SOMO: 音声認識出力の可読性向上を目的とした整文手法の提案

杉野かおり [1]　山野陽祐 [1]　河崎真琴 [1]　田森秀明 [1]　岡崎直観 [2]　乾健太郎 [3,4,5]
[1] 株式会社朝日新聞社　[2] 東京科学大学　[3]MBZUAI　[4] 東北大学　[5] 理化学研究所
{sugino-k,yamano-y,kawasaki-m3,tamori-h}@asahi.com
okazaki@comp.isct.ac.jp　kentaro.inui@mbzuai.ac.ae

概要

本研究は, 音声認識出力を可読性の高い形式へと変換する「整文」手法を提案する. 日本速記協会の「発言記録作成標準」を参考に, 語断片の除去や文法, 流暢性の向上, 簡潔化までの 4 つのステップを定義した. これにより大規模言語モデル（LLM）を用いた整文処理が可能となり, 発言に忠実な書き起こしから, 要点のみをコンパクトにまとめた要約形式まで, 多様な用途に合わせた出力結果が得られる. これにより, 音声認識出力の利活用の可能性を広げることを目指す.

1　はじめに

昨今の音声認識技術は飛躍的に性能が向上し, 会議・講演・動画字幕作成・コールセンターの顧客対応記録など, さまざまな現場で活用されている. しかし, 音声認識出力には誤認識や表記揺れの発生の問題, 話し言葉特有の言い淀み, 冗長表現, 助詞誤りが含まれ, そのままでは可読性に難があることが多い.

一方, 音声認識出力を利活用する場面においては様々なレベルの処理が求められる. 例えば音声認識出力をインタビュー記事執筆に利用する場合は,「ケバ取り」と呼ばれる最低限の処理に留め, 話し手の口癖や言葉を選ぶ様子などのニュアンスを残すよう求められることが多いが, ビジネス会議の議事録や動画の字幕作成などの用途においては, 情報の正確性を保ちつつも繰り返しや冗長な表現を排除した簡潔さが重要である.

本研究では音声認識出力の可読性を高める手法 SOMO（Stepwise Optimization for Meaningful Output）を提案する. SOMO は日本速記協会[1)]の「発言記録作成標準」[1] が策定する, 速記後の可読性を高める「整文」手法に基づいている. この整文手法は, 必要最低限の処理から徐々に可読性を高めるステップ構造を取っており, ユーザーが求める多様な処理レベルに対応できる.

本研究の貢献を以下に示す.

- 整文という概念の音声認識出力への適用：速記の整文手法を, 音声認識出力に応用した.
- 整文の指針および指示文の策定：冗長性除去から文法性・流暢性向上, 簡潔化までのステップを指針化した.
- 多様な用途に対応した処理方法の提案：必要に応じて各ステップを選択・適用することで, 用途に合った処理レベルのテキスト出力を実現した.
- 評価指標の適用による有効性の検証：BERTScore や ROUGE を用いた定量的評価によって, SOMO の有効性を検証した.

2　提案手法 SOMO

SOMO の特徴は以下の 3 点である.

1. 整文処理を 4 ステップに分解し, 各ステップで達成すべき指針を明確にした.
2. 各ステップ, 各指針を LLM が解釈可能な具体的な指示としてプロンプト化した.
3. ステップ単位の処理により, 用途に応じた柔軟な整文レベルの制御を実現した.

以降, 各ステップの詳細と処理フローについて述べる.

2.1　整文処理ステップ

SOMO は以下の 4 ステップで定式化される. なお各ステップは, 定められた指針を達成するために必要な処理内容を LLM への指示として定義する.

