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

岩田耕平[†] 伊藤慶明[†] 小嶋和徳[†]
石亀昌明[†] 田中和世[††] 李時旭[†††]

パソコン・ハードディスクレコーダの普及にともない，ユーザが見たい場面を容易に検索できる機能が必要不可欠となってきている．そこで本論文では，すでに収録されている音声をテキストおよび音声で検索する音声文書検索について，有効となる音響モデルとその処理方式を検証し，実際に放送されたテレビ番組に対し提案方式を適用した結果を報告する．本論文では従来の大語彙音声認識システムにおける未登録語に対処するため，音素や音声記号に相当するサブワードモデルを基本単位とする語彙制約のない検索方式を採用する．提案システムは，新しいサブワードモデルを用いる点およびサブワードモデル間の音響的な距離を利用する点に特徴がある．サブワードとして，従来の triphone モデルと，音素を時間軸上で精緻化した 2 つの新たなモデル（半音素モデル，Sub-Phonetic Segment モデル）を比較した．検索性能の比較実験により，一般的な triphone モデルの性能よりも F 値 (%) による評価において 8 ポイント以上向上し，新しいモデルの優位性を確認できた．また，サブワードモデル間の音響距離の導入により F 値が 6 ポイント以上向上し，音響距離導入の有効性も確認できた．本システムを，平成 16 年に発生した新潟中越地震の際にテレビ放送された安否確認放送に適用し，音声認識システムの辞書に登録されていない固有名詞を多く含む音声の検索実験を行った結果，システムの有用性，利用可能性を確認できた．

An Investigation of New Subword Models and Subword Phonetic Distance for Vocabulary-free Spoken Document Retrieval System

Kohei Iwata,[†] Yoshiaki Itoh,[†] Kazunori Kojima,[†]
Masaaki Ishigame,[†] Kazuyo Tanaka[††] and Lee Shi-wook[†††]

According to the recent spread of personal computers and video hard-disc recorders, a new function is needed such that it is easy for a user to identify the scene that a user wants to watch in video data. For this purpose, this paper investigates effective acoustic models for a spoken document retrieval system by a text and a speech query, and the system is applied to an actual TV broadcasted program. To deal with out-of-vocabulary words of a large vocabulary speech recognizer, we adopt a vocabulary-free spoken document retrieval method based on subword models such as phonemes and some phonetic symbols. The proposed method is characterized by using new subword models and phonetic distance between the subword models. We use two types of new models that are more sophisticated models than triphone models on the time axis. The models are called demi-phone models and Sub-Phonetic Segment models. We conducted some experiments for evaluating the retrieval performance between these models. The F-measure of proposed subword model becomes 8% higher than that of traditional triphone models. Furthermore, the introduction of phonetic distances between subword models improves the retrieval performance becomes 6% higher. We applied the system for retrieving the safety information in actual TV broadcast news of Niigata-Chuetsu earthquake in 2004, and confirm the effectiveness and possibility of the proposed system.

1. はじめに

近年，パソコンにおいてマルチメディアを扱う環境が充実し，ハードディスクレコーダ等が一般家庭に普及したことにともない，ビデオデータが容易かつ頻繁に利用されるようになった．ビデオデータは今後，従来のビデオテープに代わりハードディスク上のディ

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

ジタルデータとして保存・利用されていくと予想される．さらにハードディスクの大容量化にともないビデオデータは長時間保存することが可能となり，将来的には興味のある番組はすべて予約録画，あるいはすべての番組を録画しておき，録画データの中で見たい番組・カットのみ鑑賞するというビデオスタイルに変わっていくことが想定される．このようなビデオスタイルにおいては，長時間保存したデータの中から自分の見たい場面を検索できる機能が必要不可欠となる．電子的に提供されるテレビの番組表から録画されたビデオデータのタイトルやトピックを知ることは可能であるが，どのような内容か，あるいはキーワードが実際にいつ話されたのか等の詳細は一般的に入手することが困難である．また，ユーザが関心のある特定のキーワードが話された部分を検索・特定する機能は提供されていない．

ユーザが見たい・聞きたいビデオの特定の区間を検索するためには，テキストまたは音声を検索語とする音声文書検索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，大語彙連続音声認識（以下では語彙制限型連続音声認識と表記）の結果を利用するシステムが 1 つの代表的方式である[1),2)]．しかし，この方法では辞書に登録されていない語句（未登録語）が検索語として与えられた場合，その言葉が話されている区間を特定することは困難となる．人名や地名，専門的技術用語といった特殊な単語は一般に未登録語となりやすいが，これらはビデオデータを特徴付ける語句となることが多く，利用者から検索語として与えられる頻度が高いと考えられる．実際，付録に示すように，検索サイト Yahoo[3)] における年間検索キーワードの上位の単語のうち，約 4〜5 割が音声認識デコーダ Julius の汎用辞書および Web 辞書（Rev.3.4.2 とともに提供されている辞書．約 6 万語彙）に登録されていない．このため音声文書検索システムには未登録語対応が必要不可欠といえる．一方，辞書に登録されている単語であれば語彙制限型連続音声認識システムの結果を利用することができる．そこで，検索語として与えられた単語が辞書に登録されている場合には語彙制限型連続音声認識システムの結果を検索し，辞書に登録されていない場合は本論文で提案する語彙フリーな音声文書検索方式を利用すればよいと考える．本論文では認識の単位を単語とするのではなく，音素あるいはさらに細かい単位とすることによって，あらゆる語句の検索を可能とする語彙フリーな検索システムの構築を目指す．これらの単語よりも細かい単位を本論文においてはサブワードと呼ぶ．また，サブワード単位の音声認識をサブワード認識と表記する．関連研究として語彙制限型連続音声認識と音節ベースでのインデックスを組み合わせることによる未知語に頑健な音声文書検索手法が提案されている[4),5)]．この研究は音声認識デコーダに未登録語判定機能を付加し，未登録語と思われる区間に対して音節列での認識結果を利用するものである．しかし，未登録語を辞書登録単語と誤判定する割合が高く，未登録語を検索語として与えた場合に誤判定区間は候補として提示されない．

