End-to-end ASR to jointly predict transcriptions and linguistic annotations

大町 基†

1 はじめに

本稿では、著者の論文 "End-to-end ASR to jointly predict transcriptions and linguistic annotations" (Omachi et al. 2021)（以下、原論文）のテーマの着想から NAACL2021 の採録に至るまでの経緯を述べる。著者は音声処理の研究開発に従事しており、著者よばれものが異なる読者もいるものと思う。そのような読者にとっても、本稿が研究を進めていく上で参考となれば幸いである。

2 論文について

2.1 アイデアの着想

音声認識は、音声から発話内容の書き起こしを推定する技術であり、音声対話システムや音声ユーザーオンタフェース (VUI: Voice user interface) を構築する上では重要な役割を果たす。かつては音響モデルや言語モデルなどの複数のモデルを組み合わせた HMM 音声認識 (Gales and Young 2008) が主流であったが、近年ではそれらのモデルを単一のニューラルネットワークにより実現する End-to-End (E2E) 音声認識の研究が盛んにおこなわれている (Karita et al. 2019b; He et al. 2019)。HMM 音声認識では、多様な声調に対応するためにモデルサイズが大きくなるので計算リソースが豊富な計算機が必要となる (藤田 2019)。一方、E2E 音声認識は、モデル量子化などを応用することで、計算リソースが豊富ではないスマートフォンなどのデバイスでも動作が可能となる (He et al. 2019)。さらに、一部のベンチマークにおいては、E2E 音声認識の性能が HMM 音声認識を上回ることが報告されている (Karita et al. 2019a)。このことから、E2E 音声認識の実用化も進んでおり、今後も研究開発が活発におこなわれていくものと思われる。著者もまた、音声認識の研究開発に従事する一人として E2E 音声認識の实用化を目指し日々奮闘している。

さて、原論文 (Omachi et al. 2021) では、図 1(b) に示したように、E2E 音声認識の枠組みで発話文の表記列に加えて素性（読み、品詞）を同時に推定する方式を提案した。現在提案され
End-to-end ASR to jointly predict transcriptions and linguistic annotations

図 1 提案法のシステム構成。従来法では、E2E 音声認識と単語分割や読み・品詞推定などの自然言語処理を組み合わせているのに対し、提案法は E2E 音声認識と自然言語処理を同時に実現する。

ている多くの E2E 音声認識は発話文の表記列のみを出力するが、素性を同時に出力するという点が原論文のアイデアである。このアイデアは、E2E 音声認識の出力として読みの情報も同時に出力できると有益なことがあるのではないかという疑問から生まれた。例えば、音声検索システムで「日本橋」という駅を調べたいとする。E2E 音声認識では表記しか推定できないので、発話者の意図した入力が「にほんばし」と「にっぽんばし」のどちらの駅を指しているかを特定できない。しかし、E2E 音声認識で読みの情報も同時に得られるならば、発話者が調べたい駅を特定するのも容易になる。

さらに議論を進めていくと、E2E 音声認識で表記・読みだけでなく品詞などの素性も同時に推定するという原論文のアイデアが生まれた。このアイデアは、音声対話システムや VUI などのシステムに E2E 音声認識を応用する際にも有益である。多くのシステムでは、図 1(a) のように、E2E 音声認識により得られる表記列に対して単語分割や読み・品詞推定などの自然言語処理を適用し、意味理解のための素性を抽出する。そのため、意味理解を正確におこなうためにには、E2E 音声認識と後続の自然言語処理の両方の性能も考慮する必要がある。また、スマートフォンなどのデバイスへの応用を考えると、自然言語処理をおこなうための辞書などに必要なメモリやストレージはできる限り削減できることが望ましい。一方、提案法は、単一のニューラルネットワークにより音声認識と単語分割、素性抽出を実現するので、自然言語処理に必要な計算資源を削減できる。また、言語情報だけでなく音響的な情報も利用できるので、言語情報のみでは単語分割や読み・品詞推定が難しい文章（e.g., 同型異音語の読み）に対しても頑健に動作することが期待できる。
2.2 実現方法の検討

アイデアが決まると、次に具体的な実現方法について検討した。具体的な方式を決めるために、(1)E2E 音声認識の枠組みで複数種類の系列を推定する枠組み、(2) その枠組みで用いるニューラルネットワークのモデル構造の 2 つの観点で技術調査を実施した。

