4P-1

複数サブワード認識結果統合による音声中の検索語検出の精度向上 －複数の音響モデル・言語モデルの利用－

斉藤 裕之† 伊藤 慶明† 小嶋 和徳† 石亀 昌明† 田中 和世†† 李 時旭†††
†岩手県立大学ソフトウェア情報学部 ††筑波大学 †††産業技術総合研究所

1. はじめに

近年記録媒体の大容量化に伴い大量の音声データから目的区間を検索する機能が求められている．このため，音声中の検索語検出(Spoken Term Detection：以下 STD)に関する研究が盛んに行われるようになった．STD では未知語の検索語へ対応するためサブワード認識結果を用いる方法が代表的である．しかし，既知語と比べ未知語では検索精度が大きく劣る．本研究では未知語の検索語に対する STD の精度向上目指す．

先行研究[1][2]では，複数の言語モデル(以下 LM)を用いて複数のサブワード検索結果を生成し，それらを統合することで検索性能が上昇することが確認された．本研究では音響モデル（以下 AM）を JNAS に限らず，JNAS，CSJ 各々から作成した複数の音響モデルを併用し複数の検索結果を統合することで検索性能の向上を図る．

2. サブワード単体による検索性能

2.1. サブワード認識結果を用いる STD 方式

音声ドキュメントを認識し，その認識結果からサブワード系列のデータベースを事前に作成する．検索語が未知語の場合，検索語のサブワード系列と音声ドキュメントのサブワードデータベースとを連続 DP 法により照合し，候補発話区間を特定する．

2.2. サブワードモデル単体の検索性能

本稿では，サブワードモデルとして monophone と triphone を用い，学習データとして新聞記事読み上げ音声コーパス(JNAS)と日本語話し言葉コーパス(CSJ)を用いて AM と LM を構築する．サブワード単体の検索性能を表 1 に示す．表中上半分は AM，LM を同一コーパスから学習時の精度，下半分は別コーパスから学習時の精度である．

評価指標には Mean Average Precision(MAP)を，評価用データ，検索語は[3]の簡易評価版を用いた．表中の SW とはサブワードを指す．

表 1 より CSJ を学習に用いることによって，モデル数の多い triphone の AM，LM の学習が改善したため triphone を用いた際の検索精度が向上したと考える．CSJmonophoneAM のみ精度が低下したが，これは monophone の音素数が 43 と少なく，CSJ により過学習となったと推察する．

表 1：サブワードモデル単体の検索性能

SW	AM	LM	MAP(%)
monophone	JNAS	JNAS	55.04
triphone	JNAS	JNAS	57.03
monophone	CSJ	CSJ	51.31
triphone	CSJ	CSJ	70.08
monophone	JNAS	CSJ	56.13
triphone	JNAS	CSJ	66.93
monophone	CSJ	JNAS	52.04
triphone	CSJ	JNAS	64.16

3. 複数検索結果の統合

3.1. 複数検索結果の統合

検索結果の統合には(1)式を用いる[2]．サブワードを s とし，AM，LM 学習用コーパスに a,l を用いた際のある発話区間 u の距離を d(s,a,l,u)とする．但し，本稿では s は monophone と triphone，a,l は JNAS と CSJ とする．N は統合する各サブワードの検索結果の数である．

各サブワードの検索結果 i における候補発話区間 u に対して重み $weight_i$ を乗じた線形和をとることで統合距離 D(u)を求め，候補発話区間を決定する．

$$D(u) = \sum_{i=1}^{N} weight_i \times d(s,a,l,u) \qquad (1)$$

4. 評価実験・考察

4.1. 実験条件

実験では簡易評価用セット[3]49 講演，13 時間分と，50 個の検索語を用いた．なお，評価用デ

An Improvement for Spoken Term Detection using Recognition Results of Plural Subwords -Using of Plural Acoustic Models and Language Models-
Hiroyuki Saito†, Yoshiaki Itoh†, Kazunori Kojima†, Masaaki Ishigame†, Kazuyo Tanaka††, Shi-Wook Lee†††
†IwatePrefectural University, Faculty of Software,††Tsukuba University, †††AIST

