

アイヌ民話アーカイブに対する音声認識

松浦 孝平[†]・三村 正人[†]・河原 達也[†]

本稿では、アイヌ民話（ウウェペケレ）の音声認識に関する我々の取り組みについて述べる。まず、2つの博物館から提供されたアイヌ語アーカイブのデータを元に、沙流方言を対象としたアイヌ語音声コーパスを構築した。次に、このコーパスを用いて注意機構モデルに基づく音声認識システムを構成し、音素・音節・ワードピース・単語の4つの認識単位について検討した。その結果、音節単位での音声認識精度が最も高くなることがわかり、話者クローズド条件と話者オープン条件のそれぞれについて、音素認識精度で93.7%と86.2%、単語認識精度で78.3%と61.4%を実現した。音声認識精度が話者オープン条件において大幅に低下する問題に対して、CycleGANを用いた教師なし話者適応を提案した。これは、学習データ内の話者の音声から認識対象話者の音声への写像をCycleGANに学習させ、学習データ内の音声を全て認識対象話者風の音声に変換するものである。本手法によって最大で相対60.6%の音素誤り率の改善を得た。さらに、日本語とアイヌ語が混合した音声における言語識別についても検討を行い、音素認識と単語認識を用いた構成で一定の識別性能を達成できることを示した。

キーワード：音声認識, 言語識別, アイヌ語

Automatic Speech Recognition for the Archive of Ainu Folklores

KOHEI MATSUURA[†], MASATO MIMURA[†] and TATSUYA KAWAHARA[†]

In this article, our work on the speech recognition of Ainu folklores (Uwepeker) is described. First, we constructed an Ainu speech corpus for the Saru dialect based on the data provided by two museums that had constructed the Ainu archive. Next, we built an automatic speech recognition (ASR) system based on an attention-based encoder-decoder model, and compared four recognition units of phones, syllables, word pieces, and words. With the syllable unit, we achieved a phone recognition accuracy of 93.7% and 86.2%, and word recognition accuracy of 78.3% and 61.4% for the speaker-closed and speaker-open conditions, respectively. To address the problem of significant degradation in the speaker-open condition, an unsupervised speaker adaptation method using a CycleGAN is proposed. In this method, mapping of the speaker's voice in the training data to the target speaker's voice is learned by a CycleGAN, that converts all speech in the training data into the target speaker's

[†] 京都大学, Kyoto University

speech. This method reduced the phone error rate by up to 60.6%. In addition, we investigated language identification in Japanese and Ainu mixed speech and realized reasonable performance by cascading phone and word recognition modules.

Key Words: *Automatic Speech Recognition, Language Identification, Ainu*

1 はじめに

アイヌとは北海道・樺太・千島列島に住む民族であり、独自の文化と言語を持っているが、これらは19世紀後半から行われた同化政策の影響で急速に失われていった。これに対して、20世紀後半からアイヌ文化保護活動が活発に行われており、その過程で多くの口頭伝承の音声収録されてきた。このような録音資料はアイヌ文化を理解するうえで重要な役割を果たすものであるが、アイヌ語に関する専門知識を持った人材の不足からその大半は未だ書き起こされておらず、十分に活用されていないというのが現状である。そこで、アイヌ語に対する音声認識システムを構築することが強く求められているが、これまで本格的な研究は行われていない。

近年、音声認識技術は大規模コーパスと深層学習の導入によって劇的な進歩を遂げ、実用的な水準に達している (Anmol et al. 2020; Lüscher et al. 2019)。その代表的なもので、現在最も用いられている DNN-HMM ハイブリッドモデル (Hinton et al. 2012) は、音響モデル、言語モデル、発音辞書からなる階層構造を持っている。一方で、音響特徴量列から直接ラベル列へと変換する End-to-End モデル (Chorowski et al. 2015) がその単純な構造と応用の容易さから活発に研究されており、ハイブリッドモデルと同等以上の性能を達成しつつある。

しかしながら、これらの深層学習を適用するためにはかなり大規模な学習データが必要となるため、低資源言語において実現することは難しい。本研究で構成するアイヌ語音声コーパスは40時間の音声データからなるが、これは『日本語話し言葉コーパス (CSJ)』 (Maekawa et al. 2000) や英語の LibriSpeech コーパス (Panayotov et al. 2015) などと比較して10分の1以下であり、アイヌ語もまた低資源言語に分類される。低資源言語の音声認識のために、表現学習 (Hermansky et al. 2000; Tóth et al. 2008) やマルチリンガル学習 (Cui et al. 2015; Thomas et al. 2013; Grézal and Karafiát 2014) が検討されている。表現学習では、主要言語の大規模コーパスで学習された多層パーセプトロンを特徴抽出器として使用する。マルチリンガル学習では、認識対象でない言語のデータで学習データの量を補完して音声認識モデルを学習させる。これらの手法はアイヌ語音声認識においても有用であることが予想されるが、アイヌ語音声コーパスは話者数の少なさと話者毎のデータ量の偏りという特徴を持っており、上記の手法を単純に適用できない。

また、アイヌに関する一次資料は日本語とアイヌ語が混合した音声であるが、高い音声認識性能を得るためにはアイヌ語の発話区間をあらかじめ抽出しておく必要がある。音声データにおける言語識別の従来手法として、フォルマントに基づくもの (Leonard and Doddington 1974),

音素認識モデルと言語モデルを組み合わせたもの (Kirchhoff and Parandekar 2001), 音響特徴量列から直接言語ラベルを出力するもの (Cai et al. 2019) などが存在するが, 日本語アイヌ語混合音声には一人の話者が複数の言語を流暢に話すという点で上記の研究対象より難度が高い。

本稿の構成を以下に記す。まず, 我々は白老町アイヌ民族博物館と平取町アイヌ文化博物館から提供されたアイヌ語アーカイブのデータを元にアイヌ語音声コーパスを構成する。次に, 本コーパスを用いたアイヌ語音声認識において, 音素・音節・ワードピース・単語の4つの認識単位を比較する。実験は, 学習セットと評価セットで話者が同一である話者クローズド条件と, 話者が異なる話者オープン条件で行う。話者オープン条件での認識性能の低下を緩和するために, CycleGANを用いた声質変換技術による教師なし話者適応を提案する。最後に, 日本語とアイヌ語が混合した音声における言語識別について検討を行う。

2 アイヌ語と従来研究

2.1 アイヌ語について

アイヌとは, 主に北海道・樺太・千島列島に住み, 狩猟採集を生業としてきた日本の民族の一つであり, 19世紀中ごろには約2万人存在していたとされる (Hardacre 1997)。しかし, 19世紀後半における明治政府の同化政策を経てその母語話者の数は急速に減少し, 2009年には国際連合教育科学文化機関 (ユネスコ) によって北海道アイヌ語は「極めて深刻」な状況にあると分類されるに至っている (Alexandre 2010)。このような状況に対して, 20世紀後半以降アイヌの民話や歌謡が積極的に収集されている。例えば, 国立アイヌ民族博物館 (旧: アイヌ民族博物館) は約700時間のアイヌに関する録音音声を保持している。このような音声データはアイヌ文化の理解をより一層深めるための手がかりとなり得るが, その大部分は書き起こされておらず, 十分に研究されていないのが現状である。

2.1.1 アイヌ語の概説

アイヌ語は膠着的, 複統合的, 抱合的な特徴を示し (ブガエワ 2014), 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, 系統不明の孤立言語である。アイヌ語は大まかに北海道アイヌ語, 千島アイヌ語, 樺太アイヌ語の三つに分類され (田村 1988), それぞれがさらに細かい方言の分類を内包する。本節では, 北海道アイヌ語に属する沙流方言について解説を行う。

音声認識の観点からアイヌ語に関して重要な特徴を述べる。アイヌ語の音節には開音節 (CV など) と閉音節 (CVC など) の双方があるが, いずれにおいても音節頭と音節末に許容される子音の数は最大で1であり, 英単語の *strike* における ‘*str*’ のような子音群は存在しない (田村 2013)。例えば, *apkas* (「歩く」) は ‘*ap*’ と ‘*kas*’ の2つの閉音節から構成される。また, アイヌ語には複合語が多く存在する。例えば, *atuykorkamuy* (「ウミガメ」) という単語は *atuy*

(「海」), *kor* (「所有する」), *kamuy* (「神」) という 3 つの要素 (形態素) に分解でき, なおかつ各々に対応する単語が存在する (田村 1996).

