Vol.2018-SLP-125 No.10
2018/12/10

音声中の検索語検出におけるドキュメントの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

金子大祐[†1]　小嶋和徳[†1]　李時旭[†2]　伊藤慶明[†1]

近年，長時間の音声ドキュメント中から特定のシーンを検索する音声中の検索語検出(STD: Spoken Term Detection)の研究，およびクエリを音声で入力する SQ-STD(Spoken Query STD)の研究が盛んに行われている．DNN(Deep Neural Network)を音声認識に用いることで認識率が大幅に向上することが報告されており，SQ-STD においても DNN を使用することで高精度な検索が可能となった．本稿では SQ-STD における検索時間削減と検索精度向上のための方式を提案する．本稿で提案するドキュメント最優化系列方式では，DNN から得られる事後確率から入力音声の各フレームにおける確率最大値に対応した状態番号を抜き出した系列，最尤系列に変換し予め保持することで，検索の際の計算量の削減を行う．また，検索結果の上位候補の再度照合方式により，検索時間の増加を抑えつつ検索精度の向上を図る．これらの方式のより SQ-STD の検索時間削減および検索精度向上を図る．ドキュメント最尤系列化方式による実験の結果，Posteriorgram 照合と比較して検索語 1 つあたりの検索時間が 90%以上削減できたとともに，112GB のメモリ空間使用量が削減できた．また，上位候補の再度 Posteriorgram 照合方式により P@N の評価指標において，検索精度の向上が確認できた．以上の結果から，SQ-STD に対してドキュメント最尤系列化が有効であることを確認した．

キーワード：音声中の検索語検出，SQ-STD，ポステリオグラム，最尤系列化

Improvement of Search Accuracy and Search Time in Query-by-Example by Using Maximum Likelihood Sequence and Rescoring Top Candidates

DAISUKE KANEKO[†1]　KAZUNORI KOJIMA[†1]
SHI-WOO LEE[†2]　YOSHIAKI ITOH[†1]

This paper proposes a maximum likelihood sequence method of retrieval documents obtained from DNN (Deep Neural Network) and a method of rescoring the top candidates of retrieval results again. Improvement search time and accuracy by using the above two methods.

Keywords：Query-by-Example，Posteriorgram，Maximum Likelihood Sequence

1.　はじめに

近年，記憶デバイスの大容量化や，WEB 上における動画投稿サイト利用の一般化により，動画・音声データを活用する場面が増加している．これに伴い，長時間の動画・音声データ中から特定のシーンを検索する技術が必要とされている．この検索機能の実現のため，音声データ中から検索語（クエリ）を検索する，音声中の検索語検出 (STD：Spoken Term Detection) に関する研究が盛んに行われている．国立情報学研究所が主催する情報検索システムのワークショップである NTCIR Workshop が 2011 年から 2016 年にかけて開催され、STD における検索速度の高速化や検索精度の向上など，様々な観点から評価された[1-6].

STD とは，検索対象である音声ドキュメント内からクエリが発話されている区間を検出するタスクである．STD では音声認識システムを用いて音声ドキュメントを予め音声認識し，その認識結果に対してクエリの検索をテキスト処理で行う方式が一般的である．照合には連続動的計画法(CDP: Continuous Dynamic Programming)が多くの場合用いられる．クエリとドキュメントにおける各発話区間の距離を短い順に並べたものを検索結果として出力する．

2014 年に開催された NTCIR Workshop 11 から音声をクエリとする SQ-STD(Spoken Query STD)の評価がされており，近年のスマートフォン等の普及も追い風になり，研究が活発に行われている．SQ-STD の方式としては，音声クエリを音声認識システムによりテキスト化し，前述のテキストクエリに対する STD システムと同様に検索するのが一般的である．その局所距離には Edit Distance がよく用いられるが，サブワード間の音響的類似度を用いた音響距離を導入することにより高精度が図られた[7].

