Vol.2019-AAC-9 No.34
2019/3/10

音声認識を用いた字幕作成システムの改良

秋田 祐哉[1]　上乃 聖[2]　三村 正人[2]　河原 達也[2]

概要：我々は，効率的な字幕付与を実現するために，音声認識を用いたオンラインの自動字幕作成システムを運用している．本システムの特徴の 1 つに，より精度の高い音声認識のために，サーバ上で認識用のモデルを対象の音声に適応させて認識を実施できることがある．これまでの本システムの運用において，適応したモデルをユーザ端末上の音声認識でも利用できるようにする機能の要望があった．また，本システムの運用を行っている間に音声認識の技術はいっそう進展し，最近では全てのモデルを 1 つのニューラルネットワークに統合した，End-to-End 型音声認識が有望となってきている．本稿では，我々の字幕作成システムについて，これらの実装を行って利便性や性能を改善したので報告する．

Improvement of Automatic Captioning System using Speech Recognition Technology

Yuya Akita[1]　Sei Ueno[2]　Masato Mimura[2]　Tatsuya Kawahara[2]

Abstract: We have been running an online captioning system with the automatic speech recognition (ASR) technology for efficient captioning of speech materials. One of the key characteristics in the system is the capability of automatic adaptation of ASR models to the target speech. During operation of the system so far, we received requests from users to provide adapted models for ASR-based captioning on user PC. In the meanwhile, the ASR technology has significantly advanced in recent years, and the state-of-the-art End-to-End framework, where all ASR models are integrated into a single neural network, is considered promising. In this report, we will describe the implementation of these functionalities to the system.

1. はじめに

講義や講演などの音声を，事後に，あるいはリアルタイムに書き起こすことは，コストの大きな作業である．これは，人手による書き起こしがそもそも時間を要する作業であることに加えて，講義・講演のような専門性の高い内容では，話題を理解できる作業者でないと正確な聞き取り・書き起こしが難しいことによる．このような熟練した作業者の数が限られていることが，字幕付きのコンテンツを普及させ，またリアルタイムの字幕（文字通訳）をより多くの場面で提供するための障害となっている．

これに対して，音声認識技術を用いて自動的に音声を書き起こして字幕や文字通訳に利用する取り組みが進んできている．音声認識を用いることで，短時間に音声を全て書き起こすことができる．音声認識を用いた自動的な字幕作成は，たとえば YouTube[*1] で行われており，アップロードされた映像にサーバ側で字幕を付与させることができる．また，音声認識による文字通訳としては，UD トーク [*2] やこえとら [*3] などのアプリケーションが用いられている．近年では音声認識を必ずしもその場の端末で行う必要はなく，ネットワークを通じた，いわゆるクラウドで行うことも一般的であり，たとえば Google[*4] や Microsoft[*5] などがアプリケーション開発者向けのサービスを提供している．

音声認識では，特殊な表現や専門用語などの認識を実現するためには，あらかじめ認識のモデル（言語モデル）を対

[1] 京都大学 大学院経済学研究科
Graduate School of Economics, Kyoto University
[2] 京都大学 大学院情報学研究科
Graduate School of Informatics, Kyoto University

[*1] https://www.youtube.com/
[*2] http://udtalk.jp/
[*3] http://www.koetra.jp/
[*4] https://cloud.google.com/speech-to-text/?hl=ja
[*5] https://azure.microsoft.com/ja-jp/services/cognitive-services/speech-to-text/

象の音声に適応させておかなければならない．そこで我々は，認識対象に合わせてモデルをカスタマイズした上で音声認識を実行できる，字幕の自動作成システムを開発して運用してきた [1]．我々のシステムは，専門家・技術者でない利用者でも字幕の作成が行えるよう，アップロードされた音声と関連テキストから自動的に音声認識をセットアップ・実行して字幕を作成するサーバシステムである [*6]．また，このシステムを活用して，サーバ側でリアルタイムに音声認識を実行して文字通訳に利用する枠組みも開発した．