2.1.2 アイヌ口頭伝承の分類

アイヌ口頭伝承は大きく次の 3 つに分類される. ユカラとカムイユカラは節をつけて歌われるのに対し, ウウェペケレは単調に語られる. アイヌ語音声資料の大部分はこのウウェペケレが占めるとされている.

ユカラ (英雄叙事詩): 節をつけて語られる英雄の物語

カムイユカラ (神謡): ユカラ を神の視点から歌ったもの

ウウェペケレ (民話, 散文説話): 会話調で語られる物語

2.1.3 表記体系

アイヌ語にはもともと固有の筆記体系が存在しないが, 現在は『アコロイタッ』 (北海道ウタリ協会 1994) に範示される表記法に基づいて記述されるのが一般的である. この表記法では, {*a, c, e, h, i, k, m, n, o, p, r, s, t, u, w, y*} の 16 種類のローマ字が用いられる. これらのローマ字をここでは便宜上「音素」と呼ぶ. アイヌ語話者が日本語を発した場合, 上記に加えて {*b, d, g, z*} が使用されることがある. さらに, 人称接続を表す {=} が用いられる他, 発音の脱落を表す {-, _} などが表記されることもあるが, これらは明示的に発音されない. 本稿では, アイヌ語の文から空白記号と {=} 記号によって分割される単位を「単語」と定義する. 以下にアイヌ語の文とその構文解析の例を一つ挙げる.

原文	<i>mos=an _hine inkar'=an</i>				
日本語訳	目が覚めて見ると				
解析	<i>mos</i>	<i>=an</i>	<i>hine</i>	<i>inkar</i>	<i>=an</i>
	起きる	4sg.	そして	見る	4sg.

ここで **4sg.** は不定人称単数を表すが, 特に引用文中の自称として使用されており, 単なる一人称 (物語の語り手自身) とは区別される.

2.2 アイヌ語における言語処理と音声認識

言語処理の分野において, アイヌ語を扱った研究はこれまでに数件存在している. 瀬沼と相澤 (Senuma and Aizawa 2018) は, 『アイヌ神謡集』 (Yukie 1923) に含まれるアイヌ語のテキストにおいて, 単語間の関係を表現する 依存関係木の アノテーションを行った. Nowakowski (Nowakowski et al. 2019) らは, 『アイヌ語会話字典』から抜粋された 2500 程度のアイヌ語のフ

レーズや文章にアノテーションを付与した学習データを作成し、サポートベクトルマシンによる品詞タグ付けツールを開発した。アイヌ語の音声認識は、低資源言語音声認識の研究の一環として、Anastasopoulos ら (Anastasopoulos and Chiang 2018) によって試みられている。約 2.5 時間のアイヌ語の民話が使用されているが、アイヌ語の音声認識は主な目的ではなく、また音素誤り率は 40% 程度と実用的なレベルには至っていない。

3 アイヌ語音声コーパス

3.1 コーパスの概要

本研究で使用する音声コーパスは、白老町アイヌ民族博物館¹（以下、「白老」と表記）と平取町アイヌ文化博物館²（以下、「平取」と表記）から提供されたデータに含まれる民話（ウウエペケレ）から構成される。提供されたデータのうち英雄叙事詩（ユカラ）や神謡（カムイユカラ）は数が少なく、なおかつ 2.1.2 節で述べたように民話とは大きく異なる性質を持つため、本研究の対象にはせずコーパスにも含めないことにした。表 1 に、提供されたデータに含まれる民話の一部とその日本語訳を具体例として例示する。

提供された民話は合計 114 個で、約 40 時間の音声データとその書き起こしからなっており、書き起こしは 2.1.3 節で述べた表記方法に従っている。単語を区切るスペースと {=} が入れられており、アーカイブ閲覧の利便性の点から 5 から 15 語程度の意味的なまとまりで改行されている。話者数が 8 名ときわめて少数であり、かつデータ量が話者ごとに大きく偏っているが、この特徴は収録当時の各話者の知識量や体調などに収集できるデータ量が依存したためである。また、白老の音声データは 1980 年代に収録されている一方、平取のデータは 1969 年に収録されている点で、話者ごとに収録時期に大きな時間的隔たりが存在し、録音環境も統一されてい

原文	日本語訳
<i>Samormosir mosir noski ta</i>	隣の国の真ん中で
<i>a=kor hapo i=resu hine</i>	母が私を育てて
<i>oka=an pe ne _hike</i>	暮らしていました。
<i>kunne hene tokap _hene</i>	夜も昼も
<i>yam patek i=pareoyki</i>	栗だけで養われ
<i>yam patek a=e kusu</i>	栗だけを私は食べていたので
<i>somo hetuku=an pe ne kunak</i>	大きくならないと
<i>a=ramu a korka</i>	思っていました

表 1 話者 KM による説話『ポロシルンカムイになった少年』より一部抜粋

¹ <http://ainugo.ainu-museum.or.jp/>

² <http://www.town.biratori.hokkaido.jp/biratori/nibutani/>

ない。収録は自宅などでカセットデッキを用いて行われており、周囲の騒音が多いわけではないが、SN 比は高くない。

3.2 データの調整

音声認識の学習を容易にするために、提供されたデータを加工した。まず、テキストから 2.1.3 節で述べた文字セットの内、{=, _} の補助記号を以下の例のように除去する。

原文	<i>uymam'=an wa isam=an _hi okake ta</i>
調整後	<i>uymam=an wa isam=an hi okake ta</i>

人称接続を表す記号 {='} も補助記号であるが、アイヌ語のテキストに幅広くみられ、なおかつ文法的に重要な情報を保持しているため、除去しないことにした。

音声認識モデルを学習するにあたり、音声データは適切な長さに分割されている必要がある。提供された時点で音声データは意味のまとまりに従って分割されていたが、発声の途中に分割点があるなど音声認識に不向きな分割がなされていたため、無音区間で区切られた単位 (IPU) で再分割を行った。以下、この IPU を便宜上「発話」と呼ぶ。

3.3 コーパスの分析

表 2 に、アイヌ語音声コーパスにおける 8 名の話者ごとのデータ量、民話の数、発話 (IPU) の数、単語数を示す。各話者には便宜上 ID を与えている。このうち、話者 KM と UT の元データは白老から、その他 6 名のデータは平取から提供されたものである。

提供された元データにおける話者ごとのデータ量の偏りはコーパスにも引き継がれており、例えば話者 KM の発話量はコーパス全体の半分以上を占めている。さらに、話者 KM は 1 発話当たりの平均単語数が 15.8 語であるのに対し話者 NN は平均 6.0 語であるなど、話者ごとにデータの性質も大きく異なっている。

一方で、各話者の発話時間は最低でも一時間半と、いずれの話者に関してもある程度多くの音声データが存在している。

話者 ID	KM	UT	KT	HS	NN	KS	HY	KK	合計
発話時間 (h:m)	19:41	7:15	3:14	2:06	1:45	1:43	1:37	1:35	38:55
発話時間 (%)	50.6	18.6	8.3	5.4	4.5	4.4	4.1	4.1	100.0
民話の数	29	26	20	8	8	11	8	7	114
発話数	9170	3610	2273	2089	1572	1302	1220	1109	22345
単語数	144984	45865	26424	16357	13636	12730	12575	14676	287247

表 2 話者ごとのデータ量

4 アイヌ語音声認識システム

本節では、アイヌ語音声コーパスを用いた音声認識システムを構成するに際して、音素、音節、ワードピース、単語の4つの認識単位の比較と、マルチリンガル学習の検討を行う。音声認識モデルには注意機構モデルのエンコーダに CTC 損失関数を付加した End-to-End モデル (Hori et al. 2017) を用いる。