1） https://sokki.or.jp/

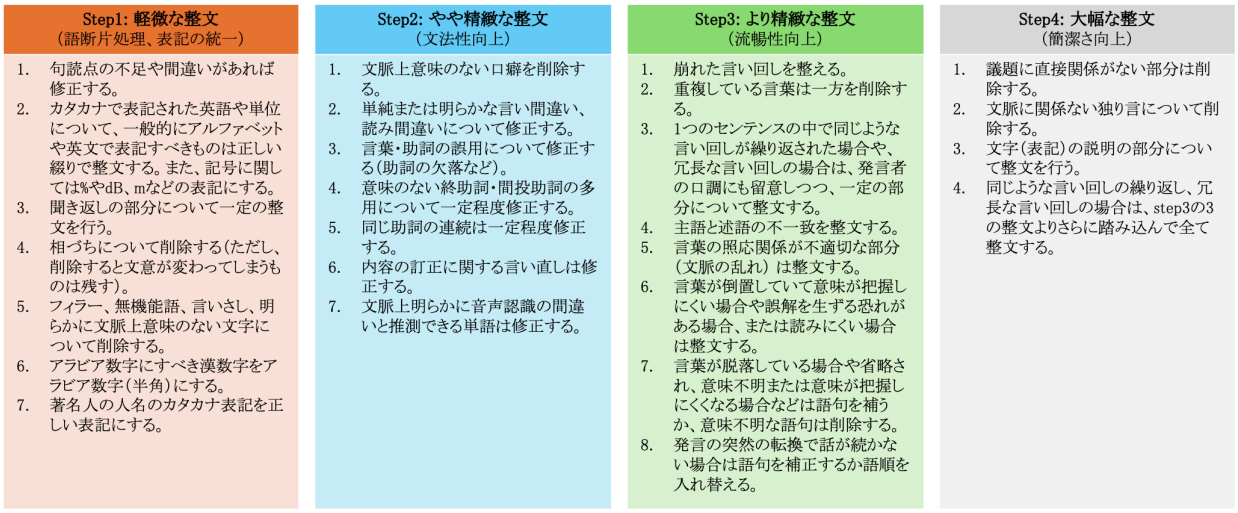

図 1　音声認識出力の可読性を高めるための指針と指示内容

Step1. 軽微な整文（語断片処理, 表記の統一） 句読点の付与, 無意味な語断片の除去, 不要なフィラーや相槌, 言い直しの除去, 表記揺れの統一等, 最低限のテキスト処理を行う. 例えばインタビュー記事を執筆する場合のような, 発言のニュアンスが大事な場合に有用である.

Step2. やや精緻な整文（文法性の向上） 発話特有の口語的な並べ方を整え, 主語・述語・目的語の対応関係を明確化する. 誤認識による文法的乱れを補正し, 書き言葉として成立する文に整える. 例えば議会の議事録のような, 可読性を高めつつも発言に忠実な記録をするときに有用である.

Step3. より精緻な整文（流暢性の向上） 前後文脈を考慮し, 接続詞や助詞の適切な挿入, 冗長な表現の短縮, 文構造の再編成を通じて, より自然な流れをもつテキストへと改善する. 例えばコールセンターでのやりとりの記録のような, 発言情報を担保しつつ, 高い可読性が求められる場合に有用である.

Step4. 大幅な整文（簡潔さの向上） 要点を明確化し, 不要な情報を整理・削除することで, 全体としてコンパクトで要点をつかみやすいテキストに仕上げる. 例えば動画字幕やビジネスシーンの議事録など, 正確で端的な記録が必要な場合に有用である.

2.2　プロンプト設計と処理方法

図 2 に, SOMO におけるプロンプトの設計方法と推論フローを示す. まず, 音声認識出力に対して適用する, 整文ステップ（Step1〜Step4）の指針に基づいた指示文を用意し, これらを連結してプロンプトとし, LLM に入力する（プロンプトの詳細は付録 A 参照）. LLM は与えられたプロンプトに従い, 各ステップに応じた整文処理を施し, 可読性の高いテキストを出力する.

