

アイヌ語に対する音声合成モデルの学習と評価

花田 佳文^{1,a)} 中田 琉稀² 林 克彦¹

概要：日本の先住民族言語であるアイヌ語は消滅の危機にあり、その音声合成の研究はほとんど行われていない。本研究では、北海道アイヌ語の音声データ計約 12.85 時間を整備し、ラテン文字表記を文字単位で入力として複数の音声合成モデル (Matcha-TTS, VITS, F5-TTS) を学習した。Matcha-TTS と VITS については学習済みモデルからのファインチューニングとスクラッチからの学習を 5 分割交差検証で比較し、自動評価と自然性の主観評価 (MOS) を行った結果、ファインチューニングした Matcha-TTS が最良であり、自然音声に最も近い品質が得られた。一方、ファインチューニングの効果はモデルにより異なり、自動評価と主観評価によるモデルの順位は概ね一致した。本研究は、管見の限りアイヌ語の音声合成に関する初めての系統的な報告である。

キーワード：アイヌ語, 音声合成

Training and Evaluation of Speech Synthesis Models for the Ainu Language

Abstract: The Ainu language, an indigenous language of Japan, is critically endangered, and its speech synthesis has rarely been studied. We constructed a corpus of approximately 12.85 hours of Hokkaido Ainu speech and trained three speech synthesis models (Matcha-TTS, VITS, and F5-TTS) with character-level Latin-script input. For Matcha-TTS and VITS, we compared fine-tuning from pretrained models with training from scratch using five-fold cross-validation, with objective metrics and a subjective naturalness evaluation (MOS). The fine-tuned Matcha-TTS performed best and came closest in quality to natural speech, while the effect of fine-tuning varied across models, and the model rankings from the two evaluations were broadly consistent. To our knowledge, this is the first systematic report on speech synthesis for the Ainu language.

Keywords: Ainu language, Text-to-Speech

1. はじめに

世界には約 7,000 の言語があるとされているが、その相当数が消滅の危機にある。Krauss [1] は、次の 100 年で世界の言語の 90% が消滅する可能性を指摘している。一方で、世界各地でこうした消滅危機言語の話者や継承者自身による言語の記録、教育などの言語復興活動が続けられており、こうした活動を支える手段の一つとして音声技術の活用が試みられている。中でも、テキストから音声を生成する音声合成は、音声を伴わないテキストの発音の提示や

聴取素材の作成を可能にし、言語の学習・継承を支えうる。近年の音声合成は、大規模な音声コーパスを用いた深層学習によって高品質な合成音声を実現している。しかし、利用できる音声データが限られる言語では、こうした技術の適用が遅れている。

音声資源の乏しい言語では数時間から数十時間規模のデータで音声合成を試みた研究があるものの [2], [3], 十分な品質に達していない例が見られる。こうした低資源条件では、どのモデルをどのように学習すれば品質を確保できるかが定まっておらず、大規模事前学習モデルの言語適応 [4] やスクラッチからの学習など複数の選択肢がある。しかし、どのモデル・学習法が有効かはデータ量や言語、モデルの構造に依存するため、これらの知見が音声資源の乏しい言語にそのまま当てはまるとは限らず、対象とする

¹ 東京大学 大学院総合文化研究科 言語情報科学専攻
Department of Language and Information Sciences, Graduate School of Arts and Sciences, The University of Tokyo

² 東京大学 教養学部
College of Arts and Sciences, The University of Tokyo

^{a)} yhanada@g.ecc.u-tokyo.ac.jp

言語のデータで検証する必要がある。

こうした音声資源の乏しい言語の一つに、本研究が対象とするアイヌ語がある。アイヌ語は、北海道で話されている日本の先住民族言語であり、かつては樺太、千島などでも話されていた。日琉語族とは異なる独立した言語である。言語系統的には孤立した言語とされ、複数の接辞を語根に接続し、複雑な内容を一語で表現する複統合性や、動詞語根に名詞を接続し、生産的に複合動詞を生成することができる抱合性を持つ言語である。また、音声の面でも日本語と異なり、閉音節を多用するなどの特徴を有する [5]。

アイヌ語は北海道アイヌ語、樺太アイヌ語、千島アイヌ語に大別されるが、樺太アイヌ語と千島アイヌ語は 20 世紀中にそれぞれ話者が途絶えている [6]。北海道アイヌ語も 20 世紀以降、母語話者が著しく減少したものの、完全に話者が途絶えたわけではなく、北海道を中心に各地で復興の取り組みが続けられており、話者や学習者は増加傾向にある [5], [6]。なお、本稿では以降、特に断らない限り北海道アイヌ語をアイヌ語と呼ぶ。

一方で、アイヌ語学習者が任意のテキストの発音を音声で確認できる手段は、ほとんど存在しない。研究面では、アイヌ語は音声認識の研究が先行しており [7]、音声合成についても同グループが Tacotron 2 や VITS による試行に言及しているが [8]、手法・データ・評価を明示した報告は管見の限り存在しない。さらに、他の言語的少数派の事例と同様に [3]、アイヌ語も母語話者の高齢化と著しい減少により、合成音声の品質を評価できる話者や熟達した学習者を十分に確保することが難しく、評価方法自体が課題となる。このように利用・研究・評価のいずれの面でも課題があるものの、音声合成はアイヌ語の利用・学習の機会の拡大と話者や学習者の言語権に資するものであり重要である。