本論文では，サブワード認識結果に基づく語彙フリー音声文書検索方式の性能向上を目指して，2 つのアプローチを提案し，その有効性を検証する．1 つは，音素等よりも時間的に精緻なモデル化を行ったサブワードモデルを導入すること，他の 1 つはサブワード系列どうしの照合時に局所距離としてサブワード間音響的距離を利用することである．サブワードとしては，音素よりも時間的に精緻化したセグメントとして今回新たに開発した環境依存半音素モデル（triphone を時間的に半分にしたもの．以下半音素と略記）と，音声の音響的特徴を詳細に区分化した音素片モデル（*Sub-Phonetic Segment*，以下 SPS と略記）[6)] を音声検索に導入することを提案する．以下，それぞれを導入した理由について述べる．

語彙制限型連続音声認識システムでは，音素を単位として monophone，triphone 等が利用されているが，性能面から triphone が使われることが多い．検索システムの場合も triphone を用いた研究が行われているが[7)]，時間的な整合性をとらえるうえで，時間軸上でより精緻なモデルが性能向上に有効であると考えた．たとえば，音節よりも音素が時間軸上で詳細なサブワードの単位となる．「イワテ」を表す場合，音節では 3 つ，音素では 5 つ（2.1 節表 1 参照）のサブワードで表記される．1 つのサブワードの欠落/誤認識が起こった場合，サブワード数の少ない音節をサブワードとした方が大きな影響を受ける．サブワードの認識性能は後述するように高くないため（5〜8 割程度），時間軸上でより詳細なモデル化を行うことで，これらの認識誤りにロバストになると考えた．また，時間軸上でのサブワードの並びが多いほど，脱落挿入が多いサブワード系列どうしのスポッティング的な照合には有利になると考えた．一方フレーム単位に処理する方式[8)] は，記録量，検索時間のオーバヘッドが大きく，また不特定話者対応が難しいと考えられる．そこで本論文では，サブワードモデルとして音素より時間的に精緻なモデル化を行った半音素と SPS を利用する．環

境依存の半音素とは前後の音素を考慮したうえで，各音素を時間軸上で 2 つに分割したものである．それぞれのモデルは Hidden Markov Model（HMM）で構成する．本論文では，triphone および半音素，SPS を用いた検索性能について比較評価を行い，適切なサブワードモデルについての検討を行う．

本検索システムでは，音声クエリと音声データベース（ビデオ音声）それぞれについてサブワードモデルで認識を行い，検索語のサブワード系列と音声データベースのサブワード系列とを連続 DP により照合する．本論文では，連続 DP による照合の際の局所距離に音響的距離を導入する方式を提案する．サブワードモデルを用いた関連研究には，edit distance を用いた研究[9]，検索語やデータベース中に出現するサブワードモデルの頻度を用いることによってターゲットとなる区間を特定する研究[10),11)] 等がある．これらのアプローチでは，サブワード認識の結果に検索性能が大きく依存する．本論文で提案するアプローチではサブワード間の音響距離を用いることでサブワードの誤認識の影響を低減できると考える．2 つのサブワード間の音響距離はサブワード HMM の統計量から事前に距離マトリックスとして計算しておき，サブワード列の照合時にはこの距離マトリックスを参照して検索を行う．

本論文では，平成 16 年に発生した新潟中越地震の際にテレビで実際放送された安否確認放送に対して本検索システムを適用し評価を行う．地震直後には有線電話，携帯電話がつながらない状態が続き，被災者との連絡をとれなかったため，安否確認放送が延々と放送された．安否確認放送では，災害地の内外から連絡を取りたい人に対するメッセージをアナウンサが伝え続けるもので，該当者が放送時にテレビを見ていなければ情報は伝わらず，検索機能が必要とされた．語彙制限型連続音声認識システムを用いた検索では，多くの氏名と地名からなるこの放送に対しては十分機能しないと想定される．そこで本論文で提案するサブワードモデルを利用した検索方式を適用し，多数の固有名詞が含まれるビデオデータに対してシステムの有効性および実利用性を検証する．

本論文では 2 章で提案する語彙フリーの音声文書検索システムを概説し，比較検討する複数のサブワードモデルとサブワード間の音響的距離について解説する．3 章で各サブワードモデルにおける検索性能について実験および考察を行う．4 章では，提案する音声文書検索システムの災害放送への適用とその評価実験について述べる．最後にまとめを述べる．

2. 語彙フリー音声文書検索システム

2.1 システム概要

提案する語彙フリーの音声文書検索システムの概要を**図 1** に示す．

録画したビデオ等の音声データはあらかじめサブワード認識を行いサブワード列のデータベース（DB）とする．このデータは番組ごとあるいはコマーシャル部をカットした部分番組ごと，各音声部ごと等に区切られ，それぞれの区間に対応するサブワード系列があると仮定する．検索語はキーボード等によるテキスト，または音声で与えることを想定する（本論文におけるテキストクエリとは，検索語の読みをキーボードより入力する際に得られる音節系列を変換規則により変換したサブワード系列を意味し，これをテキストクエリと呼ぶものとする）．検索語が音声で与えられた場合は音声 DB 同様のサブワード認識によりサブワード列に変換する．検索語のサブワード列と DB 中の全サブワード列間で連続 DP によるサブワード列照合を行う．照合の結果，距離が局所最小となる区間のうち，距離が小さい順に候補としてユーザにそのビデオクリップを提供する．

