同時通訳研究のための大規模音声データベースとその利用

遠山仁美[1]　　松原　茂樹[2]
名古屋大学大学院情報科学研究科[1]
名古屋大学情報連携基盤センター[2]

1.　はじめに

大規模同時通訳音声データの活用は、コーパス言語学的観点から通訳学にアプローチすることを可能にする。しかしながら、通訳音声データの収集は、テキストデータや単一言語 データと比べ、必要となるコストも格段に高く、通訳研究に利用可能な音声資源を、研究者個人のレベルで大量に整えることは難しい。一方、高い品質を備えた音声情報処理技術を開発する上で、音声データベースの使用が不可欠であり、近年、大量の音声データの収集を目的とした大規模プロジェクトがいくつか推進されている。音声翻訳機での活用を目指し、通訳音声が収集されており、それらの多くは通訳研究で利用することも可能である。英語音声と日本語音声を対象に講演及び会話の同時通訳音声を収集しており、人手による音声の文字化、ならびに、談話タグの付与を実施している。その規模は音声データにして約 182 時間、文字データにして約 100 万語に達しており、世界最大の同時通訳コーパスと位置づけられる。通訳研究の発展のために、研究者間で言語資源を共有し、活用していくことが重要であり、これまでにデータ共有のための環境整備、及び、談話タグの充実を推進してきた。また、近年、配布準備の整ったデータ「CIAIR 同時通訳データベース Ver.10」の配布を行っている。

本発表では、「CIAIR 同時通訳データベース Ver.1.0」が多くの研究者によって効果的に利用されるために、データの利用環境について説明する。また、通訳者の発声タイミングの分析や訳出パターンの分類など、本データベースを使用した研究事例も生まれており、本発表ではこれらの成果について紹介することにより、データの応用可能性について論じる。

2．データベース構築の目的

名古屋大学統合音響情報研究拠点（CIAIR）では、マルチリンガルコミュニケーション支援環境の実現を目指し、話し言葉翻訳技術の向上、ならびに、通訳理論の構築を目的に、5年間にわたり、同時通訳データベースを構築してきた。データベースは独話および対話を対象として構成されており、全体で約 182 時間の音声を収録し、音声の文字化、視覚化、および、言語分析を完了している。文字化データのサイズは単語数にして約 100 万語に達し、世界最大の同時通訳データベースと位置付けられる。大規模な同時通訳音声データを駆使し、通訳理論研究にアプローチする試みは、コーパス言語学的な手法を通訳学に応用する先駆的な研究と位置づけられ、情報工学のみならず、言語学、通訳教育、認知科学な

ど、より広範な研究分野で多面的に活用され、研究領域を超えた意見交換を行い、総合的に進展していくことが望ましいと考えている。

3．データベースの設計

大規模データベースの収集は、それに費やされる膨大なコストを勘案すると、将来における幅広い利用可能性を考慮し、汎用性を備えた設計が要求される。本データベースは、多様なデータを収集するために、独話（monologue）および、対話（dialogue）の同時通訳音声をいくつかの日常的なトピックを設定して収録した。対象言語は英語と日本語とし、その双方向音声を収録した。本データベースの収録内容を表 1 に示す。

表 1．CIAIR 同時通訳データベースの収録様式

項目	詳細
談話形態	独話　　対話
対象言語	英語　　日本語
通訳スタイル	同時通訳
メディア	音声　　文字

4．データの収録

名古屋大学 CIAIR では、実音響環境下での音声データを収集することを重視しており、収録は教室レベルの録音環境を採用した。また、同時通訳者にとって、話者の発話だけでなく、表情や振る舞いも重要な情報となるため、通訳者は話者をガラス越しに観察できる通訳専用のブースに入り、通常行われる同時通訳とほぼ同じ環境下で収録を行った。収録を通して全て同一のスタンドマイクを使用し、話者と通訳者の音声を、サンプリング周波数 16kHz、16 ビットでデジタル化し、デジタルオーディオテープ（DAT）に複数チャンネル環境で収録した。また、同時通訳者は、第一線で活躍しているプロの通訳者を起用し、高い通訳レベルを保証している。

