一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

J-HARD-TTS-EVAL-v2: 高難度日本語音声合成コーパスの構築と大規模システム評価

ホワンウェンチン† 今井 柊平†† 榎本 悠久†† 栗原 清††† 齋藤 佑樹††††

† 名古屋大学
†† Parakeet 株式会社
††† 株式会社松尾研究所/東京大学大学院
†††† 東京大学/産業技術総合研究所
E-mail: †wen.chinhuang@g.sp.m.is.nagoya-u.ac.jp

あらまし 近年のテキスト音声合成（text-to-speech；TTS）は高い自然性を実現しているが，入力テキストに応じた自然な韻律の生成能力を検証するには，評価目的に即したテキストの設計が重要である。先行研究の J-HARD-TTS-EVAL-v1 は，語句の繰り返しや脱落などの系列整合性を評価するために構築された。本研究では，評価対象をテキストに応じた韻律の自然性にも広げ，8 種類のサブセット，各 20 文の計 160 文からなる J-HARD-TTS-EVAL-v2 を構築し，オープンウェイト 20 システムとクローズウェイト 2 システムの計 22 システムを評価した。日本語母語話者 50 名による聴取実験では，発話自体の不自然さと，発話としては自然でも文脈に合わない棒読みを区別するため，テキストに応じた韻律の自然性を重視した独自の 7 段階尺度を用い，評点分布と高評価の割合に基づいて結果を分析した。Gemini 3.1 Flash TTS が全体で最も高い評価を得た一方，多くのシステムでは感情表現や繰り返しを含む発話に課題が見られた。また，日英コードスイッチング音声の評価では全体順位と異なる傾向が見られ，サブセット別評価の重要性が示された。デモページ：https://j-hard-tts-eval-v2.wenchinhuang.com/?lang=ja

キーワード テキスト音声合成、聴取実験

1. はじめに

テキスト音声合成（text-to-speech；TTS）システムの評価では，評価対象とする能力，評価に用いるデータ，合成条件，および評価方法を選定する必要がある。TTS 技術の発展に伴い，これらの選択も変化してきた。2024 年に提案された NaturalSpeech では、単一話者の評価条件において，合成音声と肉声の主観評価結果に統計的に有意な差が認められなかったことが報告されていた [1]。一方，学習時に含まれない話者の参照音声を用いて合成を行うゼロショット（zero-shot）TTS にも関心が集まっており，参照音声に含まれる話者性や発話スタイルなどを再現する能力が評価されている。評価方法についても，聴取実験による主観評価に加え，音声認識（automatic speech recognition; ASR）結果に基づく単語誤り率，話者埋め込みのコサイン類似度，および UTMOS [2] などの主観評価予測モデルが，低コストで再現可能な客観評価指標として広く用いられている。また，Voice Arena(注1)や Artificial Analysis(注2)など，大規模な選好評価を行う第三者の評価プラットフォームも登場している。こうした合成条件や評価方法の多様化に加え，評価したい能力を引き出すテキストの設計も重要である。

TTS には，音質，明瞭性，話者性だけでなく，構文，句読点，文の意味などに応じた自然な韻律を生成することが求められる。この能力の向上に向けて，事前学習済みテキスト埋め込みの導入 [5]〜[7] など言語表現を強化する研究が進められてきた。一方，広く用いられている LJSpeech [8]，VCTK [9]，LibriTTS [10] は，主に書籍や新聞などの読み上げ音声に基づくコーパスであり，特定の韻律現象を体系的に検証するために設計されたものではない。また，豊かな表現を含む参照音声の特徴を再現する能力と，入力テキストから手掛かりを捉えて適切な話し方を生成する能力は区別する必要がある。そこで，評価したい言語現象を意図的に含むテキストを用いることで，後者の能力をより直接的に検証できると考えられる。BASE TTS [11] や EmergentTTS-Eval [12] では，複雑な構文や多様な表現を含む高難度テキストを用いた評価が提案されている。

日本語についても，JSUT [13]，JVS [14]，Hi-Fi-CAPTAIN [15] などの公開音声コーパスが，単一話者，多話者，対話調の音声合成研究を支えてきた。日本語では，複数の文字体系や文脈に依存する漢字の読みに加え，表記に明示されないアクセント型や，語の結合に伴うアクセントの変化などが自然な音声生成の課題となる。近年では，こうした言語固有の課題を検証する評価資源も整備されている。Joyo Kanji Yomi Benchmark [16] とその拡張を行った Parakeet 版(注3)は，漢字の読みの正確さを対象

(注1)：https://voicearena.com/
(注2)：https://artificialanalysis.ai/text-to-speech/arena
(注3)：https://huggingface.co/datasets/Parakeet-Inc/joyo-kanji-yomi-benchmark-parakeet