近年の音声認識においては，DNN（Deep Neural Network）を用いて HMM の状態出力確率を計算することにより，従来の GMM と比べ，音声認識率が大幅に向上することが確認され[8]，現在では DNN を用いた音声認識が一般的である．

STD の検索精度は，事前処理にあたる音声認識の精度に左右されるため，DNN を用いた音声認識システムにより STD についても検索精度の向上が実現された[9].

†1 岩手県立大学
Iwate Prefecctural University.
†2 独立行政法人 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology

SQ-STD システムでは，音声クエリ・ドキュメントを音声認識システムにより認識し，出力された認識結果同士を CDP 照合する方式と，音声をフレームレベルの特徴量で比較する方法等がある．MFCC(Mel-Frequency Cepstrum Coefficients) 等の音声特徴量を用いた場合，不特定話者間での検索精度は低くなるため，音声特徴量を不特定話者の特徴量ベクトルにする必要がある．DNN 等に音声特徴量をフレーム毎に入力し状態毎の事後確率を出力する Posteriorgram[10,11]とし，音声クエリ・ドキュメントを Posteriorgram に変換した上で CDP 照合する方式が代表的である．Posteriorgram に変換したクエリとドキュメント同士の CDP 照合が行われる．事後確率は GMM で求められるが，DNN・RNN (Reccurent Neural Network) 等を用いることで高い検索精度が得られている[12].

Posteriorgram 照合の場合，DNN，RNN は入力として音声特徴量を入力し，出力ノードは HMM(Hidden Markov Model)を構成する状態を単位とすることが一般的である．従来，triphone の状態であれば出力 3000 程度の出力ノードに設定されることが多い．Posteriorgram 照合では，出力確率を用いたフレームレベル照合を行う際に長い検索時間を要してしまう問題がある．この問題に対し先行研究では，フレームレベル状態積上方式[13]を提案し，検索精度をある程度保ったままフレーム毎の内積計算を省略して検索時間を削減した．一方，この方式では，検索対象となる長時間の音声ドキュメントから得られる出力確率をそのままメモリで処理するため，事前に求めたフレーム毎の事後確率を保持しておく必要がある．例えば NTCIR10 のデータでは後述するように 100GB 以上の，大量のメモリ空間が必要になる．

そこで本稿では，フレームレベル状態積上方式を発展させ，音声ドキュメントから得られた事後確率の最尤系列化を行い，float 型(4byte)である事後確率値を short 型(2byte)に変換するとともに，約 3000 次元の出力ノード数の 1 ノードだけを抽出することでメモリ空間使用量および検索時間を削減する方式を提案する．また，検索結果の上位候補に対して再度照合を行い，低下した検索精度の改善を行う方式についての提案も行う．

2. SQ-STD の概要

2.1 DNN の概要

音声認識では一般的に DNN と HMM を併用する DNN-HMM が使用される．DNN の入力ノードは通常，FBANK (log-Filter Bank)や MFCC 等の特徴量を複数フレーム連結して入力する．出力層の各ノードは予め HMM の各状態と対応づけられており，出力ノードから HMM の状態の事後確率が得られる．この系列は Posteriorgram と呼ばれる．

2.2 Posteriorgram を用いた SQ-STD システム

音声ドキュメントと音声クエリの Posteriorgram を用いてフレームレベルの照合を行う．Posteriorgram は DNN のほかに GMM や RNN を用いることで求めることが可能である．照合は一般的に，CDP を用いる．音声ドキュメント・クエリの Posteriorgram 同士で CDP 照合を行う．CDP を行う際の局所距離は式(1)を用いる．

$$\mathrm{D}(P_i, P_j) = -log_{10}\left(\sum_{k=0}^{N} P_i(k) \cdot P_j(k)\right) \qquad (1)$$

P_iは音声ドキュメント i フレームの事後確率ベクトル，P_jは音声クエリ j フレームの事後確率ベクトルで，N は事後確率ベクトルの次元数である．2 つの事後確率ベクトル間の内積を求め，負の対数を取ったものを局所距離として扱う．Posteriorgram を用いた照合は音声クエリをテキスト化する方式と比べて検索精度は高いが，1 フレーム毎に約 3000 の内積をクエリのフレーム数分求めるため，計算コストが大きい．

