大規模音声基盤モデルを用いた長期録音資源のクリーン化と音声検索精度の向上

静岡大学　工学部数理システム工学科
甲斐充彦

１．　はじめに

最近ではコンピュータで音声言語情報を扱う自動音声認識や音声検索の技術が進歩し、世界の主要な言語に対して精度が改善され実用化が進んでいる。一方で、日本国内の様々な地域で収録された雑談や地方議会の会議録音、古い時代に収録された歴史的な証言音声のような収録音声を対象とする場合、自動音声認識や音声検索の精度はまだ極めて低い。そのため、過去から引き継がれる長期的かつ大規模な録音データから発言内容を検索するような用途での音声言語処理技術の活用はまだ困難な状況である。

そのような現状の理由の一つとして、音声言語技術の核となるAI技術で知識獲得のために必要とされる学習用データが、特に地域特有の諸方言に対しては圧倒的に不足していることが挙げられる。一般に音声言語処理のAI技術では知識獲得のために音声とその書き起こしのデータが大量に必要となるが、古くから残された録音には書き起こしが付与されたものは極めて少なく、また録音品質にばらつきがあることが多い。そこで、書き起こしがまだ少ない諸方言や品質が多様な録音資源にもこのようなAI技術の適用範囲を広げて、自動音声認識や検索の学習データとして活用できるように工夫することが必要となる。

このような問題に対処するため、長期にわたって収録された全国各地の方言音声データを活用し、AI技術の活用によって録音品質や方言の違いに左右されにくい音声基盤モデルの構築と、音声認識および検索システムへの応用を試みた。そして、最先端の大規模音声基盤モデルからこれらの目的にチューニングした諸方言音声認識モデルを用いた音声認識と検索の評価実験によってその効果を検証した。

２．　長期的で大規模な録音資源活用のための課題と解決のアプローチ

これまで構築されてきた音声言語処理のAI技術のほとんどは標準語を対象としたものである。一般的なAI技術は大規模な学習用データで学習された基盤モデルを用いており、音声言語処理用の基盤モデルの構築では音声とその書き起こしの対応がとれたデータを大規模に必要とするためである。そのため、書き起こしがある音声データが十分な量で蓄積されていない言語（低資源言語）や方言に対しては、音声データのみまたは他言語の知識を活用してどのように効率よく音声言語の基盤モデルを構築するかがこの分野で重要な研究課題となっている。

このような問題に対して近年有望とされるアプローチの一つとして、大規模な事前学習済み音声基盤モデルの活用がある。自己教師あり学習(SSL)は、大量のラベルなしデータから潜在的な表現を学習する技術であり、 自動音声認識をはじめ主要な音声処理アプリケーションに応用したときに少量学習データでチューニングするだけで目的に特化して大規模学習した

モデルと同等の性能を発揮できるほどの有効性が示されている[1] [2]。事前に訓練されるSSL モデルとしては様々な環境や言語で録音された大量のラベルなしデータセットを活用することができるため、前述のような課題に対して頑健性が高いモデルを実現できる可能性がある。実際に、我々の先行研究[3]では大規模多言語 SSL モデルを用いて方言識別と ASR のマルチタスク学習を行うことで ASR の性能を大幅に向上させている。

一方で、SSL ベースのモデルとは対照的に、ASR、機械翻訳、言語識別などの複数のタスクに対して低品質なラベルを含む教師ありデータを極めて大規模かつ同時に利用した End-to-end 型モデルの学習法についても有効性が示されており、世界の主要言語に対してトップクラスの性能を示すことが知られている。この方法で実現された多言語対応の大規模音声言語モデル Whisper[4]では、日本語に対しても高い精度を示すことが示されている。しかし、地域によって大きく異なる多様な日本語方言への適応性は明らかにされていない。