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2026 by IEICE

表 1　各サブセットの例文。Normal は MiniMax-Speech [3] で提案された TTS-Multilingual-Test-Set の日本語 100 文から 20 文を選出した。Repetition と Rhyme は J-HARD-TTS-EVAL-v1 [4] をそのまま流用した。

サブセット名	例文
Normal	週末は実家に帰省する予定なんです。
普通文	最近始めたヨガのレッスンがとても気に入っています。
Rhyme	母の派手な発表の反響は果てしなく、その判断が早い発展の始まりとなった。
押韻	見識ある賢者の見解を検討した結果、権利の主張は謙虚なのが賢明だ。
Syntactic Complexity	警備員が先週確認したリストに載っていなかった不審な男が突然走り出したため会場は騒然となりました。
係り受けが複雑な長文	弟が拾ってきた猫が追いかけ回していたネズミが逃げ込んだ穴は予想以上に深くて暗いものでした。
Compound Nouns	独立行政法人高齢者障害者雇用支援機構職業能力開発総合大学校の入学試験に、多くの若者が挑戦した。
複合語	太平洋沿岸地域津波避難計画策定ガイドラインの改定に伴い、住民を対象とした説明会が公民館で開かれた。
Punctuations	いやだーー！ビュッフェの予約を忘れたなんて！
句読点	え……あ……そ、そうですね……少し考えを整理します。
Repetition	1 時、2 時、3 時、4 時、5 時、6 時、ご飯を食べて、8 時、9 時、と、時計の針は進んでいく。
繰り返し	あの、あのあの、あのあの、あのあの、あのあの一、すみません、今なんて？
Code-switching	研究室の MacBook に USB をつないで、Python の log を確認してから、GitHub に結果を送ります。
言語切り替え	会議室の予約は済ませました。The projector and microphones are already prepared.
Emotions	約束を破るなんて最低だ。もう君の顔なんて、二度と見たくもないよ。
感情	やったあ！発表会で優勝できたよ！嬉しくて、ぴょんぴょん飛び跳ねちゃいそうだ！

としている。一方，J-HARD-TTS-EVAL-v1 は，短い発話，語句の繰り返し，押韻，および途中で打ち切られた文を用いて，自己回帰型 TTS における繰り返し，脱落，追加などの系列整合性を評価する [4]。これらの読みや系列整合性に着目した評価に加え，テキストに応じた自然な韻律を生成できるかを，多様な言語現象にわたって検証するための評価資源が求められる。

本研究では，J-HARD-TTS-EVAL-v1 を基に，評価対象をテキストに応じた韻律の自然性にも広げた J-HARD-TTS-EVAL-v2 を構築する。本コーパスは 8 種類のサブセット，各 20 文の計 160 文からなる。さらに，オープンウェイト 20 システムとクローズウェイト 2 システムの計 22 システムを対象に，日本語母語話者 50 名による聴取実験を実施する。発話自体の不自然さだけでなく，自然でも文脈に合わない棒読みや，テキストに適した自然な表現を捉えるため，それぞれの状態を評点の基準として明示した独自の 7 段階評定尺度を設計する。得られた評点の分布と高評価の割合に基づく分析により，システム間およびサブセット間の違いを示すとともに，全体順位だけでは捉えられないシステムごとの得意・不得意を明らかにする。

2. コーパス構築

J-HARD-TTS-EVAL-v1 では，EmergentTTS-Eval に含まれるような多様な表現を評価する高難度テキストが抜けているため，それを参考に押韻・繰り返しに加えて新しく 6 つのサブセットを構築した。J-HARD-TTS-EVAL-v2 の構築にあたり，通常の読み上げ文に加え，構文，語の構成，句読点，言語の切り替え，感情表現などに関する特徴を持つ文を作成した。各サブセットの例文を表 1 に示す。

J-HARD-TTS-EVAL-v2 では，テキストに応じた韻律の自然性を多面的に検証するため，v1 のサブセット構成を見直した。v1 から押韻と繰り返しの 2 サブセットを継承し，通常文，係り受けが複雑な長文，複合語，句読点，言語切り替え，感情の 6 サブセットを新たに導入した。追加する評価観点の設計には，EmergentTTS-Eval [12] などの高難度テキストを用いた評価を参考にした。各サブセットの例文を表 1 に示す。

2.1 サブセットの構成

Normal（普通文）　日常的な内容を扱う通常の読み上げ文からなる。他のサブセットとの比較に用いる基準として設けた。

Rhyme（押韻）　押韻した文からなる。類似した音の並びを含む文において，自然なリズムや抑揚を維持できるかを評価する。

Syntactic Complexity（係り受けが複雑な長文）　複雑な係り受け構造を持つ長文からなる。文の構造や意味のまとまりに応じて適切な位置にポーズを挿入し，自然なまとまりとして発話できるかを評価する。