4.1 認識単位の比較

End-to-End 音声認識モデルは認識単位をある程度自由に設定することが可能である。単語単位で認識を行った場合、推論時に長期的な文脈を考慮することができるが、語彙サイズが大きくなるため学習が難しい。逆に、音素や文字単位の場合は語彙サイズが小さくなるが、ラベル長が長くなり長期的な文脈を掴むことが難しくなってしまう。以上のトレードオフの関係を考慮し、アイヌ語音声認識に最適な単位を検討する。認識単位の候補は、音素、音節、ワードピース、単語の4通りであり、表3に各単位で表記した文の例を示す、また、各認識単位に関する説明を以下に述べる。

音素単位 2.1.3 節で述べた通り、ここではアイヌ語を表記するときに使用されるローマ字を便宜上音素とする。また、人称接続を表す記号‘=’と単語境界を表す特殊記号‘<wb>’を出力候補に加え、認識結果を単語単位へと復元できるようにする。

音節単位 アイヌ語における音節は、V, CV, VC, そして CVC の4通りを取りうる。ここで、‘C’ と ‘V’ はそれぞれ子音と母音を表す。2.1.3 節に示される文字セットの内、音素 {a, e, i, o, u} が母音であり、その他が子音に分類される。以下に、単語から音節へと分割するための手続きを示す。音節の項と同様に、人称接続を表す‘=’と単語境界を表す‘<wb>’を含める。

- (1) 1文字の単語はそのまま音節とする。
- (2) 連続するCとVが存在する場合、そこに音節境界を挿入する。

$$R^*\{CC, VV\}R^* \rightarrow R^*\{C-C, V-V\}R^*$$

$$R := \{C, V\}$$

原文	<i>a=saha i=kokopan wa</i>
音素	<i>a = s a h a <wb> i = k o k o p a n <wb> w a</i>
音節	<i>a = sa ha <wb> i = ko ko pan <wb> wa</i>
ワードピース	<i><wb>a = saha <wb>i = ko ko p an <wb>wa</i>
単語	<i>a = saha i = <unk> wa</i>
日本語訳	姉はダメだと言って

表3 各認識単位の例

- (3) 3つ以上の音素からなる部分列について, 語頭のVの直後に音節境界を挿入する.

$$VCR^+ \rightarrow V-CR^+$$

- (4) CV か CVC が残るまで, 左から右へ CV の直後に音節境界を挿入する.

$$(CV)^*\{CV, CVC\} \rightarrow (CV-)^*\{CV, CVC\}$$

ワードピース単位 単語を部分列へと分割する別の方法として, Byte Pair Encoding (BPE) (Sennrich et al. 2016) と 1-gram 言語モデル (Kudo and Richardson 2018) がある. 前者は, 頻出の記号ペアを新しい一つの記号に繰り返し置き換えていくことで語彙サイズを調整する. 後者はテキストの生成確率が最大になる分割方法を探索する. 本稿では後者の 1-gram 言語モデルによるワードピースへの分割を採用し, オープンソースのソフトウェアである SentencePiece トークナイザ³を利用する. 表3におけるワードピース単位の分割例では単語境界 '`<wb>`' が後続の部分列に接続している場合もあるが, これは本ソフトウェアの仕様である. ワードピースの構成方法は音節の構成方法とは独立であり, ワードピースが音節の連結であるとは限らない.

単語単位 コーパスの書き起こしは既に '=' 記号と空白記号で単語単位に分割されている. 本研究では, '=' 記号も1つのラベルとして扱うため, 例えば表3に見られるように, '`a=saha`' というフレーズは {'`a`', '=', '`saha`'} の3つのラベルとして扱われる. また, 語彙サイズを減らすために, 低頻出語は未知語を表す特殊記号 '`<unk>`' で置換する.

4.2 マルチリンガル学習

マルチリンガル学習は, 目的言語以外の言語の音声コーパスを転用し, 学習データ量を補完して音声認識モデルを学習させる手法である. 英語などの主要言語の大規模コーパスを用いる場合 (Çetin et al. 2008) だけでなく, 低資源言語コーパスのみであってもマルチリンガル学習によってデータを増強することは有効であり, これは DNN-HMM ハイブリッドモデル (Cui et al. 2015; Thomas et al. 2013; Grézl and Karafiát 2014) と End-to-End モデル (Toshniwal et al. 2018; Cho et al. 2018) の両方で示されている. 例えば, Toshniwal ら (Toshniwal et al. 2018) はベンガル語などの10の低資源言語を単一の注意機構モデルで学習したところ, 別個に学習した場合と比較して相対約20%の単語誤り率の改善を達成している.

本稿でもマルチリンガル学習について検討を行うことにし, アイヌ語以外の言語として日本語を採用する. これは, 日本語とアイヌ語では, 音素セットが同じであることや, 子音群 (例えば英語の '`strike`' における '`str`') が存在しないことなど共通点が多くみられるためである. このため, 日本語はアイヌ語とのマルチリンガル学習に適していると考えられる. 比較のために英語コーパスを用いたマルチリンガル学習についても実験を行う.

³ <https://github.com/google/sentencepiece>

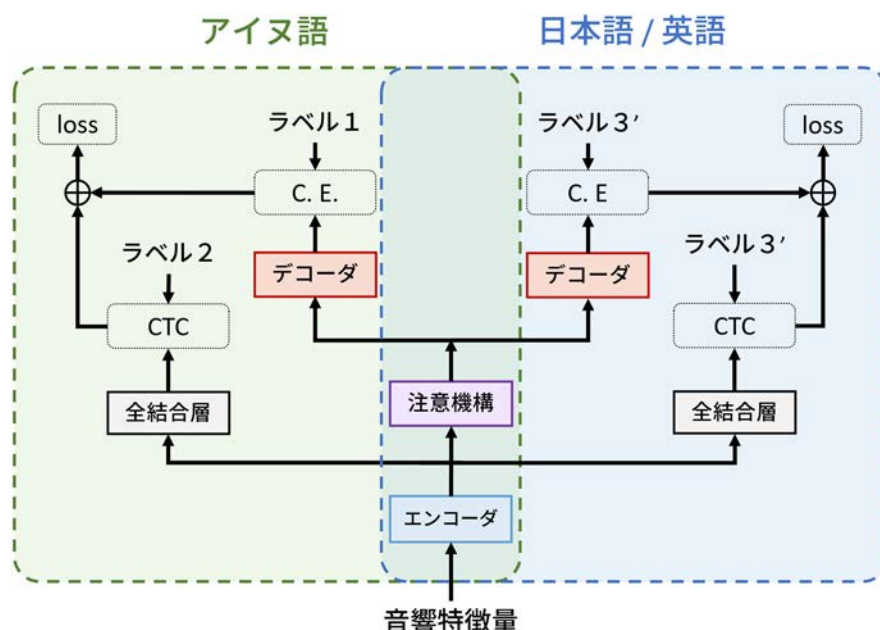


図 1 マルチリンガル学習における音声認識モデルの概略図. C. E. は交差エントロピ損失関数を表す.

マルチリンガル学習における音声認識モデルは図 1 のように構成する. エンコーダと注意機構は言語間で共有されるが, デコーダは言語ごとに別途用意する. 例えば, アイヌ語音声コーパス由来の学習データが入力された場合は図の左側の部分を用いて認識と学習を行う.

4.3 実験条件

実験は, 話者クロズド条件と話者オープン条件の 2 通りについて行う. 話者クロズド条件では, 学習セットと評価セットに含まれる話者が同じである. 具体的には, 各話者から民話を二つずつ選び, 開発セットと評価セットとし, それ以外のデータを学習セットとする. これによって, 開発セットは 2 時間 23 分の 1585 発話となり, 評価セットは 2 時間 48 分の 1841 発話となった. 話者オープン条件では, 評価用話者のデータを除く全てのデータを学習セットとして使用する. ハイパーパラメータなどの調整は話者クロズド条件における開発セットを用いるため, 話者オープン条件において開発セットは存在しない. また, 話者 KM や UT のデータを除くと学習データが大幅に小さくなってしまったため, この 2 話者については話者オープン条件の実験を行わない.