そこで本研究では、アイヌ語のテキスト音声合成の初期的な検討として、一般に入手できるアイヌ語の音声データ (計約 12.85 時間) を整備し、構造の異なる複数の音声合成モデル (Matcha-TTS, VITS, F5-TTS) を同一条件で学習・比較する。また、既存の学習済みモデルからのファインチューニングとスクラッチからの学習を比較し、限られたデータ量のもとでの両者の効果を、5 分割交差検証と自動評価および主観評価により検討する。本研究は、アイヌ語の音声合成を複数モデルの比較と自動・主観評価により検討した、管見の限り初めての報告である。

なお本研究は、音声教材への利用など、学習用途に向けた発音の再現を主眼とする。そのため、節回しを伴わない uepeker (散文説話) と日常会話の音声のみを学習に用い、kamuyyukar (神謡) など節回しを伴う韻文の音声は対象としない。

以上の比較を通じて、本研究で明らかにするのは次の 3

表 1 アイヌ語の主要子音素 (/ / 内は音素, [] 内は IPA)

	両唇	歯茎	硬口蓋	軟口蓋	声門
閉鎖音	/p/ [p]	/t/ [t]	/c/ [t͡ʃ]	/k/ [k]	
摩擦音		/s/ [s]			/h/ [h]
はじき音		/r/ [r]			
鼻音	/m/ [m]	/n/ [n]			
接近音	/w/ [w]		/y/ [j]		

点である。(1) 限られたデータのもとでアイヌ語音声合成がどの品質まで到達できるか、(2) 事前学習済みモデルからのファインチューニングの効果はモデルによらず得られるのか、(3) 評価者の確保が困難な言語において自動評価指標は主観評価の代わりになりうるか。

2. アイヌ語の音声的特徴

本章ではアイヌ語の音声的特徴について述べる。前章の通り、アイヌ語は日本語とは異なる独立した言語である。標準的な表記が存在せず、時代や研究者によってカタカナ、キリル文字、ラテン文字で表記されてきたが、本稿では『アコロ イタク』 [9] に倣ったラテン文字での方式で表記する。

2.1 音素

アイヌ語は北海道内において、方言間の語彙の差は見られるものの、音素の面では大きな差はない [5]。また、使われる音素についても比較的少なく、母音は日本語と同様に /a, e, i, o, u/ の 5 種類、子音は表 1 のとおり、/p, t, k, c, s, h, m, n, r, y, w/ の 11 種類である。

このうち、アイヌ語の /u/ は /o/ の音にも近く、古い文献では /aynu/ 「人間」を “Aino” と記載している例も見られる*1。一般的には、日本語のウが非円唇の [u] であるのに対し、アイヌ語の /u/ は口を突き出し、舌の奥を持ち上げる円唇後舌狭母音 [u] とされる [10]。なお、樺太アイヌ語には短母音と長母音の区別があるが、北海道アイヌ語においてはこの区別はない [5]。

アイヌ語の子音について、上記 11 個の子音に加え、研究者によってはこれに声門閉鎖音を加える場合もあるが、本研究では中川 [5] に従い、音素とはみなさない。アイヌ語は閉鎖音において有声・無声の対立がなく、/p/ と /b/, /t/ と /d/ などは弁別性を持たない。また子音のうち、/c/ と /h/ は音節末には立たない [5]。

2.2 音節

アイヌ語の音節構造は母音 (V) と子音 (C) から成り、表 2 の通り、(C)V(C) の組み合わせによる計 4 種類である [5]。つまり、音節中には母音は必ず含まれるが、子音を含まない音節もある。CCVC のような子音連続を持つ音節は存

*1 Pfizmaier, A. (1854) *Vocabularium der Aino-Sprache* など

在しない。また (C)V(C) のような母音連続は (C)V 部と V(C) 部が結合した結果として現れるため、/V:/のように長母音として発音するのではなく/VV/をそれぞれ別の音として発音する。例えば、ooho「深い」の音節は o-o-ho であり、/o:ho/とは発音しない。

表 2 アイヌ語の音節構造と語例

音節構造	例
V	a「座る」, e「～を食べる」
CV	ru「道」, ku「弓」
VC	at「紐」, uk「～を取る」
CVC	sap「下りる」, sat「乾いている」

2.3 異音と音素交替

各音素は生起環境に応じて異音をもち、その条件づけの方向は音素によって異なる。まず母音については、狭母音/i, u/(まれに/e/)が無声子音に挟まれると無声化することがあり、たとえば、/cise/「家」は[t̚ɕise], /seta/「犬」は[setḁ]と発音される。ただし日本語の無声化とは分布が異なり、si・ciの母音が最も無声化しやすい一方、tu・pi・puの母音は無声化しない[10]。

次に子音について、表3の通り子音音素の主な異音とその生起条件をまとめる。このうち、閉鎖音/p, t, k/は音節末で開放を伴わない閉鎖音[p̚, t̚, k̚]となる。/r/は音節頭では基本的に[r]で実現されるが、音節末では先行する母音と同質の超短母音を伴う[r̥]として実現される[5]。例えば、/pirka/「良い」は[pir̥ka], /arpa/「行く」は[ar̥apa]のような発音となる*2。