Compound Nouns（複合語）　複数の語が結合した複合語を含む文からなる。語の結合に伴うアクセントの変化を適切に反映し，一本調子にならず自然なピッチで発話できるかを評価する。漢字を主体とする複合語，カタカナを主体とする複合語，長い複合語，および中程度から短い複合語を複数含む文を作成し，表記と長さに多様性を持たせた。

Punctuations（句読点）　感嘆符，疑問符，三点リーダ，かぎ括弧などを含む文からなる。記号を文脈に応じて解釈し，適切なポーズ，イントネーション，感情表現の強弱，および発話の切れ目を生成できるかを評価する。例えば，疑問符を含む文では通常の疑問に加えて皮肉や困惑を，感嘆符を含む文では命令や強い感情を扱う。また，三点リーダによるためらいや思考の中断，かぎ括弧による引用や強調も含めた。

Repetition（繰り返し）　同じ語句の繰り返しを多用する文からなる。反復部分を含む発話において，文脈に応じた自然なリズムやポーズを生成できるかを評価する。

Code-switching（言語切り替え）　日本語と英語が混在する文からなる。言語の切り替えに伴う声質やピッチのつながりが自然であるかを評価する。文中での切り替え，文間での切り替え，

および短い表現を付加する形式の切り替えを含めた。文間の切り替えについては，日本語から英語，英語から日本語の両方向を扱う。

Emotions（感情） 文の意味から感情が読み取れる文からなる。明示的な感情ラベルを与えることなく，テキストの内容に応じた声色や抑揚を生成できるかを評価する。対象とする感情は，怒り，喜び，恐怖，悲しみ，嫌悪，驚き，絶望，安堵，感嘆，混乱，羞恥，軽蔑，決意，愛情，退屈，嫉妬，焦燥，郷愁，感謝，期待の 20 種類とした。各感情について 1 文ずつ作成し，計 20 文を収録した。

これらのサブセットは，評価したい特徴に基づいて分類したものであり，各文が持つ言語的・韻律的な特徴は相互に排他的ではない。例えば，感情を表す文には句読点による韻律の手掛かりも含まれ得る。

2.2 構築手順

Normal は，MiniMax-Speech [3] で提案された TTS-Multilingual-Test-Set (注4) に含まれる日本語 100 文から 20 文を選出した。Rhyme および Repetition は，J-HARD-TTS-EVAL-v1 [4] の該当サブセットを変更せずに用いた。残る 5 種類のサブセットは，大規模言語モデル（large language model；LLM）を用いて作成した。具体的には，各サブセットの評価目的と求める条件を記述したプロンプトを作成し，Gemini 3.1 Pro に入力して候補文を生成した。生成された候補文が，意図した特徴を含んでいるか，また文として自然で意味が通じるかを確認した。品質が不十分な場合は，プロンプトの条件や指示を修正して再度生成した。この手順を繰り返し，各サブセットの目的に沿った文を生成できるようになった段階で構築を終了した。

3. 大規模システム評価

3.1 評価対象システム

構築したコーパスを用いて，日本語の音声合成に対応する 22 システムを評価した。システム一覧を表 2 に示す。評価対象は，モデル重みが公開されているオープンウェイト 20 システムと，商用サービスとして提供されているクローズウェイト 2 システムからなる。同一系列でもモデル規模やバージョンが異なる場合は，それぞれ独立した評価対象として扱った。

3.2 音声合成と候補選定

3.2.1 合成条件

参照音声による条件付けシステムでは，参照音声を入力とするゼロショット合成を行った。一部参照音声に対応しないシステムでは，システムが提供する固定話者を用いた。参照音声は，Common Voice Corpus [29] 21.0 の日本語セットから選定した。無音区間を除去した後の長さが 5 秒以上かつ UTMOS スコアが 2.5 以上の音声から UTMOS スコアの高い順に第一著者が聴き，発話の明瞭さ，雑音の有無，書き起こしの正確さ，および自然な発話であるかを確認した。最終的に，女性話者による音声 10 サンプルと男性話者による音声 10 サンプルの計 20 サンプルを選定した。全サブセットで同じ 20 サンプルを使用し，各サブセット内の 20 文との対応関係をサブセットごとに個別にシャッフルした。

