2-2　音声コミュニケーション技術
2-2　Speech Communication Technology
2-2-1　多言語音声コーパス
2-2-1　Multilingual Speech Corpora

水上 悦雄　加藤 宏明
MIZUKAMI Etsuo and KATO Hiroaki

統計的機械学習手法を用いる音声認識モデルにおいては、学習データとしての“音声コーパス”が必要となる。NICTでは、第4期中長期計画において、生活会話における日英中韓をはじめとした15言語の実用レベルの音声翻訳技術を開発し、現在の第5期中長期計画においては、ビジネスシーンにおける実用的な自動同時通訳技術の開発を目指している。本稿では、これらの音声翻訳技術の一つの要素技術である音声認識の開発及び改善のために、多言語の音声コーパスをどのように設計し、構築してきたのか、について報告する。

In speech recognition modeling using statistical machine learning methods, “speech corpus” is necessary as training data. During the 4th mid-long-term plan, NICT developed a practical speech-to-speech translation technology for daily conversations supporting 15 languages including Japanese, English, Chinese and Korean and in the current 5th mid-long-term plan, we aim to develop a practical-level automatic simultaneous interpretation system that can be used for business. In this paper, we report the designing and development process of constructing the multilingual speech corpora used for developing and improving speech recognition—one of the elements of speech-to-speech translation technology.

1　まえがき

近年のAI技術のコア技術となる深層学習を主とした機械学習によるモデル構築においては、学習対象となるデータの適切な選定とデータ量がそのモデル性能に大きく影響を及ぼす。情報通信研究機構（NICT）で研究開発を推進している自動音声翻訳（speech-to-speech translation）技術においてもそれは変わらない。現状のスタンダードな音声翻訳技術は、音声認識・機械翻訳・音声合成の三つの技術の結合技術として達成されているが、このうち、本稿で取り上げる音声コーパス（speech corpus）は、音声認識技術のための学習データであり、音声データとその書き起こしデータから成る。音声認識（automatic speech recognition）は、音声をテキスト（文）に変換するシステムであり、一連の音（音素）の入力に対して、もっともらしい文字を推定するためのモデル、すなわち“音響モデル”と、一連の文字の入力に対して、もっともらしい単語列、文章を推定するためのモデル、すなわち“言語モデル”の二つのモデルの統合によって実現されている。このうち、音声コーパスは、音響モデル学習のための学習データと言える*1。本稿ではNICTで開発している音声認識のための多言語音声コーパス（multilingual speech corpora）の概要を記述するとともに、その開発の歴史について述べる。

2　NICT多言語音声コーパス

2014年に総務省が掲げたグローバルコミュニケーション計画（以降、GCPと記述する。）に基づき、NICTではその第4期中長期計画において、東京オリンピックを見据え、インバウンドあるいはアウトバウンドの利用者が、来日時、あるいは訪問国において、互いが互いの母語でのコミュニケーションを可能とするよう

*1　書き起こしは言語モデルの学習にも用いるが、それだけでは十分ではないので、Web等から収集したより大規模なテキストコーパスをベースとするのが一般的である。

な、基礎的な会話における実用レベルのいわゆる“言語の壁を超える”ための音声翻訳性能を達成した。続く第5期中長期計画においては、2025年の大阪万博に向けて、ビジネスシーンでの実用レベルでの自動同時通訳のための研究開発を推進している（以降、GCP2025と記述する。）。音声認識技術の側面から見た、GCPとGCP2025の目標の違いは、前者が、生活会話（交通機関や店舗、病院、公共施設等を利用する際に交わされる会話）を対象とし、機械への音声入力を前提とした比較的改まった発話スタイルでの、一人の話者による、一発話ごとに区切られた音声を認識対象としているのに対して、後者は、自動同時通訳の導入が期待されるような、ビジネスシーンで行われる、プレゼンテーションや打合せ、会議などで交わされる様々な発話スタイルでの、複数話者による、連続した音声に対する逐次認識を目標とする点である。これらの音声認識技術の開発目標にあわせて、音声コーパスを設計しなければならない。

