1R-07

音声中の検索語検出における検索精度向上のための フレームレベル照合方式

皆川玲緒† 小嶋和徳† 伊藤慶明†
岩手県立大学†

1. はじめに

記憶媒体の大容量化と共に，音声データを含む大量のデータ中から特定のシーンをクエリ（検索語）で検索する音声中の検索語検出（STD：Spoken Term Detection）の研究が活発に行われている[1-2]．任意のクエリでの STD を実現する方法として，サブワード間[3-4]，状態間[3]，フレームレベル照合[5]がある．サブワード間照合では，検索対象である音声データを予め音声認識システムで自動認識し，その認識結果のテキスト系列をサブワード系列として保持しておく，このサブワード系列とクエリのサブワード系列とで連続 DP(Dynamic Programming)[6]照合により検索を実現する．状態間照合では，サブワードを HMM の状態に展開した上で照合を行うことで，サブワード間照合よりも詳細な単位で照合を行うことができる．フレームレベル照合方式では，音声認識システムの音響モデルとして用いられる DNN(Deep Neural Network)[7]の出力である事後確率ベクトルを用いてフレームレベルでより詳細な照合を行う．本研究では，検索精度の向上を目指し，フレームレベル状態系列照合（事後確率距離）方式を提案する．提案方式ではフレームレベルでの連続 DP に用いる局所距離を事前に構築済みの音響距離から，後述するように音声データの事後確率を用いた距離に変更した．音声データ中の同一の最尤状態番号であっても事後確率は異なっており，その事後確率値を距離に反映することでより詳細な照合を行い，検索精度の向上を期待する．

2. 先行研究

はじめにでも述べた通り，任意のクエリの STD を実現する 3 つの方式について概説する．

2.1. サブワード間照合

音声データの自動認識結果とテキストクエリを各々サブワード系列に変換して，事前に構築済みのサブワード間音響距離[3]を用いて連続 DP 照合を行い，検索結果を求める．

Improving of Retrieval Accuracy for Spoken Term Detection in Frame Level Matching Method
†Minakawa Reo, Kojima Kazunori, and Itoh Yoshiaki・Iwate Prefectural Univercity

2.2. 状態間照合

サブワード間照合と同様に自動認識結果とテキストクエリから各々のサブワード系列を得る．このサブワード系列を状態系列に自動変換し，事前に構築済みの状態間音響距離[3]を用いて連続 DP 照合を行い，検索結果を求める．

2.3. フレームレベル状態系列間照合

音声データに対し，各フレームで DNN から得られた事後確率が最大となる状態Sをそのフレームの状態とする．テキストクエリを状態に自動変換する．音声データとテキストクエリ共に状態系列となる．音声データがフレームレベルのまま状態系列間で連続 DP 照合を行い，検索結果を求める．

3. 提案方式

本研究では，STD の検索精度向上のため，連続 DP 時に音声データの事後確率を用いて距離化したフレームレベル状態系列照合（事後確率距離）方式を提案する．STD における連続 DP の特性として，クエリの状態等の系列数が音声データ中のクエリが発話されている区間の状態等の系列数と合わない場合に検索精度が下がってしまう．テキストクエリを自動変換した状態系列と比べ，音声データのフレームレベルの状態系列は，細かい時間単位の状態系列であるため，検索精度が低下してしまう．そこで，テキストクエリの状態系列を疑似的なフレームレベルの状態系列に変換する．先行研究[5]では，各状態における継続フレーム数を，学習データの強制アライメント結果より求め（平均 2.7），テキストクエリを状態系列に変換した後，同じ状態番号を複数個（2~4 つ）並べることにより疑似的なフレームレベルの状態系列とした．

提案方式では，クエリを[5]と同様に，同じ状態番号を複数個ずつ並べ疑似的なフレームレベルの状態系列とし，連続 DP の際の局所距離を状態間の音響距離を用いず，次のような事後確率を用いた距離を求める．クエリの状態系列の各状態に対応するように音声データの事後確率を抽出し，クエリの状態系列順に音声データの事後確率値を疑似的に並び替える．クエリが発話された区間であれば，対角線上に高い確率値が並ぶ．クエリの i 番目の状態 S_i とし，音声データ