また、語や形態素が連続する境界では、表4に示すような音素交替が生じる[5]。

2.4 アクセント

アイヌ語は、特に北海道南西部においては、低から高への上昇が弁別性を持つ高低アクセント(上がり核アクセント)を持つ。アクセントが付く箇所は、その語の第一音節が開音節ならば第二音節から高くなり、第一音節が閉音節ならば第一音節が高くなるのが基本規則である[6]。なお、静内方言や北海道東部のアイヌ語においてはアクセント対立のない変種も存在する[5]。例えば、第一音節が開音節であるkamuy「神」は第二音節にアクセントが付くため、/kamúy/と発音し、第一音節が閉音節であるsinrit「祖先」の場合は、第一音節にアクセントが付くため、/sínrit/と発音する。例外として、húci「おばあさん」のように、基本規則に合わない語は記載者によってはテキストにアクセント記号を残すことがある[5]。

以上のように、アイヌ語のラテン文字表記は音素との対応が概ね一対一である。よって、文字をそのまま入力とし

*2 アイヌ語カナ表記ではこの/r/の異音を区別するために/ピリカ/, /アラバ/のように小文字のラリレロを用いる。

表 3 子音音素の主な異音と生起条件

音素	IPA	生起条件	例
/h/	[h]	基本	/horkew/「オオカミ」 [hor̥okew]
	[ç]	/i/の前	/ayupihi/「我が兄」 [ajupici]
	[ɸ]	/u/の前	/husko/「昔」 [ɸusko]
/s/	[s]	基本	/sinrit/「祖先」 [sínrit]
	[ç]	/i/の後	/cis/「泣く」 [t̚ɕiç]
	[s̚]	/a, e, o, u/の後	/as/「立つ」 [as̚]
/n/	[n]	基本	/nankor/「だろう」 [nan̥kor̥]
	[m]	/p, m/の前	/anpe/「～であるもの」 [ampe]
	[ŋ]	/k/の前	/honkor/「妊娠する」 [hoŋkor̥]
/r/	[r]	音節頭	/rap/「羽」 [rap̚]
	[r̥]	音節末	/pirka/「良い」 [pir̥ka]
/-p/	[p̚]	音節末	/sap/「下りる」 [sap̚]
/-t/	[t̚]	音節末	/sat/「乾いている」 [sat̚]
/-k/	[k̚]	音節末	/sak/「～を欠く」 [sak̚]

表 4 形態素境界における主要な音素交替の例

交替	例	実現形
-t + i- → -ci-	mat「女」 + ikor 「宝物」	/macikor/「女の宝物」
-n + s- → -ys-	pon「小さい」 + seta「犬」	/poyseta/「子犬」
-n + y- → -yy-	pon「小さい」 + yuk「鹿」	/poyyuk/「子鹿」
-r + t- → -tt-	kor「持つ」 + tuki 「杯」	/kottuki/「彼の杯」
-r + c- → -tc-	kor「持つ」 + cise 「家」	/kotcise/「彼の家」
-r + n- → -nn-	kor「持つ」 + nankor「だろう」	/konnankor/「彼が持つだろう」
-r + r- → -nr-	kor「持つ」 + rusuy「したい」	/konrusuy/「彼がほしい」

ても、音素入力に近い情報がモデルに与えられると期待できる。一方で、母音の無声化、音節末の/r/の異音や[p̚, t̚, k̚]といった無解放閉鎖音、形態素境界の音素交替は表記に現れないため、モデルはこれらの音声実現を音声のみから学習する必要がある。この点は5章の誤り分析で検証する。

3. テキスト音声合成のアーキテクチャ

3.1 テキスト音声合成の処理フロー

テキスト音声合成(Text-to-Speech; TTS)は、テキストを入力として音声波形を生成するタスクである。確率的な枠組みでは、テキストと音声波形の対から音響モデルを学習する学習部と、学習済みモデルを用いて任意のテキストから音声を生成する合成部から成る(図1)[11]。学習部では既知のテキストWと音声波形Xから音響モデル入を学

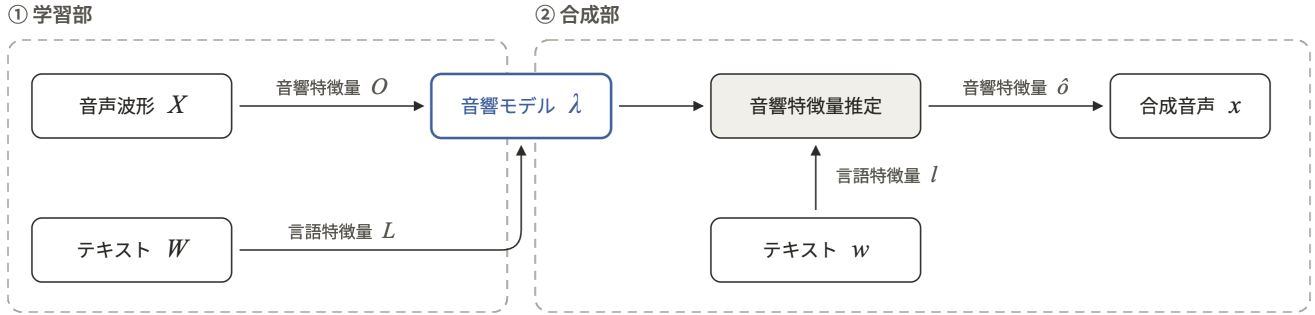


図 1 テキスト音声合成の処理の流れ (文献 [11] を参考に作成)

