音声・音響信号のセグメント符号化に基づく高速で制約の少ない音声・音楽検索方式

Fast and Flexible Approach for Speech and Audio Segment Retrieval Based on Segmental Decoding of Speech and Audio Signals

田中　和世（Kazuyo TANAKA, Dr. Eng.）

筑波大学図書館情報メディア系教授 知的コミュニティ基盤研究センター長

(Faculty of Library, Information and Media Science, Professor, University of Tsukuba)

電子情報通信学会フェロー　日本音響学会　人工知能学会　IEEE 会員

著書：「音声工学」（森北出版）「視覚認知と聴覚認知」（オーム社）等

研究専門分野：音声分析・音声変換・音声認識などの音声情報処理　音響信号分析・認識

あらまし　音声認識システムは 1990 年代以降、著しい性能向上を達成してきたといえるが、その基本的枠組みは学習用データの大量収集や言語モデル更新などに代表される環境条件への適応化であり、負担の重い処理を必要とする方式である。一方、多くの音声アプリケーションを想定すると大規模な投資が見合うものは少なく、より負担の軽い音声認識システムの構築方式が、ニーズが高いと見込まれる。筆者らは、こうした観点から大量データの収集やデータの更新・メンテナンスに依存することが少ないシステム構築技術の開発に力を注いでいる。本稿では、まず現行の音声認識システムの構成と課題を考察し、次に筆者らが進める音声認識方式の特徴について述べ、その応用システムとしての音声語検索システム、また、この考え方を拡張したカバー楽曲名リアルタイム同定システムについてその概要を紹介する。

１．はじめに

最初に、本研究の背景となる音声認識技術とこれを組込んだ実用的な応用システムの開発と課題について考察する。音声認識技術を組込んだ実用製品はすでに 30 年以上の歴史をもつが、早くから将来に期待される技術として取り上げられてきたわりには、世の中で広範囲に使用された実用システムは存在して来なかった。（音声認識・合成など音声応用製品の動向調査は、筆者も関係する社団法人電子情報技術産業協会(JEITA)・音声入出力専門委員会[1]で 20 年以上に亘り実施してきている。）このような状況が最近になって変化しつつある。いわゆるスマートフォンへの音声認識機能の搭載で多くの一般のユーザにも身近な機能としてテストされるようになった。筆者の周囲のユーザにも、これなら十分に有用性があるという感想がかなり多い。国内では、これより 10 年程先行して携帯電話やカーナビに音声認識機能が搭載されてきたが、性能や機能性・操作性の問題もあって、あまり使用されて来なかった。それらと今回との違いは、音声認識性能がこの 10 年ほどでかなり改善されたことも一つであるが、大きな違いはその使用条件や使用目的にある。例えば、性能だけで言えば、カーナビに搭載して十分に有効に機能することは現在の技術でも難しい。すなわち、どのような目的でどのような使われ方をされるのかが、音声認識機能が有用か否かに大きく係わってくる。

音声認識機能を搭載した商用システムは現在でも多数あるが、コストに見合って効果を発揮しているシステムは、これまでのところ非常に限定的な目的で狭い範囲の応用システムである。先に挙げたスマートフォンへの搭載もそのコストを明示的に回収することを期待したものとはいえない。例えば、音声認識ソフトに値段をつけてその開発費用を回収し利潤をあげるのは必ずしも簡単ではない。その原因の１つは、音声認識システムを作成するには、使用目的に適合した音声や言語データの収集とデータのメンテナンスという重い負担が必要となる点にある。これは、現在、主流となっている音声認識手法がモデルの学習に用いる音声・言語データや使用環境に強く依存すること、言い換えれば、音声・言語データあるいは使用環境により強く

依存すればするほどより高い性能を引き出すことができるというパラダイムに乗っている点にある。

そこで筆者らは、データに依存することが少ないシステム構築技術の開発、端的にいえば「軽い音声認識システム」の開発に力を注いでいる。これによって大量データの収集やデータの更新・メンテナンスの負担を軽減できることになる。

本稿では、まず現行の音声認識システムの構成と課題を考察し、次に筆者らが進める音声認識方式の特徴、その応用システムとしての音声検索システム（Spoken Term Detection System：音声語検出システム）、またこの考え方を拡張した（カバー曲を含む）楽曲名リアルタイム同定システムについてその概要を紹介する。

