プロジェクトページ

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







HUANG Wen-Chin（名大）・今井柊平・榎本悠久（Parakeet(株)）・
栗原清（(株)松尾研/東大）・齋藤佑樹（東大/産総研）

第28回音声言語シンポジウム兼第13回自然言語処理シンポジウム
2026年11月25日

本研究の要約

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

v2だよ！

最先端日本語TTSシステム22個を
日本語母語話者50人に
評価してもらったよ！

日本語TTSに着目し、
8 種類の高難度テキスト
を作ったよ！

イメージ

- **これはもう簡単なので…**

この料理の作り方を教えて
いただけませんか。



- **こういうのを作りたいよね**

いやだーー！ビュッフェの
予約を忘れたなんて！



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。



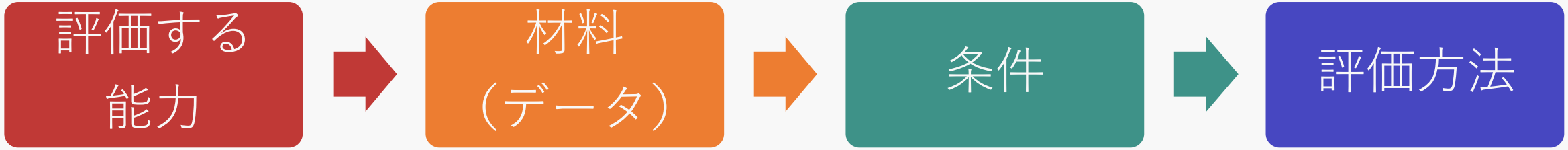
目次

- **背景**
- **コーパス構築：サブセット一覧＆定義**
- **大規模システム評価**
  - 評価したシステム
  - 合成方式、候補選定
  - 聴取実験
    - 聴取者属性
    - 評価尺度
    - 評価結果

目次

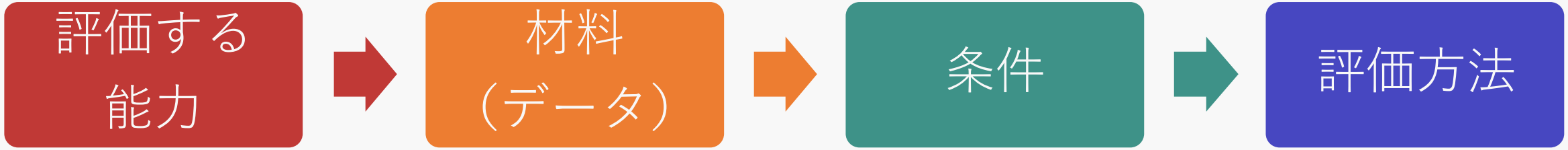
- 背景
- コーパス構築：サブセット一覧＆定義
- 大規模システム評価
  - 評価したシステム
  - 合成方式、候補選定
  - 聴取実験
    - 聴取者属性
    - 評価尺度
    - 評価結果

TTS評価に必要な要素 (1/2)



- **Neural TTS初期**（Tacotron [Wang+; '17]）**～中期**（VITS [Kim+; '21]）**：単一話者TTS**
  - 評価する能力：**自然性**
  - 材料（データ）：LJSpeech [Ito; '17] などの単一話者データセット
  - 条件：単一話者TTS（学習話者の声を再現する）
  - 評価方法：平均オピニオン評点方式の聴取実験による主観評価
- **NaturalSpeech [Tan+; '24]：単一話者TTSにおいて、肉声と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**

TTS評価に必要な要素 (2/2)



- **現在・主流（VALL-E [Wang+; '23] などのLLMベースTTS以降）：ゼロショットTTS**
  - 評価する能力：**明瞭度、話者類似度**
  - 材料（データ）：LibriTTS [Zen+;' 19]などの多話者データセット
  - 条件：ゼロショットTTS（学習話者以外の参照音声を用いて合成）
  - 評価方法：聴取実験による主観評価より、ASR結果に基づく**WER**、**話者埋め込みのコサイン類似度**、**UTMOS** [Saeki+' ;22] などの主観評価予測モデル
- **最近：Voice ArenaやArtificial Analysis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**

着目点：「自然な韻律を生成する」能力

- **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**
  →材料が必要！



- 評価したい言語現象を意図的に含むテキストを用意する必要がある
- BASE TTS [Lajszczak+; ‘24]、EmergentTTS-Eval [Manku+; ‘25]：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。

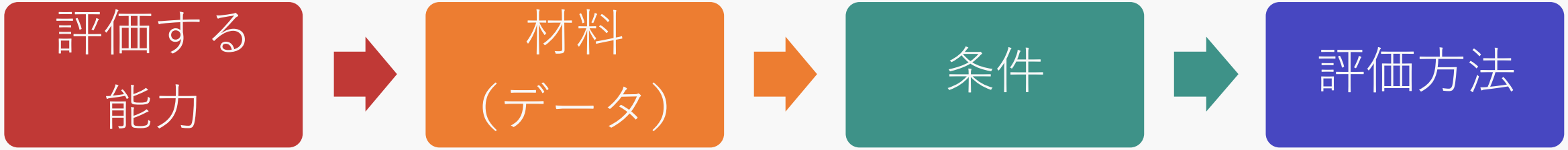
- LJSpeech、VCTK、LibriTTSなど常用データ：
  書籍や新聞などの読み上げ音声
  →豊かな表現を含んでいない

日本語TTS常用ベンチマーク

ベンチマーク名	テキストドメイン	話者数	時間/発話数
JSUT basic 5000 [Sonobe+; ‘17]	Wikipedia + TANAKA	1	10/5k
JVS [Takamichi+; ‘19]	Wikipedia	100	24/10k
HiFi-Captain (JP) [Okamoto+; ‘23]	対話調	2	44/33k
sbintuitions/Joyo-Kanji-Yomi-Benchmark [Liu+; ‘26]	常用漢字	N/A	N/A
J-HARD-TTS-EVAL [Imai+; ‘25]	short / repetition / rhyme / continuation	N/A	N/A