本研究では、現時点で利用できる国内の諸方言音声として最大規模の諸方言音声コーパス（COJADS）[7]を用いる。このコーパスは国立国語研究所が 2019 年から公開を始めたものであるが、収録された音声は文化庁が 1977 年から 1985 年に行った各地方言収集緊急調査で収録された音声など長期的に収録された音声を集めたもので、国内では貴重な音声言語資料の一つとなっている。一方で、このコーパスは音声・言語学的な分析の観点から整理されたものであるため、音声と書き起こしが細かな単位で時間的に対応付けされたものではない。そのため、音声基盤モデルの学習用データとして直ちに利用できるものではなく応用例はなかった。また、今日の音声基盤モデルの構築で一般的に利用される学習用音声のデータ量が一言語当たり数百時間を超える程度であるのに対して、このコーパスは 2021 年公開版の時点で全国地域あわせて約 60 時間の規模しか含まない。更に、地域によって言語的特徴に大きな違いがあるため、限られた量のデータを活用するために地域の違いを考慮した学習方法を工夫する必要があった。これらの背景から、本研究では以下に挙げる複数の問題解決に向けて取り組んだ。

第一に、録音品質や条件が異なる全国地域の長期録音データを活用するうえで、地域による録音品質の違いの影響を調査し、録音品質を改善する方法として提案されている音声復元モデルを応用して改善を図った。第二に、様々な日本語の方言に対応する統一的な ASR モデルを構築する効果的な方法に焦点を当てた。地域ごとに言語的な特徴が大きく異なる日本語諸方言の特質と少資源言語としての側面に対する既存の有効なアプローチを考慮し、2 種類の大規模な多言語事前学習モデルを活用した ASR モデル構築を試みた。一つは SSL 大規模事前学習モデルから構築する ASR モデル[3]で、もう一つは OpenAI 社が公開する Whisper という多言語大規模 ASR モデルであり、これらを COJADS コーパスで適応学習する。さらに ASR モデルからのテキスト出力、または SSL モデルからの音声特徴を利用した音声検索語検出(QbE-STD)システムを構築し、方言の検索タスクへの提案モデルの適用可能性を示す。我々はこの COJADS コーパスを活用した例としては初めて、のべ 60 時間分程度のデータを用いて諸方言をカバーする自動音声認識モデルの構築を開始している[3]。そのような包括的なモデルを用

いた音声検索の応用例はないため、まずは他の最先端の多言語音声認識モデルを比較対象として有効なアプローチと現状の性能を確認した成果となる。

３．　音声復元モデルを用いた全国諸方言の長期収録音声のクリーン化

前述のように本研究で用いる諸方言音声コーパス（COJADS）は、地域ごとにし収録された年代が異なり、録音環境は統一されていない。このコーパスには、全国の代表的な地域で複数の談話が収録されており、談話別に収録された地域ラベルが記録されている。初めに本研究では、この地域ラベルをもとに全国の収録音声データを 8 地域に分類し、音声による地域識別モデルを学習することによって、地域による音声特徴の自動分類性能を実験的に確かめた。本研究で利用した諸方言音声データの内訳を表 1 に示す。

表 1　利用した日本語諸方言音声データの内訳

	発話数	のべ時間数（時間）
方言識別モデル学習用	54403	27.7
音声認識（ASR）モデル学習用	69860	29.8
評価用データ	1962	0.83

実験では、我々の先行研究[3]の例に基づいて、自己教師あり学習（SSL）により大規模多言語音声で事前学習された音声基盤モデル（XLSR）をベースとして方言識別モデルを構築した。その結果、8 地域識別モデルによる発話単位での地域識別の正解率は 79.8%となった。音声信号の入力のみで地域識別性能が非常に高い理由として、地域による収録環境の違いの影響が予想されたため、発話がされていない録音部分（非発話部分）のみをランダムに抽出して同様の地域識別モデルを学習し、非発話部分のみでの評価を実施した。その結果、地域識別率は 45.9%となり、ランダムに識別した場合（チャンスレート）の 12.5%（1/8）を大きく上回ることが分かり、収録環境の違いの影響が大きいことが明らかとなった。