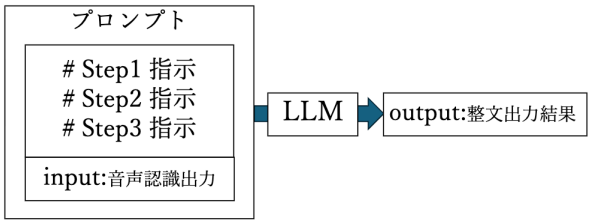

図 2　SOMO の処理方法（Step3 の場合）

このように SOMO の特徴は, LLM への指示を中心に据え, 整文ステップを柔軟に選択・適用できるフレームワークとして機能する点にある.

3　実験設定

3.1　データセット

本実験では, 自社収集の音声データを自動音声認識システムで得た出力を利用した. この出力に対してアノテーターが Step1~4 の各指示を適用して書き換え, 各ステップの結果を得た. よって各サンプルは, 音声認識出力と, Step1~4 それぞれの整文結果からなる. 音声認識出力の各サンプルの平均文字数は 500 文字, Step1 は 498 文字, Step2 は 490 文字, Step3 は 469 文字, Step4 は 453 文字だった. また品質保証の観点から, 評価用データは「発言記録作成標準」を定めた日本速記協会に作成を依頼した.

表 1　音声タイプの内訳 (%)

音声タイプ	訓練データ 1,800 件	評価データ 100 件
インタビュー	60.3	75.0
会議	18.5	10.4
記者会見	15.1	12.5
演説	4.6	2.1
その他	1.6	0.0

3.2　データの内訳

作成したデータセットのそれぞれのサンプルに対しては 5 つの音声タイプを人手によりアノテーションした. その内訳を表 1 に示す.

3.3　実装方法

実験では, gemma-2-2b-jpn-it[2] をベースモデルとして, 以下の 3 方式を比較した.

Baseline　付録 A のように, 一般的な話し言葉・書き言葉変換処理として考えうる簡素な指示文をプロンプトとし, LLM で処理する. ステップごとに細かく定義された指示による提案手法の有効性を確認するため, 各ステップでは指示を変えず, 単一の指示文で比較する.

SOMO　図 2 のように, 必要なステップまでの処理をプロンプト内に列挙し, LLM で推論する.

SOMO-FT　訓練データによるファインチューニングにより, 各ステップを処理するためのモデルをあらかじめ構築する. 推論は SOMO と同様の手順で行う. 実験における主要なパラメータ設定として, Epoch 数は 3, 学習率は 5e-4, バッチサイズは 4 とした. 推論時のデコーディング手法としてビームサーチを使用し, ビームサイズは 5 とした.

3.4　評価指標

BERTScore　正解と出力結果の意味的類似度を評価し, 各ステップで必要な情報が削除されていないか, あるいは不必要な情報が過度に付与されていないかを測定する.

ROUGE　正解と出力結果の表層的な語の一致度を評価する. 過度な書き換えが起きていないか, また語順がどれぐらい保たれているかを測定するため, bi-gram での評価とした（ROUGE-2）.

