コーパス構築における発話アライメントの現状

著者	石本 祐一
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	30-37
発行年	2017
URL	http://doi.org/10.15084/00001455

コーパス構築における発話アライメントの現状

石本 祐一 (国立国語研究所コーパス開発センター) *

Present Condition of Automatic Alignment of Utterance Transcription for Speech Corpus Development

Yuichi Ishimoto (National Institute for Japanese Language and Linguistics)

要旨

音声コーパスの構築にあたり、音声信号に対し発話・音韻・韻律などの各種ラベルを付与する必要がある。これらのラベルは音声分野の知識を有した作業者による目視や聴音を基に付与されることがほとんどであり、大規模コーパス構築において大きな負担となっている。特に近年研究対象となることが多い自発発話では、言い誤りや言い澱み、曖昧な発声などの現象が頻繁に生じるため、自動ラベリングを困難にしている。本稿では、転記テキストのラベリングに焦点を絞り、既存の音声認識によるシステムを応用した自動アライメントの現状について報告する。自発発話が収録されている「日本語話し言葉コーパス (CSJ)」および「日本語日常会話コーパス (CEJC)」を用いてシステムの性能評価を行い、自動アライメントの今後の課題について述べる。

1. はじめに

音声コーパスを様々な研究分野で活用することを考慮すると、音声信号から読み取れる情報が種々のラベルとして付与されていることが望ましい。例えば、言語研究では使用されている文法や語彙に着目するために単語境界や品詞などの形態論情報が求められるし、会話研究では形態統語的な情報以外に発話中のポーズや発話タイミングも重要となる。音声学的研究においてはイントネーションやアクセントなどの韻律情報が必要となるし、音声工学的研究では言語情報に加えて基本周波数やスペクトルなどの音響特徴量が用いられる。他にもパラ言語的研究では感情や態度といった発話に対する印象評価が必須となる。このように研究の目的によって音声コーパスに求められる要素が異なることから、コーパスを幅広い研究分野に供するためには付与するラベルの充実がコーパス構築における重要課題となる。

しかし、これまでに公開されている音声コーパスにそのような種々のラベルが付与されていることはほとんどない。これはラベリングに対する負担が非常に大きいためである。ラベルの多くは音声・言語分野の知識を持った作業者により人手で付与される必要があり、コンピュータによる自動解析が利用できる一部のラベルについても最終的には人手による修正が不可欠であることが多い。このラベリングの負担を軽減しコーパス構築を容易にするためには、コン

* yishi@ninjal.ac.jp

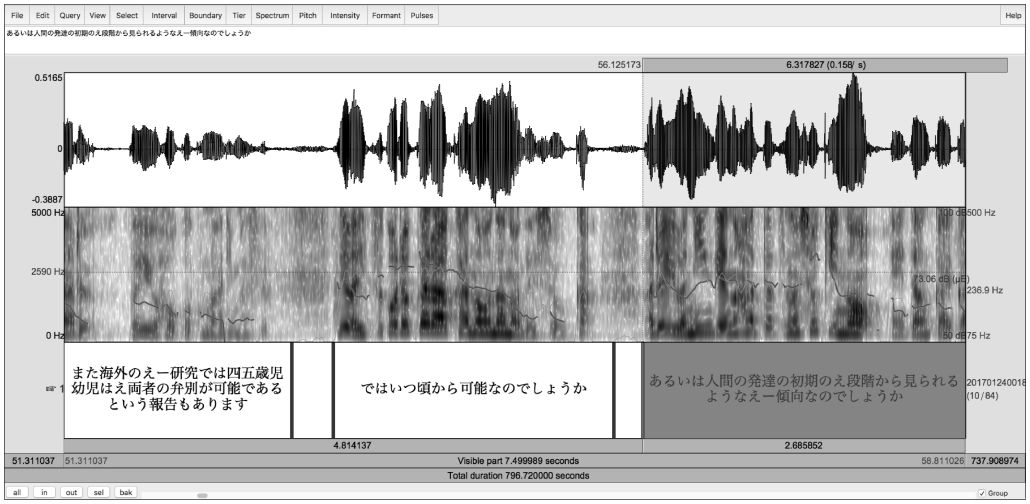

図 1　Praat による発話開始・終了時刻のアノテーション

ピュータによるラベリングの自動化が適用される範囲を広げるほかない。

本稿では、音声コーパスに付与されるラベルのうち発話を文字で書き起こしたテキスト（以下、転記テキスト）に焦点を絞り、音声データへの転記テキストの配置について、コンピュータでの自動処理における現時点での実用可能性について報告する。

