論文 / 著書情報
Article / Book Information

論題(和文)	音声認識の今後の研究課題
Title(English)	
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	NHK技研R&D, Vol. , No. 65, pp. 10-15
Citation(English)	, Vol. , No. 65, pp. 10-15
発行日 / Pub. date	2001, 1

技術動向

音声認識の今後の研究課題

古井貞熙

音声認識技術は，コンピューターシステムへの音声コマンド入力や，さまざまな音声ドキュメントの文字化などへの応用を目的として，近年大きく進展しているが，残されている課題も多い。それらについて解説するとともに，今後の展望を述べる。特に，種々の変動を含む話し言葉に対応できる適応的言語モデルおよび音響モデルの構築，そのための話し言葉の大規模コーパス*の作成，音声から話題語を抽出し，話し手の意図を理解し，要約する機能の実現の重要性などに焦点を当てる。

＊ コーパス
データの集まり。特に原稿のような言語データの集まりで，データベースのようには体系化されていないものをいう。

1. はじめに

音声は，人と人との最も自然で，基本的，容易かつ効率的な情報交換手段である。人とコンピューターとの情報交換においても，コンピューターから情報を受け取るには一般にディスプレイが最も効率的であるが，コンピューターへの入力手段としては，音声が使えれば極めて効率的になると期待される。電話を用いて，音声でコンピューターシステムと対話し，コンピューターシステムから種々の情報を得るサービスが，米国を中心に急速に広がりつつある。今後，PDAのような携帯端末によるインターフェースが主流になってくると，キーボードを装備するわけにいかないので，音声入力の役割が一層大きくなってくると予想される。

日本語文書入力に関しては，優れた使いやすいワープロ（ソフト）が普及しているとはいえ，かな・漢字変換を必要とするため，英語のようにしゃべる早さに近い速度でタイプ入力するのは極めて難しい。音声認識技術によって音声を自動的に文字化することができれば，単に音声ワープロとしての実現形態だけでなく，会議の議事録や講演録の作成，放送ニュースのアーカイブ化への利用など，数多くの応用が考えられる。日常生活の中で，会議，討論，講演，講義など，音声で伝達される情報は極めて多く，その重要性はマルチメディアの世の中でも大きく変わっていない。これらの音声ドキュメントを録音して蓄えることは容易であるが，そのままでは聞き直すしかないので，その利用は極めて難しい。これらを音声認識技術によって，テキスト，要約などの利用しやすい形に変換することができれば，検索も容易になり，その有用性は極めて大きいであろう。

音声認識の研究は，1950年代から現在まで半世紀にわたって，世界中の研究者によって進められてきた。近年，統計的パターン認識理論に基づく技術の発展と，コンピューターの高速化に支えられて，音声認識技術は大きく進展しているが，まだその能力には限界があり，多くの研究課題が残されている。今後は，これらの課題を着実に解決していくことにより，話し言葉に対する音声認識の基本的性能を高めるだけでなく，話し手の意図を理解し，要約したりする機能へと技術の高度化を進めていくことが期待されている。

2. 話し言葉の音声認識の難しさ

現在の音声認識技術で，過去のニュース原稿で学習した言語モデルを用いて，放送音声を音声認識すると，スタジオでアナウンサーが原稿を見ながら発声した音声であれば，90%以上の認識率が得られる。しかし，解説，現場からの中継，対談などの音声に対しては，大幅に認識率が低下する。この原因は，普通の人の発声では個々の音の発音が必ずしも明確でないこと，雑音が重畳していること，原稿を読んでいない自発的な音声の言語的構造は，言語モデルの学習に用いたニュース原稿の文とはかなり異なることなどである。

