Vol.2019-SLP-130 No.16
2019/12/6

民話を対象としたアイヌ語音声コーパスと End-to-End 音声認識

松浦 孝平[1]　上乃 聖[1]　三村 正人[1]　坂井 信輔[1]　河原 達也[1]

概要：我が国を構成する民族の一つであるアイヌは独自の言語を用いて文化を口頭で伝承してきたが，アイヌ語は母語話者の数が減り消滅の危機に瀕している。口頭伝承を録音することでアイヌ文化の保存が図られているが，その大部分は書き起こされておらず十分に活用されていない。我々は，アイヌ民族博物館と二風谷アイヌ文化博物館から提供されたデータをもとに，民話を対象に音声認識の研究を行っている。本稿では End-to-End モデルによる音声認識器の構築について報告する。モデルには注意機構と Connectionist Temporal Classification（CTC）を組み合わせたものを用いる。音素，音節，ワードピース，単語の各認識単位によって学習したモデルによる認識性能を比較し，単語認識精度と音素認識精度の両方について音節単位が最も高いという知見を得た。話者クローズの場合，各話者について 3 時間程度の学習データがあれば，単語認識精度で 80%以上，音素認識精度で 90%以上となることがわかった。話者オープンの場合，話者によって大きく異なるが，単語認識精度は平均的に 60%程度（音素認識精度は 85%程度）となった。また，日本語コーパスとのマルチリンガル学習の導入も行い，話者オープンの場合に効果を確認した。

1. はじめに

音声認識研究は、大規模なデータベースの構築と深層学習の導入により、飛躍的な発展を遂げて、実用的な認識性能を実現した。しかし、これは英語や日本語のように数千時間規模の書き起こしのあるデータベースがあることが前提で、主要言語に限定される。したがって今後の研究課題の一つは、このように大規模データベースが構築できない言語を対象とした音声認識である。

世界には５千以上の言語が存在するが、そのうち半数以上は死語になる危険性をはらんでいるとされている。日本には先住民族としてアイヌがいるが，明治以降その文化は急速に失われ、2009 年には UNESCO がアイヌ語を「極めて深刻」な消滅危機言語に認定した。一方で、アイヌ語を保存する活動も活発に行われており、消滅危機言語の中では例外的に大量の録音資料が作成・保存されている。しかし、その大半が書き起こしされておらず、十分に活用されていない。その書き起こしにはアイヌ語の知識を要するため、従事できる人が限られている。したがって、この音声アーカイブを（母語話者がほとんどいない状況で）効率的に書き起こすことが求められており、我々はこれを対象とした音声認識システムの研究開発を行っている。

白老町にあるアイヌ民族博物館[*1]（以下，白老）と平取町立二風谷アイヌ文化博物館[*2]（以下，平取）から提供された音声データを元に，音声認識研究のためのコーパスを構築した。アイヌ口頭伝承には民話と歌謡があるが，まずは民話を対象とすることにした。

音声認識のアプローチとして，近年 End-to-End モデルが提案され，従来手法である DNN-HMM ハイブリッドモデルと同等以上の認識精度を達成しつつある [1–3]。このモデルは大部分がニューラルネットワークにより構成されており，従来手法のような複雑な階層構造を持たず，なおかつ言語に関する専門知識を必要としないという利点を持つ。本研究では，学習モデルに Sequence-to-Sequence モデル [4] の一種である注意機構モデル [5, 6] を採用し，Connectionist Temporal Classification [7, 8] を用いた補助タスクと合わせて学習を行う。本稿ではアイヌ語音声認識に対して最適な認識単位について検討する。また，データ量の不足を補うために，日本語コーパスを活用することも検討する。

本稿の構成を以下に示す。まず第２章でアイヌ語の概要を記し，第３章で音声認識のためのコーパスの構成について述べる。第４章で End-to-End 音声認識について概観し，実験に使用したモデルについて説明をする。第５章では評

[1] 京都大学大学院 情報学研究科

[*1] http://ainugo.ainu-museum.or.jp/
[*2] http://www.town.biratori.hokkaido.jp/biratori/nibutani/

