1W-06

音声中の音声検索語検出におけるフレーム圧縮手法および複数の深層学習モデルのスコア統合による検索精度・検索速度・メモリ量の改善

畠山和望† 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，音声データを含む大量のデータ中から特定シーンを音声で検索する音声中の音声検索語検出(SQ-STD: Spoken Query - Spoken Term Detection)の研究が盛んに行われている．SQ-STDの代表的な方法であるPosteriorgram照合[1]では高い精度が得られるが，検索時間が長く，メモリ使用量が大きい．これらの削減を目的とし，最尤系列化照合方式[2][3]が提案されたが，検索精度が低下した．

本稿では，照合の際に使用するPosteriorgramのフレームを圧縮するblank-cut方式とフレーム重複排除方式を提案し，検索時間とメモリ使用量の削減を目指す．また，複数の異なる深層学習モデルに提案方式を適用し，その結果を統合することで検索精度・検索速度・メモリ使用量のバランスの良い方式の実現を目指す．

2. 先行研究

2.1. Posteriorgram照合

音声クエリと検索対象の音声データに対してフレーム毎に抽出した特徴量をDNN-HMM(Deep Neural Network - Hidden Markov Model)に入力すると，triphone HMMの各状態等に対応する事後確率が得られる．全ての状態の事後確率(事後確率ベクトル)をフレーム順に並べた行列をPosteriorgramと呼ぶ．Posteriorgram照合方式[1]では事前に作成した音声データのPosteriorgramと，与えられる音声クエリのPosteriorgramを照合し，検索を行う．出力をtriphoneに対応させた場合は約3,000次元の事後確率ベクトル同士の内積に対して負の対数をとり局所距離化するため，長い検索時間と大きなメモリ使用量が問題となる．

2.2. 最尤系列化方式

音声クエリ最尤系列化方式[2]では，音声クエリのPosteriorgramを作成し，各フレームにおける事後確率値が最大となる状態番号を最尤状態として各フレームに最尤状態番号を対応させる．これにより，約3,000次元のPosteriorgramを1次元の最尤状態番号系列に変換する．音声データのPosteriorgramの各確率値は予め負の対数をとり局所距離行列としておき，その行列を音声クエリの最尤状態系列を用いて参照することで内積計算を行わずに照合可能となり，検索時間を削減できたが，メモリ使用量は大きいままだった．一方，音声データ最尤系列化方式[3]は音声データのPosteriorgramを予め最尤状態系列に変換し，参照・照合を行うため，検索時間とメモリ使用量を削減したが，検索精度はPosteriorgram照合に比べて低下した．本稿では音声クエリ・音声データ最尤系列化方式をそれぞれクエリ最尤・データ最尤と呼ぶ．

2.3. フレーム強制統合方式

Posteriorgramの圧縮を目的として，Posteriorgramの隣接する2フレーム分の事後確率ベクトルを平均する方式が提案された[4]．本稿ではフレーム強制統合方式(2FFI: 2 frame forced integration)と呼ぶ．Posteriorgramの1フレームとその次のフレームの事後確率ベクトルの平均をとり，得られた事後確率ベクトルを新たなPosteriorgramに格納していく．これにより，フレーム数が1/2となるため照合時間が削減できたが，若干の精度低下につながった．

2.4. スコア統合方式

クエリ最尤・データ最尤は情報量の減少により，Posteriorgram照合に比べて検索精度が低下したため，複数の深層学習モデルを用いて並列で照合を行い，1つの候補に対して得られる複数の照合スコアを線形和統合する方式が提案され，精度の向上が図られた[5]．また，BLSTM (Bidirectional Long Short Term Memory)とESPnetの3つのモデルを用いて4種のスコア統合を行うことで，従来のPosteriorgram照合等を上回る検索精度が示された[4].