2. 転記テキストのアノテーション

音声コーパスの構築においては発話に関わる様々な情報がラベルとして付与される。そのひとつである転記テキストは音声から文字への単なる書き起こしにとどまらず

- 発話単位
- 発話内の時間関係 (ポーズ)
- 発話間の時間関係 (発話の重なりや発話間の空白時間)
- 韻律・非言語情報 (強調や笑いなど)
- 非流暢性 (言い誤りやフィラーなど)

などの情報を表している。コーパスに付与される形態論情報や詳細な韻律情報といったその他のラベルはこの転記テキストを基にするため、コーパスの基盤となるものである。

しかし、転記テキストのアノテーション作業は転記基準を熟知した作業者による手作業によるところが大きく、コーパス構築における初期の問題となっている。例えば、比較的容易な発話の開始・終了時刻の認定においては、波形やスペクトログラムが表示される音声分析ソフトウェア（図 1）を用いて、実際の音声を聞き波形を見ながら数 ms 単位での調整が必要となる。つまり、発話位置を探し転記テキストを開始・終了時刻に合わせ調整（アライメント）する作業だけで発話の実時間の数倍・数十倍の時間が費やされることになり、このような作業が自動化されるだけでもコーパス構築の負担軽減が期待できる。

3. 音声認識を用いた転記テキストの自動アライメント

音声情報処理研究において、検索対象の語に適合する音声データの位置を特定する「音声ドキュメント検索」と呼ばれる問題がある (秋葉 2010)。音声ドキュメント検索は (1) 音声認識と (2) 音声と認識結果との関連づけを組み合わせた技術であり、音声ドキュメント検索が実用化されれば、その応用でコーパス構築における転記テキストの書き起こしおよびアライメント作業の自動化も可能となるであろう。しかし、実環境に存在する雑音の影響や自発発話の非流暢性などの問題から日常場面での音声認識の精度はまだ不十分である。そこで本項では、発話を書き起こしたテキストがすでに存在する状態を仮定し、テキストと音声とを関連づけることで発話位置を認定する「転記テキストのアライメント」の自動化について検討する。

3.1 自動字幕作成システム

書き起こしテキストデータから映像・音声内の位置を特定する既存システムとして、音声認識を用いた自動字幕作成システム (秋田ほか 2015, 河原ほか 2016) が公開されている。このシステムは、音声ファイルや映像ファイルを入力とし、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して字幕として提示できるようにする目的で構築されており、実際に放送大学の講義の字幕付与に利用されている。また、音声認識結果をそのまま書き起こしテキストとして用いるのではなく、あらかじめ入力されたテキストに対して音声を同期させる（テキストに音声の時刻を付与する）「同期限定モード」があり、上述の転記テキストの自動アライメントを行うシステムとしての利用が期待できる。ただし、字幕作成に特化したシステムであるため、発話終了時刻は重視されていない。そこで、本稿ではアライメントについて発話開始時刻だけを取り上げることとする。

3.2 データ

すでに転記テキストが付与されているコーパスデータを用い、自動字幕作成システムによるアライメントの結果と比較することで、システムによる自動アライメントの可能性を探る。

データは、日本語話し言葉コーパス (CSJ)(Maekawa et al. 2000) と日本語日常会話コーパス (CEJC)(小磯ほか 2015) から抜粋して用いた。