自発的な話し言葉の音声認識の難しさには，次のようなものがある。

(a) 助詞などの言葉の脱落・省略，間投詞（「あのう」「えー」など）を含む不要語の付加，倒置

(b) 言い間違い，言いよどみ，言い直し，繰り返し

(c) 言語モデルには含まれていない新しい言葉（未知語）の発声

これらの問題への対処法としては，

(i) 単語やフレーズを単位としたスポッティング*1あるいは検出法と，発声確認のアプローチ

(ii) 余計な言葉をスキップしながら文仮説とマッチングする方法

(iii) 未知語検出アルゴリズム

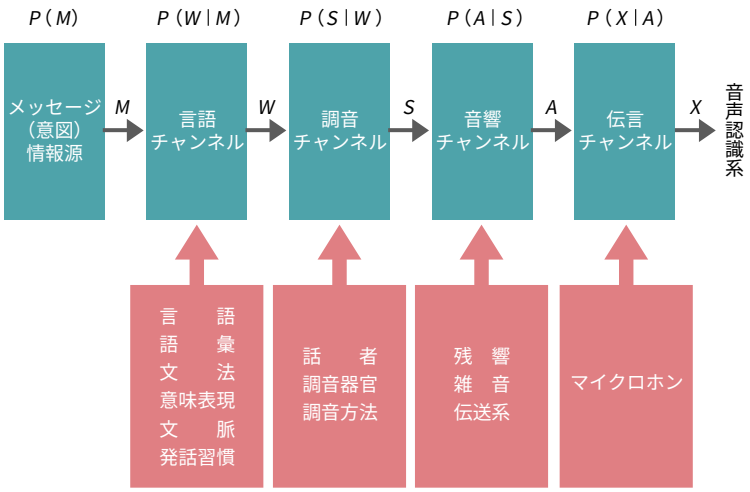
などが研究されている。ただし，現在の中心的方法である統計的言語モデルと整合する有効な方法はまだない。自発的な話し言葉の音声認識では，すべての言葉を文字にしようとするこれまでのアプローチを超えた，新しいアプローチが必要になると思われる。

＊1 スポッティング
普通に話された音声の中から，キーワード部分のみを音声認識すること。

3. 音声認識から音声理解へ

3-1. 意図理解の枠組み

1図に，通信理論のアプローチからの，音声生成モデルを示す[1]。メッセージ（意



1図　通信理論のアプローチからの音声生成モデル

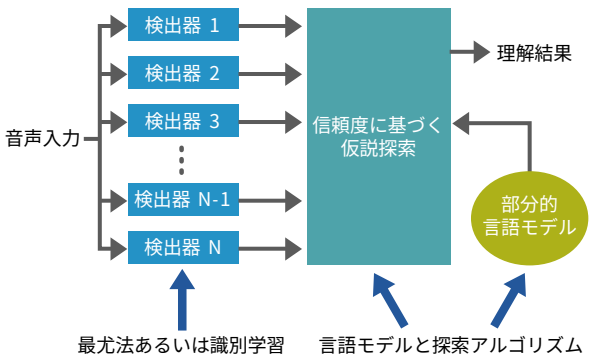
図）情報源で，意図したメッセージMが生成され，言語チャンネルによって単語系列Wに変換される。同じメッセージを表すにも，言語や語彙の違いをはじめとして種々の方法があるので，言語チャンネルは，各言語について代表的な単語列を中心として確率的に機能する。単語列Wは，調音チャンネルによって，音の系列Sとして実現される。調音器官は話者によって異なり，同じ話者でもまったく同じ波形を繰り返すことはできないので，調音チャンネルも確率的に機能する。音の系列Sは話者の口から放射され，部屋の音響特性が畳み込まれたのち，音響入力Aとしてマイクロホンに到達する。この過程をここでは音響チャンネルと呼んでいる。音響入力信号Aは，マイクロホンによって電気信号に変換され，何らかの歪みを受けて，音声認識系にXとして入力される。この最後の過程を，ここでは伝達チャンネルと呼んでいる。以上の各チャンネルには，不確実性あるいは変動があり，$P(W|M)$，$P(S|W)$，$P(A|S)$，$P(X|A)$の確率分布で表現される。$P(M)$はメッセージの事前分布である。

現在の音声認識システムでは，ベイズ決定理論に基づくパターン認識理論に従って，$X=x_1, x_2, \cdots, x_T$が与えられたときの事後確率$P(W|X)$を最大にする単語列$W=w_1, w_2, \cdots, w_k$が選ばれる[2]。これからの話し言葉を対象とした音声認識システムでは，発声者の意図Mを抽出することを目標とするべきであろう。このためには，これまでのように$P(W|X)$を最大化するのではなく，1図に示したモデルで，$P(M|X)$を最大化するMを選ぶことが必要である[1]。これにより，単純に$P(W|X)$を最大化するよりも，Wの正解精度を上げることができる可能性もある。この定式化は，これまでに試みられてきた種々の音声認識方法を包含し，かつ新しい方法を示唆する。例えば，条件付き確率$P(W|M)$の表現法には，これまでに研究された話題別の言語モデル，キャッシュ(cache)モデル，単語の共起関係など種々の方法が適用できる。