表 2 評価対象システム。下線が引かれたシステムはクローズウェイトシステムである。

システム	公開年月	モデルサイズ
VALL-E-X*[a] [17]	2023.08	300M
ElevenLabs Multilingual v2†[b]	2023.08	—
XTTS-v2 [18]	2023.11	467M
CosyVoice2 [19]	2024.12	500M
Zonos-v0.1[c]	2025.02	1.6B
Chatterbox Multilingual[d]	2025.09	500M
T5Gemma-TTS [20]	2025.12	5B
MagpieTTS 357M†[g]	2025.12	357M
Qwen3-TTS 0.6B [21]	2026.01	0.6B
Qwen3-TTS 1.7B [21]	2026.01	1.7B
Fish Audio S2 Pro [22]	2026.03	4B
X-Voice [23]	2026.04	400M
Sarashina2.2-TTS [16]	2026.04	800M
Irodori-TTS[e]	2026.04	500M
OmniVoice [24]	2026.04	800M
VoxCPM2 [25]	2026.04	2B
Gemini 3.1 Flash TTS†[f]	2026.04	—
MOSS-TTS 1.5 [26]	2026.05	8B
LFM2.5-Audio-1.5B-JP†[h]	2026.06	1.5B
Chatterbox Multilingual V3[d]	2026.06	500M
Zonos2 [27]	2026.06	8B
dots.tts [28]	2026.06	2B

[a] 非公式の公開実装 https://github.com/Plachtaa/VALL-E-X を用いた。
[b] https://elevenlabs.io/blog/elevenlabs-comes-out-of-beta-and-releases-eleven-multilingual-v2-a-foundational-ai-speech-model-for-nearly-30-languages
[c] https://github.com/Zyphra/Zonos
[d] https://huggingface.co/ResembleAI/chatterbox
[e] https://github.com/Aratako/Irodori-TTS
[f] https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-tts-preview
[g] https://huggingface.co/nvidia/magpie_tts_multilingual_357m
[h] https://huggingface.co/LiquidAI/LFM2.5-Audio-1.5B-JP
†：固定話者，*：非公式実装。M/B は百万/十億パラメータ，—は非公開を表す。

3.2.2 候補生成と選定

評価対象の多くは LLM に基づく TTS システムであり，推論時のサンプリングによって，同一の入力に対しても生成結果が異なる。このばらつきを考慮し，ハルシネーションの少ない音声を聴取実験に用いるため，複数の候補音声を生成し，ASR 結果に基づいて評価音声を選定した。J-HARD-TTS-EVAL-v1 [4] を参考に，各システムで 1 テキストにつき 5 個の候補音声を生成した。ただし，Gemini 3.1 Flash TTS は 1 テキストにつき 1 個のみ生成した。各候補音声を IBM Granite Speech 4.1 2B(注5)，Qwen3-ASR 1.7B [30]，ReazonSpeech ESPnet および ReazonSpeech NeMo(注6) の 4 種類の ASR システムで書き起こした。各書き起こしと入力テキストとの間で文字誤り率（character error rate; CER）を計算し，4 種類の ASR 結果のうち最小の CER を当該候補音声の CER として採用した。5 個の候補音声のうち，CER が最も低い

(注4)：https://huggingface.co/datasets/MiniMaxAI/TTS-Multilingual-Test-Set

(注5)：https://huggingface.co/ibm-granite/granite-speech-4.1-2b
(注6)：https://github.com/reazon-research/reazonspeech

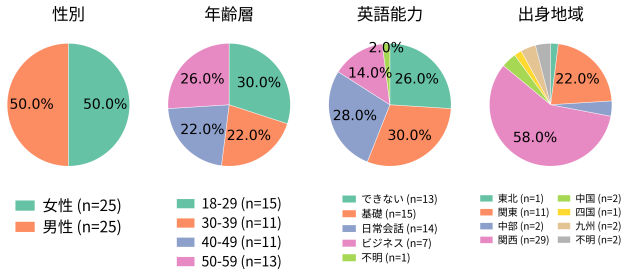


図 1　聴取実験参加者の属性（50 名）。

表 3　テキストに応じた韻律の自然性の評定尺度。

スコア	説明
−3	著しく劣化し，発話として成立していない。
−2	発話として成立するが，機械的な声質や雑音などにより不自然。
−1	発話として成立するが，話し方がやや不自然。
0	発話としては自然だが，文に合わない棒読み。
+1	文に合った話し方を試みているが，不自然さが目立つ。
+2	文に合った話し方が概ね自然だが，改善の余地がある。
+3	文に合った話し方で，人間が自然に読んでいるように聞こえる。

候補を選定した。CER が同じ候補が複数ある場合は，UTMOS スコアが最も高い候補を選定し，両指標が同じ場合は候補番号が最も小さいものを採用した。以上の手順で選定した計 3,520 音声を以降の実験に用いた。

3.3　クラウドソーシングによる聴取実験

合成音声の韻律の自然性を評価するため，クラウドソーシングによる聴取実験を実施した。参加者の募集および実験の実施は株式会社インターグループ(注7)に委託し，日本語母語話者 50 名が参加した。参加者の属性を図 1 に示す。計 3,520 の評価対象音声をそれぞれ 5 名が評価する設計とし，総設問は 17,600 とした。全設問を 176 音声ずつの 100 セットに分割し，各参加者に 2 セットを割り当てた。したがって，参加者 1 名当たりの評価音声数は 352 である。