システムの公開・運用を始めてから 5 年あまりが経過し，この間に本システムはいくつかのプロジェクトで使用されてきた [2], [3] [*7]．また，リアルタイム字幕についても，実際にいくつもの学会・シンポジウムの講演で提供を行ってきた [4]．この間に寄せられたユーザの要望として，システムで適応させた言語モデルをユーザ側の端末で使用したいというものがあった．これに対して言語モデル提供機能・辞書編集ツールを開発して提供を開始したので，本稿で報告する．

また，システムの運用開始以降，音声認識は深層学習の技術を用いて大きく進展した．本システムでも従来から一部に深層学習（ディープニューラルネットワーク）のモデルが含まれているが，音声認識のモデルをすべてニューラルネットワークに統合した，いわゆる End-to-End 型の音声認識の研究が現在盛んに進められている．本システムでも既存の枠組みに加えて End-to-End 型音声認識を搭載したので，あわせて報告する．

2. 字幕自動作成システム

2.1 事後的な字幕の作成

まず，本システムの従来からの構成について説明する．本システムでは，ユーザにより収録された講義・講演や討論などの音声・映像に対して，事後的に字幕を付与することを想定している．**図 1** にシステムの利用の流れを示す．まず，ユーザがこれらのコンテンツを字幕サーバにアップロードする．音声・映像に加えて，言語モデルを話題に適応させるために，コンテンツの話題と関連するテキスト（たとえば講演予稿やスライド）もアップロードすることができる．字幕サーバではコンテンツからの音声の抽出および検査が行われ，ユーザの指定や関連テキストに応じて自動的に音声認識システムが構成された上で認識処理が実行される．音声と同期した字幕ファイルがサーバ上に出力されるとともに，これらにアクセスするためのアドレスがユーザに通知される．

従来型の音声認識は，音響モデル・言語モデル・単語辞書（発音辞書）と，これらを用いて実際に認識を行う音声認識エンジンからなる．本システムでは，講演や討論など，

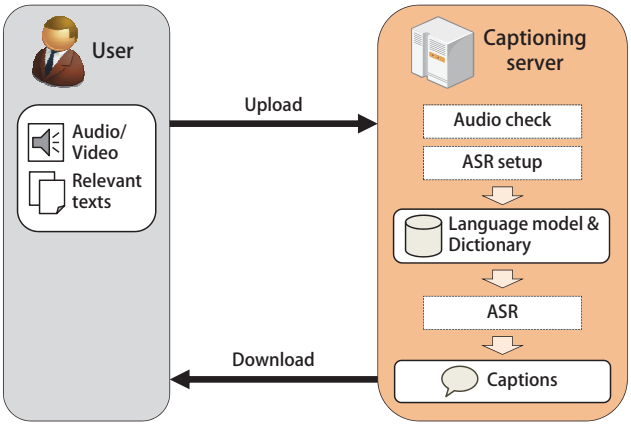

図 1 システムの利用の流れ

表 1 モデルのプロファイルの一覧

名称	講演	スピーチ	討論
学習データ	CSJ（学会講演）	CSJ（模擬講演）	国会音声
音響モデル	DNN-HMM		
言語モデル	単語 Trigram		

コンテンツの種類に応じていくつかの音響モデル・言語モデルの組み合わせをプロファイルとして用意しており，ユーザは実際に音声認識に利用するプロファイルをコンテンツのアップロードの際に選択することができる．現時点でのプロファイルの一覧を**表 1** に示す．たとえば，講演には『日本語話し言葉コーパス』（CSJ）の学会講演データから学習したモデル [5], [6] を，討論には国会音声・会議録から学習したモデル [7] を用意している．音響モデルは，入力音声（特徴量）に対する音素単位の状態遷移を隠れマルコフモデル（HMM）で表現し，HMM の各状態における出力確率を従来の統計モデルではなくディープニューラルネットワーク（DNN）で求める，いわゆるハイブリッド型の DNN-HMM モデルである．言語モデルは，前 2 単語を条件として次の単語の発生確率を定めて単語の予測に使用する，単語 Trigram（3-gram）モデルである．本システムでは音声認識エンジンとして Julius[*8] を使用している．