習し、合成部では与えられたテキスト w に対応する音声波形 x を生成する。この過程は、言語特徴量を抽出する前処理、音響モデルの学習および音響特徴量の予測、波形を生成するボコーダの各処理に分解される [11]。ただし近年のニューラル TTS では、これらの処理の一部を統合して学習することが多く [11]、本稿で扱う 3 モデルも統合の仕方が互いに異なる (3.5 節)。以下、各処理を順に述べる。

3.2 前処理

前処理は、テキストから音響モデルの入力となる言語特徴量を抽出する処理である。学習時は音声に対応するテキスト W から言語特徴量 L を、合成時は入力テキスト w から言語特徴量 l を得る [11]:

$$\hat{L} = \arg \max_L P(L | W), \quad \hat{l} = \arg \max_l P(l | w). \quad (1)$$

具体的には、入力文字列の正規化、単語の同定 (トークン化)、音素への変換 (Grapheme-to-Phoneme; G2P) を経て、得られたシンボル列を言語特徴量として用いる。日本語のように文字列から読みが一意に定まらない言語では形態素解析による単語の同定が必要となるが、アイヌ語はわかち書きされているため単語分割は自明である。英語の CMUdict や日本語の UniDic のような発音情報を備えた辞書が整備された言語と異なり、アイヌ語にはこうした発音辞書が現状存在しないため、変換はもっぱら規則ベースとなる。アクセントや韻律境界 (ポーズ) のように規則化できる情報の付与もこの段階で行われる。なお本稿では、アイヌ語のラテン文字表記をそのまま文字単位で入力とするため、G2P は行わない。

3.3 音響モデル

音響モデル λ は、言語特徴量と音響特徴量 (メルスペクトログラム等) の対応を表す確率モデルである。学習部では、音声波形 X から抽出した音響特徴量 O ($\hat{O} = \arg \max_O p(X | O)$, 音響特徴量抽出) と言語特徴量 L を用いて、音響モデルを学習する [11]:

$$\hat{\lambda} = \arg \max_{\lambda} p(\hat{O} | \hat{L}, \lambda). \quad (2)$$

合成部では、学習済みの音響モデル $\hat{\lambda}$ のもとで、言語特徴量 l から音響特徴量を予測する (図 1 の音響特徴量推定):

$$\hat{o} = \arg \max_o p(o | \hat{l}, \hat{\lambda}). \quad (3)$$

ニューラル音声合成では、この予測はエンコーダ、アライメント、デコーダから成るニューラルネットワークで実装されることが多い。

エンコーダは、入力のシンボル列を実数値埋め込みベクトル列に変換し、文脈を反映した表現を得る [11]。アライメントは、シンボル列と音響特徴量の時間方向の対応づけである。継続長予測器が各シンボルの継続長 (フレーム数) を予測し、シンボルの表現をその回数だけ複製することで、音響特徴量と同じ時間解像度に展開する [12]。デコーダは、展開された表現に話者の埋め込みを条件として与え、音響特徴量を出力する。メルスペクトログラムは、短い時間区間ごとの周波数成分の強さを、聴覚特性に沿ったメル尺度の帯域へまとめた音響特徴量である。話者の情報は話者ごとの埋め込みベクトルとして学習され、デコード時にこれを参照することで目的の話者の声質が反映される。

3.4 ボコーダ

ボコーダは、予測された音響特徴量 $\hat{o}$ から音声波形 x を生成する ($\hat{x} \sim p(x | \hat{o})$) [11]。メルスペクトログラムでは位相 (各周波数成分のずれ) の情報と周波数方向の細かい構造が失われているため、ボコーダはこれらを補いながら波形を復元する。

3.5 比較するモデル

Matcha-TTS [12] は、条件付き Flow Matching により音響特徴量であるメルスペクトログラムを生成する非自己回帰モデルであり、波形の生成は事前学習済みの外部ボコーダ (HiFi-GAN) に分離した構成である。VITS [13] は、変分自己符号化器と正規化フロー、敵対的学習を組み

合わせ、テキストから波形までを単一のモデルで直接生成する end-to-end モデルであり、音響特徴量の予測と波形生成を統合した構成である。実装には Piper^{*3}を用いた。F5-TTS [14] は、多言語音声コーパス Emilia のうち英語・中国語の約 9.5 万時間で事前学習された非自己回帰モデルであり、Diffusion Transformer を用いた Flow Matching に基づき、参照音声を与えるゼロショット合成を実現している。継続長予測器を持たず、発話全体の長さのみを指定して生成する点で、3.3 節の一般的な構成とは異なる。本稿で取り上げる 3 つのモデルは、外部ボコーダを用いる軽量の非自己回帰音響モデル、敵対的学習による end-to-end モデル、大規模事前学習を前提とする基盤モデル、という異なる構造であり、いずれも音声資源の乏しい言語の音声合成研究において言及されたことのあるモデルである。構造の異なるこれらを同一条件で比較することで、アイヌ語の音声合成におけるモデルの違いやファインチューニングの影響を検証できると考える。