CSJ からは

- 学会講演 2 名分（男女各 1 名）
- 模擬講演 2 名分（男女各 1 名）
- インタビュー対話 2 対話分（インタビュイー男女各 1 名）

を用い、学会講演発話、模擬講演発話、インタビュアーの発話、インタビュイーの発話の 4 タイプについてシステムのアライメント結果を調べた。インタビュー対話をインタビュアーとインタビュイーに分けたのは、インタビュアーの発話はフィラーや相槌が多く、インタビュイーの発話とは異なる傾向をみせると考えられたためである。システムへの入力には音声と転記テキストを用いる。CSJ では話者ごとに近接マイクを配置して音声を収録しているため、音声は雑音の非常に小さいクリアな音質となっている。テキストについては CSJ に付与されている転記テキストから転記記号を全て取り除いた上で節単位絶対境界または強境界を発話区切りと

表 1　CSJ に対する発話開始時刻の推定数

	学会講演	模擬講演	対話	
			インタビュアー	インタビュイー
正解数	185	190	317	247
推定数	185	190	309	245
検出率	100.0%	100.0%	97.5%	99.2%

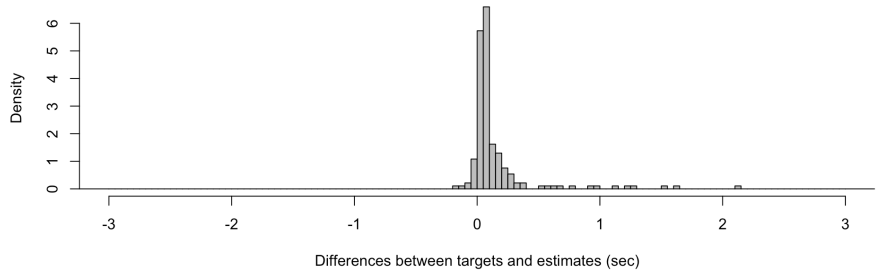

図 2　CSJ の学会講演における発話開始時刻の推定誤差

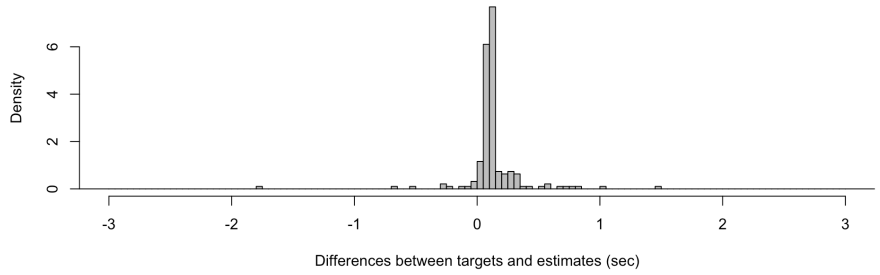

図 3　CSJ の模擬講演における発話開始時刻の推定誤差

して設定した。なお、自動字幕作成システムでは講演・スピーチ・討論の 3 つの音声認識モデルが選択できるが、講演モデルは CSJ の学会講演、スピーチモデルは CSJ の模擬講演のデータにより構築されており、CSJ データに対してそれぞれ対応する音声認識モデルを選ぶことで理想的な環境でのシステム出力とみなすことができる。

CEJC はまだ構築が済んでおらず公開されていないが、作業者による転記テキストのアライメントが完了したデータから

- 環境音の大きい飲食店内の女性 2 名の対話（以後、会話 1）
- 環境音のほとんどない室内の女性 2 名の対話（以後、会話 2）

の 2 会話を用いた。会話 1 の話者 2 名（以後、話者 A、話者 B）と会話 2 の話者 2 名（以後、話者 C、話者 D）のそれぞれについてシステムのアライメント結果を調べた。CEJC では話者ごとに IC レコーダを配置して収録しているため、システムへの入力には各話者の IC レコーダの音声を用いた。ただし、周囲の環境によって雑音や BGM、他者の音声などが入り込んでおり、話者の音声は必ずしもクリアではない。入力テキストには、書き起こしテキストを音響的

