

音声検索技術とその応用

成田知宏*
花沢利行*
岩崎知弘*

Voice Search Technology and Application

Tomohiro Narita, Toshiyuki Hanazawa, Tomohiro Iwasaki

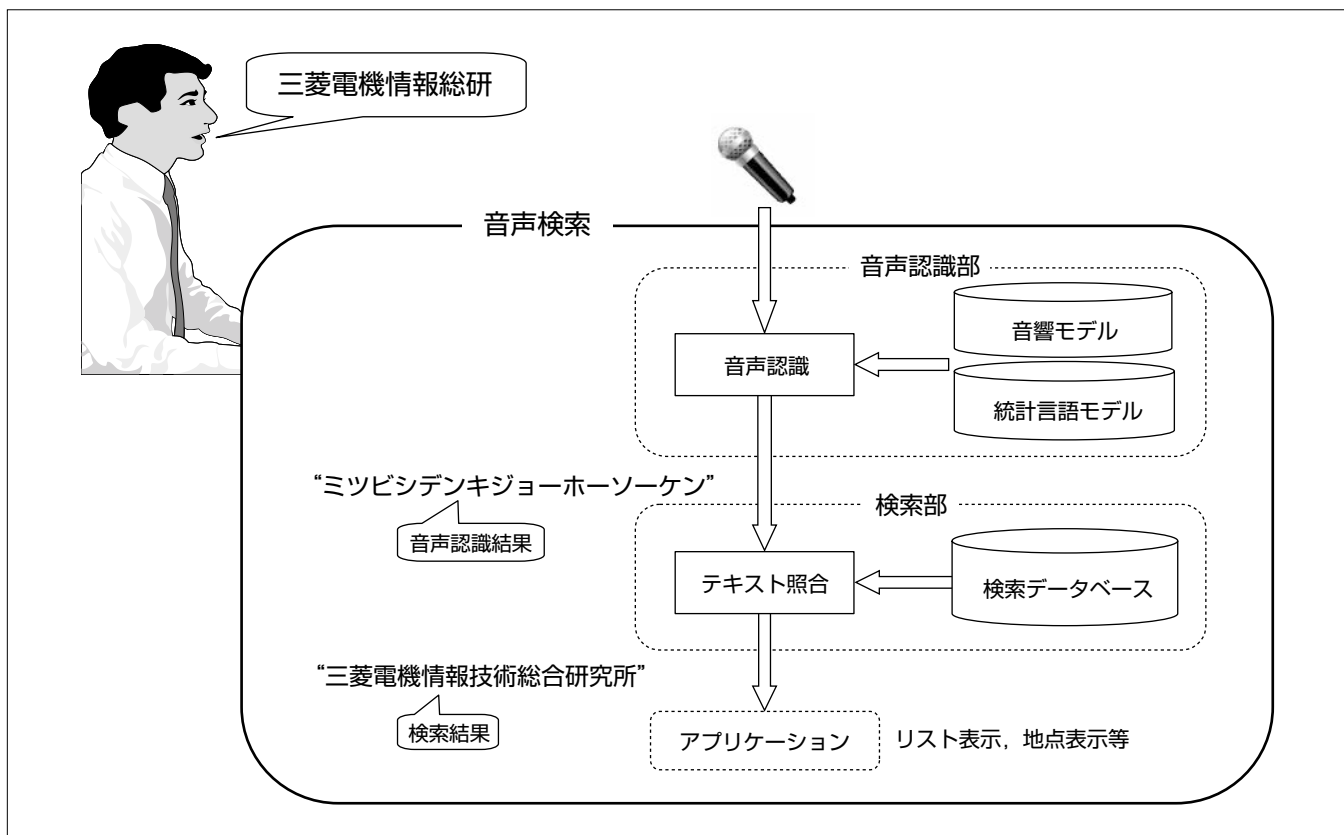
要 旨

音声は人間にとって最も自然な入力手段であり、近年カーナビゲーションやスマートフォン等のコンシューマー向けの製品に音声認識機能が搭載されており身近なものになりつつある。また、製品の高機能化に伴い操作が複雑化していることも、音声による入力手段が注目される要因となっている。特に子供や高齢者等キーボード入力に慣れていないユーザーにとって音声による情報検索(以下“音声検索”という。)は利便性が高く、所望の情報を得るまでの時間を短縮できる。しかし、ユーザーは検索対象の正式名称を覚えていないことが多いため、正式名称以外の略称や通称に対応する必要がある。三菱電機ではこの課題に対し、どの音に近いかという音響的な探索と、どの言葉に近いかという言語的な探索を分離するアプローチに基づく音声検索方式を開発した。

