End-to-End 音声認識と音声からのダイレクト点訳への応用

小林彰夫 [1)]，大西淳児 [2)]

筑波技術大学　産業技術学部　産業情報学科 [1)]
筑波技術大学　保健科学部　情報システム学科 [2)]

要旨：本稿では，情報保障で用いられる音声認識技術について，その基礎から最近の動向までを概観する。続いて，深層学習に基づくend-to-end（エンド・ツー・エンド）音声認識を用いた，音声からのダイレクト点訳に関する筆者らの研究を紹介する。日本では，音声からの点訳による盲ろう者の情報アクセシビリティの向上が期待されている。日本の点字は，日本語音声の特徴を反映したかな文字が基盤であり，音声認識との親和性が高い。筆者らは，ニューラルネットワークを用いて音声から点字を直接生成することを試みた。また，音声認識と点訳を組み合わせた手法と比較した。

キーワード：盲ろう，点字，音声認識，end-to-end，ニューラルネットワーク

1．はじめに

スマートフォンの普及により，私たちは日常的にさまざまなアプリケーションを使うようになった。とりわけ音声認識アプリケーションは，スマートフォンで捉えた音声を即座に文字に変換できることから，身近な情報保障手段として利用した経験のある読者もいることだろう。一方で，音声認識技術は，かな漢字のような視覚的な文字を生成するだけではなく，点字のような触覚による文字の生成にも利用可能である。本稿では，筆者らが最近取り組んでいる音声認識を用いた点訳手法 [1] について，音声認識の基礎を説明したうえで，わかりやすく紹介する。

2．音声認識

本章では，音声言語と音声認識に関する基本的な説明を行う。紙数の制限上，以下では専門的な学術用語の多用を避け，正確さよりも直観的な理解を優先して記述している。図版も同様である。音声認識の詳細に興味のある読者は，大学院生向けではあるが文献 [2,3,4] といった成書を参考にしていただきたい。

2.1　音声言語の階層構造

音声言語は，図 1 に示すように最も基本的な単位を音素として，語や文を上位とする階層的な構造によって表される。音素とは意味の異なりを表す音の最小の単位である。例えば，「愛（あい）」と「青（あお）」という 2 つの語は「い」と「お」の音の違いによって意味の違いが生じる。「い」と「お」はそれぞれ母音を表す音素 /i/ および /o/ で表される。日本語には母音や子音を表すさまざまな音素があり，定義によってその異なり数が変わるが，音声認識で使われる音素はおよそ 40 種類である。音素は音と文字を結びつける記号であって，音素の並びが具体的な文字や語と対応づけられ，語が連結して句や文，パラグラフといったより複雑な構造を生み出していく。

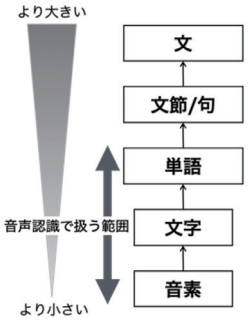

図1　音声言語の構造

2.2　調音と音響分析

人間が音声を生成するさまを指して調音運動とよぶ。声帯を震わせる，口唇を動かす，あるいは口形を変えるなどの運動によって，さまざまな音声が生み出される。調音運動の結果は物理的な音声波形であり，さまざまな高さ（周波数）と強さを持った音が含まれる。どのような高さ・強さの音が含まれているかは，音声全体を短い区間（フレーム）に分割して分析し，音響特徴を得ることでわかる。分析した音響特徴のフレームを時間軸に沿って並べたものがサウンドスペクトログラム（声紋）である（図 2）。スペ

クトログラムを観察すると，音の高さ・強さによる濃淡のパターンがみて取れる。これらのパターンは音素ごとに，話し手によらない共通した特徴を持っている。この特徴を音素，または文字や語に対応づけることが音声認識である。

