A Study on Automatic Alignment of Utterance Transcription for Japanese Dialect Speech

方言音声に対するテキスト自動アライメントの試み

石本 祐一 (国立国語研究所コーパス開発センター) *

A Study on Automatic Alignment of Utterance Transcription for Japanese Dialect Speech

Yuichi Ishimoto (National Institute for Japanese Language and Linguistics)

要旨

音声コーパスの構築にあたり、音声に対する発話・音素・韻律などの各種ラベル付与が作業者の大きな負担となっている。この負担軽減を目的としてラベリングを自動化する試みが行われており、音声認識技術を利用した転記テキストの自動アライメントシステムがすでにコーパス構築の補助として稼働し始めている。しかし、システムの音声認識部を構成する音響モデル・言語モデルが標準語を基に設計されていることから、現在のところは標準語を主とした音声へのシステム利用にとどまっており、標準語とは異なる特性を持ちうる方言音声に対してはシステムの有効性が不明である。そこで本稿では、方言音声に対する転記テキストの自動アライメント性能について調べた結果について報告し、方言音声コーパスの構築におけるテキスト自動アライメントシステムの実用可能性について述べる。

1. はじめに

音声コーパスには様々なラベルが付与されていることが望ましいが、ラベル付与は音声・言語分野の知識を持った作業者による人手作業によるところが大きく、その負荷の高さがコーパス構築における問題となっている。そこで、国立国語研究所コーパス開発センターでは、コーパス構築の負担を軽減するべく、コンピュータによるラベル付与の自動処理の検討を進めている。その中で発話を文字で書き起こしたテキスト（以下、転記テキスト）の音声データへの配置（アライメント）については、人手での修正作業という後処理がある程度求められるものの、ほぼ実用に足ることがわかってきた。これは、音声認識による書き起こしをタイムスタンプ付きで出力する自動字幕作成システム (秋田ほか 2015, 河原ほか 2016) を応用するものであり、音声に加えて転記テキストも入力に用いることで実用的な精度で転記テキストに対応する発話の開始時刻を推定することができる (石本 2017)。現在では実際に、『日本語日常会話コーパス (CEJC)』(小磯ほか 2017) の構築にこのテキスト自動アライメント手法が活用されている。

しかしながら、この手法があらゆる音声コーパス構築においても有効であるかはまだ明確でない。特に、音声認識システムに用いられている音響モデル・言語モデルと異なる性質の音声に対しては音声認識がうまく働かず、アライメントの精度が低下する可能性がある。例えば、上記の自動字幕作成システムは主に『日本語話し言葉コーパス』(国立国語研究所 2006) を基

* yishi@ninjal.ac.jp

表 1　対象データ

ID	地域	話者数	長さ (秒)	発話数	ID	地域	話者数	長さ (秒)	発話数
1	北海道	3	72.667	35	29	奈良	2	105.222	65
3	岩手	2	62.145	33	30	和歌山	3	82.648	59
4	宮城	3	77.027	54	31	鳥取	2	75.704	51
5	秋田	3	62.042	35	32	島根	3	95.392	53
6	山形	4	72.030	40	33	岡山	2	60.893	44
7	福島	3	68.520	23	36	徳島	3	62.405	32
8	茨城	3	60.565	32	37	香川	2	110.048	66
9	栃木	3	62.566	36	38	愛媛	2	60.692	42
12	千葉	3	62.987	54	39	高知	3	72.211	42
15	新潟	4	61.023	36	41	佐賀	3	60.082	37
16	富山	3	63.329	60	42	長崎	2	217.650	88
18	福井	3	80.031	50	43	熊本	2	64.845	35
22	静岡	4	96.930	54	44	大分	3	72.857	35
24	三重	4	66.485	37	45	宮崎	2	80.697	38
25	滋賀	3	75.174	44	46	鹿児島	2	62.904	30
26	京都	4	71.046	46	47_1	沖縄 A	2	67.384	43

に音響モデルおよび言語モデルを構築しているため、いわゆる標準語への適応が中心であり、方言音声に対しては言語モデルが適さなかったり、外国語を母語とする日本語学習者の音声には音響モデルがうまく適合しないことが考えられる。しかし、そのような音声に対してもテキスト自動アライメントシステムが効果的に活用できるのであれば、コーパス構築の負担軽減に大きく貢献できる。

そこで本稿では、現在構築作業が進められている『日本語諸方言コーパス (CJD)』(木部ほか 2017) へのテキスト自動アライメントの導入を考慮し、日本全国の様々な地域の方言音声に対するテキスト自動アライメント性能の検証を行った結果について報告する。