4. 実験

4.1 データ整備

4.1.1 コーパス

本研究では、アイヌ語沙流方言、千歳方言、静内方言を収録した 3 つのコーパス (表 5) を用いた。なお、本研究は音声教材など学習用途に向けて、節回しを伴わない uepeker (散文説話) [5] や日常会話における発音の再現を主眼としているため、これらを学習・音声合成の対象としている。一方、伝承者や説話ごとに異なる節回しを伴う kamuyyukar (神謡) や yukar (英雄叙事詩) などの韻文は学習・評価から除外した。

4.1.2 音声の前処理

利用する音声データはコーデック・サンプリングレート・チャンネル数が一様でなかったため、学習の入力を揃えるべく、すべての音声を 24kHz、モノラル、16bit の PCM WAV へ変換して統一した。統一後の音声を対応するテキストと発話単位で対応づけ、計 12.85 時間、11,560 件の発話データを学習と評価に用いた。発話単位への分割は、文単位で音声、テキストが切られている『トピック別アイヌ語会話辞典』と『アイヌ語音声コーパス』では文単位の分割を用いた。一方、説話単位でファイルが分かれており、文単位の区切りのない『アイヌ語口承文芸コーパス』では強制アライメントにより分割を行った。

なお話者ごとのデータ量には偏りがあり、利用したデータのうち、KM 氏の発話が全体の約 7 割を占める。

4.1.3 テキストの正規化

音声に対応するテキストは『アコロ イタク』[9] に倣ったラテン文字表記を用い、次の正規化を施した。編集上の

注記や日本語訳、句読点・引用符、人称接辞と語幹間の境界記号である=は除去し、大文字は小文字へ統一した。また、一部の資料に付与されているアクセント記号は除去し、アクセントやポーズなどの韻律に関するラベルも付与していない。モデルへの入力とは正規化後の文字列を用いた。

4.2 実験条件

4.2.1 交差検証

評価には 5 分割交差検証を用いた。全 11,560 発話を 5 つの集合に分け、各集合を順にテストデータとした。すなわち、1 つの集合 (全体の 20%, 2,312 発話) をテストデータ、残りを訓練データ (60%, 6,936 発話) と検証データ (20%, 2,312 発話) とした。分割は発話単位で行ったため、同一話者・同一説話の発話が訓練とテストの双方に含まれる。したがって本評価が測るのは、学習した話者における未見テキストに対する合成品質であり、未知話者への汎化性能は本評価の対象外である。

分割はすべての TTS モデルで共通とし、同一の条件で比較した。各 fold では検証データの MCD が最小となるチェックポイントを選択し、テストデータで評価し、fold 間の平均と標準偏差を報告した。この選択は 3 モデルすべてに同一の基準で適用した。なお選択基準が MCD であるため、選択されたチェックポイントが他の指標に対して最適であるとは限らない。

4.2.2 学習の設定

ファインチューニングでは、Matcha-TTS は LJSpeech (英語、約 24 時間) で学習された公開チェックポイント、VITS は Piper の公開学習済みモデル en-US-lessac-medium (英語)、F5-TTS は Emilia の英語・中国語データ (約 9.5 万時間) で学習された F5TTS.Base を初期値とした。Matcha-TTS と VITS については、同一のアーキテクチャを乱数初期化から学習するスクラッチからの学習も行った (表 6 ではそれぞれ FT, scratch と表記)。なお F5-TTS は大規模事前学習を前提とする基盤モデルであるため、スクラッチからの学習は行わない。

Matcha-TTS と VITS は、話者 ID を条件として与える多話者モデルとして学習し、合成時にはテスト発話の話者 ID を指定した。F5-TTS は話者 ID を用いず、参照音声によって話者を条件付けた。F5-TTS の推論では、参照音声としてテスト発話と同一の話者 ID をもつ訓練データ中の発話を与えた。

エポック数は Matcha-TTS で 200、VITS でファインチューニング 100・スクラッチからの学習 600、F5-TTS で 30 とした。バッチサイズは Matcha-TTS で 32、VITS で 64 (いずれも発話単位)、F5-TTS で 38,400 フレームである。このエポック数は、各モデルの検証性能が収束する点の観測に基づき設定した学習の上限である。検証性能が最良と

^{*3} <https://github.com/rhasspy/piper> (2026 年 7 月 19 日閲覧)

表 5 使用したコーパスと話者 (時間は前処理後に学習・評価に用いた実時間)

コーパス	話者	方言	形式	サンプリング周波数	チャンネル	時間
トピック別アイヌ語会話辞典 [15]	KS	沙流	mp3	44.1 kHz	2	1.85 h
アイヌ語口承文芸コーパス [16]	KK	沙流	mp3	44.1 kHz	2	0.15 h
	OI	千歳	mp3	44.1/48 kHz	2	0.58 h
アイヌ語音声コーパス [17]	KM	沙流	wav (16 bit PCM)	44.1/48 kHz	1	8.59 h
	OS	静内	wav (16 bit PCM)	44.1/48 kHz	1	1.68 h
合計						12.85 h