そこで、関連研究で近年提案されていた音声復元モデル[8]を適用することで更にこの問題の改善の効果を検証した。このモデルは、劣化した音声とその書き起こしを入力として明瞭性が高い音声を生成する生成モデルとして大規模な音声データと書き起こしから学習されている。本研究で用いる諸方言音声データは書き起こしを含むため、これを利用して収録音声データのクリーン化を行うことで前述の地域識別モデルの学習の前処理の改善を図った。その結果、このクリーン化を行った音声に対して非発話部分に対する識別率は 21.3%となり、長期収録音声データで録音品質が不均一であったことの影響は大きく改善された。一方、地域識別モデルの学習と評価では、クリーン化前の識別率 79.8%に対して、クリーン化後は 37.9%になった。これは、クリーン化後の条件では本来注目すべき音声言語的な地域特徴と関係ない部分の情報が削減された効果を含むが、音声復元モデルによる音声言語情報の劣化の影響の可能性も予想された。そこで、関連して ASR モデルの構築と評価の実験を行ったと

ころ、クリーン化前の音声データに対して文字誤り率が 35.7%であったのに対して、クリーン化後は 41.9%に悪化した。これは、音声復元モデルが標準語の音声データを中心に学習されていることから、学習に含まれていない方言特有の言語特徴（書き起こし）やアクセント特徴を含む音声で品質低下が生じたことが大きな要因の一つと考えており、今後は音声復元モデルの改善が必要と考えている。

４．　諸方言音声コーパスを活用した ASR 用音声基盤モデル構築

現在の最先端の自動音声認識技術を支える要素として、音声認識モデルを学習するためのデータが重要となる。通常、数百時間から数千時間を超える規模の音声録音とその書き起こしとの対応が取れたデータが必要となる。しかし、諸方言の話し言葉に対してその正確な文字化は一般的でなく、諸方言をカバーする書き起こしとして利用できるデータは限定的である。そのため、前述の諸方言音声コーパス（COJADS）を活用する。このコーパスは、全国の自然な方言音声を含む、実環境で録音された談話形式の音声から構成され、書き起こしはカタカナのみである。モデルの適応学習には前述の表 1 のデータを利用し、学習と評価用データの間では話者の重複はない。

諸方言の ASR モデルの構築のため、図 1 に示す 2 種類の大規模な事前学習の音声基盤モデルを適応学習するアプローチを採用する。以下の節では、比較する大規模事前学習モデルの適応学習のアプローチ（図 2）について説明する。