の j フレーム目との局所距離 $d(i, j)$を以下の式(1)で求める．式(1)の $P_j(S_i)$は音声データの j フレーム目の状態 S_iに対応した事後確率である．

$$d(i,j) = -\log_{10}\left(P_j(S_i)\right) \quad (1)$$

このように，局所距離はフレーム毎に異なる事後確率値を用いるため詳細な照合が可能となる．

4. 評価実験

4.1. 実験条件

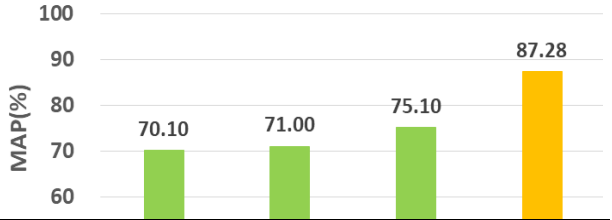
先行研究との比較のため，音声データの認識には音声認識エンジン Julius[8]を使用し， DNN-HMM を用いた．追加実験では音響モデルとして BLSTM-HMM を使用した．音響モデル，言語モデルに用いる学習データは CSJ の偶数講演（約 287 時間，1,255 講演）を用いた．音響モデルは各 3 状態の triphone で構成し，状態共有によって 3,009 状態とした[9]．先行研究のサブワード間照合方式，状態間照合方式，テキストクエリにおけるフレームレベル状態系列間照合の音響距離は DNN の事後確率に基づく Confusion Matrix を用いて構築した音響距離 (DNN-CM) を用いた[9]．

4.2. テストセット

評価用のテストセットは NTCIR-10 Formal run[2]を使用した．検索対象の音声データは SDPWS 104 講演（約 29 時間，40,746 発話），クエリは NTCIR-10 で提供された 100 個（講演音声中に発話あり）を用いた．評価指標は MAP(Mean Average Precision)を用いた．

4.3. 実験結果

先行研究であるサブワード間照合方式，状態間照合方式，フレームレベル状態系列間照合方式と提案方式であるフレームレベル状態系列照合（事後確率距離）方式の検索結果を図 1 に示す．提案方式の検索時間は並列処理を行った結果であり，並列処理を行っていない結果は括弧内に示した．図 1 より，先行研究で最も MAP が高かったフレームレベル状態系列間照合方式の 75.10%よりも提案方式では 87.28%と 12.18pt 向上したが，検索時間が 1.00 秒から 1.69 秒となり，メモリ使用量も音声データが SDPWS 104 講演の場合約 1,000 万フレームあり，各フレームの事後確率を保持するため， 114.63GB と Posteriorgram 照合[10]と同量となった．



	サブワード間	状態間	フレーム間	提案方式
検索時間(s)	0.03	0.18	1.00	1.69(4.18)
メモリ量 (GB)	0.24	0.03	0.05	114.63

図 1 照合方式の検索結果

表 1 音響モデルを変更した提案方式の検索結果

音響モデル	DNN	BLSTM
MAP(%)	87.28	89.27

追加実験として提案方式の DNN の音響モデルを BLSTM に変更した場合の検索精度を表 1 に示す．音声データの事後確率ベクトルのデータ量は DNN と BLSTM でフレーム数と状態数が同じであるため，メモリ使用量と検索時間は同じになる． 表 1 より，DNN の音響モデルの場合よりも BLSTM の音響モデルの場合の方が MAP が 89.27%と 1.99pt 向上した．

NTCIR-10[2]で最良であった MAP 67.5%と比較して，21.77pt 向上，論文等で発表された NTCIR-10 Formal run の最良の結果である状態間照合方式[11]の MAP 78.4%で，そのリスコアリング方式の MAP 81.7%と比較しても 7.5pt 高くなった．以上より提案方式の有効性を確認できた．

5. まとめ

本稿では，音声データの事後確率から求めた距離を用いて連続 DP を行うフレームレベル状態系列照合（事後確率距離）方式を提案し，検索精度の向上を図った．音響モデルに BLSTM を用いた検索精度が 89.27%となり，先行研究と比べ高い検索精度が得られ，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] 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”, 情報処理学会論文誌, vol.48, no.5, pp.1990-2000 (2007)
[4] 西崎博光他,“音声認識誤りと未知語に頑健な音声文書検索手法”, 信学論, vol.J86-D-II, no.10, pp.1369-1381 (2003)
[5] 紺野良太他，“音声中の検索語検出におけるフレームレベル状態系列間照合方式”，信学技報, vol. 115, no. 146, SP2015-37, pp. 7-12 (2015)
[6] 古井貞熙，“音声情報処理”，森北出版 (1998)
[7] G.E. Hinton et al,“A fast learning algorithm for deep belief nets”, Neural Computation, Vol.18, No.7, pp.1527-1554 (2006)
[8] 大語彙連続音声認識エンジン Julius, http://julius.sourceforge.jp/
[9] 紺野良太他，“音声中の検索語検出における Deep Neural Network の出力確率を用いた音響距離構築方式”， 信学論, Vol.J100-D, No.8, pp.798-807 (2017)
[10] 小原真人他，“音声中の音声検索語検出における Posteriorgram 照合の検索時間削減方式”， 日本音響学会春季講演論文集，2018.
[11] 丹治遥他，“音声中の検索語検出におけるクエリの関連語を利用したリスコアリング方式”，情報処理学会論文誌, vol.61，no.1，pp.103-112 (2020)