本実験では，発話としての成立度と，テキストに応じた話し方の適切さ・自然さを一つの尺度で評価するため，独自の 7 段階評定尺度（−3 から +3）を用いた。著しい音声の劣化などにより韻律の評価が困難な場合も想定し，負の評点には音声の劣化や発話の不自然さを，0 には発話としては自然だが文脈に合わない棒読みを，正の評点にはテキストに合った話し方の実現度を対応させた。各評点の定義を表 3 に示す。評価では，間，リズム，抑揚の適切さを重視し，強い感情表現は文脈上必要な場合にのみ求めた。

実験プロセスは，基本情報の入力，トレーニングセッション，および正式テストの 3 段階から構成した。基本情報として，性別，年齢層，英語能力，および出身地域を収集した。トレーニングセッションでは，各評点に対応する音声例を 1 件ずつ計 7 件提示し，読み上げ対象テキスト，評価基準，および評点の理由を併記した。参加者には，韻律やテキストに合った話し方に注目し，読み間違いや軽微な雑音を過度に重視しないよう説明した。全例の再生終了を正式テストの開始条件とし，正式テスト中もトレーニング内容を再確認できるようにした。正式テストでは，各音声と読み上げ対象テキストを提示した。問題文，指示，および各評点の説明は，サブセットごとの評価目的に合わせて設定した。サブセットの提示順は表 1 の記載順に固定し，各サブセット内の 22 システムの音声の提示順はセットごとにランダム化した。再生回数に制限は設けなかった。各設問には任意のコメント欄を設け，明らかな問題がある場合や評定に迷った場合に必要に応じて記入するよう案内した。Code-switching では追加質問として，当該文脈で英語部分の発音は「日本語アクセントに近い発音」と「ネイティブ英語に近い発音」のどちらがより自然であると考えるかを尋ねた。

(注7)：https://www.intergroup.co.jp/

3.4　評価結果

3.4.1　システム別結果

本実験の評定尺度は順序尺度であり，隣接する評点間の距離が等しいとは限らない。そこで，評点間の等間隔性を仮定する平均値（いわゆる平均オピニオンスコア）による比較ではなく，評点の分布と，+2 以上の評点が占める割合を用いて結果を分析した。表 3 に示すとおり、+2 および +3 はテキストに合った話し方が概ね自然，または人間らしく自然であると判断された評点に対応する。したがって、この割合を要約指標として用いた。以下では，評点を Y とし，この割合を $P(Y \geq 2)$ と表記する。

図 2 に，各システムの全評点の分布を示す。まず、Gemini 3.1 Flash TTS は +3 が 64.4%，+2 が 22.5% を占め，$P(Y \geq 2)$ は約 86.9% と，全システム中で最も高かった。また，負の評点と 0 の割合はいずれも小さく，不自然な発話や平板な話し方として評価される例が少なかった。一方，同じクローズウェイトシステムである ElevenLabs Multilingual v2 は，$P(Y \geq 2)$ が最も低く，クローズウェイトシステム間でも大きな違いが見られた。オープンウェイトシステムでは，Irodori-TTS が最も高い $P(Y \geq 2)$ を示し，+3 が 38.9%，+2 が 30.6% を占めた。ただし，同システムでも 0 が 8.5%，+1 が 13.8% を占め，テキストに合った自然な話し方の実現には改善の余地がある。他の多くのシステムでも，0 と負の評点が一定の割合を占めており，文脈に応じた表現と発話自体の自然性の両面で課題が残る。

また，日本語向けに開発・調整されたシステムであっても，Irodori-TTS, T5Gemma-TTS, Sarashina2.2-TTS, および LFM2.5-Audio-1.5B-JP の結果には大きな違いが見られた。モデル規模についても，500M の Irodori-TTS がより大規模な複数のシステムを上回っており，本評価ではモデル規模のみからテキストに応じた韻律の自然性を説明できるとは言えない。

3.4.2　サブセット別結果

図 3 に，各システムとサブセットの組合せにおける $P(Y \geq 2)$ を示す。Gemini 3.1 Flash TTS は，全サブセットで 70% を超え，特に Emotions では 95% に達し，幅広い条件で高い評価を得た。同サブセットで 50% を超えたのは Gemini 3.1 Flash TTS, Irodori-TTS, T5Gemma-TTS, および Zonos2 の 4 システムのみであり，他の多くのシステムでは，テキストの意味に応じた自然な感情表現に課題が見られた。