3. 提案方式

本節では，フレームレベル状態積上方式のメモリ空間使用量を削減するためにドキュメント最尤系列化方式を提案する．低下した検索精度の向上を図るために提案方式により得られた検索結果の上位候補の再度高精度検索方式について提案し，各方式について詳述する．

3.1 ドキュメント最尤系列化方式

ドキュメント最尤系列化方式では，検索対象である音声ドキュメントの Posteriorgram を求め，各フレームにおける事後確率の最大値となる状態番号を抽出する．この系列をドキュメント最尤系列と呼び，予め保持する．具体的な最尤系列への変換方法を図 3.1 に示す．縦軸は音声ドキュメントのフレーム番号，横軸は各フレームにおける出力ノードに対応する状態番号を示す．各フレームの最大事後確率値に注目し，図中の赤字で示す最大確率値に対応する状態番号を取得し，最尤系列に順に格納していく．例えば図 2.1 の 1 フレーム目の最大事後確率が 0.14 であった場合，その最大事後確率値の状態番号 2 が抽出され，最尤系列の 1 フレーム目にこの 2 が格納される．

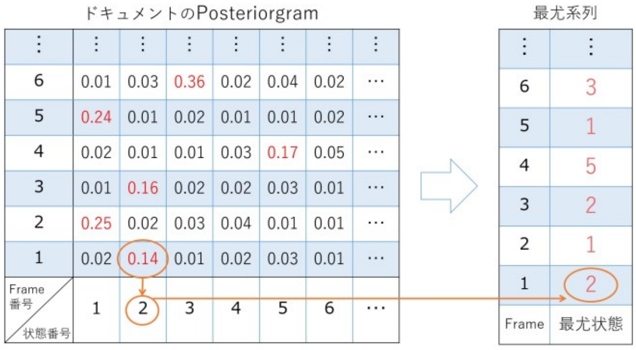

図 3.1 ドキュメント最尤系列への変換

この処理を全てのフレームに対して行い、最尤状態の番号の系列に変換していく．このように Posteriorgram の各フレームにおける最大確率値を抽出することにより，ドキュメント最尤系列はその音声を表現する最も尤もらしい状態番号の系列となる．

音声クエリが与えられると，クエリもドキュメントと同様に Posteriorgram に変換する(図 3.2 左)．クエリとドキュメントで類似しているフレーム系列があった場合，その対応する対角線上の事後確率は高い値を示すため，ドキュメント最尤状態番号に対応するクエリの事後確率値をセットする．図 3.1 の音声ドキュメントの最尤状態の番号系列を図 3.2 右では横軸に配置し，縦軸をクエリのフレーム番号として，クエリと音声ドキュメントの局所距離行列を以下の手順で作成する．

① 音声ドキュメントの各フレームの状態番号を抽出する．(ex. 図中の赤字の 1 と緑字の 2)

② その状態番号に対応するクエリの Posteriorgram 中の状態の確率値をクエリの全フレーム分を距離行列上にコピーする．(ex. 図中の赤、緑の四角の囲み) この段階では事後確率行列であるが，ドキュメント中にクエリが含まれる区間は図中の赤字のように高い事後確率値が対角線に並んでいくため，CDP 照合を行えば対角線上を進み，トータルで高い累積確率値が求められ，クエリを含む区間を特定できる．

③ 事後確率行列を以下の式(2)により局所距離行列とする．

$$D(s_{(i)},j) = -log_{10}(P_{(i,j)}) \qquad (2)$$

以上が局所距離行列作成のイメージであるが，実際には図のような配列は作成せず，クエリの Posteriorgram の事後確率値を式(2)による局所距離行列にするのみである．CDP 照合実行時にこの局所距離行列を参照する形で扱っている．例えば，音声ドキュメントのフレーム i の状態を$s_{(i)}$とし，音声ドキュメントのフレーム j とすると，クエリのフレーム j の局所距離はクエリの局所距離行列$(s_{(i)},j)$を参照するのみとなる．

