Vol.2011-SLP-88 No.5
2011/10/28

音声検索実用化の現状と課題

大淵 康成, 神田 直之[†1]

社会の様々な場面で、大規模な音声データが蓄積されるようになっている。これらのデータを有効活用するため、音声認識技術を活用した「音声検索」への期待が高まっている。既に様々な分野で実用化が進められている中で、いわゆる音声認識率のみならず、様々な観点での性能を高めることが求められている。本稿では、音声検索の様々な実用化例を紹介するとともに、処理速度やデータサイズなど、音声検索を取り巻く多様な評価要素について、現状と課題を述べる。

Practical Applications of Spoken Term Detection: Current Status and Issues

Yasunari Obuchi and Naoyuki Kanda[†1]

Accumulation of large-scale speech data has been becoming popular. Spoken term detection (STD), which is an application of speech recognition technology, is an important tool to utilize these data. STD applications are spreading into various fields, and their performances are evaluated not only by the recognition rate, but also by many other factors. In this paper, we introduce some examples of the practical use of STD, and describe the current status and issues of those factors, such as the processing speed and data size.

1. はじめに

コンピュータによる音声の認識は、単純な認識精度という観点では、いまだ人間には及ばない。しかし、低コストで大量の作業が可能であるというコンピュータの強みを活かすことができる状況においては、人間による作業を代替させることにより、大きな利益を得ることができる。例えば、数百時間の音声データを聴取して特定の音声を抽出するという作業は、人間がやれば数十万円の人件費が必要になってしまうが、コンピュータであれば僅かな電気代だけで済む。仮にコンピュータによる処理が完璧では無いとしても、人間による数時間の後処理で対応することが可能であれば、大きなコスト削減となる。

このように、大量に蓄積された音声データの自動処理、なかでも特定のキーワード発話を見つけ出す音声検索[⋆1]の枠組みに注目が集まっている[1),2)]。音声検索技術による大量の蓄積音声データの活用は、現在人間によって行われている聴取業務のコストを削減するのみならず、これまで解析不可能として廃棄されていた様々な音声データの再活用を可能にし、新たなビジネスチャンスを開拓するためのキー技術ともなりうる。

本稿では、音声検索の様々な応用例を紹介するとともに、そうした実用システムに求められる様々な性能要件を示す。音声検索精度そのものに関しては、必ずしも100%に近い値が要求されるとは限らない一方で、大規模データを扱うシステムならではの多様な要求仕様が存在する。以下では、そうした要件を満たすための技術的な取り組みについても述べる。また、音声検索技術の発展・普及により、今後さらに拡大が予想される応用場面についても検討したい。

2. 音声検索の基本構成

音声検索システムの代表的な構成を、図1に示す。一般的に、大規模データに対する検索を行う際には、検索対象となるデータを登録する際に、インデキシングと呼ばれる処理によりインデクス（索引）データを作っておく。このインデクスデータを活用することにより、検索実行時に、検索クエリが与えられてから極めて短い時間で検索を実行することが可能になる。なお、インデクスデータのみにより検索が実行可能な場合であっても、人間による聴取確認のため、圧縮した音声データを保持しておくことが多い（図の一番上のフロー)。

音声検索システムの実装方式はいくつかあるが、代表的なものの一つが、大語彙連続音声認識とテキスト検索の組合せである。図1で “Search” と書かれたブロックの中の、1番上の処理フローがこれに対応する。この方式で用いるテキストデータを作成するため、インデキシング時には、特徴抽出～尤度計算～認識の一連の流れにより、入力音声を完全にテキス

†1 日立製作所中央研究所
Central Research Lab., Hitachi Ltd.

⋆1 「音声検索」という語は、「音声データを対象とする検索」と「音声でクエリを入力する検索」の両方の意味で使われるが、本稿が扱うのは前者である。正確を期する場合には、”Spoken Term Detection”の訳語として「音声中の検索語検出」という表現を用いる場合もあるが、本稿では、ビジネスシーンでの使用を想定し、「音声検索」で統一する。