開発した音声検索方式は下図のように、音声認識部と検索部で構成している。音声認識部は音響モデルと統計言語モデルを用いて入力音声を変換する処理を行う。検索部は音声認識結果の文字列を検索クエリとして、検索データベースからクエリに合致する情報を検索する処理を行う。この方式は次の特長を持っている。

- (1) 多少の認識誤りがあっても検索可能
- (2) 部分的な発話や倒置表現でも検索可能
- (3) 組み込み機器の限られたメモリ量、演算量で大規模なデータを検索可能

この方式は、カーナビゲーションにおける施設名検索に適用している。今後マニュアル検索、テレビの番組表検索等に広く応用が期待されている。



音声検索方式

音声認識部では音響モデルと統計的言語モデルを用いた音声認識を行い、入力音声を変換する。検索部では音声認識部で得られた文字列を検索クエリとして検索データベースから所望の情報を検索する。この図のように“三菱電機情報総研”という略称から“三菱電機情報技術総合研究所”を検索することができる。

1. ま え が き

モバイル端末や車載端末のストレージ容量の増加，通信インフラのブロードバンド化が進むにつれて，大規模な情報を音声で検索したいという要求が高まっている。特に子供や高齢者等キーボード入力に慣れていないユーザーにとって音声検索は利便性が高く，所望の情報を得るまでの時間を短縮できる。

本稿では音声検索を構成する音声認識部及び検索部について述べ，応用例と今後の展望を述べる。

2. 音声検索の課題とアプローチ

音声検索の課題として，ユーザーが検索対象の正式名称を覚えていない，覚えていたとしても広く普及している略称や別名が用いられるという問題がある。例えば，“三菱電機情報技術総合研究所”は略称の“三菱電機情報総研”と呼ばれることが多い。この場合，従来方式では正式名称の“三菱電機情報技術総合研究所”のみが音声認識語彙として登録されるため，正しい音声認識結果を得ることができない。この問題に対しては，事前に様々な言い換え表現を作成しておくアプローチ⁽¹⁾があるが，認識語彙数の増加に伴い言い換え表現の追加・整備コスト及び演算量が増加するという問題があった。

これらの問題に対して当社では，どの音に近いかという音響的な探索と，どの言葉に近いかという言語的な探索を分離するアプローチを採用している。音響モデルと統計言語モデルを用いて音声から文字列への変換を行う音声認識処理と，音声認識のポスト処理としての検索処理に分離することで，語彙数が大幅に増加した場合でも演算量の増加を抑制することが可能になった。また，認識処理と検索処理を分離したことで，多少の音声認識誤りに対しても頑健で，部分的な発話や倒置表現でも検索可能な機能を実現している。

3. 音声検索方式

図1に開発した音声検索方式の構成を示す。音声認識部では音響モデルと統計言語モデルを用いて音声認識を行い，入力音声を文字列に変換する。検索部では音声認識部で得られた文字列を検索クエリとしてデータベースの検索を行う。検索部は音声認識部とは独立な構成となっており，タッチパネルなどの音声以外の入力結果に対しても検索が可能である。これより先，音声検索対象が施設名の場合を例に音声認識部と検索部の詳細について述べる。

3.1 音声認識方式

音声認識部で用いる音響モデルは音響特徴量の統計量を表したもので隠れマルコフモデルと呼ばれる確率モデルで表現される。統計言語モデルは認識対象語彙のデータであり，単語間の接続関係を表現したものである。