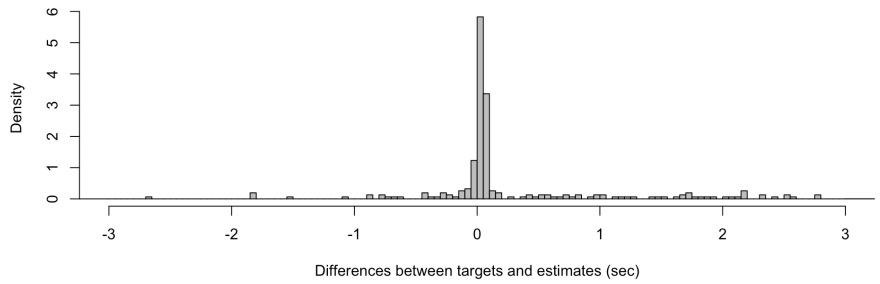

図 4　CSJ のインタビュー対話（インタビュアー）における発話開始時刻の推定誤差

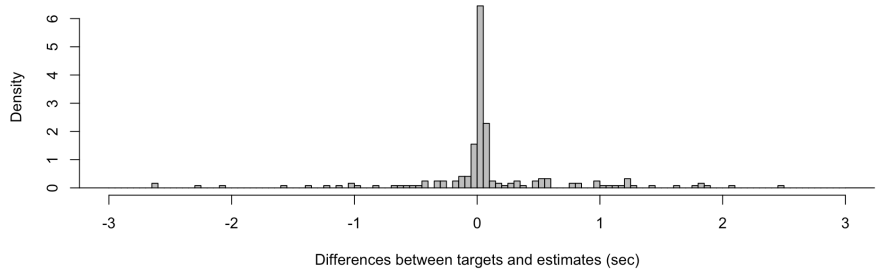

図 5　CSJ のインタビュー対話（インタビュイー）における発話開始時刻の推定誤差

な切れ目や韻律的な切れ目で区切った「短い発話単位」(Den et al. 2010) を基にした単位を用いた。そのため、CSJ よりも短い発話が多いデータとなっている。また、システムの音声認識モデルはスピーチのみを使用した。

3.3　結果

自動字幕作成システムでは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、発話位置の推定ができないこともある。表 1 に CSJ のデータにおいて発話開始時刻を推定できた発話数を示す。

学会講演、模擬講演ではすべての発話に対して発話開始時刻を推定できているが、インタビュー対話については少数ながらも推定できていない発話があった。推定されなかった発話は「うん」「うーん」「ええ」「はー」といった波形振幅が小さく 1 発話の長さが短い発話がほとんどであった。ただし、同様の発話であっても推定されているものもあるため、小さく短い発話がまったく推定できないわけではない。むしろ、97% 以上の発話が推定できていることから、非常に高い検出精度をシステムが有しているといえる。

次に、コーパスにあらかじめ付与されている発話開始時刻を正解値として、システムで推定された発話開始時刻との差を推定誤差として算出した。図 2–5 に CSJ のそれぞれの発話タイプにおける推定誤差のヒストグラムを示す。ヒストグラムの bin 幅は 50ms とした。±3 秒以上の誤差を生じた発話も存在したが、少数であるため図示の対象外としている。

図 2,3 からわかるように学会講演、模擬講演に対しては推定誤差が非常に小さくなっており、ほとんどが ±300ms 程度の範囲におさまっている。これは非常に高い精度で発話開始時刻を

表 2　CEJC に対する発話開始時刻の推定数

	会話 1		会話 2	
	話者 A	話者 B	話者 C	話者 D
正解数	656	798	994	1014
推定数	651	788	974	930
検出率	99.2%	98.7%	98.0%	91.7%

