[サンプル版] 音声AI 投資判断レポート

対象：専門業務で使用される日本語音声

音声：3ファイル・合計23分13秒

比較：3つの音声認識モデル

---

1. 結論

今回は、モデルAを基準に改善を進めることを推奨します。

3つの音声認識モデルを同じ音声で比較した結果、**モデルAが文字起こし全体で最も高い精度**となりました。

モデル	全体CER
モデルA	**9.5%**
モデルB	20.7%
モデルC	25.9%

※CERは文字誤り率です。低いほど、正解データとの差が少ないことを示します。

今回の音声では、モデルB・Cへの変更を優先する根拠は得られませんでした。

一方で、モデルAにも業務上重要な言葉の取りこぼしが残っています。

そのため、現時点では大規模な追加学習には進まず、

モデルAを基準に、まずは学習を伴わない改善を小さく検証する

ことを推奨します。

今回の判断

モデル変更 → 現時点では優先しない

設定・語彙による改善 → 次に検証する

ファインチューニング → 現時点では判断を保留

本番導入 → 追加の実音声で評価してから判断する

---

2. なぜ「最も精度の高いモデル」を選ぶだけでは不十分なのか

全体の文字起こし精度と、業務上重要な言葉の認識精度は一致しませんでした。

文字起こし全体では、モデルAが大きく上回りました。

一方で、特定の重要項目だけを見ると結果が変わります。

例えば今回、ある専門用語の評価では次の結果になりました。

モデル	認識できた項目
モデルA	11 / 19
モデルB	**12 / 19**
モデルC	8 / 19

モデルBは、文字起こし全体のCERではモデルAを大きく下回っています。

それでも、この専門用語に限ればモデルBが1項目多く認識しました。

ただし、評価対象は19項目と少ないため、

「モデルBの方が専門用語に強い」と一般化できる結果ではありません。

重要なのは、

> 文字起こし全体がどれだけ正確か

と、

> 業務上、間違えると困る言葉をどれだけ正しく認識できるか

を**分けて評価すること**です。

---

3. 業務上重要な項目を個別に評価

今回は、業務上重要な項目を4つに分けて確認しました。

カテゴリ	モデルA	モデルB	モデルC
専門用語	58%	**63%**	42%
固有名詞	**86%**	77%	48%
数値	**91%**	87%	66%
否定・重要表現	**47%**	29%	12%
合計	**80%**	73%	49%

総合的にはモデルAが最も良い結果でした。

ただし、もう一つ重要な結果があります。

158項目のうち23項目は、3モデルすべてで正しく認識できませんでした。

つまり、

モデルAをBやCに変更するだけでは、解決しない可能性があります。

23項目の内訳は次のとおりです。

- 専門用語：5項目
- 固有名詞：6項目
- 数値：4項目
- 否定・重要表現：8項目

---

4. 何が原因なのか

この時点では、原因を断定しません。

3モデルすべてが同じ箇所で間違えたからといって、

「学習データが足りない」

「モデルの性能が悪い」

とは限りません。

考えられる要因には、

- 専門用語への対応
- 発音
- 録音環境
- 前後の文脈
- 音声の分割方法
- 音声認識の設定
- 評価方法

などがあります。

今回の診断だけで、原因を一つに特定することはできません。

そこで、**確認できた事実と、これから検証する仮説を分けて整理します。**

確認できたこと

23項目は、3モデルすべてで正しく認識できなかった。

考えられる原因

専門用語への対応、音声条件、文脈、認識設定など。

次に確認すること

条件を一つずつ変え、改善するか、逆に新たな誤りが増えないかを確認する。

---

5. 次に何を試すか

いきなりファインチューニングには進みません。

今回の結果から、次の順番で検証することを推奨します。

1. モデルAを基準にする

今回の音声では、文字起こし全体の精度、重要項目の総合結果ともにモデルAが最も良い結果でした。

2. 学習なしでできる改善を試す

まずは、

- 専門用語の候補提示
- 語彙指定
- 音声認識設定の調整

など、小さく試せる方法を検証します。

改善した項目だけを見るのではなく、

新たな誤りを増やしていないか

も確認します。

3. 未使用の音声で再評価する

今回使った音声で改善しても、本番環境で同じ効果が出るとは限りません。

異なる話者や録音環境の音声を使い、改善効果を再確認します。

4. それでも性能が不足する場合に追加学習を検討する

追加学習は、

- 学習による改善が見込める原因が残っている
- 未使用音声でも性能不足が確認された
- 費用に見合う業務効果が期待できる

という条件が揃った段階で検討します。

---

6. この診断で分かったこと / まだ分からないこと

分かったこと

今回の音声と評価条件では、

- モデルAの文字起こし精度が最も高い
- 重要項目の総合結果もモデルAが最も良い
- 一部の専門用語ではモデルBがモデルAを上回った
- 23項目は3モデルすべてで正しく認識できなかった
- 現時点ではモデル変更を優先する根拠はない
- まずは学習なしでできる改善を試すべき

ということが分かりました。

まだ分からないこと

今回の診断だけでは、

- 他の話者でもモデルAが最も良いか
- 本番環境全体で十分な性能が得られるか
- 設定変更でどこまで改善できるか

- ファインチューニングでどこまで改善できるか
- 追加投資に見合う業務効果が得られるか

までは判断できません。

今回は1話者・23分13秒の音声による評価のため、**他の話者や環境でも同じ結果になるとは限りません。**

---

7. 推奨する次のステップ

段階	実施すること	次へ進む条件
1. 評価基準を確定	評価対象・正式表記を確認	評価条件が確定
2. 小規模な改善を検証	語彙・設定などを個別に試す	改善と悪化を把握
3. 未使用音声で再評価	異なる話者・環境で確認	必要な性能基準を満たす
4. 実運用を検証	修正時間・確認負荷・費用を測定	業務で許容できる
5. 次の投資を判断	継続・変更・中止を比較	費用と効果の根拠が揃う

この順番で検証したうえで、

追加開発を進めるのか。

モデルを変更するのか。

追加学習するのか。

あるいは、追加投資をしないのか。

を判断することをお薦めいたします。

まとめ

今回の推奨

モデルAを基準に、検証を継続します。

まずは学習を伴わない改善を小さく試し、 その効果を未使用の実音声で確認します。

現時点では、ファインチューニングへの投資判断は保留します。

音声AI 投資判断レポート

実際の音声データを使って、

「どのモデルが最も良いか」だけではなく、

「次に何を試すべきか」まで整理します。

株式会社デジロウ