また、目的が音声翻訳である以上、互いに自分の母語での音声入力が前提となるため、対象とする全言語の音声認識モデルの開発が必要となり、そのための多言語の音声コーパスが必要となる。GCPにおいては、訪日インバウンド対象国、日本からのアウトバウンド対象国の公用語に鑑み、その対象を日本語、英語、中国語、韓国語、タイ語、ベトナム語、インドネシア語、ミャンマー語、フランス語、スペイン語（以降、GC10言語と記述する。）とし、その後、ブラジルポルトガル語、フィリピン語を追加し、さらに追加言語として、その対象をネパール語、クメール語、モンゴル語にまで拡張した（以降、GC15言語と記述する。）。続く、GCP2025においては、対象言語としてはGC15言語を継承しつつ、近年国内外で高まる経済安全保障上の観点から、ロシア語、アラビア語、ドイツ語、イタリア語、ヒンディー語、ウクライナ語を追加し、本稿執筆時点において、計21言語を対象とした、音声認識モデル開発のための音声コーパスを構築することとなった。以下では、NICTが開発を推進している、多言語音声コーパスの開発について、その詳細を述べる。

2.1　コーパス設計

音声コーパスは、その目的に依存して、大きく分けて、ドメイン（ジャンル）、話者属性、発話スタイル、音響環境の四要素を如何に適応的かつバランスよく収集するかがポイントとなる。前述のように、音声コーパスは主として音響モデル構築のための学習データである。そのため、当該言語で発話され得る、多様な人による、多様な音素のパターンが、しかるべき頻度で出現する必要があるが、どんな人の、どんな発話にも対応可能なモデル、ということになると、有限のコーパスでそれを実現することは難しい。そのため、ある程度目的や対象を絞ってコーパスも収集することが現実的となる。以下に、NICTにおいて開発段階に応じて現在までに構築してきた多言語音声コーパスの種別と、それぞれの設計思想及びその開発経緯について述べる。

(1) 模擬会話・独話音声コーパス

前述のように、GCPにおいてNICTが目指したのは、訪日外国人が日本国内で旅行や生活をする際に、母語で、日本人接遇者とコミュニケーションする、あるいは、邦人が海外に赴いた際に、日本語で同様にコミュニケーションすることを支援するための音声翻訳技術である。そのため、開発当初の設計では、買物場面、病院窓口、公共機関窓口、災害場面、チケット購入及び公共交通機関窓口での会話を想定して、客と接客者、患者と病院関係者、申請者と公共機関窓口担当者のような二者間の一連の会話を模した「模擬会話」を様々な状況を想定して収録するという形をとった。実際の場面ではこの二者間では異なる言語が使用されるわけであるが、インバウンドとアウトバウンドで双方の立場があり得るため、それぞれの言語でこれらを収録すれば、両方の立場の人の音声入力に対応できることになる。想定している発話スタイルは、初対面の人同士で話される程度の、フランク過ぎず、機械へのコマンド入力でもないような発話スタイルと言える。また、話者バランスも、当該言語の対象国におけるインバウンド対象層の男女比や年齢比だけでなく、方言話者比なども考慮した配分設計を行った。音響環境としては、実環境を想定しつつも、突発雑音や他者の声が大きく入るような環境での収録を避けるようにした。

模擬会話は、二者による交互の対話であることで、より自然な会話を収録することができるという利点がある一方で、二者がその場にいなければならないという制約がある。まずその二者の予定を合わせる必要がある上に、当日になってどちらかが来られないことも少なくなく、その管理コストは多大である。最初から原稿が用意されていれば、相手がいることを想定して自分の発話だけをすればよく、後にこの「独話」形式の会話も収録することとした。原稿として用いたのは、NICT先進的翻訳技術研究室が構築した多言語パラレルコーパス[1]である。多言語パラレルコーパスは、機械翻訳を目的としているため、対訳の対応関係がつきやすいように、会話と言いながらも直訳的な印象を受ける文スタイルとなっている。そのため、各言語で自然に発話できるように修正した原稿を用い、さらに現場でも各自が話しやすいように適宜変更して発話するようにし、原稿をそのまま読み上げるようなことは禁