音声認識結果には各単語の推定時刻が付与されているので，これを表示のタイミングとして，認識文からなる字幕ファイルを作成する．この際，音声認識結果には文や節の境界は与えられていないため，句読点の自動推定 [8] を用いて字幕の行に分割する．また，フィラーや口語表現，文末表現などの冗長部分を削除・修正するため，自動整形手法を適用する [9]．

2.2 リアルタイムの字幕作成

前節で述べた字幕作成システムは，収録された音声に対する事後的な処理である．一方，このシステムを応用して，講義・講演の会場で情報保障のためにリアルタイムに字幕

[*6] http://caption.ist.i.kyoto-u.ac.jp/
[*7] http://gclip1.grips.ac.jp/video/
[*8] https://julius.osdn.jp/

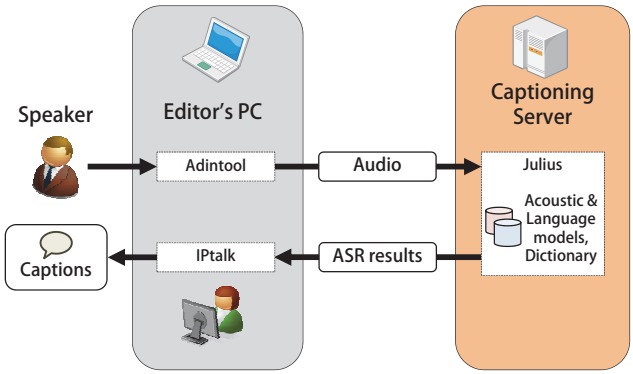

図 2　リアルタイム字幕作成システムの構成

を作成・表示するシステムも構築した．このシステムの構成を図 2 に示す．本システムは，講義・講演会場で作業者が入出力・編集に使用する PC と，音声認識を行うサーバから構成される．講師の音声は PC に入力され，Julius 付属の音声入力ツールである Adintool によって発話検出・セグメンテーションを行ったのち，サーバ側の Julius にネットワーク経由で送信される．サーバではあらかじめ対象の講義・講演用に構成された音響モデル・言語モデルを使用して音声認識を行い，この結果をネットワーク経由で作業者の PC に送信する．なお，ここでは音響モデル・言語モデルとして表 1 に挙げた講演プロファイルに相当するものを主に使用している．GPU を用いてデコードすることにより，DNN-HMM 音響モデルでも実時間以下の処理時間で音声認識を行っている．作業者の PC では，PC 要約筆記で一般的に用いられている IPtalk[*9] を字幕の編集・表示ツールとして使用する．

3. 適応言語モデルの提供

本システムはサーバ側で認識処理を行うよう構成されている．しかし，実際の利用場面，特に文字通訳の場合は，通信環境の問題やプライバシー・機密の保護の問題から，このようなオンライン（クラウド）のサービスを利用できないことがある．この場合はユーザ側の PC で音声認識を実行することとなるが，この際に認識対象に適応したモデルを用意することは，一般のユーザには容易ではない．1 節で述べたアプリケーションやクラウドサービスでも，ユーザが専門用語などを音声認識の単語辞書に登録したり，あるいは言語モデルの適応を行えるものもあるが，ユーザ PC のみによる実行時はこれらを利用できず，またユーザがモデルを入手して使用することもできない．そこで，このようなニーズのために，本システムでは Julius 向けの適応言語モデル・単語辞書を作成して提供する機能を公開する．