なるエポックはモデルによって異なったため、モデルおよび学習方法により上限を変えた。学習率などその他のハイパーパラメータは各実装の既定値を用いた。

4.2.3 自動評価

自動評価には、メルケプストラム歪み (MCD; Mel Cepstral Distortion), 基本周波数 (F0) の二乗平均平方根誤差 (RMSE), 有声無声判定の F1 値 (VUV F1; Voiced/Unvoiced F1), 話者類似度 (SECS; Speaker Encoder Cosine Similarity) の 4 指標を用い [2], [18], 各テスト発話の合成音声に対応する自然音声と比較した。MCD は、スペクトル包絡から求めたメルケプストラム係数の距離であり、音色や分節音の実現が自然音声にどれだけ近いかを表す。算出では両音声のフレームを DTW で対応づけた (pymcd を使用)。F0 RMSE は声の高さの再現誤差であり、WORLD^{*4}で推定した両音声の F0 の有声フレーム列を先頭から順に対応づけ、フレーム対ごとの誤差を二乗平均平方根を取った値を Hz 単位で算出した。VUV F1 は、フレームごとの有声無声判定について自然音声を正解とした F1 値であり、母音・鼻音などの有声区間と摩擦音・無音などの無声区間の並びの再現度を表す。SECS は、合成音声の話者性が自然音声の話者にどれだけ近いかを表し、話者認識モデル ECAPA-TDNN^{*5}で抽出した話者埋め込み間のコサイン類似度として算出した。

各モデルの学習は実装標準のサンプリングレート (Matcha-TTS と VITS は 22.05kHz, F5-TTS は 24kHz) に従うため、評価では合成音声をすべて 24kHz に再標準化したうえで指標を算出した。

4.2.4 主観評価

主観評価では、テスト集合から選んだ 30 発話について、自然音声と各構成の合成音声を、どの構成かを伏せてランダムに呈示し、自然性を 5 段階 (1: 非常に悪い~5: 非常に良い) で評定した。30 発話は 5 つの fold から 6 発話ずつ無作為に選び、各発話はその発話をテストデータとする fold のモデルで合成した。呈示にあたっては全条件の音声を 24kHz に統一しピーク音量を正規化して、サンプリング

レートや音量が構成の手掛かりとならないようにした。

評価者は第一著者 1 名である。第一著者は 15 年以上のアイヌ語の学習歴を有し、アイヌ語文の音読・聴解が可能な上級学習者であり、分節音の実現や音素配列の誤りを聴取で判別できる。評定の発話間の平均と標準偏差を、平均オピニオン評点 (MOS; Mean Opinion Score) として報告する。

4.3 実験結果

事前学習済みモデルからファインチューニングした Matcha-TTS が、MCD・SECS・MOS の 3 指標で最良であった (表 6)。F0 RMSE については VITS のファインチューニングが最良であった。最良構成の MCD は 5.937, MOS は 3.28 であり、自然音声 (MOS 4.23) に最も近い品質であった。なお自然音声の MOS も 4.23 と天井には達していない。これは学習・評価に用いた音声が高齢話者からのフィールド録音であり音質が一様でないことによると考えられ、各構成の MOS はこの自然音声を上限とする相対値として解釈する必要がある。また、VUV F1 は Matcha-TTS と VITS の 4 構成でいずれも 0.75 前後と同等であり、有声無声の時間構造の再現にはモデル・学習法による差はほとんど見られなかった。

ファインチューニングの効果はモデルによって異なった。Matcha-TTS ではスクラッチからの学習 (MCD 6.212, MOS 2.60) に対してファインチューニング (MCD 5.937, MOS 3.28) で明確に改善し、話者類似度も 0.650 から 0.688 へ向上した。一方 VITS では、スクラッチからの学習 (MCD 6.206) とファインチューニング (MCD 6.254) が自動指標ではほぼ同等であり、その差は fold 間の標準偏差の範囲に収まった。

F5-TTS はファインチューニングのみを実施したが、自動評価・主観評価のいずれでも 3 モデル中で最下位であった (MCD 9.171, SECS 0.399, MOS 1.13)。他の 2 モデル (MCD は 6 程度以下, SECS は 0.65 前後) と比べて MCD が大きく、話者類似度も低い。VUV F1 も 0.619 と唯一大きく低く、有声無声の時間構造の再現においても他の 2 モデルに明確に劣った。主観評価においても、ほとんどの合成音声が悪質に聞こえているか、合成できているものでもアイヌ

^{*4} <https://github.com/JeremyCCHsu/Python-Wrapper-for-WORLD-Vocoder> (2026 年 7 月 19 日閲覧)

^{*5} <https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb> (2026 年 7 月 19 日閲覧)

表 6 評価結果 (平均 ± 標準偏差)

モデル	学習法	MCD↓	F0 RMSE↓[Hz]	VUV F1↑	SECS↑	MOS↑
自然音声 (GT)	—	—	—	—	—	4.23 ± 1.28
Matcha-TTS	scratch	6.212 ± 0.132	24.97 ± 2.02	0.750 ± 0.027	0.650 ± 0.020	2.60 ± 1.38
	FT	5.937 ± 0.087	24.71 ± 1.98	0.749 ± 0.021	0.688 ± 0.018	3.28 ± 1.28
VITS	scratch	6.206 ± 0.149	23.86 ± 1.47	0.754 ± 0.024	0.639 ± 0.023	2.28 ± 1.13
	FT	6.254 ± 0.185	23.07 ± 1.26	0.761 ± 0.015	0.649 ± 0.021	2.00 ± 0.98
F5-TTS	FT	9.171 ± 0.141	40.63 ± 4.97	0.619 ± 0.030	0.399 ± 0.034	1.13 ± 0.51