2. 方言音声のテキスト自動アライメント

2.1 対象データ

表 1 に本稿で取り扱う方言音声データの種類および地域ごとのデータ量を示す。これらは、『全国方言談話データベース 日本のふるさとことば集成』(国立国語研究所 2002) に収録されている音声のうち、CJD 構築のために人手で転記テキストのアライメントがすでに行われている部分である。地域によって作業の進捗状況が異なるため、それぞれ 60 秒から 200 秒程度と地域ごとに異なる長さの音声データになっている。なお、音声フォーマットはサンプリング

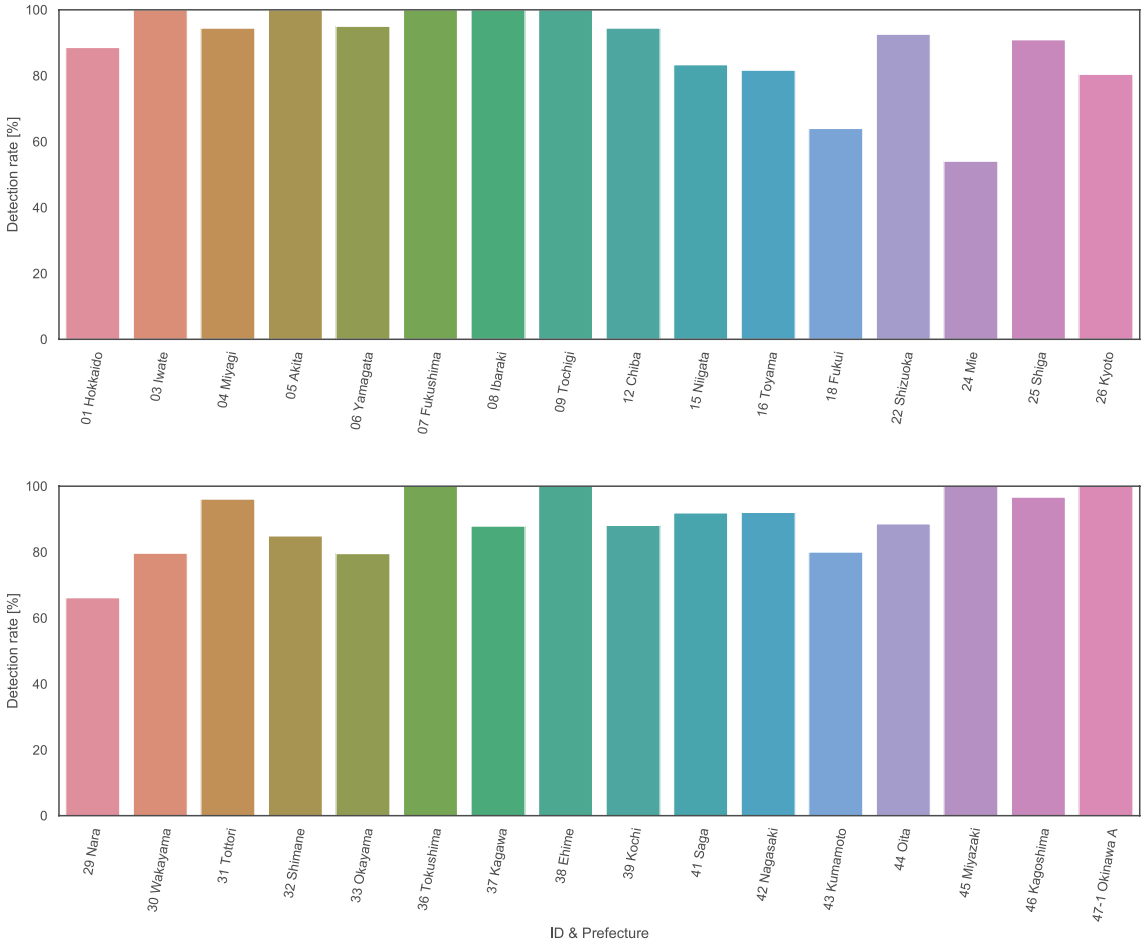

図 1　発話開始時刻推定における検出率

周波数 16 kHz、量子化ビット数 16 bit の PCM である。音声の内容は各地域の方言会話であるが、複数話者の会話が 1 チャンネルに収録されているため発話音声の重複が生じている箇所があり、この重複箇所においては音声認識性能の低下が予想される。また、転記テキストは 200 ms の無音区間を境界とする間休止単位で区切られているため、本稿では間休止単位を発話単位として用いることとする。

2.2　結果

前述の自動字幕作成システムを用いてテキスト自動アライメントの精度を調べる。なお、自動字幕作成システムは字幕用の処理を目的としているため、現時点では発話開始時刻のみの推定となっている。そのため、発話開始時刻に焦点を絞り、音声データおよび転記テキストから各発話の開始時刻を求めるアライメント処理を行った。

2.2.1　検出率

すべての入力テキストに対してアライメントが行われるわけではなく、システムが発話を検出できない場合は、発話位置が推定がされないこともある。各地域のデータに対し発話開始時刻を推定できた発話数の割合（検出率）を図 1 に示す。