実験条件によって各認識単位のクラスサイズは多少異なるが, 音素単位は 25 クラス前後, 音節単位は約 500 クラスとなった. 単語単位認識の際, 開発セットを用いた予備実験に従い出現回数が 2 回未満のものを未知語ラベル ‘unk’ で置き換えた. この時, 全データにおける語彙サ

イズは 4710 であったが、学習セットに限ると話者クローズ条件下で語彙サイズは 3911 となった。評価セットに出現する語彙のうちこれでカバーされない異なり語の数は 654 で、未知語率（未知語の出現割合）は 4.3% であった。話者オープン条件の場合、未知語率は話者によって異なり、話者 KM, UT, KT, HS, NN, KS, HY, KK の順にそれぞれ 6.7%, 3.3%, 4.9%, 5.9%, 5.7%, 5.9%, 5.1%, 3.5% であった。ワードピース単位認識における語彙サイズは事前実験によって 500 に決定した。このとき、語彙には音素が全て含まれたため、評価セットにおいて未知語となる単語は発生しなかった。

実験条件の詳細を以下に述べる。入力音響特徴量は、連続する 3 つのフレームの 40 次元対数メルフィルタバンク特徴量をスタッキング (Wöllmer et al. 2011) して 120 次元としたものである。ここで、各フレームの窓幅は 25 ミリ秒であり、シフト幅は 10 ミリ秒である。音声認識モデルのエンコーダは 5 層の双方向 LSTM (Schuster and Paliwal 1997) (BiLSTM) からなっており、注意機構付きデコーダは 1 層の単方向 LSTM である。各層は 320 個のノードをもつ。

音声認識の注意機構モデルの学習は、単調性が保証される CTC とマルチタスク学習を行うことにより安定することが知られている (Hori et al. 2017) ため、モデルの損失関数は、下式 (1) のように、注意機構モデルのデコーダにおける交差エントロピー関数 ($\mathcal{L}_{\text{AED}}$) と CTC 損失関数 ($\mathcal{L}_{\text{CTC}}$) の重み付き線形和とした。

$$\mathcal{L}_{\text{all}} = \lambda \mathcal{L}_{\text{AED}} + (1 - \lambda) \mathcal{L}_{\text{CTC}} \quad (1)$$

開発セットを用いて重み λ を 0.0~1.0 の間でチューニングしたところ、0.3~0.8 の間で大きな性能の違いは見られなかったため、本節では双方が同じ重みとなる 0.5 とした。全ての実験において、CTC 損失関数の計算には音素ラベルを使用した。

マルチリンガル学習において、日本語には JNAS コーパス (Itou et al. 1999) (話者 320 名、約 80 時間) を、英語には WSJ コーパス (Paul and Baker 1992) (280 名、約 70 時間) を使用した。これらの正解ラベルにはいずれも音素単位を採用した。

4.4 実験結果

表 4 に、話者クローズド条件と話者オープン条件における単語誤り率 (WER) と音素誤り率 (PER) を示す。話者クローズド条件において WER が 21%, PER が 6% 程度、話者オープン条件においては WER が 39%, PER が 14% 程度が達成できている。両条件において、最も認識精度が高かったのは音節単位で認識を行った場合であり、音節が音響的なマッチングを保ちながらアイヌ語としての言語的な制約のある程度モデル化できていることが示唆される。

単語単位認識における PER は他の認識単位よりも高くなっているが、これは下記のように単語単位モデルの出力にしばしば ‘<unk>’ が現れるためである。この例での単語単位モデルの PER は 30% になる。一方で、音節単位モデルは未知語に対して誤りを含みつつも類似のラベル

話者クローズド条件										
	認識単位	KM	UT	KT	HS	NN	KS	HY	KK	ave.
WER	音素	22.2	28.5	24.2	28.6	27.2	30.6	40.4	36.1	27.9
	音節	13.2	18.4	19.6	29.4	26.7	26.7	38.9	29.0	21.7
	ワードピース	14.4	20.0	21.6	25.0	27.1	23.2	37.8	42.5	22.3
	単語	14.7	19.6	21.3	32.9	27.1	24.6	40.7	31.2	23.1
PER	音素	10.7	16.3	7.9	5.6	7.4	13.6	10.1	14.8	11.1
	音節	3.2	6.9	4.4	7.7	7.9	9.5	9.4	10.7	6.3
	ワードピース	4.7	8.0	5.2	6.7	8.4	6.8	10.4	12.6	7.1
	単語	11.2	12.9	12.6	24.0	17.1	15.4	27.0	20.1	15.9

話者オープン条件										
	認識単位	KM	UT	KT	HS	NN	KS	HY	KK	ave.
WER	音素	—	—	38.8	40.5	41.9	53.1	35.9	54.7	43.4
	音節	—	—	33.4	37.8	37.3	47.2	32.0	48.6	38.6
	ワードピース	—	—	58.4	37.2	38.6	47.9	32.6	48.8	45.7
	単語	—	—	34.0	49.0	39.4	48.9	31.5	84.3	46.6
PER	音素	—	—	14.9	13.9	15.9	21.4	11.2	27.0	17.1
	音節	—	—	10.7	12.6	13.5	16.5	10.3	22.0	13.8
	ワードピース	—	—	41.5	14.1	15.9	19.3	11.5	23.6	23.6
	単語	—	—	24.6	39.9	29.6	33.1	20.4	67.0	34.8

表 4 各認識単位における，話者クローズド条件と話者オープン条件における単語誤り率 (WER; %) と音素誤り率 (PER; %) の一覧.

列を出力でき，PER を 5% に留めている．

正解	<i>i okake un a unuhu a onaha</i>
音素単位モデル	<i>piokake un a unuhu a onaha</i>
単語単位モデル	<i><unk> un a unuhu a onaha</i>

WER は音素誤り率よりも大幅に高くなっている．これは，2.1.1 節でも述べたようにアイヌ語には複合語が多く含まれるため，一つの単語が長くなることと，単語境界の位置をしばしば誤ることの二つに起因する．例えば下記の例では，PER は 0% であるが，WER は 57% である．

正解	<i>nen poka apkas an mak an kusu</i>
認識結果	<i>nenpoka apkas an makan kusu</i>

音節単位認識におけるマルチリンガル学習の実験結果を表 5 に示す．これから，マルチリン

	学習データ	話者クロード条件	話者オープン条件
WER	アイヌ	21.7	38.6
	+ JNAS	21.1	34.8
	+ WSJ	21.3	35.8
	+ 両方	21.4	34.7
PER	アイヌ	6.3	13.8
	+ JNAS	6.0	11.7
	+ WSJ	6.0	12.1
	+ 両方	6.0	11.2

表 5 マルチリンガル学習を行ったときの PER(%) と WER(%)

ガル学習は話者オープン条件で効果的であることがわかる。日本語コーパス (JNAS) を加えた場合の改善幅の方が英語の場合よりも大きく、これは 4.2 節で述べた日本語とアイヌ語の類似性によるものと考えられるが、最も改善幅が大きかったのは両言語を使用した場合であり、この時 WER は相対的に約 10%改善した。

5 CycleGAN を利用した教師なし話者適応

3.3 節で述べた通り、アイヌ語音声コーパスにはわずか 8 名の話者しか含まれておらず、しかも話者ごとのデータ量の分布の偏りが大きい。この結果、4.4 節に示した通り、話者オープン条件における認識性能の大幅な低下が生じた。この問題を「話者スパース問題」と呼ぶことにし、本節ではこの問題について取り組む。

多くの話者のデータを収集することが困難なのは、低資源言語音声認識において一般に当てはまる。例えば、Griko コーパス (Boito et al. 2018) は話者 9 名、Mboshi コーパス (Rialland et al. 2018) は 3 名、Basaa コーパス (Hamlaoui et al. 2018) は明記されていないものの数名とされている。このため、話者スパース問題はアイヌ語を含む低資源言語の音声認識において重要である。

