1W-05

音声中の音声検索語検出における音声データの最尤および上位の状態系列の利用による検索精度向上

皆川玲緒† 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)および検索語(クエリ)を音声で入力するSQ-STD(Spoken Query STD)の研究が活発に行われている[1-2]．代表的なPosteriorgram照合[3]では，事後確率ベクトル同士の内積計算により，高い検索精度が得られる一方計算コストが高い．我々が提案した最尤系列照合[4-5]では，事後確率ベクトルの次元圧縮によって内積計算を省略し，検索時間の削減を実現したが検索精度の低下が見られた．そのため平均事後確率ベクトル(APPV：average posterior probability vector)圧縮方式[6]を提案し検索精度の低下を抑え，検索時間とメモリ使用量の削減を実現した．本稿ではAPPV圧縮方式を改良した上位n件状態番号利用方式を提案する．提案方式では各フレームの事後確率値のうち1位からn位までの状態番号の確率値を利用することで2位以降の情報で補完し検索精度の向上を目指す．

2. 先行研究

2.1. Posteriorgram照合

DNN-HMM型の音声認識システムでは一般的に，音声データを1フレーム毎に特徴量に変換し，DNNに入力するとHMMで定義されたN個の状態等の事後確率ベクトルが得られ，フレーム毎に時間順で並べた行列をPosteriorgramと呼ぶ．Posteriorgram照合では，Posteriorgramを音声データ，音声クエリの両方で求め，連続DP(Dynamic Programming)[7]等で照合を行い，検索結果を求める．

2.2. 音声クエリ(音声データ)最尤系列化照合

音声クエリ(音声データ)の各フレームの事後確率ベクトルのうち，最大確率を示す状態番号をそのフレームの最尤状態番号とし，最尤状態番号のフレーム毎に並べた最尤系列を求める．音声データ(音声クエリ)のPosteriorgramは事前に局所距離化しておくことで，Posteriorgram照合での内積計算を省略し，局所距離の参照のみとできるため，検索は高速になったが，検索精度が低下した．

2.3. APPV圧縮方式

APPV圧縮方式では，音声データのPosteriorgramから最尤系列の中で同じ最尤状態番号を持つフレームの事後確率ベクトルについて，状態毎に平均ベクトルを事前に求め，各フレームの事後確率をこの平均事後確率ベクトルに置き換える．これによりPosteriorgramを状態毎の平均事後確率ベクトル(APPV)行列に圧縮することができる．このAPPV行列を事前に距離化しておくことで照合時に距離行列を参照するのみとなるため，検索時間を最尤系列照合と同程度に抑えたままメモリ使用量の削減を実現している．APPV行列を圧縮する単位は全講演，講演毎，発話毎と分けて作成することができる．

3. 提案方式

提案方式である上位n件状態番号利用方式では，最尤の状態の確率値のみではなく確率値が大きい状態にも有益な情報が含まれていると考え，上位n件の確率値を利用するものである．図1上側が示すように音声データのPosteriorgramから各フレームの事後確率値が1位，2位，…，n位と高い状態番号を抽出し(図1では2位まで)，これをフレーム毎に並べた上位n件状態番号系列を保持しておく．

図1のように音声クエリの最尤系列と音声データの上位n件状態番号系列を用いて，APPV行列を参照し，局所距離を求める．局所距離を求める際，上位の状態番号の情報の方が重要だと考え，線形重みとフレーム毎の事後確率から重みを算出した確率重みの2つの重み付け方法を提案する．

線形重みの場合，1位：2位：…：n位に対して，n：n-1：…：1の重みで各距離の積和を求める．線形重みでは2byte×(n-1)×音声データのフレーム数分のメモリ使用量の増加となる．

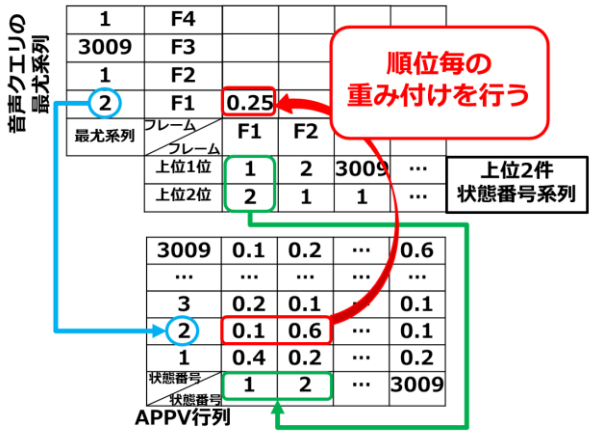
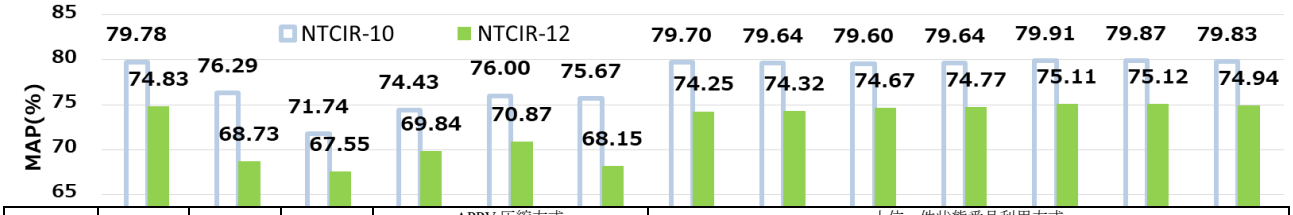


