3-2 「言葉の壁」から解放された万博体験の実現
3-2　A Language-Free Expo Experience for All

PAUL Michael　今村 賢治　王 暁林　東山 翔平　内山 将夫　藤本 雅清
岡本 拓磨　菊池 武文　塩飽 裕彦　香山 健太郎*1
PAUL Michael, IMAMURA Kenji, WANG Xiaolin, HIGASHIYAMA Shohei, UTIYAMA Masao, FUJIMOTO Masakiyo, OKAMOTO Takuma, KIKUCHI Takefumi, SHIAKU Hirohiko, and KAYAMA Kentaro

グローバルコミュニケーション計画 2025 に基づき、大阪・関西万博での活用を目指して、多言語翻訳技術の高度化及び高精度、高品質、低遅延の多言語同時通訳システムの実現のための研究開発に取り組んだ。

同時通訳技術としてはチャンク単位の入力分割や文脈処理、マルチエンジン翻訳技術等を、音声認識技術としてはハイブリッド型及び E2E 音声認識モデル等を、音声合成技術としてはニューラル音声合成モデル等を開発し実装するとともに、これらのためのコーパスを構築した。また、音声マルチスポット再生技術については、研究開発に加え、システムのコンパクト化や同時通訳技術と連携したシステム構築を行った。

そして、システムの出展や技術移転を行うとともに、民間企業とも連携して、実証実験を実施した。

これらの結果、民間企業によるサービスが万博で活用されるとともに、音声マルチスポット再生技術を用いたシステムも出展されるに至った。

In line with the Global Communication Plan 2025, we have pursued the development of multilingual translation technologies to realize a language-barrier-free experience at the Osaka-Kansai Expo. Our work included the advancement of simultaneous interpretation systems emphasizing high accuracy, quality, and low latency. Key achievements include chunk-based input segmentation, context-aware translation, and multi-engine machine translation. In speech recognition, we developed both hybrid and end-to-end models, while in speech synthesis, neural voice synthesis models were implemented. We also constructed dedicated corpora to support these systems and developed a compact, multi-spot voice playback system integrated with our simultaneous interpretation framework. Through demonstration deployments and collaboration with private companies, our technologies have led to real-world applications, with several services and systems—including those using multiple sound spot synthesis technology—set to be showcased at the Expo.

1　まえがき

情報通信研究機構（NICT）の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（ASTREC）では、世界の「言葉の壁」をなくし、グローバルで自由な交流を実現することを目的としたグローバルコミュニケーション計画に基づき、多言語音声翻訳技術の研究開発及び社会実装を推進してきた。2014 年に総務省が策定した「グローバルコミュニケーション計画」[1]（以下、「GC 計画 2020」という。）の取組により、NICT の多言語翻訳技術の研究開発及び社会実装が進展したことを受け、2020 年 3 月には、2025 年に向けた AI による「同時通訳」の実現など多言語翻訳技術の更なる高度化を推進する目的で、総務省施策グローバルコミュニケーション計画 2025 [2]（以下、「GC 計画 2025」という。）が発表された。

近年、訪日・在留外国人数は、コロナ禍により一時落ち込んだものの、ともに年々増加傾向である。訪日

＊1　**2** はパウル、今村、王、東山、内山が、**3** は藤本が、**4** は岡本が、**5** は岡本、菊池、塩飽が、**1**,**6**–**8** は主に香山が執筆した。

外国人数は2024年時点で、在留外国人数は2022年末時点でコロナ禍前の人数を超えており、今後も増加・多国籍化が見込まれる。さらに、コロナ禍を経て、人々の価値観や行動様式等が大きく変化し、デジタルシフト及び制度改革が加速してきている。それにより、移動を伴わないオンライン会議が増加し、グローバルな会議への参加機会も増加していると考えられる。

また、ASTRECでは、GC計画2020の下、東京2020オリンピック・パラリンピック競技大会を一つのマイルストーンとして多言語翻訳技術の研究開発及び社会実装を推進した。この取組により翻訳精度の向上や対応言語の拡大を実現し、多様な翻訳サービスが実用化・普及して、同大会で活用されるとともに行政手続・医療・交通・観光等の様々な分野で活用されている。一方で、本技術は発話開始から発話終了までを一区切りとした文章を翻訳する、いわゆる「逐次翻訳」であり、1対1での対面の短い対話の場面で有効に利用されているが、ビジネスや国際会議での議論の場面も含め、多言語での会議対応や十分なコミュニケーションが可能な「同時通訳」へのニーズも大きい。

そこで、GC計画2025では、多言語翻訳システムの更なる普及・発展を目指すことに加え、2025年日本国際博覧会（以下「大阪・関西万博」という。）に向けて、産学官連携により、同時通訳技術及びこれと様々な技術とを組み合わせたシステムを段階的に実現するとともに、各種見本市やパビリオン等での技術の利活用も通じて、社会実装を推進することが掲げられた。これを受け、多言語翻訳技術の更なる高度化によりAIによる「同時通訳」を実現するための研究開発を実施する、総務省のICT重点技術の研究開発プロジェクト「多言語翻訳技術の高度化に関する研究開発」（2020～2024年度）を民間企業5社とともに受託した。そして、さらに民間企業3社を加えた合計9団体による「総務省委託・多言語翻訳技術高度化推進コンソーシアム」を結成して研究開発及び社会実装に取り組み、その成果が大阪・関西万博で活用されるに至っている。

これらの活動と並行して、NICTでは、異なる方向に異なる音声を届ける「音声マルチスポット再生技術」の研究開発を実施している。本技術は、多言語の同時通訳結果を聴衆にタイムラグなく一斉に伝えることに適していることから、同時通訳技術と連携させた様々なシステムを構築し、視察時等のデモや展示会への出展、一般向け施設での実証実験を実施してきた。大阪・関西万博にも、本システムを出展した。

本稿では、これらの技術の研究開発、実証実験及び社会実装、そして大阪・関西万博での活用・出展について詳述する。

2　同時通訳技術

同時通訳は、話者の発言を聞きながら、ほぼリアルタイムで別の言語に変換する作業であり、通訳者は高度な集中力と瞬発力、そして両方の言語に関する深い知識が求められる。しかし、通訳者の負担が大きく、時間的制約、言語間の差異、そして専門的な知識や用語の壁などの問題が発生し、日常生活で気軽に利用することはできない。

機械翻訳はニューラル機械翻訳（NMT）[3]の登場により、翻訳品質が大幅に向上し、実用性も向上した。近年は、書き言葉については文書翻訳など、長い文章の翻訳が可能になっている。話し言葉についても、同時通訳のように、話されている途中でも翻訳を開始できるような技術が開発されている。

本節では、NICTで研究開発している同時通訳システム[*2]の翻訳部分について解説する。このシステムは、講演等の話し言葉を、話を中断させることなく、高品質かつ低遅延で翻訳するものであり、また、英語、中国語、アジア言語など、15言語（表1）に対応した多言語システムである。

遅延を少なく翻訳するためには、発話途中で翻訳を行えばよいが、一般的には、十分な文脈を聞かないうちに翻訳を行うと、翻訳品質は低下する。翻訳品質を落とさずに遅延を少なくするために、NICTの同時通訳システムでは、チャンクという単位で翻訳を行っている。これは、通訳者が同時通訳を行う際に処理する単位を模倣したもので、数単語程度の、文より短い単位である（図1）。

図2は、本システムの構成である。同時通訳器は、大きく入力分割、翻訳、後処理（要約等を含む）から成り立っており、それぞれの処理は言語コーパスから学習したモデルを使用している。以下、**2.1**では、モデル学習のための言語コーパスについて述べる。続く**2.2**

表1　対象15言語

コード	言語	コード	言語
ja	日本語	mn	モンゴル語
en	英語	my	ミャンマー語
es	スペイン語	ne	ネパール語
fp	フィリピン語	pt_BR	ブラジルポルトガル語
fr	フランス語	th	タイ語
id	インドネシア語	vi	ベトナム語
km	クメール語	zh	中国語
ko	韓国語		

＊2　本稿で扱う技術は、正確には同時通訳（simultaneous interpretation）というより、同時翻訳（simultaneous translation）と呼ぶ方が正しいが、本稿では多くの人に親和性が高い同時通訳という用語で統一する。

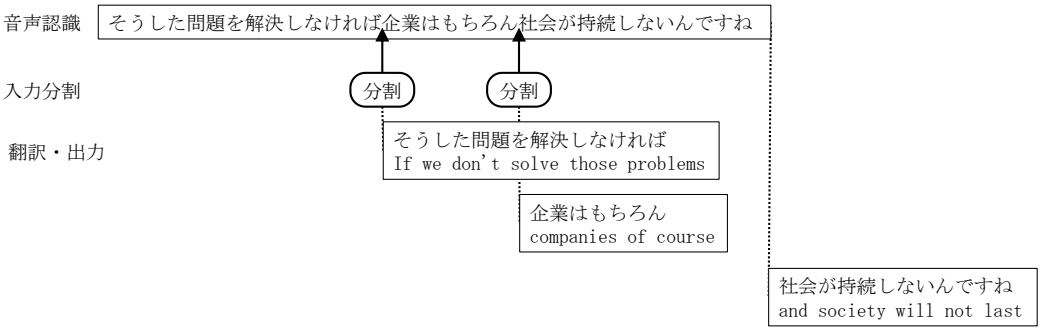


図 1　チャンク分割型同時通訳のイメージ

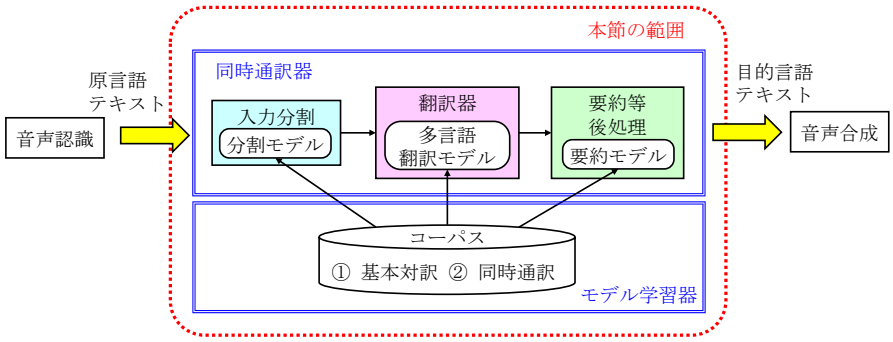


図 2　同時通訳の構成

では、同時通訳システムで使用されている多言語対応の入力分割処理とマルチエンジン翻訳、会話の流れを理解するための文脈処理を述べる。**2.3** では、同時通訳システムのユーザビリティ向上方法について述べる。

2.1　言語コーパス

まず、基本的な翻訳品質を確保するため、GCP コーパスと呼ぶ多言語対訳コーパスを使用した [4][5]。これは、グローバルコミュニケーション計画に基づいて模擬対話を中心に作成したコーパスである。主たるターゲットとして、来日する外国人との音声翻訳を実現するために開発したため、医療、防災、ショッピング、観光、その他のドメインをカバーするように設計されている。多言語化は、日本語を外国語に翻訳することで実現した。