E2E 音声認識の枠組みで複数種類の系列を推定する枠組みを調査すると、それは「対数モデル」「conditional chain に基づく対数モデル」「一対一モデル」の 3 種類に分類できることがわかった。対数モデルは図 2(a)に示すように、複数の系列を独立に推定する。モデルの学習は個々の系列が互いに独立であるという条件付き独立性の仮定のもとでおこなわれる。しかし、表記と読みは必ずしも独立ではないため、このモデルは本研究の用途としては望ましくない。また、独立した系列間のアライメントを取る必要があるという点も応用する上では望ましくない。例えば、「X から Y まで」という入力が与えられた際に、X と Y の読みをそれぞれ正しく得るためには、表記と読みのアライメントの正確性が求められる。 conditional chain に基づく対一モデルは、条件付き独立性の仮定が不要だが、図 2(b) に示すように、各系列を別々に推定するため、やはりアライメント処理が必要となる。対一モデルは、図 2(c) に示すように、それぞれの系列の要素をアライメントした系列を推定する。条件付き独立性の仮定に基づかず、アライメント処理が不要という点で本研究の用途としてふさわしいモデルであると考えられる。そこで、対一モデルを用いて複数種類の系列を推定する先行研究を調査した。

一対一モデルを用いて複数種類の系列を推定する研究としては、英単語のスペルと単語そのものを推定するもの (Audhkhasi et al. 2018) や、単語と対応する固有表現を推定するもの (Ghannay et al. 2018) などが提案されており、モデルとしては、connectionist temporal classification (CTC) (Graves et al. 2006) が主に用いられていた。しかし、CTC は出力シンボル間の条件付き独立性の仮定に基づく手法であり、本研究にそのまま応用するのは適切ではない。そこで、本研究では、

\[ X \] (音声) \rightarrow \text{ニューラルネットワーク} \rightarrow \text{明日、は、晴れ (Sequence 1)} \]

(a) 一対多モデル

\[ X \] (音声) \rightarrow \text{ニューラルネットワーク} \rightarrow \text{アシタワハレ (Sequence 2)} \]

(b) conditional chain に基づく一対一モデル

\[ X \] (音声) \rightarrow \text{ニューラルネットワーク} \rightarrow \text{明日、は、晴れ、ハレ (Sequence 1 & 2)} \]

(c) 一対一モデル（提案法）

図 2 複数系列を推定する E2E 音声認識の枠組み
End-to-end ASR to jointly predict transcriptions and linguistic annotations

正解: ピッチ/ピッチ/名詞 ひと/助詞 スペクトラ/スペクトラ/言い易さ スペクトル/スペクトル/名詞 …
従来法: ピッチ/ピッチ/名詞 ひと/助詞 スペクトラ/スペクトラ/UNK/UNK …
提案法: ピッチ/ピッチ/名詞 ひと/助詞 スペクトラ/スペクトラ/言い易さ スペクトル/スペクトル/名詞 …

(a) 言い渋みを含む発話

正解: その/ソノ/連体詞 後/ゴ/名詞 音楽/オンガク/名詞 番組/バンガミ/名詞 が/ガ/助詞 …
従来法: その/ソノ/連体詞 後/アト/名詞 音楽/オンガク/名詞 番組/バンガミ/名詞 が/ガ/助詞 …
提案法: その/ソノ/連体詞 後/ゴ/名詞 音楽/オンガク/名詞 番組/バンガミ/名詞 が/ガ/助詞 …

(b) 同型異音を含む発話

図 3 提案法による表記・読み・品詞の同時推定結果の例。X/Y/Z のうち、X, Y, および Z はそれぞれ
表記、読み、および品詞を表す。音声認識と自然言語処理に基づく単語分割と読み・品詞推定を組
み合わせた従来法では正しく推定できない発話が、提案法では正しく推定できている。

究では出力シンボル間の依存関係を考慮した Transformer (Vaswani et al. 2017) とその拡張
である Transformer-CTC (Karita et al. 2019b) を用いて表記と素性を推定することを試みた。
Transformer と Transformer-CTC は E2E 音声認識で広く用いられているモデルである。