表 1: 各話者のデータ詳細

話者 ID	KM	UT	KT	HS	NN	KS	HY	KK	合計
説話の数	29	26	20	8	8	11	8	7	114
発話時間	19:40:58	7:14:53	3:13:37	2:05:39	1:44:32	1:43:29	1:36:35	1:34:55	38:54:38
発話数	9170	3610	2273	2089	2273	1302	1220	1109	22345

価実験の詳細と結果を示す。

2. アイヌ語の概要

2.1 アイヌ語と表記体系

アイヌ語は膠着語であり文法面では日本語との類似点も見られるが，系統不明の孤立語である。特徴として，閉音節をもつこと，頻繁に複数の単語が結合し新たな単語が構成されること，動詞によって取ることのできる目的語の数が定まっていることなどが挙げられる。

現在，アイヌ語の大部分は北海道ウタリ協会が編集したアイヌ語学習の参考書『アコロイタク』 [9] で範示される表記法に基づいて記述されている。この表記法では，17 種類のアルファベット {a, c, e, h, i, k, m, n, o, p, r, s, t, u, w, y, ’} で書き表される。これらは発音とほとんど一対一対応しているため，本研究において便宜上「音素」と呼ぶことにする。なお，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。加えて，人称接続を表す {=} が含められる他，発音の脱落を表す {_, __} などが表記されることもある。これらは明示的に発音されない。以下に具体例を挙げる。

原文	uymam’=an wa isam=an __hi okake ta
訳	私が交易に行っていなくなった後で

2.2 アイヌ語資料の分類

アイヌ語は大きく，北海道アイヌ語，千島アイヌ語，樺太アイヌ語の 3 つに分類でき，各々はさらに詳細に区分される。白老と平取のデータに含まれるアイヌ語音声はいずれも北海道アイヌ語に属する沙流方言である。

アイヌ口頭伝承は歌謡のユカラ（英雄叙事詩），カムイユカラ（神謡）と，会話調で語られる民話のウエペケレ（散文説話）の 3 種類に分類される。本研究では音声認識が比較的容易と思われるウエペケレ（散文説話）を対象とする。

3. 音声認識用コーパス

3.1 話者と説話の数

本研究で構築したコーパスに含まれる各話者の説話数と発話時間の詳細を，発話時間の長い順に並べて表 1 に示す。本コーパスに含まれる話者数は，白老のデータから 2 名（話者 KM と話者 UT）と平取のデータ（その他の話者）から 6 名の計 8 名である。発話時間には大きなばらつきがあり，話者 KM が単独で約半分を占める一方で，話者 HS，HY，KK，KS，NN はいずれも全体の 5 %程度である。収録の性質上，このような話者数の少なさやデータ量の偏りといった特性が生じる。

以下に実際の説話の一部を掲載する。

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

民話 「ポロシルンカムイになった少年」より一部抜粋

3.2 データの加工

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 節で述べた文字のうち，{=} 以外の記号 {’, _, __} を除く。例えば，原文は次のように処理される。

原文	uymam’=an wa isam=an __hi okake ta
処理後	uymam=an wa isam=an hi okake ta

また，提供されたデータは意味的な単位ごとに区切られてタイムスタンプが付与されていたが，発話の途中に意味的な単位の境界が置かれる例が多くみられた。音声認識モデルの学習には無音区間で区切られた単位（IPU）の方が有用であるため，人手でアノテーションを行った。これにより各話者の発話数は表 1 のようになった。

4. End-to-End 音声認識

4.1 End-to-End モデル

End-to-End モデルは音響特徴量から直接文字列や単語列を推定する手法である。従来の DNN-HMM ハイブリッドモデルと比較して単純な構成であり，音響モデルと分離して，発音辞書や言語モデルを設計・学習する必要がない。End-to-End モデルの実現方法には Connectionist Temporal Classification（以下，CTC）を用いるものや Sequence-to-Sequence（以下，seq2seq）モデルの拡張であ