以上の方式により，事後確率ベクトル同士の内積計算を行わずに，クエリとドキュメントの距離行列を容易に求める（参照する）ことができる．2.2 の Posteriorgram を用いた照合では局所距離に Posteriorgram 同士の内積計算を行っていたが，本方式では内積計算を行わずに事後確率を参照するだけで局所距離を求めることができるため計算コストが少ない．さらに，先行方式のフレームレベル状態積上方式とは異なり，クエリではなくドキュメントを最尤系列に変換しているため，検索対象が約 30 時間の音声データ量であれば 100GB 以上（4byte/次元×3000 次元/フレーム×100 フレーム/秒×30 時間×3600 秒/時間）のメモリ空間使用量が約 6MB（4byte/次元×3000 次元/フレーム×100 フレーム/秒×5 秒/1 クエリ）に削減可能であり，検索時間の削減も期待できる．

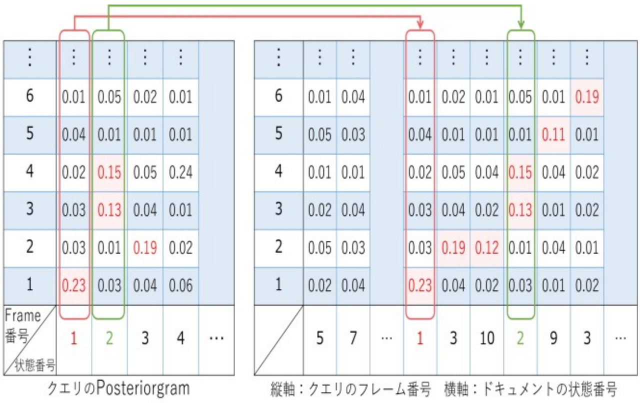

図 3.2 クエリ Posteriorgram のイメージ

最尤系列化による検索結果

発話区間番号	CDP距離
312	0.02
15	0.05
124	0.11
100	0.24
29	0.26
201	0.46
197	0.88
269	1.04
88	1.09
71	1.52
⋮	⋮

Posteriorgram 照合

高精度照合による検索結果

発話区間番号	CDP距離
312	0.01
100	0.04
29	0.06
124	0.13
15	0.16
201	0.46
197	0.88
269	1.04
88	1.09
71	1.52
⋮	⋮

図 3.3 上位候補の高精度検索結果のイメージ

3.2 最尤系列化方式検索結果の上位候補の高精度検索方式

先行方式であるフレームレベル状態積上方式や 3.1 の最尤系列化を行うと，Posteriorgram 同士の照合と比較して情報量が削減されるため，検索精度低下の問題が生じる．最尤系列化により検索した結果，正解候補が上位にあると仮定し，検索結果の上位候補のみをフレームレベルの Posteriorgram で再度照合を行うことにより，検索時間を抑えつつ検索精度の改善を図る．具体的な流れを図 3.3 に示す。

例えば，上位 5 件の再照合を行う場合，上位 5 件に対応する音声ドキュメントの Posteriorgram を読み込み, Posteriorgram 照合を行う．図 3.3 では赤文字になっている候補番号が正解候補区間を示している．3.1 の最尤系列化方式では下位にランクされた正解候補区間が，Posteriorgram 照合により上位にリランキングされる．

本方式では，上位候補の少量の発話区間に対してのみ Posteriorgram 照合を行うことにより、検索時間を抑えつつ，最尤系列化方式の検索精度向上が期待できる．

4. 評価実験

4.1 実験条件

学習データは，CSJ[11]中の 2,702 講演から評価に使用する Core 177 講演を除いた計 2,525 講演のうち，偶数講演(1,255

講演，約 287 時間)を使用した．使用特徴量は，FBANK40 次元とΔ，ΔΔの計 120 次元である．DNN は，入力層，5 層の隠れ層，出力層の計 7 層構造のモデルを学習した．入力特徴量は FBANK 120 次元の特徴量に前後 5 フレームを追加した合計 11 フレームの 1,320 次元とした．出力層は triphone の各状態と対応している．triphone は状態共有を行い 3,009 状態とした．その他の条件は表 4.1 および表 4.2 に示す通りである．