GCP コーパスは、以下の特徴を持つ。

- 模擬対話を中心に構成しており話し言葉を多く含んでいる。
- しかし、人手で作文したものであるため音声言語の特徴であるフィラーや言い直しを含んでいない。
- 発話者やドメインなどの追加情報を含んでいる。

これにより、話し言葉翻訳の基本的品質を確保した。

次に、同時通訳に適した翻訳モデル・入力分割モデルの学習・評価を可能にするため、日本語を中心とした多言語の同時通訳コーパスを構築した。対象とした翻訳方向は、日本語→表 1 の日本語を除く 14 言語と、7 言語 (en,fp,fr,ko,th,vi,zh) →日本語である。

同時通訳コーパスを構築する方法の一つに、人間の話者・通訳者の音声を書き起こす方法も考えられるが、コーパス構築のための時間やコスト、翻訳モデルへの学習効果も考慮し、本コーパスの構築方法として、次の方法を採用した [7]。

(1) 通訳対象の実際の講演・会話の書き起こしテキストまたは講演・会話を模したシナリオテキストとする。なお、同時通訳システムの中心的な利用需要を想定し、会話はビジネス会話に限定した。講演は、様々な分野・話題についての講演を含んでいる。

(2) 人間の通訳者が、原文が読み上げられる状況を想定しながら、同時通訳において一度に翻訳を行う意味的なまとまり (チャンク) に原文を分割し、チャンクごとに順送り方式で原文を翻訳先言語へ翻訳した同時通訳文を作成する。その際、原文中の語句に対し、原文の内容を極力 100 % カバーするように訳出することとする。

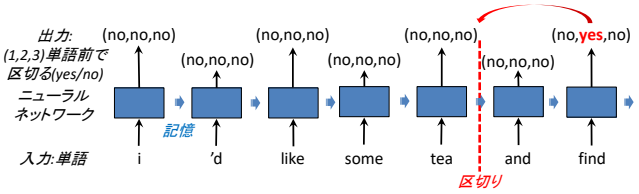


図３　可変数の後続単語を考慮する区切り推定

表２　文脈処理のためのコンテクストタグの種類

場面種別	コンテクストタグ	対象言語
話者	日本人、外国人	全言語
分野	ビジネス、医療、防災、教育、自治体、観光、ショッピング、スポーツ、交通機関、その他	全言語
省略	わたし、あなた	日本語
性別	女性、男性	タイ語

表３　翻訳エンジンの種類

種類	エンジン	翻訳モデル
ニューラル機械翻訳	汎用NT(GPMT)	・５４言語ペア対応の多言語翻訳モデル
ニューラル機械翻訳	適応NT(TSEG)	・GPMTモデルに基づいたドメイン適応訓練モデル
ニューラル機械翻訳	汎用UM(UNIV)	・３２ｘ３１言語の翻訳を1モデル
大規模言語モデル 機械翻訳	汎用UM(RWKV)	・多言語対応の文脈理解を伴う翻訳モデル

2.2　同時通訳器

本節では、同時通訳器本体の仕組みについて解説する。**2.2.1** ではオンライン入力分割の手法、**2.2.2** では同時通訳向けの文脈処理、**2.2.3** では翻訳品質を最適化するマルチエンジン翻訳技術について述べる。

2.2.1　入力の分割

図２で示しているように、自動化された同時通訳システムでは、通常、自動音声認識と機械翻訳という、2つの基本的な自然言語処理技術を統合する必要がある。自動音声認識が生成する文字起こしには、文の区切りが存在しない一方で、機械翻訳が正確に機能するためには、文単位やチャンク単位での入力が渡される必要がある。これを解決するため、オンライン入力分割を用いる。オンライン入力分割とは、自動音声認識が出力する文字列をリアルタイムで文またはチャンクに分割することで、自動音声認識と機械翻訳の間にあるギャップを自然に埋める手法である。

オンライン入力分割問題の難しさは、主に２つの要因に起因する [9][10]。一つめは、自動音声認識が出力する句読点のない単語列に対して、適切に入力を分割することは容易ではないこと、二つめは、現在の位置に文の区切りがあるかどうかは後続の単語に依存し、最適な待機単語数は文脈によって異なるということである。

一つめの課題に対しては、再帰ニューラルネットワークという、言語モデルで使用されるアーキテクチャで対処した。このモデルはオンラインで動作し、各時刻において単語を入力として受け取り、文またはチャンク分割のための区切り情報を返す。また、ネットワークのメカニズムにより、現時刻までの入力を記憶している。

二つめの課題に対しては、ニューラルネットワークモデルを訓練する際、後続単語を考慮して、「$\{1,2,3\}$ 語前に区切りがあったか」という３つの区切り判定を同時に学習させることで解決した [10]。図３の例では、「find」という単語を入力した後、最初の「Yes」信号が出力される。これは２語前に対応しており、チャンクが２語前、つまり「tea」という単語の後で終了したと判断されたことになる。

2.2.2　文脈処理

翻訳処理にはニューラルネットワークに基づく方式を採用し、対訳文を学習することによって翻訳を可能にしている。この方式では、入力に「タグ」という形で情報を付加することで、同じ入力の意味解釈を切り替え、様々な出力バリエーションを生成することができる。NICT は、この機能を文脈処理に適用した。機械翻訳における文脈処理とは、言葉の前後関係や文の流れを理解し、会話の場面（誰がどんなトピックについて話すか）と今までの会話での発話内容を把握することで、より自然な翻訳を目指す技術である。

同時通訳システムで使用されている翻訳モデルでは会話の場面を考慮できるように、表２のコンテクストタグを学習リソースに付与し、標準の翻訳モデルの適応訓練を行った。これにより、指定された文脈に合わせた翻訳ができるようになった。

文脈処理に関しては、会話の場面以外にも、今までの会話での発話内容を把握することが重要である。機械翻訳では、長らく１文単位の翻訳方式が主流であったが、近年のニューラル機械翻訳では、複数の文にまたがる文脈の利用が可能になった [6]。同時通訳システムでは、過去の会話とその翻訳結果を記憶して、次に入力された発話を翻訳する時、文外文脈として供給する。この情報は、入力された文だけで意味が曖昧な場合、会話の流れを参照して適切な表現を補完する。

2.2.3　マルチエンジン翻訳

マルチエンジン対応の翻訳技術は、複数の翻訳エンジンを並列に使用し、オンラインで適切なエンジンを選択することで、より自然で高品質な翻訳を可能にする技術である。表３に同時通訳システムで使用可能な翻訳エンジンとその翻訳モデルを示す。

汎用 NT（GPMT）とは基本対訳コーパスで学習されている高品質な NMT モデル、適応 NT（TSEG）は GPMT 翻訳モデルの学習リソースに同時通訳用コー

パスを加え、かつ、2.2.2 のコンテクストタグ情報を付加して適応訓練された、文脈処理可能な翻訳モデル、汎用 UM（UNIV）の翻訳モデルは複数の言語ペアの学習リソースを組み合わせて、1 つの NMT を学習させたユニバーサル翻訳モデルである。

ニューラル機械翻訳の技術に加え、大規模言語モデル（LLM）に基づく機械翻訳手法にも多くの注目が集まっている [8]。汎用的な言語理解と文脈処理に優れ、自然で柔軟な訳文生成が可能という特徴を持つ。同時通訳システムでは、RWKV の LLM モデルを使用している。

複数の翻訳エンジンから、最も自然で文脈にあっている翻訳結果を選択するため、入力発話と折り返し翻訳結果の類似度を算出し、最も類似するエンジンの翻訳結果を採用している。折り返し翻訳とは、翻訳結果を入力言語に再翻訳する技術で、各エンジンの逆方向の翻訳モデルを用いて、翻訳された発話を入力言語に翻訳している。類似度は、入力発話と折り返し翻訳結果の単語列をそれぞれベクトル化し、両者のコサイン類似度で算出している。

2.3 自動同時通訳のユーザビリティ

同時通訳システムのレスポンスタイム（発話から訳出までの時間）が遅くなるほど、会話の流れが進まないため、自動同時通訳の使いやすさ（ユーザビリティ）が低下する。

リアルタイム応答を提供する実用的な同時通訳システムを実現するために、NICT で研究開発している同時通訳システムは、文単位と同時にチャンク単位での翻訳処理も行う。

文単位の分割モデルをベースに、同時通訳用コーパスのチャンク区切り情報を用いて、文単位より短い単位のチャンクを分割するように各言語の分割モデルの適応訓練を行うことにより、低遅延かつ実用的な精度の自動同時通訳を実現した。

しかし、チャンク単位の情報量は文単位より少なく、チャンク単位での翻訳品質は文単位の翻訳品質よりも低いため、チャンク単位と文単位の翻訳処理を並列に実行する本システムでは、チャンク区切りと文区切りが一致している場合、文全体の全体の訳し直し（連続するチャンク翻訳結果を文単位の翻訳結果に置き換える）も行っている。

このような技術により、同時通訳を開始するまでの時間間隔を短くし、早いタイミングで聴き手に情報を与えるとともに、自動で聞き手にわかりやすい最適な通訳結果を出力し続ける同時通訳システムを実現した。

3 音声認識技術

音声認識（ASR: Automatic Speech Recognition）技術は、入力された音声を文字（テキスト）にメディア変換する技術であり、長年にわたって基礎技術及び応用技術の研究開発が推進され、研究成果が積み上げられてきた。近年では深層学習技術（Deep learning）[11]–[13] の台頭により、その性能が飛躍的に改善し、直近の 10 年程の間に音声認識技術の実用化が極めて急速に進展した。現在では、スマートフォンやタブレット等のモバイルデバイス、さらにはスマートスピーカに代表されるホームデバイスにおける音声検索、音声翻訳、音声対話等、我々の日常生活に密接に関連する多種多様な音声認識サービスが展開されており、現代における生活インフラの一つとしてその地位を確立しつつある。また、音声認識技術を利用した映像メディアへの自動字幕付与も急速に進んでおり、地上波放送や BS/CS 放送等の伝統的な放送メディアのみならず、動画投稿サイト等のネットメディアにおいても欠かせない技術となりつつある。

3.1 音声認識技術の変遷

音声認識技術の研究開発には、40 ～ 50 年以上にわたる歴史があり、今日に至るまでに様々な技術的変遷やパラダイム・シフトが生じてきた。その主な流れは以下のとおりであり、現在では End-to-End 音声認識と大規模言語モデルの応用技術が主流となっている。

- 様々なパターンマッチング手法の適用（1990 年代以前）[14]
- 機械学習技術の発達と統計的音声認識の確立（1990 年代～ 2010 年代）[15]–[17]
- 深層学習技術の台頭とハイブリッド型音声認識への移行（2010 年頃～ 2015 年頃）[18]–[27]
- End-to-End（E2E）音声認識への発展（2015 年頃以降）[28]–[33]
- 大規模言語モデル（LLM: Large Language Model）の登場とその応用（2020 年頃以降）[34]–[40]