短い発話，語句の繰り返し，押韻，途中で打ち切られた文
→テキストに応じた自然な韻律を生成できるかを検証するために、さらなる多様な言語現象が必要

本研究の立ち位置



- 評価する能力：テキストから自然な韻律を生成する
- 材料：８サブセットからなる高難度テキスト
  （普通文、押韻、係り受けが複雑な長文、複合語、句読点、繰り返し、言語切り替え、感情）
- 条件：ゼロショットTTS
- 評価方法：22日本語TTSを対象に日本語話者50名による聴取実験

目次

- 背景
- コーパス構築：サブセット一覧＆定義
- 大規模システム評価
  - 評価したシステム
  - 合成方式、候補選定
  - 聴取実験
    - 聴取者属性
    - 評価尺度
    - 評価結果

サブセット一覧

- 8つのサブセットに対して20文を作成

Normal	普通文
Rhyme	押韻
Syntactic Complexity	係り受けが複雑な長文
Compound Nouns	複合語
Punctuations	句読点
Repetition	繰り返し
Code-switching	言語切り替え
Emotions	感情

- 以下、各サブセットを紹介
  - 明記がはければ、構築方法：Gemini 3.1 Proで候補文を生成・確認・修正

構築方法：プロンプト様式

日本語TTSの性能を評価するためのテキストコーパスを作成してください。
これから、あなたは「＜サブセット名＞」というコーパスを作成します。
＜サブセットの説明＞
出力する際には適切に漢字を使用してください。カタカナの外来語を入れても構いませんが、出来れだけ漢字を使ってください。
変に（）や「」で文中で括らなくて大丈夫です。
文章は口語的な文章や、そうではない文章など、色々と散りばめてください。
文の意味はできる限り通じるようにしてください。
各文章の長さは、自然な速度で読み上げて～秒から～秒程度になるように調整してください。
様々な発話のバリエーションが欲しいので、以下の長さの文を作ってください。
＜詳細スペック>

（もしサンプルがあれば）
以下はサンプル文章で、これらと似ているように作成してください。
・＜サンプル＞

出力時のフォーマットは以下を守り、xxは01から20までの20文を考えてください。
|＜サブセット名＞-01.wav|発話内容|
|＜サブセット名＞-02.wav|発話内容|
...
|＜サブセット名＞-20.wav|発話内容|

なお、以下の条件を厳密に遵守してください。
・漢字を使用する場合は常用漢字のみを使用してください
・読みの曖昧性がある単語は使わないでください。例えば、「今日」は文脈次第で「きょう」「こんにち」と読めますし、他にも「昨日」は「きのう」と「さくじつ」、「描く」は「かく」と「えがく」、「17」は「じゅうなな」と「じゅうしち」など、曖昧な単語が数多くあります。これら以外にも、読みに曖昧性がある単語は使用しないようにしてください。
・全ての文章は極力単語の被りがないようにしてください。エントロピーを最大限大きくすることを意識してください。
・文章の構成の仕方についても極力被りが内容にし、エントロピーを最大限大きくすることを意識してください。
・音素バランスを意識してください。「py」「dy」「by」「gw」「kw」などの比較的出現しづらい音素を最低1回は使用してください。

サブセット：Normal（普通文）

- **評価基準：**
  表示されたテキストの読み上げとして、話し方の自然さ
- **構築方法：Minimaxai/tts-multilingual-test-setの日本語100文から20文を選出**
  - JSUT, JVSなど常用ベンチマークを避けたい
- **例文：**

> 週末は実家に帰省する予定なんです。
>
> 最近始めたヨガのレッスンがとても気に入っています。

サブセット：Rhyme（押韻）

- **評価基準：**
  似た音が繰り返される中でも、一本調子にならず自然なリズムや抑揚を保てているか
- **構築方法：J-HARD-TTS-EVAL-v1をそのまま流用**
- **例文：**

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。
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。

サブセット：Syntactic complexity（係り受けが複雑な長文）

- **評価基準：**
  句読点がない場所でも、文の意味のまとまりに応じた間や息継ぎが自然か

- **例文：**

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。
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。

サブセット：Compound nouns（複合語）

- **評価基準：**
  言葉がつながった時のアクセント変化が一本調子にならず、自然なリズムで読めているか

カタカナ	漢字	複合文長	例文
無	有	中	循環型社会形成推進基本法に基づき、地域の廃棄物処理施設ではリサイクルの徹底が義務付けられている。
無	有	超長	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。
有	無	長	グローバルサプライチェーンマネジメントシステムの最適化プロジェクトが始動し、物流コストの削減を目指す。
中〜短の複合語を複数			青年海外協力隊の活動報告会に参加して、開発途上国でのボランティア体験談を熱心に聴いた。

サブセット：Punctuations（句読点）

- **評価基準：**
  記号が文脈に合った自然な抑揚、間、語尾、感情の変化へ変換されているか

句読点	例文	スペック
？	また同じ書類をなくしたのですか、立派な管理ですね？	皮肉、呆れ
	扉の奥からグワッと音がしたけど、何ですか？？	驚き、混乱
！	その赤い線の内側へ、今すぐ下がってください！	命令、指示
	いやだーー！ビュッフェの予約を忘れたなんて！	強い感情
……	その件は、ええと……うまく言える言葉が見つかりません。	言葉探し
	階段を下りた先に、黒い布が……いや、ただの上着でした。	思考の途中で文が切れ
「」	係の女性は「まだ入らないで」と低い声で伝えました。	引用
	この展示では「静けさ」を耳で感じてもらいます。	強調

サブセット：Repetition（繰り返し）

- **評価基準：**
  同じ語句や表現が繰り返される中でも、途中で止まらず、自然な話し方として最後まで読めているか
- **構築方法：J-HARD-TTS-EVAL-v1をそのまま流用**
- **例文：**

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？
お願いお願いお願いお願いお願いお願いお願い、これが最後の頼みだからさ。

サブセット：Code-switching（言語切り替え）

- **評価基準：**
  日本語から英語、英語から日本語へ切り替わる際の声質、声の高さ、リズムのつながりが自然か

スペック	例文
文中コードスイッチング	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。
文間コードスイッチング（日本語から英語＆英語から日本語）	I sent the revised manuscript to the coauthors last night. 返事が来るまで、関連する表を見直します。
付加コードスイッチング（感動詞や日本語の終助詞など）	初めての環境では小さな不具合が起きやすいです、Don't worry about it.

