情報処理学会論文誌 Vol. 50 No. 2 524–533 (Feb. 2009)

語彙制限のない音声文書検索における複数サブワードの統合
——検索語彙に依存した検索性能推定指標の導入

伊藤慶明[†1] 岩田耕平[†1] 石亀昌明[†1]
田中和世[†2] 李時旭[†3]

マルチメディア環境やハードディスクレコーダの普及にともない，ビデオに代表される音声文書データが大量に蓄積されるようになり，容易に検索できる機能が求められている．検索する際の検索語は特殊な言葉が用いられることが多く，検索語の語彙を制限しないことが望ましい．我々はサブワードを用いた語彙制限のない音声文書検索システムの実現を目指し，新しいサブワードモデルと，サブワードモデル間の統計的音響距離を用いた新しい検索方式を提案し，その有効性を検証した（岩田ら，2007）．サブワードとしては，音声認識で一般的な monophone，triphone モデルに比べ，音素を時間軸上で精緻化した 1/2 音素モデル，1/3 音素モデル，Sub-Phonetic Segment モデルの方が音声文書検索性能において優位であった．本論文では，これらの複数のサブワードによる検索結果の統合方式の検討を行い，検索性能の向上を図る．まず，複数サブワードの結果を単純に線形統合する方式を提案する．次に，検索語によりサブワードの検索性能が異なることから，与えられた検索語の検索性能を，検索語に含まれるサブワードモデル系列の認識性能から推定し，検索性能推定指標として導入し，統合時の結合重みとして利用する方式を提案する．日本語話し言葉コーパスを用いた音声文書検索実験を通して，複数のサブワードモデルの検索結果の単純な統合によって，平均適合率が 4.2%向上するのを検証できた．統合時に検索語の検索性能推定指標を導入し結合重みの設定を実験的に最適化すると平均適合率が 7.4%向上した．検索性能推定指標を結合重みに直接利用することで重み設定を自動化した場合でも 5.3%向上するのを確認でき，提案方式の有効性を検証できた．

Integration of Plural Subword Models for Open Vocabulary Spoken Document Retrieval —Introducing a Pre-estimated Index of Retrieval Performance according to Each Query

Yoshiaki Itoh,[†1] Kohei Iwata,[†1] Masaaki Ishigame,[†1]
Kazuyo Tanaka[†2] and Lee Shi-wook[†3]

Due to the spread of multi-media environments and digital hard disc recorders, recently large quantities of spoken document, such as video data, are being stored. Therefore, a new function to retrieve this spoken data is needed. Because the query words used for information retrieval are often special terms, the query words should not be restricted to words available in the dictionaries of speech recognition systems. We have been developing an open vocabulary spoken document retrieval system based on subword models, and have already proposed new subword models and an acoustic distance measure between these models. We demonstrated the effectiveness of those approaches (Iwata, et al., 2007). The experimental results showed that the new subword models worked better than the monophone and triphone models that are the general models used in speech recognition. The new models are more sophisticated than the triphone models in the time axis, such as the half-phone model, one-third phone model and Sub-Phonetic Segment model. This paper investigates a method for integrating the results obtained from these plural subword models. We first propose the introduction of a pre-estimated index of retrieval performance according to each query word, because the retrieval performance of a given query depends on the type of subword used. The simple linear integration of the plural results improved the retrieval performance 4.2% at average precision in experiments for spoken document retrieval using the Corpus of Spontaneous Japanese. The introduction of the pre-estimated index of retrieval performance according to each query word improved the retrieval performance 7.4% at the maximum average precision when integration weighting factors were optimized in the experiments. An improvement of 5.3% was achieved when weighting factors were automatically determined by directly applying the proposed pre-estimated index to the weighting factors.

†1 岩手県立大学
Iwate Prefectural University
†2 筑波大学
University of Tsukuba
†3 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology

1. はじめに

近年，パソコンやハードディスクレコーダの普及にともない，ビデオに代表される音声文書データは容易かつ頻繁に扱われるようになった．ハードディスクの大容量化は，ビデオ・音声文書データの長時間保存を可能とした．今後は長時間録画・録音されたデータの中から自分の興味ある部分を見たい・聞きたいというニーズが高まり，音声文書検索機能が不可欠となると考えられる．電子的なテレビの番組表からタイトルおよびトピックを知ることは可能であるが，詳細な内容は通常入手できず，見たいアーティスト名等のキーワードが話された部分を検索・特定することは現状では困難である．ユーザが見たい・聞きたい特定の区間を検索するためには，テキストまたは音声で検索語（単語やフレーズ）を入力して音声文書を検索する（SDR: Spoken Document Retrieval）方法がユーザにとっては最も簡便な方法と考えられる．音声文書検索の方法としては，音声認識の結果を利用する方法が代表的である[2)–6)]が，音声認識システムの辞書に登録されていない語句の検索は一般に困難である．一方，新しい言葉や，人名/地名，技術専門用語等の特殊な単語は検索語となりやすいが[1)]，このような固有名詞は音声認識システムの辞書に登録されていない場合が多い．このため我々はサブワード[⋆1]を用いて語彙制限のない音声文書検索システムの提案を行った[1)]．単語 Confusion Network 中に音素列を埋め込み辞書に登録されていない単語にも対応する方式等[7),8)]も提案されているが，本方式はサブワードのワンベストの認識結果だけを利用するもので，複数のサブワードを用いてもコンパクトに構成することが可能である．

