2,000 時間の音声データの検索語検出の検索時間削減

◎三上凌，小嶋和徳（岩手県立大），李時旭（産総研），伊藤慶明（岩手県立大）

1　はじめに

近年，HDD(Hard Disc Drive)等の記憶媒体の大容量化や，Web 上での動画投稿サイトの利用が一般的となり，音声を含む大量のビデオデータの利用機会が増加している．

保存されたデータに効率的にアクセスするために，音声から特定の言葉が話されている区間を検出する技術(STD:Spoken Term Detction)が重要と考えられる．

テキストクエリでの音声データ検索を行うフレームレベル系列照合 [1]は，クエリと音声データ（全発話）をフレームレベルで照合する手法である．高い検索精度が実現できる一方で，検索時間が長くメモリ使用量が多い．音声データの増加に比例して，検索時間及びメモリ使用量が増加するため，その削減が課題となる．

この課題に対処するため，事前検索手法[2]が提案された．事前検索手法は，予め全音素 3gram で音声データを照合しておく．ある音素の 3gram の距離が小さい発話には，その音素 3gram が話されていると仮定し，候補発話として一定数保持する．クエリに含まれる音素 3gram が話されている発話を高速に絞り込むことで，全発話と照合する手法と同等の検索精度を保ちながら検索時間を削減した．一方，保持件数の設定が困難であること，候補発話が順位付けされていないために全ての候補発話を再照合する必要があった．

本研究では、これらを解決するために 2 つの手法を提案する．1 つ目は，各発話で実際に話されている音素 3gram のみを保持するようにする．これにより無駄な候補を持たないよう，各発話の音素系列数を基準に各発話ごとに候補数を限定する． 2 つ目は，候補発話に距離情報を付与し，その距離情報を用いてクエリと候補発話の近似的な距離を求める．これにより，再照合件数を削減し，高速な検索を実現する．

本研究では 2,000 時間のテストセットを構築して評価を行う．これは，日本語の STD テストセットの中で，我々が知る限り最大の規模である．

2　関連研究

2.1　フレームレベル系列照合

Posteriorgram とは，Fig.1 に示すように，音声データをフレームごとに深層学習モデルに入力して得られる各音素に対応する事後確率ベクトルを，時系列順に並べた行列である．フレームレベル系列照合は，Posteriorgram を用い，フレームレベルでの詳細な照合を行い，高い検索精度が得られる．一方で検索時間とメモリ使用量が多い．

フレームレベル系列照合では，Posteriorgram 中の事後確率を．負の対数により距離化し，Posteriorgram と同じ大きさの距離行列を構築しておく．クエリ入力後，クエリは音素系列に変換される．クエリの音素系列に対し，メモリ上の距離行列を参照して，連続 DP により累積距離を求め，クエリの話されている区間を特定する．この最小累積距離をクエリと発話の距離とする．

sh	0.1	0.1	0.1	…	0.1
…	…	…	…	…	…
i	0.1	0.0	0.0	…	0.1
a	0.6	0.9	0.5	…	0.2
frame	1	2	3	…	N

Fig. 1 Posteriorgram のイメージ

“Search Time Reduction of Spoken Term Detection in 2,000 hours of speech data” Ryo Mikami, Kazunori Kojima (Iwate Prefectural Univercity), Shi-Wook Lee(National Institute of Advanced Industrial Science and Technology), Yoshiaki Ito(Iwate Prefectural Univercity)

e	1.5	1.2	1.7	2.1	7.1	7.1	0.1	…	3.2
t	4.9	2.3	0.9	4.5	4.9	0.1	2.5	…	4.5
a	0.1	5.1	2.1	0.1	0.1	1.9	5.2	…	3.7
w	1.1	2.9	0.1	3.1	2.3	2.3	3.2	…	3.9
i	3.1	0.1	4.1	1.2	4.1	4.1	1.4	…	4.1
Frame	1 (a)	2 (i)	3 (w)	4 (a)	5 (a)	5 (t)	6 (e)	…	N (u)

Fig. 2 連続 DP での最小累積距離が得られるパスのイメージ

2.2 事前検索手法

事前検索手法では，事前に構築した全ての音素 3gram の候補発話の索引を用いて，全発話からクエリ中の音素 3gram が話されている発話を絞り込み，得られた候補発話にのみ再照合することで，検索時間とメモリ使用量を削減する．