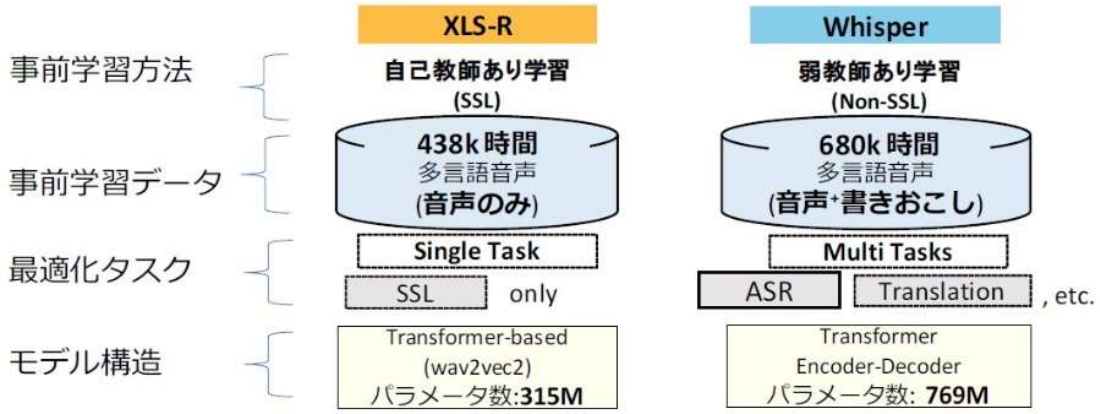

図 1　比較する大規模事前学習の音声基盤モデルのおもな仕様の違い

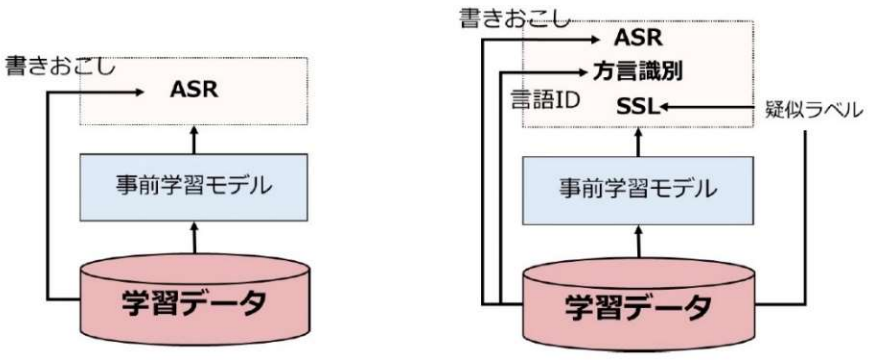

図 2　比較する 2 つの音声認識（ASR）モデル構築のアプローチ

４．１． SSL 大規模事前学習モデルを用いた諸方言適応 ASR モデル構築

XLSR[5]は、SSL モデル wav2vec 2.0[1]をベースとしたモデルであり、53 言語にわたる 56,000 時間のラベルなし音声データを使用して事前学習されている。ASR タスクのためのモデルを構築するために、最も単純なアプローチは、ASR ラベルを持つ少量のターゲットドメインデータのみを使用して事前訓練されたモデルを追加学習する方法である（図 2 左)。先行研究[3]では、マルチタスク学習による多段階適応（図 2 右）を適用することで、日本語方言の ASR 性能を向上させる方法を提案した（以下、3 段階学習と呼ぶ)。この方法は、SSL 損失、書き起こしとのアライメントを評価するための CTC 損失、および方言音声が発話された地域を特定する方言識別（DID）タスクのためのクロスエントロピー損失を最小化する。本手法は、標準的な日本語データと方言データの両方を学習に利用ことで、学習データ不足の問題を解決する。本研究では上述の SSL モデル XLSR の他、事前学習の音声データ量が大幅に増えた SSL モデルの XLS-R [6]（図 1 左）を用いた諸方言適応モデルの構築による実験も行った。 XLSR では約 2 時間、XLS-R では約 49 時間の日本語音声を含むため、諸方言音声への適応学習ではそれらの違いの影響を確認することができる。

４．２．多言語大規模 ASR モデル Whisper の諸方言適応 ASR モデル構築

Whisper は、弱教師付き学習[4]によって事前に訓練されたオープンソースの ASR モデルである。弱教師付き学習とは、機械的にアノテーションされた低品質のラベルを含む、大量のラベル付きデータを使ってモデルを学習する手法である。Whisper は、ASR だけでなく、翻訳、音声区間検出、アライメント、言語識別などのマルチタスク学習のために、ウェブから収集された 68 万時間の多言語音声を使用して訓練されている。Whisper には、パラメータ数の異なる 5 つのモデル（tiny、base、small、medium、large）がある。本研究では、Whisper-medium モデルを用いて、方言データのみを用いて諸方言 ASR モデルを構築する。このモデルは言語ラベルが固定されているため、前述の図 2 左の ASR タスク単独の適応学習のみを採用する。