5.1 関連研究と提案手法

話者スパース問題を緩和する手法はこれまでにいくつか提案されている。その一つが、評価データの音声認識結果を疑似的に正解ラベルとみなし、学習済みの音声認識モデルのファインチューニングを行う自己適応学習である。落合らは、音声強調と音声認識を同時に行うモデルにおいて、ファインチューニングを行う際にどの部分のパラメータを固定するべきかについて検討している (Ochiai et al. 2018)。Meng らは、カルバック・ライブラー距離による正則化と話者敵対学習を提案し、ファインチューニング時の過学習を緩和しようと試みている (Meng et al. 2019)。しかしながら、音声認識結果には誤りが多く含まれており、ファインチューニングの際

にその影響を受けるため大幅な認識性能の改善には至っていない。

自己適応学習の他には, i-vector を音響特徴量に付加する手法がある (Audhkhasi et al. 2018). 各話者の特徴を表現する i-vector は話者適応に有用であるが, その計算には不特定話者の平均的な音声を表現する Universal Background Model (UBM) が必要であり, 話者数の少ないコーパスでは実現が難しい。

4.2 節で述べたマルチリンガル学習も低資源言語の音声認識で広く用いられているが, 認識対象の言語でないデータを使用して音声認識モデルを学習させるため mismatches がある。

本節では, 他の言語コーパスにも誤りを含む学習データにも依らない教師なし話者適応の手法を提案する。提案手法の手続きは以下の通りである。まず, 学習データに含まれる音声と評価データに含まれる音声の間の写像を声質変換モデルで学習する。次に, この声質変換モデルを用いて学習データに含まれる音声を全て認識対象話者風の音声に変換する。最後に, 元の音声と変換された音声の両方を使用して音声認識モデルを学習する。以上の手続きによって, 認識対象話者にマッチした音声データを正しいラベルとともに音声認識モデルに与えることができる。本提案手法は, 声質変換モデルの性能が十分に高くなければ実現できないが, 近年 CycleGAN (Radford et al. 2015; Zhu et al. 2017) に基づく高品質かつ学習時にパラレルデータを必要としない声質変換モデルが提案されており (Kaneko and Kameoka 2018; Kaneko et al. 2019), 本研究ではこれを採用する。

5.2 実験条件

本節の実験では話者 KS か NN のいずれかを認識対象の未知話者と仮定し, その他の 7 名を既知の話者とする。これに加えて, 話者 KM の 1 名のみが既知であると仮定する極端な例についても実験を行う。以上に挙げた 4 通りの実験条件の詳細な情報を表 6 に示す。ただし, ALL/XX は話者 XX を除く 7 名全員を表し, 便宜上各実験条件の ID として KM-KS, KM-NN, ALL-KS, ALL-NN を付与する。提案手法には認識したい話者の音声はどの程度必要かを調査するため, CycleGAN は学習セットに含まれる音声の全てと認識対象話者の音声の一部 (1, 5, 10, 20, 30 分。ただし, 音声認識の評価セットは含めない。) を用いて学習する。

CycleGAN へ入力する音響特徴量は, 窓幅 25 ミリ秒, シフト幅 10 ミリ秒で抽出した 40 次元の対数メルフィルタバンク (LMFB) である。CycleGAN-VC2 の元論文 (Kaneko et al. 2019) で

実験 ID	KM-KS	KM-NN	ALL-KS	ALL-NN
既知話者	KM	KM	ALL/KS	ALL/NN
データ量 (h:m)	19:41	19:41	37:11	37:10
認識対象話者	KS	NN	KS	NN

表 6 アイス語音声コーパスにおける 4 通りの実験条件

使用される Mel-cepstral 係数 (MCEPs) 特徴量とは異なっていることに注意されたい。

音声認識モデルには 3 章と同じものを採用し、実験条件は 4.3 節で述べたものと同じである。アイヌ語における認識単位は 4.4 節の結果に基づき音節単位とする。式 (1) における λ は本実験条件であらためて開発セットを用いてチューニングを行い、0.8 とした。比較のために、自己適応学習とマルチリンガル学習についても実験を行う。自己適応学習による話者適応において、使用する評価話者の音声の量を 1, 5, 10, 20, 30 分の場合について音声認識モデルのファインチューニングを行う。マルチリンガル学習に関するモデルと実験条件は 3 章で述べたものと同じであり、JNAS と WSJ の両方を使用する。

5.3 実験結果

表 7 に各手法の PER を示す。表のうち、自己適応学習と提案手法における数値は学習時に使用する認識対象話者の音声量を 1, 5, 10, 20, 30 分と変化させた時の最良の値である。全実験条件において提案手法が自己適応学習とマルチリンガル学習よりも PER が改善している。実験条件 KM-KS と KM-NN では、認識対象話者の音声を 30 分使用した時に、PER がそれぞれ 45.4% から 17.8% (60.6% の相対改善率)、そして 42.6% から 18.3% (57.0% の相対改善率) へと改善している。実験条件 ALL-KS と ALL-NN におけるベースラインの PER は、実験条件 KM-KS と KM-NN のものよりも大幅に改善しており、これから学習データに含まれる話者数の重要性が再確認できる。実験条件 ALL-KS と ALL-NN では既にベースラインモデルの認識性能が高いものの、提案法によって性能はさらに改善している。具体的には、認識対象話者の音声を 20 分使用した時に、実験条件 ALL-KS では 33.7% の相対改善率、実験条件 ALL-NN では 35.9% の相対改善率を達成している。本実験において、評価データを CycleGAN の学習に使用していないため、認識対象話者の異なる音声についても同様の改善が得られることが期待される。なお、ALL-KS 及び ALL-NN においてデータを 30 分まで増やすと提案手法と自己適応学習において PER が上昇する傾向がみられるが、これはオーバーフィッティングによるものと考えられる。

図 2 に、提案手法と自己適応学習の二つの手法について、モデルの学習時に使用する認識対

実験 ID	KM-KS	KM-NN	ALL-KS	ALL-NN
ベースライン	45.4	42.6	15.9	13.8
自己適応学習	45.6	38.8	14.7	12.0
マルチリンガル	30.4	33.6	13.2	11.1
提案手法	17.8	18.3	10.5	8.8

表 7 話者適応の効果 (PER%)。ただし、自己適応学習と提案手法における数値は学習時に使用する認識対象話者の音声量を 1, 5, 10, 20, 30 分と変化させたときの最良の値である。

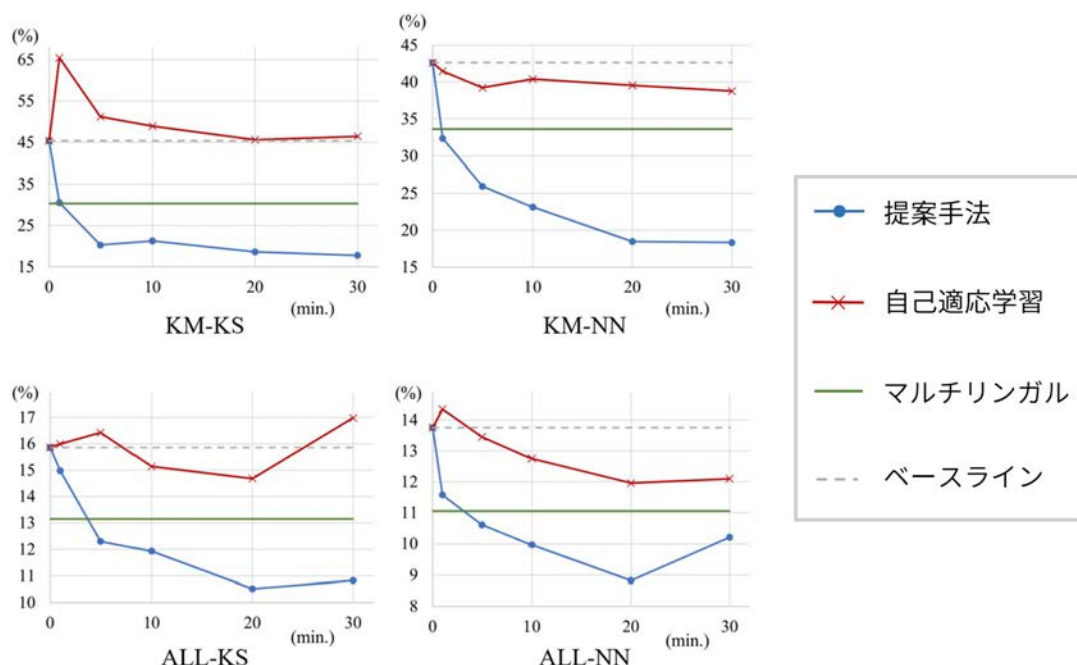


図 2 認識対象話者の発話の使用量を 1, 5, 10, 20, 30 分と変化させたときの PER(%) の変化. ただし, マルチリンガル学習は認識対象話者の音声を使用しないため一定値となる.

