei He, Sheng Zhao, and Michael Zeng. CoVoMix: Advancing zero-shot speech generation for human-like multi-talker conversations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 100291–100317. Curran Associates, Inc., 2024. doi: 10.52202/079017-3183. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b5fd95d6b16d3172e307103a97f19e1b-Abstract-Conference.html.

Leying Zhang, Yao Qian, Xiaofei Wang, Manthan Thakker, Dongmei Wang, Jianwei Yu, Haibin Wu, Yuxuan Hu, Jinyu Li, Yanmin Qian, and Sheng Zhao. CoVoMix2: Advancing zero-shot dialogue generation with fully non-autoregressive flow matching. In *Advances in Neural Information Processing Systems*, volume 38, pp. 128915–128938. Curran Associates, Inc., 2025. doi: 10.52202/085713-4293. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/bb71b5567ee985e0a4cee54ade19275c-Abstract-Conference.html.

Leying Zhang, Bowen Shi, Haibin Wu, Bach Viet Do, and Yanmin Qian. Jastin: Aligning llms for zero-shot audio and speech evaluation via natural language instructions. *arXiv preprint arXiv:2605.04505*, 2026a.

Xueyao Zhang, Chaoren Wang, Huan Liao, Ziniu Li, Yuancheng Wang, Li Wang, Dongya Jia, Yuanzhe Chen, Xiulin Li, Zhuo Chen, and Zhizheng Wu. SpeechJudge: Towards human-level judgment for speech naturalness. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8acd02d0c55edc3b47a7f9aa42b9cfae-Abstract-Conference.html.

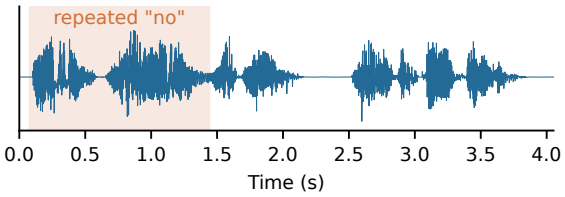
Zhixian Zhao, Shuiyuan Wang, Guojian Li, Hongfei Xue, Chengyou Wang, Shuai Wang, Longshuai Xiao, Zihan Zhang, Hui Bu, Xin Xu, Xinsheng Wang, Hexin Liu, Eng Siong Chng, Hung-yi Lee, and Lei Xie. The ICASSP 2026 HumDial challenge: Benchmarking human-like spoken dialogue systems in the LLM era. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21850–21852. IEEE, 2026. doi: 10.1109/ICASSP55912.2026.11460911. URL https://ieeexplore.ieee.org/document/11460911.

Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. ZipVoice: Fast and high-quality zero-shot text-to-speech with flow matching. *arXiv preprint arXiv:2506.13053*, 2025. URL https://arxiv.org/abs/2506.13053.

A CASE STUDY

Humanlike-100 sample `25_MiniMax` illustrates acoustic measurement, evidence construction, and AudioLLM judgment, linking scores and explanations to word timings, prosody, and segment evidence.

1. Speech and target text. "No, no, no, not now. Why is the screen freezing?" The waveform below marks the three aligned occurrences of "no" (0.08–1.44 s).



2. Constructing evidence from numerical observations. The first "no" spans [0.08, 0.60] s, with prominence 1.000 and boundary score 1.522; the next "no" begins at 0.63 s. Stage 2 applies these deterministic rules:

Measurement	Rule	Resulting evidence
Duration: $0.60 - 0.08 = 0.52$ s	$0.3 \leq d < 0.7$	`duration_tag: long`
Following gap: $0.63 - 0.60 = 0.03$ s	$0 < g < 0.1$	`gap_after_tag: short`
Prominence: 1.000	$0.9 < p \leq 1.2$	`stress_tag: strong`
Boundary: 1.522	$b > 1.2$	`boundary_tag: very_strong`
Adjacent identical tokens; $g = 0.03$ s	Same token and $g < 0.1$	`repeat_with_next: true`
Boundary: 1.522	$b > 0.8$ or $g \geq 0.3$	`segment_break: true`

The strong boundary splits the segment despite the short gap, leaving only the first "no": [0.08, 0.60] s, tempo tag `slow`, and repeat flag `yes`. Global summaries include `speed_tag: medium` and `prominence_tag: scattered`. The words "now" (1.68–2.20 s) and "freezing" (3.27–3.83 s) have prominence values 1.080 and 1.131, respectively, both mapped to `strong`. These fields localize timing and emphasis. Stage 3 receives the original audio, target text, and evidence under the human-likeness prompt.

