4Q-04

音声中の検索語検出における状態数の異なる複数の深層学習モデルを用いた検索精度の向上

西野将弘† 小嶋和徳† 李時旭‡ 伊藤慶明†
岩手県立大学† 産業技術総合研究所‡

1. はじめに

近年，音声中の検索語検出(STD：Spoken Term Detection)の研究が盛んになっている[1-2].

SQ-STD(SQ : Spoken Query)は，STD に音声クエリを与えるもので，この音声クエリと検索対象の音声データを音声認識システムでテキスト化し，サブワード等の細かな単位にした上で，テキストレベルで CDP(Continuous Dynamic Programming)を用い照合し検索を行う方式が代表的である．最近の音声認識システムで用いられる DNN(Deep Neural Network)を用い，その出力の事後確率ベクトルを用いてフレームレベルで照合を行うことで，高い検索精度が得られた[3]. この DNN では音声データのフレーム毎の音声特徴量を入力すると，一般に triphone HMM の各状態の事後確率が得られる．この状態数分のベクトルを事後確率ベクトルと呼び，この事後確率ベクトルをフレーム順に並べたものを Posteriorgram と呼ぶ．音声クエリと音声データの Posteriorgram 同士を用いて CDP 照合を行い検索を行う．CDP 照合時の局所距離は，音声クエリと音声データの 2 つの事後確率ベクトルの内積計算を行い，その負の対数をとって求める．Posteriorgram 照合は高い検索精度が得られるが，事後確率同士の約 3000 次元の内積計算がフレーム毎に必要となり検索時間を要す．これに対し，音声クエリ/音声データの最尤系列化[4-5]では，音声クエリあるいは音声データのいずれか一方の Posteriorgram を約 3000 次元から最尤の状態番号系列の 1 次元に次元圧縮し，他方の事後確率値を利用することで内積計算を省略し検索時間を削減した．一方，情報量が減少するため，検索精度の低下が問題になった.

先行研究[6]では，DNN，BLSTM(Bidirectional Long Short Term Memory)，CTC(Connectionist Temporal Classification)の複数の深層学習モデルを音声データ最尤系列化方式に適用した．モデル毎に検索結果が得られ，これらのスコアを統合することで情報を補完し検索精度の向上を実現したが，Posteriorgram 照合の精度に達しなかった.

そこで本研究では，DNN の出力に対応する事後確率が monophone，triphone，単語と異なったものに対応した複数の深層学習モデル・機構 (以降，モデルとする) を用い，後述するように異種のサブワードから得られる異種の情報を用いることで[7]情報を補完し，検索精度の向上を図る.

2. 提案手法

本研究では，DNN，BLSTM，CTC，Hybrid CTC/Attention[8]を用いる．Hybrid CTC/Attention の学習は ESPnet[9]を用いた．フレーム毎に得られる事後確率の状態数はモデルにより異なり，DNN・BLSTM は triphone と対応し，状態共有を行うことで 3,009 状態．CTC は monophone と対応し，HMM の 1 モデルを 1 状態として，41 個の音素と blank ラベルを合わせた 42 次元．Hybrid CTC/Attention は単語に対応し，学習データの書き起こし文の単語に対応した 3212 次元となる.

上記の通り，状態数が異なる異種のモデルで検索を行うことで入力特徴量に対し得られる特徴が増えるため，結果を統合する際に効果的な情報の補完が可能になると考える[7].

なお，CTC の事後確率は 42 次元と小さく，必要メモリ容量と検索精度の観点から音声クエリ最尤系列化方式を用いた．CTC は特性上，最尤系列に状態番号 0(blank ラベル)が連続する場合が多い．検索時間削減のため，音声クエリの最尤系列は状態番号 0(blank ラベル)が連続した時 1 つに圧縮する．圧縮による検索精度の低下は 0.31pt と小さかった.

1 つのモデルで照合すると，各発話に対してスコアが得られるため（NTCIR の評価方式），各発話に対してモデルごとに照合スコアが得られる．これらのスコアを線形和することで統合し統合距離を求める．ある発話に対して 2 つのモデルから得られた照合スコアをD_1，D_2とし，式(1)により新たなスコアD_{new}を求める．線形和の統合割合は重みαで表し，$0 \leq \alpha \leq 1$ とする.

$$D_{new} = \alpha D_1 + (1-\alpha) D_2 \qquad (1)$$

Improving Retrieval Accuracy in Query-by-Example by Using Multiple Deep Learning Models with Different Number of States.
†Nishino Masahiro, †Kojima Kazunori, ‡Lee Shi-wook, and †Itoh Yoshiaki,
†Iwate Prefectural Univercity, ‡AIST

3. 評価実験

3.1. 実験条件

学習に用いる音声は最も良い検索精度が得られたデータ量とし，DNN，Hybrid CTC/Attention は CSJ 2,525 講演(約 560 時間)．BLSTM，CTC は CSJ 2,702 講演(約 600 時間)とした．

入力特徴量は，ESPnet を用いて学習した Hybrid CTC/Attention はフィルタバンク 80 次元にピッチを追加した 83 次元．それ以外はフィルタバンク 120 次元とし，前後 5 フレームを追加した 1320 次元とした．検索情報の測定には，CPU に Intel Core i7-6700K, GPU に NVIDIA GeForce GTX 1080, RAM 16GB を搭載したマシンを使用した．

3.2. テストセット[2]