サブワードとは単語よりも細かい単位で，音素や音節がこれに該当し，語彙制限型連続音声認識システムでは monophone や triphone が一般に用いられている．一方，1 章で述べたように連続する音声データの検索の際には時間軸上での整合が重要になるため，精緻なモデルが有効ではないかと考え，triphone モデルとともに，時間軸上でより精緻なモデル化を行ったモデルをサブワード検索に導入することを提案する．本論文で導入するモデルは次節で詳述する半音素モデルと SPS モデルの 2 つである．**表 1** に「イワテ」という単語を表現する際の各サブワードモデルの表記を

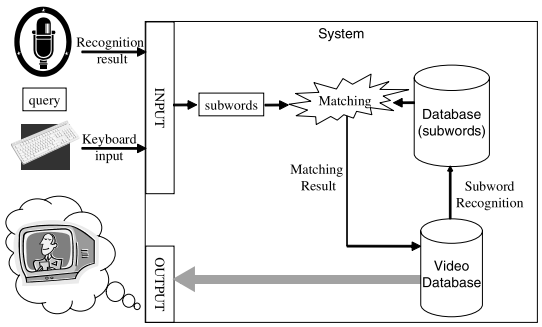


図 1 提案する音声文書検索システム概念図
Fig. 1 System image of the proposed Spoken Document Retrieval system.

表 1 各サブワードにおける「イワテ」の表記
Table 1 Subword expressions for the word “Iwate”.

Subword	Expression	Num. of subwords
音節	i wa te	3
monophone	i w a t e	5
triphone	#-i+w i-w+a w-a+t a-t+e t-e+#	5
半音素	#1i i2w i1w w2a w1a a2t a1t t2e t1e e2#	10
SPS	#i ii iw ww wa aa at tcl tt te ee e#	12

記す．表中#は前後に隣接する単語の最初あるいは最後の音素を表す．monophone および triphone は 5 つのサブワードから構成されるのに対し，半音素は 10，SPS は 12 となる．本論文では各サブワードはすべて同一状態数からなる HMM でモデル化するため，表記に必要なサブワード数が多いほど，時間的に精緻なモデル化がなされているといえる．

連続 DP によるサブワード列照合の際の局所距離は，検索語のサブワードの記号と音声 DB のサブワードの記号間の距離となる．この局所距離に前述した edit distance を用いる場合が多いが，本論文では次節で述べるサブワード HMM 間の統計的な距離（サブワードモデル間の音響的な近さ，遠さを表す距離）を導入する方式を提案する．

以降の節で，サブワードモデルの構築方式とサブワード間距離の導入方式について詳述する．

2.2 サブワードモデルとその構築方法

以下，本論文で扱うサブワードモデルとその構築方法について述べる．

2.2.1 検討するサブワードモデル

(1) **triphone モデル：**現在の語彙制限型連続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て音素をモデル化したものである．言語的に連続可能な音素の組は約 21,000，実際に音声として存在するのは約 8,000 であるといわれている[12]．今回のモデル数☆：7,956.

(2) **半音素モデル：**本論文で考案したモデルであり，前後の音素環境を考慮した 1 つの音素（triphone）を時間的に 2 つのサブワードに分割したモデルである．**図 2** に半音素モデルの概念図を示す．図中上段のモデルを triphone とすると，前半部と後半部を 2 つの半音素として表し，時間軸上で 2 倍に精緻化されたモデルとなる．たとえば，後続する音素が異なる 3 種類の triphone の L1-X+R1，L1-X+R2，L1-X+R3 の前半部は L1-X で同一ラベルとなるため triphone よりモデル数は少なくなる．X が母音の場合，前後任意の音素が付加可能であるため，母音 X に関して構成されうるモデル数は，triphone の場合，音素数 × 音素数，半音素の場合，音素数 + 音素数（前半部，後半部につく音素となるため）となる．今回のモデル数：1,333.

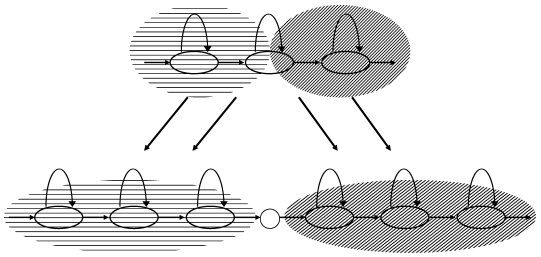


図 2 半音素モデル概念図
Fig. 2 Illustration of a new subword (demi-phone).

(3) **SPS モデル：**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[6]．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の記号を割り当てる．たとえば，表 1 の音素 a は先行音素 w との渡りの部分 wa と，a の中心部分 aa および後続音素 t の渡りの部分 at から構成され，次の音素 t は先の渡りの部分の at，破裂音の直前の無音部 tcl，t の中心部分 tt および後続音素 e の渡りの部分 te から構成される．個々の記号をそれぞれモデル化することで詳細なモデル化がなされる．モデル数：423.

音素ラベルと各サブワードモデルにおける対応関係を**図 3** に示す．上記の 3 つのサブワード系列は読みが与えられると，その音素系列から各サブワード系列へと自動的に変換可能である．

2.2.2 サブワードの音響および言語モデルの構築

本論文では，すべてのサブワードモデルは同一の状態数，自己ループを含む left-to-right 型の HMM を用い，同一の音声コーパスを利用して構築する．

☆ 今回のモデル数とは 3.1 節で述べる学習データ中に出現するモデル数を意味する．

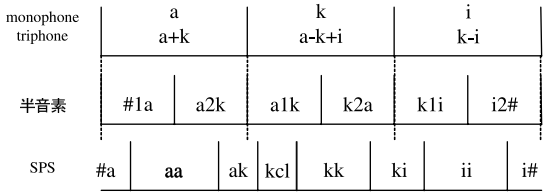


図 3 音素ラベルと各サブワードモデルにおける対応関係
Fig. 3 Boundaries of each subword sequence for the same word.