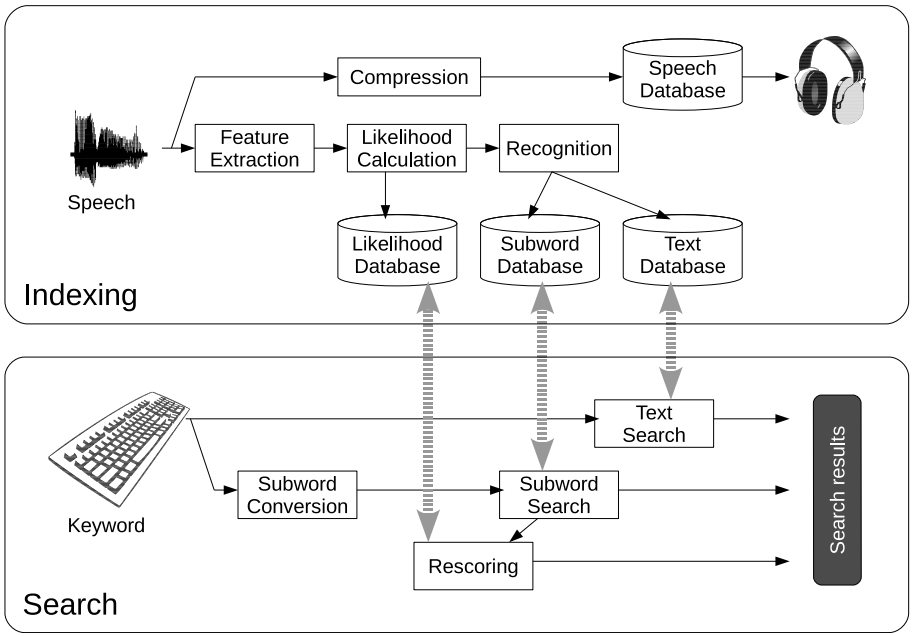

図 1 音声検索システムの基本構成図
Fig. 1 Basic Structure of Spoken Term Detection System

ト化しておく。検索実行時には、蓄積されたテキストデータに対し、一般的なテキスト検索を行う。一般に、テキストデータの検索は低処理量で実行可能なことから、検索の高速性を簡単に確保できるというメリットがある。また、良質の学習データから作成した言語モデルが利用可能な場合には、高い検索精度を得ることができる。

一方、大語彙連続音声認識に基づく手法のデメリットとして、インデキシングの処理負荷の重さと、言語モデル不適合時の性能劣化とがある。前者は、インデキシング時に実行する大語彙連続音声認識の処理負荷が重いことによるもので、マシン性能によってはリアルタイムのデータ登録ができないこともある。後者は、言語モデル適合時の高い検索精度の裏返しであり、言語モデル学習時と音声データ取得時のタスク間に不適合がある場合、性能が大きく劣化する。極端な例としては、言語モデルを構成する単語辞書に含まれない語をキーワードとした場合、原理的に検知不可能となってしまう。

これらの問題を解決するアプローチとして、サブワードに基づく方式がある。この方式では、キーワードを音素・音節などのサブワードの系列として表現し、蓄積データの中からそのような系列を検出する。この場合、インデキシング時に認識結果としてサブワード系列を保持しておくことになる。図 1 の “Search” ブロックの上から 2 番目の系列がこれに対応する。しかし、一般に音素や音節の認識精度は低く、そのままでは十分な検索精度が得られない。そこで、サブワード系列の検索により得られた候補に対し、音響モデルそのものを用いた再照合を行うことにより、検索精度を向上させる（図 1 “Search” の 3 番目のフロー）。一般に再照合は音響モデルのレベルで行うため、特徴抽出や、各モデルに対する尤度計算まではインデキシング時に終わらせておくことが可能である。なお、音響モデルレベルでの再照合は処理負荷が高いため、大規模データに対して実用的なレベルの反応時間を実現するためには、前段階でどれだけ候補の数を絞っておけるかという点が重要である。処理速度と検索精度の最適なバランスを得るために、2 種類の異なる再照合手法を組み合わせる方法もある[3]。

検索速度や検索精度の優劣の他に、サブワード方式に固有の問題として、同音異義語の区別ができないという問題がある。例えば、「京都 (kyo:to)」というキーワードから得られる検索結果に、「教徒」という語が含まれてしまうことは避けられない。さらに、「東京都」「今日と明日」などのように、複数単語にまたがるサブワード系列も検出されうる。大語彙連続音声認識方式では、前後の文脈によりこれらを識別可能である。

3. 音声検索の実用例

音声検索の対象となる大量の音声データを蓄積している代表的な状況として、コールセンターがある。近年のコンプライアンス意識の高まりにより、「言った、言わない」のトラブルを避けるため、コールセンターでの全会話音声を保存しておくケースが増えている。これらのデータを用いて、オペレータの管理・教育などを行うことも試みられており、コールセンター運営業者の 64%がリアルタイムのモニタリングと録音装置を併用しているといったデータもある[4]。しかし、管理者によるデータの聴取は負担が大きく、全データをチェックすることは不可能に近く、音声検索技術に対する期待は高い。とりわけ、コールセンターの業務は個別のタスクに特化していることから、汎用の言語モデルを用いた大語彙連続音声認識方式との相性はあまり良くない。個別のタスクに対するチューニングのコストをかけたくない場合、メンテナンス不要のサブワード方式が有効である。