表 1　NICT 多言語音声コーパスの諸特徴

	シナリオ	ドメイン	話者属性	発話スタイル	発話の長さ	音響環境
模擬会話・独話	あり	観光・生活	男女比、年代比、方言比率を考慮	比較的改まった発話	フレーズ～数文	比較的静音～実環境
発話ログ	なし	不定[1]	不定	不定[2]	比較的短文[3]	実環境
講演（フォーマル）	あり	ビジネス	実務経験者、標準語話者重視	丁寧、非流暢性低～中	長文	屋内実環境
会議（インフォーマル）	あり			多様、非流暢性中～大	短文～長文	
会話（カジュアル）	なし			多様、非流暢性大		

[1] 主な利用用途はインバウンド、アウトバウンドの現地でのコミュニケーション目的であることが推測されるが、語学学習用としても利用されているようである。
[2] 発話スタイルは、ユーザの利用用途に依存し、コマンド入力のようなものから、人に対する発話同様に、カジュアルであったり、丁寧であったり様々である。
[3] VoiceTra® は、無音を検出して入力を確定する EPD（end-point-detection）が導入されており、かつ、10 秒程度の制限時間がある。

じた。音響環境としても、模擬会話収録当初は、雑音の混入を恐れるばかりに防音対策のとられた部屋で収録されることが多く、周りの雑音が自然に入ってくるような、実環境での収録を推奨した。後述するように一部の言語を除き、GC15 言語のほとんどは、模擬会話形式よりも、この独話形式の音声コーパスによって構成されており、NICT の音声認識技術のベースとなっている。

(2) 発話ログ音声書き起こしコーパス

前述の「模擬会話・独話」形式で想定していたシチュエーションは、理想的には、当該状況で交わされる、人と人の自然な会話スタイルによるコミュニケーションである。ただし、実際に機械を介したコミュニケーションをする場合、必ずしも人同士の発話スタイルのようにはならず、かつ、音声認識や機械翻訳が長文を受理できないと考える話者―実際には、ある程度の長さの文であるほうがよいのだが―の発話は短いフレーズのようなものが多い。また、様々な音響環境に対応するために、学習データに雑音を重畳して耐雑音性を強化することもできるが、実環境で収録された音声を用いることができるなら、それに越したことはない。NICT では、GCP の実証実験として、また成果公開の一環として音声翻訳アプリ VoiceTra®（https://voicetra.nict.go.jp/）を現在も公開している。この VoiceTra® はまさに実環境における実ユーザ発話（発話ログ）が収録されているため、これを書き起こして、学習データに用いる*2 ことができれば、より現場音声に強い音声認識モデルを構築することができる。よって、この発話ログが十分にある言語においては、これを利活用することで、VoiceTra® 入力音声に対する頑健性を向上させている。特に、日本語は本稿執筆時点の 2022 年 7 月段階で、累計 1.3 億発話を超える音声入力があり、また、ミャンマー語は、2015 年 12 月の VoiceTra® 公開時点では、世界ではじめてのミャンマー語による音声翻訳が可能なアプリであったこともあり、徐々に利用者が増え、2018 年 12 月には一日 10 万発話を超え、一時期日本語の入力数を上回っていた。現在でも、ミャンマー語は日本語、英語に次ぎ、中国語を上回る音声入力数となっており、これら四言語については、音声コーパスのバランスとしてもその他の言語に比して発話ログが重きを占めている。GC15 言語の各言語で「模擬会話・独話」と「発話ログ」のコーパスのバランスは異なるものの、結果として生活会話に対する、実用レベルの音声認識精度達成に貢献している。

(3) 講演・会議・会話音声コーパス

前述のように、GCP における音声コーパスが、観光・生活シーンにおける機械への音声入力を前提とした短文認識を対象としていたのに対して、GCP2025 においては、講演音声やビジネスミーティング音声の逐次認識を対象としている。そのために必要な音声コーパスも必然的に、講演やプレゼン、打ち合わせや会議の音声となる。ドメインとしても、観光・生活ドメインから、各種業種におけるビジネスドメインに変わり、専門用語や、ビジネス用語への対応も必要になるだけでなく、ビジネスシーン特有の表現、言い回しがコーパス内に適切に含まれなければならない。また、発話スタイルとしても、事前に準備が可能な講演やプレゼンであれば、ある程度形式的で、言い淀みや言い誤りなどの非流暢要素も多くはないが、日々行われる打合せや会議においては、参加者間の関係性にも依存して、丁寧さの度合いや敬体も様々に変化するだけでなく、話者によっては、早口であったり、はっきりと発音し