意図理解の結果をデコーダーにフィードバックし，その結果に基づいて意図理解をやり直し，この結果を再びデコーダーにフィードバックするという方法で，MとWの両方の精度を上げることができる可能性もある[1]。

3-2. 発声検出に基づく音声理解のアプローチ

不要語などを多数含む話し言葉から，意味のある部分だけを抽出（意味抽出，意味検出）する方法として，単語およびフレーズの検出に基づく音声理解のアプローチが研究されている[3][4]。この原理を**2図**に示す[1]。各検出器では，あらかじめ規定され



2図　発生検出・確認に基づく音声理解システム

た各イベント（音素，単語，フレーズなど）と，その対立モデルに対する尤度(ゆうど)比に基づき，検出判定が行なわれる。モデルの学習には，通常の最尤法あるいは識別学習が用いられる。各検出器の出力は，その信頼度とともに探索部へ送られ，言語モデルとの組み合わせによって音声理解の判定が行われる。

不要語を吸収する音響モデルや，その単語やフレーズの受理を判定するための尺度を適切に構成することによって，良い性能を得ることができることが実験的に示されている。ただしこれまでの実験は，中規模程度の語彙数のシステムに対してであり，大語彙のシステムにどのように適用し，適度な処理量で，これまでの方法の性能をどれだけ上回れるかは今後の課題である。

3-3. 音声理解としての音声要約

発声者の意図Mの表現形態，すなわち音声理解の１つの形態として，音声の要約作成がある。我々は音声中から，話題を担っている重要語（話題語）をできるだけ文法的制約*2に従うように最適に選んで要約文を作成する方法として，各単語の重要度スコア，要約文の統計的言語スコア，さらに音声認識結果の信頼性をもとに，動的計画法を用いる方法を試みている[5]。重要度スコアには，情報検索の分野で用いられているtf-idf尺度*3を変形させた尺度を用いた。統計的言語スコアは，音声よりも比較的簡潔な文体となっていると考えられる新聞の記事から計算した単語トライグラムを用いた。音声認識結果の信頼性は，認識結果として得られる単語グラフから計算される各単語の事後確率を用いた。放送ニュース音声の音声認識結果を用いて，文字数にして20～70%に要約する条件で実験を行っている。

本方法は，ニュースの自動字幕作成への適用のほか，会議の議事録や講演録作成など，今後の種々の応用分野への展開が期待できる。ニュース音声，講演，講義などを音声認識技術を用いて自動的に要約してデータベース化することができれば，情報検索に限らず種々の用途に極めて有用であろう。音声メールを自動的に要約して文字化できれば，内容を一覧するために役に立つであろう。

＊2 文法的制約
音声認識において，正解候補の数を減らすのに利用する，文法に基づいた制約のこと。

＊3 tf-idf尺度
ある文書に対するキーワード（話題語）の重要度を表す尺度。tf尺度とidf尺度の積で表される。
tf尺度とは，ある文書中に出現するキーワードの出現頻度。idf尺度とは，あるキーワードが全文書中のどのくらいの文書に出現するかを表す尺度。

4. 音声の音響的・言語的変動への対処

4-1. 音響モデルの適応化

音声認識で誤認識を生ずるのは，音響モデルや言語モデルの学習に用いたデータと，実際の認識すべき音声の間に，何らかのずれ（mismatch）があるためである。音響的変動の原因には，音声（声質や話し方）の個人差（方言や年令によるものも含む）が極めて大きいこと，同じ人でも体調や精神状態（ストレス）などによって発声速度や話し方が変化することなどがある。多くの場合に，部屋の反響を含む種々の雑音が音声に加わっており，しかもそれが時間的に変化すること，さらにその上にマイクロホンや電話機，伝送系などの歪みが加わることなどもある[2]。