[*9] http://www.s-kurita.net/

3.1 リアルタイム字幕作成における適応モデルの利用

前節で述べたリアルタイムの字幕作成では，音声認識を作業者の PC で実行することで（すなわち，図 2 における字幕サーバを PC で代替して）作業者の PC のみで作成処理を行うことができる．実際に，IPtalk では Julius との連携機能が実装されており，PC に Julius の認識キット [*10] をインストールすることで，PC のみで音声認識を用いた字幕作成が可能である．

この際に使用する音声認識のモデルは CSJ を用いて学習されており，学会講演データを用いて講演向けに構成された「講演音声モデルキット」と，模擬講演データを用いて一般の話し言葉向けに構成された「話し言葉モデルキット」の 2 つが公開されている．ただし，これらは実際の認識対象には適応されていない標準的なモデルである．本システムにより適応されたモデルをダウンロードして差し替えることで，作業者 PC 上での音声認識を適応モデルにより実施できる．

本システムで適応言語モデルを得る方法は，前節で述べた音声コンテンツに対する字幕作成の要領と同一であり，音声認識と字幕作成を省略して，代わりに言語モデルを提供するものである．

3.2 適応の原理

ここでは，この適応処理がどのように行われているかを述べる．本システムでは単語 Trigram 言語モデルを使用している．これは，音声認識における言語制約を 3 単語の連鎖確率で規定するもので，この確率は言語モデルの学習テキストにおける単語列 $\{w_i\}$ の統計頻度 C から定められる．

$$P(w_3|w_1w_2) = \frac{C(w_1w_2w_3)}{C(w_1w_2)} \qquad (1)$$

ただし，学習テキストに出現しない文脈（単語履歴，(1) 式における w_1w_2）に対処するため，実際にはこの確率が平滑化（スムージング）されて用いられる．

標準の言語モデルに含まれない単語を認識したい場合，最も簡便な方法は辞書への単語登録である．ただし，これはあらかじめ特殊な単語（未知語や UNK とよばれる）を含めて言語モデルを学習しておき，後から辞書に追加された単語はすべてこの単語と見なして単語の確率を適用するものである．したがって，追加された単語は，その意味や役割にかかわらず言語モデル内の文脈が共通であり，必ずしも適切な確率を得ることができない．

これに対して，言語モデル適応は文脈も含めて確率を更新するものである．本システムでは，関連テキストが与えられている場合は，選択されたプロファイルの言語モデルに対してテキスト混合に基づく適応が行われる．すなわち，ベースとなるモデルと与えられた文書によるモデルの線形

[*10] https://julius.osdn.jp/index.php?q=dictation-kit.html

補間を，学習テキストの重み付き混合で行う．ベースモデルの学習テキストにおける頻度を C_b，適応用に与えられたテキストにおける頻度を C_t とすると，最終的な頻度 C は

$$C(w_1w_2w_3) = \lambda C_b(w_1w_2w_3) + (1-\lambda)C_t(w_1w_2w_3) \quad (2)$$

で与えられ，これを用いて (1) 式から適応言語モデルを学習している．ここで混合重み λ $(0 < \lambda < 1)$ を定める必要があるが，任意のテキストに対して最適な値の推定は難しい．本システムでは，適応用のテキストの総単語数がベースモデルの学習テキストに対して一定の程度になるよう頻度のスケーリングを行っている．

本システムでは，与えられたテキストに対してベースのモデルの学習テキストとまったく同一の前処理（形態素解析や補正処理）を行って単語列に変換した上で適応処理を行う．適応用の文書に出現する単語は，原則として全て単語辞書に登録される．これにより，新たに加える単語の出現確率が文脈に応じて変わるため，より適切な言語制約となることが期待される．なお，本システムでは，単語辞書への登録や修正のために，辞書の編集ツールもあわせて提供する．