象話者の音声量を変化させたときの PER の推移を示す. 全実験条件において, 認識対象話者の音声 が 1 分であっても一定の認識性能の向上を得られることがわかる. これは, 提案手法が認識対象話者のデータ量に対して効率的に機能することを示している. 自己適応学習は本アイヌ語音声コーパスにおいてあまり認識性能の向上に寄与しなかったが, 最初の音声認識結果に含まれる誤りが多すぎるためであると考えられ, アイヌ語以外の低資源言語においても同様の問題が生じることが予想される. また, 認識対象話者の音声 が 5 分あれば, マルチリンガル学習による改善幅を提案法が上回る.

以下に, 実験 KM-KS における認識結果の実際の改善例を挙げる. 提案手法を適用した場合, ある程度誤りは残っているものの, ベースラインの結果と比較して大幅に削除誤りが減少していることがわかる.

正解	<i>a unuhu an a onaha an hine oka an hike iskar emko un</i>
ベースライン	<i>a onaha ne okaymi ki iskar emko</i>
提案手法	<i>a ponomo an a onaha an hine oka an he ki iskar emko un</i>

6 日本語アイヌ語混合音声における言語識別

アイヌ語アーカイブの元データである一次音声資料は日本語の発話とアイヌ語の発話が混合したもの（以下、「言語混合データ」）である。図3に一つの収録音声の中で言語が切り替わる例を示す。この音源には一人のアイヌ語話者（語り手）と日本語話者（聞き手）の発話が含まれており、語り手の民話を中心に雑談や質疑応答が行われている。ここで、語り手はアイヌ語と同様に日本語も流暢に話す。

今回二つの博物館から提供されたデータは既にアイヌ語部分のみが抽出されていたが、未調査の大規模なアイヌ録音資料を処理・分類するにあたって、この抽出作業も自動化できることが望ましい。ここで、言語混合データを多人数かつ大量に整備することは困難であるため、学習には単言語コーパスのみを使用する。また、図3に例示したように、アイヌ語と日本語の発話は各々連続する傾向にあり、大局的な構造が把握できれば十分であるため、本研究では発話ごとに言語識別を行う。

6.1 関連研究と提案手法

音声データの言語識別はもともと (House and Neuburg 1977) に示される音素列の重要性に基づいて、音素認識モデルに言語モデルを組み合わせる手法が模索されてきた (Kirchhoff and Parandekar 2001; Tong et al. 2009)。しかし、DNN の登場以降、i-vector (Lopez-Moreno et al. 2016) や対数メルフィルタバンク特徴量を自己注意プーリング (Cai et al. 2018) によって固定長のベクトルに変換したものを DNN への入力とする (Cai et al. 2019) など、音響特徴量から直接言語識別を行う手法が活発に研究されている。

本稿で扱う言語混合データは一人の話者が複数の言語を流暢に話すという点で従来の言語識別タスクよりも難易度が高い。そこで、音素認識と単語認識を段階的に適用し、言語識別を行う。以下に、処理の手順を示す。まず、音響特徴量列を音素認識器にかけ、音素列に変換する (A2P)。音素列に変換することで、単言語コーパスごとの音響的な差異を緩和する。次に、音

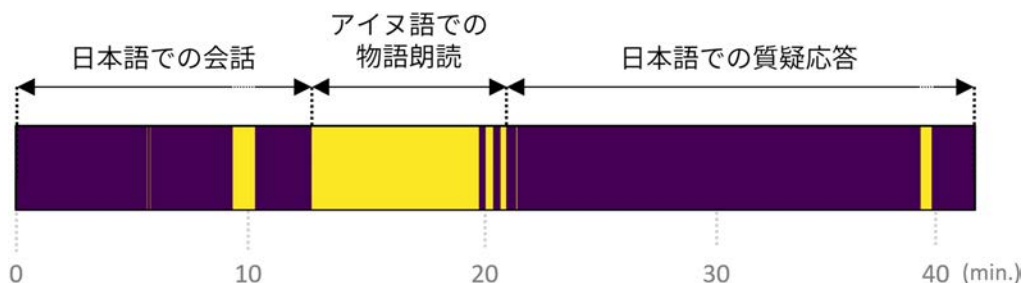


図3 日本語とアイヌ語が混合した収録音声の例。薄色の箇所がアイヌ語の発話であることを表す。

素列から単語列を復元する (P2W)。これにより、語彙的知識を利用できることが期待される。最後に、単語列の中の各言語における単語数によって多数決をとって、その発話の属する言語とする。A2P モデルはアイヌ語音声コーパスのみで学習し、P2W モデルは両単言語コーパスを用いて学習する。

6.2 実験条件

本稿では、単言語コーパスとして、アイヌ語音声コーパスから話者 KM の発話を除いたもの (20 時間) と『日本語話し言葉コーパス (CSJ)』の模擬講演 (260 時間) を使用する。参考のために言語混合データには白老から提供された話者 KM の民話を含む日本語アイヌ語混合音声 54 セッション (30 時間) を使用する。このうち 5 セッション (4 時間) が評価セット、1 セッション (0.7 時間) を開発セットとし、残りを参考実験における学習セットとする。

提案法における P2W モデルは、4 層の CNN と 5 層の双方向 LSTM からなるネットワークを CTC 損失関数を用いて学習したものである。P2W モデルは注意機構を含むエンコーダ・デコーダモデルであり、エンコーダは 4 層の CNN と 5 層の双方向 LSTM から、デコーダは 1 層の単方向 LSTM からなる。日本語の単語列は MeCab を用いて得た。音素数は 22、単語数は 50756 となった。

比較手法として、6.1 節で紹介した Cai らによる言語識別モデル (Cai モデル) についても実験を行う。Cai モデルは ResNET と双方向 LSTM によって音響特徴量を識別に適した潜在表現列に変換し、これを自己注意プーリングによって固定長の 1 つのベクトルに埋め込んだ表現を用いて言語を識別するものである。本実験では、モデルの次元数 (ResNet の出力) は 128、双方向 LSTM の層数は 3 とした。

先述の通り言語識別は発話単位で行うが、評価はフレーム単位での Macro-F1 値 (アイヌ語と日本語それぞれに対する F 値の単純平均) とした。フレーム単位の正解言語ラベルは GMM-HMM モデルを用いて、強制アライメントを行って得た。

6.3 実験結果

表 8 に各手法における言語識別結果のフレーム単位での Macro-F1 値とアイヌ語に関する適合率 (precision)、再現率 (recall)、F 値 (%) を示す。また、図 4 は推定結果を可視化したものであり、薄色の箇所がアイヌ語、濃色の箇所が日本語を表す。単言語コーパスで学習した Cai モデルでは、図 4 の下から 2 番目の結果のように、ほとんど全ての発話をアイヌ語と識別している。これは、モデルが音響条件や話者性の影響を大きく受けているためである。一方で、言語混合音声を学習データに含めると、最も下の図のように有意な識別結果が得られる。ただし、このモデル学習は特定話者 (KM) で行われており、他の話者に適用することは困難であると考えられる。以上から、Cai モデルは同一話者の言語識別を行う能力は十分に保持しているもの