3.2 音声認識の定式化と構成要素

音声認識を定式化するにあたり、まず音声認識器の入出力系列を以下のように定義する。

- 入力系列 $\boldsymbol{O} = \{\boldsymbol{o_0}, \dots, \boldsymbol{o_t}, \dots, \boldsymbol{o_T}\}$：音声信号波形、もしくは音声信号波形を音響分析して得られた、対数メル周波数スペクトルやメル周波数ケプストラム係数 [17] 等の特徴量ベクトルの系列（Tは入力の系列総数）
- 出力系列 $\boldsymbol{W} = \{\boldsymbol{w_0}, \dots, \boldsymbol{w_n}, \dots, \boldsymbol{w_N}\}$：入力系列に対応する単語、文字、サブワード等のテキスト系列

（Nは出力の系列総数）

上記の定義に基づき、入力系列が与えられた時の音声認識器が出力すべきテキスト系列$\widehat{\boldsymbol{W}}$は、次式により得られる [16]。

$$\widehat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W}} p(\boldsymbol{W}|\boldsymbol{O}) \tag{1}$$

式 (1) において$p(\boldsymbol{W}|\boldsymbol{O})$は音声認識モデルと呼ばれており、統計的音声認識以降の音声認識技術では大量の音声データを用いて機械学習技術もしくは深層学習技術により、$p(\boldsymbol{W}|\boldsymbol{O})$を構成する精密なパラメータ群を推定する。

統計的音声認識及びハイブリッド型音声認識では、式 (1) の右辺に音素（音韻を弁別する上での最小単位）の系列である$\boldsymbol{S}$という中間表現を導入して、次式の様な近似を行う [16]。

$$\widehat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W}} \sum_{\boldsymbol{S}} p(\boldsymbol{O}|\boldsymbol{S})\, p(\boldsymbol{S}|\boldsymbol{W})\, p(\boldsymbol{W}) \tag{2}$$

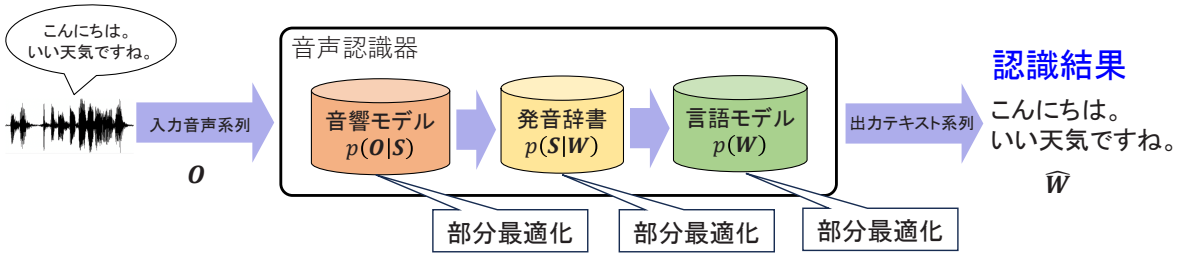
式 (2) により音声認識モデル$p(\boldsymbol{W}|\boldsymbol{O})$は、図 4 (a) に示すように音響モデル（入力系列$\boldsymbol{O}$を音素系列$\boldsymbol{S}$に変換）、発音辞書（音素系列$\boldsymbol{S}$を単語w_nに変換）、言語モデル（単語w_nを単語系列$\boldsymbol{W}$に変換）の 3 つの機能ブロックに分解される。これらの機能ブロックを個別に部分最適化して接続することにより音声認識器が構築される。

一方、図 4 (b) に示すように音声認識モデル$p(\boldsymbol{W}|\boldsymbol{O})$を機能ブロックに分割せずに、一つのモデルのみで音声認識を行う方法が2015年頃から注目され始めた。この方法は、音声認識器の端（入力側 End）から端（出力側 End）までを一つのニューラルネットワークを用いて構成することから、End-to-End（E2E）音声認識 [28]–[33] と呼ばれている。E2E 音声認識は、図 4 (a) のような機能ブロック分割を行わず、音声認識器全体を一つのニューラルネットワークで構成するため、システムの全体最適化が容易である。またシステム構成としてもシンプルになる。このような利点から、現在では全世界的に E2E 音声認識の研究開発が極めて活発に行われている。

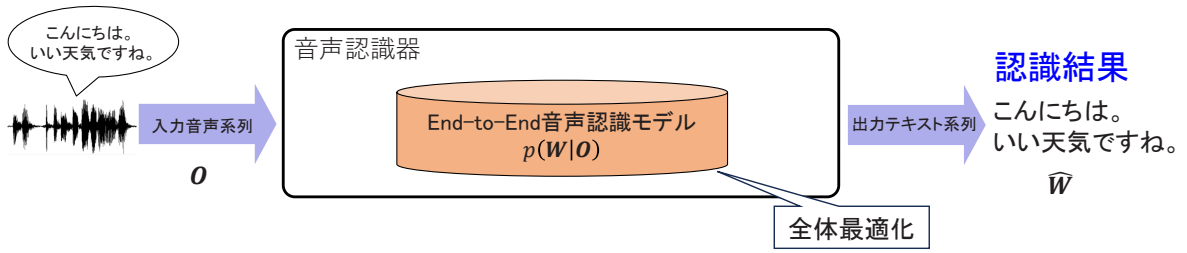
3.3　End-to-End 音声認識

E2E 音声認識は多くの場合、図 2 に示すようなエンコーダ・デコーダ（符号器・復号器）ネットワーク [41] にて構成される。

図 5 においてエンコーダは、入力された特徴量ベクトル系列$\boldsymbol{O}$を音声認識のための適切な音響特徴表現系



(a) 統計的音声認識もしくはハイブリッド型音声認識の構造



(b) End-to-End音声認識の構造

図 4　音声認識器の構成図

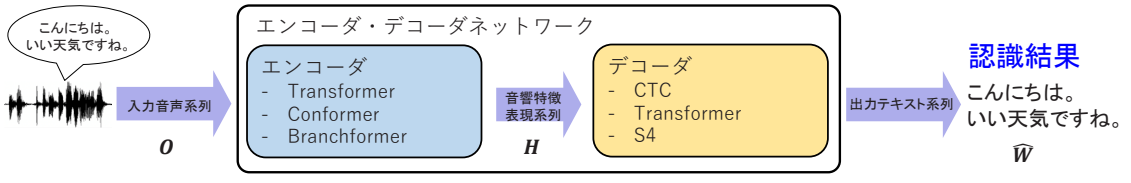


図 5　エンコーダー・デコーダーネットワークの概要

列$\boldsymbol{H}=\{\boldsymbol{h}_0,\ldots,\boldsymbol{h}_t,\ldots,\boldsymbol{h}_T\}$に変換する。エンコーダには、Bi-directional Long-Short Term Memory（B-LSTM）[42] 等 の Recurrent Neural Network（RNN） や、Transformer [43] 等のモデルが用いられる。デコーダは、エンコーダ出力である音響特徴表現系列$\boldsymbol{H}$に基づいて、入力音声に対応するテキスト列（トークン列）を生成して出力する。デコーダにはTransformer の他にConnectionist Temporal Classification（CTC）[44]、Structured State Space Sequence（S4）モデル等が用いられる [45]。

E2E 音声認識のもう一つの実装方法として、RNN-Transducer [46] がある。エンコーダー・デコーダーネットワークが入力特徴量ベクトル系列$\boldsymbol{O}$を一定時間以上観測（観測した時間分の処理遅延が発生）して処理する必要があるのに対して、RNN-Transducer はフレーム同期のリアルタイム処理が可能となっている。

3.4 LLM を活用した音声認識

E2E 音声認識の更なる発展として、LLM を活用した音声認識手法（LLM-based ASR）の研究が活発化している [34]–[40]。LLM-based ASR では、図 6 のようにE2E 音声認識モデルのエンコーダ（自己教師あり学習による事前学習モデル [47][48] 等）と、テキスト生成を担う Decoder only LLM（GPT: Generative Pre-trained Transformer [49][50] 等）との間に、アダプターと呼ばれる小規模な Feed Forward Network を挿入する構成が一般的である [40]。アダプターは、エンコーダが出力する音響特徴量を、LLM が処理可能なテキスト埋め込み表現へ変換する役割を担う。また、LLM への入力には、アダプターに加えて入力音声系列と対応するテキスト系列とのアライメント（時間的対応付け）が必要となる [51]–[53]。多くの場合、エンコーダ及び LLM のパラメータは固定され、アダプターとアライメント処理のみを学習対象とする。また、LLM には音声認識を行うように誘導するプロンプトを与えることが多い [38]。

3.5 NICT における取組

NICT では 2014 年から総務省主導で遂行された GC 計画 2020 [1] の目標である「音声翻訳技術の社会実装」、2021 年から実施された GC 計画 2025 [2] の目標である「多言語同時通訳技術の社会実装」を達成するため、高精度、低遅延で動作する多言語音声認識の研究開発を推進している。2025 年 8 月の時点で、アジア言語を中心として以下の 22 言語のハイブリッド型音声認識モデル及び E2E 音声認識モデルを開発し、音声翻訳アプリ VoiceTra™ [54] 等に実装している。また、グローバルサウス対応のため、2025 年度内にタミル語、ベンガル語の音声認識モデルの開発と実装を実施する予定である。

- 主要 4 言語：日本語、英語、中国語（簡体字、繁体字）、韓国語
- アジア言語：インドネシア語、ベトナム語、タイ語、ミャンマー語、フィリピン語、クメール語、ネパール語、モンゴル語、ヒンディー語、
- ヨーロッパ言語：スペイン語、フランス語、ドイツ語、イタリア語、ポーランド語、ロシア語、ウクライナ語
- その他：アラビア語、ブラジルポルトガル語

各言語の音声認識において、主要 4 言語については人間レベルの音声認識性能を達成しており、他の言語においても実用レベル以上の音声認識を達成している*3。

このほかにも音声認識を用いた映像メディアへの自動字幕付与システム [55] や、自動会議録生成システムの開発を推進している。自動会議録生成システムにおいては、音声認識のみを用いるのではなく、言語識別・話者識別技術 [56] と組み合わせることにより詳細な講演録、会議録を自動生成することが可能となっている。

*3 音声認識の性能基準は以下の定義に従っている。
- 人間レベル：音声認識結果を問題なく読んで理解できる
- 実用レベル：軽微な誤りがあるが音声認識結果を読んで十分に理解できる