3. Scores and accompanying reason. The three valid overall scores are 4, 4, and 4. The table compares model means with means of ten human ratings for the overall judgment and all eight dimensions.

	Overall	R&P	Para.	Non.	Iden.	Stress	Acous.	Diff.	Error
Human	4.00	4.00	4.00	4.30	4.50	4.00	4.40	3.90	3.80
HEAR	4.00	4.00	4.00	4.00	4.33	4.33	4.00	5.00	4.33

The first valid output explains its Rhythm & Pacing score as follows: "Natural variation in speed and pauses between repetitions of 'no' mimics human hesitation." This reason refers to the repeated tokens localized above. HEAR matches the human overall score of 4.00. The dimension scores provide a more detailed comparison: Rhythm & Pacing and Paralinguistic Cues match the human means, while Difficult-content Handling receives 5.00 from HEAR and 3.90 from annotators.

4. Linking local evidence to the overall judgment. For Stress & Intonation, the model writes: "Strategic emphasis on 'now' and 'freezing' matches contextual urgency; question intonation rises naturally without template-like rigidity." The named words correspond to the strong-stress tags above; the intonation description comes from the model's audio judgment. Its overall reason is: "Highly convincing natural speech with authentic disfluencies and emotional nuance; minor rhythmic regularity in repetitions prevents absolute indistinguishability from human speakers." The overall score of 4 is a separate holistic judgment: the first output's eight dimension scores average 4.125.

B Evaluation Protocols

B.1 Human Annotation Instructions

Annotators assessed perceived human-likeness from the audio using the shared overall rating anchors and eight dimension criteria in Appendix B.2. They assigned an overall rating before scoring the eight dimensions, using integers from 1 to 5. The overall rating reflects the holistic listening impression rather than an average of dimension scores.

Score	Overall rating anchor shared by human annotators and model judges
1	Clearly synthetic or mechanical.
2	Obviously TTS, with many rigid or odd cues.
3	High-quality but like reading or studio delivery, rather than natural conversation.
4	Mostly human-like, with only mild artifacts.
5	Indistinguishable from natural human speech.

Reading-like or studio-style delivery is capped at 3. The Error Correction criterion does not assign the highest score to perfectly flawless reading; Acoustic Realism assesses coherent, realistic background and space, with unnaturally clean or dead sound treated negatively. A Chinese-language summary of the shared criteria is provided in `reproducibility/humanlike100/human_annotation_instructions_zh.md`.

The real-recording manifest includes source labels such as NonverbalTTS, IEMOCAP, LibriTTS, Emilia, CoVoST2, ESD, and SeniorTalk.

Humanlike-100 rating distribution. For each of the 100 utterances, the reference overall score is the mean of ten annotators' overall ratings. Across these 100 utterance-level means, scores range from 1.0 to 5.0, with a mean of 3.487 and a standard deviation of 1.116. The intervals $[1, 2]$, $(2, 3]$, $(3, 4]$, and $(4, 5]$ contain 13, 15, 37, and 35 utterances, respectively.

Humanlike-100 confidence intervals use 5,000 bootstrap resamples of utterances.

B.2 Human-Likeness Evaluation Prompt

The following Stage 3 prompt is used for the reported human-likeness results. Audio is attached separately; `<target_text>` and `<evidence>` are replaced with the utterance text and Stage-2 evidence. Placeholder delimiters and typography are normalized for display; the exact prompt text is provided in `reproducibility/humanlike100/prompt.txt`. To apply HEAR to other general speech evaluation tasks, we use task-specific prompts and output formats for preference and MOS evaluation.

The response schema requires a numerical rating and a separate reason for each dimension; validation enforces both fields.

System message.

> You are a professional speech perception evaluation expert. Your task is to assess the perceived human-likeness of a given speech sample.

User message. Human-likeness Judging Prompt

1. Role

You are a listening expert. Judge **how human-like the speech sounds**, mainly from the **audio**.

2. Inputs

- Audio: main evidence.

- Target text: the intended script / reference transcript for this utterance (what the speaker was supposed to say). Use it to judge whether disfluencies, omissions, or corrections are natural human behavior vs. synthetic artifacts.

- JSON (tag summary from alignment analysis): auxiliary only (may be noisy / imperfect).

