連載

5分で分かる!? 有名論文ナナメ読み

増村　亮（NTT メディアインテリジェンス研究所）

Listen, Attend and Spell : A Neural Network for Large Vocabulary Conversational Speech Recognition

ICASSP (2016)

音声認識の方式を一新？

音声認識は，入力された音声をテキストに自動変換する技術であり，機械が音声言語を理解するために必須の技術である．音声は，同じ言葉（テキスト）でも，発話する人によって，そして環境などによってまったく異なるものになり，簡単な人手によるルールでこの自動変換をモデル化することは非常に難しいと感じていただけるだろう．そこで，機械学習を駆使すべきタスクとして，これまでさまざまな音声認識の方式が検討されてきており，2020 年現在で実用化されている主流の方式は，想像以上に長年のノウハウが詰まったものになっている．すなわち，音声認識という技術は，機械学習を用いれば簡単に実現できるというものではなかったというのがこれまでである．さて，本稿で紹介する論文は，長年のノウハウが詰まった音声認識方式を最新の機械学習の力で一新してしまうという流れの先駆けとなった論文である．以下では，「ノウハウが詰まった Hybrid 方式」と，この論文が提案する「独立仮定なしの End-to-End 方式」の違いを中心に紹介する．なお，類似した方式は同時期に複数発表されてきており，本論文はその代表的な 1 つであることに注意されたい．

ノウハウが詰まった Hybrid 方式

2020 年現在，世の中で使われる音声認識システムは，Hybrid 方式と呼ばれるモデル化を広く採用している．Hybrid 方式では，音声からテキストへの自動変換問題をいくつかのモデルの組合せにより実現している．具体的には，音声から音素へのマッピングをモデル化する音響モデル，音素と単語のマッピングをモデル化する発音モデル，そして単語間のつながりをモデル化する言語モデルの 3 つのモデルから構成され，この構成自体に長年のノウハウが詰まっている．さらに 2010 年頃から，個々のモデル化について深層学習の導入が進み，大きく性能改善されたことが音声認識の大きなパラダイムシフトとなり，Hybrid 型はすでに 1 つの完成形の方式となっている．

そんな Hybrid 方式であるが，いくつかの技術課題があることが知られている．1 点目は，音声認識という問題に対して全体最適化されていないということが挙げられる．すなわち，個々に最適化された 3 つのモデルの組合せは，音声認識の問題に最適であるとは限らないという点である．2 点目は，個々のモデル化で，さまざまな独立過程をおいている点である．たとえば音響モデルで用いる隠れマルコフモデルでは，「現在の音素状態は 1 フレーム前の音素状態にのみ依存する」という独立仮定，言語モデルで用いる n-gram モデルでは，「現在の単語は前の数単語にのみ依存する」という独立仮定をおいている．このような独立仮定はデコーディングの仕組みと相性が良いことから，これまで引き継がれてきたが，それが性能限界を与えていると考えられる．

独立仮定なしの End-to-End 方式

この論文では，前述の 2 点の課題を解決するために，Listen, Attend and Spell（LAS）と呼ばれ

る End-to-End 方式の手法を提案している．最大の特徴は，音響モデルも言語モデルも発音モデルも使うことなく，音声からテキストへの自動変換問題を LAS という 1 つのモデルで表現している点である（図 -1 参照）．そして，その 1 つのモデル内において，独立仮定を一切おくことなく全体をモデル化している点も大きな特徴である．

さて，詳細なモデル化の話に入っていこう．LAS では条件付き自己回帰生成モデルにより，音声（音響特徴量系列）からテキスト（単語系列）への変換を独立仮定なしに End-to-End 方式でモデル化している．具体的には，Listener と呼ばれるリカレントニューラルネットワークに基づく音声エンコーダと，Speller と呼ばれるリカレントニューラルネットワークに基づくテキストデコーダを結合することで，1 つのネットワーク構造により条件付き自己回帰生成モデルを構成している．そして，LAS の面白い点は，Listener と Speller を結びつける Attention の存在である．この Attention の処理は注意機構と呼ばれ，音声の各フレームとテキストの各単語の対応を自動で考慮できる．

モデル全体は，音声とテキストのペアデータから直接最適化でき，具体的には，音声が与えられた際のテキストの条件付き確率の最大化，勾配降下法での学習においては負の対数条件付き確率の最小化によりモデルパラメータを最適化している．一方，音声が与えられたときに学習済みのモデルパラメータを用いてテキストを推定する際は，単語単位の Left-to-Right ビーム探索に基づき，終端記号が出現するまで再帰的に単語を生成することで可変長のテキストを推定する．

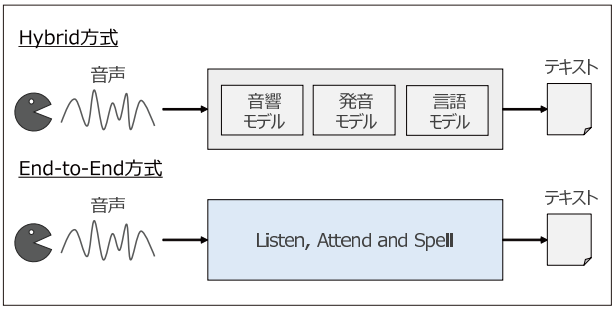

図 -1　Hybrid 方式と End-to-End 方式

最終的に，最新の Hybrid 方式の音声認識システムの性能には届かないものの，近いところまで End-to-End 方式により到達できることをこの論文では報告している．Hybrid 方式の課題を解決しているのだから，性能改善できるのではないかと思われるかもしれないが，理論上は課題を解決できていても，最適化の難しさなどからこの論文の中ではそこまで至らなかったと解釈していただくのがよいだろう．

その後の展開と現状

この論文をスタートとして，独立仮定なしの End-to-End 方式の音声認識技術が大きく進展し，今日では End-to-End 方式が Hybrid 方式を上回る性能を達成してきている．独立仮定なしの End-to-End 方式は最強の方式とも感じられてしまうが，実はまだまだ課題は残っている．現状はモデルへの入力の時点で，すでにヒューリスティックに抽出された音響特徴量系列を前提としているが，音声の波形情報からダイレクトにモデル化する方式の検討がより必要と考えられている．また，現状は 1 つの発話内に閉じて独立性仮定なしに End-to-End でモデル化することがほとんどだが，発話境界を超えてこれまで何を話してきたか，どんなシチュエーション（会話相手，場所など）で発話しているか，などまで含めたモデル化の検討は不十分である．このような未解決の課題に対して，今後どのような技術的ブレークスルーが音声認識に対して生み出されていくのか，期待は膨らむばかりである．

（2021 年 1 月 4 日受付）

増村　亮（正会員）
ryou.masumura.ba@hco.ntt.co.jp
2011 年東北大学大学院工学研究科修士課程，2016 年博士課程修了．博士（工学）．2011 年日本電信電話（株）入社．現在，NTT メディアインテリジェンス研究所特別研究員．音声認識を中心に，音声音響処理，自然言語処理，画像映像処理等のメディア処理全般の研究開発に従事．