４．３．自動音声認識としての評価実験

実験結果を表 2 に示す。表では、文字誤り率（CER）を用いて、標準日本語（ここでは CSJ コーパス）と方言（ここでは COJADS）に対する各モデルの ASR 精度を比較している。標準日本語の ASR では、SSL モデルの CER は 6%程度であるが、Whisper モデルの CER は 4.1%と最も高い精度を達成している。この結果は、事前学習データに含まれる日本語の量が多いためである。SSL モデルの事前学習では、XLSR が 2 時間、XLS-R が 49 時間の日本語を含むのに対し、Whisper は 7,054 時間のラベル付き日本語音声データを ASR タスクの事前学習に使用している。方言音声の場合、Whisper モデルは 32.9%の CER を達成し、これは 1 段階での学習（Full fine-tuning）を行った XLS-R の結果と同等である。このことは、適応データが少ない場合、1 段階学習だけでは方言音声の ASR 性能の向上には不十分であることを示唆している。SSL モデルでは、地域の方言の違いを考慮した 3 段階学習（3-step fine-tuning）を用いた ASR モデルで、約 3～4%の改善が見られ Whisper モデルよりも高い精度が得られている。

表 2 標準日本語適応モデル（CSJ 用）と諸方言適応モデル（COJADS 用）の ASR 精度の比較

Pre-Trained ASR/SSL Model	ASR Adaptation	CER[%]	
		CSJ	*COJADS*
Target speech: Standard Japanese (SJ)			
Whisper-medium	Full fine-tuning	**4.1**	49.2
XLSR	Full fine-tuning	6.5	52.7
XLS-R	Full fine-tuning	6.1	51.3
Target speech: Japanese dialects (JD)			
Whisper-medium	Full fine-tuning	-	32.9
XLSR	Full fine-tuning	-	34.1
	3-step fine-tuning	-	30.0
XLS-R	Full fine-tuning	-	32.6
	3-step fine-tuning	-	**29.2**

４．４．諸方言収録音声に対する音声検索としての評価実験

ASR 技術の有用な応用例として、音声検索（Spoken Term Detection: STD）タスクがある。比較した STD 手法の概要を図 3 に示す。第一の方法は、ASR システムを用いてターゲット音声とクエリの両方をテキストに変換し、テキスト間の動的タイムワープ（DTW）を通じてスポッティングを行うという、STD を実装する最も単純な方法を採用している。第二の方法は、テキストに変換する代わりに、ターゲットとクエリの音声特徴を直接検索する方法である。後者の方法は、リソースの少ない言語によく使われる。日本語方言については、ASR モデルの学習に用いなかった COJADS の評価データ（1962 発話）を検索対象音声とクエリ音声として用いた。クエリ音声については、方言特有の表現の抽出を試みた。COJADS の書き起こしから、統計的に方言特有の言い回しとして 12 のクエリを自動選択した。前述の第一の方法で用いる ASR モデルは第 3 節で説明したものと同じである。特徴ベース法では、SSL モデルは事前に訓練された XLSR を使用した。評価指標は検索評価で用いられる MTWV で、大きな値ほど精度が高いことを表す。

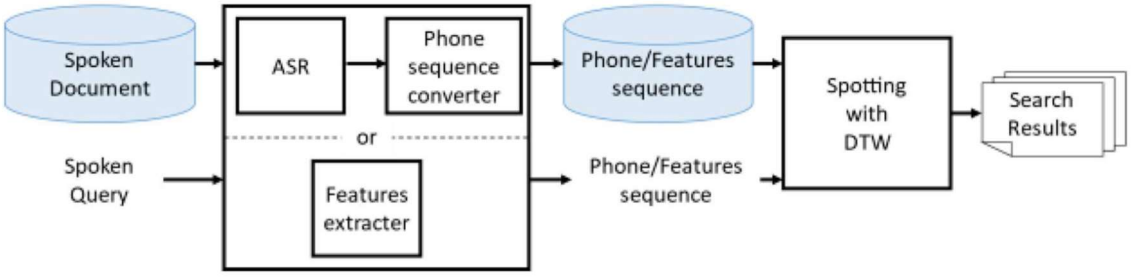

図 3 音声検索システムの構成の概要（ASR ベースまたは特徴量ベースの２つの方法）

表 3　標準語および諸方言に対する音声検索アプローチの違いによる性能比較

前処理	特徴量/モデル	ASR モデル適応	検索精度（MTWV）	
			標準語（CSJ）	諸方言（COJADS）
音声特徴量	フィルタバンク	——	0.128	0.089
	XLSR-R（SSL）	——	0.443	0.269
ASR モデル	Whisper-medium	単一言語適応	0.766	0.553
	XLSR-R（SSL）	3 段階学習	——	0.630