The JSON is a compact tag representation and may include:

- `global_tags`: coarse utterance tags for timing/prosody plus rule-derived acoustic and speaker-stability tags, including `jitter_tag`, `shimmer_tag`, `hnr_tag`, `pitch_variation_tag`, `mfcc_variation_tag`, `acoustic_realism_tag`, and `speaker_stability_tag`

- word_level: per-token timing and tags like token, start/end, duration_tag, stress_tag, boundary_tag, gap_after_tag, repeat_with_next, segment_break

- segment_level: segment summaries like tempo_tag, pause_after_tag, stress_profile, optional environment label hints (e.g., secondary_label_mode), and ECAPA-derived speaker_stability_tag; possible_timbre_drift means an abnormally low similarity between overlapping adjacent speaker windows, not a proven speaker change

Target text:

<target_text>

Evidence (auxiliary tags):

<evidence>

If evidence and audio conflict, **always trust the audio**. Target text helps interpret what was intended, but the audio remains the primary evidence.

3. Scale (1–5, integer)

1 = clearly synthetic / mechanical

2 = obviously TTS, many rigid or odd cues

3 = high-quality but like reading / studio, not natural conversation

4 = mostly human-like, only mild artifacts

5 = indistinguishable from natural human speech

Reading-like, studio-style speech should **not exceed 3**.

4. Dimensions to Rate (1–5 each, no text)

1) Rhythm & Pacing – natural variation of speed, pauses, flow vs. uniform / metronomic.

2) Paralinguistic Cues – fillers, sighs, laughter etc. natural and well blended vs. inserted / exaggerated.

3) Non-lexical Vocal Sounds – subtle breathiness, tension, micro-variations like human physiology vs. random / artificial.

4) Identity Consistency – sounds like **one** stable speaker without identity drift.

5) Stress & Intonation – emphasis and pitch follow meaning vs. flat / templated.

6) Acoustic Realism – background / space coherent and realistic vs. unnaturally clean or dead.

7) Difficult-content Handling – under long/complex content, speech remains stable and human-like vs. breaking down or changing style.

8) Error Correction – human-like monitoring and corrections; perfectly flawless reading **should not** get highest scores.

5. Overall Human-likeness (1–5, not simple average)

Question: "Would an average listener believe this is a real human speaking naturally?"

Score should reflect the **holistic impression**, not a strict mean of dimensions.

6. Output Format (JSON only)

Return **only one JSON object**, no extra text / comments / markdown.

For each dimension and overall, output:

- rating: integer 1–5

- reason: 1–2 sentences (English), based on what you hear in the audio (and optionally the JSON as auxiliary).

```
{
"evaluations": [
    { "dimension": "Rhythm & Pacing",
      "rating": 1-5, "reason": "..." },
    { "dimension": "Paralinguistic Cues",
      "rating": 1-5, "reason": "..." },
    { "dimension": "Non-lexical Vocal Sounds",
      "rating": 1-5, "reason": "..." },
    { "dimension": "Identity Consistency",
      "rating": 1-5, "reason": "..." },
    { "dimension": "Stress & Intonation",
      "rating": 1-5, "reason": "..." },
    { "dimension": "Acoustic Realism",
      "rating": 1-5, "reason": "..." },
```

```
    { "dimension": "Difficult-content Handling",
      "rating": 1-5, "reason": "..." },
    { "dimension": "Error Correction",
      "rating": 1-5, "reason": "..." }
  ],
  "overall_human_likeness": { "rating": 1-5, "reason": "..." }
}
```

B.3 General Speech Evaluation Protocols

This appendix describes the protocols used to assess HEAR's generalization to other general speech evaluation tasks (Section 5.3), including task inputs, scoring criteria, output formats, and aggregation procedures.

Exact prompts and configurations are stored by task in `reproducibility/speechjudge/`, `reproducibility/ttsds2/`, `reproducibility/speakbench/`, and `reproducibility/s2sarena/`. Shared spoken-response prompt rendering, output contracts, fusion rules, and statistics are in `reproducibility/common/response_preference/`. Executed prompt wording is preserved, including references to fields not supplied in every configuration. The TRACE baseline's source revision and local execution settings are recorded in the shared configuration and task guides.