サブセット：Emotions（感情）

- **評価基準：**
  怒り、喜び、悲しみ、驚きなど、テキストの意味に応じた感情が声色や抑揚に自然に反映されているか

感情	例文
安堵	あぁ、よかった……。無事に見つかったんだね。本当に、生きた心地がしなかったよ。
郷愁	この古い校舎の匂い、懐かしいなあ。夕暮れまで泥だらけになって遊んだ記憶が、鮮やかによみがえるよ。
焦燥	やばい、もう時間がないぞ！急いで荷物をまとめろ！このままだと全員、逃げ遅れてしまう！

目次

- 背景
- コーパス構築：サブセット一覧＆定義
- **大規模システム評価**
  - 評価したシステム
  - 合成方式、候補選定
  - 聴取実験
    - 聴取者属性
    - 評価尺度
    - 評価結果

評価したシステム (1/3)

	クローズウェイトモデル	固定話者	日本語特化	リリース	サイズ
VALL-E-X（非公式実装）				2023.8	300M
ElevenLabs Multilingual v2	V	V		2023.8	N/A
XTTS-v2				2023.11	467M
CosyVoice2				2024.12	500M
Zonos-v0.1				2025.2	1.6B
Chatterbox Multilingual				2025.9	500M
T5Gemma-TTS			V	2025.12	5B
MagpieTTS Multilingual 357M		V		2025.12	357M

評価したシステム (2/3)

	クローズウェイトモデル	固定話者	日本語特化	リリース	サイズ
Qwen3-TTS 0.6B				2026.1	0.6B
Qwen3-TTS 1.7B				2026.1	1.7B
Fish Audio S2 Pro				2026.3	4B
X-Voice				2026.4	400M
Sarashina2.2-TTS			V	2026.4	800M
Irodori-TTS			V	2026.4	500M
OmniVoice				2026.4	800M
VoxCPM2				2026.4	2B

評価したシステム (3/3)

	クローズウェイトモデル	固定話者	日本語特化	リリース	サイズ
Gemini 3.1 Flash TTS	V	V		2026.4	N/A
MOSS-TTS 1.5				2026.5	8B
LFM2.5-Audio-1.5B-JP		V	V	2026.6	1.5B
Chatterbox Multilingual V3				2026.6	500M
Zonos2				2026.6	8B
dots.tts				2026.6	2B

合成条件

- **デフォルト：参照音声によるゼロショット合成方式**
  - 一部、提供された既知話者を使用
- **参照音声データ：Common Voice [Ardila+;, ‘21] 21.0の日本語テストセット。高品質の参照音声を選ぶために、以下を行った：**
  - 自動選別：冒頭・末尾の無音除去後の長さが5秒以上 & UTMOSスコアが2.5以上
  - UTMOSスコアの高い順に試聴し、明瞭さ、雑音、書き起こしの正確さ、母語話者らしい発話かを確認・判断した上で、性別ごとに上位10サンプルを選択
- **最終的な参照音声：女性声*10 + 男性声*10 = 20サンプル。**
  - 全サブセットで同じ20サンプルを使用し、対応関係は個別にシャッフル