2.3 評価実験

提案法を実装し、有効性を確認するために音声認識および素性抽出の観点で評価実験をおこ
なった。紙面の都合により、評価実験の条件や結果の詳細は本稿では割愛するが、Transformer-
CTC モデルを利用した場合に、音声認識の性能劣化が少なく、十分な性能で素性抽出が可能で
あることを示している。実験の詳細については、原論文 (Omachi et al. 2021) を参照していただ
きたい。図 3 は提案法を用いて表記・読み・品詞を同時に推定した例を示している。比較とし
て、図 1(a) のように、E2E 音声認識の後段に単語分割および読み・品詞推定を適用した手法も
示している。言い渋みを含むような発話において、提案法は音声に含まれる無音区間を観察に
単語分割を正確に行うことができ、読みや品詞も正しく推定できる。また、同型異音を含む
場合においても、提案法は音の情報から正しい読みを確定できている。提案法は言語情報だけ
でなく音響的な情報を利用することができるので、言語情報のみに基づく方法では解くことが
難しい条件に対しても頑健に動作するという著者らの期待通りの結果が確認できた。

3 NAACL2021 の採録に至るまで

NAACL2021 では、採択結果が出される前に Reviewer のコメントに対して意見を返すことが
できる Rebuttal 期間が設けられていた。Reviewer からは、提案法の有効性が指摘された。投稿
した時点の論文では、音声認識の性能改善が目的ではなく、枠組みを提案するものであるという
点を強調していたが、研究の意義をうまく伝えることができなかった。そこで、Reviewer
4 おわりに

本稿では NAACL に採録された著者の論文の着想から採録に至るまでの経緯を概説した。本研究は、音声認識と自然言語処理を組み合わせたものであり、今後もさらなる応用が期待できる。音声認識の研究開発の従事するひとりとして、自然言語処理分野の研究者の皆様と手を取りあいながらより良いサービスを社会に提供できるように努めていきたい。

謝 辞

本研究にあたり、数多くのご助言をいただきましたヤフー株式会社 藤田悠哉氏、ジョンズ・ホプキンズ大学 渡部晋治教授（現在、カーネギー・メロン大学）、同大学 Matthew Wiesner 氏に感謝の意を表します。

参考文献

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 Proc. ICASSP, pp. 4759–4763.
藤田悠哉 (2019). Web 検索等に利用される分散型音声認識システムへのディープラーニングの実装. デジタルプラクティス, 10 (2), pp. 322–336. [Y. Fujita (2019). Web Kensaku Tou ni Riyo Sareru Bunsangata Onsei Ninshiki Sisutemu eno Dipu Raningu no Jisso. Disital Practice, 10 (2), pp. 322–336.].
Gales, M. and Young, S. (2008). Application of Hidden Markov Models in Speech Recognition. Now Foundations and Trends.
Ghannay, S., Caubrière, A., Estève, Y., Camelin, N., Simonnet, E., Laurent, A., and Morin, E. (2018). “End-To-End Named Entity And Semantic Concept Extraction From Speech.” In Proc. SLT, pp. 692–699.
Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). “Connectionist Temporal
End-to-end ASR to jointly predict transcriptions and linguistic annotations

Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks.” In *Proc. ICML*, pp. 369–376.

He, Y., Sainath, T. N., Prabhavalkar, R., McGraw, I., Alvarez, R., Zhao, D., Rybach, D., Kannan, A., Wu, Y., Pang, R., Liang, Q., Bhatia, D., Shangguan, Y., Li, B., Pundak, G., Sim, K. C., Bagby, T., Chang, S.-y., Rao, K., and Gruenstein, A. (2019). “Streaming End-to-end Speech Recognition for Mobile Devices.” In *Proc. ICASSP*, pp. 6381–6385.

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N. E. Y., Yamamoto, R., Wang, X., Watanabe, S., Yoshimura, T., and Zhang, W. (2019a). “A Comparative Study on Transformer vs RNN in Speech Applications.” In *Proc. ASRU*, pp. 449–456.

Karita, S., Soplin, N. E. Y., Watanabe, S., Delcroix, M., Ogawa, A., and Nakatani, T. (2019b). “Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration.” In *Proc. Interspeech*, pp. 1408–1412.

Omachi, M., Fujita, Y., Watanabe, S., and Wiesner, M. (2021). “End-to-end ASR to jointly predict transcriptions and linguistic annotations.” In *Proc. NAACL*, pp. 1861–1871, Online. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.-N., Kaiser, L., and Polosukhin, I. (2017). “Attention is all you need.” In *NIPS*, pp. 5998–6008.

略歴

大町 基：2009年早稲田大学理工学部コンピュータ・ネットワーク工学科
卒業。2011年同大学大学院基幹理工学研究科情報理工学専攻修士課程修了。
2017年同大学大学院基幹理工学研究科情報理工学専攻博士後期課程修了。博士
（工学）。現在、ヤフー株式会社にて音声認識の研究開発に従事。