評価用のテストセットは，[6]と同様に NTCIR-10 Formal run を使用した．検索対象の音声データは音声ドキュメントワークショップの 104 講演（約 28 時間，40,746 発話）を用いた．クエリは講演中に正解を含む 100 個を用いた．NTCIR-10 は音声クエリが存在しないため，男女各 5 人，計 10 人の 100 クエリを録音し，全 1000 発話を音声クエリとして使用した．検索精度の評価には MAP(Mean Average Precision)を用いた．

3.3. 実験結果

前述の通り，CTC は音声クエリ最尤系列化方式，その他のモデルは音声データ最尤系列化方式を用いた．Hybrid CTC/Attention は 1 フレームの特徴量から CTC と Attention の 2 つの事後確率が得られる．今回は音声クエリ(Posteriorgram)に Attention から出力された事後確率を，音声データ(最尤系列)に CTC(blank ラベル削除)から出力された事後確率を用いた時，最も良い検索精度が得られたためこれを用いた．

各モデル単体での検索結果を表 3.2 に示す．CTC，Hybrid CTC/Attention に比べ，BLSTM(70.99%)，DNN(69.77%)では高い検索精度を得られたが，両者とも約 4 秒の検索時間を要した．CTC と Hybrid CTC/Attention は最尤系列における状態番号の圧縮，削除処理の効果により短い検索時間となった．

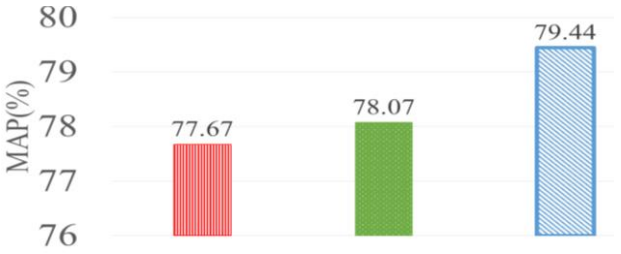
次に，各モデル単体の検索結果を統合した．統合割合は 0.1 ずつ変更し，検索精度が最大の割合(0.2:0.3:0.3:0.2)とした．結果を図 3.1 に示す．

図 3.1 の通り，DNN，BLSTM の Posteriorgram 照合は検索精度が高いが，検索時間と必要メモリ容量が大きい．全モデルの統合結果は 79.44%となり，先行手法の DNN の Posteriorgram 照合(77.67%)と比べ+1.77pt，BLSTM の Posteriorgran 照合(78.07%)と比べ+1.37pt となり，Posteriorgram 照合を上回った。

検索時間は，全モデルを並列で動かすと統合時間(0.05 秒)を加えても 4.26 秒であり，Posteriorgram 照合の 50.57 秒から 90%以上高速化した．メモリ容量も 112GB から 0.5GB へ削減した．

表 3.2 各モデル単体での検索結果
(全て最尤系列化方式)

Model	DNN	BLSTM	CTC	Hybrid CTC/Attention
MAP(%)	69.77	70.99	64.49	66.68
必要メモリ(GB)	0.02	0.02	0.45	0.01
検索時間(秒)	4.21	4.08	1.25	1.03



照合方式	DNN-Posteriorgram	BLSTM-Posteriorgram	全モデル統合
検索時間(秒)	50.57	50.14	4.26
メモリ(GB)	112	112	0.5

図 3.1 Posteriorgram 照合と
全モデル統合の検索精度

4. まとめ

本稿では，状態数の異なる複数の深層学習モデルから得られた異種のサブワード情報を用いた検索結果を，CDP スコアの線形和で統合し検索精度の向上を図った．統合方式の検索精度は 79.44%となり，DNN，BLSTM の Posteriorgram 照合と比べ高い結果が得られた．検索時間は，検索を並列で動作させ，統合処理を加えても 4.26 秒だった．必要メモリ容量は 112GB から 0.5GB と削減し，検索精度は，従来の DNN(Posteriorgram 照合)から+1.77pt の向上を達成した．

謝辞：本研究の一部は JSPS 科研費 18K11358 の助成を受けたものです．

参考文献

[1] Jonathan G. Fiscus et al, SIGIR Workshop Searching Spontaneous Conversational Speech. Results of the 2006 spoken term detection evaluation, pp. 45-50, 2007.
[2] Tomoyosi Akiba et al., Overview of the NTCIR-10 SpokenDoc-2 Task, NTCIR-10 Workshop Meeting, pp. 573-587, 2013.
[3] Masato Obara et al., Rescoring by Combination of Posteriorgram Score and Subword-Matching Score for Use in Query-by-Example, INTERSPEECH, pp.1918-1922, 2016.
[4] 岩崎瑛太郎他，“音声中の検索語検出における深層学習の事後確率を用いたクエリの最尤系列化方式”，日本音響学会春季講演論文集，2018.
[5] 金子大祐他，“音声中の検索語検出におけるドキュメント最尤系列化と上位候補の再照合方式による検索時間・精度の改善”，情報処理学会研究報告(SLP)，2018/12/10.
[6] 金子大祐他，“最尤系列化を用いた音声中の検索誤検出における複数の機械学習モデルによる検索精度の改善”，2018 年度 博士前期課程論文，2018.
[7] Shi-wook Lee et al, “Effective Combination of Heterogeneous Subword-based Spoken Term Detection Systems,” 4 pages, IEEE Spoken Language Technology Workshop (SLT), 2014-12.
[8] S. Watanabe et al, “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, 2017.
[9] S. Watanabe et al, “ESPnet: End-toend speech processing toolkit,” in Proc. Interspeech, 2018, pp. 2207–2211.