SpeechJudge-Eval. The judge receives target text and audio candidates A/B, and rates each from 1 to 10 based on prosody, pacing, articulation, and overall perceived naturalness. HEAR appends both evidence records to the same base prompt. The JSON contains `output_a_score`, `output_b_score`, and one short reason. The higher score determines preference; ties count as incorrect in binary evaluation. The run uses Transformers with thinking disabled, temperature 0.6, top-$p = 0.95$, top-$k = 20$, and a 512-token output budget. Three valid outputs support single-run, mean-score, and majority-vote analyses.

TTSDS2. The judge receives an utterance and its intended text; HEAR adds auxiliary evidence. It rates overall perceived naturalness on a five-point bad-to-excellent scale. The JSON contains `overall_naturalness` with an integer rating and a short reason. Three valid ratings are averaged before system aggregation. Local vLLM decoding settings match those of the main Humanlike-100 run described in Appendix C.2.

TRACE. Given the spoken instruction, responses A/B, and instruction transcript, the judge returns separate Content, Voice Quality, and Paralinguistics decisions in {`1`, `2`, `both_good`, `both_bad`}. Only the latter two receive HEAR evidence. The audio-only and HEAR conditions use Gemini-3.5-Flash. For HEAR, the official dataset-specific TRACE fusion rule combines audio-based Content judgments with evidence-supported Voice Quality and Paralinguistics judgments into an overall four-way preference. Dataset-specific input builders are in `reproducibility/speakbench/prompt.py` and `reproducibility/s2sarena/prompt.py`; shared fusion is in `reproducibility/common/response_preference/fusion.py`. These Gemini-3.5-Flash runs use temperature zero, top-$p = 1$, and low reasoning effort. The output budget is 4,096 tokens for SpeakBench and S2S-Arena audio-only, and 2,048 tokens for S2S-Arena HEAR.

C Reproducibility Details

C.1 Evidence Rules and Implementation

Fixed thresholds map numerical measurements to categorical evidence for the main human-likeness judge. The public repository stores the Stage-2 implementation and field exclusions in `reproducibility/common/evidence/`, per-tool environment specifications in `code/environments/`, and task-specific prompts and inference settings in the corresponding subdirectories of `reproducibility/`. Applied to saved raw observations, these rules reproduce the semantic content of all 100 main-configuration inputs.

Table 3: Numerical-to-tag rules, applied in the listed order with the stated boundary inequalities. Units: seconds for duration, aligned tokens/s for rate, and dB for HNR.

Measurement	Thresholds and output categories
Token duration / gap	$x \leq 0$: none; $0 < x < .1$: short; $.1 \leq x < .3$: medium; $.3 \leq x < .7$: long; $.7 \leq x < 1.5$: very long; $x \geq 1.5$: ultra long.
Utterance / segment duration	$x \leq 0$: none; $(0, 2)$: short; $[2, 5)$: medium; $[5, 10)$: long; $[10, 20)$: very long; $[20, \infty)$: ultra long.
Token count	$n \leq 0$: none; 1–3: few; 4–10: some; 11–30: many; $n > 30$: very many.
Speaking rate	$x < 1$: very slow; $[1, 2)$: slow; $[2, 3.5)$: medium; $[3.5, 5)$: fast; $x \geq 5$: very fast.
Prominence strength	$x \leq .3$: very weak; $(.3, .6]$: weak; $(.6, .9]$: medium; $(.9, 1.2]$: strong; $x > 1.2$: very strong.
Boundary strength	$x \leq .1$: very weak; $(.1, .4]$: weak; $(.4, .8]$: medium; $(.8, 1.2]$: strong; $x > 1.2$: very strong.
Prominent-token ratio	$x < .05$: none; $[.05, .2)$: local; $[.2, .5)$: scattered; $x \geq .5$: dominant.
Detected breath-interval ratio	$x < .1$: stable speech dominant; $[.1, .25)$: speech dominant; $[.25, .4)$: speech weakened; $x \geq .4$: transitional weakened.
HNR	$x < 5$: low/noisy; $[5, 10)$: moderate; $[10, 20)$: normal; $x \geq 20$: high/clear.
Estimated RT60	$x < .05$: abnormally short; $[.05, .1)$: very short; $[.1, 2)$: normal; $[2, 5)$: long; $x \geq 5$: abnormally long.
Mean MFCC standard deviation	$x < 1$: abnormally stable/possible synthetic; $[1, 5)$: stable; $[5, 15)$: normal; $x \geq 15$: highly variable. These categories provide heuristic descriptions of MFCC variability.
Adjacent ECAPA cosine	$x < .70$: possible timbre drift; $[.70, .80)$: unstable; $[.80, .85)$: borderline; $x \geq .85$: stable.