4. End-to-End 型音声認識

4.1 End-to-End 型のあらまし

End-to-End 型の音声認識 [10], [11], [12], [13], [14] は，従来の音声認識で個別に構築されていたモデルを 1 つのニューラルネットワークに統合して学習・認識するものである．**図 3** に従来の音声認識の枠組みと End-to-End 型の枠組みを示す．従来の音声認識では，音響モデル・言語モデルおよび単語辞書をもとにデコーダが入力音声に対して多数の仮説を展開して探索するため必然的に時間がかかり，リアルタイムに認識するためには種々の高速化の工夫，また探索空間の制限のような性能低下につながる対応が必要である．これに対して End-to-End 型の音声認識では，入力音声をディープニューラルネットワークに投入して，ネットワークの計算（推論）を行うだけで認識結果が出力されるため，実行の制御はシンプルであり，高速に認識を実行できる．これらの利点を考慮して，本システムでも End-to-End 型の音声認識を新たに搭載した．

ただし，現在のところ End-to-End 型のモデルを効率的に適応する手法は確立されていない．従来の方式のように単語辞書が単独の構成要素としてあるわけではないので，単語登録は容易ではない．また，単語 N-gram モデルのように単語列とその確率が明示的に与えられておらず，N-gram モデルで特定の確率を操作して適応を実現するようなことは，ニューラルネットワークでは難しい．これらは，原理的にはネットワークの再構成や再学習を必要とし，

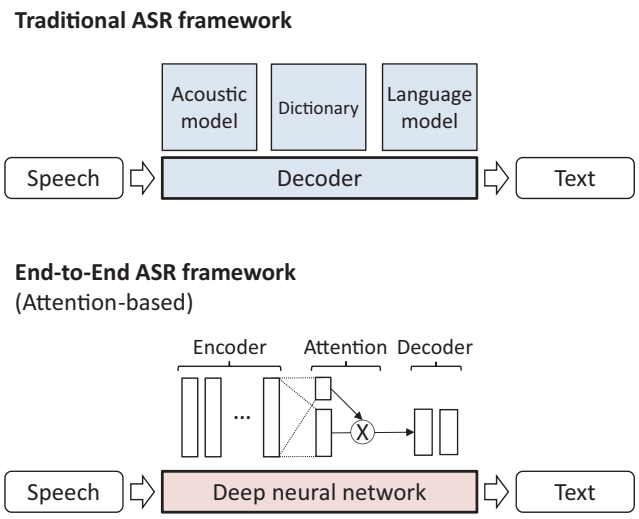

図 3 従来の枠組みと End-to-End の枠組み

その計算コストも大きい．したがって現時点で本システムでは End-to-End 型については適応が無効となっていて，将来の対応を検討している．

4.2 仕様と評価

本システムでは，図 3 にあるような，注意（attention）機構を用いた単語単位の End-to-End 型音声認識を実装した [15]．ただし，表 1 にある従来型のプロファイルとは異なり，現時点では CSJ の学会講演と摸擬講演から学習した 1 つのモデルのみである．

入力特徴量は，通常の音声認識が MFCC（メル周波数ケプストラム）であるのに対して LMFB（対数メルスケールフィルタバンク）であり，もとの周波数スペクトラムに近い特徴量である．エンコーダ部分は 5 層の双方向 LSTM で，各層は 320 ユニットから構成されている．一方，注意機構（デコーダ）部分は 1 層の単方向 LSTM で，こちらも 320 ユニットである．これらの実装は PyTorch による．

本稿では，End-to-End 型音声認識の速度を測定するため，実際の講演音声を用いて認識を行った．使用したデータは京都大学で開催されたシンポジウムで収録した音声で，講演者 2 名・計 62.6 分（実発話時間 47.9 分) である．従来の音声認識では 12.5 分を要したのに対して，End-to-End 型音声認識では 4.5 分であった．音声は一定以上の無音で区切って認識したため，合計で 768 区間が入力され，平均の長さは 3.7 秒であるが，End-to-End 型ではこれらを平均 0.35 秒で認識しており，高速に認識できているといえる．なお，ここで使用した計算機は，CPU が Xeon E-2124 (定格 3.3GHz)，メインメモリが 32GB，GPU が Nvidia Quadro P6000（メモリ 24GB）である [*11]．