インターネット上のコンテンツに対する検索の需要も、近年では高まっている。動画共有

サイトなどでは、大語彙連続音声認識を用いた自動字幕作成機能が公開されている例もある。ニュース映像など、既存の言語モデルとの相性が良いものに対しては、かなり正確な字幕が作成され、検索も高精度で行うことができる。一方、様々なユーザーによって投稿される不特定の内容の動画に対しては、高い認識率を得ることは難しい。

放送局などのように、既に大量の音声・動画データを蓄積している機関でも、音声検索に対するニーズは高い。こうした機関では、手持ちの大量データに対するインデキシングを一括して行う必要があることから、特にインデキシングの速度に対する要求が高まる傾向がある。

テレビやラジオを受信し、その音声に対して検索を行いたいという需要もある。例えば、企業が自らの会社名や製品名をキーワードとして検索を行い、会社の評判や市場動向を調査したいというようなケースである。もちろん、一般の視聴者が、大量に録画したコンテンツを検索したいという場合もある。後者については、未編集のホームビデオなどにも同様のニーズがあると思われる。

ここまでに挙げた応用例は、主として接話マイクにより収録された音声を対象としたものであるが、遠隔マイクの音声に対する検索が可能になれば、さらに応用は広がる。例えば、監視システムと組み合わせれば、注意すべき事象の検知精度を上げることができる。会議の音声をすべて録音し、音声検索を議事録作成支援に用いるという応用もある。さらには、店頭における営業活動などでも、コールセンターと同様のモニタリングができるようになるかもしれない。雑音環境下での音声認識は困難な課題であり、大語彙の自由発話を対象に高い認識率を得ることは難しいが、検索という切り口に着目することで、大きく市場が広げられると期待されている。

4. 検索実行時の性能指標

4.1 検索精度

音声検索においても、音声認識の場合と同様に、どれだけ正確にキーワードを検知できるかということが、最も重要な性能指標となることは間違いない。情報検索全般と同様に、適合率 (precision) と再現率 (recall) の二つの指標で性能を定義することができる。前者は、検知されたキーワード発話候補のうち正解の占める割合であり、後者は、対象データ中に存在するキーワード発話のうち正しく検知されたものの割合である。

サブワード方式の音声検索では、閾値の調整により検知する発話数を調整することができるため、適合率と再現率を滑らかに変化させることが可能である。この様子は、通常 ROC(Receiver Operating Characteristic) 曲線によって表される。例えば、業務における NG ワードのように、重要な発話を漏らさず検知する必要がある場合には、再現率優先の設定のもとで、検索結果を人間が再確認するという方式が採られる。一方、ウェブ検索のような情報取得目的で使用する場合には、上位候補のみをチェックするのが普通であり、適合率が重要な指標となる。また、これらいずれの場合でも、操作する人間がどれだけの時間を有しているかにより、適合率と再現率のバランスを変えられることが望ましい。

通常の大語彙連続音声認識方式では、認識結果のテキストを一意に決定するため、単一の適合率・再現率しか得ることができない。ただし、大語彙連続音声認識でも、音声認識信頼度を保持したり、複数の認識結果をネットワークなどの形で保存しておくことにより、適合率・再現率をある程度調整可能にすることもできる。

適合率・再現率は、可変の閾値をどう設定するかにより変わる指標であるが、アルゴリズム自体の性能を示す指標としては、両者の値が等しくなるような設定での値を用いることもある。この値を Break Point と呼び、このときの誤検知率を EER(Equal Error Rate) と呼ぶ。Break Point の値は、「対象データ中に含まれるキーワード数と同じ数の候補を提示した際の正解率」と表現することもできる。また、より現実的な値として、1 時間当たりの誤検出数が 0 個から 10 個の場合の再現率の平均を、FOM(Figure Of Merit)[5] と定義して用いることも多い。

なお、音声検索の精度評価を行う場合には、テスト用キーワードをどのように選ぶかが重要である。当然であるが、語彙外単語を含むかどうかにより、大語彙連続音声認識方式の性能は大きく変わる。また、キーワードの長さが偏っていたり、対象データ中の出現頻度が極端に少ない単語が含まれていたりすると、検索精度の誤差が大きくなるという問題がある。

4.2 検索速度

通常のウェブ検索と同じように、ユーザーが端末の前で操作する対象と考えた場合、音声検索システムの応答時間として許容されるのは、せいぜい 2〜3 秒程度であろう。テキストデータを対象とする場合であれば、発話時間にして数千時間相当のデータが対象であっても、この程度の応答時間は容易に実現できる。サブワード方式でも、リスコアリングを行わない場合には、同程度の処理速度が得られる。一方、リスコアリングを行う際には、2〜3 秒の応答時間内で結果を返すために、リスコアリング対象の絞込みや、リスコアリング処理の高速化を行う必要がある。筆者らはかつて、2 段階のリスコアリング方式の組合せにより、実用的な検索精度を保ちつつ、2000 時間の音声データから 3 秒以内で検索結果を返すシステムを実現している[3]。なお、検索速度を決める要因として、メモリ容量やディスクア