サブセット間にも評価傾向の違いが見られた。Normal, Rhyme, Compound Nouns では比較的高い割合を示すシステ

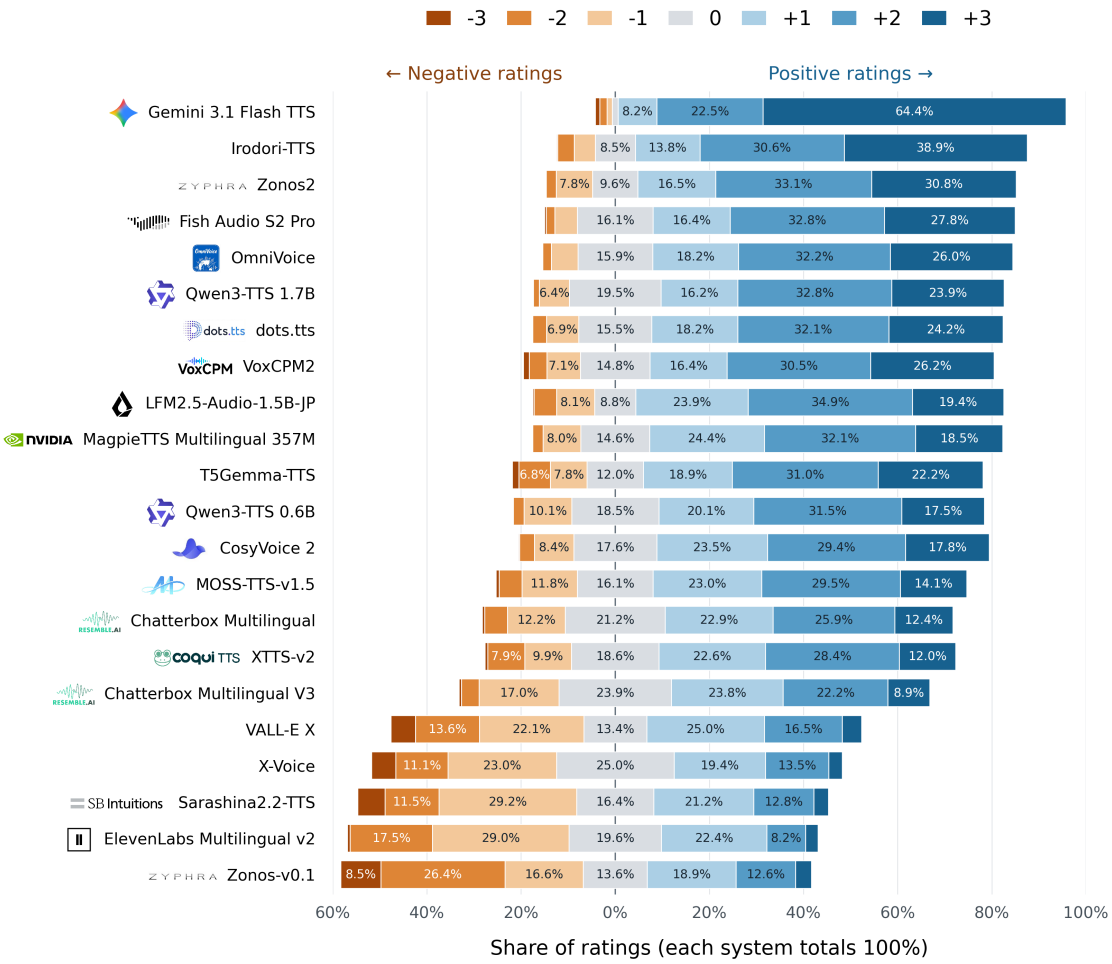


図 2　TTS システム別の順序尺度評価の分布。システムは、+2 または +3 の評価が占める割合の高い順に、上から下へ配置した。各横棒は各システムに対する評価の 100% を表し、ゼロを中心に配置した。割合が 6% 以上の区分には百分率を表示した。

ムが多い一方，Repetition と Emotions では低い割合にとどまるシステムが多かった。Syntactic Complexity でも，Normal より低い割合を示すシステムが多く，複雑な構文に応じた自然な韻律生成が課題となることが示唆される。Punctuations では，例えば T5Gemma-TTS が 77% を示すなど，全体順位とは異なるシステムごとの特徴が見られた。

Code-switching では，全体順位と異なる傾向が顕著であった。Fish Audio S2 Pro が 85% と最も高く，全体で最も高い評価を得た Gemini 3.1 Flash TTS の 73% を上回った。一方，全体順位が 2 位の Irodori-TTS は 48%，T5Gemma-TTS は 30% にとどまり，日本語向けに開発されたシステムでも，言語切り替えを伴う発話には課題が残ることが示された。このような順位の逆転は，全体の結果に加えてサブセット別の評価を行う意義を示している。言語切り替え時の具体的な問題や，各サブセットにおける詳細な誤り分析は今後の課題とする。

4. おわりに

本研究では，テキストに応じた韻律の自然性を評価するための高難度日本語 TTS 評価コーパス J-HARD-TTS-EVAL-v2 を構築し，22 システムを対象に日本語母語話者 50 名による聴取実験を実施した。今後は，サブセット別の詳細分析と，テキストに応じた韻律の自然性を自動予測するモデルの開発に取り組む。

謝辞　本研究は，JST, BOOST JPMJBY24C9, JSPS, KAKENHI 25K00143 と 26H0253 の支援を受けたものである．

文　献

[1] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He, S. Zhao, T. Qin, F. Soong, and T.-Y. Liu, “NaturalSpeech: End-to-End Text-to-Speech Synthesis With Human-Level Quality,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.46, no.6, pp.4234–4245, 2024.