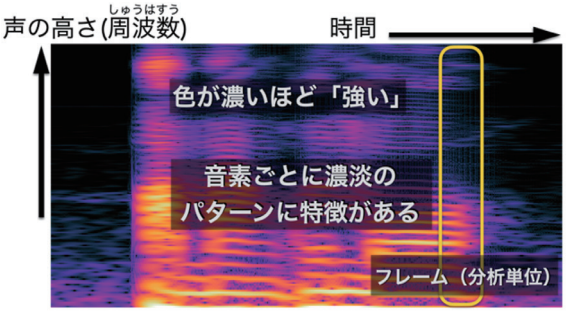

図2　サウンドスペクトログラム

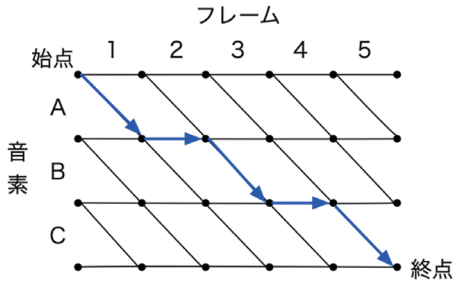

図3　フレームと音素の対応

2.3　音声認識

音声認識は，広くパターン認識とよばれる技術の一つである。音声認識における学習とは，音響特徴のフレームと該当する音素（厳密には音素を表す統計モデル・有限状態機械）とを対応づけることである。音響特徴のフレーム数と該当する音素の数は異なり，一般的にはフレーム数のほうが音素の数よりもはるかに多い。したがって，長さの異なるものどうしを対応づけることになる。フレームと音素の対応づけは，図3のように表すことができる。いま，音響特徴5フレームを3つの音素 A，B，C と対応づけるとする。まず，図の格子の縦横にそれぞれフレーム番号 1，2，…，5，音素を表すアルファベット A，B，C を振っておく。フレームと音素の対応関係は，図の格子の左上の頂点（始点）から右下への頂点（終点）をたどる経路として表すことができる。頂点から頂点へ辺上をたどる移動にはコスト（確率で表される）がかかり，音響特徴を表すフレームが音素の特徴をどれだけ反映しているか（類似しているか）に依存して移動コストが変わる。始点から終点に辿り着く経路はさまざまであるから，任意の経路についての移動コストの累計（経路のもっともらしさ）もまた異なる。移動および経路のコスト，すなわちコストを表す確率を，与えられた音響特徴と音素のペアから推定することが学習である。

音声認識を実行する際には，音響特徴の系列に対してもっともらしい音素系列を推定する（探索とよぶ）。しかし，すべての音素の組み合わせをしらみつぶしに探索することは物理的に困難なため，もっともらしい組み合わせのみを効率よく探索する手法（ビームサーチ）を用いる。音素から語，文へは言語的な情報に基づいて推定を行う。従来の音声認識では，音素を並べた音素列と単語との対応テーブル（発音辞書）用いて単語を求め，さらに単語どうしのつながりやすさ（統計的言語モデル）から文を推定する。

2.4　End-to-End 音声認識

近年，音響特徴と音素を対応づけるのではなく，深層学習の枠組みを用いて音響特徴と文字・語，あるいは音声波形そのものと文字・語を対応づける end-to-end（エンド・ツー・エンド）音声認識手法が数多く提案されている。End-to-end 手法では，図1に示した音声言語の階層構造が単純化されるなどの利点がある。例えば，従来の音声認識では，音素の系列から単語への変換で用いられる発音辞書は人手により作成されることが一般的であった。したがって，単語が10万以上ともなれば，発音辞書の作成は極めてコストが高くつく。End-to-end 音声認識であれば，音素を推定することなく文字や語を直接推定するため発音辞書は不要である。End-to-end 音声認識の利点を活かすことにより，低コストのソフトウェアの研究開発が可能となる。

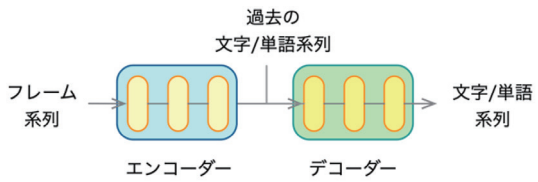

図4　エンコーダー・デコーダーモデル

End-to-end 音声認識では，その多くがエンコーダー・デコーダーモデルとよばれる方式を用いる（図4）。エンコーダー，デコーダーはそれぞれニューラルネットワークにより実装される。エンコーダーは音響特徴のフレームの系列を受け取って，音声の特徴を端的に表すような多次元のベクトルに変換する。一方，デコーダーにはさまざまな実装があるが，トランスデューサー [5] とよばれる手法では，エンコーダーから出力される多次元ベクトルと，出力済みの過去の文字・単語の系列から，当該系列に引き続く文字・単語を予測する。トランスデューサー内部では，2.3で述べたような入力と出力の対応を行って文字・単語の系列（音声認識結果）を出力する。

2.5　音声認識の性能評価

音声認識では，音素，文字，単語を単位とした誤り率に基づく評価が一般的である。いずれを単位とした場合でも，動的計画法とよばれる手法で正解と音声認識結果を対応させる。対応の結果，以下の３種類の誤りが生じうる。

① 置換誤り… 正解と対応する音声認識結果の音素，文字，単語が異なる。
② 挿入誤り… 正解に存在しない音素，文字，単語が認識結果に挿入されている。
③ 脱落誤り… 正解の音素，文字，単語が音声認識結果から脱落している。