これまでに我々は，音声文書検索に用いるサブワードとして，音素より時間的に精緻にモデル化を行った環境依存の 1/2 音素，1/3 音素，および SPS[10)] とそれぞれのモデルの構築方法を提案するとともに，これらのサブワードを構成している HMM の統計量を用いてサブワードモデル間の音響的な近さ（音響距離）を求めておき，照合時にサブワードモデル間距離として利用することを提案し，これらの方法の有効性を確認した[1),12)]．環境依存の 1/2 音素，1/3 音素とは前後の音素を考慮したうえで，各音素を時間軸方向にそれぞれ 2 つ，3 つに分割したモデルである．各モデルは同一状態数の HMM で構成しているため，音素モデルよりもそれぞれ時間軸上で 2 倍，3 倍に精緻化したモデルとなる．

これまでの研究では monophone，triphone，1/2 音素，1/3 音素，および SPS 等をサブワードとして用い，実験に用いた複数の検索語に対する平均性能で個々のモデルを評価し，新しいモデルの優位性を確認した．平均性能が高いサブワードがどの検索語においてもつねに優位な性能を示すわけではなく，検索語によっては性能が逆転することもあった．このような結果をふまえ，複数のサブワードの検索結果を統合することにより検索性能の向上が図れると考えた．本論文ではまず，複数のサブワードで並行に検索を行い，その結果を線形統合する方式を提案し，性能の向上を確認する．次に線形統合の際，各サブワードの結果に対して重みを与える必要があるが，これを次に述べる検索語の検索性能推定指標より自動的に設定する方法を提案し，その有効性を検証する．用いるサブワードによって検索語ごとに性能が異なることから，あるサブワードについて検索しやすい/検索し難い検索語があると考えられる．検索しやすい検索語とは，認識しやすい言葉すなわち認識しやすいサブワードモデルの系列であると仮定し，認識しやすいサブワードモデルの系列であれば，音声文書中でも高い精度でサブワード認識でき，検索語のサブワードモデル系列と一致しやすいと考える．そこでまず各サブワードの各々のサブワードモデルの認識性能を事前に求めておく．検索の際，検索語が与えられると，検索語は変換ルールに基づきサブワードモデル系列に変換され，この検索語サブワードモデル系列に対するサブワードの平均認識率を求める．本論文ではこの平均認識率を検索語検索性能推定指標と呼び，検索語サブワードモデル系列の認識のしやすさ・検索のしやすさを表すものとする．複数のサブワードの検索結果を統合する際，各サブワードの結果に付与する重みにこの検索語検索性能推定指標を直接利用することで重み設定を自動化し同時に検索性能の向上を図る．

本論文では，日本語話し言葉コーパスを用いて検索性能の評価実験を行い，複数のサブワードの検索結果を統合することの有効性と，検索語検索性能推定指標を直接利用する統合方式の有効性を検証する．

2 章で本音声検索方式に用いる複数のサブワードとサブワードモデルの構成方法，検索結果の統合方法についての説明を行い，3 章で各統合方式における検索性能について実験および考察を行う．最後にまとめを述べる．

2. 複数サブワードの統合方式

我々が提案する音声文書検索システムでは，サブワードレベルでの照合を行うことによって，単語辞書に依存しない語彙制限のない検索を実現する．音声文書に対してはあらかじめサブワード認識を行い，サブワードモデル列に変換しておく．サブワード認識とは，通常の連続音声認識における認識の単位である単語をサブワードモデルに置き換えて認識する方

⋆1 本論文では，原則として monophone，triphone 等をサブワード，monophone における a，triphone における a-k+i 等の個々のモデルをサブワードモデルと呼ぶことにする．

式で，認識の際の言語モデルにはサブワードモデルのバイグラム，トライグラム等を利用する．検索語がテキストまたは音声で与えられると，検索語をサブワードモデル系列に変換（音声の場合はサブワード認識）し，検索語のサブワードモデル系列と全音声文書サブワードモデル系列との照合を行い，類似度の高い（照合距離が小さい）順に候補として出力する．

本章では，本論文で用いるサブワードと，複数のサブワードの検索結果を単純に統合する方式および，統合の際必要となる結合重みに検索語の検索性能推定指標を直接用いる方式を提案する．

2.1　サブワード

本論文では monophone，triphone，1/2 音素，1/3 音素，および SPS の 5 つのサブワードを用いる（以降の表中ではそれぞれ mono，tri，1/2，1/3，SPS と表記する）．本節ではそれらの各々のサブワードについて簡単に解説する．本論文では triphone，1/2 音素，1/3 音素については，学習データに出現した物理モデルで実験を行った．検索語に新出のモデルが含まれる場合については対応する物理モデルを定義する等の対応が必要になるが，今回の実験には含まれていない．

(1)　**monophone**：音素環境非依存の音素モデル．今回のモデル数は 43.

(2)　**triphone**：現在の音声認識システムでは最も代表的なモデルで隣接する音素の影響を考慮した音素モデル．実際に音声として存在するのは約 8,000 であるとされる[13]が，今回のモデル数は学習データ中に出現した 7,956.

(3)　**1/2 音素**：1 つの triphone（音素）を前後の音素環境を考慮しながら時間的に前半と後半の 2 つの 1/2 音素に分割したモデル．今回のモデル数は学習データ中に出現した 1,333.