3.1.1 音声認識の基本方式

まず，入力信号に対してパワーやその他の情報に基づき音声区間の検出を行う。次に検出した音声区間に対して一定時間ごとに特徴量の抽出を行う。特徴量には音声スペクトルの概形を表現するパラメータが用いられる。次に音響モデルの尤度(ゆうど)計算を行う。尤度計算は入力信号に対するモデルの尤(もっと)もらしさを計算する処理である。一般的に音響モデルの最小単位は前後の音の違いを考慮し音素の特徴を表現するトライフォンと呼ばれるモデルが使用されるが，必要なメモリ量と演算量が大きいという問題がある。そこで，音素片と呼ばれる独自のコンパクトなモデルを採用することで，省メモリ化と少演算量化を実現している。最後に統計言語モデルを利用したデコード処理が行われ，入力音声に対する尤もらしさ(認識スコア)が一番高い単語列を認識結果として出力する。

3.1.2 統計言語モデルを用いた音声認識

2章で述べたように言語モデルに“東京国立近代美術館”

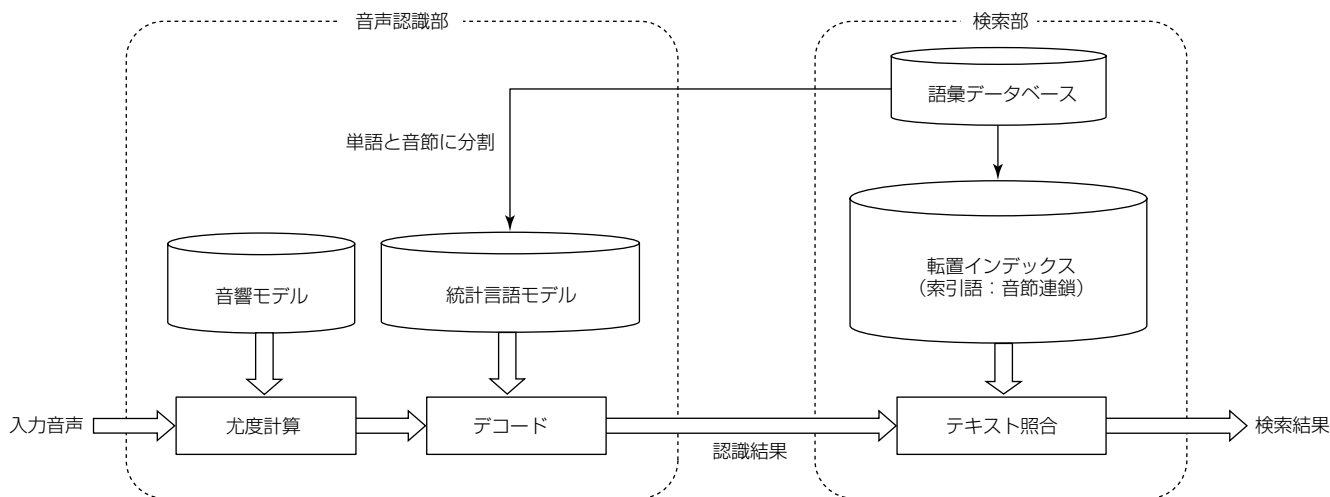


図1. 音声検索方式の構成

だけを登録した場合，“東京近代美術館”のように“国立”を省略した発話は認識できない。対策として“東京国立近代美術館”を“東京”“国立”“近代”“美術館”というように単語に分割し、各単語とともに単語間の接続規則を言語モデルに登録する方式がある。例えば，“国立”は省略可能という規則を登録すれば，“東京近代美術館”という言い換え表現も認識可能となる。しかし、多様な呼称をカバーする規則を手で整備するのは困難である。

これに対し統計言語モデルでは単語間の接続のしやすさを統計量に基づく確率で表す。例えば、先に述べた例の場合，“東京”という単語の次に“国立”や“近代”が続く確率が登録されており、次に続く可能性のある単語を予測しながら認識を行う。単語間の確率はデータベースから自動的に算出が可能である。また、データベースに出現しない単語接続に対しても、近似的な確率を与えることで多様な呼称に対応することが可能である。