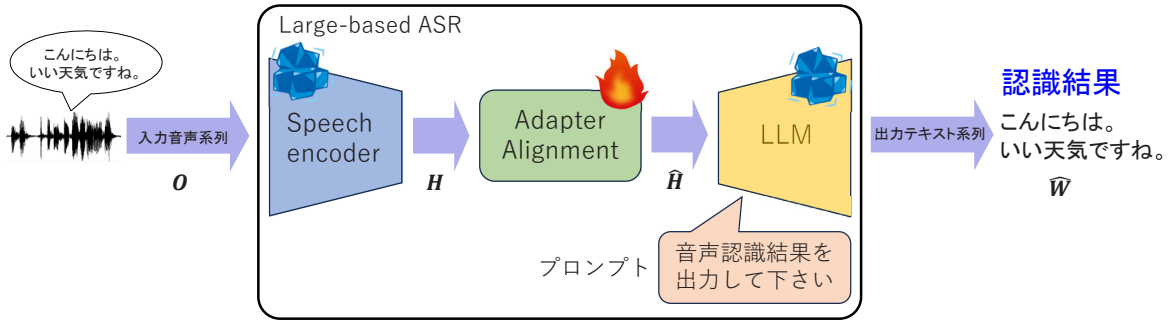


図 6 LLM-based ASR の概要

4　音声合成技術

4.1　テキスト音声合成

入力テキストを機械が自然な音声で読み上げるテキスト音声合成（Text-to-speech synthesis: TTS）は、音声コミュニケーションにおいて重要な技術の 1 つである。近年では、自動音声ガイダンス、駅等での自動アナウンス、スマートスピーカ、カーナビ、対話ロボット、生成 AI 等の様々な場面で使われており、日常的に使われる技術となっている。万博会場においても、様々な言語を話す人々が多数集まるため、**2** 及び **3** で紹介した機械翻訳及び自動音声認識とテキスト音声合成を組み合わせることにより多言語同時通訳となり、日本語話者の音声を多言語音声として伝えることが可能となる。

多言語同時通訳への応用を含めてテキスト音声合成に求められる要件は大きく 2 つある。1 つは、肉声感のある自然な音声で合成できること（高品質合成）、もう 1 つはテキスト入力を受けて即座に合成できること（高速合成）である。

NICT では、言葉の壁を超えたコミュニケーション技術を目指して、これまでに多言語テキスト音声合成の研究開発に取り組んできた [57][58]。2013 年頃からニューラルネットワークの導入により音声合成の品質は飛躍的に改善し、現在では自然音声と匹敵する品質にまで至っている [59][60]。

本節では、2023 年以降に NICT にて研究開発した高速・高品質多言語音声合成技術について紹介する。

4.2　NICT の研究開発した高速・高品質ニューラル音声合成技術

前述のとおりテキスト音声合成の音質は、ニューラルネット技術の導入により近年飛躍的に向上し肉声に匹敵するほどとなったが、膨大な計算量が大きな課題であり、ネットワークに接続されていないスマートフォンでの合成は到底不可能であるという課題があった。また、多言語同時通訳においては、話者の発話終了を待たずに次々と翻訳音声を出力する必要があるため、音声認識や機械翻訳と同様、テキスト音声合成の更なる高速化が求められる。そこで、合成品質を維持しつつ、高速生成を実現するためのニューラル音声合成モデルの研究開発に取り組んできた。

テキスト音声合成モデルは、入力テキストを中間特徴量へと変換する「音響モデル」と、中間特徴量を音声波形へと変換する「波形生成モデル」から構成される [59][60]。

ニューラル音声合成の「音響モデル」では、機械翻訳の分野や、音声認識や ChatGPT を始めとする大規模言語モデル等にも幅広く使われている Transformer [61] が主流であるのに対して [62]、近年画像識別の分野で新たに使われ始めた高速・高性能なニューラルネット ConvNeXt [63] を音響モデルに導入し、従来方式と比較して、品質を損なわず 3 倍の高速化を達成した [64]。さらに、音素スキップ接続を提案し、話速変換に頑健な音響モデルを実現している [65][66]。

また、肉声に匹敵する音声を合成可能な従来の「波形生成モデル」（HiFi-GAN [67]）を発展させる形で、信号処理方式 [68]-[70] を学習可能なニューラルネットとして表現するモデル（MS-HiFi-GAN）を 2021 年に導入し、合成品質を損なわず合成速度を 2 倍にすることに成功した [71]。そして、2023 年には同モデル（MS-HiFi-GAN）を更に高速化するモデル（MS-FC-HiFi-GAN）の開発に成功し、従来方式（HiFi-GAN）と比較して、品質を損なわず合成速度を 4 倍にすることを実現した [72][73]。

これらの成果の集大成として、上記で開発した「音響モデル（Transformer 型エンコーダ + ConvNeXt 型デコーダ）」と「波形生成モデル（MS-FC-HiFi-GAN）[73]」を用いた新しい高速・高品質なニューラル音声合成モデルを開発した（図 7）。これにより、CPU コア一つで 1 秒の音声をわずか 0.1 秒で高速合成することが可能となった（既存モデルの約 8 倍の速さ）。さらに、「波形生成モデル」のみを逐次合成する方式を実装することで（図 8）、合成品質を一切損ねることなく、ネットワークに接続されていないミドルレンジスマート

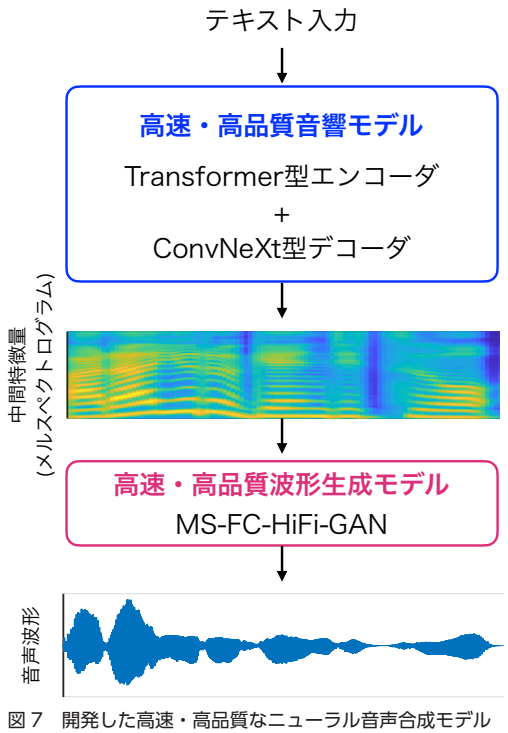


図 7　開発した高速・高品質なニューラル音声合成モデル

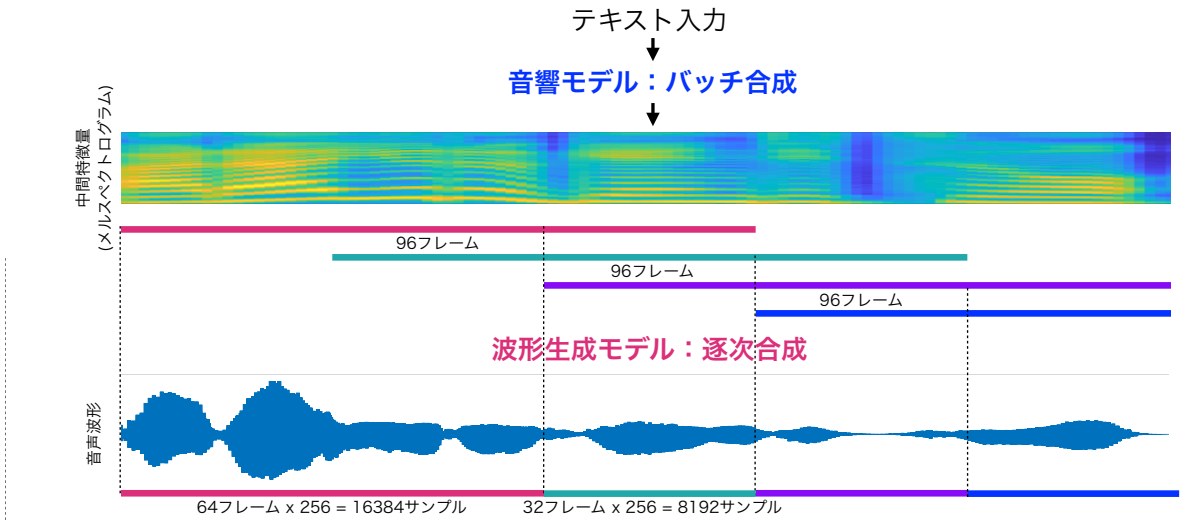


図 8　波形生成モデルのみを逐次合成することにより待ち時間短縮を実現

フォン端末上でも、テキスト入力からわずか 0.5 秒の高速生成が可能となった [74][75]。これにより、これまでのサーバ経由での合成が不要となり、インターネット通信を必要とせず、通信不要でスマートフォンや PC 等での高品質ニューラル音声合成が可能となった。また、逐次合成処理により、多言語同時通訳においても翻訳テキストを即座に合成することが可能となった。なお、日本語を入力とし、出力を英語、中国語、韓国語とする 1 入力 3 出力の多言語同時通訳デモシステムはノート PC1 台での実装が可能である [76]。

現在では、**3** で紹介した音声認識と同様、VoiceTra の 22 言語（日本語、英語、中国語、韓国語、タイ語、フランス語、インドネシア語、ベトナム語、スペイン語、ミャンマー語、フィリピン語、ブラジルポルトガル語、クメール語、ネパール語、モンゴル語、アラビア語、イタリア語、ウクライナ語、ドイツ語、ヒンディー語、ロシア語、ポーランド語）の音声には、この音声合成技術を導入し、一般公開している。

そして 2025 年 9 月 16 ～ 22 日には、大阪･関西万博における Future Life Experience にて次節で紹介する音声マルチスポット再生技術を出展し、出力合成音声には開発したニューラル音声合成モデルが用いられた。

今後は、商用ライセンスを通して、多言語音声翻訳やカーナビを始めとするスマートフォンアプリ等への社会実装を行う。

4.3　NICT における近年の研究開発

音声合成の研究を加速させるため、2023 年 8 月に高品質日英音声合成用コーパス HiFi-CAPTAIN を公開した [77]。これまでは 1 モデルにつき 1 話者の合成であったが、現在は 1 モデルで複数の話者や言語を合成可能なモデル [78][79]、入力話者の音声の発話内容を保持したまま別の話者の音声へと変換する声質変換モデル [80] 及び感情音声合成モデルの研究開発に着手している。また、ニューラルネットに基づく波形生成モデルはデータ駆動型であるため学習データ範囲外の声の高さの音声等では合成品質が下がってしまう問題に対して、従来の信号処理方式と組み合わせることにより学習データ範囲外の音声を外装可能なニューラル波形生成モデルを提案している [81]–[83]。さらに、日本語や中国語などのピッチアクセント言語において、自然言語処理 =（大規模）言語モデルを用いたアクセント辞書不要な音声合成モデルの検討も行っている [78][84]。加えて、合成音声の品質を人手ではなくニューラルネットを用いて自動評価を行う研究 [85] や、合成音声技術を安心・安全に提供するための研究開発にも取り組んでいる。

5　音声マルチスポット再生技術

5.1　多数のスピーカを用いた音の局所再生・マルチスポット再生

大阪・関西万博では、様々な言語を話す人々が多数集まるため、展示に際しては多言語対応が必須となるが、**2**, **3**, **4** で紹介した多言語同時通訳技術（音声認識＋機械翻訳＋合成音声）[86] を活用することにより、日本語話者の発話を多言語に同時通訳された音声を合成し、各言語を母語とする人々に対しても展示内容を詳細に伝えることが可能となった。

一方で、多言語同時通訳された合成音声の再生方法については課題が残る。通常、音は全方向に広がるため、これらの合成音声を通常のスピーカで同時再生（単純再生）すると、図 9（a）のように音が混ざり合ってしまい、聴取者は目的音を聞き取ることが困難にな