音素 3gram の索引は、以下の手順で構築される．まず，各音素 3gram に対して全発話との照合を行い，各々の累積距離を求める．累積距離が小さい発話ほど，その音素 3gram が実際に話されている可能性が高い．各音素 3gram について，累積距離に基づいて全発話を順位付けし，予め設定した保持件数に従って上位の発話番号のみ(4 Byte)を保持する．クエリが与えられると，クエリは音素 3gram に分割される．それぞれの音素 3gram で索引を参照し，得られた発話番号を和集合することで，候補発話が得られる．その後，候補発話のみにフレームレベル系列照合で再照合する．Posteriorgram は外部記憶装置に配置し，メモリ上には索引のみを保持するため，メモリ使用量が削減できる．Posteriorgram の読み込み時間が検索時間に含まれるため，候補発話数の増加に伴って検索時間が増加する．

3 提案手法

3.1 上位 3gram 保持方式

事前検索手法の索引中，ある発話で話されている音素 3gram を全て保持することが望ましい．一方，音素 3gram の累積距離の情報だけで音素 3gram が話されている発話を特定することは困難である．全ての音素 3gram で一律に保持件数を設定しているが，音声データ中で話される音素 3gram の出現頻度には偏りがあり，話されない音素 3gram を保持している可能性が高い．このため，メモリ使用量及び候補発話数が増加し，検索時間の増加につながる．

そこで，発話毎の上位 3gram 保持方式を提案する．各発話に対して全音素 3gram で照合し，発話毎に累積距離が小さい順に音素 3gram を並べ，発話の音素系列数を基準にその上位音素 3gram を保持する．これにより，発話で話されていそうな音素 3gram のみを保持することになる．その結果，各音素 3gram はそれぞれ異なる発話数を候補として保持する．

3.2 距離計算方式

事前検索手法で得られる候補発話にあらかじめ累積距離が付与されれば，クエリとの累積距離が推定でき，累積距離が小さい順に再照合候補とすれば，再照合件数を減らしつつ，高い検索精度を維持できると考える．距離計算方式では事前検索時の索引構築において，発話番号だけでなくそれに対応する累積距離も保持する．クエリが与えられると，クエリ中の各音素 3gram で索引を参照し，累積距離を総和することで，各発話とクエリの近似的な累積距離を算出する．これにより，各発話とクエリの累積距離を高速に算出できる．なお，クエリ中の音素 3gram が保持されていない発話における，その音素 3gram の累積距離は，十分に大きい固定値を与える．発話毎に累積距離を保持するため，索引のメモリ量は従来の 2 倍となる．

Fig.3 のように音声データ中のある発話で「坂井」というクエリが話されていなくとも，クエリ中の全ての音素 3gram が話されていた場合に近似した距離が小さくなる．このような発話を除外するために，上位の発話をクエリ全体で再照合する必要がある．

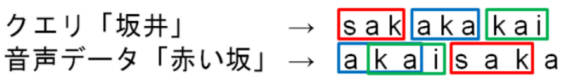

Fig. 3 音声データ中でクエリの全 3gram が話されている例

4 実験

フレームレベル系列照合，事前検索手法，及び提案手法（上位 3gram 保持方式，距離計算手法）で同じ検索精度が得られる条件で，検索時間とメモリ使用量を比較した．事前検索手法及び提案手法の索引構築及び再照合にはフレームレベル系列照合を用いた．検索精

度の評価指標には，P@1 を用いた．

4.1 テストセット

Table 1 に示すように，約 2,000 時間の音声コーパスとして公開されている LaboroTVSpeech [3]を検索対象音声データとして用い，10 クエリを設定した．Table 2 にクエリとそのクエリが出現した正解発話数を示す．

Table 1 構築したテストセット

音声データ	LaboroTVSpeech (1,985 時間，1,601,935 発話)
クエリ数	10

Table 2 設定したクエリと正解発話数

クエリ	正解発話数
茨城県教育委員会	3
IP アドレス	24
霜降り明星	26
サザン オールスターズ	33
UFO	73
大谷翔平	164
地震	1,034
ホームラン	1,102
逮捕	4,393
台風	4,434