2)　https://huggingface.co/google/gemma-2-2b-jpn-it

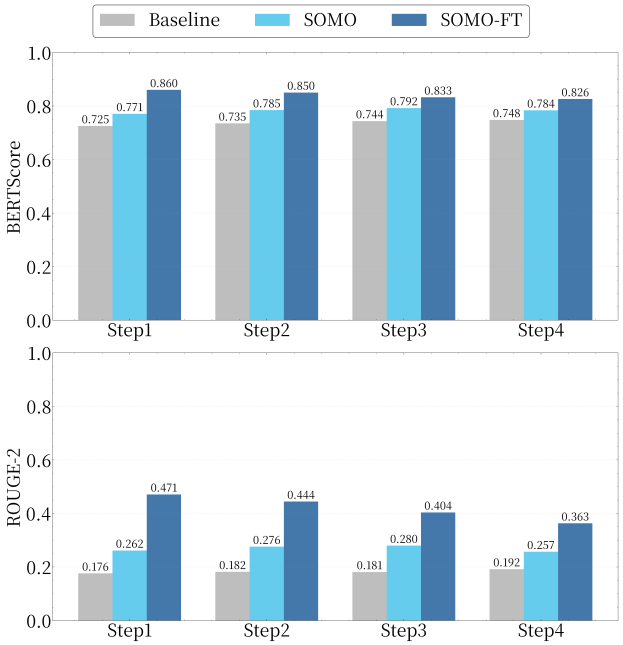

図 3　手法別の整文性能比較

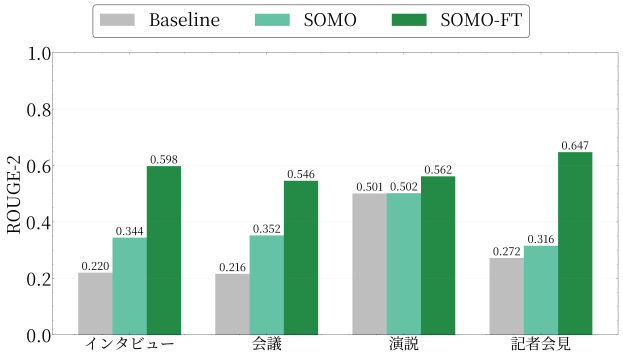

図 4　Step2 の音声タイプ別の整文性能比較

4　実験結果

4.1　SOMO の有効性

実験結果を図 3 に示す. Baseline に比べて SOMO は, 全ステップでスコアの向上を達成した. また, SOMO-FT においても全ステップで SOMO に比べてもスコアが向上した. 特に ROUGE-2 の Step1 では, 0.262 から 0.471 に上昇した. これらの結果から, 各ステップの指示が有効に機能しており, ファインチューニングによる効果も高いことが確認された.

4.2　音声タイプ別の評価

Step2 までを適用した場合の音声タイプ別の整文性能（ROUGE-2）を図 4 に示す. 4.1 節と同様, 音声タイプによらず, SOMO の有用性が確認できた. 一方, 演説は Baseline でも高いスコアを示している. こ

表 2　Step 別の整文出力結果のサンプル [a]

input	
音声認識出力	でもなんだろうやっぱりそのーえまあ資本国家資本主義って強かったと思うんですよね
Step1	軽微な整文（語断片処理, 表記の統一）
正解	でも、なんだろう、やっぱり、国家資本主義っていうのが強かったと思うんですよね。
Baseline	一方で、資本国家資本主義の強さから、
SOMO	一方で、資本国家資本主義が強かったと思います。
SOMO-FT	でもなんだろう、やっぱり資本国家、資本主義って、強かったと思うんですよね。
Step2	やや精緻な整文（文法性の向上）
正解	でも、国家資本主義っていうのが強かったと思うんですよね。
Baseline	一方で、資本国家資本主義の強さから、
SOMO	一方で、やはり資本国家資本主義が強かったと思います。
SOMO-FT	でも資本国家、資本主義って、強かったと思うんです。
Step3	より精緻な整文（流暢性の向上）
正解	けれども、国家資本主義が強かったと思うんです。
Baseline	一方で、資本国家資本主義の強さから、
SOMO	けれども、一方で、やはりその部分で資本国家資本主義が強かったと思います。
SOMO-FT	が、資本国家、資本主義が強かったと思います。
Step4	大幅な整文（簡潔さの向上）
正解	けれども、国家資本主義が強かったと思うんです。
Baseline	一方で、資本国家資本主義の強さから、
SOMO	けれども、一方で、やはり資本国家資本主義が強かったと思います。
SOMO-FT	が、資本国家資本主義は強かったと思います。

[a] サンプルは各ステップの推論結果の一部を抜粋.

れは, 演説の発話の文構造が明解で, 話し言葉特有の表現が少ないため, 具体的な指示がなくても整文が容易である可能性を示唆する. インタビューや会議, 記者会見に関しては, 発話の形式や複数人での発話が演説に比べて多いことが考えられ, SOMO-FT はファインチューニングによってその特有の表現を学習し, スコアが大幅に向上したと推察される.