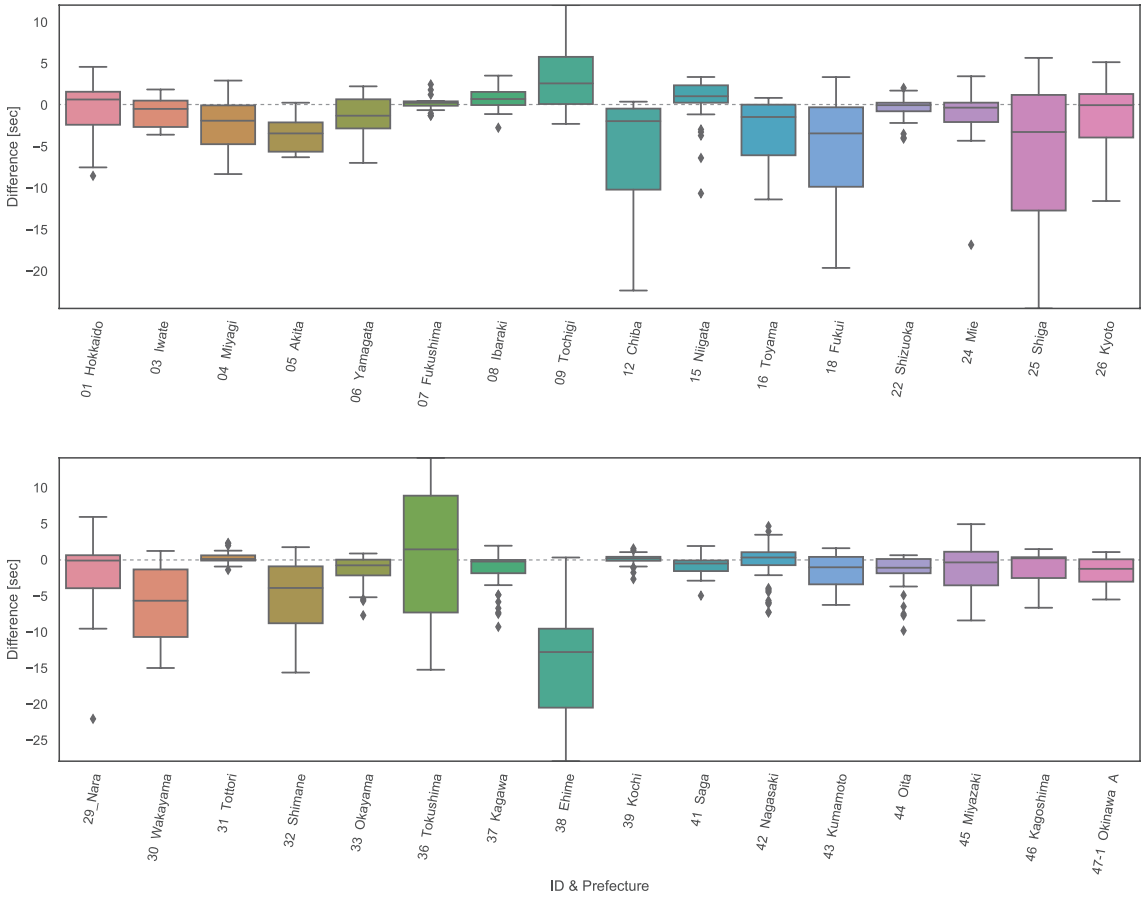

図 2　発話開始時刻の推定誤差

概ね 80% 以上の検出率を示しており、80% 以下の検出率となった地域は、福井、三重、奈良、和歌山、岡山、熊本であった。近畿地方に検出率が低い地域が比較的偏っているようにも思われるが、特に低い福井や三重に隣接する滋賀の検出率は高く、地域の位置と検出率の関係は絶対的なものではない。また、標準語の言語モデルに適合しないように思われる東北地方や九州地方については、予想に反し検出率の大幅な低下は見られなかった。

2.2.2　推定誤差

各地域における正解開始時刻と推定開始時刻との差（推定誤差）の分布を図 2 に示す。なお、人手で行った転記テキストのアライメント結果を正解時刻とみなす。値が負値である場合は推定時刻が正解時刻よりも前になっており、正値である場合は推定時刻が正解時刻よりも遅れていることを意味する。

大きく分類すると、ほとんどの発話の推定誤差が 0 に近い（すなわち正しく推定されている）地域と、一部のみ正しいものの誤りも多い地域、ほとんどの発話で正しい推定ができていない地域の 3 つに分かれている。例えば、北海道、京都、奈良、徳島などは、正しく推定されている発話もあれば大きくずれている発話もあることがわかる。また、宮城、秋田、栃木、千葉、富山、福井、和歌山、島根、愛媛といった地域は推定時刻の多くが正解時刻から大きくずれている。一方で、福島、鳥取、高知といった地域はほとんどの発話に対して正しくアライメ