候補生成とASRに基づくサンプル選定

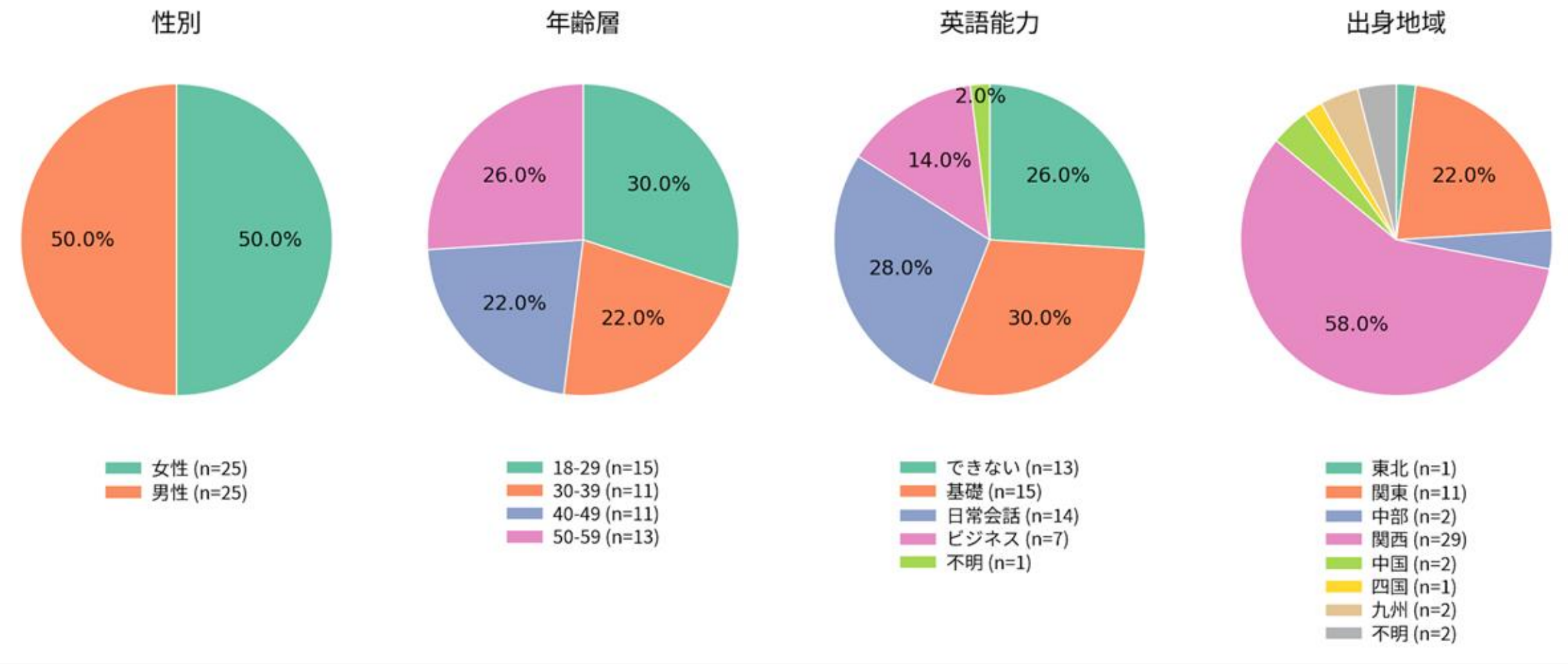
ハルシネーション対策

1. 1テキストにつき、候補サンプル5個を生成（Gemini TTSは1個）
2. 各候補を複数のASRシステムで書き起こし：
   IBM Granite Speech 4.1 2B, Qwen3-ASR 1.7B,
   ReazonSpeech ESPnet, ReazonSpeech NeMo
3. ASR結果のうち、表記に基づくCERの最小値を候補のCERとして採用
4. 5個の候補サンプルから、次の優先順位で選定
   1. CERが最も低い候補
   2. CERが同じ場合、UTMOSが最も高い候補。
   3. 両方が同じ場合、候補番号が最も小さい候補。
5. 選定した音声を聴取実験に使用

クラウドソーシングによる聴取実験

- 参加者募集・実験実施：株式会社インターグループ
- 評価対象音声：
  22システム * 8 サブセット * 20 サンプル = 3,520 サンプル
  それぞれ5人が評価 →総設問：17,600
- 日本語母語話者50人を募集→1 人当たりの評価音声数：352

受験者情報（50名）


性別
50.0%
50.0%
女性 (n=25)
男性 (n=25)
年齢層
26.0%
30.0%
22.0%
22.0%
18-29 (n=15)
30-39 (n=11)
40-49 (n=11)
50-59 (n=13)
英語能力
2.0%
14.0%
26.0%
28.0%
30.0%
できない (n=13)
基礎 (n=15)
日常会話 (n=14)
ビジネス (n=7)
不明 (n=1)
出身地域
22.0%
58.0%