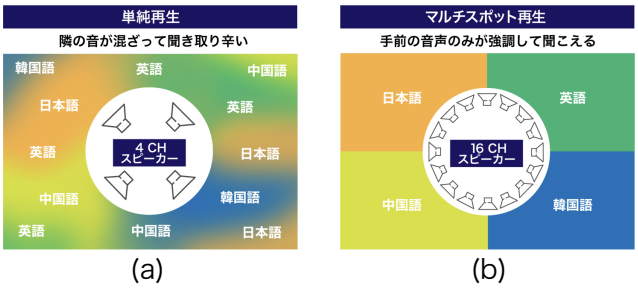


図 9　単純再生とマルチスポット再生

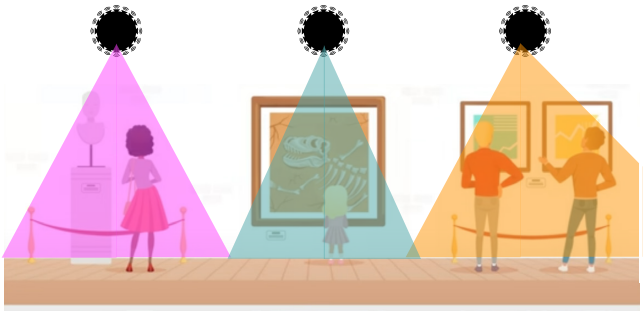


図 11　美術館での応用例

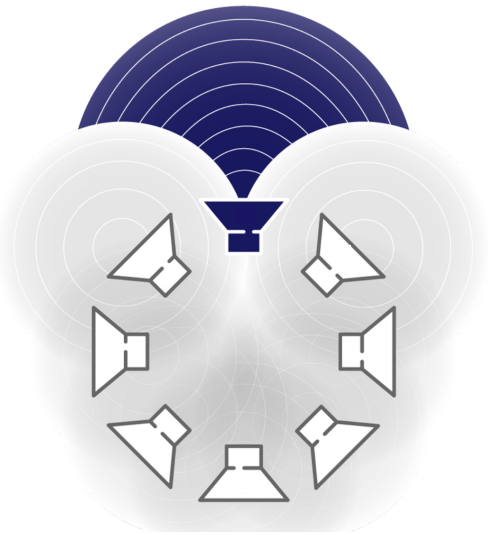


図 10　局所再生の原理

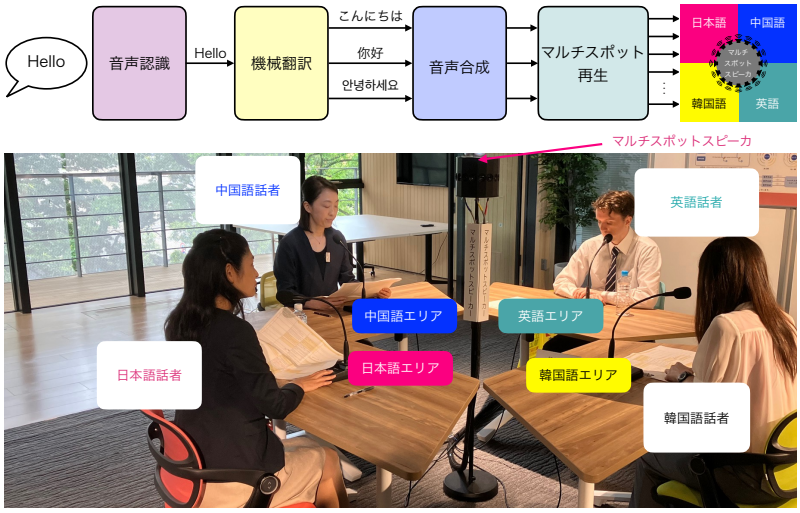


図 12　音声マルチスポット再生技術と多言語同時通訳技術を組み合わせた日英中韓 4 言語会議システム（2023 年 6 月　NICT オープンハウス 2023）

る。また、駅や空港で多言語対応を行う場合と同様に、日本語→英語→中国語のように順番にアナウンスすると、それらの音が混ざることはないが、言語の数だけ再生時間が必要となり、多言語同時通訳を実現するメリットが薄れてしまう。本節で紹介する「音声マルチスポット再生技術」（以下、本節において「本技術」という。）は、これらの問題を解決し、多言語で同時通訳された合成音声を、その言語を必要とする人々に同時に提供することができる技術である。

本技術は、多数のスピーカを用いて、音が聞こえる空間と、その音が聞こえない（または別の音が聞こえる）空間を創出することを可能にする [87]。本技術の原理（ここでは端的な紹介に留めており、正確なものではない。）としては、ノイズキャンセルの原理と同様に、目的方向以外の方向に広がった音を、別のスピーカから再生された音で打ち消すことにより、目的方向にのみ目的音を届ける「局所再生」が可能となり（図 10）、この局所再生を方向ごとに重ね合わせることで、異なる方向に異なる音を届ける「マルチスポット再生」を実現している（図 9（b））。なお、多数のスピーカを用いた音の局所再生技術としては、音響コントラスト法 [88]、音圧マッチング法 [89]、これらを組み合わせた方式 [90]、振幅マッチング法 [91] 等が提案されているが、NICT では直線や円形に配置した多数のスピーカを用いた空間フーリエ変換 [92] に基づく局所再生方式及びマルチスポット再生方式 [93][94] を提案し、音響コントラスト法や音圧マッチング法よりも高精度な制御性能を達成している。

本節の冒頭で触れたように、本技術と多言語同時通訳技術を融合させることで、日本語話者の発話が即座に翻訳され、英語、中国語、韓国語等の合成音声を別々の方向に同時に届けることが可能となる。言い換えれば「音を時間ではなく空間で分ける」ことが可能であり、これにより、複数の音声が混ざり合って聞き取りづらくなる問題やそれらを順番にアナウンスした際に時間的なロスが生じる問題を解決することができる。また、本技術は多言語対応の用途のみならず、観光施設やエンターテインメント施設、美術館等での同時解説音提示（図 11）、車内における活用等も期待できる。さらに、2025 年 3 月には世界防災フォーラムに出展し、緊急時における多言語での避難誘導への応用も期待されている。

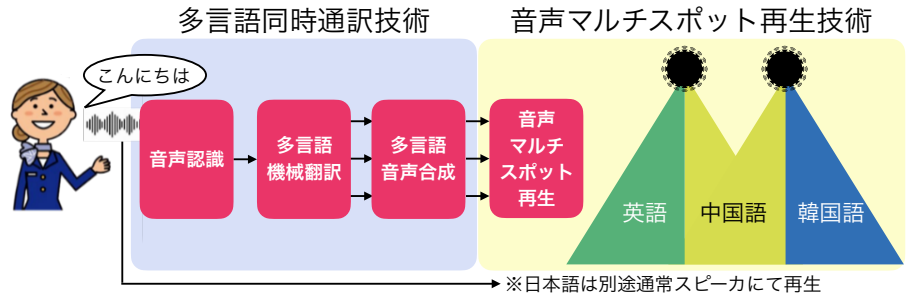


図 13　多言語同時通訳と音声マルチスポット再生技術を用いた実証実験（2025 年 1 月　海遊館での実証実験）

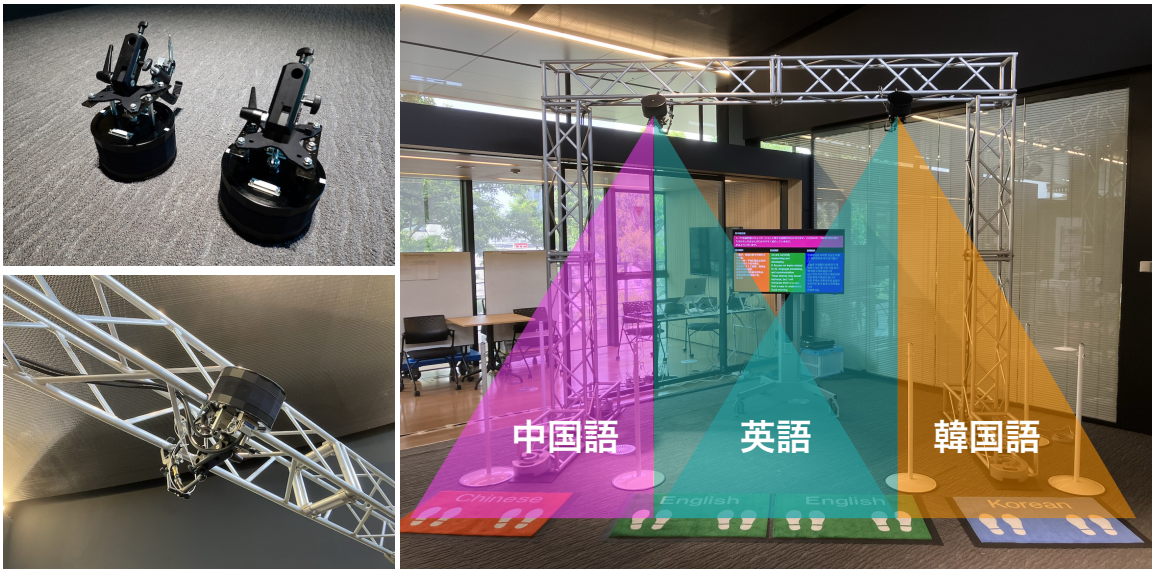


図 14　円形スピーカを用いた直線移動型マルチスポット再生デモシステム（2025 年 6 月 NICT オープンハウス 2025）

5.2　社会実装に向けたこれまでの取組

本技術の社会実装に向けた取組としては、第一に、技術紹介を行うホームページ [95] や動画 [96][97] を公開し、社会実装を目指すに当たってパートナーとなる企業の探索を行っていることが挙げられる。第二に、スーツケース 1 つで持ち運び可能な 16 チャネル円形スピーカを用いたデモシステムを実装し [98]、CEATEC 等の大規模展示会におけるデモ展示や水族館等における実証実験を実施しており、本技術の有効性の確認及びフィードバックのヒアリングを行っていることが挙げられる。本節では、その代表例について紹介する。

2023 年 6 月の NICT オープンハウス 2023 では、本技術と多言語同時通訳技術を組み合わせた日英中韓 4 言語会議システムのデモ展示を行った（図 12）。これは、日本語話者の発話が英語、中国語、韓国語に同時通訳され、それらの合成音声が音声マルチスポット再生により各言語の話者に届くというデモ展示である。これにより、各言語の話者は自身の母語で話し、通訳された母語を聞くことにより異なる母語の話者と会議を行うことが可能となることを提示した。