２．音声認識技術の現状と課題

現在普及している音声認識システムの構成を示すと、**図１**のようになる。図において左側が（未知）入力音声で、右端がその認識結果で、（日本語であれば）仮名漢字表記の文が出力される。認識課程のレベルとしては下位から順に音響特徴抽出、音声記号レベルの認識（確率計算）、単語レベルの認識（確率計算）、文レベルの認識（確率計算）となる。処理の流れ自体は上位からの仮説（候補）に対する尤度ないし事後確率を行い、上位の返すという形になる。ここで重要な役割を果たすのはそれらの確率計算の基になる音響モデル、単語辞書、言語モデルである。現行方式による音声認識性能は基本的にはこの３要素がどの程度適切にできているかに懸っているといえる。

音響モデルは、音声を記述する音声記号に相当する記号の物理音響的モデルであり、認識性能向上のためには最も重要な役割を果たすといえる。音声認識では、この音響モデルに HMM（Hidden Markov Model）と呼ばれる確率モデルが採用されている。このモデルを精度よく作成（モデルのパラメータを推定）するには、大量の音声サンプルの収集が必要になる。また、言語モデルは語の並び順を共起確率によって表すもので、言語文法に相当している。これもこのモデルを作成するには大量の言語テキストコーパスが必要となる。とくに音声認識の場合には、性能を上げるためには認識タスクとして扱う分野を限定してテキスト文書データや音声の書起しテキストデータにより作成する必要がある。さらに多くの場合、人名や固有名詞は日々、新語が登場したり、語と語の関係も変化する（たとえば、首相や大統領が交代すれば人名との共起関係も変化するなど）。したがって、辞書や言語統計データの更新・再計算が必要となる。

以上をまとめると、認識精度を向上させるための方策としては、以下の２つがある。

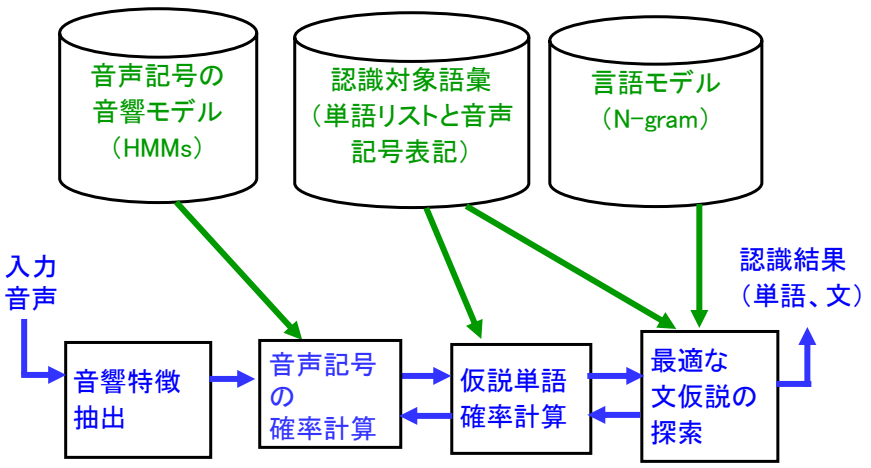

図１　一般的な現行の音声認識システムの構成を示すブロック図

① 音響モデルをその収録環境に適応させる。話者に適応させるとか、周囲環境、回線環境などに適応させる。
② 言語モデルを認識対象となる話題ドメインに適応させる。また、認識対象辞書にない単語（未知語）をできるだけ少なくする。

要するに、それぞれの使用環境に合わせれば良いというのが一般的な考え方であり、この適応処理をできるだけ少ないデータで効果的に実行することが必要となる。しかし、現実には、認識精度を追及したシステムではこの作業に注力せざるを得ず、高コスト化の原因となっている。

そこで筆者らはより軽い負担で応用システムに組込むことができる音声認識システムの実現を目指して、以下のような技術の開発に取り組んでいる[2],[3]。

イ） 比較的少ない音声データから高い性能を得ることできる音響モデルの開発
ロ） 辞書や言語モデルに依存することが少ない認識技術
ハ） ポータビリティの高いシステム構成

３．汎用音声符号系への符号化に基く音声認識方式の提案

上に述べたようなシステムの開発のポイントは、できるだけ実データ（言いかえると環境条件）に依存することが少ない方式・手法の開発である。およそ新しいアプリケーションを構築するに当たっては、いまある認識システム（ソフトウエア）をそのまま使用すればよいということは少ない。何かしら修正を加える必要が出てくることが多い。このときの負荷が全体として小さいものがよいということになる。

