6N-04

音声中の音声検索語検出における 平均事後確率ベクトル圧縮方式の検索精度改良

横田平志† 小嶋和徳† 眞田尚久† 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)や STD にクエリを音声で与える SQ-STD(SQ : Spoken Query)の研究が盛んになっている[1-2]．Posteriorgram 照合方式は，SQ-STD を実現する代表的な方法である．この方式では，音声データと音声クエリそれぞれに対しフレームごとの特徴量を，DNN(Deep Neural Network)の入力とし，triphone HMM の各状態の事後確率を得る．この全状態の事後確率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系列としたものが Posteriorgram となる．Posteriorgram 同士をフレームレベルで照合する Posteriorgram 照合により高い精度が得られている．一方，Posteriorgram 照合方式では，事後確率ベクトルの次元数が数千となり，メモリ使用量が多い．また，局所距離を求める際，内積の計算コストが高く検索時間を要する．音声クエリ最尤系列化[3]では，各フレームで事後確率ベクトルの要素の中で最も事後確率が高い要素(状態)を最尤状態とし，そのフレームにその最尤状態番号を対応させ内積計算をせずに局所距離を求めることで，検索時間を削減した．一方，照合時のメモリ使用量は Posteriorgram 照合と同等で大容量が求められた．これに対し，我々は音声データの Posteriorgram を状態番号ごとに圧縮することにより，メモリ使用量を削減する平均事後確率ベクトル（APPV : Average Posterior Probability Vector）圧縮方式[4]を提案し，メモリ使用量の削減を実現したが，検索精度が若干低下した．

APPV 圧縮方式では各状態の APPV は各フレームに対応付けられた最尤状態番号をもとに作成される．最尤状態番号は DNN の学習モデルから出力された各状態の事後確率から求めており，最尤状態番号が誤っている場合は検索精度の低下につながると考えられる． そのため，本稿では複数の学習モデルから求めた最尤状態番号を用いて APPV を改善し，検索精度向上を目指す．

2. 提案方式

本稿は APPV の改善方式を提案する．BLSTM と DNN の 2 種を用いたため，スコア統合を行うことにより検索精度向上を目指す．ここで 2 種のモデルを用いるときに精度の高い BLSTM の APPV のみを用いる方式を提案する．以降それぞれについて説明する．

(1) APPV の改良方式

まず，各学習モデルを用いて，音声データの最尤系列を求める．次に各フレームの最尤状態番号を比較し，異なっている場合に以下 2 つの処理により APPV を改良する方式を提案する．

① 一致：状態番号が異なるフレームは削除し，一致するフレームのみ用いる
② 併用： 状態番号が A と B と異なる場合には，APPV の構築時にそのフレームを A にも B にも用いる

一致方式では 2 つの最尤状態番号が異なるフレームは誤った推定をしている可能性が高いため，削除することで信頼性の高い APPV を作成できると考えた．図 1 に併用方式のイメージ図を示す．併用方式では図中の F3 のように 2 つの最尤状態番号が 2 と 1 と異なる場合，そのフレームの事後確率ベクトルは状態 1 と状態 2 の両者の APPV の要素とする．片方の最尤状態番号が誤っている場合でも，両方の APPV の要素とすることで APPV の情報量が増加し，検索精度向上を期待するものである．

(2) スコア統合

本稿では BLSTM と DNN の 2 種の学習モデルを用いたため，スコア統合を行うことができ，検索精度の向上を図る．各モデルで照合すると，各発話に対して複数のスコアが得られる．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．α は線形和の統合割合を表し，$0 \le \alpha \le 1$ とする．

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

BLSTM，DNN の APPV をそれぞれ作成し，提案方式を用いて並列で照合を行い，それぞれの照合結果を統合する．一方検索精度が高い BLSTM の APPV のみを BLSTM，DNN の最尤系列に対して兼用して照合を行い，それらを統合することにより高い精度が得られると考え，本稿ではその APPV 兼用方式を提案し，評価を行う．

3. 評価実験

3.1. 実験条件

3.2. テストセット

音声は開発データにおいて最も良い検索精度が得られたため，学習データは DNN では CSJ 2,525 講演の偶数講演（約 280 時間），BLSTM では CSJ 2,702 講演（約 600 時