(4)　**1/3 音素**：1/2 音素モデルと同様，前後の音素環境を考慮しながら 1 つの triphone を時間的に前半，中心，後半部と 3 つの 1/3 音素に分割したモデル．今回のモデル数は 1,374（1/2 音素のモデル数 + 音素数）.

(5)　**SPS モデル**：SPS は国際音声記号（IPA）に準拠した ASCII コードである XSAMPA をベースにして，この音声記号に対して音響物理的特性を考慮して分割したサブ音声セグメント符号系である[10]．1 つの音素を音響特性に応じて中心部および音素の渡り部分で別々にモデル化を行うことで詳細なモデル化がなされる．モデル数は 423.

図 1 に音素列 “k a p” について各サブワードにおける表記を記す．図中#は前後に隣接する単語の末尾あるいは先頭の音素を表す．monophone および triphone は 3 つのサブワードモデルから構成されるのに対し，1/2 音素は 6，1/3 音素では 9 のサブワードモデルで表現される．本研究では各サブワードモデルはすべて同一状態数からなる HMM でモデル化するため，必要なサブワードモデル数が多いほど，時間的に精緻なモデル化がなされているといえる．音節や音素単位のサブワードモデルを用いた場合，1 つのサブワードモデル認識誤りや脱落誤りが照合結果に与える影響は大きくなる．一方，時間的に精緻なサブワードでは，冗長性が高いため 1 つのサブワードモデルの認識誤りや脱落の影響は大きくならずにすむ．ワードスポッティング的な音声文書検索においては時間整合性が重要であり，1/2 音素や 1/3 音素のように冗長性の高いサブワードが有効に働くと考える．

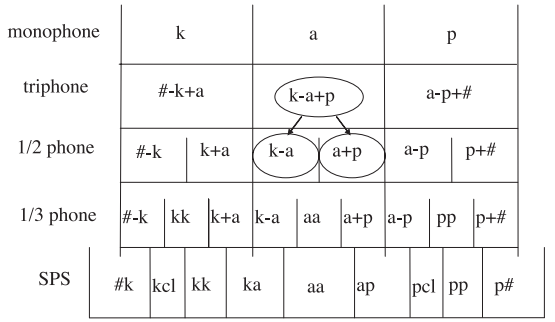

図 1　音素列 “k a p” の各サブワードによる表記例
Fig. 1　An example of the descriptions for each subword model as for a phone sequence “k a p”.

表 1　各サブワードのモデル数および認識性能
Table 1　The number of models and recognition performance for each subword.

サブワード	モデル数	PP	精度	置換	脱落	挿入
mono	43	7.91	70.80	19.17	7.41	2.62
tri	7,956	4.73	46.03	41.29	1.52	11.16
1/2	1,333	2.97	53.89	32.42	1.58	12.10
1/3	1,374	2.02	65.22	26.04	3.01	5.73
SPS	423	2.65	73.93	18.98	2.82	4.26

（PP は perplexity，精度，置換率，脱落率，挿入率は%）

参考までに**表 1** に各サブワードのモデル数，学習セットデータのサブワード言語モデルの perplexity，正解精度とそれぞれの誤認識率を示す[⋆1]．モデル間で認識の単位（各サブワードモデルの継続時間），モデル数が異なるため公平な比較分析は難しい．1/3 音素は 2 番目

⋆1 本実験結果は 3.1 節におけるサブワードモデルの accuracy を求めた際に得られたものである．

のサブワードモデルが一意に決まるため実質的には 1/2 音素と同程度の perplexity と考えられる．評価セットの perplexity は monophone で 17.4，triphone で 46.4 となり表中の性能差が推察できる．

2.2　複数サブワードの検索結果統合方式

本節では複数のサブワードの検索結果を統合する方式について詳述する．

音声文書検索において，単語単位の認識結果とサブワード単位の認識結果を複数統合することで検索性能の向上を図る手法が提案されている[7]．本研究においては複数サブワードの結果を統合することによる検索性能向上を目指す．これまでの研究結果より，検索語ごとに検索性能が高いサブワードと低いサブワードが存在し，平均性能の優劣が逆転する場合もあることから，与えられた検索語に対してサブワードごとの認識のしやすさを，検索語のサブワードモデル系列の平均認識性能（検索語検索性能推定指標）として予測し，統合時の結合重みとして利用する方式を提案する．

実験条件で述べるが，評価は検索語が話された発話区間が特定できるかで行っている．サブワードの種類ごとに検索を行った結果，以下の 2 つが得られる．

(1)　音声文書内の各発話区間のサブワードモデル系列に対し，検索語のサブワード系列が最も対応する部分区間（サブワードモデル系列）

(2)　各発話区間の DP 正規化距離の最小値（上記の対応区間の距離）

同一発話区間内における上記の部分区間は 2 種類のサブワードで位置に違いが生じていることもあるが，本論文における複数サブワードの結果の統合ではこれらの情報は用いず，同一の発話区間に対する DP 正規化距離の最小値を用いその線形結合によって統合する．