学習および照合は CPU に AMD Ryzen 7 1700 eight-core processor, GPU に NVIDIA GeForce GTX 1080, 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には, 表 4.3 に示す NTCIR-10 Formal run および NTCIR-12 Formal run を使用した. NTCIR10 では検索対象ドキュメントは音声ドキュメントワークショップ (SDPWS : Spoken Document Processing Workshop)の 104 講演（約 29 時間，40,746 発話），NTCIR12 では SDPWS の 98 講演(約 28 時間，37782 発話)を用いた．

表 4.1 特徴量抽出条件

特徴量	FBANK(40dim)＋ ΔFBANK(40dim)＋ ΔΔFBANK(40dim) (計 120 次元)
窓長	25 msec
フレームシフト	10 msec
窓関数	ハミング窓

表 4.2 DNN の学習条件

ノード数		入力層 1,320 出力層 3,009
隠れ層：5		2,048
RBM	学習係数	0.004
	Mini-batch Size	256
	Epoch	10
DNN	学習係数	0.007
	Mini-batch Size	256
	Epoch	30

表 4.3 テストセットデータ

	NTCIR-10	NTCIR-12
音声ドキュメント	SDPWS 104 講演 (約 29 時間, 40746 発話)	SDPWS 98 講演 (約 28 時間, 37782 発話)
音声クエリ	Formal run: 100 (10 名)	Formal run: 113 (10 名)

クエリ数は NTCIR10 では正解を講演中に含む 100 クエリ，NTCIR12 ではシングルタームのみの 113 クエリである．NTCIR-10 で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．NTCIR12 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)と P@N を用いた．P@N は，P@1 から P@5 までを 1 刻みで求め，評価を行った．

4.3 ドキュメント最尤系列化方式

本稿では，ベースラインの Posteriorgram 照合とクエリ最尤系列化方式（先行方式）を提案方式（ドキュメント最尤系列化方式）と比較する．NTCIR10，12 における実験結果を図 4.1, 図 4.2, 表 4.4 および表 4.5 に示す．図中の横軸は 3 つの方式について棒グラフで検索精度 MAP（左縦軸），折れ線グラフで 1 クエリあたりの検索時間（右縦軸）を示す．MAP および検索時間は話者 10 人の平均である．

両データセットにおいて Posteriorgram 照合が最も高い検索精度となった．NTCIR10 ではクエリ最尤系列化方式(73.05%)が提案方式(69.77%)を 3.3 ポイント上回ったが，NTCIR12 では提案方式(66.70%)がクエリ最尤系列化方式(65.03%)を 1.6 ポイント上回った．一方，P@N の評価指標では，両データセットにおいて, P@1 では Posteriorgram 照合と提案方式が同程度の精度になっており, P@2 以降は Posteriorgram 照合の方が平均で 3 ポイントほど高い結果となった．クエリ最尤系列化方式と比較すると，全体的に提案方式の精度が高く，P@1 では平均で 2.4 ポイント高かった．

1 クエリあたりの検索速度は NTCIR10 において，Posteriorgram 照合が 50.38 秒，クエリ最尤系列化方式が 4.01 秒，提案方式が 3.35 秒となり，NTCIR12 ではそれぞれ 46.21 秒，3.47 秒，3.04 秒となった．Posteriorgram 照合と比較すると両データセットで提案方式は 93%の時間削減および 112GB のメモリ削減が実現できた．クエリ最尤系列化方式と比べると提案方式では検索速度は 1～2 割早くなり，メモリは Posteriorgram 照合と同様に 112GB 削減できた．

提案方式の MAP が低下した原因としては，クエリ最尤系列化方式と同様に Posteriorgram 照合よりも次元数を少なくし、照合に用いる情報量を削減したためだと考える．一方，P@1 の時に Posteriorgram 照合と同等であり，上位候補に対する精度が高く, ユーザの利用時には望ましい結果と考える. また，30 時間の音声ドキュメントに対して，3009 次元の Posteriorgram 中から 1 次元のみを抽出し，short 型で数値を扱えるようになったため，100GB 以上のメモリ使用量が削減できたと考える．以上の結果から，検索時間・メモリ空間使用量の削減，および検索精度低下が抑制でき，提案方式の有効性を確認できた．