3. 提案方式

3.1. blank-cut方式(b-cut)

End-to-Endの音声認識ツールキットであるESPnet[6]におけるHybrid CTC/Attention[7]のCTC(Connectionist Temporal Classification)から作成したPosteriorgramでは，入出力の系列長を合わせるためblankラベルが挿入される．blankラベル部分の事後確率ベクトルは有効な情報を有しないと仮定し，そのフレームを削除した後に照合する方式を提案する．本稿ではblank-cut方式(b-cut)と呼ぶ．Posteriorgramの最尤状態系列において，最尤状態番号が0のフレーム(blank部分)を削除することで検索時間・メモリ使用量の削減を実現する．また，CTCのPosteriorgramは大半がblankであるため，b-cut方式により有効な情報のみが強調され，検索精度の向上にもつながると考える．

3.2. フレーム重複排除方式(FDD)

Posteriorgram照合の検索精度を維持しつつ検索時間とメモリ使用量を削減するために，フレーム重複排除方式(FDD: frame De-duplication)を提案する．FDD方式のイメージを図1に示す．Posteriorgramにおいて最尤状態系列を作成し，最尤状態番号が連続するフレーム区間の事後確率ベクトルの平均をとることで新たな事後確率ベクトルを作成し，Posteriorgramとして格納していく． 適用前に比べてPosteriorgramのフレーム数が削減されるため，検索時間とメモリ使用量の削減が可能となる．また，全フレームに対して2フレームずつ平均をとる2FFI方式と異なり，事後確率ベクトルの傾向が類似しているフレームのみを圧縮することで情報の損失を抑えられると考えた．

スコア統合方式では用いるモデル数に比例してメモリ使用量が増加するが，提案方式により検索時間とメモリ

Improvement of Retrieval Accuracy, Retrieval Speed, and Amount of Memory by a Frame Compression Method and Score Integration of Multiple Deep Learning Models in Query-by-Example.
†Kazuki Hatakeyama, †Kojima Kazunori, ‡Lee Shi-wook, and †Itoh Yoshiaki, †Iwate Prefectural Univercity, ‡AIST

使用量が削減されているため，スコア統合を行った場合でも 2，3[GB]程度の実用的なメモリ使用量に抑えられると考える．以上のように提案方式とスコア統合方式との併用によって，検索精度・検索速度・メモリ使用量のバランスの良い方式を実現する．

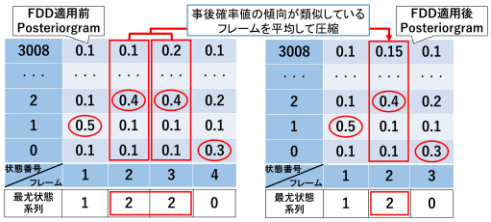

図 1 フレーム重複排除方式(FDD)のイメージ

4. 評価実験

4. 1. 実験条件とテストセット

本稿では先行研究[4]との比較のため，[4]で用いられた BLSTM と Hybrid CTC/Attention を同様の構成で用いた．学習には CSJ 2,702 講演(約 600 時間)の音声を用いた．BLSTM の入力特徴量はフィルタバンク 120 次元を前後 5 フレーム連結した計 1,320 次元とし，出力となる triphone 音響モデルの状態数は 3,009 とした．また， character，syllable，monophone を出力する Hybrid CTC/Attention の入力特徴量はフィルタバンク 80 次元にピッチ 3 次元を加えた 83 次元とし，出力次元数は character で 3,245 次元，syllable で 264 次元，monophone で 43 次元である．窓長は 25ms，フレームシフトは 10ms とした．

実験の評価には NTCIR-10 と NTCIR-12 の Formal run を使用した．各検索対象の音声データは SDPWS(Spoken Document Processing Workshop)の 104 講演(約 29 時間，40,746 発話)，98 講演(約 27.5 時間，37,782 発話)を用いた．音声クエリの数は NTCIR-10 では 100 個，NTCIR-12 では 113 個である． NTCIR-12 ではオーガナイザから提供された 10 人分の音声クエリを用いた．(NTCIR-10 の音声クエリは独自に 10 人の発話を収録) 検索精度の評価には MAP(Mean Average Precision)を用いた．実験結果では 10 人の MAP の平均を示す．検索実験には，CPU に Intel Core i5-9400 ， GPU に NVIDIA GeForce RTX 2070SP ， RAM16GB を搭載したマシンを使用した．

4. 2. 実験結果

まず，b-cut 方式および FDD 方式の評価のために NTCIR-10 における syllable と monophone の Posteriorgram 照合結果を図 2 に示す．b-cut 方式の適用により，適用前(normal)に比べて MAP が向上し，検索時間とメモリ使用量を削減した．b-cut 方式と FDD 方式の併用により，更に検索時間とメモリ使用量を削減し，NTCIR-12 でも同様の傾向が見られた．また， BLSTM のデータ最尤では適用前に比べ FDD 方式により MAP が若干低下したが，検索時間とメモリ使用量を削減した．以上の結果から syllable・monophone の Posteriorgram 照合と，BLSTM のデータ最尤を用いた計 3 種のスコア統合を行う．SQ-STD において，これまでで最も高い検索精度が示された先行研究[4]における 4 種統合結果をベースラインとして比較を行う．NTCIR-10，12 におけるベースラインの 4 種統合および提案方式適用時の 3 種統合に用いたモデル単体の照合結果を表 1 に示す．先行研究で音声データの Posteriorgram にのみ 2FFI 方式を適用したものを SD_2FFI と表記する．