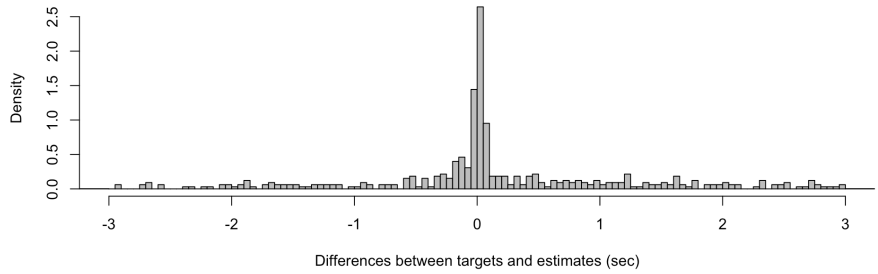

図 6　CEJC の会話 1・話者 A における発話開始時刻の推定誤差

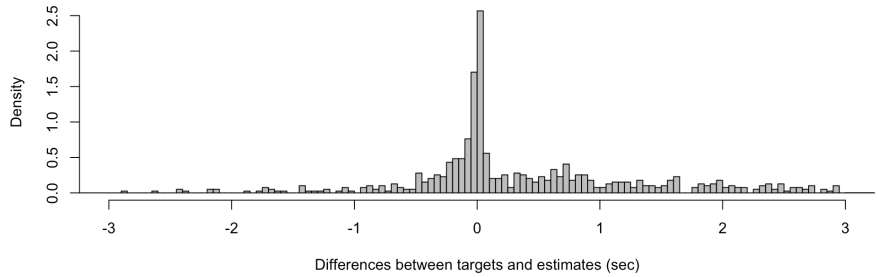

図 7　CEJC の会話 1・話者 B における発話開始時刻の推定誤差

推定できていることを示している。一方、図 4,5 に示されるインタビュー対話の推定誤差をみると、概ね ±300ms 程度におさまっているがなかには ±1,2 秒程度のズレが生じているものもあり、学会講演や模擬講演よりも精度が低下している。このような大きな誤差が生じる発話を個別にみると、ほとんどが「うん」や「うーん」といった上述の推定できなかった発話と同種のものであった。インタビュイーとインタビュアーの間で推定誤差の傾向に大きな違いは見られないが、これはインタビュアーに多いフィラーや相槌が検出不能としてある程度除かれた後の評価であるためと考えられる。

CEJC のデータにおける推定数と推定誤差についても同様に調べた。表 2 をみると、会話 1 の話者 A, B および会話 2 の話者 C に対しては 98% 以上という高い検出率を示した一方で、会話 2 の話者 D に対しては 92% 弱の検出率となった。会話 1 では環境音が大きく入り込み雑音があるにもかかわらず検出不能な発話が少ないことになる。しかし、図 6,7 で示される推定誤差からわかるように誤差が大きい発話も多く現れ、高精度の推定できているとはいえない。環境音の小さい会話 2 の結果を示す図 8,9 からも同様に推定誤差が大きく、特に検出率が低