2025 年 1 月には、NTT データ カスタマサービス株式会社により、本技術を用いた実証実験（総務省委託）が大阪市の水族館「海遊館」で実施された [99]。本実証実験では、解説員が水生動物の様子を日本語で解説し、同時通訳された英語、中国語、韓国語の合成音声を、本技術によって異なる方向に提供された。体験者からは、「ほかの言語の音声が混ざることなくはっきりと聞こえる」といった声が多数あり、本技術の有効性を確認することができた（図 13）。なお、実験現場は水生動物がいる水槽の上部（バックヤード）にあり、物を落としたりすることが許されない状況であるため、指向性超音波スピーカや多言語対応のための解説用ヘッドセット等が使えないという課題があったが、本技術によりそれらが解決できることが示された。また、従来、円形スピーカを用いたマルチスポット再生においては、図 9（b）に示されるように聴取者がスピーカの周りを歩き回って音を聞く「円形分割再生方式」としていたが、本実証実験では、図 14 のとおり円形スピーカを壁面に縦方向に取り付けることにより、聴取者は頭上から降り注ぐ音を聞き、左右に移動することによりマルチスポット再生を体験できる「直線エリア分割型再生方式」を採用した。「円形分割再生方式」は、考えられる

本技術の応用先候補が限定的であったが、「直線エリア分割型再生方式」は、本技術の応用先候補を広げることにつながっており、画期的な再生方式であると言える。

2025 年 4 月には、多言語同時通訳も含めてノート PC1 台で動作可能であり、かつリュックサック 1 つで持ち運び可能なコンパクトデモシステムの試作に成功し [76]、これまで以上にどこにでもデモシステムを持ち運べるようになった（図 15）。2025 年 8 月には、このコンパクトデモシステムを音声分野のフラッグシップ国際会議 Interspeech 2025 における Speech Science Festival [100] へ出展（招待）し、国際的にもアピールすることに成功した。

さらに、2025 年 9 月 16 日～ 22 日には、大阪・関西万博における Future Life Experience にて音声マルチスポット再生技術を出展し、数多くの方に本技術を体験いただくことができた。

現在はこのコンパクトデモシステムを用いて様々な場所での実証実験を行い、本技術の応用先候補を検討するとともに、更なる高精度な再生方式の研究開発にも取り組んでいる [79][101]。

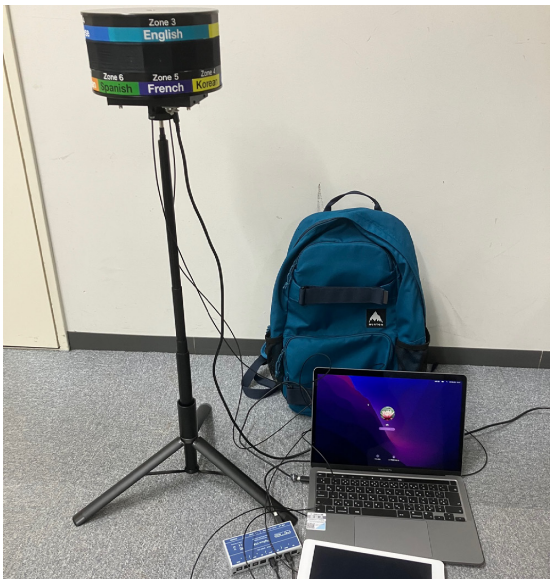


図 15　リュックサック 1 つで持ち運び可能なコンパクトデモシステム

6　出展・実証実験・技術移転

2 の同時通訳技術、**3** の音声認識技術、**4** の音声合成技術、**5** の音声マルチスポット再生技術について、日本語又は英語のプレゼンテーションを同時通訳している動画の放映又はその実演デモ（図 16）を、それぞれの時点の最新の技術を用いて多くの展示会で実施し、同時通訳技術をアピールするとともに、その音声出力にマルチスポットスピーカを活用することにより、あわせて音声マルチスポット再生技術をアピールした。

2023 年度は、大阪・関西万博の運営主体である（公社）2025 年日本国際博覧会協会（以下「万博協会」という。）や関連府省庁からの依頼を複数回受け、G7 群馬高崎デジタル・技術大臣会合、G7 富山・金沢教育大臣会合、G7 堺・大阪貿易大臣会合、国際連合主催のインターネット・ガバナンス・フォーラム京都 2023 に出展した。これらの展示では、我が国の関係閣僚のみならず諸外国の閣僚も展示に訪れ、一部は報道でも取り上げられるなど、特に広く周知された。

これらの技術は大阪・関西万博のほか、CEATEC や NICT オープンハウス、けいはんな R&D フェア等のイベントにも毎年出展し、特に 2024 年の CEATEC2024 では、日本語又は英語による NICT の 10 件の展示内容の紹介をリアルタイムで 3 言語に同時通訳するデモ（日→英中韓又は英→日中韓）を実施し、継続的に技術力の周知に努めている。（図 17）。

また、NICT では、200 以上の会員から成るグローバルコミュニケーション開発推進協議会（以下「GCP 協議会」という。）の事務局をつとめ、活動の企画・運営を行っている。この GCP 協議会を産学官連携の拠点として活用し、会員を主な対象とした各種会合や一般を対象とした自動翻訳シンポジウムにおいて、多言語翻訳技術や同時通訳技術及びその社会での活用に関する講演を行うことにより、社会実装への機運を高めた。

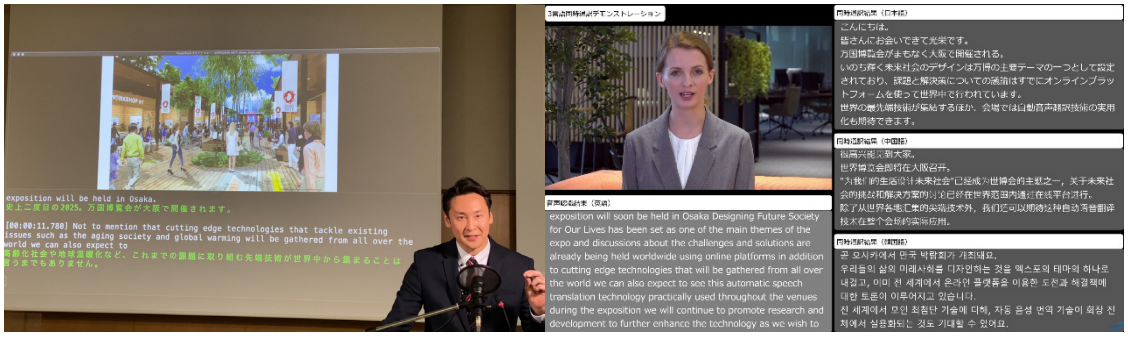


図 16　同時通訳デモシステム

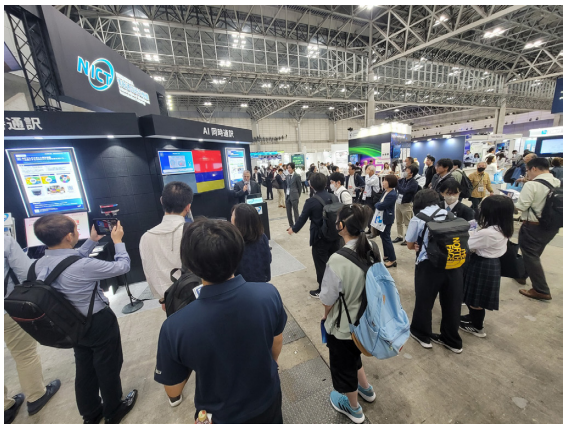


図 17　CEATEC2024 における実演デモ

「総務省委託･多言語翻訳技術高度化推進コンソーシアム」においても、成果の社会実装・利活用に向けた活動を行った。GCP 協議会会員に対しては、その成果である同時通訳 API、基本アプリ、UI デザインルールの試験提供を実施するとともに、GCP 協議会の Web サイトにおいて UI デザインルールを広く一般に提供した。また、外国人就労支援･自治体･ビジネス会議･遠隔協業などの分野における実証実験を実施した。さらに、クラウド型及びスタンドアロン型のシステムを構築し、大阪・関西万博におけるセミナーやシンポジウム等のシーンを想定した、登壇者の発話音声を同時通訳して聴講者に向けて字幕を表示する実証実験が行われた。これらの実証実験は、大阪・関西万博の紹介イベントや半導体分野の大規模な国際展示会等の実利用に近い環境で、大阪・関西万博の運営主体である万博協会の協力も得て民間企業主導で複数回実施され、実運用に向けた課題の抽出とそれに基づくシステムの改良が行われた。

NICT が開発した技術の特許出願や技術移転も進み、同時通訳のコア要素（チャンク翻訳、音声合成、音声認識・言語識別技術）に関連する 44 件の特許出願及び PCT 出願を行った（2024 年度末時点）。また、TOPPAN（株）の「MeeTra」が 2024 年 6 月に商用サービスを開始されるなど、スタンドアロン型同時通訳技術に関して民間企業から商用サービスが始まっている。

7 大阪・関西万博における民間企業主体の多言語翻訳技術活用状況

大阪・関西万博では、NICT の技術を活用した民間企業の以下のサービスが活用されている（図 18）。

1. 1 対 1 の多言語の会話シーンで利用される翻訳アプリ「EXPO ホンヤク」（TOPPAN ホールディングス（株））
2. 万博協会主催のガイドツアー等 1 対複数人で利用できるサービス「EXPO ホンヤク Remote」（TOPPAN ホールディングス（株））
3. 万博協会主催のセミナー等で利用される、発表内容を同時通訳しリアルタイムで字幕表示を行う「EXPO 同時通訳システム」（クラウド版及びスタンドアロン版、TOPPAN ホールディングス（株））
4. 会場内で流れるアナウンスを同時通訳技術により多言語で配信する「テキストアナウンス」（SoundUD コンソーシアム）
5. バーチャル会場での多言語チャットコミュニケーション（（株）NTT ドコモ、（株）みらい翻訳）
6. アバター対話等の最新技術を用いた先進的な UI/UX による「同時通訳体験ブース」（期間限定、SoundUD コンソーシアム）

1.～5. のサービスは、万博会場を未来社会のショーケースに見立て、先進的な技術やシステムを取り入れ未来社会の一端を実現することを目指す「未来社会ショーケース事業」における「デジタル万博」の「自動翻訳システム」として位置付けられており、全期間を通じて活用される。

「EXPO ホンヤク」「EXPO ホンヤク Remote」は逐次翻訳技術を活用したサービスである。「EXPO ホンヤク」は、大阪･関西万博の全来場者や、公式参加国、民間パビリオン出展者などが無料で使用可能なアプリであり、App Store 及び Google Play から無料でダウンロードが可能である。30 言語（音声翻訳は 13 言語）に対応しており、万博に特化した専門用語も搭載されていて、様々な利用シーンにおけるスムーズな対応を実現するものである。

「EXPO ホンヤク Remote」は、手持ちのスマートフォン等の端末から二次元コードで Web アプリにアクセスし、言語選択するだけで、ガイドの話す言語を

図 18　大阪・関西万博における NICT の技術の活用

希望の言語に翻訳して聞き取ることができるサービスで、日本語を含む 13 言語に対応している。

「EXPO 同時通訳システム」は同時通訳技術を活用したサービスで、クラウド版では、登壇者の発表内容が、会場スクリーンへの字幕や視聴者の端末へ、自動同時通訳されて表示される。オンライン配信にも対応しており、会場外の視聴者もアクセス可能である。日本語を含む 6 言語に対応している。また、スタンドアロン版は、クラウド版と同機能の環境をローカルの PC 内に構築し、オフラインで自動同時通訳技術を提供するものであり、日英の 2 言語に対応している。