4.3　出力結果の観察

整文出力結果のサンプルを表 2 に示す. 各手法による推論結果を観察したところ, 主に以下のような特徴があった. まず, 正解と各出力結果を比較すると, いずれにおいても文意は大きく変わらないことが分かった. また, Baseline の出力結果は, 全ステップのうち Step4 の正解に近く, 文を簡潔にする傾向が強いことが見てとれる. これは図 3 で示した, Step4 における Baseline の ROUGE-2 が他のステップに比べて高くなる実験結果とも一致する. Step3 の SOMO では, 音声認識出力にはない「その部分で」という言葉を補完することで, 文の流暢性を向上させている. SOMO-FT の, 特に Step1~2 は, 指示された整文内容に厳密に従いつつ, 音声認識出力のニュアンスを保持しており, 結果として正解データに最も近い仕上がりを示している. これは, ファインチューニングにより, 指示範囲を超えた処理をせず, 音声認識出力に忠実な整文を学習できていると考えられる.

5　関連研究

音声認識後の処理としては, フィラー除去や句読点付与, 逆テキスト正規化のみを行う限定的なアプローチ [2], あるいは音声データから可読性の高い書き言葉を直接生成する End-to-End モデルを用いた手法 [3, 4] が提案されてきた. また, 庵ら [5] は日本語の話し言葉から書き言葉変換コーパスを構築し, 多岐にわたる変換要素を同時に考慮する試みを行っている. しかし, これらはいずれも単一形態の出力を目指すことが多く, 用途に応じて可読性・冗長性の度合いを可変的に調整するには不十分であった.

本研究は, どの順番でどの指示までを適用すべきかをステップごとに明確に定義し, 処理レベルの異なる出力を可能にした. これにより, 発言に忠実な記録から, 簡潔なテキストに至るまで, 同一の音声認識出力を多目的に再利用可能となる. このような仕組みは従来の関連研究には見られず, 本研究の独自性を示す重要な要素である.

6　おわりに

本研究は, 日本速記協会の整文手法にヒントを得た 4 つのステップを定義し, 音声認識出力を用途に応じた可読性の高いテキストに整文する SOMO を提案した. さらに, BERTScore や ROUGE を用いた定量評価により, その有効性を示した. 今後は, LLM などを活用した評価方法の検討, より高度な言語モデルの活用, リアルタイム処理能力の向上, 多言語への適用可能性などの拡張を検討する予定である. 各ステップの指示の順序が整文品質に与える影響についても調査したい.

また, 今回は整文というタスクの評価において, BERTScore や ROUGE といった既存の指標を使用したが, 整文に適した評価手法については検討できていない. 例えば, LLM を使用してより人手評価に近い評価が得られるような方法等も今後検討していきたい.

謝辞

本研究は日本速記協会の知見から着想を得たものであり, 研究への理解と音声書き起こしタスクへの適用における助言, 評価データセットの作成などに協力いただきました.

また, 訓練データセットの作成においては株式会社はちのへ東奥朝日ソリューションに, アノテーション作業への助言は三菱 UFJ リサーチ&コンサルティング株式会社に協力いただきました.

参考文献

[1] 日本速記協会. 発言記録作成標準　第 8 版. 公益社団法人 日本速記協会, 2021.

[2] Thai-Binh Nguyen, Le Duc Minh Nhat, Quang Minh Nguyen, Quoc Truong Do, Chi Mai Luong, and Alexander Waibel. Adapitn: A fast, reliable, and dynamic adaptive inverse text normalization. In **ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 1–5, 2023.

[3] 三村正人, 河原達也. 国会会議録のための音声から書き言葉への end-to-end 変換. 自然言語処理, Vol. 30, No. 1, pp. 88–124, 2023.

[4] Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In **Proceedings of the Ninth Conference on Machine Translation**, pp. 1301–1317, 2024.

[5] 庵愛, 高島瑛彦, 増村亮. 日本語文章のための話し言葉・書き言葉変換コーパス. 言語処理学会 第 26 回年次大会, 2020.