ここで，ある候補区間に対し i 番目のサブワードにおける DP 正規化距離，そのサブワードの距離に付与する重みをそれぞれ $dist(i)$，$weight(i)$ とすると，当該候補区間の統合距離 D は以下の式 (1) で求める．なお，統合するサブワードの種類数は N 種類とする．

$$D = \sum_{i=1}^{N} weight(i) \times dist(i) \tag{1}$$

$$\text{ただし，} \sum_{i=1}^{N} weight(i) = 1 \tag{2}$$

(1)　**単純線形結合方式**

各サブワードの距離に付与する重み $weight(i)$ は事前に定めておき式 (1) により線形結合することで統合する．重みは検索語が変わっても変化させず固定値を用いる．

(2)　**検索語検索性能推定指標の優劣を用いた統合方式**

検索語に含まれるサブワードモデル系列から，これらのサブワードモデルの平均認識性能を表す検索性能推定指標を算出する．検索性能推定指標はサブワードおよび検索語に依存し，この指標が高いサブワードを重視して用いた方が有利として，より大きな重みを与える．

まず，学習データとは異なる音声データに対して各サブワードで連続サブワード認識を行い，その結果からすべてのサブワードにおける各サブワードモデル s についての認識性能 $accuracy(s)$ をあらかじめ算出・確定しておく．与えられた検索語 q は N_q 個のサブワードモデル（$s_1, s_2, \ldots, s_{N_q}$）を含むとすると，サブワード i における検索語 q の検索性能推定指標 $prfmnc'(i,q)$ は，式 (3) のように N_q 個のサブワードモデルの認識率の積を N_q 乗根して求める．この指標は各サブワードモデルの認識性能の平均値とすることもできるが，この認識率が HMM の尤度と比例関係にあると仮定し，単語全体の尤度は個々の尤度の積になることから，今回は N_q 個のサブワードモデルの認識率の積を N_q 乗根して求めることとした．

$$prfmnc'(i,q) = \sqrt[N_q]{\prod_{k=1}^{N_q} accuracy(s_k)} \tag{3}$$

異なるサブワードではモデル数および認識の単位（1 つのモデルあたり継続時間）が異なるため，表 1 に示したように平均サブワードモデル認識率に差が生じる．式 (3) で得られた $prfmnc'(i,q)$ を式 (4) により平均サブワード認識率で正規化することでサブワード i における検索語 q の正規化検索性能推定指標とする．式中 $ave\_accuracy(i)$ はサブワード i の全サブワードモデルの平均認識率を示す．

$$prfmnc(i,q) = \frac{prfmnc'(i,q)}{ave\_accuracy(i)} \tag{4}$$

正規化した検索性能推定指標をサブワードごとに求め，指標の大きいサブワードから順に大きい重みを与えていく．本項の方式では重みは事前に設定しておく（ex. 3 つのサブワードを統合する場合，0.6，0.3，0.1 等と設定しておき，指標の大きいサブワードの結果に大きい重みを与える）．

(3)　**検索語検索性能推定指標に基づく重み設定方式**

本項では，検索語に対してサブワードごとの認識のしやすさを表す正規化検索性能推

定指標を式 (1) の結合重みに直接反映することで，重み設定の必要ない統合方式を提案する．与えられた検索語に対するサブワードごとの正規化検索性能推定指標を式 (4) によって求め，サブワード i の結果に与える重みは，式 (5) に示すように他のサブワードの正規化検索性能推定指標との割合として直接利用する．割合としているため，式 (2) の条件は満足される．

$$weight(i,q) = \frac{prfmnc(i,q)}{\sum_{j=1}^{N} prfmnc(j,q)} \tag{5}$$

3. 評価実験

3.1 学習用・評価用音声データと HMM

monophone，triphone，1/2 音素，1/3 音素，SPS の 5 種類のサブワードを用いて評価を行う．各サブワードモデルは Left-to-right 型，3 状態 16 混合の同一のモデル構成とする．triphone の状態数が多いため，triphone についてのみ 2,000 状態に状態共有を行った．音響モデルは新聞記事読み上げ音声コーパス JNAS[11] に収録されている男性話者 153 名，約 150 文を用いて学習を行った．サブワードモデルの言語モデルは Palmkit を利用し，JNAS のテキストデータをサブワードモデル系列に変換しサブワードモデルのバイグラム，トライグラムを構築した．サブワード認識の際に用いる辞書は各サブワードモデルを単語として考え，すべてのサブワードモデルで構成する．たとえば SPS の場合，辞書の語彙は SPS のサブワードモデル数の 423 個となる．

検索性能推定指標を得るための各サブワードモデル s の認識性能 $accuracy(s)$ は学習データとは異なる音声コーパス ETL-DB[15] を用いてあらかじめ算出した．

評価用の検索対象となるデータは，日本語話し言葉コーパス（CSJ）[14] に収録されているコアと呼ばれる 49 講演，約 13 時間分の連続音声データを用いた．

検索語と音声 DB との連続 DP による照合の際には，サブワードモデル間距離としてサブワードモデル間の統計的な音響距離（音響的な近さ）を用いる[1]．2 つのサブワードモデルの HMM において同一位置の状態からそれぞれ 1 つずつ分布を取り出し分布間距離を求め，その中で最小となる分布間距離をその状態間の距離とし，すべての状態間の距離の平均をサブワードモデル間の音響距離として定義する．