ただし、検索対象が数十万語の規模になると、検索対象に含まれる単語数も数万語のオーダーとなるため、音声認識時のメモリ量と演算量が問題となる。そこで、各単語をより小さな単語に分割して表現する方式⁽²⁾を採用している。例えば，“美術館”は，“びじゅつ”と“かん”という単語で表現する。単語の総数は施設名中での出現頻度の高い上位N個に制限する。単語の連鎖で表現できない場合は、音節の連鎖で表現する。例えば，“近代”は，“き”“ん”“だ”“い”という音節の連鎖で表現する。このように音節を単位とすることで、検索対象の施設数が増加しても認識単語数を一定数に保ち、メモリ量及び演算量に制限のある組み込み機器でも大規模な音声検索を可能にしている。

3.2 検索方式

3.2.1 テキスト照合に基づく音声検索

3.1.2項で述べたように音声認識部では施設名を単語と音節に分割して認識する。認識結果は登録施設名と完全一致しない場合が生じるので、認識結果の音節列と最も良くマッチする候補名を検索する方式⁽²⁾を採用している。検索部では図2に示す転置インデックスを用いて、各施設に対する検索スコアを算出し、スコア最大の施設を検索結果とする。索引語としては図に示すとおり2音節の連鎖を用い、索引語ごとに、その索引語を含む施設のID番号を保持する。

例えば、認識結果の音節列が“しいばら”の場合，“しい”“いば”“ばら”の3種類の音節連鎖が存在する。次に転置インデックスを参照して、これらの音節連鎖を含む施設名に対し、音節連鎖1個ごとにスコアを1点加算する。図2の例では、施設ID=2“しいばらだいす”のスコアが3となり、検索結果の第一位として出力される。このように転置インデックスを参照することで、高速にスコアを計算することができる。なお、音節連鎖を索引語とする利点は、単語や

語彙データベース

施設名ID	施設名の音節列
0	よ こ は ま す う ら し あ
1	よ こ は ま ぎ ん こ う
2	し い ば ら だ い す



自動生成

転置インデックス

索引語： 音節連鎖	施設名ID
よ こ	0, 1, ...
こ は	0, 1, ...
は ま	0, 1, ...
し い	2, ...
い ば	2, ...
ば ら	2, ...
：	...

図2. 転置インデックス

単語連鎖と比較して音声認識誤りに頑健なことである。例えば，“東京近代美術館（とーきょーきんだいびじゅつかん）”という発話に対し、音声認識部の出力結果が“東京近代美術館（とーきょーきんだいぶじゅつかん）”となった場合、単語単位では誤認識した“美術館”が“美術館”とはマッチせずスコアに寄与しないが、音節単位なら“じゅつ”“つか”“かん”の3個の音節連鎖が“びじゅつかん”とマッチするため、正解の“東京近代美術館”のスコアに寄与できる。

3.2.2 必須(ひつす)形態素の検証

3.2.1節で述べた検索方式の問題点の一つとして不自然な検索結果が湧き出す場合がある。例えば，“清水砂丘（しみずさきゅう）”という発話の音声認識結果が“しみずさき”のように誤認識を含む場合，“A書房弘前清水3丁目店（/えー/しょ/ぼー/ひ/ろ/さ/き/し/み/ず/さ/ん/ちょー/め/て/ん/）”のほうが，“清水砂丘（/し/み/ず/さ/きゅう/）”よりもマッチする箇所(下線部)が多い(/は音節の区切りを表す)。このため、前者が正解よりも上位候補として誤検索されるという問題点があった。

“A書房”などのブランド名を含む施設を検索する場合、ブランド名は省略しないと考えられる。そこで、施設名ごとにその施設を検索する際に省略しないと考えられる文字列(以下“必須形態素”という。)を抽出した必須形態素辞書を事前に作成しておき、検索結果のうち必須形態素を含む施設名に対し、入力音節列中にその施設の必須形態素が含まれるか否かを検証する方式⁽³⁾を開発した。そして、必須形態素が含まれない場合、検索スコアに事前に定めたペナルティを付与する。なお、必須形態素の検証には置換、挿入及び脱落を許す文字列マッチングを用い、誤認識で音節表記が完全一致しない場合でも検証可能とする。