Aggregation and localization. Word timestamps and derived gaps are rounded to three decimals. A segment ends when the following gap is at least 0.3 s or the boundary score exceeds 0.8. Consecutive identical tokens, ignoring case, receive a repeat flag for gaps below 0.1 s. Global strength averages available prominence measurements; the global boundary tag discretizes the maximum boundary value. Global prominence counts tokens with stress or boundary scores of at least 1.0. Segment prominence divides the summed counts of strong-stress and strong-boundary tokens by segment token count; a token may contribute to both. The breath-interval ratio is summed detected breath duration divided by utterance duration.

Each segment's secondary event label maximizes summed YAMNet scores over top-five frame labels timestamped within the segment, excluding "Speech". ECAPA uses 6-s windows and a 3-s hop, localizing transitions to consecutive windows' 3-s overlap. Each segment receives the most severe overlapping transition tag; global speaker stability records the most severe transition. Unavailable acoustic and speaker measurements use `not_assessable`. Original timing and tool-observation records accompany the three evidence levels.

C.2 Tools, Versions, and Inference Settings

The Humanlike-100 tool configuration specifies audio preprocessing, pretrained extractors, and numerical settings such as analysis windows and hop sizes. The main HEAR runs share the saved extraction settings, evidence rules, and scoring rubric; runtime and decoding settings are recorded separately for each AudioLLM.

The pipeline resamples tool inputs to mono 16-kHz audio. Whisper large-v3 supplies the alignment transcript. MFA 3.4.1 uses English or Mandarin resources; wavelet analysis provides prominence and boundaries. The environment uses Librosa 0.11.0, Praat 6.4.52, pycwt 0.5.0b0, and TensorFlow 2.21.0 for YAMNet v1 (0.96-s windows, 0.48-s hops). Respiro-en uses `respiro-en.pt`; ECAPA-TDNN uses `pretrain.model`. Sidon v0.1 enhancement is restricted to the ECAPA branch. Per-tool envi-

ronment files are under `code/environments/`; task-specific model and decoding settings are stored in `reproducibility/humanlike100/configurations.json` and the external-task configuration files.

Qwen3-Omni-30B-A3B-Thinking runs with Python 3.10.20, PyTorch 2.13.0, Transformers 5.15.0, vLLM 0.27.1, and qwen-omni-utils 0.0.9. Its human-likeness run uses four-way tensor parallelism, bfloat16, batch size 8, temperature 0.6, top-$p = 0.95$, top-$k = 20$, repetition penalty 1.05, and a 32,768-token context limit. The initial 8,192-token output budget rises to 16,384 after a length failure. Outputs must satisfy the exact JSON schema, contain all eight dimensions and an overall score, use integer ratings in $[1, 5]$, and provide nonempty reasons. The first three valid outputs are averaged; failures are retained and schema validity determines retries.

C.3 Baseline Outputs and Implementation Records

Table 4 specifies each baseline's score. On Humanlike-100, utterance-level scalar baselines and TRACE reuse the same predicted score for correlations with the overall human rating and each human-rated dimension.

Table 4: Baseline outputs and implementations verified from saved records and runner configurations. Citations identify model families; checkpoint details are listed separately.