音声コーパスの書き起こし音素列から各々のサブワード系列を求め，このサブワードラベルをもとに各サブワードモデルについて学習を繰り返す．混合数 1 のサブワード HMM を構築し，混合数を増加させて学習を行い，最終的なサブワード HMM を構築する．

認識時に用いる言語モデルは，上記のサブワード系列に変換した学習データを用いて，サブワードバイグラムおよびサブワードトライグラムとする．

2.3 サブワード間距離

本検索システムでは，検索語と音声 DB との連続 DP によるサブワード列照合の際に，サブワード 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）をサブワードモデル間距離として定義する．サブワード HMM の各状態には一般の HMM 同様，複数の分布が与えられている．サブワード間の間違いやすさは，同一の順番にある状態間（Ex. 2 つのサブワードそれぞれの 1 番目の状態間）の最も類似する分布の影響が大きいと考えられる．そこでサブワード間音響距離は，2 つのサブワード間の同一の順番にある状態間であらゆる分布の組合せにおいて，分布間の統計的な距離が最も小さくなる距離を用いて定義する．

各サブワードは N 状態からなる HMM で構成され，各状態に M 個のガウス分布を持つ混合ガウス分布モデルとする．2 つのサブワード間の距離計算手順を以下に記述する．ただし，p，q を 2 つのサブワードモデル，s_i^p をサブワードモデル p の i 番目の状態，$c_{i,j}^p$ をサブワードモデル p，i 番目の状態の j 番目の分布とする．

(a) サブワードの選択：未処理の 2 つのサブワード (p, q) を選択して (b) へ．未処理のサブワードの組合せがなければ距離マトリックスを出力して終了する．

(b) 状態間距離の計算：サブワード p，q の i 番目の状態（$1 \leq i \leq N$）について (b-1)，(b-2) により分布間距離を求めた後，(b-3) で状態間距離を決定する．まず $i = 1$ として，

(b-1) 分布の選択：状態 s_i^p と s_i^q 中の未処理の 2 つの分布（$c_{i,j}^p$，$c_{i,k}^q$）を選択して (b-2) へ．未処理の分布の組合せがなければ (b-3) へ．

(b-2) 分布間距離の計算：分布間距離を測る代表的な距離尺度として Kullback-Leibler Divergence [13] や Mahalanobis 距離，Bhattacharyya distance [14] などが考えられるが，本論文では多くの研究や我々の予備実験においても優位性が示されている Bhattacharyya distance を利用する．ここでサブワード p，状態 i の j 番目の分布 $c_{i,j}^p(x)$ を構成する L 次元無相関ガウス分布を $g_j(x)$ その第 ℓ 次の平均と分散をそれぞれ $\mu_{j\ell}$，$\sigma_{j\ell}^2$ とすると，2 つの分布間距離 $d_B(c_{i,j}^p(x), c_{i,k}^q(x))$ は式 (1) で定義される．

$$\begin{aligned} & d_B(c_{i,j}^p(x), c_{i,k}^q(x)) \\ & = -\log \int \sqrt{g_j(x) g_k(x)} dx \\ & = \frac{1}{4} \sum_{\ell=1}^{L} \left\{ \frac{(\mu_{j\ell} - \mu_{k\ell})^2}{\sigma_{j\ell}^2 + \sigma_{k\ell}^2} + \log \frac{(\sigma_{j\ell}^2 + \sigma_{k\ell}^2)^2}{4\sigma_{j\ell}^2 \sigma_{k\ell}^2} \right\} \end{aligned} \tag{1}$$

(b-3) 状態間距離の決定：式 (2) により 2 状態 s_i^p と s_i^q との状態間距離 $d_s(s_i^p, s_i^q)$ を求める．ここで M 個の混合分布から構成されているので，$M \times M$ 通りの分布間距離を (b-2) で計算し，全組合せ中の最小値を状態間距離とする．$i = N$ であれば (c) へ，それ以外であれば次の状態の処理のため $i = i + 1$ として (b-1) へ．

$$d_s(s_i^p, s_i^q) = \min_{1 \leq j,k \leq M} d_B(c_{i,j}^p(x), c_{i,k}^q(x)) \tag{2}$$

(c) サブワード間距離の計算：サブワード p と q のサブワード間距離 $d(p, q)$ は，式 (3) に示したように状態間距離の平均値として計算する．(a) へ戻る．

$$d(p, q) = \frac{1}{N} \sum_{i=1}^{N} d_s(s_i^p, s_i^q) \tag{3}$$

サブワード間音響距離算出はサブワード間の組合せと分布間の組合せが多く計算コストが大きいため，全サブワード間の距離を事前に計算したものを距離マトリックスとして保存しておき，検索時に距離マトリックスを読み込む．

2.4 サブワード列照合方式

本システムでは音声 DB と検索語のサブワード系列間の照合方式に連続動的計画法（連続 DP：*Continuous*

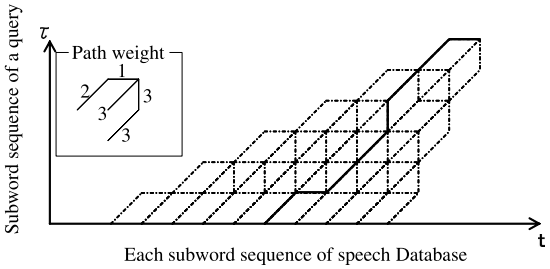


図 4 連続動的計画法
Fig. 4 Continuous Dynamic Programming.