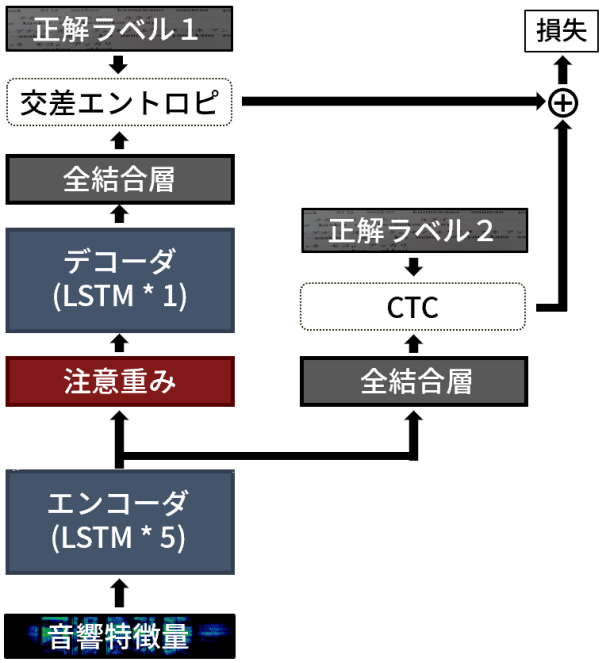

図 1: CTC と注意機構を組み合わせたモデル

る注意機構モデル等が挙げられる。

CTC は時間フレーム毎に RNN (Recurrent Neural Network) で分類される記号列を縮約することで識別を行う。その際に，ブランク記号（ϕ）を導入するとともに，連結する同一記号をまとめることで縮約する。具体例を次に示す。

$$\mathrm{aab}\phi\mathrm{bbccc} \rightarrow \mathrm{abbc}$$

ここで，入力される音響特徴列 $\boldsymbol{X}$ の下で正解ラベル列 $\boldsymbol{L}$ が出力される確率を次のように定義する。（$|\boldsymbol{L}| \leq |\boldsymbol{X}|$）

$$p(\boldsymbol{L}|\boldsymbol{X}) = \sum_{\substack{\boldsymbol{\Pi} \in \mathcal{B}^{-1}(\boldsymbol{L}) \\ |\boldsymbol{\Pi}| = |\boldsymbol{X}|}} p(\boldsymbol{\Pi}|\boldsymbol{X}) \qquad (1)$$

$\mathcal{B}$ はラベル系列を縮約する関数であり，$\mathcal{B}^{-1}(\boldsymbol{L})$ は縮約すると L になる系列の集合を表す。この確率 $p(\boldsymbol{L}|\boldsymbol{X})$ の対数に -1 をかけたものを CTC による損失と定義する。

seq2seq モデルは，系列長の異なる 2 つの系列を写像する方法であり，入力をフレーム毎に分散表現に変換するエンコーダと分散表現から記号列の出力を行うデコータの 2 つの RNN から構成される。これにより，異なる長さの入出力を扱うことが可能となっている。エンコーダが各フレームを処理した分散表現をすべて保存しておき，どの時点の分散表現が現在のデコードにとって重要かを表す注意重みとの内積をとってデコーダに送るのが注意機構モデルである。

注意機構モデルと CTC を組み合わせることで音声認識性能が向上することが報告されている [10, 11]。本研究では双方向 LSTM (Long Short-Term Memory) [12] を用いて，図 1 のように注意機構と CTC でエンコーダを共有させたモデルを実装した。

表 2: 各認識単位における正解ラベルの例

原文	a=saha i=kokopan wa
音素単位	a = s a h a ⟨wb⟩ i = k o k o p a n ⟨wb⟩ w a
音節単位	a = sa ha ⟨wb⟩ i = ko ko pan ⟨wb⟩ wa
WP 単位	⟨wb⟩a = saha ⟨wb⟩i = ko ko p an ⟨wb⟩wa
単語単位	a = saha i = ⟨unk⟩ wa
日本語訳	姉さんはだめだと言って

4.2 認識単位

従来の DNN-HMM ハイブリッドモデルはフレーム毎に音響特徴量から音素状態を認識し，それを発音辞書と言語モデルを用いて単語列に変換する。一方で，End-to-End モデルでは認識単位について自由度があるため，音響特徴量から直接文字や単語を推定する研究がなされている [13–15]。その他，文字や単語といった既存の単位によらず，単語を頻出の部分列に分割するワードピースモデルが提案されている [16]。