Baseline	Output and implementation
WER	Negative ASR error rate using Chinese characters and lower-cased Latin alphanumeric word spans; transcription from saved acoustic observations.
SQUIM (Kumar et al., 2023)	VERSA `torch_squim_mos` (subjective MOS), with a fixed non-matching real reference selected without scores from the same CJK/Latin text bucket.
DNSMOS (Reddy et al., 2022)	VERSA `dns_overall` evaluated at a 16-kHz input rate.
UTMOS (Saeki et al., 2022)	`utmos22_strong`, checkpoint `utmos22_strong_step7459_v1.pt`, loaded through SpeechMOS v1.2.0.
UTMOSv2 (Baba et al., 2024)	`fusion_stage3`, five folds, five crop predictions per fold, averaged; silence removal disabled. The saved plan fixes revision and checkpoint hashes.
NISQA (Mittag et al., 2021)	VERSA `nisqa_mos_pred`, the overall MOS output from `nisqa.tar`.
Audiobox Aesthetics (Tjandra et al., 2025)	VERSA `audiobox_aesthetics_CE/CU/PQ`: Content Enjoyment, Content Usefulness, and Production Quality, reported separately.
Auto-ATT (Wang et al., 2025b)	Official `AGI-Eval/Auto-ATT` revision `5dd1bb0`; Qwen2-Audio-7B-based evaluator with mono 16-kHz input and deterministic one-token generation. The official Good/Fair/Bad prompt produces the scalar HLS $3P(\text{Good}) + 2P(\text{Fair}) + P(\text{Bad})$. All 100 utterances yielded valid scores in one pass.
TTSDS2 (Minixhofer et al., 2026)	PyPI `ttsds==2.1.1`; public data revision `873d63a`; Environment disabled; 11 default metrics aggregated with equal Generic/Prosody/Speaker/Intelligibility weights; 20 systems per domain, 880 metric–system–domain outputs, and rating-count-weighted human MOS.
TRACE (Chandra et al., 2026)	TRACE commit `9cb2a59`; Gemini-2.5-Flash with textual blueprints and intended text. Humanlike-100 uses a single-utterance human-likeness prompt, temperature 0, top-$p = 1$, an 8,192-token budget, and provider-default thinking. We average the overall rating from three scheduled repeats, retaining the first schema-valid response for each repeat. This scalar is reused against each human dimension.

For Humanlike-100, the TTSDS2 metric compares the five generated utterances from each of the ten TTS systems with the 50 real reference recordings. We correlate the resulting ten system-level metric scores with the mean human ratings for each system's generated utterances.

Auto-ATT is a trained specialist baseline. Its configuration and prompt are retained in `reproducibility/humanlike100/baselines/auto_att/`, with saved per-item outputs under `evaluation_results/humanlike100/predictions/`. TRACE receives text-only input and uses its own backbone and task-specific prompt; its comparison with HEAR evaluates the complete judge configurations. The Gemini-2.5-Flash Humanlike-100 run covers all 100 utterances with 300 valid outputs. Its configuration and prompt are retained in `reproducibility/humanlike100/baselines/trace/`, with saved predictions in `evaluation_results/humanlike100/predictions/trace.csv`.

C.4 HEAR-H Data Resource

HEAR-H provides reusable generation tasks for analyzing the speech behaviors assessed by HEAR. Humanlike-100 measures evaluator agreement with human ratings; HEAR-H supplies prompt–target pairs for synthesizing speech under specified conditions. Its three subsets target expressive delivery, integration of non-verbal events, and stable delivery of demanding content, enabling focused comparisons of speech generators.

Task coverage. The resource contains 2,702 prompt–target pairs in Chinese, English, and mixed Chinese–English. The disjoint *Expressive* subset (1,102 pairs; `basic-v1`) covers emotional dialogue and narrative speech, emphasizing emotion, prosodic variation, and continuity. The *Paralinguistic* subset (1,106 pairs) specifies non-verbal events in the target text, supporting analysis of their realization, placement, and integration with speech. The *Hard* subset (494 pairs) includes classical Chinese, terminology-rich English, and Chinese–English code-switching, examining delivery under linguistic demands. Separating basic expressive ability, optional non-verbal generation, and robustness to difficult content supports comparisons within each capability, accommodating differences in what TTS systems are designed to generate.

Generation inputs. We reorganize public speech recordings into prompt–target pairs: an initial segment and its transcript provide conditioning context, and the continuation text specifies the content to synthesize. Each pair contains prompt audio, its transcript, and target text, including event tags in the Paralinguistic subset. The same inputs can be supplied to different TTS systems for comparisons within each scenario. The data are available at `https://huggingface.co/datasets/dinosaaaur/HEAR-HSet`.

D Additional Evaluation Results

D.1 Complete Dimension-Wise Pearson Results

Table 5 complements the dimension-wise results in Table 1 with Pearson correlations, using the same samples, evaluation levels, and dimension order. For scalar metrics and TRACE, the same scalar score is correlated with each human-rated dimension. AE denotes Audiobox Aesthetics: CE is Content Enjoyment, CU is Content Usefulness, and PQ is Production Quality.

Table 5: Pearson correlations for Overall and all eight dimensions.