東北 (n=1)
関東 (n=11)
中部 (n=2)
関西 (n=29)
中国 (n=2)
四国 (n=1)
九州 (n=2)
不明 (n=2)

評価尺度

テキストに合った自然さと話し方を重視する、独自の－3〜＋3の7段階評価尺度

- ＜0： 劣化した音声や不自然な発話
- 0： 発話としては自然だが、文脈に合わない棒読み
- ＋1／＋2： 表現を試みているが不自然／概ね自然で適切
- ＋3： テキストに合った、人間らしく自然な話し方

トレーニングセッション

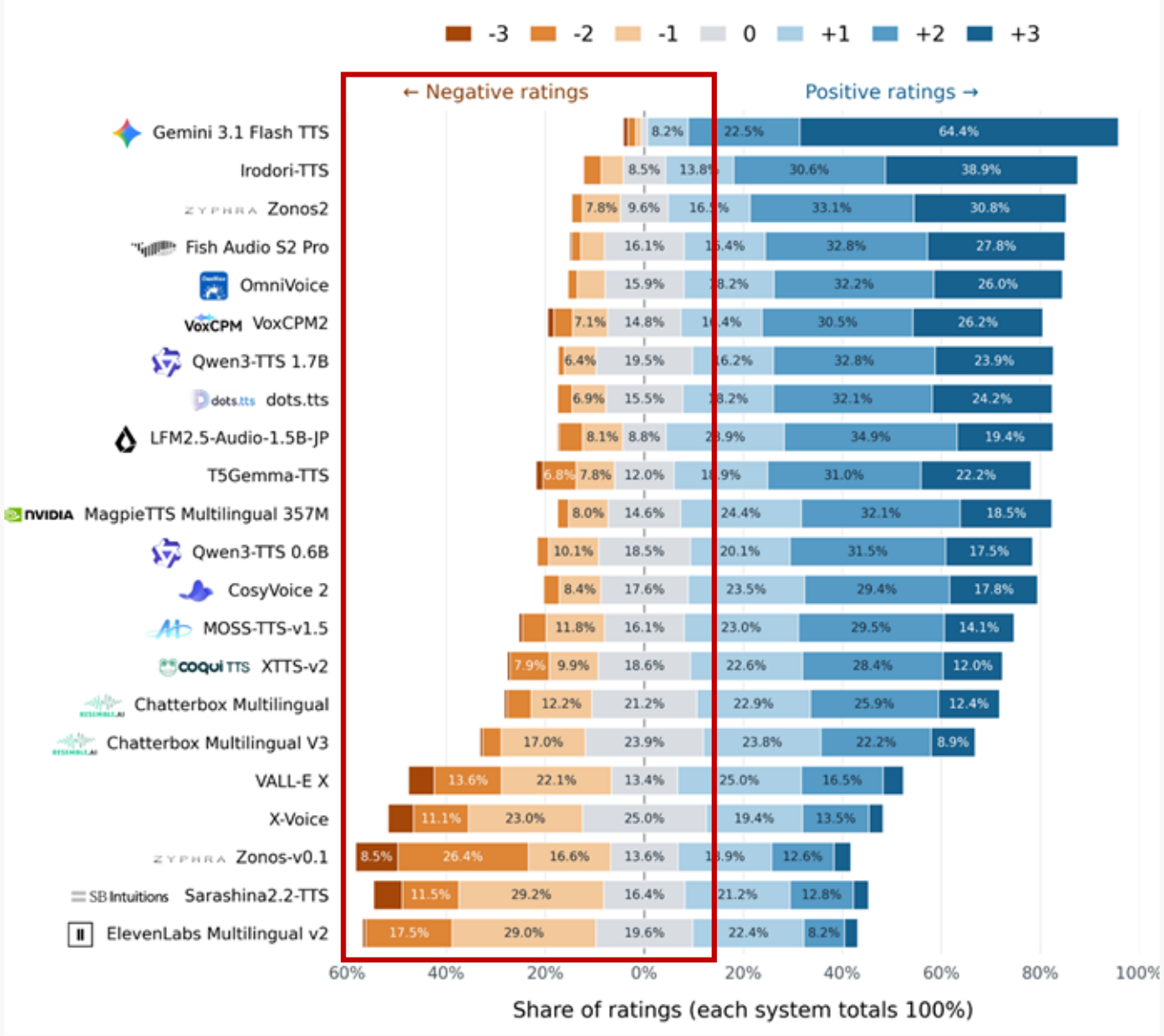
スコア	音声例	読み上げテキスト	説明
-3		目の前で彼女が誘拐された時、私は怖くて身動きもできなかった。	発話として成立していない（雑音がひどい、明瞭度が低い）
-2		あんたは誰だと思っているの？生意気な態度は許せない。	発話として成立しているが、不自然（機械的な声質、軽度な雑音）
-1		あの人がヴァンパイアだなんて信じられない！本当に血を吸っているの？	発話として成立しているが、話し方がやや不自然
0		最近の私の人生は何もかもが停滞していて、どうしようもなく悲しい。	発話としては自然だが、棒読み（文に合わない平板さや不自然さがある）
1		うわ、この音楽、騒がしすぎて集中できないよ。	文に合った話し方をしようとしているが、不自然さが目立つ
2		イエーイ！今日は手頃な価格で美味しい食事を楽しめたんだ！	文に合った話し方がある程度自然だが、まだ少し改善点がある
3		ふん、いい加減にしろよ！いつもダメ出しばかりで何も褒め言葉がないんだ。	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