本稿においてもアイヌ語 End-to-End 音声認識における最良の認識単位を模索する。ただし，認識結果の活用の容易さを考えて単語単位に復元できることを条件とする。候補は音素，音節，ワードピース (WP)，単語単位の 4 種類とする。各々の具体例を表 2 に示し，その詳細を以下に述べる。

4.2.1 音素単位

2.1 節で記した通り，本研究ではアイヌ語が記述されているアルファベット自体を音素として扱う。また，単語境界（‘⟨wb⟩’）と人称接続をあらわす ‘=’ も同時に学習させることで，表 2 の「原文」に示されるような単語単位の表記を復元することを可能にする。

4.2.2 音節単位

アイヌ語の音節は CV か CVC からなる。（ここで，C は子音，V は母音を表す。）任意の単語は，以下の手続きによって自動的に音節に分割される。

(1) 一文字の単語はそのままとする
(2) C や V が隣接する場合，そこで分割する
(3) 部分列が V から始まる場合，その直後で分割する。ただし部分列が VC のみの場合は分割しない
(4) CV か CVC が残るまで，残った各部分列の先頭から CV を切り離してゆく

例えば ‘esirkirap’（「難儀する」）と言う単語は次のように音節へと分割される。

$$\mathrm{esirkirap} \rightarrow \mathrm{esir\text{-}kirap} \rightarrow \mathrm{e\text{-}sir\text{-}kirap} \rightarrow \mathrm{e\text{-}sir\text{-}ki\text{-}rap}$$

ただし，これによって常に正しい音節に分割されるわけではない。例えば，‘isermakus’（「無事を祈る」）という単語は上記の手続きによって ‘i-ser-ma-kus’ と分割されるが，正しくは ‘i-ser-mak-us’ と発音されるべきである。本研究では前者を「音節」とみなして認識単位とする。また，音素単位と同様の理由で ‘⟨wb⟩’ と ‘=’ を認識単位に含める。

4.2.3 ワードピース単位

単語をサブワードに分割するその他の手法として，BPE (Byte Pair Encoding) [17] や Unigram 言語モデル [18] によるワードピースへの分割が挙げられる。前者は貪欲的に分割を決定していくのに対し，後者は EM アルゴリズムを用いて各部分列の出現尤度を最大化するよう分割を決定する。本研究では後者を採用し，外部ツールの Sentencepiece（`https://github.com/google/sentencepiece`）を用いてテキストをワードピース単位へと変換する。この時，表 2 の例のように ⟨wb⟩ と単語の部分列が接合して一つの認識単位となることがある。

4.2.4 単語単位

元の書き起こしテキストを空白で区切ったものを単語単位とする。ただし，‘=’ は独立した一単語として扱う。また，コーパス内における低頻度語の学習が困難なため，出現回数が 2 回未満のものを特殊ラベル ‘⟨unk⟩’ で置き換える。例えば表 2 において ‘a=saha’ は ‘a’ と ‘=’ と ‘saha’ の 3 単語として扱い，‘kokopan’ は ‘⟨unk⟩’ で置き換える。

4.3 マルチリンガル学習

少資源言語のモデル学習法として，データ量の多い他の言語のコーパスを転用してモデルを学習させるマルチリンガル学習が提案されている [19, 20]。本稿では日本語コーパスの活用を検討する。表 1 に載せた 8 名の話者はアイヌ語と日本語の二重言語話者であり，今回準備したコーパスに含まれるアイヌ語は少なからず日本語の影響を受けていると考えられる。このため，日本語によるデータ拡張が有効であると期待できる。

具体的には，注意機構モデルのデコーダを言語ごとに用意し，エンコーダと注意重みの計算部分を共有させて学習を行う。日本語コーパスには新聞記事読み上げコーパス（JNAS） [21] を使用する。