このためには、システム構成のブロック化とブロック間の処理の独立性およびそれらの間のインタフェースの規格が明確であることが望ましい条件の１つである。システム開発の開始時に技術者が基礎理論や蓄積された手法・技術について学習する手間をできるだけ少なくするには、知識をマニュアル的な記述として提供できることが望ましい。

音声認識システムを実データと切り離して開発することはもちろん困難であるが、環境条件が変わるごとに大量のデータの収集を必要とするのは避けたい条件である。この要素は、もちろん程度問題ということになるが､データ収集とデータの更新･メンテナンスは、製品のコストの問題と直接関係するので重要な要素である。

筆者らが開発してきた音声認識方式は、すべての信号レベルの音声を一旦、音声記号のような記号レベルに符号化・変換し、この変換された記号領域において音声認識などのシステムを構築するというものである。**図２**にその基本的枠組みを示す[2]。ここで、従来システム（すなわち現在の主流）の方式との違いは、この枠組みでは音響処理領域と記号処理領域(図では SPS domain と表示)を分離することができる点にある。

この方式では、音響領域から記号領域へ送られるものは音声信号に対応した記号系列のみであり、それらに付随する尤度のような変量は送られない。これによって音声信号は記号列テキストとして蓄えられることになる。例えば、記号系を音素とした場合、通常の認識技術による変換では、変換後音素系列には多くの認識誤り（変換誤り）が含まれていることになるが、これは、記号領域における距離計算等の処理によりこの誤りを吸収する処理を行う。本方式における記号列テキストはもともと人が読むことを前提としないプリミティブな符号列であり、音響レベルの変動を反映した記号系列である。

従来方式と提案方式の処理形態の違いを**図３**(a), (b)に示す。従来方式図（a）では、記号領域での仮説（記号列）が音響領域まで下りて行き、仮説（個々の記号列）に対する尤度が上に返される。すなわち、音響領域は記号領域から来る仮説に依存し、記号領域は音響領域の音響モデルに依存している。これに対して、提案方式図(b)では、記号領域から降りてくる情報はなく、上げられるのは、音響領域で変換された記号列のみである。記号領域でも記号列間距離計算に音響モデルが利用されるが、これは音響領域で使用されたものとは（独立に作成された）別の音響モデルが使用される。

このような処理方式によれば、以下のような利点が考えられる。

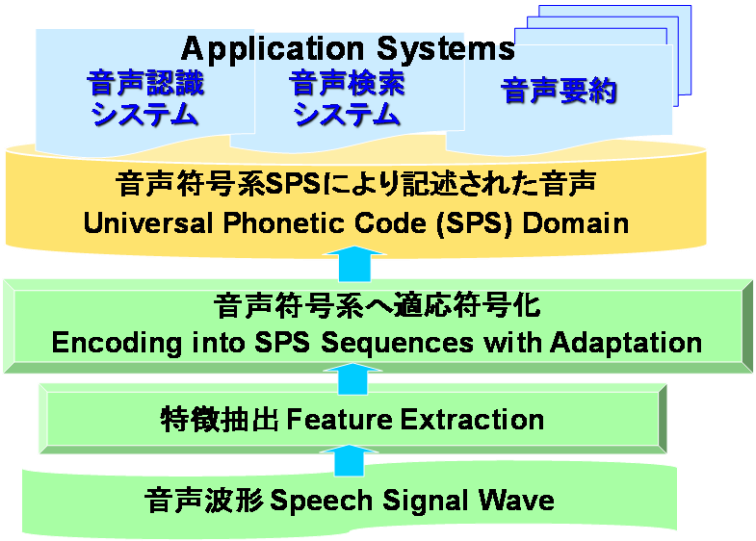

図２　本研究で採用している音声認識・検索システムの構成

①記号領域に蓄積される記号列は（もっとも）プリミティブな系列として保持される。また、それらを適当にインデックス化することで、高速な検索等を可能にすることができる。
②自由度の高い仮説に対して、その距離（確からしさ）を計算することができる。
③音響量に比べ効率的に送信することができ、分散処理に有利である。