このような必須形態素の検証処理によって、認識結果の音節列“しみずさき”は“A書房弘前清水3丁目店”の必須形態“えーしょぼー”を含んでいないため、検索スコアにペナ



図 3. HDDカーナビゲーションシステムNR-HZ001

ルティが付与されることになり、正解の“清水砂丘”より上位に検索されることを抑制できる。

4. 音声検索の応用例

ここまで述べた音声検索技術の応用例と今後の展開について述べる。

4.1 カーナビゲーションの音声50音検索

音声50音検索とは、音声で施設名を入力し検索を行う機能であり、当社のHDD(Hard Disk Drive)カーナビゲーションシステム“NR-HZ001(図3)シリーズ”に搭載されている⁽⁴⁾。従来のカーナビゲーションシステムでは50音検索の入力画面(図4)上でタッチパネルから1文字ずつ施設名の読みを入力する必要がある入力に時間が掛かっていた。さらに、走行中はタッチパネル操作が禁止されるため、停車中以外は入力することができなかった。NR-HZ001では音声50音検索を導入することで、所望の施設を検索するまでの時間を短縮するとともに、走行中での50音施設検索を可能にしている。この製品の音声50音検索には次の特長がある。

- (1) 都道府県単位で最大100万件規模の超大語彙の施設検索が可能
- (2) 部分表現や倒置表現等、正式名称以外の表現での施設検索が可能
- (3) 追加発話による施設の絞り込みが可能

追加発話による絞り込みの例を次に述べる。まず、“三菱電機”と発話すると“ミツビシデンキ”を検索クエリとして検索が行われ、検索結果がリスト表示される。次に絞り込みのために“研究所”を発話すると、“ミツビシデンキ”と“ケンキュウジョ”を検索クエリとして再検索が行われ、施設名のリストが更新される。このように複数の発話によって逐次的に施設を検索することができる。

4.2 今後の展開

本稿で述べた音声検索技術は次のような様々な製品や情報システムに適用可能である。



図 4. NR-HZ001の50音検索画面

- (1) 製品マニュアルの検索
- (2) 番組表の検索
- (3) コールセンターでの顧客情報の検索
- (4) 点検業務における業務データの検索

ただし、これらの製品・システムは各々機能要件が異なるため適用の際には別途検討が必要である。例えば、使用時の騒音環境、使用可能なメモリ量及び演算量、要求性能、応答速度等が要検討項目に該当する。今後はこれらの製品・システムに対する音声検索のニーズを探るとともに、必要な要素技術の検討を進める予定である。

5. む す び

音声検索を構成する技術要素と応用例について述べた。大規模データの検索では、入力時間及び入力自由度の面から音声は最も適した入力手段であると考えられる。今後は様々な製品やシステムへの展開を検討するとともに、意図理解技術や音声対話技術を統合することで、より直感的で自由な音声検索の実現を目指す所存である。

参 考 文 献

- (1) 岩崎知弘，ほか：カーナビゲーションシステムの音声インタフェースー現状と将来展望ー，三菱電機技報，78，No. 9，613～617（2004）
- (2) 花沢利行，ほか：統計言語モデルによる音声認識とテキスト照合に基づく大語彙音声検索，日本音響学会研究発表会講演論文集，61～62（2009）
- (3) 花沢利行，ほか：必須形態素の検証によるテキスト検索の改善，電子情報通信学会総合大会論文集D-14-7（2011）
- (4) 成田知宏，ほか：カーナビゲーション向け音声認識の新たな展開，三菱電機技報，84，No. 9，495～498（2010）
- (5) 岡登洋平，ほか：統計言語モデルによる音声認識とテキスト照合に基づく大語彙音声検索の精度改善検討，日本音響学会研究発表会講演論文集，99～100（2010）