ントができている。このように推定誤差に関しては、特定の地方に偏って誤差が大きくなるような傾向は見られない。反対に誤差の小ささに着目すると、九州・沖縄地方は他の地方に比べて安定して推定誤差が小さい傾向があるように思われる。

2.3 考察

検出率は低いものの推定誤差が小さい三重のデータを詳細に見ると、ある程度の大きさの暗騒音がデータ全体に入っており、音声の大きさが小さい箇所の検出に失敗している傾向にある。一方、雑音に対して相対的に音声が大きい発話に対しては概ね正確に推定できており、雑音の有無が自動アライメント性能に影響を与えていることがうかがえる。雑音のみが問題であれば、観察された暗騒音はほぼ定常音であるため、雑音除去の前処理を行うことで検出率の改善が期待できる。

また、検出率は高いが推定誤差が大きい秋田や愛媛のデータにおいては、会話の冒頭付近の大きな推定誤差がのちの発話の推定時刻に影響しているようであり、入力に用いた転記テキストの記述順と発話音声の出現順の整合性を重視しすぎた結果として全体的なズレが生じている可能性がある。すなわち、音声に転記テキストを強制的に割り当てる手続きによって本来とは異なる発話音声にテキストが誤って付与されることになり、大きな誤差が生じていると考えられる。音声データの時間が長いほどこのズレは広範囲に影響することが考えられるため、適切な長さに音声データを分割してからアライメントを行うことが必要になるかもしれない。

全体に目を向けると、地方によるアライメント性能の低下は顕著には現れていない。これは標準語との言語的な差異がアライメント性能に大きな影響を与えていないことを示しており、方言音声に対してテキスト自動アライメントが適応できないのではないかという当初の懸念を払拭するものである。もっとも、CEJC を対象にしたアライメントの性能評価では検出率が 98% で推定誤差が概ね ±1 秒程度であった (石本 2017) ことを考えると、まだ無条件でシステムを適応できるものではなく、低い検出率や大きな推定誤差が生じる原因についてさらに分析し明らかにする必要がある。また、誤差が十分に小さいとは言えないことから、後処理において作業者による修正作業は必須であろう。それでも、すべて人手で付与する労力を考慮すると、方言のような標準語とは異なる音声のコーパス構築においてもテキスト自動アライメントによるラベル付与の負担軽減の効果はあり、本システムの活用はコーパス構築の効率化を推し進める仕組みになりうると考える。

3. おわりに

本稿では、音声コーパス構築における負担の軽減を目指して、日本の様々な地域の方言音声に対する転記テキストの自動アライメント性能の検証を行なった。その結果、標準語からの言語的な違いがアライメント性能に影響を与えることはほとんどなく、方言音声に対してもシステムによるテキスト自動アライメントが実現できることが示された。システムの後処理として作業者によるラベル修正作業は必要となるものの、作業者との連携を適切に設計することにより自動アライメントシステムは方言音声のコーパス構築においても有効に活用できると考えられる。

謝　辞

本研究は、国立国語研究所共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」および「大規模日常会話コーパスに基づく話し言葉の多角的研究」の支援を受け、国立国語研究所共同研究プロジェクト「日本の消滅危機言語・方言の記録とドキュメンテーションの作成」による成果を利用して行われたものである。

文　献

秋田祐哉・三村正人・河原達也 (2015).「音声認識を用いた講義・講演の字幕作成・編集システム」　情報処理学会研究報告　2015-SLP-108(2) 巻, pp. 1–6.

河原達也・秋田祐哉・広瀬洋子 (2016).「自動音声認識を用いた放送大学のオンライン授業に対する字幕付与」　情報処理学会研究報告　2016-AAC-2(5) 巻, pp. 1–4.

石本祐一 (2017).「コーパス構築における発話アライメントの現状」　言語資源活用ワークショップ 2016 発表論文集, pp. 30–37.

小磯花絵・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉 (2017).「『日本語日常会話コーパス』の構築」　言語処理学会第 23 回年次大会発表論文集, pp. 775–778.

国立国語研究所 (2006).「日本語話し言葉コーパスの構築法」　国立国語研究所報告:124.

木部暢子・佐藤久美子・中西太郎・中澤光平 (2017).「『日本語諸方言コーパス』の構築について」　言語資源活用ワークショップ 2016 発表論文集, pp. 57–68.

国立国語研究所 (2002).『全国方言談話データベース日本のふるさとことば集成』　国書刊行会　1–20 巻.

関連 URL

音声認識を用いた自動字幕作成システム

`http://caption.ist.i.kyoto-u.ac.jp/`