3.2 実験条件

講演音声に対しパワーを用いて無音区間が 300 ms 以上続いたときに音声区間を区分化し，その音声区間に対して事前にサブワード認識を行っておく．音声の分析条件は**表 2** に示す．分析時のフレームシフトは 10 ms または 5 ms とした．時間的に精緻なモデルに関しては，10 ms では十分な学習データが確保できず性能が低下することが予備実験によって確認されたため，1/2 音素，1/3 音素，および SPS の 3 種類のサブワードに対しては 5 ms フレームシフトとした．一方音素モデルの monophone および triphone ではフレームシフトを 5 ms とすると状態数を倍にした場合でも性能が低下したため 3 状態のまま 10 ms とした．

CSJ を検索するための検索語は，各講演を特徴付ける専門用語のうち，検索対象データに複数回（3 回〜50 回）出現する語句 50 個を抽出した．**表 3** に実験に用いた検索語の例とその出現回数を示す．なお，本システムでは検索語をテキストおよび音声で与えることが可能であるが，本実験では検索語は音声ではなくすべてテキストで与える．

音声文書検索ではユーザは検索語が話された区間を検索し，聴取・確認すると想定し，区分化された発話区間に検索語が含まれていれば正解と見なす．このため区間中の検索語の位置整合については評価していない．

表 3 に示したとおり，検索語ごとに出現回数が異なり，総合評価としての precison-recall グラフが描けないため，次式に示す平均適合率（Average precision）を評価尺度として用いる．

表 2　音声分析条件
Table 2　Speech analysis conditions.

Sampling	16 KHz，16 bits
Feature vector	MFCC（12 dim）+ Δ MFCC（12 dim）+ ΔLogPow
Window	Hamming window
Window length	256 points（16 msec）
Frame shift	10 msec（monophone，triphone） 5 msec（1/2 音素，1/3 音素，SPS）

表 3　検索語ごとの出現回数（回）と平均適合率（%）(サンプル)
Table 3　Appearance frequency (times) and average precision rate (%) for each query (5 samples).

検索語	出現回数	サブワード				
		mono	tri	1/2	1/3	SPS
マスク値	8	81.34	87.90	100.00	89.17	100.00
周波数伸縮	8	75.71	64.08	87.95	66.67	90.89
ニュース	48	56.24	30.75	38.09	64.95	48.25
セグメント統計量	3	67.33	57.60	79.17	100.00	66.76
声道長	9	39.68	19.90	19.94	50.66	75.80

$$\text{precision}(k) = \frac{\sum_{i=1}^{k} \delta_i}{k} \tag{6}$$

$$\text{Average precision} = \frac{1}{D_q} \sum_{k=1}^{R_q} \delta_k \times \text{precision}(k) \tag{7}$$

検索語 q についての候補を類似度の高い（DP 正規化距離が小さい/検索性能推定指標が大きい）順に並べ，第 k 位の候補が正解区間である場合は $\delta_k = 1$，不正解区間である場合には $\delta_k = 0$ とする．したがって式 (6) の precision(k) は，k 位まで候補を出力したときの precision となる（分子は正解数となる）．検索語 q に適合する正解区間数（正解数）を D_q，正解が最後に現れた順位を R_q とすると，平均適合率を表す式 (7) の右辺は $\delta_k = 1$ すなわち正解が出現する D_q 回の precision の平均値となる．たとえば正解区間が 3 つある検索語の場合，3 つの正解区間が検出されたときの 3 つの precision 値を平均したものになり，湧き出し誤りの比率を含めた指標となっている．

3.3　実験結果と考察

本節では，個々のサブワードの検索性能を示すとともに用いるサブワードの種類数を $N = 2$〜5 として 2.2 節に述べた統合方式各々について評価を行い，その結果と考察について述べる．

3.3.1　単一サブワードを用いた結果

表 4 に単一サブワードによる検索性能結果を示す．1/2 音素，SPS モデルの結果が CSJ の検索においては性能が高かった．triphone の性能が高くならなかったのは，モデル数が多く，今回用いた学習データのみからでは音響モデル/言語モデルが十分でなかったと推察する．検索語および話者によって検索性能の差が大きかった．特に検索語については同音異義語や類似発音語句の頻出（ex. “尤度” と “粒度”）が原因と考えられる．

表 4 の右側が示すように，検索対象となる音声データに対し，1 位の候補（DP 正規化距離が最小）の正解率はすべてのサブワードにおいて 70%を超えた．表 3 に示すように，各サブワードの平均適合率の優劣が検索語により逆転することから，サブワードにより得意な（検索精度の高い）検索語と苦手な（検索精度の低い）検索語が存在すると想定できる．これにより，複数のサブワードの結果を統合する手法が有効に働くと考えられる．

表 4　単一サブワードを用いた CSJ 検索性能
Table 4　Retrieval performance for CSJ using single subword.