４．音声表記記号と音響モデル

音響モデルの学習・作成において、もし、学習サンプルデータが無限にあれば、その分布を関数的に近似する必要もなく、音響モデルの作成問題も必要なくなる。つまり、この問題において前提となるのは、適度な実現可能な負担（コスト・手間）で、性能の良い音響モデルを作成するということである。この前提に立てば、ある認識性能結果が提出されたとき、学習データが不十分なため「正しい」結果が出ていないというような議論は必ずしも適切ではない。 すなわち、学習データの規模という項目は無条件に実現できる項目ではなく、性能を実現するための条件の１つである。どのような音声表記単位を採用し、どのようなタイプの音響モデルを作成するかという問題もこの条件と密接に関係している。

筆者らは、1980 年代より、「音素片」と呼ぶ音声表記単位を提案してきた[7]。その後、これを IPA（XSAMPA）をベースにして拡張し、SPS（Sub-Phonetic Segment）と呼んだ[6]。この記号系は、音声の音響現象をコンパクトな系列にセグメント化したものである。大まかには定常的区間と遷移区間を別のセグメントに分割したものであり、破裂音や破擦音などについてはその性質を特徴付けるセグメントに分割しているが、特殊な概念付けなどを採用してはいない。例を以下に示すと、

(i) TIMIT-DB にある英文
She had your dark suit in greasy wash water all year.

(ii)上の文について音声サンプルの XSAMPA 表記（TIMIT－DB より、使用記号を変更）
h# S i h E dcl dZ @ r dcl d

音声・音響信号のセグメント符号化に基づく高速で制約の少ない音声・音楽検索方式

Fast and Flexible Approach for Speech and Audio Segment Retrieval Based on Segmental Decoding of Speech and Audio Signals