クセス速度が重要であることは言うまでも無い。上記の例では、サブワードのインデクスをすべてメモリに読み込むだけでなく、音響尤度データをソリッドステートディスク (SSD) に置くことにより、ランダムアクセスの高速化を図っている。しかし、コストの観点で SSD の使用が難しいケースもあり、実装には更なる工夫が求められている。

なお、音声検索は並列化が比較的容易なタスクであり、大量のデータを N 個のサブセットに分割し、N 台のマシンで検索を行うことにより、高速化を実現することができる。

5. インデキシング時の性能指標

5.1 インデキシング速度

最も単純な音声検索システムでは、インデキシングの速度は Real Time Factor (RTF) に換算して 1.0 以下であれば良い。これは、あるサーバーに絶え間なく入力される音声データを、遅延なくインデクス化することができれば良いという考え方である。しかし、実用場面においては、しばしば当該サーバー上で別のプロセスが実行されることもあり、ある程度のマージンが必要である。更に、コールセンターなどでは、単一のサーバーで複数回線を扱うこともあり、RTF0.5 以下の性能が求められる場合もある。また、既に蓄積済の大量データを対象に検索を行いたい場合には、インデキシング速度の差がそのまま応答時間の差となって現れる。一般に、数十種類程度のサブワードを対象とするインデキシングは、数万単語を対象とする大語彙連続音声認識のインデキシングの数倍は早く、様々な条件下での適用性に優れている。一方で、マルチコアマシンや GPU などの活用による音声認識の高速化の研究[6] も進んでおり、制約条件に応じた使い分けが可能である。

5.2 インデクスサイズ

大規模コールセンターなどにおいては、日々数十から数百時間分の音声データが蓄積されており、ストレージシステムの廉価化が進んだ今日においても、データ削減はなお重要な課題である。人間による聴取用の音声は 8kbps 程度まで圧縮可能であり、それに比べてインデクスの容量が大きくなる場合には、削減が求められることも多い。大語彙連続音声認識方式で用いるテキストデータは、プレーンテキストでも 1 分あたり数百〜数千バイト程度であり、まったく問題にならない。これは、サブワード単位の認識結果でも大きくは変わらない。一方、リスコアリングのための音響尤度データをすべて保持しておく場合、フレーム数×モデル状態数の数値データを保持する必要があり、例えばフレームレート 100Hz で、状態数 2000 の音響モデルの尤度をすべて 4 バイトで保持した場合、800kB/sec となってしまう。実際には、無音区間を除いたり、閾値以下の尤度を持つ状態はすべて縮約したり、尤度を表すビット数を減らすなどして、5〜10kB/sec 程度まで削減することが可能であるが、それでも容量が問題になるようなケースでは、サブワード＋リスコアリング方式の適用は難しくなる。

6. おわりに

大量の音声データの中から特定のキーワード発話を検知する、音声検索技術の実用化が進んでいる。インターフェースとしての音声認識に比べると、精度という観点での要求が若干緩くなる反面、大量のデータを扱うための高速性や、蓄積するインデクスデータのサイズなど、様々な側面での改良が求められる。既に蓄積されている音声データの活用に加え、音声検索技術の活用を前提とした新しいアプリケーションを普及させていくためには、こうした改良を続け、誰にでも扱いやすいシステムを提供していくことが必要である。学会においても、認識精度を上げるだけでなく、様々な側面での改良を進めるような研究が活発に行われることを期待したい。

参考文献

1) NIST Information Access Division: Spoken Term Detection Portal, http://www.itl.nist.gov/iad/mig/tests/std/
2) 西崎博光他: Spoken term detection のためのテストコレクション構築とベースライン評価, 情報処理学会音声言語情報処理研究会, SLP-81-13, 2010.
3) Kanda, N., et al.: Open-Vocabulary Keyword Detection from Super-Large Scale Speech Database, *Proc. IEEE MMSP 2008*, Cairns, Australia, 2008.
4) コールセンター白書 2011, 月間コンピューターテレフォニー編集部編, (株) リックテレコム, 2011.
5) Rohlicek, J.R., et al.: Continuous Hidden Markov Modelling for Speaker-Independent Word Spotting, *Proc. IEEE ICASSP 1989*, Glasgow, Scotland, 1989.
6) Dixon, P, et al.: Recent Functionality Improvements to the T^3 Speech Decoder, 日本音響学会 2009 年秋季研究発表会講演論文集, 3-1-10, 2009