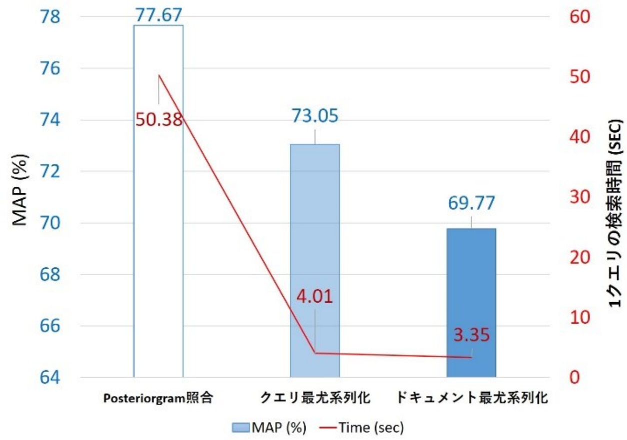

図 4.1 NTCIR10 におけるドキュメント最尤系列化方式の検索時間・精度の結果

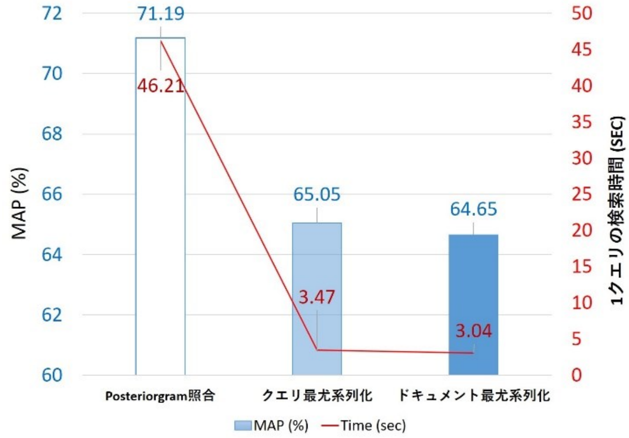

図 4.2 NTCIR12 におけるドキュメント最尤系列化方式の検索時間・精度の結果

表 4.4 NTCIR10 によるドキュメント最尤系列化方式の P@N 指標における実験結果 (%)

	MAP	P@1	P@2	P@3	P@4	P@5
Posteriorgram 照合	77.67	84.80	83.75	81.77	79.53	76.26
クエリ最尤系列化	73.05	81.60	81.10	78.03	75.70	72.50
ドキュメント最尤系列化	69.77	84.80	82.55	79.63	75.96	71.96

表 4.5 NTCIR12 によるドキュメント最尤系列化方式の P@N 指標における実験結果 (%)

	MAP	P@1	P@2	P@3	P@4	P@5
Posteriorgram 照合	71.19	79.74	78.32	75.22	71.57	67.68
クエリ最尤系列化	65.05	77.70	73.66	69.85	66.03	62.64
ドキュメント最尤系列化	66.70	79.30	77.88	74.07	70.14	66.36

4.4 最尤系列化方式検索結果の上位候補の高精度検索方式

NTCIR10，12 におけるドキュメント最尤系列化方式の検索結果に対して，上位候補の 10，50，100 件を Posteriorgram で再照合し，MAP と P@N(N=1,2,3,4,5)を求めた．その結果を図 4.3 および図 4.4 に示す（50 件の時と，N=2,4 の場合は割愛）．縦軸は MAP および P@N の精度を示し，4 本の棒グラフは左から Posteriorgram 照合，ドキュメント最尤系列化方式，上位 10 件の Posteriorgram 再照合方式，上位 100 件の Posteriorgram 再照合方式を示している．

NTCIR10 では，ドキュメント最尤系列化方式の MAP(69.76%)から，上位 10，100 件の Posteriorgram 再照合によって 71.17%, 75.65%と向上したが，MAP では Posteriorgram 照合に及ばなかった．NTCIR12 でも同様で，ドキュメント最尤系列化方式の MAP(66.70%)から上位 10 件，100 件の Posteriorgram 照合により 66.89%，70.31%，と向上した．