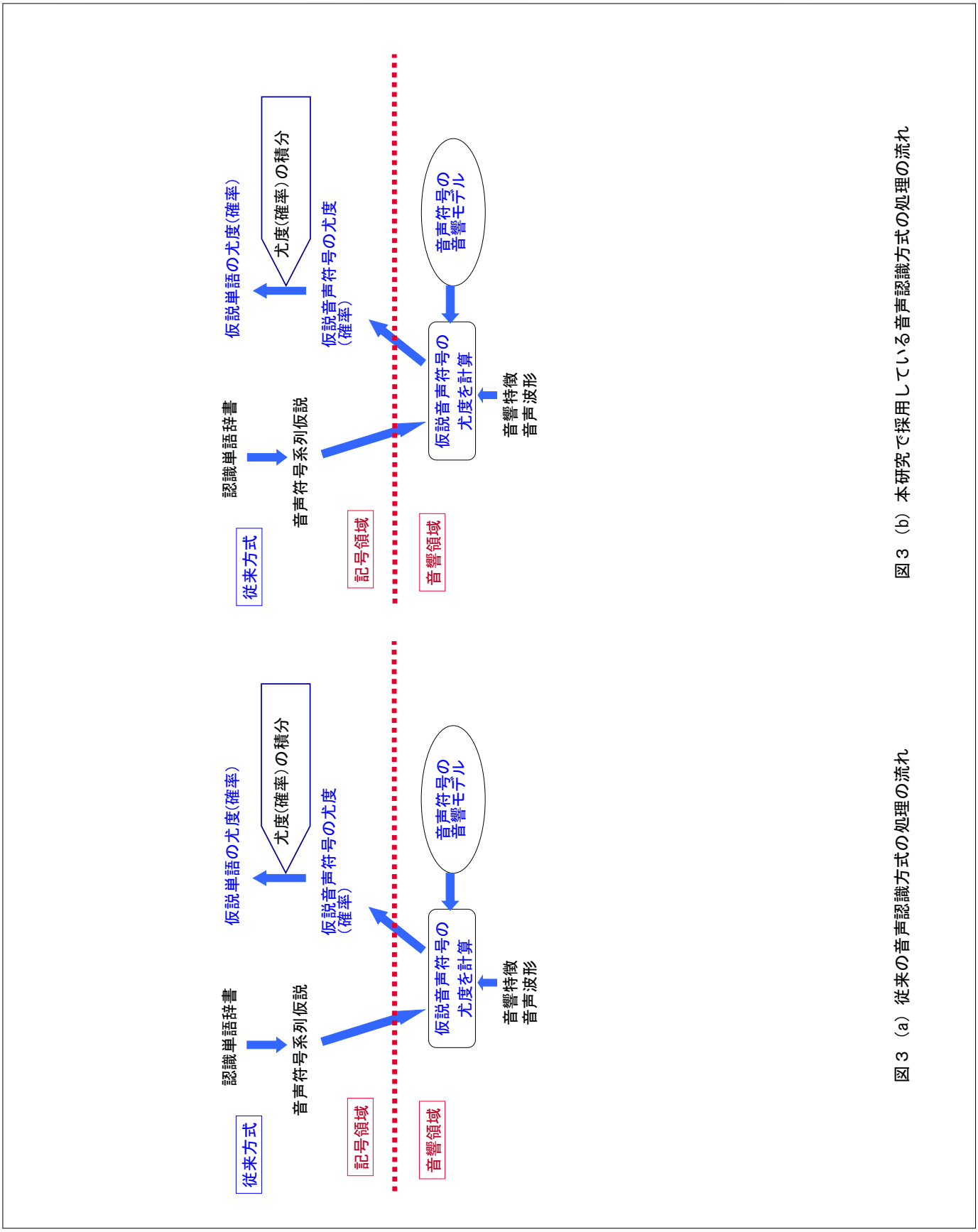

図３（a）従来の音声認識方式の処理の流れ

図３（b）本研究で採用している音声認識方式の処理の流れ

音声・音響信号のセグメント符号化に基づく高速で制約の少ない音声・音楽検索方式

Fast and Flexible Approach for Speech and Audio Segment Retrieval Based on Segmental Decoding of Speech and Audio Signals

A kcl k s u q N gcl g r i z i w OO S PAU w OO dcl d @ q OO l j I @ r h#

(iii)上の記号列の SPS

h# #S SS Si ii ih hh hE EE EdZ dcl ddZ dZ@ @@ @r rr rd dcl dd dA AA Ak kcl kk ks ss su uu uq qq qN NN Ng gcl gg gr rr ri ii iz zz zi ii iw ww wO OOO OS SS S# PAU #w ww wO OOO Od dcl dd d@ @@ @q qq qO OOO Ol ll lj jj jI II I@ @@ @r rr r# h#

日本語音声についての SPS セグメントの種類はおよそ 500 個程度である。この数は（認識システムで用いられている拡張された）音素の約 40 種、および（3 音素組の種類数となる）トライフォンの 8000～12000 種の中間である。したがって、これまでの実験的検討によれば、比較的少ない学習サンプルによってその音響モデルを作成することができる。

図４は、音声の基本認識単位として上記の符号系のいずれを採用するかが、システムの認識性能にどのように影響するかを検証した実験結果の１つである[8]。

この実験では、ユーザの発声したキーワードをデータベース音声から検索抽出するシステム（spoken term detection）の性能を試験したものである。音響モデルの学習データは JNAS（新聞記事文章コーパス）の音声サンプルセットで以下に示す通りである。

［データ量］文音声サンプル（28152、52 時間）、
音節種類数（225）、音素種類数（43）、
トライフォン種類数（7241）、
SPS 種類数（411）

ここでは、再現率（横軸）―精度（縦軸）の 2 次元上に性能を示している。図において、右上方向に（折れ線）グラフがある方が高性能であることを示している。この実験規模では、筆者らの提案する SPS がもっとも高い性能を示している。（一般に性能が高いといわれているトライフォンは制約がきついので、テストデータと学習データが良く適合する場合には高い性能を示すと推測される。）