集計方法

- **本実験の評定尺度：隣接する評点間の距離が等しいとは限らず**
  - 等間隔性を仮定する平均値（いわゆるMOS）は不相応
- **評点の分布と，+2 以上の評点が占める割合を使用**
  - +2, +3：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

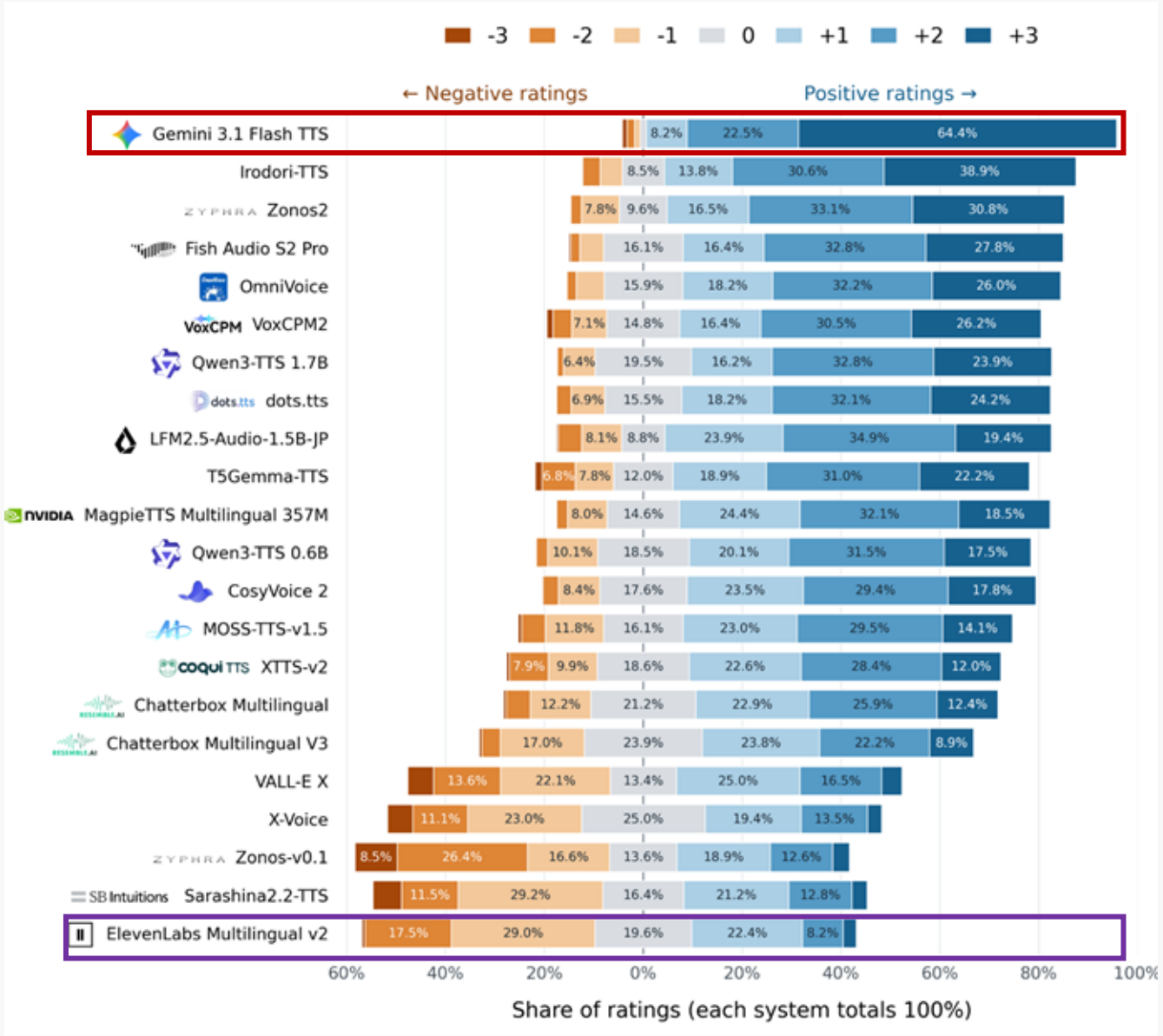
評価結果：システム別

- 多くのシステムは 0 と負の評点が一定の割合
- →文脈に応じた表現と発話自体の自然性の両面が課題



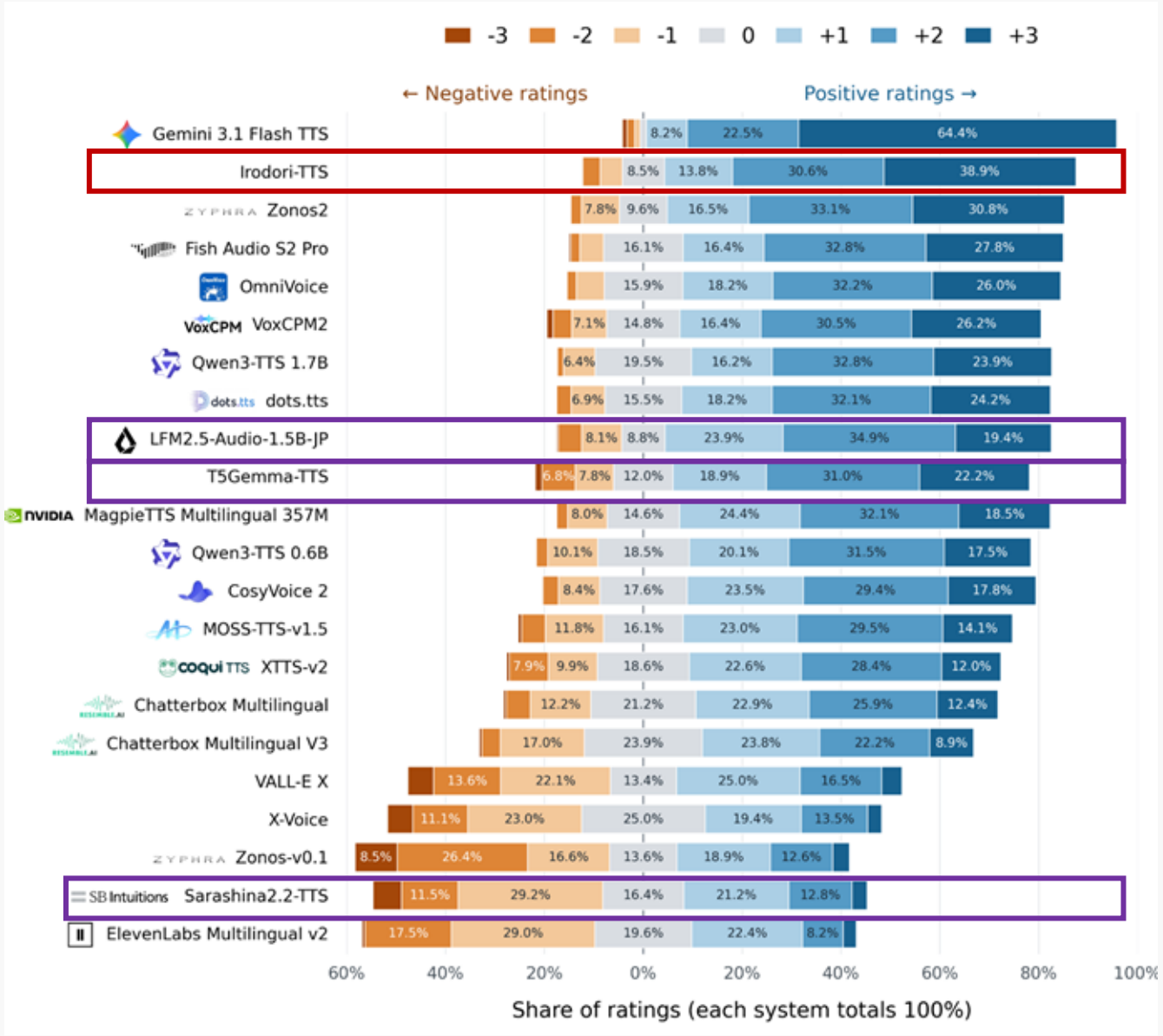
評価結果：システム別

- 首位：
  Gemini 3.1 Flash TTS
  - $P(Y \geq 2) = 86.9\%$
  - $P(Y \leq 2)$（不自然な発話や平板な話し方）の割合は小さく
- 末位：
  ElevenLabs Multilingual v2
- →クローズウェイトシステム間でも差異が顕著