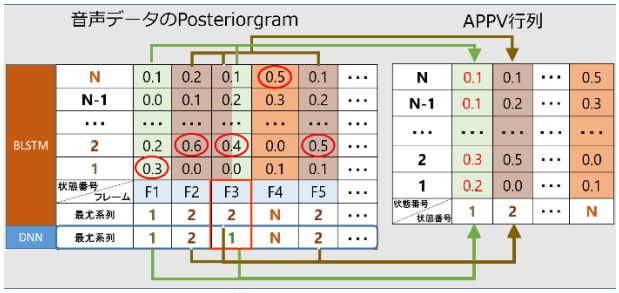

図 1 併用方式のイメージ図

Improving Retrieval Accuracy of Average Posterior Probability Vector Method in Query by Example.
†Nishino Masahiro, †Kojima Kazunori, †Takahisa Sanada, ‡Lee Shi-wook, and †Itoh Yoshiaki,
†Iwate Prefectural Univercity, ‡AIST

表 1　改良実験の結果(NTCIR10)

照合方式	クエリ最尤	全講演			講演毎		
		従来	一致	併用	従来	一致	併用
MAP	76.29	74.42	73.76	75.13	76.00	61.73	76.77
メモリ	114	0.05			3.16	2.86	3.16
TIME	1.9						

表 2　スコア統合の実験結果(NTCIR12)

照合方式	クエリ最尤	全講演			講演毎		
		従来	一致	併用	従来	一致	併用
MAP	68.72	69.84	68.73	70.86	70.87	54.08	71.74
メモリ	107	0.05			2.94	2.68	2.94
TIME	1.8						

間)に用い，入力特徴量は，フィルタバンク 120 次元とし，前後 5 フレームを追加し 1320 次元とした．検索時間の測定には，CPU に Intel Core i7-4770，GPU に NVIDIA GeForce TITAN, RAM 16GB を搭載したマシンを使用した．

テストセット

評価用のテストセットは，NTCIR-10 Formal run と NTCIR-12 Formal run を使用した．それぞれ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 29 時間，40,746 発話)，98 講演(約 27.5 時間，37,782 発話)を用いた．クエリは NTCIR-10 では講演中に正解を含む 100 個，NTCIR-12 ではシングルタームシングルタームのみの 113 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 1,000 発話を音声クエリとした．NTCIR-12 ではオーガナイザーが提供した 10 人分のクエリを使用した．検索精度の評価には MAP(Mean Average Precision)を用いた．

3.3.　実験結果

APPV 行列は全講演で 1 つ，講演毎に 1 つ，発話毎に 1 つの 3 つの圧縮単位が考えられる．NTCIR-10，NTCIR-12 の結果では，全講演に比べて講演毎では MAP が向上したが，講演毎と発話毎を比較すると MAP が低下し，データ量も増加したため本稿では全講演，講演毎で評価する．

APPV 圧縮方式，改良方式の NTCIR-10，NTCIR-12 の検索結果を表 1，表 2 に示す．表中の単位は，MAP は%，メモリは GB，TIME は秒とした．従来の APPV 圧縮方式と比較すると，改良方式の一致方式では，MAP は全講演，講演毎それぞれ NTCIR-10 で 0.66pt (74.42 → 73.76），14.27pt (76.00 → 61.73)，NTCIR-12 で 1.11pt (69.84 → 68.73), 16.79pt (70.87 → 54.08) 低下した．メモリ使用量では，全講演では同じだったが，講演毎では 0.3GB 程度減少した．異なる最尤状態番号を持つフレームを削除した結果，APPV とその情報量が減少し，MAP が低下したと考える．圧縮単位を全講演とした場合，NTCIR10 で 5 種，NTCIR12 で 6 種の状態の APPV が減少していた．

改良方式の併用方式では APPV 圧縮方式と比較すると，検索時間，データ量は同じまま，MAP は全講演，講演毎それぞれ NTCIR-10 で 0.71pt (74.42 → 75.13)，0.77pt (76.00 → 76.77)，NTCIR-12 で 1.02pt (69.84 → 70.86)，0.87pt (70.87 → 71.74) 向上した．メモリ量，検索時間は APPV 圧縮方式と同じだった．