５．語彙制約のない音声語検出システムの開発

第３、４章で述べた枠組み沿って開発している応用システムとして、語彙制約のない音声語検出システム

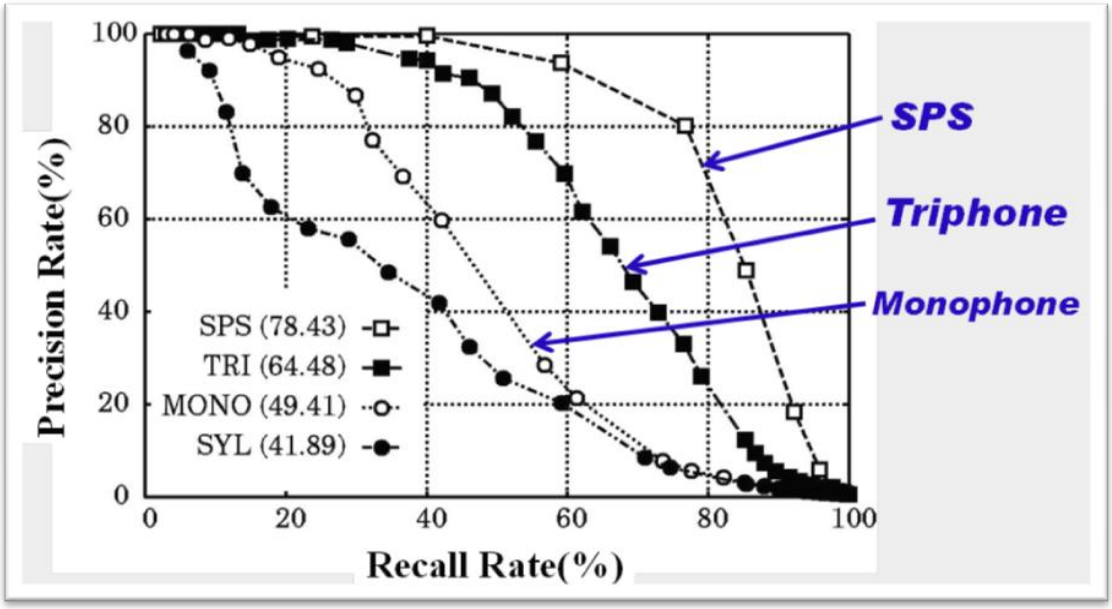

図４　音声認識における基本認識単位（音声符号系）と音声検索性能比較

(spoken term detection system)がある[3]-[5]。このシステムは従来のシステムの範疇としては音声検索に類似するが、その機能は、**４章**で述べたように、文書検索ではなく発話された音声語のピンポイントの検出である。

第３章で述べた枠組みにしたがって、検索される側の音声 DB を一旦、音声記号レベル（SPS）へ変換する。音声から SPS 記号列に変換する際は、既存のテキストコーパスを利用して SPS に関する bigram を用いているが、それ以外には辞書的なデータは利用していない。すなわち、語彙制約のない（open-vocabulary）の検索が可能である。記号領域においては、ブロック化転置インデックス法により、予め検索用インデックスを作成しておき、高速なキーワード候補検出を実行する[6]。さらに SPS 符号間距離行列を利用した距離計算に基づいて候補の絞り込みを行うこともできる。

システム構築においては、基本的には、SPS の音響モデルを学習するための音声データが必要である。精度を上げるためには、この部分を使用環境に合わせる必要がある。しかし、一旦、作成すれば、言語モデルなどを持たないので、以後のメンテナンスはほとんど必要ないといえる。**２章**で述べたように、通常の大語彙音声認識システムには単語辞書と言語文法が不可欠であるが、検索では、人名、地名、曲名など固有名詞がキーワードとなることが多く、辞書中には存在しない未知語が多数現れるため、大語彙音声認識技術を単純に適用するのは困難を伴う。

開発している応用システムは、動画コンテンツに含まれる音声トラックの音声を検索対象として、ユーザの発したキーワード（音声またはテキスト）に適合する動画をピンポイントに検索し、ユーザに提示するものである。**図５**（a）にその概要をブロック図で示す。現在、産業技術総合研究所で開発したシステム（ネット上のニュース放送映像を音声キーワードまたは文字テキストで入力して検索する試作システム）が以下の Web サイトに載っており、テストすることができる。

http://www.voiser.jp/

このシステムを使用したときの出力画面例を**図５**(b)に示す。このシステムではキーワードに一致した音声が入っている区間をスコア順にサムネイルで表示し、画面上で選択した映像区間を再生できる。こうした応用では、どのような用途に利用するかなどを考慮したユーザインタフェースの作り込みが重要になる。

６．カバー曲を含む演奏ストリームからの楽曲名検出システム

音声と同様な音響情報である音楽の検索にも、本稿で述べた方式と手法が適用できる。具体的には、音響情報を音楽データに適合するような音響セグメント系へ符号化した上で、セグメント間の距離と系列間のマッチングに基づいて、入力音楽データの検索を行うシステムである。ただし、部分楽曲を入力してその曲名を出力するシステムは、すでにいくつか製品化されており、ここでは同様な機能を持つシステムの開発には触れない。