5. 評価実験

5.1 データの区分

本稿では，学習時と評価時の話者集合が同じである場合（以下，「話者クローズ」）とそうでない場合（以下，「話者オープン」）の 2 通りについて実験を行う。後者は未知の話者の音声データを想定したものである。話者クローズの実験における開発セットと評価セットには各話者から 1 説話ずつ選んだ。この時，開発セットは 1585 発話の 2 時間 23 分，評価セットは 1841 発話の 2 時間 48 分となった。話者オープンの場合は学習データに含まれない話者の全てのデータを評価データとする。ただし，データ量の多い話者 KM と UT を学習から除くと，モデルの学習が困難なため，これらは評価に用いない。

5.2 実験条件

入力には，40 次元の対数メル尺度フィルタバンクを 10ms 毎に求めたものを 3 つずつフレームスタッキング [22] した 120 次元の音響特徴量を使用する。

双方向 LSTM の隠れ状態数は 320×2 次元とし，注意機構モデルにおけるエンコーダは双方向 LSTM 5 層，デコーダは 1 層とする。LSTM はバイアスを 0，重みを He [23] に基づいて初期化する。全結合層や注意機構に含まれるアフィン変換はバイアスを 0，重みを一様分布 $\mathcal{U}(-0.1, 0.1)$ で初期化する。また，過学習への対策として減衰率 10^{-5} の重み減衰 (Weight Decay) [24] と $\mathcal{B}e(0.2)$ に従うドロップアウトを行い [25]，パラメータは Adam [26] によって更新する。初期の学習率は 10^{-3} であり，30 エポック目と 35 エポック目で学習率を 10^{-1} 倍しつつ全部で 40 エポック学習させる。ミニバッチサイズは 30 とし，発話長で昇順にソートした。学習を安定させるため 12 秒を超える長さの発話は学習データから取り除いた。これによって，データ量はおよそ 4 分の 3 程度になった。

モデル全体の損失関数 $\mathcal{L}_{\text{all}}$ は，注意機構の損失 $\mathcal{L}_{\text{attn}}$ と CTC によって計算される損失 $\mathcal{L}_{\text{CTC}}$ の線形和

$$\mathcal{L}_{\text{all}} = \lambda \mathcal{L}_{\text{attn}} + (1 - \lambda)\mathcal{L}_{\text{CTC}} \tag{2}$$

によって求める。ここで，λ は 0.5 とした。また，全実験を通して CTC を用いた補助タスクは音素単位で学習した。

各認識単位におけるラベルのクラス数は厳密には学習データによって異なるが，テキストの開始と終了を表す特殊記号を含めて，音素単位が 25，音節単位が約 500，単語単位が約 4000（話者 KM を除いた学習データでは約 3000）となった。ワードピースの語彙サイズは開発セットによる予備実験に基づいて約 500 に設定した。

日本語とのマルチリンガル学習においては，認識単位を音節単位とした。JNAS も 12 秒を超える発話は除き，約 80 時間の学習データとした。

5.3 実験結果

表 3 に，話者クローズと話者オープン各々における各認識単位の単語誤り率と音素誤り率を示す。また，各話者の誤り率を，評価されるトークン数で加重平均をとったものを「平均」に示す。

表 3 から，話者クローズで各々 3 時間程度の学習データがあれば 80%以上の単語認識精度が得られることがわか

表 3: 各認識単位における音声認識性能の比較

		認識単位	KM	UT	KT	HS	NN	KS	HY	KK	平均
話者クローズ	単語誤り率	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**
		WP	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
	音素誤り率	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**
		WP	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	単語誤り率	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**
		WP	-	-	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
	音素誤り率	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**
		WP	-	-	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: マルチリンガル学習の結果。いずれも音節単位で認識

		話者クローズ	話者オープン
アイヌ語	単語誤り率	21.7	38.6
	音素誤り率	6.3	13.8
+日本語	単語誤り率	21.1	34.8
	音素誤り率	6.0	11.7

る。話者オープンの場合，話者によって大きな差（50%〜70%）があるが，平均的には 60%程度の単語認識精度である。音素認識精度は前者が 90%以上，後者が 85%程度である。