サブワード	平均適合率	1 位の候補の正解率（%）
mono	38.6	82.0
tri	34.0	72.0
1/2	67.4	98.0
1/3	53.5	88.0
SPS	60.2	96.0

3.3.2　単純線形結合方式の結果

まず，単一サブワードを用いた評価における上位 2 つのサブワード（1/2 音素と SPS）を単純線形結合した場合の検索性能を**図 2** に示す．結合重みは検索語によらず固定し，その値を 0.0 から 1.0 まで 0.1 刻みで変化させた．結合重み 1.0 および 0.0 はそれぞれ 1/2 音素，SPS モデルのみを用いた結果と同一である．図より 1/2 音素と SPS の 2 種類のサブワードを統合することによって検索性能が向上し，重みをそれぞれ 0.6，0.4 に設定したとき最も高い性能が得られた．

表 5 に統合するサブワードの種類数 N = 2，3，4，5 における複数サブワードの統合結果を示す．各サブワードの結果に与える重み $weight(i)$ を 0.1 単位で変化させてすべての重みとすべてのサブワードの組合せを評価し，平均適合率が最大となったサブワードと重みの組合せを示した．単一のサブワードを利用するよりも最大で 4.3%の性能の向上が得られた．単一のサブワードで性能が高いモデルの重みを大きくすると性能が高くなる傾向が見られた．本方式においては，2 つ以上のサブワードに対し適切な重みを設定することにより，検索性能が向上することを確認できた．一方で，N = 3，4，5 においては重みが 0 となる（利用されない）サブワードも存在しており，計算量と性能向上幅の観点から統合すべきサブワードを限定した方が実用的であると考える．

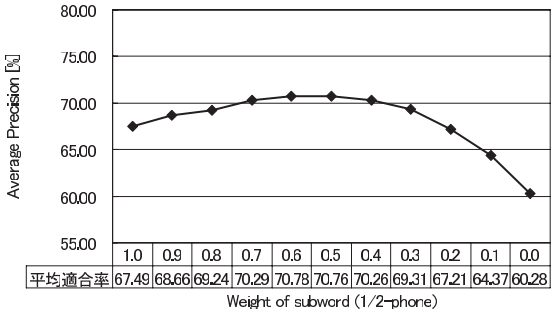

図 2　1/2 音素と SPS の線形結合による検索性能
Fig. 2　Retrieval performance by linearly integrating the 1/2 phone model and the SPS model.

表 5　線形結合による検索性能（N = 2，3，4，5 における最良の検索性能を示した組合せのみ掲載）
Table 5　Retrieval performance by linear integration (printed for a combination of the best retrieval performance at N = 2, 3, 4, 5).

N	利用するサブワード					平均適合率
2	1/2	SPS				70.78
	0.6	0.4				
3	1/2	1/3	SPS			71.66
	0.5	0.2	0.3			
4	mono	1/2	1/3	SPS		71.76
	0.1	0.5	0.2	0.2		
5	mono	tri	1/2	1/3	SPS	71.76
	0.1	0.0	0.5	0.2	0.2	

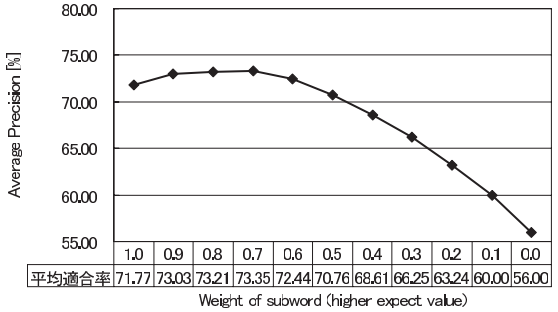

図 3　1/2 音素と SPS の検索語検索性能推定指標の優劣を用いた結合方式による検索性能
Fig. 3　Retrieval performance by integrating the 1/2 phone model and the SPS model based on superiority of a pre-estimated index of retrieval performance according to each query.

3.3.3　検索語検索性能推定指標の優劣を用いた統合方式の結果

図 3 に，前節と同様 1/2 音素と SPS を用いた場合の検索性能を示す．結合重みを 1.0 としたときは，検索語に対する検索性能推定指標が高い方のサブワードの結果のみを選択・利用する．このときの平均適合率は 71.77 となり，単純線形結合方式で同じサブワードを用いた場合の最良値より良くなった．重み 0.0 のときは検索性能推定指標の低いサブワードを選択することになり予想どおり性能は低下した．図から分かるように検索性能推定指標の高いサブワードのみを用いるよりも 2 種類のサブワードの結果を統合した方が性能が向上した．本方式において 1/2 音素と SPS を統合する際は，検索性能推定指標の高いサブワードの結果に重み 0.7 を与えたとき最も良い性能を示し，平均適合率は 73.35 となった．

表 6　検索語検索性能推定指標の優劣を用いた統合方式による検索性能（N = 2，3，4，5 における最良の検索性能を示した組合せのみ掲載）
Table 6　Retrieval performance by superiority of a pre-estimated index of retrieval performance according to each query (printed for a combination of the best retrieval performance at N = 2, 3, 4, 5).

N	利用するサブワード	検索性能推定指標順の設定重み	平均適合率
2	1/2，SPS	0.7 0.3	73.35
3	1/2，1/3，SPS	0.7 0.2 0.1	75.03
4	mono，1/2，1/3，SPS	0.7 0.2 0.1 0.0	74.89
5	mono，tri，1/2，1/3，SPS	0.4 0.4 0.1 0.1 0.0	72.48

検索語が与えられると各サブワードの検索性能推定指標を求め，その値が高いモデル順に大きい設定重みを付与する．

表 7　検索語検索性能推定指標に基づく重み設定方式による検索性能
Table 7　Retrieval performance by a pre-estimated index of retrieval performance according to each query.