Evaluator	Overall	R&P	Para.	Non.	Iden.	Stress	Acous.	Diff.	Error
Human reference†	.796	.726	.724	.722	.649	.735	.702	.657	.688
Objective metrics									
WER	.159	.101	.107	.118	.128	.142	.101	.122	.139
SQUIM	.066	.109	.073	.012	.052	-.031	.013	.063	.056
DNSMOS	-.247	-.187	-.191	-.216	-.186	-.179	-.127	-.122	-.192
UTMOS	-.279	-.283	-.216	-.222	-.154	-.180	-.085	-.134	-.207
UTMOSv2	-.290	-.248	-.183	-.210	-.197	-.125	-.078	-.152	-.221
NISQA	-.214	-.273	-.159	-.204	-.217	-.161	-.099	-.134	-.200
AE-CE	-.239	-.165	-.171	-.221	-.181	-.083	-.129	-.202	-.224
AE-CU	-.281	-.211	-.216	-.253	-.202	-.136	-.175	-.251	-.298
AE-PQ	-.331	-.218	-.265	-.283	-.251	-.170	-.233	-.260	-.317
Speech evaluators									
SpeechEval	-.235	-.178	-.136	-.239	-.074	-.155	-.084	-.185	-.206
EmergentTTS-Eval	.050	.015	.079	.094	.086	.085	.103	-.050	.051
Auto-ATT	.351	.228	.284	.265	.291	.253	.253	.238	.269
TRACE (Gemini-2.5-Flash)	.356	.216	.332	.267	**.328**	.284	.242	.158	.176
AudioLLM backbones with and without HEAR									
Qwen2.5-Omni (audio-only)	-.014	.195	.133	.157	-.041	.173	.022	.079	.151
+ HEAR	.119	.221	.196	.250	.104	.201	.094	.196	.231
Qwen3-Instruct (audio-only)	.260	.240	.218	.207	-.023	.187	-.065	.091	.245
+ HEAR	.341	.345	.331	.241	.131	.295	.114	.242	.157
Qwen3-Thinking (audio-only)	.197	.337	.308	.275	.208	.220	.224	.160	.166
+ HEAR	**.477**	.440	**.470**	**.417**	.238	**.448**	.250	.239	**.318**
Gemini-3-Flash (audio-only)	.147	.249	.280	.202	-.058	.119	.076	.056	.142
+ HEAR	.333	.448	.413	.256	-.049	.306	.255	**.290**	.216
Gemini-3.5-Flash (audio-only)	.317	**.449**	.367	.319	.029	.355	.181	.283	.253
+ HEAR	.354	.398	.435	.373	.090	.359	**.258**	.198	.300
System-level distributional evaluation ($n = 10$; separate scope)									
TTSDS2‡	.071	-.023	-.140	-.175	.319	-.005	.056	-.017	.036

† Human reference: mean leave-one-annotator-out correlation with the remaining annotators; an empirical measure of inter-annotator agreement. ‡ TTSDS2: system-level correlation ($n = 10$, five utterances per system); all other evaluator rows use utterance-level correlation ($n = 100$). Scalar metrics, Auto-ATT, and TRACE reuse one score across human dimension targets, as in Table 1.

D.2 Additional Acoustic Observations

Experimental setup. We separately add ten groups of acoustic observations to the HEAR configuration fixed before Humanlike-100 evaluation. Evidence conversion, the judge, the scoring prompt, and score aggregation remain unchanged. Each configuration is evaluated on the same 100 utterances by averaging the first three valid outputs.

Table 6: Additional-observation experiments on Humanlike-100 ($n = 100$). Each row adds the named observations to the fixed HEAR pipeline.

Additional observations	r	ρ
None (HEAR)	**.477**	**.524**
Acoustic artifacts	.331	.384
pYIN	.361	.390
CREPE	.303	.371
eGeMAPS	.372	.438
WavLM	.347	.407
HuBERT	.335	.372
SSL combination	.346	.433
Pitch ensemble	.339	.402
Artifacts + eGeMAPS	.343	.398
All candidate tools	.306	.359

Results and interpretation. All ten additions yield lower Pearson and Spearman correlations than the fixed HEAR configuration (Table 6).

E LIMITATIONS

Speech evaluation must accommodate diverse languages, speaking styles, and listening contexts. Human judgments can also vary across listeners, making consistent evaluation across these settings challenging. Although HEAR improves agreement with human ratings in the evaluated conditions, its broader applicability remains to be established. Future work should extend evaluation to more diverse speech and listener populations and examine the robustness and interpretability of AudioLLM-based judgments.