これらのシステムでは、TOPPAN ホールディングス(株)の協賛サービスである「EXPO 多言語用語集」が活用されており、専門用語の表現が統一されることによって、大阪・関西万博にまつわる会話の正確かつスムーズな翻訳が実現され、コミュニケーション上の混乱を防いでいる。

「テキストアナウンス」は、万博会場の屋内外で流れるアナウンスを、手持ちのスマートフォンなどの端末に設定された言語に翻訳し、テキストで確認することができるサービスで、二次元コードを読み取って Web サイトにアクセスするだけで、専用アプリ不要で利用することができる。日本語と英語を基本とし、最大 7 言語に対応している。

「バーチャル会場での多言語チャットコミュニケーション」は、オンライン空間上に 3DCG で再現された夢洲会場のバーチャル会場で、テキストチャットによるコミュニケーションを翻訳するもので、日本語を含む 14 言語に対応している。

「同時通訳体験ブース」は、テーマウィーク「未来のコミュニティとモビリティウィーク」(5 月 20 ～ 24 日)において経済産業省が主催する「福島復興展示」や、5 月 27 日～ 6 月 2 日の Future Life Experience 期間展示にて設置されたもので、多言語翻訳技術とキャラクターを融合させ、スクリーンに投影されたキャラクターを通訳者として、異なる言語の人同士が円滑な会話やコミュニケーションを行えるというものである。

8 おわりに

GC 計画 2025 の下、NICT では、同時通訳技術を実用化レベルまで高めるとともに、音声認識技術、音声合成技術それぞれの高度化を実現し、音声マルチスポット再生技術との連携も進めた。また、多数の実証実験、多様な広報活動を実施して、NICT の技術の技術移転及び社会実装を進め、民間企業と連携した社会実証により大阪・関西万博で NICT の技術を用いた様々なサービスが提供されるに至っている。

これら技術の更なる高度化を進めるとともに、今回の大阪・関西万博での実績も活用し、技術移転及び社会実装をより一層進めていく。

謝辞

本稿の一部は、総務省の「ICT 重点技術の研究開発プロジェクト（JPMI00316）」における「多言語翻訳技術の高度化に関する研究開発」による委託を受けて実施した研究開発による成果について記載している。

音声マルチスポット再生技術の研究開発の一部は JSPS 科研費 JP23K11177 の助成を受けたものである。

【参考文献】

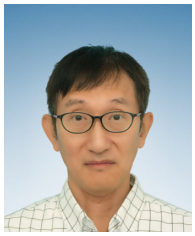
1 総務省，“グローバルコミュニケーション計画，” April 2014, https://www.soumu.go.jp/main_content/000285578.pdf
2 総務省，“グローバルコミュニケーション計画 2025，” March 2020, https://www.soumu.go.jp/main_content/000678485.pdf
3 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” CoRR, Vol.abs/1706.03762, 2017.
4 Kenji Imamura and Eiichiro Sumita, “Multilingual parallel corpus for global communication plan,” Proceedings of LREC 2018, pp.3453–3458, May 2018.
5 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 回年次大会発表論文集, pp.512–515, 2018.
6 今村 賢治, 隅田 英一郎，“対話翻訳における長距離文脈の利用，” 言語処理学会第 25 回年次大会発表論文集, pp.550–553, 2019.
7 東山 翔平, 今村 賢治, 内山 将夫, 隅田 英一郎，“GCP 同時通訳コーパスの構築，” 言語処理学会第 29 回年次大会発表論文集, pp.1405–1410, 2023.
8 東山 翔平，“大規模言語モデル時代の機械翻訳の展望，” Jxiv，2024, https://jxiv.jst.go.jp/index.php/jxiv/preprint/view/932
9 Xiaolin Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. “An Efficient and Effective Online Sentence Segmenter for Simultaneous Interpretation.” Proceedings of the 3rd Workshop on Asian Translation (WAT 2016), pp.139–148, Osaka, Japan, Dec. 12, 2016.
10 Xiaolin Wang, Masao Utiyama, and Eiichiro Sumita. “Online Sentence Segmentation for Simultaneous Interpretation using Multi-Shifted Recurrent Neural Network.” Proceedings of the 17th Machine Translation Summit, pp.1–11, Dublin City University, Dublin, Ireland, Aug. 19–23, 2019.
11 L Y. LeCun, Y. Bengio, and G. E. Hinton, “Deep learning,” Nature, vol.521, no.7553, pp.436–444, May 2015.
12 I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning,” MIT Press, 2016.
13 麻生 英樹，安田 宗樹，前田 新一，岡野原 大輔，岡谷 貴之，久保 陽太郎，ボレガラ ダヌシカ，神嶌 敏弘，“深層学習 Deep Learning，” 近代科学社, 2015.
14 L. Rabiner and C. Schmidt, “Application of dynamic time warping to connected digit recognition,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.28, issue 4, pp.377–388, Aug. 1980.
15 C. M. Bishop, “Pattern recognition and machine learning,” Springer, 2006.
16 F. Jelinek, “Statistical methods for speech recognition (Language, speech, and communication),” MIT Press, 1998.
17 河原 達也，“音声認識システム，” オーム社, 2006.
18 K. Vesely, A. Ghoshal, L. Burget, and D. Povey, “Sequence-discriminative training of deep neural networks,” Proceedings of Interspeech ’12, pp.2345–2349, Aug. 2013.
19 A. Graves, N. Jaitly, and A. Mohamed, “Hybrid speech recognition with deep bidirectional LSTM,” in Proceedings of ASRU ’13, pp.273–278, Dec. 2013.
20 久保 陽太郎，“音声認識のための深層学習，” 人工知能，29 巻，1 号，pp.62–71，Jan. 2014.
21 H. Sak, A. Senior, and F. Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling,” Proceedings of Interspeech ’14, pp.338–342, Sept. 2014.
22 D. Yu and L. Deng, “Automatic speech recognition: A deep learning approach,” Springer, 2015.
23 D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, “Purely sequence-trained neural networks for ASR based on lattice-free,” Proceedings of Interspeech ’2016, pp.2751–2755, Sept. 2016.
24 S. Watanabe, M. Delcroix, F. Metze, and J. R. Hershey, “New era for robust speech recognition - Exploiting deep learning,” Springer, 2017.
25 U. Kamath, J. Liu, and J. Whitaker, “Deep learning for NLP and speech recognition,” Springer, 2019.
26 高島 遼一，“Python で学ぶ音声認識，” インプレス, 2021.
27 久保 陽太郎，“機械学習による音声認識，” コロナ社，2021.
28 神田 直之，“音声認識における深層学習に基づく音響モデル，” 日本音響学会誌，73 巻，1 号，pp.31–38，Jan. 2017.
29 渡部 晋治，堀 貴明，“音声言語理解のための音声認識，” 電子情報通信学会誌，vol.101，no.9，pp.885–890，Sept. 2018.
30 渡部 晋治，久保 陽太郎，“深層学習が支える音声認識技術，” 電子情報通信学会誌，vol.105，no.5，pp.392–396，May 2022.
31 林 知樹，“End-to-End 音声処理の概要と ESPnet2 を用いたその実践，” 日本音響学会誌，76 巻，12 号，pp.720–729，Dec. 2020.
32 河原 達也，“音声認識技術の変遷と最先端 －深層学習による End-to-End モデル－，” 日本音響学会誌，74 巻，7 号，pp.381–386，July 2018.
33 R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter and S. Watanabe, “End-to-End Speech Recognition: A Survey,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.32, pp.325–351, Oct. 2023.
34 M. Wang, I. Shafran, H. Soltau, W. Han, Y. Cao, D. Yu, and L. El Shafey, “Speech-to-text adapter and speech-to-entity retriever augmented LLMs for speech understanding,” arXiv preprint arXiv:2306.07944, June 2023.
35 J. Wu, Y. Gaur, Z. Chen, L. Zhou, Y. Zhu, T. Wang, J. Li, S. Liu, B. Ren, L. Liu, and Y. Wu, “On decoder-only architecture for speech-to-text and large language model integration,” Proceedings of ASRU ’23, Dec. 2023.
36 S. Ling, Y. Hu, S. Qian, G. Ye, Y. Qian, Y. Gong, E. Lin, and M. Zeng, “Adapting large language model with speech for fully formatted end-to-end speech recognition,” Proceedings of ICASSP ’24, pp.11046–11959, April 2024.
37 W. Yu, C. Tang, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, “Connecting Speech Encoder and Large Language Model for ASR,” Proceedings of ICASSP ’24, pp.12637–12641, April 2024.
38 Y. Fathullah, C. Wu, E. Lakomkin, J. Jia, Y. Shangguan, K. Li, J. Guo, W. Xiong, J. Mahadeokar, O. Kalinli, C. Fuegen, and M. Seltzer, “Prompting large language models with speech recognition abilities,” Proceedings of ICASSP ’24, pp.13351–13355, April 2024.
39 Z. Chen, H. Huang, A. Andrusenko, O. Hrinchuk, K. C. Puvvada, J. Li, S. Ghosh, J. Balam, and B. Ginsburg, “SALM: Speech-augmented Language Model with In-context Learning for Speech Recognition and Translation,” in Proceedings of ICASSP ’24, pp.13521–13525, April 2024.
40 F. Verdini, P. Melucci, S. Perna, F. Cariaggi, M. Gaido, S. Papi, S. Mazurek, M. Kasztelnik, L. Bentivogli, S. Bratières, P. Merialdo, and S. Scardapane, “How to Connect Speech Foundation Models and Large Language Models? What Matters and What Does Not,” arXiv preprint arXiv:2409.17044, Sept. 2024.

