

# モデル基本情報

モデル	開発元	Github / 紹介記事リンク
reazonspeech-nemo-v2	Reazon Research (日本)	<a href="#">Hugging Face</a> <a href="#">紹介記事</a>
ElevenLabs Scribe	ElevenLabs (USA)	<a href="#">公式サイト</a> <a href="#">紹介記事</a>
Whisper (large-v3)	OpenAI (USA)	<a href="#">GitHub</a> <a href="#">紹介記事</a>

# 最新高精度ASRモデル比較

## 対象モデル

- reazonspeech-nemo-v2
- ElevenLabs Scribe
- Whisper (large-v3)

## 比較の視点

- 精度・性能（文字起こし・話者分離・リアルタイム性）  
※注: WER/CERは**低いほど良い**（エラー率が低い＝精度が高い）
- 技術的特徴（アーキテクチャ・学習データ）
- 利用環境（API・ライセンス・導入容易性）
- 実適用事例・ユーザ評価
- 最新研究動向と展望

# 文字起こし精度比較 (英語)

※注: 数値が**低いほど良い** (エラー率が低い = 精度が高い)

データセット	nemo-v2	Scribe	Whisper (large-v3)
LibriSpeech test-clean	—	—	2.7%
LibriSpeech test-other	—	—	5.2%
FLEURS 英語テスト	—	3.4%	4.7%
Common Voice 英語	—	6.7%	9.0%

## ポイント

- WhisperとScribeはトップレベルの高精度
- Scribeが実環境データでWhisperを僅かに上回る

# 文字起こし精度比較（日本語）

※注: 数値が**低いほど良い**（エラー率が低い＝精度が高い）

モデル	日本語認識性能
reazonspeech-nemo-v2	CER 7～9% (日本語特化で最高精度)
ElevenLabs Scribe	WER 5%以下推定 (多言語モデル中最高水準)
Whisper (large-v3)	日本語含む99言語対応も、やや劣る傾向

## ポイント

- 日本語では、reazonspeechが最も精度が高い
- 多言語モデルとしてScribeは優秀

# 話者分離（話者識別）精度比較

モデル	話者分離性能
reazonspeech-nemo-v2	単体では非対応（NVIDIA NeMoツールキット併用で対応可能）
ElevenLabs Scribe	<b>最大32人識別可能</b> 、非音声イベントタグ付
Whisper (large-v3)	単体非対応、外部diarization連携が必要

## ポイント

- Scribeが統合型で実用的かつ高評価
- Whisperやnemo-v2は柔軟だが追加実装が必要

# リアルタイム処理能力比較

モデル	リアルタイム性	必要環境
reazonspeech-nemo-v2	非常に高速（ストリーミング対応）	□ーカルGPU推奨
ElevenLabs Scribe	クラウドAPI提供（バッチ処理、ストリーミング未対応）	クラウド利用
Whisper (large-v3)	GPUあれば実時間程度（軽量版あり）	□ーカルGPUまたはAPI

## ポイント

- nemo-v2 はオンプレミスでの高速リアルタイム処理に最適
- Whisperは大規模なため環境に依存
- Scribeはバッチ処理専用（低遅延版開発中）

# 技術的特徴比較①

## reazonspeech-nemo-v2

- **アーキテクチャ:** Conformer (Encoder:Fast Conformer構造。畳み込みサブサンプリングによる8倍ダウンサンプリングやLongformer型の局所注意機構の組み合わせ) + RNN-T (Decoder:サブワード単位の出力 (SentencePieceトークン数約3000) )
- **高速化:** Fast Conformer、8倍ダウンサンプリング、長時間音声対応
- **学習:** 日本語TV音声35,000時間を使用

## ElevenLabs Scribe

- **アーキテクチャ:** 正式な技術仕様は非公開 (Transformerベース多言語モデル (99言語対応) が推測)
- **統合機能:** 話者分離・非音声イベント検出タグ付け
- **学習:** 独自大規模音声テキスト対を利用 (詳細非公開)

# 技術的特徴比較②

## Whisper (large-v3)

- **アーキテクチャ**: Encoder-Decoder Transformer (約15億パラメータ)
- **学習**: Web音声68万時間、Noisy Student方式
- **特長**: 多言語ASR + 翻訳、オープンソース・MITライセンス

## ポイント

- nemo-v2は最新の高速Transformer応用
- Scribeは多言語対応と統合機能が特徴
- Whisperは大規模データによる汎用性とオープン性



# 利用環境・実装面比較

モデル	提供方法	ライセンス	導入容易性
reazonspeech-nemo-v2	OSS (ローカル)	Apache 2.0	GPU必要、カスタマイズ自由
ElevenLabs Scribe	クラウドAPI	SaaS/API従量課金 \$0.40/音声1時間	インフラ不要、企業向け迅速導入
Whisper (large-v3)	OSS/API両方	MIT/API従量課金 \$0.36/ 音声1時間	OSSは自由度高、APIは手軽

## ポイント

- OSSは自由度が高いが環境整備が必要 (nemo-v2, Whisper)
- API型は即時導入可能で企業利用に最適 (Scribe, Whisper API)

# 実世界の適用事例・評価

- reazonspeech-nemo-v2
  - 国内企業（例：コールセンターシステム）で実証済み
  - SNSで日本語精度・高速性が高評価
- ElevenLabs Scribe
  - メディア業界や国際会議での採用事例あり
  - 多言語高精度、話者分離機能が好評
  - クラウド依存やバッチ処理の制限あり
- Whisper (large-v3)
  - 医療、報道、議事録作成など幅広く採用
  - オープンソースとしてデファクトスタンダード

# 最新研究動向と今後の展望

- **大規模事前学習:** Whisper以降、各社で精度向上が加速
- **軽量化技術:** Fast Conformer、Zipformer等で高速・軽量化が進展
- **ストリーミングASR:** リアルタイム字幕・翻訳向けの研究が活発化
- **統合型ASR:** 話者分離、音声イベント検出の統合モデルの研究
- **マルチモーダル連携:** 映像・テキスト連携を含む総合音声AIの進化

# 結論とモデル選択指針

利用シーン	推奨モデル
日本語特化で高精度・高速リアルタイム	reazonspeech-nemo-v2
多言語対応・話者識別・クラウド導入	ElevenLabs Scribe
自由度高いOSSで汎用ASR・翻訳統合	Whisper (large-v3)

- 各モデルは用途に応じたメリットを持つ
- 今後さらに精度・速度の向上が期待される