Dynamic Programming）を用いる．連続 DP の傾斜制限は**図 4** に示した非対称型傾斜制限を用いる．

検索語の τ 番目のサブワードとデータベースの t 番目のサブワードまでの累積距離 $G(\tau,t)$ は以下の式 (4) で与えられる．式中の $d(p_\tau,q_t)$ は検索語の τ 番目のサブワード p_τ とデータベースの t 番目のサブワード q_t との局所距離を意味しており，この局所距離に 2.3 節で述べたサブワード間音響距離を用いる．連続 DP 中，局所距離はメモリー上のサブワード距離マトリックスを参照する．なお検索語ごとにサブワード数が異なるため，式 (4) で求めた累積距離を検索語のサブワード数で正規化する．これを連続 DP の正規化距離とする．

$$G(\tau,t)=\min\begin{cases}G(\tau-1,t-1)+3\cdot d(p_\tau,q_t)\\ G(\tau-2,t-1)+3\cdot d(p_{\tau-1},q_t)\\ \quad+3\cdot d(p_\tau,q_t)\\ G(\tau-1,t-2)+2\cdot d(p_\tau,q_{t-1})\\ \quad+d(p_\tau,q_t)\end{cases}\tag{4}$$

3. 評価実験

3.1 学習用・評価用音声データ

各サブワードモデルの検索性能を評価するため比較実験を行う．

すべてのサブワードモデルは状態数 3 の HMM で構成し，日本音響学会の新聞記事読み上げ音声コーパス（JNAS：*Japanese Newspaper Article Sentences*）[15] を利用して音響モデル，言語モデルを作成した．検索実験に用いるモデルの混合数は距離マトリックス作成時の所要時間の関係から 16 とした．

言語モデルは，単語ではなくサブワードのバイグラム，トライグラムを用い，JNAS の各サブワード表記から学習した．なお，サブワード認識の際用いる辞書は，すべてのサブワードで構成し，たとえば SPS の場合，辞書の語彙は SPS のモデル 423 個となる．

表 2 音響分析条件
Table 2 Conditions of feature extraction.

Sampling	16 KHz，16 bits
Feature vector	MFCC（12 dim）+ Δ MFCC（12 dim）+ power
Window	Hamming window
Window length	256 points (16 msec)
Frame interval	10 msec / 5 msec

評価用データは学習用の音声コーパスとは異なる電子技術総合研究所の単語データベース（ETL-DB）[16] を利用する．DB 中の 1,542 語 10 人分の単語セットを用いる．各単語の音素数は 4 から 12 であった．あらかじめすべての単語について音声認識デコーダ Julius を用いてサブワード認識しておく．

本 DB に含まれる音声データの音響分析条件等を**表 2** に示す．分析時のフレームシフトは 10 msec/5 msec とした．語彙制限型連続音声認識を行う際に一般的に用いられるフレームシフトは 10 msec であるが，SPS および半音素は時間的に精緻なモデルであるため，10 msec では各状態の継続時間と十分な学習データが確保できないことが予備実験によって確認された．そこでこれら 2 つのサブワードモデルに対しては 5 msec のフレームシフトを用いる．一方 triphone モデルの学習に 5 msec のフレームシフトで抽出した特徴量を用いた場合，性能の向上が見られなかったため，triphone モデルについては 10 msec のフレームシフトを用いる．

3.2 実験条件

本実験ではモデル間の性能比較を目的とし評価を単純にするため，連続音声中の検索語の検索ではなく，3.1 節で示した単語 DB の全 1,542 単語，10 人分の音声を検索対象データとして検索単語との照合実験とした．

検索語はテキストで入力する場合（テキストクエリ）と音声で入力する場合（音声クエリ）との 2 通りを想定した．テキストクエリの場合，ルールに従いサブワードの系列に変換する．検索のターゲットとなる音声単語はそれぞれ 10 人が発声しているので各単語につき 10 個の正解音声単語が存在する．テキストクエリは 1,542 語（1,542 個），テキストクエリに対する検索対象データ数は 1,542 語 $\times$ 10 回 $=$ 15,420 個となる．一方，音声クエリの場合，10 人中 1 人の話者の音声データを検索語とし，残りの 9 人のデータを検索対象 DB としたため，各話者についての性能が得られる．ある話者を検索話者とした場合，検索語数は 1,542 個，検索対象データ数は検索語発話者のデータを除く 1,542 語 $\times$ 9 回 $=$ 13,878 個となる．音声クエリの場合には各話者が検索話者となりうるため 10 回

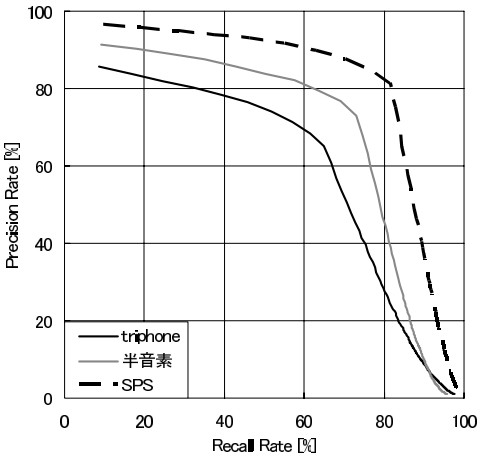


図 5 サブワードモデルによる検索性能の比較
Fig. 5 Performance curves for the three types of subword models.