[2] Takaaki Saeki and Detai Xin and Wataru Nakata and Tomoki Koriyama and Shinnosuke Takamichi and Hiroshi Saruwatari, “UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022,” Proc. Interspeech, pp.4521–4525, 2022.

[3] B. Zhang, C. Guo, G. Yang, H. Yu, H. Zhang, H. Lei, J. Mai, J. Yan, K. Yang, M. Yang, et al., “MiniMax-Speech: Intrinsic Zero-Shot Text-to-Speech with a Learnable Speaker Encoder,” arXiv preprint arXiv:2505.07916, 2025.

[4] 今井柊平，榎本悠久，金子剛士，中村泰貴，“系列整合性評価に特化した高難度日本語テキスト音声合成コーパスの検討，” 日本音響学会研究発表会講演論文集，pp.1–Q–48，Sept. 2025.

[5] T. Hayashi, S. Watanabe, T. Toda, K. Takeda, S. Toshniwal, and K. Livescu, “Pre-Trained Text Embeddings for Enhanced Text-to-Speech Synthesis,” Proc. Interspeech, pp.4430–4434, 2019.

[6] Y. Jia, H. Zen, J. Shen, Y. Zhang, and Y. Wu, “PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS,” Proc. Interspeech, pp.151–155, 2021.

[7] Y.A. Li, C. Han, X. Jiang, and N. Mesgarani, “Phoneme-Level BERT

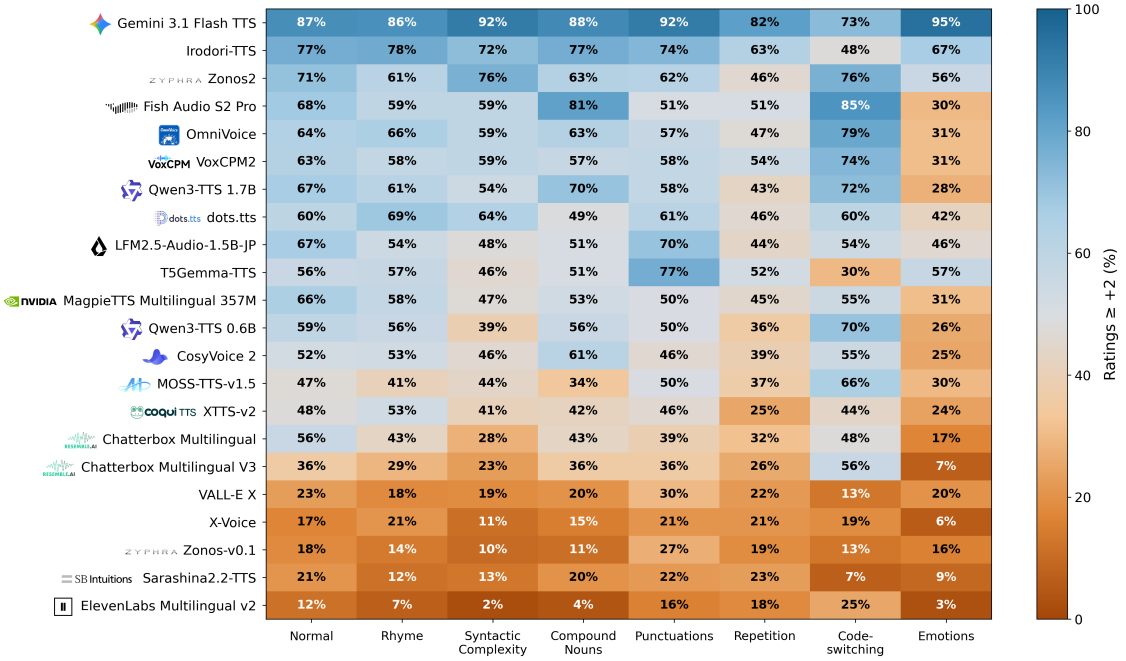


図 3　システムとサブセットの各組み合わせにおける、+2 または +3 の評価が占める割合 $P(Y \geq +2)$。各セルは百分率を表示した。システムは、全 8 サブセットを通じた $P(Y \geq +2)$ の高い順に上から下へ配置し、図 2 と同じ並び順とした。

for Enhanced Prosody of Text-to-Speech with Grapheme Predictions," Proc. ICASSP, pp.1–5, 2023.