＊2　VoiceTra® 利用ユーザの皆様には、音声翻訳技術開発のために入力データの利活用に許諾の上、利用いただいている。

ない単語があったりと、多様な非流暢要素が増大することとなり、そのような特徴が適度に含まれるコーパスでなければならない。よって、理想的には、実際のビジネス現場の音声を収録することができれば、まさに目的に応じた音声コーパスとなるのだが、機密情報保護の観点からも、収録に協力いただける企業は多くは望めず、モデル学習に必要なデータ量に対して、現実的とは言えない。

よって、GCP2025 の最初の段階では、各種業界における講演や会議のシナリオを作成して、それを当該業界の実務経験者がそれぞれの訳を演じて収録をする形式をとった。しかしながら、それだけでは、実際のビジネスシーンで交わされるような多様な特徴を有する音声への対応は十分ではないので、シナリオのない、自発発話で構成される会話音声が必要になる。そのため、テーマだけを設定して、二者、あるいは三者で自由に会話をしてもらうような会話音声も収録することとした。当初は、シナリオのある講演（“フォーマル”スタイル）を主として、シナリオのある会議（“インフォーマル”スタイル）、シナリオのない会話（“カジュアル”スタイル）の順にコーパス比率を下げる設計としていたが、その後の検討により、フォーマルやインフォーマルよりも、カジュアルスタイルの音声が比較的、講演や会議の音声認識に効果的であることが示唆された*3。よって、現在では、カジュアルの比率を重視したコーパス設計となっている。

表 1 に、「模擬会話・独話」「発話ログ」「講演（フォーマル）」「会議（インフォーマル）」「会話（カジュアル）」音声コーパスの諸特徴をまとめる。

2.2　アノテーション仕様

音声コーパスは、音声データとその書き起こしから成る、と最初に述べた。この際、音響モデル学習の観点からは、実際の発話における、各音素の正解ラベルとしての文字転記であればよい、ということになる。しかしながら、それは、例えば、日本語であれば、すべて仮名で書き下せばいい、というものではなく、言語モデル学習への援用ということを考えても、最終的な「単語列の出力」という音声認識の目的からも、当該言語の正書法、日本語であれば、漢字仮名交じりの理解可能なテキストとして記述する必要がある。また、例えば、二桁以上の数字の場合、算用数字で記述することが一般的ではあるものの、例えば、「123」と書いた場合に、これを「いち、に、さん」と読むのか、「ひゃくにじゅうさん」と読むのかが特定できないため、そのような、テキストだけでは読みが特定できない単語列に対する、実際の発話にあわせた読み方も合わせて記述することが望まれる。

NICT では、音声認識技術のための学習データであることを前提として、過剰なタグの使用は避けつつも、最低限の全言語共通のアノテーション仕様を定めて、書き起こしを作成している。その主な内容は以下である。

- 各言語の標準表記（正書法）で書き起こす。
- フィラーや感嘆詞はマークする（例：[<fil>/ あのー]）。
- 読みが特定できない、外来語表記や二桁以上の数字に関しては、標準表記とスペルアウトを併記する（例：[123/one two three]、[DX/ デジタルトランスフォーメーション] など）。
- 一文の終わりには終止符を打つ（通常句点を用いないタイ語等の言語も必要）。
- 言い誤りや訛り発音などは、併記する（例：[<unk>/ 実発音]、[正表記 / 実発音] など）。
- 無視できない非言語音はマークする（例：雑音 [<nos>]、笑い [<lgh>] など）。

実際には、言語特有の事情があり、ある程度はそれらを尊重した仕様となっているが、言語間で基本仕様を統一することは、モデル学習プログラムの言語共通部分のメンテナンスコストを下げる、という利点もある。

書き起こし精度は、認識モデル性能に直結するため、大量の音声データに対して、統一的にアノテーション仕様に従った記載がなされているかが極めて重要となるが、それには検査方法論上の問題も伴い [2]、作業コストとのバランスが求められる。また、GCP2025 においては、自由会話も対象としなければならないため、例えば、自由会話では頻繁に出現する相槌をマークしたり、多発するフィラーや言い誤りに対する書き起こし方への規定をより明確化する必要があり、現在でもアノテーション仕様は継続して更新している。