ここで提案するシステムは、入力がカバー曲を含む演奏ストリーム（時間的長さに制約はなく、多くは即興的に演奏される多数の部分楽曲の連続信号）から、

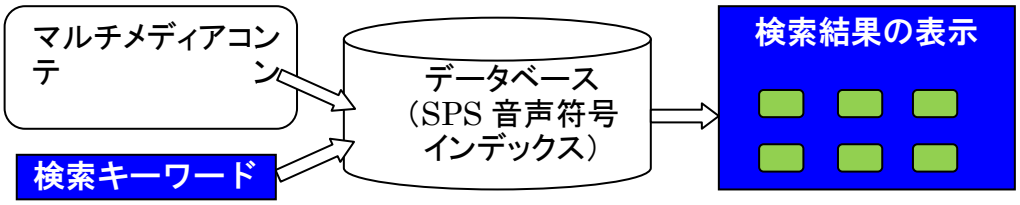

図５(a)　検索キーワードを入力してビデオデータ等の検索を行う音声語検索システム

図５（ｂ）試作システム www.voiser.jp により出力された検索結果例の画面

これに含まれる楽曲名をほぼリアルタイムに検出する機能を有するものである。**図６**は、その機能の概要を示したものである[7],[8]。近年、音楽のライブ演奏の配信などが盛んになりつつあり、本システムの機能はこのような状況における音楽の著作権処理を促進する観点からも有用であると考えられる。**図７**にシステムのブロック構成を示すが、基本的には音声処理の場合と同様である。現時点では、特徴抽出等に不十分な面があり、満足できるだけの性能は得られていないが、上記のような応用のためには、必ずしも高い性能が無くてもその抑止効果が期待できる。