語の音声として聴取できないものが多く、MOS は 1.13 と著しく低く、実用性に乏しい。

5. 考察

5.1 共通する誤りの分析

主観評価では、モデルや学習方法によらず共通する誤りが現れた。最も目立ったのは、音節末の /r/ の脱落である。Matcha-TTS と VITS のいずれでも、またスクラッチからの学習とファインチューニングのいずれでも生じた。音節末子音の脱落は特定のモデルや学習方法に固有の問題ではなく、本研究の設定に共通して現れる弱点である。

2 章で述べたとおり、音節末の /r/ は先行母音と同質の超短母音を伴う異音として実現され、その実現の仕方は表記に現れない上に、音響的にも微弱で短い区間である。このような区間の再現は少量データでの学習では特に難しく、実際に最も崩れやすい部分として現れたといえる。

また、/t, k, p/ の無解放閉鎖音についても、安定しなかった。Matcha-TTS では、音節末の /p/ が消える例と、音節末の /t/ が強く出すぎた結果、期待される [t] が [t] として開放を伴って実現される例があった。頻度は /r/ より低かった。また、一部の合成音声では語がまるごと別の語のように聞こえる崩れも観察された。

5.2 ファインチューニングの効果

学習方法による差も現れた。Matcha-TTS はスクラッチからの学習では出力が崩れて聴き取れない発話がいくつかあったが、ファインチューニングではそうした破綻はなかった。ファインチューニングが出力を安定させたといえ、自動評価・主観評価で Matcha-TTS のファインチューニングがスクラッチからの学習を上回った結果とも一致する。一方 VITS は、どちらの学習でも /r/ の脱落を中心とする似た傾向を示し、学習方法による差は自動評価・主観評価とも小さかった。

この差の要因は本研究のデータからは特定できないが、モデルの構造の違いによる可能性が考えられる。Matcha-TTS は波形生成を事前学習済みの外部ボコーダに委ねており、適応の対象が音響モデルに限られるのに対し、VITS は敵対的学習を含む end-to-end モデルの全体を少量データで適応し直す必要がある。

F5-TTS の品質の低さは、英語・中国語による事前学習に含まれないアイヌ語へ適応するには、今回のデータ量と全体のファインチューニングという適応方法では不足していたためと考えられる。このほか、参照音声の与え方や、アイヌ語のラテン文字表記がモデルの語彙にとって新規である点も品質低下に寄与した可能性がある。実際、F5-TTS をルーマニア語へ適応した先行例 [4] はモデル本体を凍結し入力側に軽量なアダプタを追加する方法を採っており、少量データでの全体のファインチューニングが有効に機能しなかった本研究の結果と整合する。

以上の結果は、事前学習が合成品質の改善につながるかどうか、モデルの構造と事前学習の内容の双方に依存することを示している。Matcha-TTS と VITS のファインチューニングは、いずれも英語の単一話者データで学習されたモデルを初期値としている。それにもかかわらず、Matcha-TTS では改善が見られ、VITS ではほとんど見られなかった。この対比は、事前学習で得た表現が品質の改善につながるかどうか、モデルの構造に依存することを示す。一方、大規模な事前学習を持つ F5-TTS が機能しなかったことは、事前学習の内容と適応方法の影響を示している。すなわち、音声資源の乏しい言語でのモデル選択は、事前学習済みモデルを使うというだけでは足りず、モデルの構造と、何をどう事前学習したモデルかまで含めて選ぶ必要がある。

5.3 自動評価と主観評価の関係

自動評価と主観評価によるモデルの順位は概ね一致し (Matcha-TTS > VITS > F5-TTS)、音声資源の乏しい条件においても、モデル選択のような粗い粒度の判断においては自動評価指標が主観評価の代替となりうることが示唆された。ただし本研究の主観評価はアイヌ語を理解できる第一著者を評価者とした 1 名による参考値であり、評価者間のばらつきを見積もれないため統計的な一般化はできない。また、評価はデータ量の多い KM 氏の発話に強く影響されており、話者間の品質差は本評価では分離できていない。

5.4 学習用途への含意

最後に、本研究の主眼である学習用途の観点では、最良

構成の MOS 3.28 は聴取可能な品質に達している一方、自然音声 (4.23) との差はなお大きい。特に音節末子音の脱落は、発音の手法としての利用では誤った発音の定着につながりうるため、実際に発音資料としての利用に向けては音節末の子音の再現の改善が最優先の課題である。

なお、本研究で学習したモデルおよび合成音声は、話者の特定が可能なほど元の話者の声色を再現していたため、公開していない。この判断の背景は次章の匿名化の課題で述べる。

6. 今後の課題とまとめ

本研究では、北海道アイヌ語の音声データ約 12.85 時間を整備し、3つの音声合成モデルを同一条件で比較した。その結果、(1) 事前学習済み Matcha-TTS のファインチューニングにより自然音声に最も近い品質が得られる一方、その差は依然残ること、(2) 合成品質はモデルと学習法の組み合わせに大きく依存し、ファインチューニングの効果はモデルの構造によって異なること、(3) 自動評価と主観評価の順位が概ね一致し、評価者の確保が困難な言語でも自動指標で開発を進められる見込みがあること、を示した。