3　NICT 多言語対訳辞書

上述までに、音響モデルのための「音声コーパス」について述べてきたが、音声認識にとっては、言語モデルのための「テキストコーパス」も、単語列の推定には重要な役割を果たす。本稿においては、紙幅の関係でこれについては詳しく述べないが、学習データとしてのコーパス内に存在しない語彙は原則出力されることはないため、音響的に似た別の語彙が出力され、いわゆる誤認識の原因の一つとなる。これは、音声認識の

*3　音声認識モデルのベースとなっている生活会話の音声コーパスと「フォーマル」や「インフォーマル」では、発話スタイルとしての差異がそれほど大きくないことが要因の一つと考えられる。

ための音声・テキストコーパスのみならず、機械翻訳のための対訳コーパスも、音声合成のための音声・テキストコーパス*4 も、コーパス内に出現する語彙数に限りがある、あるいは低頻度のために確率的に出力単語として選択されにくいため、語彙拡張のための単語辞書が必要となる。そこで NICT では音声翻訳のための多言語対訳辞書の整備も進めている。専門用語や固有名詞などは、別に辞書として作成して、それらにカテゴリ（クラス）と発音を付与して登録しておくことで、コーパス内に存在する同カテゴリ（例えば、「動植物クラス」としてのサクラやヒマワリ）の出現確率を利用して、コーパス内には存在しない、あるいは低頻度の単語（例えば、エノコログサ）が候補として選択されやすくなるようにしている。現在、GC15 言語の音声認識モデルについては、この仕組みを導入しており、GCP においては、国内の観光スポット名や、市町村名、医療用語、防災用語等、観光・生活ドメインの GC15 言語間対訳辞書として整備し、さらに GCP2025 においては、ビジネス用語、専門用語の対訳辞書を整備し、機械翻訳、音声合成の辞書としても共有し、音声翻訳全般の精度向上に役立てている。

4 むすび

本稿では、NICT で開発を進めている、多言語音声コーパスの概要と、その設計、開発の歴史などについて紹介した。開発にあたっては、その品質を確保するための様々な取組を行っており [2]、その甲斐もあって、GC15 言語については、数百～数千時間規模の音声コーパスの構築を達成しており、現在も GCP2025 の開発目標に向けて、精力的にコーパスの開発を進めている。その一方で、音声翻訳の技術も日進月歩であり、例えば、音響モデルと言語モデルの区別のない、入力音声と出力単語列を直接結ぶようなフレームワーク（end-to-end 音声認識）や、音声認識結果を機械翻訳に入力するのではなく、入力音声を直接、目標言語に翻訳するフレームワーク（end-to-end 音声翻訳）なども各国で研究開発が進められている。音声コーパス開発の課題としては、それらの新しいフレームワークにあわせて都度、適応的なコーパス設計をするだけでなく、様々な用途に利活用可能な、汎用的なコーパス設計も求められていると考えている。

＊4　音声認識のための音声コーパスが、多様な人による、多様なスタイルの大量の音声データから成るのに対して、音声合成のための音声コーパスは、一人のプロの発声者による、正確な発音の、一定量（音素列の網羅性は必要）の音声データから成る。

【参考文献】

1　今村 賢治，隅田 英一郎，“グローバルコミュニケーション計画のための多言語パラレルコーパス，”言語処理学会第 24 回年次大会発表資料集，pp.512-515，2018.

2　水上 悦雄，榎本 成悟，テオリン アクセル エリック，加藤 宏明，河井 恒，“多言語音声コーパスの人－機械品質検査手法，”言語処理学会第 24 回年次大会発表資料集，pp.817-820，2018.

水上 悦雄（みずかみ えつお）

ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター
先進的音声技術研究室
主任研究技術員
博士（理学）
音声言語コーパス、コミュニケーション科学、対話研究
【受賞歴】
2008 年　社会言語科学会 2007 年度徳川宗賢賞 萌芽賞

加藤 宏明（かとう ひろあき）

ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター
先進的音声技術研究室
主任研究員
博士（工学）
音声言語コーパス、聴覚・音声コミュニケーション
【受賞歴】
1995 年　日本音響学会 第 12 回粟屋潔学術奨励賞