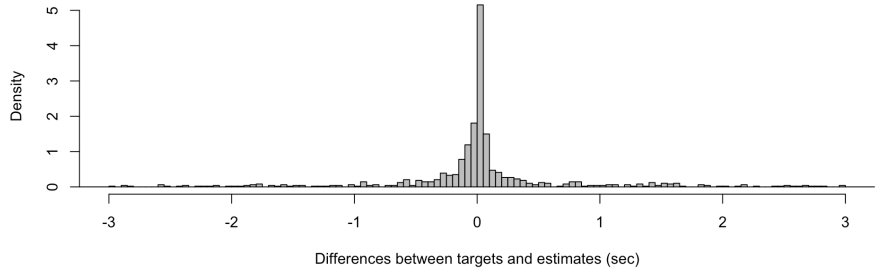

図 8　CEJC の会話 2・話者 C における発話開始時刻の推定誤差

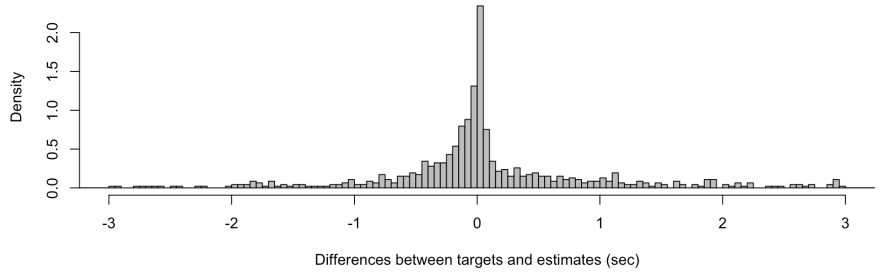

図 9　CEJC の会話 2・話者 D における発話開始時刻の推定誤差

かった話者 D に対しては誤差が大きい発話が話者 C よりも多くみられる。話者 D の検出不能の発話には「うーん」のようなフィラーだけではなく 3 秒程度のある程度の長さの発話も含まれているが、総じて大きさが小さい発話であった。また、会話 1,2 ともに推定誤差が大きい場合は全く異なる発話を指し示していることになるが、ひとつの発話の推定時刻がずれることにより後続の発話の推定時刻を誤る箇所がみられた。

3.4　考察と今後の課題

CSJ の学会講演や模擬講演において高精度で発話開始時刻を推定できているのは、音声認識の性能が大いに関係していると考えられる。すなわち、高い認識率を示す環境であれば、システムを用いた転記テキストの自動アライメントはほぼ実用的な段階に入っているといえよう。

しかし、CEJC の会話 1 のように環境音が大きい場合は検出率は高いものの推定誤差が非常に大きくなった。これはその環境音を誤って発話として認識してしまうことが原因と考えられる。また、CEJC の会話 2 のように環境音が小さい場合でも推定誤差が大きくなることがある。これはマイク位置が対象話者から離れていることにより対象話者以外の音声が入り込み、非対象話者の音声を誤って認識していることが理由のひとつとして挙げられる。以上のことから、雑音・非対象話者を含む音声に対する認識器の耐雑音性向上がシステムの適用範囲を広げるための重要な要素になっている。もっとも正しく推定できている発話も多数あることから、現段階のシステムの性能でも自動アライメントに加えて作業者による後処理を施すことを考慮すれば、コーパス構築の負担軽減には十分に役立つ状況であるといえる。

今回利用した自動字幕作成システムでは耐雑音のために振幅が小さい信号を認識対象外にす

るように構成していると考えられるが、その結果、自発会話で多く現れるフィラーや相槌への対応が難しくなっている。日常場面での収録においては収録環境の設定に制約があり理想的な収録音声を得ることが困難であることから、転記テキストの自動アライメントを推し進めるためには耐雑音性を高めるとともに小さな音も正確に認識するようなシステムの改善が必要であろう。

4. 終わりに

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、音声データへの転記テキストの自動アライメントについて現時点での実用可能性について検討した。コーパス構築を目的としたものではないものの、すでに実用されている音声認識による自動字幕作成システムを応用することで、ある程度の自動アライメントが可能であることが示された。この結果を基にコーパス構築で求められる特性を考慮した自動アライメントシステムの作成を進める予定である。

謝　辞

本研究は国立国語研究所共同研究プロジェクト「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」により行われたものである。また、「音声認識を用いた自動字幕作成システム」の使用を許可いただいた京都大学 河原達也教授、秋田祐哉講師に感謝いたします。

文　献

秋葉友良 (2010).「音声ドキュメント検索の現状と課題」 情報処理学会研究報告 2010-SLP-82(10), pp. 1–8.

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), 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), pp. 1–4.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara (2000). “Spontaneous speech corpus of Japanese.” *Proc. LREC2000*, pp. 947–952.

小磯花絵・石本祐一・菊池英明・坊農真弓・坂井田瑠衣・渡部涼子・田中弥生・伝康晴 (2015).「大規模日常会話コーパスの構築に向けた取り組みー会話収録法を中心にー」 人工知能学会研究会資料 SIG-SLUD-B5(01), pp. 37–42.

Y. Den, H. Koiso, T. Maruyama, K. Maekawa, K. Takanashi, M. Enomoto, and N. Yoshida (2010). “Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme.” *Proc. LREC2010*, pp. 2103–2110.

関連 URL

Praat: doing phonetics by computer

`http://www.fon.hum.uva.nl/praat/`

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/`