アイヌ語 TTS の研究は萌芽段階であり、学習データの増強、入力表現の改善、評価体制の整備、倫理面の配慮など、多くの課題が残されている。まずデータの増強について、本研究で整備した以外にも利用可能な音源は存在する。今後は、公開されている音源についても利用許可を得た上で訓練データに含め、音声認識や自動アライメント技術などを用いて学習データを増強する必要がある。

また、本研究はラテン文字表記をそのまま入力としたが、考察で示した音節末子音の脱落や異音の崩れに対しては、その音素や交替を明示する表記を導入することが有効であると考えられる。アイヌ語の音声の特徴を反映した G2P の設計が今後の課題である。

評価については、今回の主観評価は評価者を募ることができなかったため、アイヌ語を理解できる第一著者のみが実施した。呈示順のランダム化と構成の秘匿によりバイアスの低減を図ったが、今後は複数のアイヌ語上級者による評価体制の整備が必要である。

最後に、実用面では合成音声の匿名化が重要な課題である。本研究では、スタイル変換を行わない場合に話者の特定ができるほど元話者の声色に似た合成音声生成されることを確認しており、前章で述べたモデル・合成音声の非公開の判断もこの懸念に基づく。同様の指摘はハイダ語やオジブウェ語などの先住民言語でもなされている [2]。アイヌ語音声処理の先行研究でも、学術目的であっても同意なく音声合成を行うことへの懸念が指摘されており [8]、話者の特定につながらない学習方法や音声の匿名化手法の整備が必須である。

謝辞 本研究は JSPS 科研費 JP24K02993 の助成を受けたものです。

参考文献

- [1] Krauss, M.: The world's languages in crisis, *Language*, Vol. 68, No. 1, pp. 4–10, 1992.
- [2] Wang, S., Yang, C., Parkhill, M., Quinn, C., Hammerly, C. and Zhu, J.: Developing multilingual speech synthesis system for Ojibwe, Mi'kmaq, and Maliseet, *Proc. NAACL-HLT 2025 (Short Papers)*, 2025.
- [3] Alper, M., Basu, S., Bennett, N., Kessler, R., Ravana, S. V. and Todd, W. F. K. S.: Towards low-resource text-to-speech generation for Indigenous Pacific Northwest languages: The case of Haida Xaad Kíl, *Proc. ICSNL 60*, 2025.
- [4] Chivoreanu, R. G. and Boros, T.: F5-TTS-RO: Extending F5-TTS to Romanian TTS via Lightweight Input Adaptation, arXiv preprint, arXiv:2512.12297, 2025.
- [5] 中川 裕: アイヌ語広文典, 白水社, 2024.
- [6] 田村 すゝ子: アイヌ語の世界 新装普及版, 吉川弘文館, 2013.
- [7] Matsuura, K., Ueno, S., Mimura, M., Sakai, S. and Kawahara, T.: Speech Corpus of Ainu Folklore and End-to-end Speech Recognition for Ainu Language, *Proc. LREC 2020*, pp. 2622–2628, 2020.
- [8] 河原 達也, 松浦 孝平: 言語の多様性とアイヌ語の音声言語処理, 日本音響学会誌, Vol. 81, No. 1, pp. 35–41, 2025.
- [9] 北海道ウタリ協会: アコロ イタク, クルーズ, 1994.
- [10] Shiraishi, H.: Phonetics and Phonology, In: Bugaeva, A. (ed.), *Handbook of the Ainu Language*, De Gruyter Mouton, pp. 772–804, 2022.
- [11] 全 炳河: テキスト音声合成技術の変遷と最先端, 日本音響学会誌, Vol. 74, No. 7, pp. 387–393, 2018.
- [12] Mehta, S., Tu, R., Beskow, J., Székely, É. and Henter, G. E.: Matcha-TTS: A fast TTS architecture with conditional flow matching, *Proc. ICASSP 2024*, 2024.
- [13] Kim, J., Kong, J. and Son, J.: Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech, *Proc. ICML 2021*, 2021.
- [14] Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K. and Chen, X.: F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching, *Proc. ACL 2025 (Long Papers)*, 2025.
- [15] 国立国語研究所: トピック別 アイヌ語会話辞典, (<https://ainu.ninjal.ac.jp/topic/>) (参照 2026-07-19).
- [16] 中川 裕, ブガエワ アンナ, 小林 美紀, 吉川 佳見: アイヌ語口承文芸コーパス—音声・グロス付き—, 国立国語研究所, 2016-2024. (<https://ainu.ninjal.ac.jp/folklore/>) (参照 2026-07-19).
- [17] 東京外国語大学アジア・アフリカ言語文化研究所: AA 研アイヌ語資料公開プロジェクト, (<https://ainugo.aa-ken.jp/>) (参照 2026-07-19).
- [18] Lv, Y., Li, H., Yan, Y., Liu, J., Xie, D. and Xie, L.: FreeV: Free Lunch For Vocoders Through Pseudo Inversed Mel Filter, *Proc. Interspeech 2024*, 2024.