41 I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in Proceedings of NIPS ’14, vol.27, Dec. 2014.
42 J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” Proceedings of NeurIPS ’15, pp.577–585, Dec. 2015.
43 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Proceedings of NeurIIPS ’17, Dec. 2017.
44 A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” Proceedings of ICML ’06, pp.369–376, June 2006.
45 K. Miyazaki, M. Murata, and T. Koriyama, “Structured State Space Decoder for Speech Recognition and Synthesis,” in Proceedings of ICASSP ’23, June 2023.
46 Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S. Chang, K. Rao, and A. Gruenstein, “Streaming end-to-end speech recognition for mobile devices,” Proceedings of ICASSP ’19, pp.12–17, May 2019.
47 A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised earning of Speech Representations,” Proceedings of NeurIPS ’20, pp.12449–12460, Dec. 2020.
48 W. Hsu, B. Bolte, Y. H. 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, Oct. 2021.
49 T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., “Language Models are Few-Shot Learners,” Advances in Neural Information Processing Systems, vol.33, pp.1877–1901, 2020.
50 OpenAI, “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, March 2024.
51 Š. Sedláček, S. Kesiraju, A. Polok, and J. Černocký, “Aligning Pre-trained Models for Spoken Language Translation,” arXiv preprint arXiv:2411.18294, Nov. 2024.
52 R. Ma, T. Chen, K. Audhkhasi, and B. Ramabhadran, “LegoSLM: Connecting LLM with Speech Encoder using CTC Posteriors,” arXiv preprint arXiv:2505.11352, May 2025.
53 Ruchao Fan, Bo Ren, Yuxuan Hu, Rui Zhao, Shujie Liu, and Jinyu Li, “AlignFormer: Modality Matching Can Achieve Better Zero-shot Instruction-Following Speech-LLM,” arXiv preprint arXiv:2412.01145, Dec. 2024.
54 多言語音声翻訳アプリ VoiceTra, https://voicetra.nict.go.jp
55 総務省,“聴覚障害者放送視聴支援緊急対策事業,” March 2019, https://www.soumu.go.jp/menu_news/s-news/01ryutsu09_02000228.html
56 沈 鵬、Lu Xugang,“言語識別･話者識別技術,” 情報通信研究機構研究報告, vol.68 no.2, pp.39–46, Dec. 2022. https://www.nict.go.jp/publication/shuppan/kihou-journal/houkoku68-2_HTML/2022U-02-02-05.pdf
57 志賀 芳則, 河井 恒,“多言語音声合成システム,” 情報通信研究機構季報, vol.58, no.3, pp.19–24, Sept. 2012.
58 Y. Shiga, J. Ni, K. Tachibana, and T. Okamoto, “Text-to-Speech Synthesis,” Speech-to-Speech Translation, pp.39–52, 2020.
59 岡本 拓磨,“ニューラルネットワークに基づく音声波形生成モデル,” 音響誌, vol.78, no.6, pp.328–337, June 2022.
60 岡本 拓磨,“ニューラル音声合成技術,” 情報通信研究機構研究報告, vol.68, no.2, pp.47–55, Dec. 2022.
61 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Proc. NIPS, Dec. 2017, pp.5998–6008.
62 Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “FastSpeech 2: Fast and high-quality end-to-end text to speech,” Proc. ICLR, May 2021.
63 Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” Proc. CVPR, pp.11976–11986, June 2022.
64 T. Okamoto, Y. Ohtani, T. Toda, and H. Kawai, “ConvNeXt-TTS and ConvNeXt-VC: ConvNeXt-based fast end-to-end sequence-to-sequence text-to-speech and voice conversion,” Proc. ICASSP, April 2024, pp.12456–12460.
65 T. Okamoto, Y. Ohtani, S. Shimizu, T. Toda, and H. Kawai, “Challenge of singing voice synthesis using only text-to-speech corpus with FIRNet source-filter neural vocoder,” Proc. Interspeech, pp.1870–1874, Sept. 2024.
66 T. Ogura, T. Okamoto, Y. Ohtani, E. Cooper, T. Toda, and H. Kawai, “Phoneme-level duration controllable neural text-to-speech with phoneme embedding skip connection and modified Gaussian duration modeling,” IEEE Access, vol.13, pp.118369–118380, 2025.
67 J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” Proc. NeurIPS, pp.17022–17033, Dec. 2020.
68 T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, “Subband WaveNet with overlapped single-sideband filterbanks,” Proc. ASRU, pp.698–704, Dec. 2017.
69 T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, “An investigation of subband WaveNet vocoder covering entire audible frequency range with limited acoustic features,” Proc. ICASSP, pp.5654–5658, April 2018.
70 T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, “Improving FFTNet vocoder with noise shaping and subband approaches,” Proc. SLT, pp.304–311, Dec. 2018.
71 T. Okamoto, T. Toda, and H. Kawai, “Multi-stream HiFi-GAN with data-driven waveform decomposition,” Proc. ASRU, Dec. 2021, pp.610–617.
72 T. Okamoto, H. Yamashita, Y. Ohtani, T. Toda, and H. Kawai, “WaveNeXt: ConvNeXt-based fast neural vocoder without iSTFT layer,” Proc. ASRU, Dec. 2023.
73 H. Yamashita, T. Okamoto, R. Takashima, Y. Ohtani, T. Takiguchi, T. Toda, and H. Kawai, “Fast neural speech waveform generative models with fully-connected layer-based upsampling,” IEEE Access, vol.12, pp.31 409–31 421, 2024.
74 https://www.nict.go.jp/press/2024/06/25-1.html
75 T. Okamoto, Y. Ohtani, and H. Kawai, “Mobile PresenTra: NICT fast neural text-to-speech system on smartphones with incremental inference of MS-FC-HiFi-GAN for low-latency synthesis,” Proc. Interspeech, pp.997–998, Sept. 2024.
76 T. Okamoto and M. Kono, “Simultaneous speech translation integrated compact multiple sound spot synthesis system on a laptop carried out with a backpack,” Proc. Interspeech, pp. 3539-3540, Aug. 2025.
77 T. Okamoto, Y. Shiga, and H. Kawai, “Hi-Fi-CAPTAIN: High-fidelity and high-capacity conversational speech synthesis corpus developed by NICT,” https://ast-astrec.nict.go.jp/en/release/hi-fi-captain/, 2023.
78 T. Ogura, T. Okamoto, Y. Ohtani, E. Cooper, T. Toda, and H. Kawai, “GST-BERT-TTS: Prosody prediction without accentual labels for multi-speaker TTS using BERT with global style tokens,” Proc. Interspeech, pp.444-448, Aug. 2025.
79 T. Okamoto, “Speech masking system based on spatially separated multiple TTS maskers with a compact circular loudspeaker array,” Proc. ASRU, Dec. 2025. (accepted, in press)
80 T. Okamoto, T. Toda, and H. Kawai, “E2E-S2S-VC: End-to-end sequence-to-sequence voice conversion,” Proc. Interspeech, pp.2043–2047, Aug. 2023.
81 K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, and H. Kawai, “Harmonic-Net: Fundamental frequency and speech rate controllable fast neural vocoder,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol.31, pp.1902–1915, 2023.
82 Y. Ohtani, T. Okamoto, T. Toda, and H. Kawai, “Fast neural vocoder with fundamental frequency control using finite impulse response filters,” IEEE Trans. Audio Speech Lang. Process., vol.33, pp.1893–1906, 2025.
83 Y. Ohtani, T. Okamoto, T. Toda, and H. Kawai, “Voice factor control using FIR-based fast neural vocoder for speech generation applications,” Proc. ASRU, Dec. 2025. (accepted, in press)
84 T. Ogura, T. Okamoto, Y. Ohtani, E. Cooper, T. Toda and H. Kawai, “Mora-level prosody prediction for text-to-speech using Japanese BERT without accentual labels,” Proc. ICASSP, April 2025.
85 E. Cooper, T. Okamoto, Y. Ohtani, T. Toda, and H. Kawai, “Layer-wise analysis for quality of multilingual synthesized speech,” Proc. ASRU, Dec. 2025. (accepted, in press)
86 https://youtu.be/uyTRd5Hu6hw
87 岡本 拓磨,“スピーカアレイを用いたマルチスポット再生技術の理論と実装,” 音響誌, vol.81, no.10, pp.711-718, Oct. 2025.

88 J.-W. Choi and Y.-H. Kim, “Generation of an acoustically bright zone with an illuminated region using multiple sources,” J. Acoust. Soc. Am., vol.111, no.4, pp.1695–1700, April 2002.
89 O. Kirkeby and P. Nelson, “Reproduction of plane wave sound fields,” J. Acoust. Soc. Am., 94, 2992–3000, 1993.
90 T. Lee, L. Shi, J. K. Nielsen, and M. G. Christensen, “Fast generation of sound zones using variable span trade-off filters in the DFT-domain,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol.29, pp.363–378, 2021.
91 T. Abe, S. Koyama, N. Ueno, and H. Saruatari, “Amplitude matching for multizone sound field control,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol.31, pp.656–669, 2023.
92 E. G. Williams, “Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography,” London: Academic Press, 1999.
93 T. Okamoto, “Analytical methods of generating multiple sound zones for open and baffled circular loudspeaker arrays,” Proc. WASPAA, Oct. 2015.
94 T. Okamoto and A. Sakaguchi, “Experimental validation of spatial Fourier transform-based multiple sound zone generation with a linear loudspeaker array,” J. Acoust. Soc. Am., vol.141, no.3, pp.1769–1780, March 2017.
95 https://ast-astrec.nict.go.jp/MultipleSoundSpotSynthesis/
96 https://youtu.be/fTyYs6AqtNM
97 https://youtu.be/G1qkR-B40PM
98 https://www.jst.go.jp/pr/jst-news/backnumber/2024/202407/pdf/2024_07_p12-13.pdf
99 https://www.nict.go.jp/publicity/topics/2025/01/15-1.html
100 https://www.interspeech2025.org/science-fest
101 T. Okamoto, “SFC-L1: Sound field control with least absolute deviation regression,” Proc. WASPAA, pp.1-4, Oct. 2025.

PAUL Michael （ぱうる みひゃえる）
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
先進的翻訳技術研究室専門
主任研究技術員
博士（工学）
音声翻訳

【受賞歴】
2015 年　アジア太平洋機械翻訳協会 AAMT 長尾賞受賞



今村 賢治 （いまむら けんじ）
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
先進的翻訳技術研究室
主任研究員
博士（工学）
音声翻訳

【受賞歴】
2019 年　情報処理学会　第 241 回 自然言語処理研究会　優秀研究賞
2012 年　言語処理学会　論文賞

王 暁林 （わん しゃおりん）
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
先進的翻訳技術研究室
主任研究技術員
博士（工学）
音声翻訳

東山 翔平 （ひがしやま しょうへい）
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
先進的翻訳技術研究室
研究員
博士（工学）
自然言語処理

【受賞歴】
2024 年　言語処理学会第 30 回年次大会 委員特別賞
2021 年　the 7th Workshop on Noisy User-generated Text (W-NUT) Best Paper Award
2021 年　言語処理学会　論文賞

内山 将夫 （うちやま まさお）
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
先進的翻訳技術研究室
室長
博士（工学）
自然言語処理
音声翻訳

【受賞歴】
2020 年　第 2 回オープンイノベーション大賞 総務大臣賞
2016 年　第 31 回電気通信普及財団賞（テレコムシステム技術賞）
2014 年　アジア太平洋機械翻訳協会 AAMT 長尾賞受賞

藤本 雅清 （ふじもと まさきよ）
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
研究マネージャー
博士（工学）
音声音響信号処理、音声認識、機械学習

【受賞歴】
2015 年　IEEE ASRU ‘15 Best Paper Award Honorable Mention
2011 年　情報処理学会 平成 22 年度 山下記念研究賞
2003 年　日本音響学会 第 20 回 粟屋 潔学術奨励賞

岡本 拓磨 （おかもと たくま）
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
研究マネージャー
博士（情報科学）
音場制御、音声合成

【受賞歴】
2022 年　日本音響学会 第 9 回学会活動貢献賞
2018 年　日本音響学会 第 57 回佐藤論文賞
2012 年　日本音響学会 第 32 回粟屋 潔学術奨励賞

菊池 武文（きくち たけふみ）
イノベーションデザインイニシアティブ
主任

塩飽 裕彦（しあく ひろひこ）
ユニバーサルコミュニケーション研究所
総合企画室
企画戦略グループ

香山 健太郎（かやま けんたろう）
経営企画部
統括
博士 (工学)
人工知能