図1 上位n件状態番号利用方式の距離抽出法(上位2位まで)

Using Maximum Likelihood and Highly Ranked State Series of Speech Data to Improve Retrieval Accuracy of Spoken Query Spoken Term Detection
†Minakawa Reo, †Kojima Kazunori, ‡Lee Shi-wook, and †Itoh Yoshiaki ・ †Iwate Prefectural Univercity, ‡AIST



メモリ使用量 (GB)	Posteriorgram 照合	音声クエリ最尤系列化照合	音声データ最尤系列化照合	APPV 圧縮方式			上位 n 件状態番号利用方式						
							線形重み		確率重み				
				全講演	講演毎	発話毎	n=13 (N10best)	n=14 (N12best)	n=13	n=14	n=200 (N10best)	n=100 (N12best)	n=25
N10	114.64	114.64	0.02	0.05	3.16	50.71	0.28	0.30	0.78	0.83	11.46	5.75	1.46
N12	107.14	107.14	0.02	0.05	2.96	47.33	0.27	0.28	0.73	0.78	10.72	5.37	1.37

図 2　SQ-STD における NTCIR-10 と NTCIR-12 の結果比較

確率重みでは，大きな確率を示す状態には大きな重みを与えるべきとの考えのもと，事前に各フレームで上位 n 件までの事後確率の合計P_nを求め，各事後確率に$\frac{1}{P_n}$を乗じて重みとする．重みの合計は 1 となる．フレーム毎に重みが変わるため，確率重みの場合，状態番号毎の n 件までの重みの保持が必要となり，線形重みと比べメモリ使用量は 3 倍となるが，Posteriorgram に比べると大きくない．

4.　評価実験

4.1.　実験条件

事前実験で最も良い検索精度が得られたため，BLSTM の学習データに CSJ 2,702 講演(約 600 時間)を用い，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 i5-9400, GPU に RTX 2070SP, RAM 16GB を搭載したマシンを使用した．

4.2.　テストセット

評価用のテストセットは，NTCIR-10 Formal run[2](N10)と NTCIR-12 Formal run[8](N12)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 N10 では講演中に正解を含む 100 個，N12 ではシングルタームのみの 113 個を用いた．N10 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．N12 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．

4.3.　実験結果

Posteriorgram 照合，音声クエリ最尤系列化照合，音声データ最尤系列化照合，APPV 圧縮方式における 3 つの圧縮単位（全講演，講演毎，発話毎）の結果，および提案方式である上位 n 件状態番号利用方式における 2 つの重み付け方法（線形重み，確率重み）について N10 と N12 の結果を図 2 に示す．それぞれで MAP が最も高くなった場合と確率重みで Posteriorgram 照合の MAP を超えた場合(n=25)を示した．提案方式では全講演で圧縮した APPV 行列を用いた．今回，検索時間は Posteriorgram 照合以外の方式では理論上同じ値となるため図 2 で省略した．Posteriorgram 照合では N10 で約 24.08 秒，N12 で約 21.79 秒，Posteriorgram 照合以外では N10 で約 3.96 秒，N12 で約 3.59 秒であった．

図 2 より N10，N12 とともに先行研究で MAP が最も高かったのは Posteriorgram 照合だった．提案手法で線形重みと確率重みを比べると線形重みでは n=13，14 で最も高い MAP が得られるが，確率重みでは n=200，100 で最も高い MAP が得られ，線形重みよりも高い MAP が得られた．下位の低い確率値も有効に利用できたためと考える．メモリ使用量は線形重みで約 99%削減，確率重みで約 90%削減できた．確率重みで n=25 の場合，N10，N12 とも Posteriorgram 照合より高い MAP が得られた．一方，メモリ使用量は 1.5GB 以下と実用可能と考える．

5.　まとめ

本稿では，APPV 圧縮方式における上位 n 件状態番号利用方式を提案し，低メモリ化を図った．先行研究の Posteriorgram 照合と比べ高精度を維持したままメモリ使用量を約 90%以上削減し，提案方式の有効性を確認することができた．

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] 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] 岩﨑瑛太郎他, “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.
[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”, SLP, 2018.
[6] Takashi Yokota et al:Reduction of Posteriorgram of Speech Data by Compressing Maximum likelihood state sequence in Query-by-example , APSIPA ASC, (2020).
[7] 速水悟他, “連続 DP による連続単語認識実験とその考察”, 電子情報通信学会論文誌, Vol.J67-D No.6 pp.677-684, 1984.
[8] Tomoyoshi Akiba et al, Overview of NTCIR-12 Spoken&Doc-2 Task, NTCIR-12 Workshop Meeting, pp.167-179, 2016.