の検索実験を行う．本論文において掲載する結果は 10 人の話者の検索性能のうち，中間の性能を示した話者の結果を示す．

各検索語に対しては類似度の高い順（連続 DP の正規化距離が小さい順）に候補として出力し，式 (5) で定義される Precision-Recall グラフおよび F 値で評価を行う．式において *Num. of relevant word retrieved* はある順位 R までに正しく検出された単語数を表している．*Total num. of word retrieval* はある順位 R までの検出数（本実験においては $1{,}542 \times R$），*Total num. of relevant word* は音声 DB に含まれる全検出対象単語数（テキストクエリに対しては 15,420，音声クエリに対しては 13,878）を表している．検索語ごとに候補を出力するため，1,540 個ずつ出力される．したがって，結果を示すグラフはある順位まで出力した際の Precision rate，Recall rate を表している．F 値は Precision および Recall の調和平均を表している．本論文ではシステムの性能を測る指標として F 値の最大値を算出する．この値は Precision，Recall ともに高い値である箇所を示し，グラフにおいて最も右上に近い点を示している．

$$\text{Precision} = \frac{\text{Num. of relevant word retrieved}}{\text{Total num. of word retrieval}}$$

$$\text{Recall} = \frac{\text{Num. of relevant word retrieved}}{\text{Total num. of relevant word}}$$

$$\text{F 値} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (5)$$

3.3 結果および考察

3.3.1 サブワードによる検索性能の比較

図 5 に triphone，半音素，SPS の各サブワードの検索性能を示す．各モデルともテキストクエリ，サブワード間音響距離を用いたものである．

表 3 各サブワードの言語モデルの perplexity およびサブワード認識率
Table 3 Perplexity of language model and recognition rate for each subword model.

	triphone	**半音素**	**SPS**
Perplexity	4.73	2.97	2.65
Recognition Rate	55.89	65.08	77.84

検索性能において最も優れているのは，SPS モデルであった．同様に時間的に精緻なモデルである半音素も triphone より検索性能は向上しており，時間的に精緻なモデル化が有効であることが確認できる．半音素や SPS モデルが triphone よりも優れた性能を示した理由として主に 2 つの要因が考えられる．まず，はじめにで述べたように，時間軸上での精緻化によりロバスト性が向上したためと考える．次に，**表 3** に言語モデルのサブワード perplexity およびサブワード認識率を示すが，半音素や SPS はモデル間の接続の複雑さを抑えることができ，triphone モデルより認識率が高くなったと考える．SPS と半音素間で検索性能に差が出たが，要因としては SPS が音響物理的特性を考慮してモデル化・分割したのに対し，半音素モデルは各音素を機械的に前半部と後半部に分割したモデルであること，半音素より SPS がより時間的に精緻であることが考えられる．

半音素は時間軸上の状態数の点で，6 状態の triphone モデルと類似している．そこで半音素と同条件のもとで 6 状態の triphone モデルを構築し，2,000 状態となるよう状態共有し再学習を行ったうえで同様の評価を行った．その結果，半音素および 3 状態の triphone よりも低い性能となった．この理由は明確ではないが，50,000 状態の学習となり十分な学習が行われなかったこと，遷移確率が適切に学習できなかったことが原因と推定される．

3.3.2 音響的距離導入の有効性

サブワード間距離に 2.3 節で定義した音響距離を用いた場合と，edit distance を用いた場合とで検索性能の比較を行った．**図 6** に SPS モデルを用いた場合の結果を示す．

サブワード間距離に音響距離を用いることによって検索性能が向上することが分かる．ここでは SPS モデルの結果のみをグラフ化したが，**表 4** に示したように半音素や triphone においても，音響距離を導入することにより検索性能が向上するという同様の傾向が得られ，サブワード間の音響的距離の有効性が確認

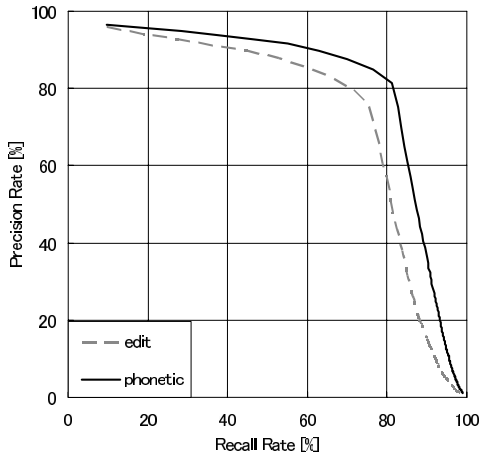


図 6 サブワード間距離による性能比較
Fig. 6 Performance curves for the edit and phonetic distances.

表 4 各実験条件下の検索性能（F 値）
Table 4 F-measure of each evaluation condition.

Query	triphone	半音素	SPS
Text (phonetic)	65.0	73.1	81.3
Text (edit)	47.8	63.2	75.3
Speech (phonetic)	53.3	64.8	71.3

できる．

3.3.3 テキストクエリと音声クエリ

音声 DB に対する検索語として，テキストおよび音声を入力した場合の検索性能を表 4 に示す．サブワード列照合には本論文で提案したサブワード間音響距離を利用した．音声クエリについては 10 人について評価を行い，10 人の F 値の平均を示している．検索語はテキストとした方が高い性能が得られた．表 4 のとおり，半音素や SPS においては音声を検索語とした場合でも，triphone のテキストクエリの検索性能と同等の検索性能が得られた．

4. 災害放送への適用

4.1 背　　景

提案手法の有効性を確認するため，多数の低頻度語（固有名詞等）を含む実際のテレビ放送に対して検索を行う．

平成 16 年 10 月に新潟中越地震が発生した．同年 12 月にはスマトラ沖地震が発生し，それぞれ深刻な被害を与えた．新潟中越地震では，被災者の親族，友人等の関係者は被災者の安否を確認しようとしたが，有線電話，携帯電話による連絡は回線が混雑・混乱し非常に困難であった．そこで，テレビを通じて被災者および関係者に対し，“東京都の山田太郎さんから山古志村の山田花子さんへ「大丈夫ですか，後で連絡ください」”といった多くのメッセージが放送され続けた．この安否情報を伝えるテレビ放送はいったん録画し，録画したビデオデータから，必要とする部分だけを検索する機能が望まれた．