A　プロンプト

Baseline

あなたは優れた日本語のライターです。input された音声認識テキストを、可読性の高い文に修正してください。解説は含めず、書き換え結果のみを output としてください。

input: はいあの朝日新聞社ですえー本日は百人の方に手伝ってもらもらって整文の方の作業を進めてもらいますなるほど２時間は予定のをしていますね作業時間の方はてら画面見えてないよはい題材はあのーこちらです。
output:

Step1

あなたは優れた日本語のライターです。input された音声認識テキストを、以下の step の手順に従って可読性の高い文に修正してください。解説は含めず、書き換え結果のみを output としてください。

Step1
軽微な整文（語断片処理、表記の統一）
1. 句読点の不足や間違いがあれば修正する。
2. アラビア数字にすべき漢数字をアラビア数字（半角）にする。
3. 著名人の人名のカタカナ表記を正しい表記にする。
4. 無機能語、言いさし、明らかに文脈上意味のない文字について削除する。
5. 相づちについて一定の整文を行う。6. 聞き返しの部分について一定の整文を行う。
7. カタカナで表記された英語や単位について、一般的にアルファベットや英文で表記すべきものは正しい綴りで整文する。また、記号に関しては%や dB,m などの表記にする。
・「」は使わない。
・input の末尾が文の途中で切れていても、その末尾は修正せずそのままにする。

input:(同上)
output:

Step2

あなたは優れた日本語のライターです。input された音声認識テキストを、以下の step の手順に従って可読性の高い文に修正してください。解説は含めず、書き換え結果のみを output としてください。

Step1
(同上)
Step2
やや精緻な整文（文法性向上）
1. 単純または明らかな言い間違い、読み間違いについて整文する。
2. 言葉・助詞の誤用について整文する（助詞の欠落など）。
3. 意味のない終助詞・間投助詞の多用について一定程度整文する。
4. 同じ助詞の連続は一定程度整文する。
5. 内容の訂正に関する言い直しは整文する。
6. 文脈上、意味のない口癖を削除する。
7. 文脈上明らかに間違いと推測できる、一般的な名詞などの間違いは整文する。
・input の末尾が文の途中で切れていても、その末尾は修正せずそのままにする。

input:(同上)
output:

Step3

あなたは優れた日本語のライターです。input された音声認識テキストを、以下の step の手順に従って可読性の高い文に修正してください。解説は含めず、書き換え結果のみを output としてください。

Step1
(同上)
Step2
(同上)
Step3
より精緻な整文（流暢性向上）
1. 主語と述語の不一致を整文する。
2. 言葉の照応関係が不適切な部分 (文脈の乱れ) は整文する。
3. 言葉が倒置していて意味が把握しにくい場合や誤解を生ずる恐れがある場合、または読みにくい場合は整文する。
4. 言葉が脱落している場合や省略され、意味不明または意味が把握しにくくなる場合などは語句を補うか、意味不明な語句は削除する。
5. 崩れた言い回しを整える。
6. 重複している言葉は一方を削除する。
7. 発言の突然の転換で話が続かない場合は、語句を補正するか、語順を入れ替える。
・input の末尾が文の途中で切れていても、その末尾は修正せずそのままにする。

input:(同上)
output:

Step4

あなたは優れた日本語のライターです。input された音声認識テキストを、以下の step の手順に従って可読性の高い文に修正してください。解説は含めず、書き換え結果のみを output としてください。

Step1
(同上)
Step2
(同上)
Step3
(同上)
Step4
大幅な整文（簡潔さ向上）
1. 議題に直接関係がない部分は削除する。
2. 文脈に関係ない独り言について削除する。
3. 文字（表記）の説明の部分について整文を行う。
4. 同じような言い回しの繰り返し、冗長な言い回しの場合は、step3 6 の整文よりさらに踏み込んで全て整文する。
・input の末尾が文の途中で切れていても、その末尾は修正せずそのままにする。

input:(同上)
output: