小特集　―音声の多様性とその応用―

言語の多様性とアイヌ語の音声言語処理 *

河原達也，松浦孝平（京都大学）**

1. はじめに

世界には数多くの民族があり、数多くの言語が存在する。民族・言語ともに数千あるといわれるが、その大多数がいわゆる少数民族・少数言語である。UNESCO の World Atlas of Languages によると、現在約 7000 の言語が使われているが、その多くが消滅の危機にあるとされている。ちなみに日本では、アイヌ語が「極めて深刻」、琉球諸島と八丈島の 7 言語（方言）が「重大な危機」または「危険」とされている。過去の政府の同化政策や、テレビやインターネットの普及、ビジネスのグローバル化により、その国のいわゆる「標準語」や「共通語」を使うことが多くなったためである。話者の多くは高齢であるので、このまま何の対策も講じなければ消滅してしまう。

言語はコミュニケーションの手段であるが、その民族や地域の文化を伝承する役割も果たしている。実際にアイヌの文化は口頭伝承で伝えられてきたし、舞踊や祭祀においても歌や詞がつけられることが多い。それらを、日本語や英語に翻訳しても、その背景を含めて完全に伝えられるものではない。例えばアイヌ語では、熊を kamuy、蛇を tannekamuy、シャチを repunkamuy、ふくろうを kamuy-cikappo と呼ぶが、これらを神(kamuy)またはその化身としてみなしていることがわかる。単に「森で蛇に遭遇した」ではそのようなニュアンスは伝わらない。言語の消滅は文化の消滅を意味するのである。

そのため、２０世紀の後半から、（日本に限らず）少数言語を保存・継承、さらには再興しようという動きが、民間そして国を巻き込んで起こっている。まずは、口頭伝承を録音・アーカイブ化することに始まり、その言語を博物館や公共のスペースで使用したり、若い世代にも学習・会話してもらう機会を設けることが挙げられる。

このような取組みのために、音声言語処理技術を研究開発・活用できないかというのが、本記事の主旨である。音声認識に代表される音声言語処理は大きな発展を遂げて、様々な実用化がなされ、一定の条件では「人間レベル」の性能に近づいているが、このような高性能・実用的な音声認識が実現されているのは 100 言語程度である（例えば Google の Speech-to-Text は 125 言語、Open AI の Whisper は 99 言語）。これは、使用者人口などのニーズ（⇔少数言語）だけでなく、学習データのリソース（⇔低資源言語）の制約によるものである。少数言語は必然的に低資源であり、大規模なデータに依存した機械学習に基づく現代の音声言語処理システムを構築することは容易でなく、かなりの挑戦的課題といえる。

しかしながら、この問題に取り組む研究者も増えている。ISCA (International Speech Communication Association) と ELRA (European Languages Resource Association) に、SIGUL (Special Interest Group on Under-resourced Languages)が設立され、定期的に Spoken Language Technologies for Under-resourced Languages (SLTU) ワークショップが開催されている。また 2019 年には、UNESCO 主催で International Conference on Language Technologies for All (LT4All) が開催され、2025 年にも企画されている。著者らは、約５年前にアイヌ語の音声認識の研究に着手した。アイヌ語に関する知識は皆無であったが、アイヌ語の専門家や保全に関わる様々な方々の協力を頂いて、当初の想定以上に進めることができた。その経験・知見をふまえて、研究テーマや現状の技術、今後の方向性について述べる。

---

* Diversity in languages and spoken language processing of Ainu.

** Tatsuya Kawahara, Kohei Matsuura (Kyoto University) email: kawahara@i.kyoto-u.ac.jp

2. 応用分野と問題設定

まず、どのようなニーズがあるのか挙げて、それに必要な音声言語処理と問題設定を述べる。

2.1 口頭伝承アーカイブの処理

昔の話を話者が生存しているうちに、話してもらい収録した音源は世の中に多数ある。アイヌ語については、3.2 節で述べる『アイヌ語アーカイブ』が、琉球諸語についても『沖縄伝承話データベース』等が構築されている。

ただし、これらのデータのうち、書き起こしがされて、音声との対応付け・時間情報付与が行われているのは一部である。したがって、この（半）自動化を行うための音声認識技術が求められる。書き起こしのための音声認識には高い精度が必要であるが、書き起こしが与えられた上で音声との対応付け（アライメント）を行うにはそれほどの精度は必要でない。ただし、数十分の音声から十秒程度の文単位に区分化する必要がある。

さらに、まだアーカイブ化が行われていない手つかずの音源も多数ある。インタビュー形式で行われていることも多く、対象言語以外（例えば日本語）で話されている区間も多いので、話者認識・言語認識とそれに基づく区間分割（セグメンテーション）が必要となる。音声には、節がつけて歌われていたり、手拍子などでリズムをとっている場合もある。

2.2 教育・展示コンテンツの作成

博物館における展示の説明文や、教材テキストにおいて、読み上げ音声が求められ、そのための音声合成技術が求められる。アイヌ語のように母語話者がほとんどいない場合は、専門家であっても、正しい韻律がわからないことがある。そのための参考になるとよい。琉球諸語のように、母語話者による大規模な音声収録が可能な場合、高品質の音声合成が可能となり、音声教材の作成も可能になると期待される。

2.3 言語学習支援システム

主要言語の学習支援と同様のシステムが想定される。語彙の訓練の他に、発音練習やさらには日常会話などの簡単な音声対話を行うシステムも考えられる。サミ語の会話[1]やマオリ語の発音練習[2]に関する研究が行われている。音声認識や音声合成の組合せで実現できるが、日本人が話す英語のように、母語としない話者を扱う必要がある。主要言語と異なり、母語話者のデータがきわめて少ないため、発音をチェックするモデルの構築はきわめて困難である。

3. 低資源言語の音声認識

次に、少数言語、すなわち低資源言語の音声認識の方法について述べる。なお、音声合成については、1 名の話者でもある程度の分量の音声データを収録できれば主要言語と同様の方法で構築可能である。

3.1 転記・認識単位

少数言語の転記には、表音文字が用いられることが多い。そもそも言語固有の正書法が存在しなかったり、研究者が当該言語の母語話者でない場合が多いためである。英語の音声認識で用いられるラテンアルファベットを用いた擬似音素が用いられることが多いが、様々な曖昧性が生じるので、IPA（International Phonetic Alphabet）の音声記号を用いることも考えられる。

アイヌ語の場合は、音素は日本語とほぼ共有しているが、閉音節（子音-母音-子音）があるので、カタカナで表記するのは無理があり（閉音節の最後の子音に半角カナを用いる表記もある）、ローマ字表記されることが一般的である。

琉球諸語の場合は、日本語と音節構造（いわゆる五十音）は同じであるが、日本語にない音素が多数存在する。カタカナで表記できなくもないが、IPA のような音声記号を用いて転記される。

音声認識（及び音声合成）の単位は、この転記に用いられる文字（概ね音素に相当）に基づいて定義される。単純にはその文字（音素）を単位とする。この連鎖単位を BPE（Byte-Pair Encoding）や unigram 言語モデルに基づいて学習することも考えられるが、データ量が少ないと信頼できる学習ができない可能性もある。そこで、音節のような単位を用いることも考えられる。

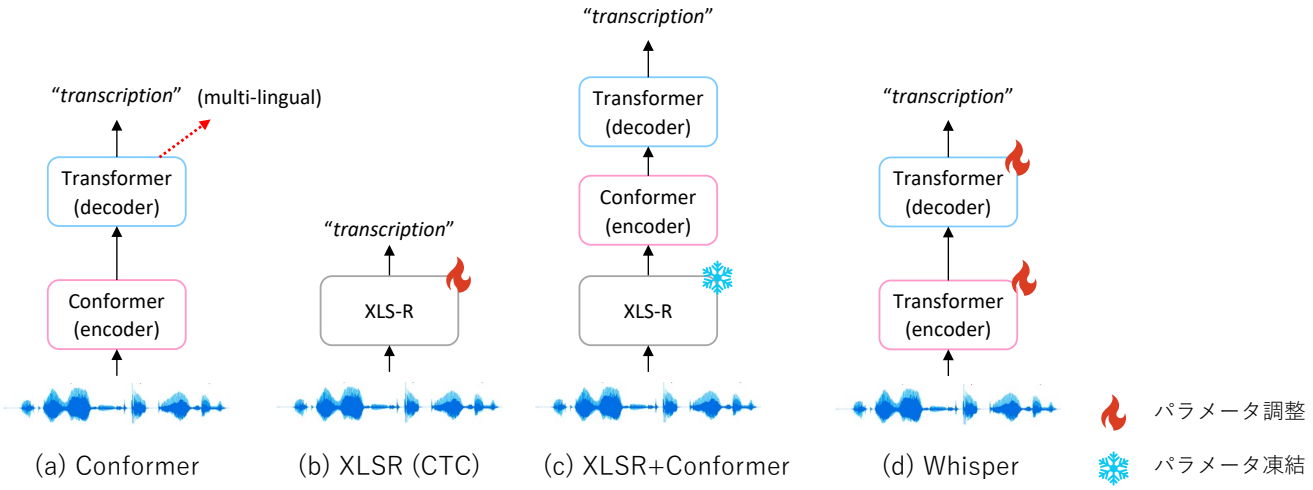

図 1 低資源言語に対する主な音声認識手法（4.3 節の実験で使用）

3.2 音響モデルの多言語学習

音声から音素や音節を認識するモデル（end-to-end モデル[3]ではエンコーダに相当し、CTC (Connectionist Temporal Classification) 損失[4]で音声認識が可能）を音響モデルと呼ぶことにして、これを少量のデータで効率的に学習することが最大の課題である。

初期の最も一般的な方法は、大規模な学習データが存在する関連する言語のデータを併用するものである。音響モデル（エンコーダ）を共通にして、最後の出力の部分だけ言語固有のものにする（図 1(a)）。データ量の大きくない複数の言語の音響モデルをまとめて学習する設定が本来の多言語（マルチリンガル）学習であるが、特定の低資源言語の音声認識のために関連する高資源言語のデータを活用するのも有用であり、転移学習と捉えられる。アイヌ語と日本語、マオリ語とニュージーランド英語のように、現代では多くの少数言語の話者はその国の主要言語の話者であることが多い。したがって、（本来そうであったかは疑問であるが）音素の発音も共通点が多いことが期待される。データ量や音素の類似性などの条件にもよるが、ほとんどの場合に効果が示されている。

3.3 大規模事前学習モデルの活用

近年、大規模なラベルなしのデータで自己教師付き学習(SSL: Self-Supervised Learning)されたモデルに基づく音声言語処理の研究が盛んになっている[5]。音声認識においても、少量のデータで学習効果が高いことが示されている。特に、多数の言語で学習された XLS-R[6]は、低資源の言語にも有用であることが示されている。

XLS-R のような SSL による大規模事前学習モデルの音声認識への活用には 2 通りの方法がある。1 つは、このモデルに認識単位に相当する出力層を付加して、CTC 損失によりモデルをファインチューニングする方法である（図 1(b)）。この場合も、モデルのトランスフォーマー層のみをファインチューニングし、CNN (Convolutional Neural Network) 層は固定する場合が多い。この方法は簡潔で、特にデータ量が少ない場合に有効である。著者らの研究でも、クメール語の音声認識[7]やアイヌ語の音声認識[8]において、5〜10 時間程度の学習データで性能が収束することが示されている。もう一つの方法は、SSL による大規模事前学習を表現学習とみなし、このモデルを特徴抽出器として用い、さらに Conformer のようなエンコーダ・デコーダモデルに基づく音声認識モデルを構成するものである（図 1(c)）。この場合は、事前学習モデルのパラメータは固定する。

これに加えて、SSL ではなく、多言語の大規模なデータで（弱）教師付き学習された Whisper[9]のようなモデルを活用することも考えられる。Whisper 自体はトランスフォーマーのエンコーダ・デコーダモデルに基づく認識器であるので、これをファインチューニングする（図 1(d)）。

図 2 アイヌ語アーカイブのユーザインタフェース ©国立アイヌ民族博物館,公益財団法人アイヌ民族文化財団

3.4 その他の手法

音声合成や音声変換により多様な話者のデータの増強[10,11]や適応[12]を行うことも検討されている。

また、音素ラベルの情報に加えて、調音素性（調音位置や調音様式）のような低レベルの情報を活用することが考えられる。各音素は調音素性の組合せで記述できるので、ニューラルネットワークの線形層で階層的に表現可能である[13]。

4. アイヌ語アーカイブの処理

4.1 アイヌ語

アイヌ民族は北海道、南樺太、千島列島に先住し、19 世紀中頃には約 2 万人いたとされる。しかし、明治政府の北海道開拓と同化政策により、母語話者の数が大きく減少し、2009 年には UNESCO により「極めて深刻」な消滅の危機にあると認定されるに至っている。

アイヌ語は膠着的、複統合的な特徴を示し、日本語とはいくつかの類似点や語彙の借用はあるものの、系統不明の孤立した言語である。アイヌ語は、大まかに北海道アイヌ語、樺太アイヌ語、千島アイヌ語の 3 つに分類され、各々さらに細かい方言がある。本稿では、最も多くのデータが利用可能な沙流方言を主に扱っている。

アイヌ語には、開音節と閉音節の両方があるが、音節頭と音節末の子音の数は最大 1 つである。つまり、子音を C、母音を V とすると、音節は V, CV, VC, CVC のいずれかになる。後述する『アイヌ語アーカイブ』では、表音表記は北海道ウタリ協会編纂の『アコロイタク』のものに準じている。母音 V は{a, i, u, e, o}の 5 つで、子音 C は{k, s, t, n, h, m, y, r, w, c, p}である。発音の脱落を表す“_”と人称接続を表す“=”も用いられる。概ね単語の単位で、空白で区切られる。

（例） Wakka ci=ku
水 私たち・飲む

4.2 アイヌ語アーカイブ

アイヌ語の口頭伝承の収録は、1970 年頃以降、個人レベルから市町村レベルにおいて進められた。中でも、白老町のアイヌ民族博物館では、大規模な『アイヌ語アーカイブ』（図 2）が構築され、2020 年のウポポイの開設にあわせて国立アイヌ民族博物館に承継された。収集された音声は日本語部分も含め 670 時間にも及び、書き起こしも進められているが、アーカイブとして公開に至っていたのは（著者らが研究開発に着手した）2018 年時点で数十時間であった。なお後述の音声認識実験で併用した平取町の二風谷アイヌ文化博物館で収集された音声アーカイブ（約 24 時間）も現在はこちらからも参照できる。

アイヌ口頭伝承には大きく以下の３つがある。

- ウウェペケレ（民話、散文説話）：散文調で語られる人間視点の物語
- ユカラ（英雄叙事詩）：節をつけて語られる英雄（超人）の物語
- カムイユカラ（神謡）：節をつけ、折り返しのフレーズを持つ神視点の物語

以降に述べる音声認識の研究では、ウウェペケレを対象としている。

4.3 アイヌ語音声認識

本節では、著者らが行ってきたアイヌ語の音声認識の実験結果の概要を示す。上記の両博物館から提供頂いたアーカイブの音声データからウウェペケレのみを選択して用いた。概略を表 1 に示す。詳細は文献[8][12]を参照されたい（ただし、学習セットと評価セットの分割の条件は本稿と異なる）。最も着目すべきは、話者の数が少ないことに加えて、データ量に大きな偏りがあることである。話者 KM のみで約 60%、話者 UT とあわせると 80%以上を占めている。これは低資源言語の音声認識においてみられる課題である。高齢者の自然発話で、収録条件もそれほどよいわけでないが、話者にクローズドな条件で行えば、95%に近い文字（音素）認識率を実現できる[8]。ここでは表 1 に示すように、話者オープンの条件（評価 1）、及び方言も異なる条件（評価 2）で実験を行った。

認識に用いたモデルは以下の通りである（図 1 参照）。

(1) Conformer：4 層 CNN＋12 層エンコーダ＋6 層デコーダのモデルをスクラッチで学習
(2) XLSR: XLS-R[6]を CTC でファインチューニング。300M と 1B のモデルを用いた。
(3) XLSR+Conformer: XLS-R[6]を固定した特徴抽出器として用い、(1)の Conformer を接続して学習。300M のモデルを用いた。
(4) Whisper: トランスフォーマーに基づくモデル全体を学習データでファインチューニング。言語タグは英語<|EN|>を使用した。Small(241M)と Large-v2(1.54B)を比較した。

表 1 音声認識実験に使用したデータの概要

	学習	評価 1	評価 2
方言	沙流	沙流	静内
話者	4 名 (KM, UT, KT, HS)	2 名 (HY, KK)	1 名 (OS)
時間	32.2 時間	3.2 時間	16.0 時間

表 2 音声認識結果（文字誤り率%）

	パラメータ数	評価 1	評価 2
Conformer	31M	10.7	22.2
XLSR 300M	317M	11.0	19.5
XLSR 1B	966M	9.5	17.6
XLSR+Conformer	351M	8.4	16.0
Whisper small	241M	7.8	15.9
Whisper large	1.54B	6.8	14.8

いずれの手法においても、認識の単位は SentencePiece を用いて得られた 500 トークンのサブワードとした。ただし、以前の研究[8]では、音節を単位とする方が高い性能を得ている。

結果を表 2 に示す。30 時間もの学習データがあったので、標準的な Conformer モデルも学習できたが、大規模データで事前学習されたモデルの方が高い性能を得ている。特にこのデータ量では、XLSR に Conformer を加えた方が、単純に CTC だけのモデルよりも有意に高い性能となった。一方、Whisper のファインチューニングにより、さらに高い性能が得られた。方言の異なる話者での評価 2 においては、全般に誤り率が 2 倍程度に悪化するが、モデル間の比較の傾向は同じである。なお、このデータは SN 比がかなり悪いことも認識率が低い要因である。

次に、学習データ量を変化させた場合の認識精度の変化を表 3 に示す。Conformer の学習は 10 時間より少ないデータでは難しいことがわかる。これに対して事前学習モデルのファインチューニングは 5 時間程度でほぼ収束しているが、Whisper の方が一貫して高い性能となっている。

表 3　学習データ量と認識精度(文字誤り率)の関係

	1h	5h	10h	32h
Conformer	77.4	41.9	16.9	10.7
XLSR 300M	75.5	16.3	13.6	11.0
XLSR 1B	25.8	12.2	10.6	9.5
Whisper small	21.0	10.4	9.0	7.8
Whisper large	22.3	9.5	8.0	6.8

これらの音声認識の精度は、アイヌ語がわかる人が希少な現状では価値があるが、実際に書き起こしに使えるかは今後の検証による。

4.4　音声とテキストとのアライメント

ある程度の音声認識ができれば、音声とテキストの時間同期（アライメント）が可能になる。音声とその書き起こしが与えられたときに、単語や文の単位で時間情報を付与することを考える。これは、アーカイブ上（図 2）でテキストをブラウジングしたり検索した上で、当該箇所の音声を頭出しで聴取することを可能にするもので、長い音声のアーカイブでは不可欠な機能である。

DNN-HMM に基づくモデルでは、与えられたテキストに対応する HMM（Hidden Markov Model）を構成して、Viterbi アルゴリズムによりアライメントをとることができた。これは、言語モデル(bigram)の値を極端な値（テキストにある単語の連鎖のみほぼ 1, 他はほぼ 0）に設定することでも近似的に実現できた。しかし、長時間の音声ではこれらの方法は現実的でない上に、end-to-end モデルではそもそも言語モデルが明示的にない。

そこで、通常の音声認識を行った上で、認識結果のテキストと正解の書き起こしテキストのアライメントをとる方法が実用的である（図 3）。この場合、時間情報が文字単位の分解能になり、挿入・脱落や置換もあるが、文レベルの区分化であれば概ね問題ない。対応がとれたら、認識結果に付与されている時間情報をコピーする。エンコーダ・デコーダモデルでは、Cross-Attention の重みに時間単調性の保証がない上に、広範囲に及ぶ場合があるので、CTC で当該文字が出された時間フレーム（複数フレームで連続した場合は中央値）を抽出する。

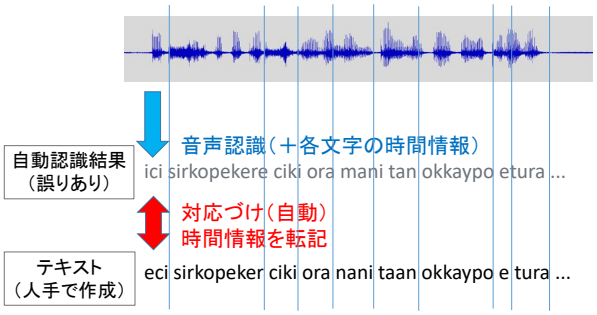

図 3　音声とテキストのアライメント処理

試行する中でわかったのであるが、アイヌ語と日本語が混じっている音声データでも単一のモデルでこの処理は十分可能である。

これにより、『アイヌ語アーカイブ』で書き起こしが作成されていた音声データの大半を自動処理することができた。1 時間の音声データに対して、人手で文単位の時間付与作業を行うと丸 1 日要していたそうで、アーカイブの公開に大きな貢献ができた。現在は公開データが約 300 時間に増えている（ただし日本語の発話区間も多い）。

4.5　音声合成

アイヌ語の音声合成も試みている。Tacotron 2[14]や VITS[15]などのモデルを使用した。話者 KM のように 20 時間くらいのデータがある話者もいるが、録音品質があまりよくないのが問題となるので、雑音除去などの前処理が必要となる。

これにより、書き起こしが残されているが、音源が逸失した音声を「再現」することが可能になった。また、アイヌ語でスピーチを行う際の参考資料としても用いられている。

アイヌ語アーカイブの場合、話者全員が高齢者で、大半が女性である。このように年齢や性別に偏りがある傾向は、消滅の危機にある言語では概ね該当すると考えられる。教材や様々なコンテンツに使用することを考えると、多様な話者で生成できることが望ましい。そのため、音声変換と組み合わせる方法などを検討している。この話者の偏りの問題は、音声認識においても深刻である。

一方で昨今の生成 AI 全般において、元データの著作権や肖像権の扱いが問題になっている。アイヌ関係者も本件について敏感になっており、たとえ学術目的であっても、同意を得ずに音声合成を行うことは、遺族のみならず、アイヌ関係者全

体から強い反発が予想されることも肝に銘じておく必要がある。

5. おわりに

アイヌ語は少数言語の中でも、（関係者の努力により）例外的に収録音声データの量が多く、話者数と話者層の偏りを除くと低資源言語ではないかもしれない。ただし、沙流方言・静内方言の話者しか扱っていないので、他の方言にどの程度適用できるかは不明である。