単語単位認識の音素誤り率が話者クローズ，オープンに関わらず著しく低い。その他の単位の場合は未知語に対して認識誤りが含まれていても類似した音素列を出力できるのに対し，単語単位で学習した場合には未知語が特殊ラベル ‘⟨unk⟩’ で置き換わることが理由であると考えられる。例えば下記のような出力結果が実際に見られるが，音素誤り率を計算すると音節単位では 5.0%であるのに対して単語単位では 30.0%となる。（音節単位の認識結果は単語単位に変換している。）

正解	i okake un a unuhu a onaha
音節単位	piokake un a unuhu a onaha
単語単位	⟨unk⟩ un a unuhu a onaha

単語誤り率と音素誤り率の両方について音節単位が最も高い精度となった。音響的なマッチングの安定性と言語的制約の両面から音節単位が適当であると考えられる。

また，音素誤り率に対して単語誤り率が大幅に高くなっている。第 2 章で述べた通り，アイヌ語は頻繁に複合語を構成する。複合語を構成する単語は語形が変化しないことも多いため，モデルが複数の単語を複合語として出力するか否かで混乱する場合がある。実際に，音素誤りは無いが単語列として誤っている次のような出力例が散見される。

正解	nen poka apkas an mak an kusu
出力例	nenpoka apkas an makan kusu

最後に，マルチリンガル学習を行った場合の各話者の加重平均を表 4「+日本語」の行に示す。話者クローズの場合は改善幅がわずかだが，話者オープンの場合は単語認識精度で相対的に 10%の改善が見られた。マルチリンガル学習は特に話者オープンの条件下で有効であることがわかる。

6. おわりに

本研究では，まず白老町のアイヌ民族博物館と平取町立二風谷アイヌ文化博物館から提供されたデータを元に，音声認識のためのコーパスを整備した。注意機構と CTC を用いた End-to-End モデルをこのコーパスを使用して学習し，音素単位，音節単位，ワードピース単位，単語単位の各々について，話者クローズと話者オープンの場合における音声認識性能を比較した。この結果，音節単位が最も高い精度を得ることがわかり，話者クローズの場合は約 80%の単語認識精度となった。話者オープンの場合だと 60%程度まで低下したが，これは日本語とのマルチリンガル学習によって多少改善された。今後は，性能の低い話者の改善

を図る予定である。

謝辞 本研究で使用したアイヌ語のデータは，白老町のアイヌ民族博物館と平取町立二風谷アイヌ文化博物館から提供されたものである。また，アイヌ語に関して様々なご助言をいただいた札幌学院大学人文学部人間科学科の奥田統己教授に謝意を表する。

参考文献

[1] Chiu, C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, K., Jaitly, N., Li, B., Chorowski, J. and Bacchiani, M.: State-of-the-art Speech Recognition With Sequence-to-Sequence Models, *CoRR*, Vol. abs/1712.01769 (online), available from ⟨http://arxiv.org/abs/1712.01769⟩ (2017).

[2] Lüscher, C., Beck, E., Irie, K., Kitza, M., Michel, W., Zeyer, A., Schlüter, R. and Ney, H.: RWTH ASR Systems for LibriSpeech: Hybrid vs Attention - w/o Data Augmentation, *CoRR*, Vol. abs/1905.03072 (online), available from ⟨http://arxiv.org/abs/1905.03072⟩ (2019).

[3] Han, K. J., Prieto, R., Wu, K. and Ma, T.: State-of-the-Art Speech Recognition Using Multi-Stream Self-Attention With Dilated 1D Convolutions (2019).

[4] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to Sequence Learning with Neural Networks, *CoRR*, Vol. abs/1409.3215 (online), available from ⟨http://arxiv.org/abs/1409.3215⟩ (2014).

[5] Chorowski, J., Bahdanau, D., Cho, K. and Bengio, Y.: End-to-end continuous speech recognition using attention-based recurrent nn: First results, *NIPS 2014 Workshop on Deep Learning, December 2014* (2014).

[6] Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P. and Bengio, Y.: End-to-end attention-based large vocabulary speech recognition, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4945–4949 (online), DOI: 10.1109/ICASSP.2016.7472618 (2016).