スコア統合結果の比較を図 3 に示す．統合割合は 0.1 ずつ変更し，MAP が最良となる統合割合の結果を示す．統

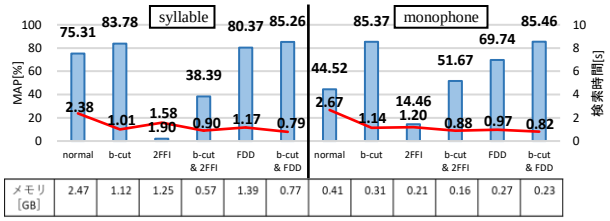

図 2 Posteriorgram 照合での提案方式の評価(NTCIR-10)

表 1 スコア統合に用いた照合の単体性能

		ベースライン[4]				提案方式適用時		
照合方式		データ最尤			クエリ最尤	データ最尤	Posteriorgram照合	
モデル (適用方式)		BLSTM (SD_2FFI)	character	syllable	mono phone	BLSTM (2FFI&FDD)	syllable (b-cut&FDD)	monophone (b-cut&FDD)
N10	MAP	69.92	71.86	78.00	77.96	71.49	85.26	85.46
	検索時間	1.93	0.17	0.18	0.37	0.68	0.79	0.82
	メモリ	0.013	0.0056	0.0057	0.41	0.011	0.77	0.23
N12	MAP	66.48	72.37	75.74	79.51	68.56	82.69	83.50
	検索時間	1.75	0.15	0.16	0.31	0.61	0.67	0.71
	メモリ	0.012	0.0052	0.0053	0.38	0.010	0.72	0.21

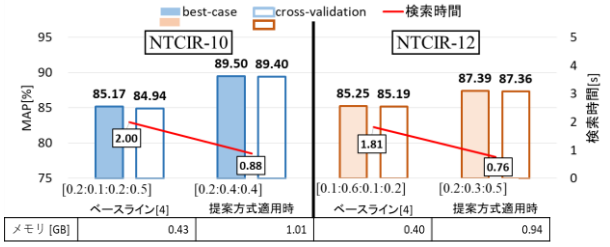

図 3 スコア統合結果の比較

合時間は NTCIR-10，12 において，3 種統合で 0.06s，0.05s，4 種統合で 0.07s，0.06s だった．

提案方式適用時の統合結果をベースラインと比較すると，NTCIR-10 では MAP が 4.33pt(85.17%→89.50%)向上し，検索時間を 1.12s(2.00s→0.88s)短縮した．NTCIR-12 においても同様の傾向が見られた．メモリ使用量は増加したが，2GB 未満であり，実用可能と考える．

5. まとめ

本稿では，Posteriorgram のフレームを圧縮する b-cut 方式とフレーム重複排除方式を提案し，検索時間とメモリ使用量を削減しつつ複数のモデルから得られたスコアを統合することで検索精度の向上を図った．b-cut 方式とフレーム重複排除方式を 3 種のモデルに適用してスコア統合を行い，検索精度が NTCIR-10 で 89.50%，NTCIR-12 で 87.39%と，これまでで最良の検索精度が得られた．更に，検索時間は NTCIR-10 で 0.88s，NTCIR-12 で 0.76s となり，それぞれ約 1s 短縮し，提案方式の有効性が確認できた．

謝辞：本研究の一部は JSPS 科研費 21K12611 の助成を受けて実施した．

参考文献

[1] 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.
[2] 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”，音講論，2018.
[3] 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”，SLP，2018.
[4] 西野将弘他，“異種・複数の深層学習モデルを用いた音声中の検索語検出方式の高精度・低メモリ化”，音講論，2021.
[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複数の機械学習モデルによる検索時間・精度の改善”，音講論，2019.
[6] Shinji Watanabe et al, “ESPnet: End-to-End Speech Processing Toolkit,” INTERSPEECH, pp.2207-2211, 2018.
[7] Shinji 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.