新潟中越地震の際には，本論文で対象とするテレビ放送のほかにも，デジタル放送によるデータ放送や，web 上で情報提供もなされていた．これらの情報が利用できれば，本システムよりも確実に情報を得ることが可能であると考えられるが，災害時に web 等が利用できない事態も想定され，複数の連絡手段および情報確認手段を確保することが重要であると考える．また，本システムは音声入力が可能なことから，だれもがより容易に検索クエリを与えることが可能であると考える．

新潟中越地震における災害安否確認放送では，人名や地名が放送内容の主要部分であり，必然的にこれらの固有名詞が検索のキーワードとなる．通常の語彙制限型連続音声認識システムで実際の放送データを認識してみたところ，固有名詞の誤認識が多く，そのまま利用することは困難であった．そこで，本論文で提案したサブワードモデルを用いた音声文書検索方式を適用することで，人名や地名等の固有名詞を多く含むビデオデータに対しての有用性を検証する．

4.2 実 験 条 件

実験データとして，新潟中越地震の際のテレビ放送を 3 倍モードでビデオテープに録画・録音したものを利用した．ビデオデータをパソコンに取り込んで MPEG-1 形式に変換し，その音声データを表 2 で示した分析条件に従って特徴量を抽出し，サブワード認識を行った．評価を容易にするため，ビデオデータはメッセージごとに事前に分割処理を行い，システムが選んだメッセージに検索語があれば正解とした．各メッセージの長さの平均は約 10 秒であり，triphone 認識の結果では約 100 の triphone を含んでいた．

本システムでは自分が呼びかけられている箇所を探すという場面を想定し，検索語としてビデオデータに 1 回のみ出現する人名を用いることとした．テキストクエリは 44 種類の人名の音節系列をサブワード系列に自動変換したものを与えた．音声クエリは男性話者（学生および教員）22 人を，11 人ずつ 2 グループに分け，各グループで先の人名を 22 種類ずつ（計 44 種類）孤立発声した．検索語はコンピュータやその他の雑音が含まれる通常の研究室内で収録した．検索性能は 100 件（約 40 分）のビデオデータに対してテキストクエリ 44 個，音声クエリ 484（11 人×22 種類×2 グループ）

個の検索語を与えて評価を行った．

今回の実験での検索語には，音声認識デコーダ Julius に付属の汎用辞書に登録されている人名もあったが，サブワード照合方式の性能を評価するという観点から，辞書に登録されている場合もサブワード照合した場合の性能で評価を行った☆．

本実験でのサブワードは triphone および 3 章の実験において最も高い検索性能を示した SPS モデルを用いた．

4.3 結果および考察

図 7 および **表 5** に検索性能結果を示す．図には音声クエリを与えた際の Precision-Recall グラフを，表には上位候補の Recall rate のみを示す．また，図には参考までに SPS の全クエリに対する正規化距離を昇順に並べた際のグラフも追加した．順位ごとに出力した場合の方が性能面で有利になり，正規化距離とともに順位情報が重要であることが分かる．表中の Rank は提示する候補数を表し，Average は正解を見つけるまでに必要な平均候補数を意味する．R 位までの候補を抽出し，正解が C 個だった場合の Recall rate は $C/484$，Precision rate は $C/(484 \times R)$ となる．Precision rate については Recall rate および順位から算出できるため，表では割愛した．

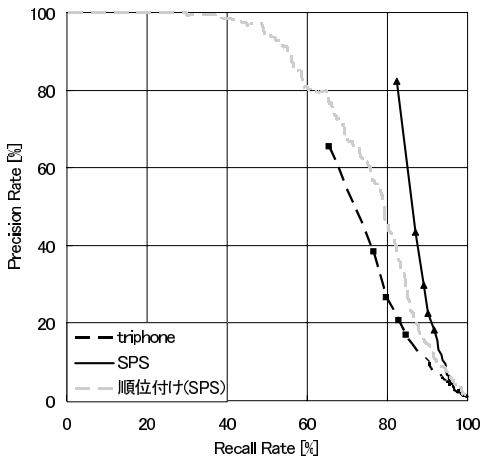


図 7 災害放送に対する検索性能（音声クエリ）
Fig. 7 Performance of SDR for TV broadcasting (speech query).

表 5 災害放送に対する検索性能（再現率）
Table 5 Performance of SDR for TV broadcasting (Recall rate).

	triphone		SPS	
Rank	**Text**	**Speech**	**Text**	**Speech**
1	79.55	65.50	93.18	82.44
3	97.73	79.75	97.73	89.05
5	97.73	84.71	97.73	91.74
Average rank	1.39 位	4.52 位	2.02 位	3.60 位

☆ 今回用いた人名のうち，名前が辞書に登録されている割合は，54.55%であった（音素系列が同じ場合，辞書に登録されているとした）．

3.3 節での結果同様，音声クエリを与えた際の検索性能は SPS モデルを用いた場合の方が triphone を用いた場合を上回った．SPS のテキストクエリを与えた場合は最も類似度の高い候補のみの出力で 90%を超える Recall rate が得られた．SPS のテキストクエリの平均順位が triphone のテキストクエリよりも劣った結果となっているが，これは 44 種類の検索語のうち，1 種類のみ SPS では検出しにくい語句が含まれていたためである．また音声クエリを与えた場合でも 100 件のデータに対して，5 個の候補を出力すると 91.7%の Recall rate が得られており，平均すると 3.6 回でユーザが見たい部分を特定できることを示している．本検索システムがなければ平均 50 件のデータを見なければならず，ユーザにとって本システムの利用価値が高いと判断することができる．