[7] Graves, A., Fernández, S., Gomez, F. J. and Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, *ICML* (2006).

[8] Graves, A. and Jaitly, N.: Towards End-To-End Speech Recognition with Recurrent Neural Networks, *ICML* (2014).

[9] 北海道ウタリ協会: アコロイタㇰ アイヌ語テキスト 1，クルーズ (1994).

[10] Hori, T., Watanabe, S., Zhang, Y. and Chan, W.: Advances in Joint CTC-Attention based End-to-End Speech Recognition with a Deep CNN Encoder and RNN-LM, *CoRR*, Vol. abs/1706.02737 (online), available from ⟨http://arxiv.org/abs/1706.02737⟩ (2017).

[11] Watanabe, S., Hori, T., Kim, S., Hershey, J. R. and Hayashi, T.: Hybrid CTC/Attention Architecture for End-to-End Speech Recognition, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 11, No. 8, pp. 1240–1253 (online), DOI: 10.1109/JSTSP.2017.2763455 (2017).

[12] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Comput.*, Vol. 9, No. 8, pp. 1735–1780 (online), DOI: 10.1162/neco.1997.9.8.1735 (1997).

[13] Li, J., Ye, G., Zhao, R., Droppo, J. and Gong, Y.: Acoustic-To-Word Model Without OOV, *CoRR*, Vol. abs/1711.10136 (online), available from ⟨http://arxiv.org/abs/1711.10136⟩ (2017).

[14] Chan, W., Jaitly, N., Le, Q. V. and Vinyals, O.: Listen, Attend and Spell, *CoRR*, Vol. abs/1508.01211 (online), available from ⟨http://arxiv.org/abs/1508.01211⟩ (2015).

[15] Li, J., Ye, G., Das, A., Zhao, R. and Gong, Y.: Advancing Acoustic-to-Word CTC Model, *CoRR*, Vol. abs/1803.05566 (online), available from ⟨http://arxiv.org/abs/1803.05566⟩ (2018).

[16] Schuster, M. and Nakajima, K.: Japanese and Korean voice search, *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5149–5152 (2012).

[17] Sennrich, R., Haddow, B. and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, *CoRR*, Vol. abs/1508.07909 (online), available from ⟨http://arxiv.org/abs/1508.07909⟩ (2015).

[18] Kudo, T. and Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, *CoRR*, Vol. abs/1808.06226 (online), available from ⟨http://arxiv.org/abs/1808.06226⟩ (2018).

[19] Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P., Weinstein, E. and Rao, K.: Multilingual Speech Recognition with A Single End-To-End Model, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2018).

[20] Cho, J., Baskar, M. K., Li, R., Wiesner, M., Mallidi, S. H. R., Yalta, N., Karafiát, M., Watanabe, S. and Hori, T.: Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling, *CoRR*, Vol. abs/1810.03459 (online), available from ⟨http://arxiv.org/abs/1810.03459⟩ (2018).

[21] Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K. and Itahashi, S.: JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research, *Journal of the Acoustical Society of Japan (E)*, Vol. 20, No. 3, pp. 199–206 (online), DOI: 10.1250/ast.20.199 (1999).

[22] Tian, X., Zhang, J., Ma, Z., He, Y. and Wei, J.: Frame Stacking and Retaining for Recurrent Neural Network Acoustic Model, *CoRR*, Vol. abs/1705.05992 (online), available from ⟨http://arxiv.org/abs/1705.05992⟩ (2017).

[23] He, K., Zhang, X., Ren, S. and Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, *CoRR*, Vol. abs/1502.01852 (online), available from ⟨http://arxiv.org/abs/1502.01852⟩ (2015).

[24] Krogh, A. and Hertz, J. A.: A Simple Weight Decay Can Improve Generalization, *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 4*, Morgan Kaufmann, pp. 950–957 (1992).

[25] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *Journal of Machine Learning Research*, Vol. 15, pp. 1929–1958 (online), available from ⟨http://jmlr.org/papers/v15/srivastava14a.html⟩ (2014).

[26] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *CoRR*, Vol. abs/1412.6980 (online), available from ⟨http://arxiv.org/abs/1412.6980⟩ (2014).