様々なモデルを用いて STD システムの音声検索性能を評価した結果を表 3 に示す。標準日本語 ASR モデルには Whisper を使用し、単一方言 ASR モデルには XLS-R（3 段階 finetuning）と Whisper を使用した。その結果、ASR モデルに入力する音声特徴表現としてよく知られているフィルタバンクを用いた検索の精度は、標準日本語、方言ともに最も低く、SSL を用いた検索がより効果的であることが示された。SSL を用いた検索では、標準日本語、方言ともに ASR を用いた検索が最も精度が高かった。Whisper ベースの標準日本語 ASR システムでは、CER は文書で 10.4%、クエリで 12.6%であったのに対し、XLSR ベースの諸方言 ASR システムでは、CER は文書で 29.3%、クエリで 15.1%であった。文書の CER に有意な差があるにもかかわらず、同様の検索精度が達成された。この結果から、XLS-R ベースの諸方言 ASR システムは、少なくとも既知の単語については、標準的な日本語検索に近い精度を達成できる可能性があることが示唆された。さらに、Whisper ベースの諸方言 ASR システムの CER は文書で 33.9%、クエリで 22.6%であり、XLS-R と比較して検索精度が低かった。

５．まとめ

本研究では、長期収録された音声資料を活用するための音声言語処理技術の実現のため、大規模事前学習された音声基盤モデルを活用する方法を提案し、その効果を確認した。音声品質が不均一な状況に対して音声復元モデルを適用することの効果を確認し、課題を明らかにした。また、大規模事前学習された最先端の音声基盤モデルから日本語諸方言に対応する自動音声認識システムを構築し、その性能を比較した。その結果、大規模多言語音声で事前学習された音声基盤モデルから方言識別タスクと同時に適応学習する方法は、諸方言の音声認識性能を向上させる効果的なアプローチであることが示された。さらに、音声検索実験により、既知の単語クエリに対しては、現在の ASR 精度で標準日本語に近い検索性能を達成できることが確認された。今後の課題としては、Whisper ASR モデルへの地域ラベルを考慮した学習法の導入や、方言クエリを増やした場合の検索性能の調査などが挙げられる。

謝辞

本研究を遂行するにあたり、公益財団法人天野工業技術研究所から多大なご支援を頂きました。また、本研究の実験全般において研究室修士学生の遠山拓実氏の協力を得ました。ここに記して謝意を示します。

参考文献

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems, Vol. 33, pp.12449-12460, 2020.

[2] W.-N. Hsu, B. Bolte, Y.-H. Hubert Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 29, pp. 3451–3460, 2021.

[3] S. Miwa and A. Kai. Dialect Speech Recognition Modeling using Corpus of Japanese Dialects and Self-Supervised Learning-based Model XLSR. In Proc. INTERSPEECH, pp. 4928–4932, 2023.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning, pp. 28492–28518. PMLR, 2023.

[5] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In Proc. INTERSPEECH, pp. 2426–2430, 2021.

[6] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In Proc. INTERSPEECH, pp. 2278–2282, 2022.

[7] COJADS. https://www2.ninjal.ac.jp/cojads/index.html. (Accessed on 2024/08/24)

[8] Y. Koizumi et al., Miipher: A Robust Speech Restoration Model Integrating Self-Supervised Speech and Text Representations, arXiv:2303.01664, 2023.