[8] K. Ito and L. Johnson, "The LJ Speech Dataset," 2017.

[9] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," 2017.

[10] H. Zen, V. Dang, R. Clark, Y. Zhang, R.J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," Proc. Interspeech, pp.1526–1530, 2019.

[11] M. Lajszczak, G. Cámbara, Y. Li, F. Beyhan, A. Van Korlaar, F. Yang, A. Joly, Á. Martín-Cortinas, A. Abbas, A. Michalski, et al., "BASE TTS: Lessons from building a billion-parameter Text-to-Speech model on 100K hours of data," arXiv preprint arXiv:2402.08093, 2024.

[12] R.R. Manku, Y. Tang, X. Shi, M. Li, and A. Smola, "EmergentTTS-Eval: Evaluating TTS Models on Complex Prosodic, Expressiveness, and Linguistic Challenges Using Model-as-a-Judge," Proc. NeruIPS, vol.38, Main Conference, 2025.

[13] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," arXiv preprint arXiv:1711.00354, 2017.

[14] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," arXiv preprint arXiv:1908.06248, 2019.

[15] T. Okamoto, Y. Shiga, and H. Kawai, "Hi-Fi-CAPTAIN: High-fidelity and high-capacity conversational speech synthesis corpus developed by NICT," https://ast-astrec.nict.go.jp/en/release/hi-fi-captain/, 2023.

[16] L. Liu, S. Zhu, K. Washizaki, R. Yoneyama, H. Jeon, M. Zhao, Y. Fujita, H. Shi, N. Yoshida, Y. Gao, et al., "Sarashina2.2-TTS: Tackling Kanji Polyphony in Japanese Speech Generation via Data Scaling and Targeted Data Synthesis," arXiv preprint arXiv:2606.25369, 2026.

[17] S. Chen, C. Wang, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers," IEEE TASLP, vol.33, pp.705–718, 2025.

[18] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, "XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model," Proc. Interspeech, pp.4978–4982, 2024.

[19] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang, et al., "CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models," arXiv preprint arXiv:2412.10117, 2024.

[20] C. Arata and K. Kurihara, "T5Gemma-TTS Technical Report," arXiv preprint arXiv:2604.01760, 2026.

[21] H. Hu, X. Zhu, T. He, D. Guo, B. Zhang, X. Wang, Z. Guo, Z. Jiang, H. Hao, Z. Guo, et al., "Qwen3-TTS Technical Report," arXiv preprint arXiv:2601.15621, 2026.

[22] S. Liao, Y. Wang, S. Liu, Y. Cheng, R. Zhang, T. Li, S. Li, Y. Zheng, X. Liu, Q. Wang, et al., "Fish Audio S2 Technical Report," arXiv preprint arXiv:2603.08823, 2026.

[23] R. Xu, Q. Liu, H. Li, Y. Chen, Z. Niu, Y. Yang, J. Zhao, K. Li, B. Sisman, Q. Cheng, et al., "X-Voice: Enabling Everyone to Speak 30 Languages via Zero-Shot Cross-Lingual Voice Cloning," arXiv preprint arXiv:2605.05611, 2026.

[24] H. Zhu, L. Ye, W. Kang, Z. Yao, L. Guo, F. Kuang, Z. Han, W. Zhuang, L. Lin, and D. Povey, "OmniVoice: Towards Omnilingual Zero-Shot Text-to-Speech with Diffusion Language Models," arXiv preprint arXiv:2604.00688, 2026.

[25] Y. Zhou, G. Zeng, X. Liu, X. Li, R. Yu, J. Gui, J. Wu, Z. Wang, X. Shen, R. Ye, et al., "Voxcpm2 technical report," arXiv preprint arXiv:2606.06928, 2026.

[26] Y. Gong, B. Jiang, Y. Zhao, Y. Yuan, K. Chen, Y. Jiang, C. Chang, D. Hong, M. Chen, R. Li, et al., "Moss-tts technical report," arXiv preprint arXiv:2603.18090, 2026.

[27] G. Clark, S. Mejjoute, M. Osman, G. Close, and B. Millidge, "ZONOS2 Technical Report," arXiv preprint arXiv:2606.24320, 2026.

[28] S. Lian, C. Li, B. Li, H. Wang, D. Zheng, J. Tian, Y. Ma, C. Zhang, and K. Yu, "dots.tts Technical Report," arXiv preprint arXiv:2606.07080, 2026.

[29] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," Proc. LREC, pp.4218–4222, May 2020.

[30] X. Shi, X. Wang, Z. Guo, Y. Wang, P. Zhang, X. Zhang, Z. Guo, H. Hao, Y. Xi, B. Yang, et al., "Qwen3-ASR Technical Report," arXiv preprint arXiv:2601.21337, 2026.