琉球諸語については、多様な方言をカバーする『沖縄伝承話データベース』や『しまくとぅばアーカイブ』が構築されている。方言間の差異も考慮しながら、少数言語をどのようにモデル化するかは、今後の大きな研究課題である。

6. 謝辞

本研究の実施にあたり、アイヌ語に関する指導から音声認識・合成の評価まで多大な協力を頂いた札幌学院大学の奥田統己先生に深く感謝する。本研究には、三村正人研究員（現在 NTT）及び李在詠君の貢献も大きい。アイヌ語アーカイブデータの提供及び研究の議論を頂いた国立アイヌ民族博物館、平取町立二風谷アイヌ文化博物館の方々に深く感謝する。本稿を閲読頂いた京都大学名誉教授の壇辻正剛先生に感謝する。

文　献

[1] K. Jokinen. Researching Less-Resourced Languages - the DigiSami Corpus, Proc. LREC 2018.
[2] C. Watson, P. Keegan, M. Maclagan, R. Harlow and J. King. The motivation and development of MPAi, a Maori Pronunication Aid. Proc. Interspeech, 2017.
[3] 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, Vol.74, No.7, pp.381--386, 2018.
[4] A.Graves, S.Fernandez, F.Gomez, and J.Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. Proc. ICML, 2006.
[5] 河原達也, 三村正人. 大規模事前学習モデルに基づく音声認識. 日本音響学会誌, Vol.79, No.9, pp.455--460, 2023.
[6] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. Proc. Interspeech, 2022.
[7] K. Soky, S. Li, C. Chu, and T. Kawahara. Domain and language adaptation using heterogeneous datasets for wav2vec2.0-based speech recognition of low-resource language. Proc. ICASSP, 2023.
[8] 松浦孝平, 三村正人, 河原達也. アイヌ民話アーカイブに対する音声認識. 自然言語処理, Vol.28, No.3, pp.824--846, 2021.
[9] A. Radford, J-W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision, arXiv:2212.04356, 2022.
[10] C. Du and K. Yu. Speaker Augmentation for Low Resource Speech Recognition. Proc. ICASSP, 2020.
[11] M. Bartelds, N. San, B. McDonnell, D. Jurafsky, M. Wieling. Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation. Proc. ACL, 2023.
[12] K.Matsuura, M.Mimura, S.Sakai, and T.Kawahara. Generative adversarial training data adaptation for very low-resource automatic speech recognition. Proc. Interspeech, pp.2737--2741, 2020.
[13] 李在詠, 河原達也. 調音属性に関する知識の埋め込みによるアイヌ語音声認識の改善. 情報処理学会研究報告, SLP-149-12, 2023.
[14] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, RJ Skerry-Ryan, R. A. Saurous, and Y. Agiomyrgiannakis, Y. Wu. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. Proc. ICASSP, 2018.
[15] J.Kim, J. Kong, and J. Son. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. Proc. ICML, 2021.

河原達也
1989 年京都大学大学院工学研究科修士課程修了. 現在、京都大学情報学研究科教授. 音声情報処理, 特に音声認識及び対話システムに関する研究に従事. 京大博士（工学）. ASRU 2007 General Chair, ICASSP 2012 Local Arrangement Chair, APSIPA ASC 2020 General Chair, SIGDIAL 2024 General Chair, 言語処理学会理事, 情報処理学会理事, APSIPA T-SIP 編集委員長, ISCA 理事・事務総長, APSIPA 会長を歴任. IEEE Fellow.

松浦孝平
2021 年京都大学大学院情報学研究科修士課程修了. 現在, 同研究科博士後期課程在学中、および日本電信電話株式会社人間情報研究所研究員. 日本音響学会会員.