ータは AM・LM の学習用とは異なるデータである．

4.2. 評価実験

4.2.1. 2種のサブワード検索結果統合

表 1 内で示した単体検索結果から 2 種を選び統合した．(1)式の重みは今回の実験では双方均等に 0.5 を与えた．

表 2 には精度が高かった上位 4 通りの組み合わせを，表 3 には先行研究[2]の精度（AM に JNAS を学習したもののみを用い，LM を変更した結果を統合）を示す．

表 2：2 種の検索結果統合時の検索性能

結果 1 (SW:AM:LM)	結果 2 (SW:AM:LM)	MAP (%)
mono:JNAS:JNAS	tri:CSJ：CSJ	73.23
mono:JNAS:CSJ	tri:CSJ:CSJ	73.17
tri :JNAS :CSJ	tri:CSJ:CSJ	72.53
tri : CS J :JNAS	tri:CSJ:CSJ	72.41

表 3：先行研究の検索性能

結果 1 (SW:AM:LM)	結果 2 (SW:AM:LM)	MAP (%)
mono:JNAS:JNAS	mono:JNAS:CSJ	57.48
tri :JNAS :JNAS	tri :JNAS:CSJ	66.78

表 2 と単体の性能(表 1)を比較すると，検索結果を統合することにより単体時よりも検索精度が向上し，単体時の最良値から最大 3.15%精度が改善した．表 3 の全ての組み合わせに CSJtriphone を用いており，CSJtriphone 導入の効果は大きいと判断できる．

mono:JNAS:JNAS と tri:CSJ:CSJ を統合した結果が最も高い精度となった．異なるコーパスを学習，異なるサブワードモデルの結果を統合したことで，多様な結果が相補的に働いたためと推察する．また JNAS は学習データ量が CSJ より少なく，JNASTriphone の AM,LM の学習(特に LM の学習)が十分に行えなかった点が要因と考える．

表 3 の先行研究と比較すると,単体時の最良値から最大で 6.45%の精度改善となった．この結果から，本提案手法が精度改善に有効であったと評価できる．

4.2.2. 3種の検索結果統合

3 種の検索結果を統合した結果を表 4 に示す．

単体での精度上位 3 種と，2 種の統合で最も高い精度となった mono:JNAS:JNAS と tri:CSJ:CSJ に，単体で高い精度だった tri:JNAS:CSJ を追加して統合した．(1)式の重みは今回の実験では全て均一に 0.33 を与えた．表中では JNAS を J，CSJ を C と表記する．

今回実験した二通りでは，どちらも 2 種の検索性能の最良値 73.17 を上回っており，最大 1.07%の精度改善となった．上位 3 種を統合した方が高い精度となっており，単体の検索精度が高いモデルを用いた方が精度向上への寄与が大きいと考える．

表 4：3 種検索結果統合時の検索性能

結果 1 (SW:AM:LM)	結果 2 (SW:AM:LM)	結果 3 (SW:AM:LM)	MAP (%)
tri:C:C	tri:J:C	tri :C:J	74.30
tri:C:C	tri:J:C	mono:J:J	73.95

5. おわりに

本稿では，複数の音響モデル，言語モデルを用いて得られた複数の検索結果を統合する方法を提案し，実験を通して提案方式により STD の精度向上を確認できた．

本稿で用いたサブワードは monophone と triphone のみであるため，今後 1/2 音素，1/3 音素並びに SPS などのサブワードを用いることによって更なる精度改善を図りたい．重みの動的決定法は今後の課題とする．

参考文献

[1]岩田耕平他, サブワードを用いた音声文書検索における複数サブワードの統合, 電気情報通信学会技術研究報告, 107 巻 116 号, pp.13-18, 2007

[2]小野寺悠二他, 複数のサブワード・言語モデルを用いた音声中の検索語検出の高精度化, 第 4 回音声ドキュメント処理ワークショップ講演論文集, 2010

[3] 伊藤慶明他,音声中の検索語検出のためのテストコレクション構築 -中間報告-, 情報処理学会研究報告, Vol.2009-SLP-78 No.4, 2009.

[4]名取賢他,任意語彙発話音声検索のための複数の認識モデルを利用した音節遷移ネットワークの構築,日本音響学会, 2009 年秋季研究発表会講演論文集, 1-R-27, pp.205-206,2009.9