N	利用するサブワード					平均適合率
2	1/2	SPS				72.92
3	1/2	SPS	1/3			73.12
4	1/2	SPS	1/3	mono		72.86
5	1/2	SPS	1/3	mono	tri	70.95

表 6 に N = 2，3，4，5 における複数サブワードの統合結果を示す．表中の「検索性能推定指標順の設定重み」は指標が高いサブワード順に左側から大きな重みを与えることを示している．重みを 0.1 単位で変化させすべての重み，すべてのサブワードの組合せで評価し，平均適合率が最も良くなった組合せを示している．多くの組合せにおいて検索語検索性能推定指標を導入することにより，単純線形結合方式より高い性能を得ることができた．1/2 音素，1/3 音素，SPS を検索性能推定指標の高い順に重みを 0.7，0.2，0.1 に設定したときが最も高い性能となり，平均適合率は 75.03 で，単一サブワードを用いるより 7.4%向上した．単純線形結合方式よりも高い性能が得られており，検索語検索性能推定指標を導入する有効性を確認できた．統合するサブワードの種類を 4 つ以上にする（重みを 0.1 以上与える）と，性能の低いモデルまで利用することになり性能の低下が見られた．

3.3.4　検索語検索性能推定指標に基づく重み設定方式の結果

表 7 に N = 2，3，4，5 における本提案方式の検索性能を示す．検索語検索性能推定指標の優劣で評価する前節の方式の場合と同様，1/2 音素，1/3 音素，SPS の 3 種類のサブワードを統合した場合に最も良い性能を示し，平均適合率は 73.12 となった．単一サブワードを用いるより 5.3%性能が向上した．前節の方式と比較すると検索性能は若干低かったが，

たとえば N = 3 の場合，検索語検索性能推定指標の優劣で評価する統合方法では 66 通りのサブワード・結合重みの組合せを評価しその最良値であるのに対し，本統合方式では重みは自動的に設定される．今回は各サブワードの検索性能の推定指標の割合を用いたため，検索性能の低いサブワードの結果も検索結果に影響を与えてしまったと考えられる．前項の方式では，検索性能推定指標の小さいサブワードに 0.0 に近い重みを与えるようになっており，検索性能推定指標の高いサブワードに大きな重みを与えるように，あるいは検索性能推定指標の低いサブワードの結果は利用しないように重み設定方法を変更することにより，検索性能はさらに向上すると考える．

図 4 に検索語ごとの precision-recall グラフの例を示す．上図は「合成音声」，下図は「エイチエムエム」を検索語とした結果である．triphone と monophone を比較すると検索語「合成音声」に対しては triphone が優位であるが，「エイチエムエム」では monophone の方が優位になっている．Proposed は 3.3.4 項で示した重み設定を自動化する提案手法で，上述したように総合的にはサブワード単体で用いるより提案手法が優位であると確認できるが，部分的にはサブワード単体で用いる方が良い場合もあり，改善の余地があると考える．

今後の課題として，3.3.3，3.3.4 項では検索語 50 語に対する結合重みの検討を行ったが，検索語の音韻的なバランスにより重みが左右される可能性があり，Leave-one-out 等の cross-validation で評価し，上述したようなサブワードの取捨選択を含む最適な重み自動設定方式を検討したいと考える．

4. おわりに

本論文ではサブワードを用いた語彙非依存の音声文書検索システムの性能向上を目指し，一般的な monophone，triphone のほか，すでに有効性を確認した 1/2 音素，1/3 音素，および SPS の 5 つのサブワードによる検索結果を統合する方式を提案した．検索語が与えられた際にその検索語の検索性能の推定指標を求め，その推定指標を検索結果統合の際に利用することで性能向上を図った．CSJ を用いた検索実験の結果，複数サブワードを利用することで性能向上を確認するとともに検索語の検索性能推定指標の導入により，統合の際に重み等のパラメータ設定することなく性能向上を実現できた．今後は統合時の最適な重みの自動設定方法の開発および，複数の検索語による検索について研究を進める必要があると考える．

謝辞　本研究の一部は文部科学省科学研究費補助金基盤（C）No.20500096 を受けて実施された．

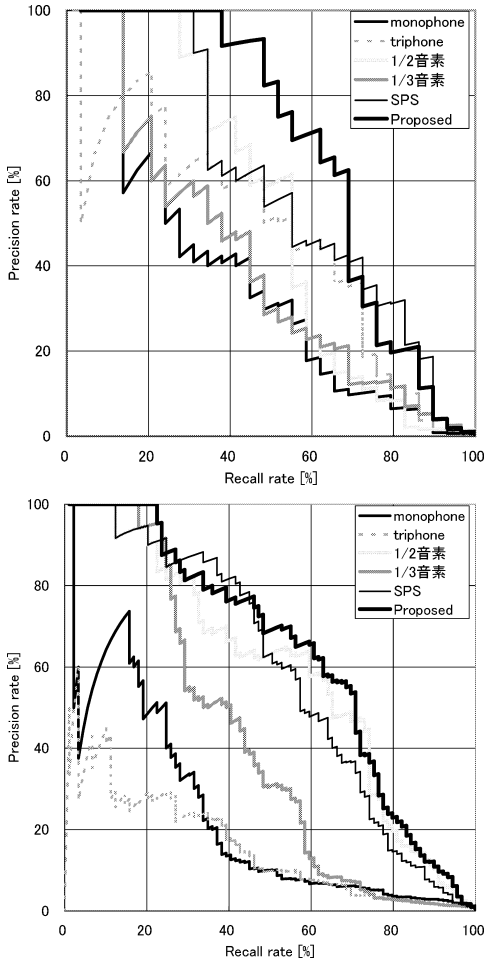

図 4　検索語ごとの precision-recall（上側：検索語「合成音声」，下側：検索語「エイチエムエム」）
Fig. 4　Precision-recall graph for each query (Upper: query of “synthesized speech”, Lower: query of “HMM”).