	Cai モデル	Cai モデル +混合音声	提案手法	提案手法 +スムージング
適合率	32.6	87.6	91.3	74.1
再現率	74.55	46.5	26.5	59.1
F 値	44.5	60.7	41.1	65.8
Macro-F1	44.5	81.4	63.1	76.3

表 8 言語識別の結果. 値はフレーム単位での Macro-F1 値とアイヌ語に関する適合率・再現率・F 値 (%)

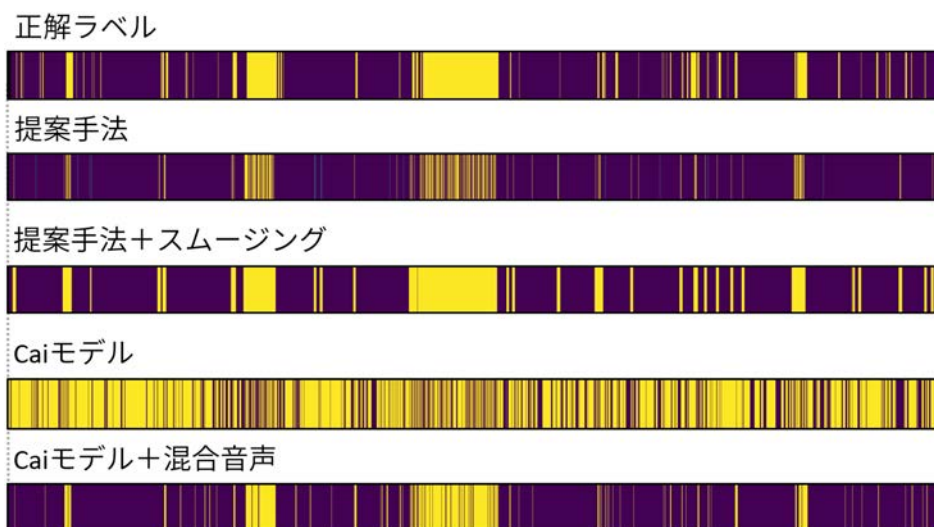


図 4 言語識別結果を可視化したもの. 薄色がアイヌ語を表し, 評価セットに含まれる 5セッションを結合して表示している.

の, その性能はコーパスの音響特性や話者に大きく依存し, 単言語コーパスのみでの学習は不可能であることがわかる.

提案手法による言語識別結果は, 学習時に言語混合音声を使用していないにもかかわらず高い性能を示しており, コーパス間の音響特性に頑健であることがわかる. 表 8 を見ると, アイヌ語に関する適合率が高いが再現率が低く, 図 4 のようにアイヌ語と識別される箇所が疎らになっている. 図 3 に例示したように, 本言語混合音声においてアイヌ語の発話は連続する傾向にあり, かつ提案手法の適合率が高いため, 結果をスムージングすることによって表 8 に示すようにさらに Macro-F1 値を高めることができる. 具体的には, 各フレームについて前後 4000 フレームの結果を元にスムージングを行った.

7 結論

本稿ではアイヌ語アーカイブを対象とした音声認識の取り組みについて述べた。まず、アイヌ語音声コーパスを整備し、注意機構モデルを用いて音声認識システムを実現した。その中で、音節単位での認識が最良の結果となることがわかり、話者クローズド条件と話者オープン条件において、それぞれ 6.3%と 13.8%の PER, 21.7%と 38.6%の WER を達成した。

次に、話者オープン条件において音声認識性能が大幅に低下する話者スパース問題に対して、CycleGAN に基づく声質変換を用いた教師なし話者適応を提案した。提案手法によって、話者オープン条件における PER は相対 35%から 60%改善し、従来手法の自己適応学習やマルチリンガル学習による改善幅を大きく上回った。

最後に、日本語アイヌ語混合音声における言語識別を検討した。学習データに単言語コーパスのみを用いるという制約の下で、音素認識と単語認識を経由する提案手法により一定の識別性能が得られることを示した。

謝 辞

本研究において大変有用なアドバイスをくださった札幌学院大学の奥田統己教授に深く感謝の意を表す。また、アイヌ語の音声データを提供してくださった国立アイヌ民族博物館（旧白老町立アイヌ民族博物館）と、平取町立二風谷アイヌ文化博物館に深い感謝の意を表す。

参考文献

- Alexandre, M. C. N. (2010). *Atlas of the World's Languages in Danger*, 3rd edn. Paris, UNESCO Publishing.
- Anastasopoulos, A. and Chiang, D. (2018). “Tied Multitask Learning for Neural Speech Translation.” In *Proceedings of NAACL*, Vol. 1, pp. 82–91.
- Anmol, G., James, Q., Chung-Cheng, C., Niki, P., Yu, Z., Jiahui, Y., Wei, H., Shibo, W., Zhengdong, Z., Yonghui, W., and Ruoming, P. (2020). “Conformer: Convolution-augmented Transformer for Speech Recognition.” In *Proceedings of INTERSPEECH*, pp. 5036–5040.
- Audhkhasi, K., Kingsbury, B., Ramabhadran, B., Saon, G., and Picheny, M. (2018). “Building Competitive Direct Acoustics-to-Word Models for English Conversational Speech Recognition.” In *Proceedings of ICASSP*, pp. 4759–4763.
- Boito, M. Z., Anastasopoulos, A., Lekakou, M., Villavicencio, A., and Besacier, L. (2018). “A small Griko-Italian speech translation corpus.” In *Proceedings of SLTU*, pp. 36–41.

- ブガエワ・アンナ (2014). 北海道南部のアイヌ語. 早稲田大学高等研究紀要, **6**, pp. 33–76. 児島康宏, 長崎郁 訳. [A. Bugaeva (2014). Southern Hokkaido Ainu. Waseda Daigaku Koto Kenkyu Kiyo, 6, pp. 33–76. Y. Kojima and I. Nagasaki (Trans.)]
- Cai, W., Chen, J., and Li, M. (2018). “Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System.” In *Proceedings of Odyssey*, pp. 74–81.
- Cai, W., Danwei, C., Huang, S., and Li, M. (2019). “Utterance-level End-to-end Language Identification Using Attention-based CNN-BLSTM.” In *Proceedings of ICASSP*, pp. 5991–5995.
- Çetin, O., Plauché, M., and Nallasamy, U. (2008). “Unsupervised Adaptive Speech Technology for Limited Resource Languages: A Case Study for Tamil.” In *Proceedings of SLTU*, pp. 98–101.
- Cho, J., Baskar, M. K., Li, R., Wiesner, M., Mallidi, S. H., Yalta, N., Karafiát, M., Watanabe, S., and Hori, T. (2018). “Multilingual Sequence-to-Sequence Speech Recognition: Architecture, Transfer Learning, and Language Modeling.” In *Proceedings of SLT*, pp. 521–527.
- Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). “Attention-Based Models for Speech Recognition.” In *Proceedings of NIPS*, pp. 577–585.
- Cui, J., Kingsbury, B., Ramabhadran, B., Sethy, A., Audhkhasi, K., Cui, X., Kislal, E., Mangu, L., Markus Nussbaum-Thom, Picheny, M., Tüske, Z., Golik, P., Schluter, R., Ney, H., Gales, M. J. F., Knill, K. M., Ragni, A., Wang, H., and Woodland, P. (2015). “Multilingual Representations for Low Resource Speech Recognition and Keyword Search.” In *Proceedings of ASRU*, pp. 259–266.
- Grézl, F. and Karafiát, M. (2014). “Combination of Multilingual and Semi-supervised Training for Under-resourced Languages.” In *Proceedings of INTERSPEECH*, pp. 820–824.
- Hamlaoui, F., Makasso, E.-M., Müller, M., Engelmann, J., Adda, G., Waibel, A., and Stüker, S. (2018). “BULBasaa: A Bilingual Basaa-French Speech Corpus for the Evaluation of Language Documentation Tools.” In *Proceedings of LREC*, Miyazaki, Japan.
- Hardacre, H. (1997). *New Directions in the Study of Meiji Japan*. Brill, Leiden New York.
- Hermansky, H., Ellis, D. P. W., and Sharma, S. (2000). “Tandem Connectionist Feature Extraction for Conventional HMM Systems.” In *Proceedings of ICASSP*, Vol. 3, pp. 1635–1638.
- Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups.” *IEEE Signal Processing Magazine*, **29** (6), pp. 82–97.
- 北海道ウタリ協会 (1994). アコロイタク. クルーズ. [Hokkaido Utari Kyokai (1994). Akor Itak. CREWS.]
- Hori, T., Watanabe, S., Zhang, Y., and Chan, W. (2017). “Advances in Joint CTC-attention