[*11] ただし実際の認識の実行にはこれほどのメモリは使用しない．

5. おわりに

本稿では，我々が運用しているオンラインの字幕作成システムにて行った改善について報告した．ユーザ端末でもより精度の高い音声認識を実行するために，適応言語モデルの提供を開始し，また最新の音声認識技術である End-to-End 型システムを実装した．今後も本システムの運用を通じて，システムの改善に努めていきたい．

謝辞 IPtalk 開発者の日本遠隔コミュニケーション支援協会 栗田茂明氏，Julius 開発者の名古屋工業大学 李晃伸先生に深く感謝申し上げます．本研究の一部は科学研究費補助金（課題番号 16H02847）および学術研究助成基金助成金（課題番号 18K11354）によって行われた．

参考文献

[1] 秋田祐哉，三村正人，河原達也：音声認識を用いた講義・講演の字幕作成・編集システム，情報処理学会研究報告，2015-SLP-108-2 (2015).
[2] 石本祐一：方言音声に対するテキスト自動アライメントの試み，国立国語研究所言語資源活用ワークショップ発表論文集，P-4-08 (2018).
[3] 鈴木泰山，内山雄司，青木保一，相良毅，秋田祐哉，河原達也，竹田香織，増山幹高：音声認識技術の活用による国会審議映像検索システムの実現，情報処理学会研究報告，2014-SLP-103-6 (2014).
[4] 平賀瑠美，秋田祐哉：音声自動認識による字幕情報保障トライアル，情報処理学会研究報告，2016-AAC-1-6 (2016).
[5] 三村正人，河原達也：話し言葉音声認識タスクにおける音素誤り最小化学習（MPE）の効果，日本音響学会秋季研究発表会講演論文集，3-Q-8 (2007).
[6] 三村正人，河原達也：CSJ を用いた日本語講演音声認識用 DNN-HMM の構築，日本音響学会秋季研究発表会講演論文集，1-P-42b (2013).
[7] 秋田祐哉，三村正人，河原達也：会議録作成支援のための国会審議の音声認識システム，電子情報学会論文誌，Vol. J93-D, No. 9, pp. 1736–1744 (2010).
[8] 秋田祐哉，河原達也：講演に対する読点の複数アノテーションに基づく自動挿入，情報処理学会論文誌，Vol. 54, No. 2, pp. 463–470 (2013).
[9] Neubig, G., Akita, Y., Mori, S. and Kawahara, T.: A monotonic statistical machine translation approach to speaking style transformation, *Computer Speech and Language*, Vol. 26, No. 5, pp. 349–370 (2012).
[10] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y.: Attention-based models for speech recognition, *Proc. Advances in Neural Information Processing Systems*, pp. 577–585 (2015).
[11] Chan, W., Jaitly, N., Le, Q. and Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, *Proc. ICASSP*, pp. 4960–4964 (2016).
[12] Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P. and Bengio, Y.: End-to-End attention-based large vocabulary speech recognition, *Proc. ICASSP*, pp. 4945–4949 (2016).
[13] Audhkhasi, K., Ramabhadran, B., Saon, G., Picheny, M. and Nahamoo, D.: Direct acoustics-to-word models for English conversational speech recognition, *Proc. Interspeech*, pp. 959–963 (2017).
[14] Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A. and Ochiai, T.: ESPnet: End-to-End speech processing toolkit, *Proc. Interspeech*, pp. 2207–2211 (2018).
[15] Ueno, S., Inaguma, H., Mimura, M. and Kawahara, T.: Acoustic-to-word attention-based model complemented with character-level CTC-based model, *Proc. ICASSP*, pp. 5804–5808 (2018).