参 考 文 献

1) 岩田耕平，伊藤慶明，小嶋和徳，石亀昌明，田中和世，李　時旭：語彙フリー音声文書検索手法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).
2) Garofolo, J.S., Auzanne, C. and Voorhees, E.M.: The TREC Spoken Document Retrieval Track: A Success Story, *RIAO 2000*, *Content-Based Multimedia Information Access* (2000).
3) 西崎博光，中川聖一：音声キーワードによるニュース音声データベース検索手法，情報処理学会論文誌，Vol.42, No.12, pp.3173–3184 (2001).
4) 松下雅彦，中川聖一，西崎博光，宇津呂武仁：音声入力による Web 検索のためのキーワード認識・抽出法の改善，情報処理学会研究報告，2004-SLP-051 (2004).
5) 桐山伸也，広瀬啓吉，峯松信明：話題知識を導入した文献検索音声対話システム，信学論，Vol.J85-D-II, No.5, pp.863–876 (2002).
6) Miller, R.H.D., Klever, M., Kao, C., Kimball, O., Colthurst, T., Lowe, A.S., Schwartz, R. and Gish, H.: Rapid and Accurate Spoken Term Detection, *INTERSPEECH*, pp.314–317 (2007).
7) Hori, T., Hetherington, I.L., Hazen, T.J. and Glass, J.R.: Open-vocabulary spoken utterance retrieval using confusion networks, *ICASSP* (2007).
8) Moreau, N., Kim, H.-G. and Sikora T.: Phonetic Confusion Based Document Expansion for Spoken Document Retrieval, *INTERSPEECH*, Vol.2, pp.1593–1596 (2004).
9) 岡　隆一，西村拓一，張　建新，伊原正典：フレーム特徴の音素記号化に基づく語彙に依存しない音声検索，信学論，Vol.D-II, No.6, pp.764–775 (2003).
10) Tanaka, K., Kojima, H., Fujimura, N. and Itoh, Y.: Constructing speech processing systems on universal phonetic codes accompanied with reference acoustic models, *International Conference on Pattern Recognition*, Vol.III, pp.728–731 (2002).
11) Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K. and Itahashi, S.: JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research, *J. Acoust. Soc. Japan* (*E*), Vol.20-3, pp.199–2006 (1999).
12) Lee, S.-W., Tanaka, K., Fujimura, N. and Itoh, Y.: Evaluation of speech data retrieval system using sub-phonetic sequence, 日本音響学会講演論文集, 3-Q-3, pp.159–160 (2002).
13) 鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄：IT Text 音声認識システム，オーム社 (2001).
14) Maekawa, K., Koiso, H., Furui, S. and Isahara H.: Spontaneous Speech Corpus of Japanese, *LREC*, pp.947–952 (2000).
15) 速水　悟，田中和世：電総研の研究用音声データベース，音響学会誌，Vol.48, No.12, pp.883–887 (1992).

(平成 20 年 6 月 4 日受付)
(平成 20 年 11 月 5 日採録)

伊藤　慶明（正会員）

平元年東京大学大学院工学系研究科航空学専攻修士課程修了．同年川崎製鉄（株）に入社．平成 4 年より技術研究組合新情報処理開発機構に出向．平成 7 年より川崎製鉄（株）に復帰．平成 12 年より岩手県立大学助教授．博士（工学）．人工知能学会，日本音響学会，電子情報通信学会各会員．

岩田　耕平

平成 17 年岩手県立大学ソフトウェア情報学部卒業．平成 19 年岩手県立大学大学院ソフトウェア情報学研究科博士前期課程修了．現在（株）あいおい保険システムズ勤務．

石亀　昌明（正会員）

昭和 43 年東北大学工学部電子工学科卒業．昭和 49 年東北大学大学院工学研究科博士課程修了．同年東北大学応用情報学研究センター助手．昭和 53 年松下電送（株）入社．昭和 63 年秋田大学鉱山学部情報工学科助教授．平成 10 年岩手県立大学ソフトウェア情報学部教授．現在に至る．工学博士．電子情報通信学会，画像電子学会，日本音響学会各会員．

田中　和世

昭和 45 年横浜国立大学工学部卒業．昭和 46 年通商産業省電子技術総合研究所入所同研究所音声研究室長，総括主任研究官等を経て平成 13 年（組織再編により）産業技術総合研究所研究グループ長．平成 14 年図書館情報大学教授平成 14 年 10 月より現職（筑波大学教授）音声情報処理の研究に従事，共著『音声工学』(森北出版）等電子情報通信学会，日本音響学会，人工知能学会，IEEE 各会員，工学博士．

李　　時旭

平成 9 年韓国嶺南大学より M.Sc.（音声認識研究）．平成 13 年東京大学大学院工学系研究科情報通信工学専攻博士課程修了（工学博士）．同年産業技術総合研究所入所．現在，産業技術総合研究所情報技術研究部門研究員．デジタル信号処理，音声認識，マルチメディアデータ処理，等の研究に従事日本音響学会，韓国音響学会各会員．