4.1 独話データの収録

独話の収録では、講演者が通訳者の通訳状況を気にせず、自分のペースで発話できるよう、講演者には通訳者の音声が聞こえないようにした。一方、通訳者は通訳用ブースに入り、講演者の振る舞いが見える中で、ヘッドホンから流れる講演者の音声に対して、同時通訳する。講演の聴衆は、ヘッドホンを利用して、通訳者の音声を聴くことができる。また、模擬講演の形式を採用することにより、発話内容（トピック、話速）の制御をある程度可能にし、品質の高い同時通訳データの収録を試みた。

また、英語あるいは日本語の講演者に対し、複数の同時通訳者が通訳を行った。つまり、

１つの講演者発話ソースに対し、複数の通訳データが存在する。従って、個人に特化しない多くの通訳事例を幅広く収集したり、１つの発話に対する、複数の通訳事例を比較することが可能である。また、通訳経験年数の違いによる訳出の特徴分析などにも利用でき、データベースの汎用性を高めている。

4.2 対話データの収録

対話の収録では、英語話者と日本語話者の異言語間対話に対し、通訳の品質を高めるために、英日、日英の 2 名の同時通訳者を設置する形態をとった。また、会話における話者の発話権を確保するために、話者は相手話者の発話を通訳した結果のみを聞くことができるようにした。一方、通訳者は、対話全体の流れを把握するために、担当する話者の音声だけでなく、もう一方の話者音声も聞ける環境を設定した。また、可能な限り自然な対話を収集するため、話者役割と対話タスクの設定のみを行い、基本的には自由発話という様式で収録した。 例えば、ドメインがホテルの予約であれば、客を担当する話者には予約したいホテル名と予約人数を、ホテル側を担当する話者には、空室状況のみを提示し、自由発話で対話を進めていく。1 対話あたりの収録時間は 1 分から 16 分であり、多様なタイプの対話収録を実現している。

5. データベースの構成

本データベースの構成を図 1 に示す。本データベースは音声データファイル、文字化データファイル、環境データファイルの３つから構成されている。

- 音声データファイル

  話者の音声と通訳者の音声を多重音声ファイルとして収納している。

  - 独話音声データファイル

    １つのセッションにつき英語と日本語からなる１つの多重音声ファイルを収納した。

  - 対話音声データファイル

    話者 1 人と担当通訳者 1 人の多重音声ファイルを１ファイルとし、２つの多重音声ファイルを作成した。

- テキストデータファイル

  英語話者発話テキスト、英日通訳者発話テキスト、日本語話者発話テキスト、日英通訳者発話テキストの４つに分け、データの種類、性質ごとにディレクトリで区分して収納した。

- 環境データファイル

  独話、および、対話に対し行った全ての通訳に対し、その環境（発話トピック、話者役割（対話のみ）、話者情報、通訳者情報（経験年数））に関する情報を記述したものを１ファイルとして収納した。