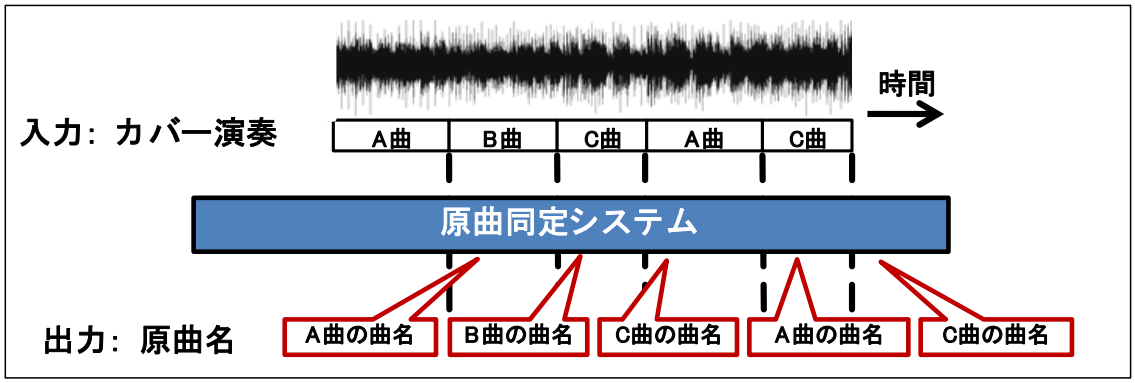

図６　カバー演奏ストリームから原曲名を同定して出力するシステムのイメージ

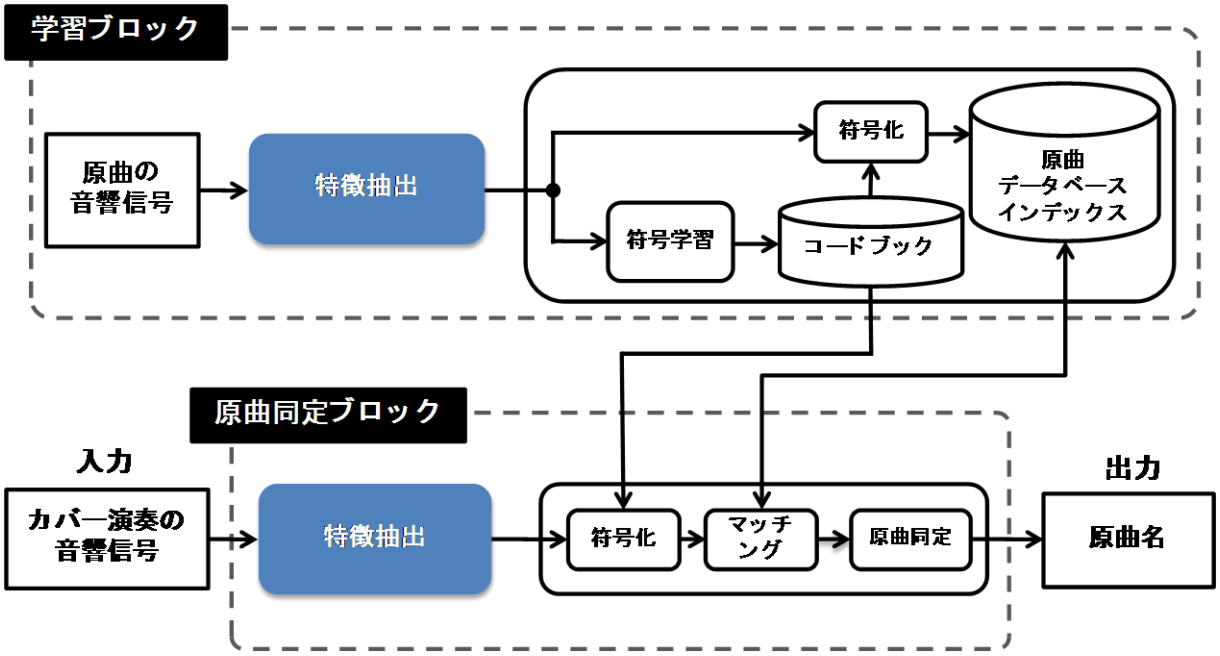

図７　図６の機能を実現するために開発中のシステムのブロック図

７．あとがき

音声認識技術を応用システムに適用する上では、その使用環境条件が非常に限定されている場合を除くと、様々な変動や外乱が混入することは避けられず、認識誤りはかなりの程度許容する必要がある。応用システムを開発するにあたっては、これを前提にそれでも有用性が確保できるというシステム全体の設計、ないし見通しが必要である。ここで紹介したシステムは、システムを構築する際に必要とされる負担が比較的少なくて済むもので、応用システムについて多くのアイデアを生むものでないかと考えている。

終わりに、本稿で述べた音声語検索システムは、産業技術総合研究所・李時旭研究員、岩手県立大学・伊藤慶明准教授との共同研究の成果であり、お二人のご尽力が大きい。

参考文献

(1) JEITA 音声入出力方式標準化専門委員会 WWW URL: http://home.jeita.or.jp/is/committee/tech-std/std/com03.html (2011/11/1).

(2) Tanaka, K., Kojima, H., Fujimura, N., Itoh, Y., “Constructing speech processing systems on universal phonetic codes accompanied with reference acoustic models,” *Proc. of International Conference on Pattern Recognition (ICPR2002)*, **Vol.Ⅲ,** pp.728-731, (Aug. 2002).

(3) 田中和世、“軽い音声認識システムの開発と課題” 電子情報通信学会技術研究報告（招待講演論文）**SP2009-48**、pp.43-47、(2009-7)。

(4) Lee, S.W., Tanaka, K., Itoh, Y., “Combining Multiple subword representations for open-

vocabulary spoken document retrieval," *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP2005)*, **Vol.1**, pp.505-508, (Mar. 2005).

(5) Itoh, Y., Tanaka, K., Lee, S.W., "An algorithm for similar utterance section extraction for managing spoken documents," *Multimedia Systems, ISSN:0942- 4962*, **Vol.10, No.5**, pp. 432-443, (May, 2005).

(6) Lee, Shi - wook; Nambu, Yoshiki; Kojima, Hiroaki, Tanaka, Kazuyo; Itoh, Yoshiaki, "Fast subword - based approach for open vocabulary spoken term detection," *Proc. of 20th International Congress of Acoustics*, Paper **No.975**(4 pages) , Sydney, (Aug., 2010).

(7) 深澤友貴、三河正彦、田中和世、"カバー演奏ストリームからのリアルタイム楽曲同定システム"、日本音響学会音楽音響研究会資料、 **Vol.28、No.5、** pp.9-13、(2009-10)。

(8) 中沢彰吾、三河正彦、田中和世、"カバー演奏ストリームからの楽曲同定のための特徴抽出手法の検討"、日本音響学会 2011 年秋季研究発表会論文集、**1-3-18、** (2011-9)。

この研究は、平成１９年度SCAT研究助成の対象として採用され、平成20年度～22年度に実施されたものです。