評価結果：システム別

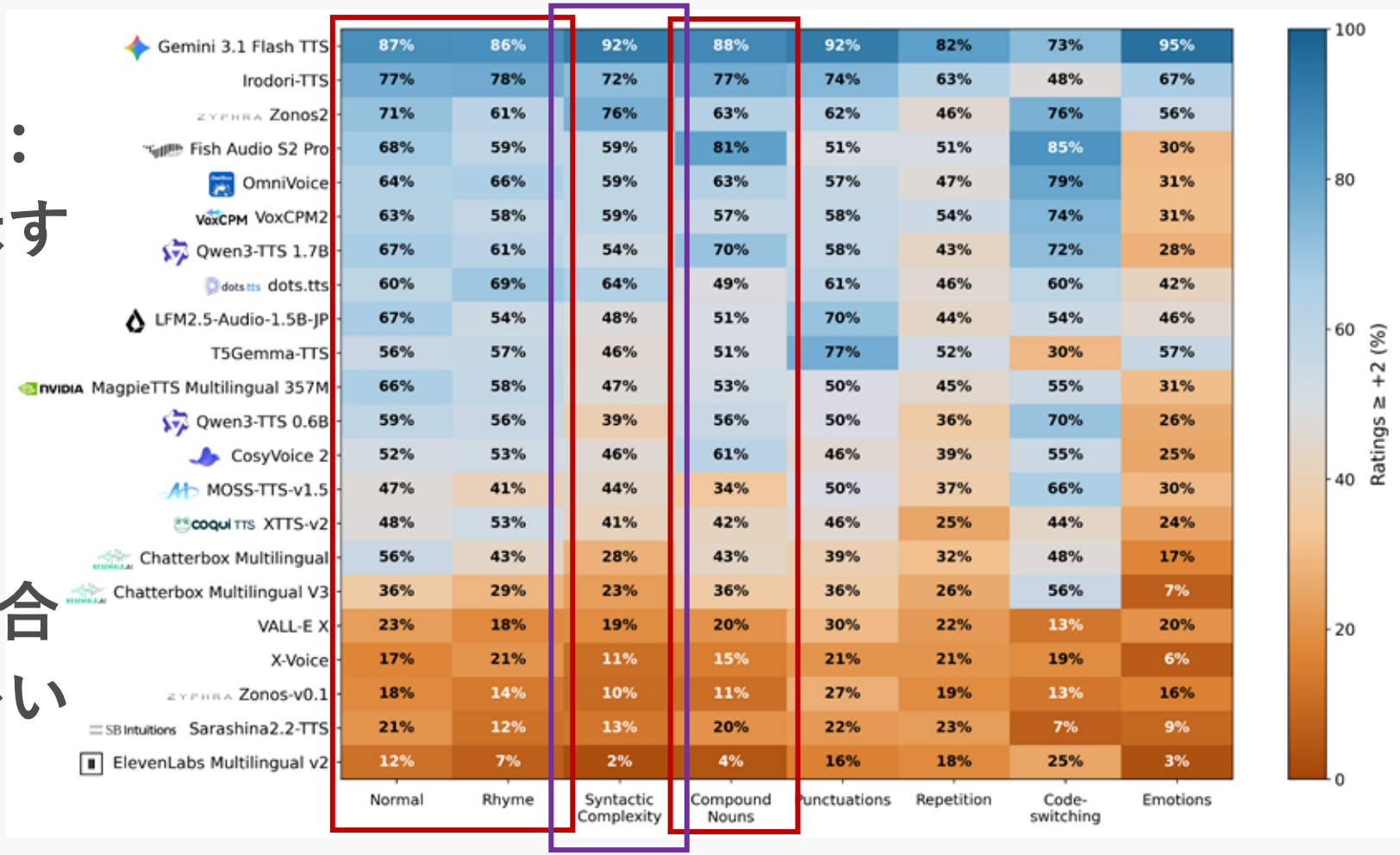
- **オープンウェイトの首位：Irodori-TTS**
  - モデルサイズ：僅か500M
    →本評価ではモデル規模のみから説明できるとは言えず
  - 0 が8.5%，+1 が13.8% を占め
    →改善の余地が有り
- **日本語向けのシステム：差異が顕著**



評価結果：システムーサブセット別

（+2 以上の評点が占める割合）

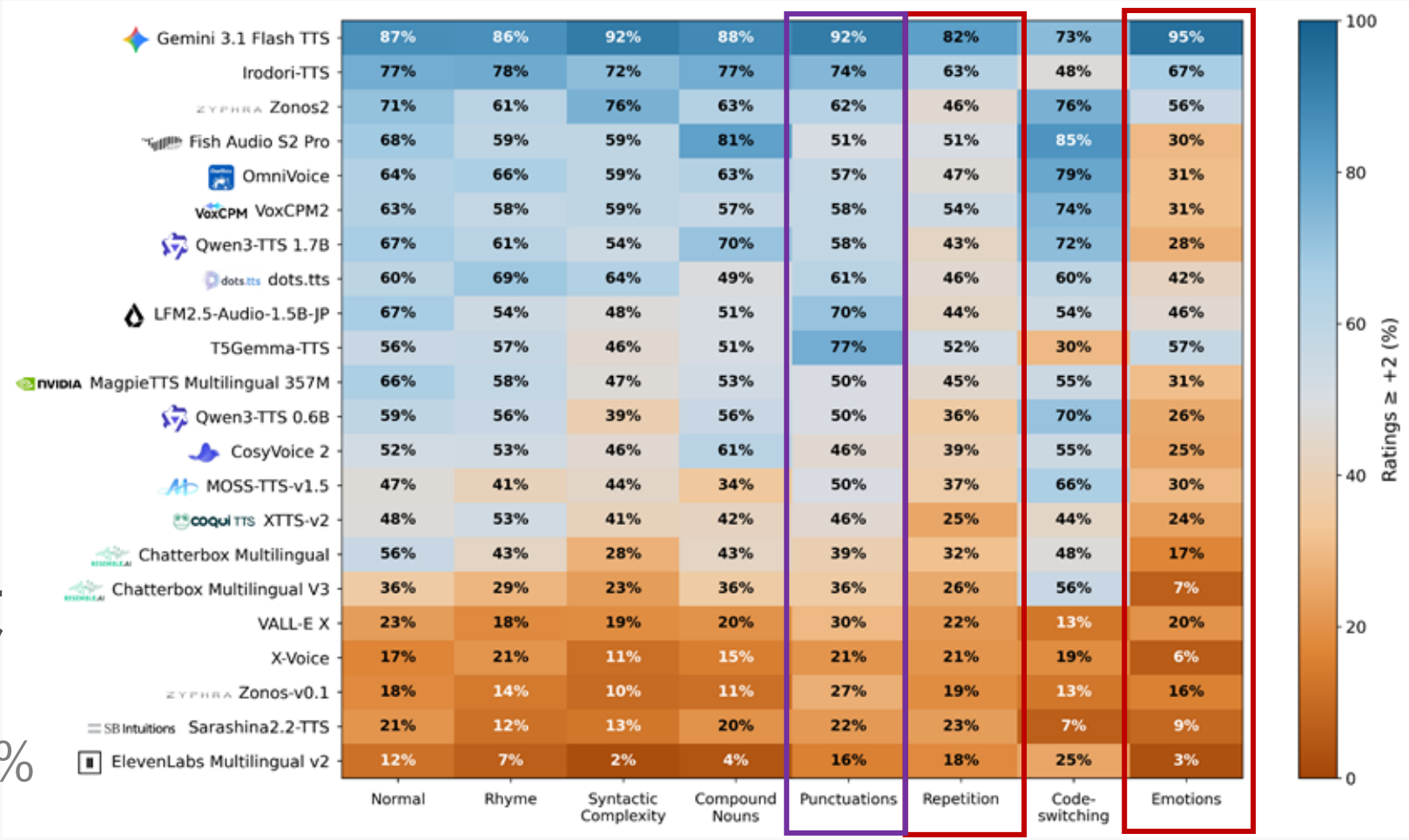
- **Normal、Rhyme、Compound Nouns：**
  比較的高い割合を示すシステムが多い
- **Syntactic Complexity：**
  Normal より低い割合を示すシステムが多い