併用方式の講演毎の MAP はクエリ最尤と比べ，NTCIR-10 で 0.48pt (76.29 → 76.77)，NTCIR-12 で 3.02pt (68.72 → 71.74) 向上した．以上の結果より，提案方式の併用方式の有効性を確認できた．

次に，クエリ最尤，APPV 圧縮方式，検索精度が向上した併用方式について BLSTM と DNN の統合結果を比較する．さらに，併用方式では BLSTM の APPV のみを BLSTM，DNN の最尤系列に対して兼用した（兼用方式）の結果を示す．DNN と BLSTM 統合割合は 0.1 ずつ変更し，NTCIR-10 では NTCIR-12 で最も高い検索精度になる割合，NTCIR-12 では NTCIR-10 で最も高い検索精度になった割合とした(NTCIR-12 の兼用方式の講演毎では 4：6, それ以外では 3：7 だった)．

表 3　スコア統合の実験結果(NTCIR10)

照合方式	クエリ最尤同士	全講演			講演毎		
		従来	併用	兼用	併用	併用	兼用
MAP	76.88	75.15	75.73	75.81	76.81	77.23	77.47
メモリ	229	0.10	0.10	0.07	5.98	5.98	3.14
TIME	2.0						

表 4　スコア統合の実験結果(NTCIR12)

照合方式	クエリ最尤同士	全講演			講演毎		
		従来	併用	兼用	従来	併用	兼用
MAP	69.26	70.58	71.49	71.61	71.69	72.39	72.63
メモリ	214	0.10	0.10	0.07	5.98	5.98	3.14
TIME	1.9						

NTCIR-10，NTCIR-12 の結果を表 3，表 4 にそれぞれ示す．併用方式は APPV 圧縮方式と比べ，検索時間とメモリ使用量は同じで，MAP は全講演，講演毎それぞれ NTCIR-10 で 0.58pt (75.15 → 75.73)，0.42pt (76.81 → 77.23), NTCIR-12 で 0.91pt (70.58 → 71.49)，0.70pt (71.69 → 72.39) 向上した．兼用方式は併用方式と比べ，検索時間は同じまま，メモリ使用量は半減し，MAP は全講演，講演毎それぞれ NTCIR-10 で 0.08pt (75.73 → 75.81)，0.24pt (77.23 → 77.47)，NTCIR-12 で 0.11pt (71.49 → 71.61)，0.24pt (72.39 → 72.63) 向上した．兼用手法の講演毎の MAP はクエリ最尤から NTCIR-10 で 0.59pt (76.88 → 77.47)，NTCIR-12 で 3.37pt (69.26 → 72.63) 向上した．

4.　まとめ

本稿では，複数の深層学習モデルから得られた最尤系列を用いて APPV 圧縮方式の精度を改良する一致方式と併用方式を提案し，スコア統合においても併用方式の有効性を示した．一致方式では APPV が減少し，APPV 圧縮方式と比べ，検索精度が低下した．一方，併用方式の検索精度は NTCIR-10，NTCIR-12 の両方で APPV 圧縮方式より高くなり，講演毎ではクエリ最尤の検索精度から平均 1.75pt MAP が向上した．スコアの統合実験では兼用方式の検索精度の向上を確認した．必要メモリ容量は平均 222GB から全講演で平均 0.07GB，講演毎で平均 3.24GB に削減し，検索精度は，クエリ最尤から講演毎で平均 1.98pt の MAP 向上を達成した．

謝辞：本研究の一部は JSPS 科研費 18K11358 の助成を受けて実施した．

参考文献

[1] Tomoyosi Akiba et al., Overview of the NTCIR-10 SpokenDoc-2 Task, NTCIR-10 Workshop Meeting, pp. 573-587, 2013.
[2] T. Akiba, H. Nishizaki, H. Nanjo and G.J.F. Jones : Overview of NTCIR-12 Spoken&Doc Task, NTCIR-12, pp. 167-179, 2016.
[3] 伊藤慶明他，“音声中の検索語検出における音声クエリ・音声ドキュメントのフレームレベル最ゆう系列化照合方式”，電子情報通信学会論文誌 D，pp919-928，2020.
[4] Takashi Yokota et al, “Reduction of Posteriorgram of Speech Data by Compressing Maximum likelihood state sequence in Query-by-example”, APSIPA ASC, pp649-653, 2020.