一方，検索語がビデオ（検索対象）中にない場合，本来であればその検索語が存在しない（棄却）とユーザに提示すべきである．検索語が語彙制限型連続音声認識システムに登録されていた場合にはその認識結果を用いることになるため，当該検索語への置換誤りが，誤動作の原因となる．語彙制限型連続音声認識において置換誤りが起こることを避けることは困難であるが，検索語以外の単語を検索語と認識する確率は低いと考えられる．このため，語彙制限型連続音声認識システムを利用した場合には高い精度で棄却をユーザに提示できると考える．一方検索語が未登録語であった場合，サブワード方式を用いることになるが，サブワード方式は類似探索であるため，現段階では類似性の高いものから順に出力されユーザに提示される．棄却方式は様々考えられるが，本論文では今後の課題としたい．

5. おわりに

本論文ではサブワードを用いた語彙制限のない音声文書検索に対して，適切なサブワードモデルとその処理方式の検討を行った．サブワードとしては，monophone，triphone より時間的に精緻な半音素や SPS を導入し，その有効性を確認した．また，照合時の局所距離としてサブワード間の音響類似度を導入することで性能向上が確認できた．実際に放送された固有名詞の多いテレビ番組を利用し，実データに対する固有名詞検索においても本手法が有効であることを検証した．

今後，検索時に最適なサブワードモデルの検討と自

動的なサブワードモデルの構成方法を検討していくとともに，より大規模な実データに対する検索実験を行い，システムの実用化を目指していきたい．

謝辞 本論文の一部は文部科学省科学研究費補助金基盤（B）(1) No.15300026 および（C）No.17500073 の支援を受けて実施された．

参 考 文 献

1) 伊藤克亘，藤井 敦，石川徹也：音声文書検索を用いたオンデマンド講義システム，情報処理学会研究報告，2001-SLP-39, pp.165–170 (2001).
2) 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).
3) http://www.yahoo.co.jp
4) 西崎博光，中川聖一：音声認識誤りと未知語に頑健な音声文書検索手法，電子情報通信学会論文誌 D-II，Vol.J86-D-II, No.10, pp.1369–1381 (2003).
5) Witbrock, M.J. and Hauptmann, A.G.: Using words and phonetic strings for efficient information retrieval from imperfectly transcribed spoken documents, *Proc. ACM Digital Libraries '97*, pp.30–35 (1997).
6) Tanaka, K. and Kojima, H.: A between-word distance calculation in a symbol domain and its applications to speech recognition, *Proc. International Conference on Neural Information Processing* (*ICONIP-97*), pp.1107–1111 (1997).
7) Ng, K. and Zue, V.W.: Phonetic Recognition For Spoken Document Retrieval, *ICASSP 1998*, pp.325–328 (1998).
8) 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，電子情報通信学会誌 D-II，No.6, pp.764–775 (2003).
9) Raghavan, H. and Allan, J.: Matching Inconsistently Spelled Names in Automatic Speech Recognizer Output for Information Retrieval, *Proc. HLT/EMNLP*, pp.451–458 (2005).
10) Moreau, N., Kim, H.-G. and Sikora, T.: Phonetic Confusion Based Document Expansion for Spoken Document Retrieval, *ICSLP*, Vol.2, pp.1593–1596 (2004).
11) Srinivasan, S. and Petkovic, D.: Phonetic Confusion Matrix Based Spoken Document Retrieval, *23rd Annual ACM Conference on Research and Development in Information Retrieval* (*SIGIR'00*), pp.81–87 (2000).
12) 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).
13) 篠田浩一，磯 健一：MDL 基準を用いた HMM サイズの削減，日本音響学会講演論文集，2-5-3, pp.79–80 (2002).
14) 小川厚徳，山口義和，高橋 敏：混合重み係数を考慮した分布間距離尺度による音響モデルの分布数削減，日本音響学会講演論文集，2-1-23 (2004).
15) 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, *The Journal of the Acoustic Society of Japan* (*E*), Vol.20, No.3, pp.199–206 (1999).
16) 速水 悟，田中和世，横山晶一，太田耕三：研究用音声データベースのための VCV/CVC バランス単語セットの作成，電子技術総合研究所彙報，Vol.49, No.10 (1985).
17) 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).

付 録

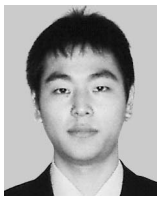
A.1 検索語の辞書登録率

表 6 に検索サイトである Yahoo[3] における 2003, 2004, 2005 年の年間検索ランキング各 30, 50, 50 位までの単語が音声認識デコーダ Julius の汎用辞書および Web 辞書に登録されていなかった割合を示す．辞書に登録されていない検索語は 4 割から 5 割存在し，辞書を利用した語彙制限型音声認識には限界があると考えられる．なお，検索語のランキング上位にはその年に話題となった商品名や人名が多かった．

表 6 検索語の辞書未登録率
Table 6 Unknown word rate of popular searching word.

Year	General dictionary	Web dictionary
2003	36%	34%
2004	40%	38%
2005	50%	48%

(平成 18 年 5 月 1 日受付)
(平成 19 年 2 月 1 日採録)



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．同年岩手県立大学ソフトウェア情報学研究科博士前期課程入学．現在に至る．日本音響学会会員．

伊藤　慶明（正会員）

平成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．AI 技術の適用研究に従事．平成 4 年より技術研究組合新情報処理開発機構に出向．実時間音声認識理解，対話システムの研究に従事．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）．実時間音声処理，音声による情報検索に関心を持っている．人工知能学会，日本音響学会，電子情報通信学会各会員．

小嶋　和徳（正会員）

平成 5 年秋田大学鉱山学部電子工学科卒業．平成 7 年同大学大学院鉱山学研究科電子工学専攻修士課程修了．平成 10 年同大学大学院博士後期課程システム工学専攻満期退学．同年岩手県立大学ソフトウェア情報学部助手．遺伝的アルゴリズムに関する研究に従事．電子情報通信学会，人工知能学会各会員．

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 年同大学大学院工学研究科博士課程修了．同年同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．信号処理，画像処理，知識工学の研究に従事．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．