- based End-to-end Speech Recognition with a Deep CNN Encoder and RNN-LM.” In *Proceedings of INTERSPEECH*, pp. 949–953.
- House, A. S. and Neuburg, E. P. (1977). “Toward Automatic Identification of the Language of an Utterance. I. Preliminary Methodological Considerations.” *The Journal of the Acoustical Society of America*, **62** (3), pp. 708–713.
- Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K., and Itahashi, S. (1999). “JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research.” *Journal of the Acoustical Society of Japan (E)*, **20** (3), pp. 199–206.
- Kaneko, T. and Kameoka, H. (2018). “CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks.” In *Proceedings of EUSIPCO*, pp. 2100–2104.
- Kaneko, T., Kameoka, H., Tanaka, K., and Hojo, N. (2019). “CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion.” In *Proceedings of ICASSP*, pp. 6820–6824.
- Kirchhoff, K. and Parandekar, S. (2001). “Multi-Stream Statistical N-Gram Modeling With Application To Automatic Language Identification.” In *Proceedings of EUROSPEECH*, pp. 803–806.
- Kudo, T. and Richardson, J. (2018). “SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing.” In *Proceedings of ACL*, pp. 66–71.
- Leonard, R. and Doddington, G. (1974). “Automatic Language Identification.” *Technical Report RADC-TR-74-200 (Air Force Rome Air Development Center, Technical Report)*.
- Lopez-Moreno, I., Gonzalez-Dominguez, J., Martinez, D., Plhot, O., Gonzalez-Rodriguez, J., and Moreno, P. J. (2016). “On the Use of Deep Feedforward Neural Networks for Automatic Language Identification.” *Computer Speech & Language*, **40**, pp. 46–59.
- Lüscher, C., Beck, E., Irie, K., Kitza, M., Michel, W., Zeyer, A., Schlüter, R., and Ney, H. (2019). “RWTH ASR Systems for LibriSpeech: Hybrid vs Attention - w/o Data Augmentation.” In *Proceedings of INTERSPEECH*, pp. 231–235.
- Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). “Spontaneous Speech Corpus of Japanese.” In *Proceedings of LREC*, Athens, Greece.
- Meng, Z., Gaur, Y., Li, J., and Gong, Y. (2019). “Speaker Adaptation for Attention-Based End-to-End Speech Recognition.” In *Proceedings of INTERSPEECH*, pp. 241–245.
- Nowakowski, K., Ptaszynski, M., Masui, F., and Momouci, Y. (2019). “Applying Support Vector Machines to POS tagging of the Ainu Language.” In *Proceedings of Computational Methods for Endangered Languages*, **2** (4), pp. 17–23.

- Ochiai, T., Watanabe, S., Katagiri, S., Hori, T., and Hershey, J. (2018). “Speaker Adaptation for Multichannel End-to-End Speech Recognition.” In *Proceedings of ICASSP*, pp. 6707–6711.
- Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). “Librispeech: An ASR Corpus Based on Public Domain Audio Books.” In *Proceedings of ICASSP*, pp. 5206–5210.
- Paul, D. B. and Baker, J. M. (1992). “The Design for the Wall Street Journal-based CSR Corpus.” In *Speech and Natural Language: Proceedings of a Workshop*, pp. 357–362.
- Radford, A., Metz, L., and Chintala, S. (2015). “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.” *arxiv:1511.06434*.
- Rialland, A., Adda-Decker, M., Kouarata, G.-N., Adda, G., Besacier, L., Lamel, L., Gauthier, E., Godard, P., and Cooper-Leavitt, J. (2018). “Parallel Corpora in Mboshi (Bantu C25, Congo-Brazzaville).” In *Proceedings of LREC*, pp. 4272–4276, Miyazaki, Japan.
- Schuster, M. and Paliwal, K. K. (1997). “Bidirectional Recurrent Neural Networks.” *IEEE Transactions on Signal Processing*, **45** (11), pp. 2673–2681.
- Sennrich, R., Haddow, B., and Birch, A. (2016). “Neural Machine Translation of Rare Words with Subword Units.” In *Proceedings of ACL*, pp. 1715–1725. Association for Computational Linguistics.
- Senuma, H. and Aizawa, A. (2018). “Universal Dependencies for Ainu.” In *Proceedings of LREC*, pp. 2354–2358.
- 田村すず子 (1988). アイヌ語. 亀井孝, 河野六郎, 千野栄一 (編), 言語学大辞典, **1**, pp. 6–94. 三省堂. [S. Tamura (1988). Ainu-go. T. Kamei et al. (Eds.), The Sanseido Encyclopedia of Linguistics, **1**, pp. 6–94. Sanseido.]
- 田村すず子 (1996). アイヌ語沙流方言辞典. 草風館. [S. Tamura (1996). The Ainu-Japanese Dictionary, Saru Dialect. Sofukan.]
- 田村すず子 (2013). アイヌ語の世界. 吉川弘文館. [S. Tamura (2013). Ainu Go No Sekai. Yoshikawa Kobunkan.]
- Thomas, S., Seltzer, M. L., Church, K., and Hermansky, H. (2013). “Deep Neural Network Features and Semi-supervised Training for Low Resource Speech Recognition.” In *Proceedings of ICASSP*, pp. 6704–6708.
- Tong, R., Ma, B., Li, H., and Chng, E. S. (2009). “A Target-Oriented Phonotactic Front-End for Spoken Language Recognition.” *IEEE Transactions on Audio, Speech, and Language Processing*, **17** (7), pp. 1335–1347.
- Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P. J., Weinstein, E., and Rao, K. (2018). “Multilingual Speech Recognition with a Single End-to-End Model.” In *Proceedings of ICASSP*, pp. 4904–4908.

- Tóth, L., Frankel, J., Gosztolya, G., and King, S. (2008). “Cross-lingual Portability of MLP-based Tandem Features - a Case Study for English and Hungarian.” In *Proceedings of INTERSPEECH*, pp. 2695–2698.
- Wöllmer, M., Schuller, B., and Rigoll, G. (2011). “Feature Frame Stacking in RNN-Based Tandem ASR Systems - Learned vs. Predefined Context.” In *Proceedings of INTERSPEECH*, pp. 1233–1236.
- Yukie, C. (1923). *Ainu Shin'yoshu*. Kyodo Kenkyusha.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks.” In *Proceedings of ICCV*, pp. 2242–2251.

略歴

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。同年，日本電信電話株式会社入社。ISCA，日本音響学会各会員。

三村 正人：2000 年京都大学大学院情報学研究科修士課程修了。京都大学情報学研究科研究員等を経て，現在，同博士後期課程在学中。IEEE，日本音響学会各会員。

河原 達也：1987 年京都大学工学部情報工学科卒業。1989 年同大学院修士課程修了。京都大学工学部助手，同助教授，学術情報メディアセンター教授を経て，現在，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。博士（工学）。IEEE Fellow。ISCA，APSIPA 各理事。情報処理学会，日本音響学会，電子情報通信学会，人工知能学会，言語処理学会 各会員。日本学術会議連携会員。

(2021 年 3 月 30 日 受付)

(2021 年 5 月 20 日 再受付)

(2021 年 6 月 24 日 採録)