4.2 音声認識モデルと学習条件

Posteriorgram を作成するための音声認識モデルには Hybrid CTC/Attention を用いた．その学習条件を Table 3 に示す．

Table 3 音声認識モデルと学習条件

モデル	Hybrid CTC/Attention
学習データ	CSJ 2,702 講演 （約 600 時間）
特徴量	83 次元 (FBANK:80 次元 +ピッチ特徴量:3 次元)
窓長	25ms
フレームシフト	10ms
出力	音素
出力次元数	43

4.3 評価用マシンの性能

評価用マシンの性能を Table 4 に示す．CPU には，Intel Core i5 11400 を利用し，シングルスレッドで実験した．

Table 4　評価用マシンの性能

CPU	Intel Core i5 11400
RAM	96GB
SSD	Sumsung 980PRO

4.4 保持件数及び再照合件数の設定

各音素 3gram の保持件数を設定するために Table 5 に示す NTCIR-10 テストセット [4]を開発データとして利用した．発話番号の保持件数を増加させながら検索精度を確認した結果，200 件保持でフレームレベル系列照合と同等の検索精度が得られた. 構築した約 2,000 時間のテストセットの音声データ長は NTCIR-10 の 71 倍であったため，14,200 保持件数とした．

提案手法の上位 3gram 保持方式では，基本的に各発話の音素系列数で変更する必要がないと考える．実際に、音素系列数を保持した結果，フレームレベル系列照合と同等の検索精度が得られた．

提案手法の距離計算手法では, NTCIR-10 テストセットで再照合件数を増加させながら検索精度を確認した結果, 12 件でフレームレベル系列照合と同じ検索精度が得られた．そのため，71 倍である，852 件を再照合件数とした．

Table 5 NTCIR-10 テストセット

音声データ	SDPWS104 講演 (28 時間，40746 発話)
クエリ数	100

4.5 実験結果

実験結果を Fig.4 に示す．事前検索手法と比較して，提案手法の上位 3gram 保持のみ適用した場合，メモリ使用量を 56%削減できたが，検索時間が 3.5 倍に増加した．検索時間が増加した原因としては，音声データで話される音素 3gram に偏りがあったためと考えられる．クエリ中の音素 3gram が，多くの発話で話されていたことで, 候補発話数が増加し，検索時間が増加したと推測する．提案手法の距離計算手法も適用することにより，事前検索手法と比較して，メモリ使用量が 11%，検

索時間が 98%削減できた. 距離計算手法では, Table 6 に示すように再照合件数が大幅に削減されたことで，0.03 秒での検索が実現できた.

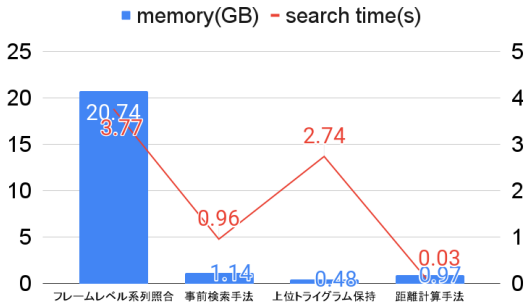

Fig. 4 検索時間とメモリ使用量の比較

Table 6 再照合件数の比較

事前検索手法	87,878
上位 3gram 保持	311,584
距離計算手法	852

5 まとめ

本研究では，事前検索手法における索引構築の精緻化と検索時間削減を目的とし，上位 3gram 保持方式，及び距離計算手法を提案した. 2,000 時間のテストセットを構築して評価した結果，上位 3gram 保持方式と距離計算手法の両提案手法を用いることにより，検索時間が 98%削減され，提案手法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] 皆川玲緒他，“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,” 情報処理学会第 84 回全国大会, 2022.

[2] H. S. e. al., “Fast Spoken Term Detection using pre-retrieval results of syllable bigrams,” Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, 2012.

[3] S. Ando, “Construction of a Large-Scale Japanese ASR Corpus on TV Recordings,” 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

[4] T. A. e. al, “Overview of the NTCIR-10 SpokenDoc-2 Task,” NTCIR-10 Workshop Meeting , 2013.