文字誤り率は，上記の３種類の文字誤りの総和（編集距離）を正解の文字数で割ることにより求められる。正解と認識結果が一致した数（正解率）ではなく，誤りの数に基づいて性能評価する理由は，挿入誤りが音声認識の性能に影響を与えるからである。挿入誤りは正解には存在しない文字や単語であり，正解率の計算では計数されないため，性能評価の指標としては適当とはいえない。

3. 音声点訳

3.1　盲ろう者と点字

聴覚と視覚に障害のある盲ろう者は，全国に 14,000 人程度いるとされる [6]。一般に盲ろうといっても，視覚・聴覚それぞれの障害に応じて情報の受発信方法は多様である。聴覚障害ののちに視覚障害となった者（ろうベース）は,元来手話や書きことばによる情報のやりとりを主体とする。視覚障害ののちに聴覚障害となった者（盲ベース）の場合，音声や点字による情報の受発信が主となる。盲ろう者への点字による情報保障を考えた場合，日本語の点字体系がかな文字をベースとすることから，音声との親和性は高いといえ，音声から点字への変換（音声点訳）は，継続的かつ安価な情報保障手段を盲ろう者に提供できる可能性がある。そこで著者らは，盲ろう者向けの情報保障手段という観点から，音声から点訳を行う手法を検討した。

3.2　点字と点訳

日本語の点字は，6 点点字に基づくかな文字および英字･数字による表記が一般的である [7]。6 点によって表現可能な文字はスペースを含めて 64 通りに過ぎないので，かな以外の英字や数字を表すために数符や外字符といった特殊な符号を併用する。例えば，かな漢字表記「6時になりました。」は図4に示すように，数符を置いて数字「6」を表記する。数符がなければ「エ」である。「ジ」は濁音を表す５の点を前置して表す。点字の表記においては，助詞「は」を「ワ」で表すことや，長音（例えば「空気（くうき）」→「クーキ」）の使い方に特徴がある。点字による

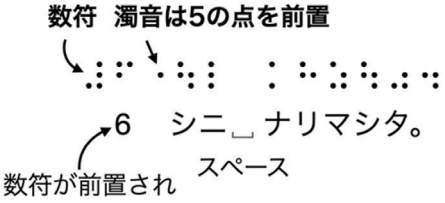

図５　点訳の例

文は，概ね文節を区切りとして空白で分かち書きを行い，構造を触読者に理解しやすくする。

3.3　音声点訳

日本語の点字の体系がかな文字を基盤としていることから，かな漢字を出力する日本語の音声認識を点字に変更すれば，盲ろう者への情報保障手段としての音声点訳が実現できる。音声点訳については，これまで研究例 [8] があるものの，日本語を対象とした研究は少ない。盲ろう者への支援として行われる音声からの点訳は，市販ソフトウェア（音声認識および点訳ソフトウェア）の組み合わせが試行されているが，学術文献は乏しく性能は明確に示されていない。本稿では，音声認識結果から点訳する方法を従来法として２段階点訳とよぶことにする。２段階点訳の場合，もとの音（音韻）に関する情報が欠落したかな漢字からの点訳となるため，読み方により意味の変わる語や固有名詞を正確に点訳できない可能性がある。また，音声認識結果からの点訳では，認識誤りを含む発話からの点訳となるため，点訳結果の品質が劣化する。前章で述べた end-to-end 手法を適用すれば，音声の持つ音韻性を活用しながら点字をダイレクトに生成可能である。図６に示すように，２段階音声点訳では，第１のニューラルネットワークで音響特徴からかな漢字出力（音声認識結果）を得たのちに，第２のネットワークで点訳を行うため構造が冗長である。一方，end-to-end 音声点訳では，１つのニューラルネットワークのみで音響特徴から直接点字を出力する。

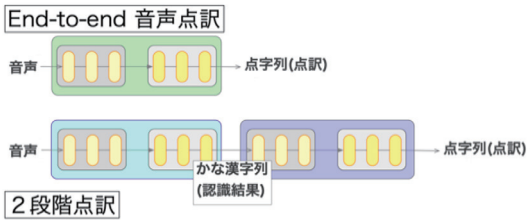

図６　End-to-End 音声点訳と2段階点訳

End-to-end 手法による音声点訳を情報保障システムの核として活用する場合，音声認識や点訳ソフトウェアを個別に用意・調整する必要はなく，安価にシステムを構成可能である。