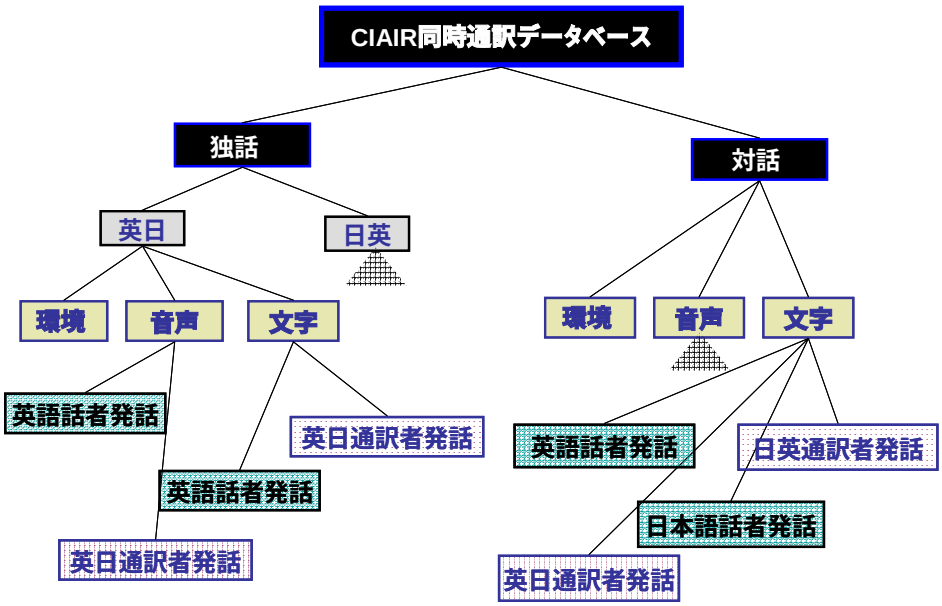

図1. データベースの構成

6. 音声データの文字化

音声データの文字化は日本語話し言葉コーパス（CSJ）の書き起こし基準に準拠した。以下に基準を示し、サンプルとして図 2 をあげる。

- 発話単位

  話者および、通訳者の音声を 200msec 以上のポーズ（無音声区間）で分割し、発話単位を定めた。

- 表記方法

  日本語音声に限り、片仮名で表記する「発音形」と漢字仮名まじりで表記される「基本形」の 2 種類で構成している。

- タグ情報の付与
  - 発話 ID

    全発話に対し通し番号を付与した。
  - 時間情報タグ

    発声の開始時刻と終了時刻を付与した。
  - 談話タグ

    話し言葉に特有の言語的現象であるフィラー（「えー」、「あのー」etc.）や言い淀みについて、談話タグを付与している。

```
0001 - 00:02:148-00:02:716 N:
Good morning<SB>
0002 - 00:02:829-00:03:073 N:<breath>
0003 - 00:03:263-00:05:388 N:
(F uh) <tongue> I'm trying to
 take a trip
0004 - 00:05:805-00:05:807 N:<FV>
0005 - 00:06:011-00:08:268 N:
today<SB> I wanna hire a car<SB>
Is that possible?<SB>
0006 - 00:19:265-00:20:212 N:
Yes、 here it is<SB>
```

```
0001 - 00:03:125-00:04:111 I:
おはようございます<SB>
オハヨーゴザイマス<SB>
0002 - 00:06:645-00:06:647 I:<FV >
0003 - 00:06:917-00:10:098 I:
  (F え)ちょっと(D う)旅行したいんですけども(D
  おわ)車借りられますか<SB>
  (F エ)チョット(D ウ)リョコーシタインデスケド
  モ(D オワ)クルマカリラレマスカ<SB>
0004 - 00:20:055-00:22:120 I:
はいこれが(F ん)免許証です<SB>
ハイコレガ(F ン)メンキョショーデス<SB>
```

図２．文字化データのサンプル　（対話―英日同時通訳）

7. データベースの利用

我々は、本データベースを利用し、同時通訳における通訳者発声タイミングや、通訳者を介したコミュニケーションの円滑さなど、通訳における個別の現象について分析を進めてきた。これら情報工学の分野のみならず、言語学、通訳学、認知科学など、多分野における、同時通訳を対象とした緻密かつ定性的な研究成果においても、分析対象となるデータが大規模であれば、さらに定量的に検証することも可能となり、有効となるのではないかと考えられる。

8.「CIAIR 同時通訳データベース Ver.1.0」の配布

現在、配布準備が整ったデータ「CIAIR 同時通訳データベース Ver.1.0」の配布を行っている。独話データは 16 講演分の話者･通訳データ（6.2 時間分）、対話データは 43 種類のトピックに関する話者･通訳者データ（4.6 時間分）で、音声データ、テキストデータ、環境情報データを合わせ、全体で約 4GB である。詳細については以下の表を参照されたい。今後、さらに、配布データ量の増加、時間情報、タグ情報、ツールの充実をはかり、随時、アップグレード版を配布していく予定である。

表2． CIAIR 同時通訳データベース Ver.1.0　独話データの詳細

講演通訳		発話タイプ	テーマ	収録時間	音声データサイズ	テキストデータサイズ
英日	英語話者	原稿参照発話	政治 経済 文化	262.5分 (4.4時間)	0.96GB	118KB
	英日通訳者					208KB
日英	日本語話者	原稿参照発話	政治 経済 文化	278.5分 (4.6時間)	1.03GB	120KB
	英語話者					179KB

表3． CIAIR 同時通訳データベース Ver.1.0 対話データの詳細

対話通訳	発話タイプ	テーマ	収録時間	音声データサイズ	テキストデータサイズ
英語話者	摸擬対話	ホテル フロント 空港 カウンタ	274.0分 (4.6時間)	1.96GB	262KB
英日通訳者					343KB
日本語話者					289KB
英語話者					203KB

<お問合せ先>
名古屋大学情報科学研究科社会システム情報学専攻
同時通訳データベース担当　sidb@el.itc.nagoya-u.ac.jp
ホームページ　　http://slp.el.itc.nagoya-u.ac.jp/sidb/