評価結果：システムーサブセット別

（+2 以上の評点が占める割合）

- **Repetition、Emotions：**
  低い割合にとどまるシステムが多い
- **Punctuations：**
  全体順位とは異なるシステムごとの特徴
  - 例：
    T5Gemma-TTS：77%

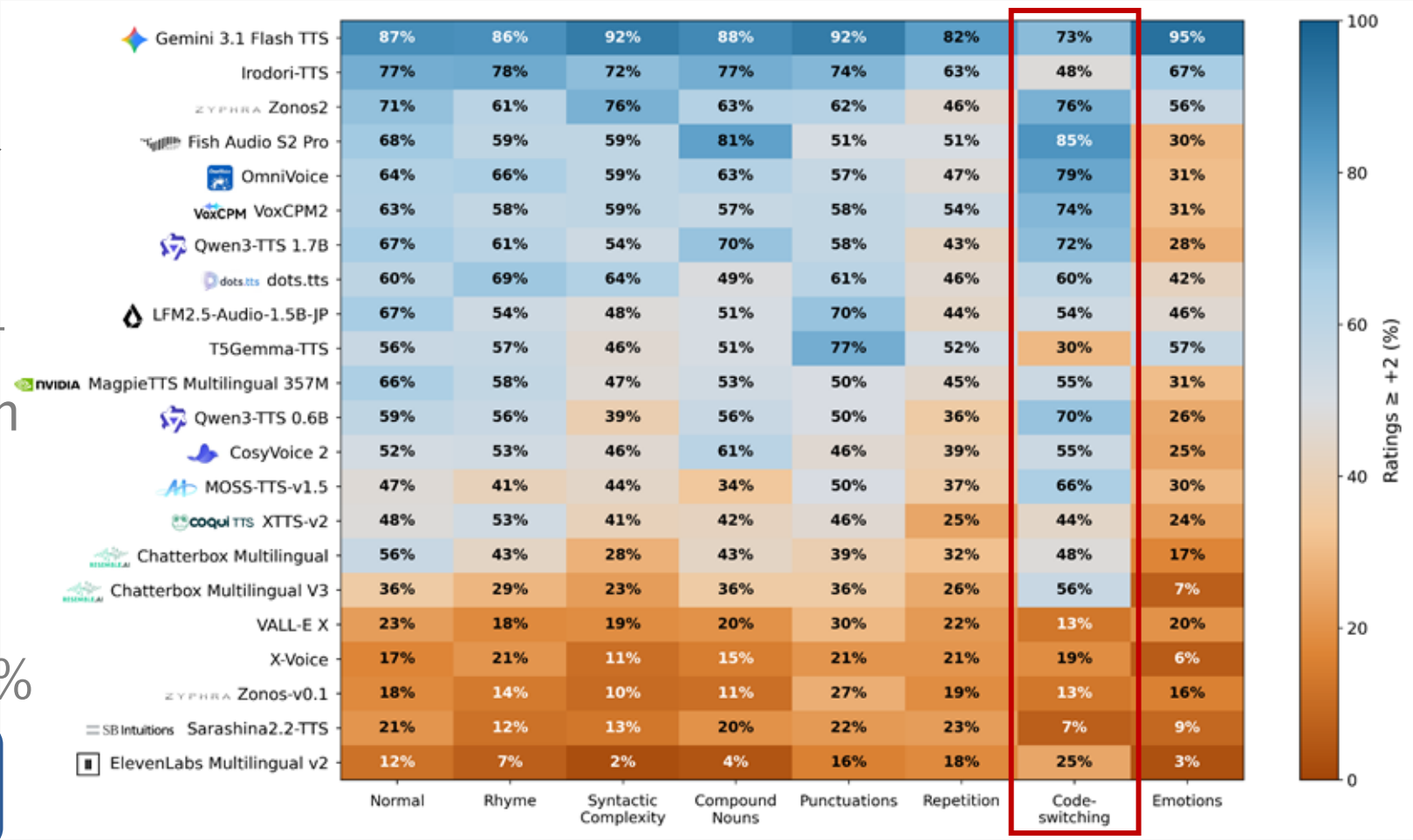


評価結果：システムーサブセット別

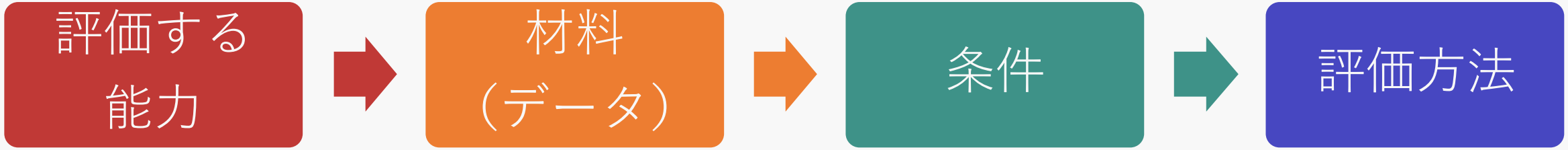
（+2 以上の評点が占める割合）

- **Code-switching：全体順位と異なる傾向が顕著**
  - 首位：Fish Audio S2 Pro（85%）が全体首位のGemini 3.1 Flash TTS（73%）を上回
  - 全体2位のIrodori-TTS：48%；T5Gemma-TTS：30%

サブセット別の評価を行う意義を示唆



Recap：本研究の立ち位置



- 評価する能力：テキストから自然な韻律を生成する
- 材料：８サブセットからなる高難度テキスト
  （普通文、押韻、係り受けが複雑な長文、複合語、句読点、繰り返し、言語切り替え、感情）
- 条件：ゼロショットTTS
- 評価方法：22日本語TTSを対象に日本語話者50名による聴取実験

今後の展望

- **サブセット別の詳細分析**
  - 特にCode-switching
- **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**

どちらもASJ2027春発表する予定！

- **DPO、GRPOなど強化学習手法によるTTS性能向上**