3.4　音声点訳実験

End-to-end 音声点訳の実験では，日本音響学会新聞記事読み上げ音声コーパス（JNAS）[9] および NHK のニュース読み上げを学習に用いた．また，同ニュース読み上げ 742 発話（かな漢字で約 39,000 文字，点字で約 73,000 文字）を使って音声認識性能を評価した。音声認識は,conformer-transducer（コンフォーマー･トランスデューサー）[10] とよばれるエンコーダー・デコーダーモデルの一形態を用いた。実験結果を表 1 に示す。表中，音声認識欄は 2 段階点訳におけるかな漢字の文字誤り率，点訳欄は点字の文字誤り率である。2 段階点訳では，音声認識結果に誤りが含まれない（表 1 の 0.0%）場合に，点訳の誤り率が 3.4% となり，end-to-end 音声点訳の 3.7% を上回る性能となった。しかし，実際には音声認識結果に誤りが含まれる（2.7%）ため，end-to-end 点訳に比べて点訳の誤りが大きくなった（5.0%）。点訳結果について一対比較による検定 [11] を行ったところ，危険率 5% で有意となった。

表1　音声点訳の実験結果（文字誤り率，%）

	音声認識	点訳
2段階点訳	2.7	5.0
	0.0	3.4
End-to-end	—	3.7

End-to-end 音声点訳結果の誤りの傾向を調べたところ，文節区切りを表すスペースの誤りが全体の 13.3%を占めた。点字のみを用いた学習では，日本語の文節レベルの情報が反映されにくく，正しく分かち書きされないと考えられ，文節境界の情報を反映した学習手法が必要だといえる。また，外字符の誤りが全体の 14.6%を占め，そのうち 55%が脱落誤りであった。これは，固有名詞に使われることの多い英文字が，外字符の脱落により正しく触読されないケースがあるということを意味している。したがって，誤読を誘発しかねない符号の脱落誤りの削減も課題である。

4. おわりに

本稿では，音声認識の基礎と音声点訳について報告した。今回の点訳実験では読み上げからの点訳に限ったが，現実の音声認識では，言いよどみを含む非流暢な自由発話を対象にしなければならない。しかし，十分な品質を備えた自由発話の点訳を学習データとして整備することは，コストが高くつく。今後は自己教師あり学習 [12] のような少量のデータからの学習手法を検討したい。また，今回の実験では逐語的な点訳により学習・評価データを揃えたが，情報保障の観点からは当事者が理解しやすい点訳を行う必要がある。当事者へのインタビューなどを通じ，点訳の理解しやすさについても検討を進めたい。

謝辞

本稿で紹介した研究の一部は JSPS 科研費 20H01716 の助成を受けたものである。

参照文献

[1] Kobayashi A, Oonishi J et al.　End-to-End Speech to Braille Translation in Japanese. IEEE International Conference on Consumer Electronics (ICCE). 2022; to appear.
[2] 篠田浩一．機械学習プロフェッショナルシリーズ　音声認識，講談社，2018.
[3] 中川聖一ほか．音声言語処理と自然言語処理（増補），コロナ社，2018.
[4] 久保陽太郎．機械学習による音声認識，コロナ社，2021.
[5] Graves A et al. Sequence Transduction with Recurrent Neural Networks. arXiv :1211.3711 [cs, stat]. 2012.
[6] 全国盲ろう者協会．盲ろう者に関する実態調査報告書，2013.
[7] 全国視覚障害者情報提供施設協会編．点訳の手引き（第4版），2019.
[8] Devi V. A. Conversion of Speech to Braille: Interaction Device for Visual and Hearing Impaired. 4th International Conference on Signal Processing, Communication and Networking (ICSCN). 2017; p. 1-6.
[9] 日本音響学会．日本音響学会 新聞記事読み上げ音声コーパス（JNAS），音声資源コンソーシアム，2006.
[10] Gulati A et al. Conformer: Convolution-Augmented Transformer for Speech Recognition. Interspeech. 2020; p. 5036-5040.
[11] Gillick L, Cox S. J. Some Statistical Issues in The Comparison of Speech Recognition Algorithms. ICASSP. 1989; p. 532-535.
[12] Baevski A et al., wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. NeurIPS, 2020.

End-to-End Automatic Speech Recognition and Its Application to Braille Translation

KOBAYASHI Akio[1)], ONISHI Junji[2)]

[1)]Department of Industrial Information, Faculty of Industrial Technology,
Tsukuba University of Technology
[2)]Department of Computer Science, Faculty of Health Sciences,
Tsukuba University of Technology

Abstract: This paper provides an overview of automatic speech recognition (ASR) technology, from the basics to recent trends such as deep neural networks. We also introduce our recent applied research on direct translation from speech into Braille using end-to-end ASR based on deep learning. In Japan, automatic Braille translation from spoken language is expected to improve information accessibility for deaf-blind people. Japanese Braille has a high affinity for automatic speech recognition because it comprises hiragana characters (kana) that strongly reflect Japanese phonetic features. Therefore, we attempted to use neural networks to translate Japanese Braille directly from speech in an end-to-end manner. We also compared our proposed approach with an existing method that combines ASR and automatic Braille translation.

Keywords: deaf-blind, braille, automatic speech recognition, end-to-end, neural network