P@N の指標で評価を行うと，NTCIR10 では P@1 の時にドキュメント最尤系列化方式でも Posteriorgram 照合と同精度が得られた．P@2 以降は再照合を行う件数を多くすることで，Posteriorgram 照合の精度に近づく傾向が確認できた．

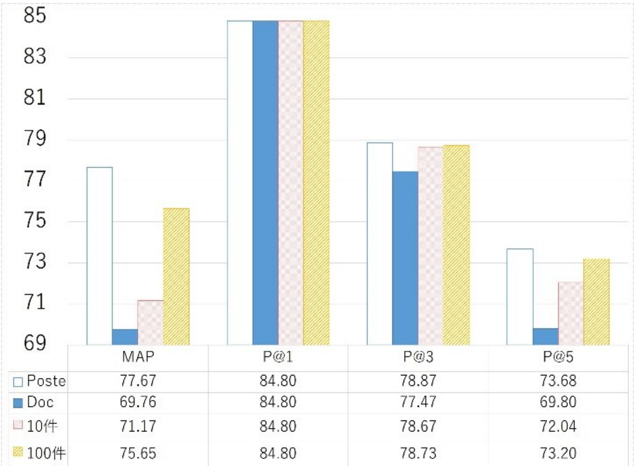

図 4.3 NTCIR10 における上位候補を Posteriorgram 照合により再度照合した時の検索精度の結果

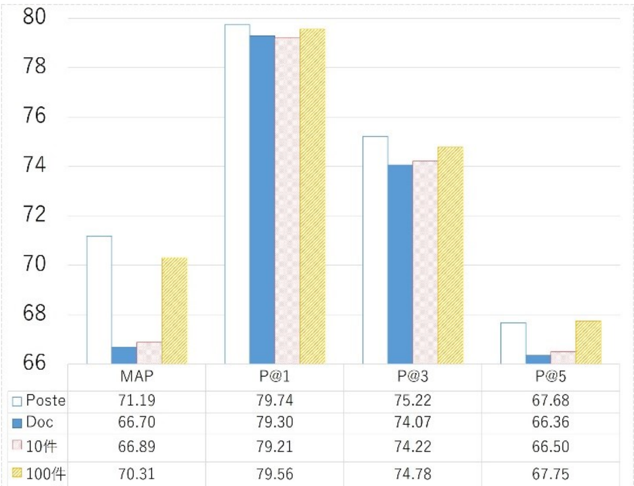

図 4.4 NTCIR12 における上位候補を Posteriorgram 照合により再度照合した時の検索精度の結果

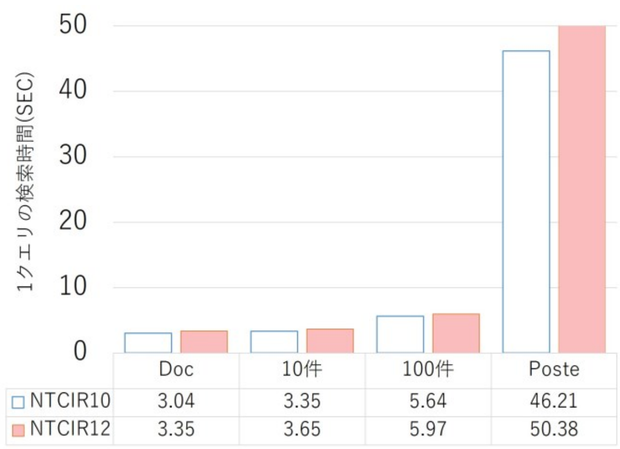

図 4.5 両データセットにおける検索速度の比較

NTCIR12 でも同様の傾向が確認でき，上位候補の再照合方式を用いることで，ほぼ全てのケースで精度が向上することを確認できた．

Posteriorgram 再照合方式を用いた時の 1 クエリあたりの検索速度を図 4.5 に示す．両データセットで同様の傾向が見られ，ドキュメント最尤系列化方式で約 3 秒要し，上位 10 件，100 件に対して Posteriorgram 照合を行うことで+0.3 秒，+2.6 秒，それぞれ処理時間が加わった（Posteriorgram 再照合を行う上位件数と検索時間の増加は線形）．

以上の結果から，少数の上位候補に対して Posteriorgram 照合を用いて再照合を行うことで, MAP が 10 件(+0.3 秒)で+1.41 ポイント，100 件(+2.6 秒)で+5.89 ポイントとなり，P@N では Posteriorgram 照合と同精度が実現できた．ユーザは 1 位の候補を聞いて確認する必要があり，その 1 件目の精度はユーザには重要と考えられるため，上位候補に対して高い精度で正解候補が得られる提案方式の有効性が確認できた．

5. おわりに

本稿では, 先行方式であるフレームレベル状態積上方式(クエリ最尤系列化方式）を発展させ，ドキュメントの Posteriorgram を最尤系列に変換し，クエリの Posteriorgram を参照する形で CDP 照合を行う方式を提案した．また，提案したドキュメント最尤系列化方式で生じた検索精度低下の問題に対して, 検索結果の上位候補を再照合する方式を提案した．

実験の結果，ドキュメント最尤系列化方式では，Posteriorgram 照合から平均で MAP が 6.2 ポイント減少したが、90%以上の検索時間の削減と 112GB 必要なメモリ容量を 6MB に削減することができた. 上位候補の高精度検索方式では，上位 100 件(+2.6 秒)で検索精度が約 6 ポイント向上し，検索精度を Posteriorgram 照合に近づけることができた．

今後は, feed-forward 型以外の深層学習モデルの導入を行う予定である．また，クエリ最尤系列化方式と提案方式の統合を行うなど，さらなる検索精度の向上を検討していきたい．

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 IR for Spoken Documents Task in NTCIR-9 Workshop, NTCIR-9 Workshop Meeting, pp. 223-235, 2011.
[3] Tomoyosi Akiba et al., Overview of the NTCIR-10 SpokenDoc-2 Task, NTCIR-10 Workshop Meeting, pp. 573-587, 2013.
[4] Tomoyosi Akiba et al, Overview of NTCIR-11 Spoken&Doc Task, NTCIR-11, pp. 350-364, 2014.
[5] Tomoyoshi Akiba et al, Overview of NTCIR-12 Spoken&Doc-2 Task, NTCIR-12 Workshop Meeting, pp.167-179, 2016.
[6] 大島聡史，伊藤慶明，“音声クエリの複数認識結果 を用いた音声中の検索語検出”, 情報処理学会研究 報告. SLP, 音声言語情報処理，2014.
[7] 岩田耕平他，“語彙フリー音声文書検索方式における新しいサブワードモデルとサブワード音響距離の有効性の検証”，情報通信学会論文誌，Vol. 8， No. 5， pp. 1990-2000，2007.
[8] Geoffrey E. Hinton et al, A Fast Learning Algorithm for Deep Belief Nets, Neural Computation, Vol. 18, pp. 1527-1554, 2006.
[9] 紺野良太 他，“DNN の事後確率から構築したサブ ワード間音響距離の STD への適用”, 日本音響学会春季研究発表会講演論文集，2015.
[10] Yaodong Zhang et al, Unsupervised Spoken Keyword Spotting via Segmental DTW on Gaussian Posteriorgrams, ASRU 2009, pp.398-403, 2009
[11] Masato Obara et al., Rescoring by Combination of Posteriorgram Score and Subword-Matching Score for Use in Queryby-Example, INTERSPEECH, pp.1918-1922, 2016.
[12] 紺野良太他，“SQ-STD における DNN 及び CTC 導入方式の検討”，日本音響学会春季講演論文集，2016
[13] 小原真人他，“音声中の検索語検出における Posteriorgram 照合の検索時間削減方式”，日本音響学会春季論文講演集，pp.83-86，2018
[14] National Institute for Japanese Language and Linguistics, “Corpus of Spontaneous Japanese, 2017.