これらの種々の音声の変動をカバーするような大量の音声データを用いて，統計的方法によりモデルを作成すれば，ある程度高い認識精度を達成することができる。しかし，集められるデータ量にはおのずから限界があり，すべての変動条件を尽くすことは不可能である。さらに，データに含まれる変動があまりに大きいと，モデルがぼけて，異なる音素や単語の物理的特徴の重なりが大きくなるため，この方法には限界がある。このため，大多数の話者に対してはうまく動作しても，少数ではあるが認識精度が極めて低い話者を生ずることがある。

これらの種々の変動に対処するため，音声認識システムに，少量の音声サンプルに基づいて，変動に自動的に追随してくれるような，適応機能をもたせる研究が広く行なわれている。一般に適応化アルゴリズムには，その規則を学習するためにあらかじめ決められた言葉を発声してもらう「教師つき学習」と，任意の音声を用いて規則を学習する「教師なし学習」がある。前者の方が規則の獲得は比較的容易であるが，ユーザーにとっては，認識装置を使う前にわざわざ特定の言葉を発声しなければならないので煩わしい。後者の方が，認識すべき音声をそのまま使って適応化できるので，ユーザーにとってはまったく意識せずに使えて便利である。極めて個人差や歪みの大きい音声に対しても，この機能が実現できるアルゴリズムを追及していくべきであろう。さらに加えて，認識装置を使っているうちに，自動的に適応化が進んでいくオンライン逐次型適応が望ましい。

4-2. 話し言葉コーパスの構築と言語モデルの適応化

上でも述べたように，スタジオでのアナウンサーのニュース音声のように比較的書き言葉に近い音声と，通常の話し言葉の間には大きなギャップがある。話し言葉の特性を明らかにし，音声認識のための言語モデルを構築するためには，話し言葉の多様な現象を含む大量なデータベース（コーパス）が不可欠である。米国では，国家プロジェクトの中で，実際の自然な会話音声を大量に録音して書き起こしたデータベースが作られ，これを用いた音声認識研究が行なわれている。我が国でも最近，科学技術振興調整費開放的融合研究推進制度による「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」プロジェクトが開始され，音声認識の高度化のための音声言語データベースの構築, 解析, 認識，要約などの研究が推進されている[(6)]。このプロジェクトをきっかけとして，放送音声などを含む多様な話し言葉のデータベースが，関係する研究者共通の財産として，継続的に集積され管理，活用されていくことが期待されている。

音声認識の対象タスクが変わるごとに，言語モデルの学習のために大量の（書き起こし）テキストを集めるのは決して容易ではない。このため，共通の基本的な言語モデルを，限られた量のテキストを用いて，異なる対象に適応化させる研究も行われている。言語モデルの適応化のためには，テキストや言語モデルの中で，対象に依存する部分と，依存しない一般的な部分を分ける必要があり，このような研究も始まっている。

5. むすび

これまで述べたように，音声認識システムには，極めて多様な応用が期待できるが，

任意の話題について，自然な話し言葉で，だれの声でも，どんな環境でも音声認識できるためには，解決しなければならない課題が多く残っている。主たる課題は，話し言葉特有の言語的・音響的変動，周囲雑音や電話音声の歪み，音声の個人差などに対するロバストなモデルとアルゴリズムの構築である。その基礎となる話し言葉の大規模コーパスの構築も極めて重要である。

今後，音声認識の研究は，単に音声を単語列に変換するだけでなく，発声者の意図の理解，話題語（フレーズ）抽出，内容の自動要約へと展開し，さらに幅広い応用が開けてくると期待される。

コンピューターの小型化，高性能化，低価格化により，21世紀の始めにはユビキタス（遍在）計算とウエアラブル計算環境の時代が来るといわれている[7]。このような環境下では，キーボード入力はなじまないので，音声認識がヒューマンインターフェースの基本的手段の１つとして広く使われるようになると期待される。この際，音声認識の多くは，個人が身につけ，個人に特化したマイクロホンおよびコンピューターで行われるようになり，そのシステム構成は，これまでとは大きく異なった形になると予想される。雑音などの環境情報は常時モニターでき，タスクに依存した知識は個人のコンピューターへ即座にダウンロードできるので，タスクへの適応性とロバスト性の向上が期待できる。ただし，それらの実現のためには，解決しなければならない課題も多く，今後の多角的な研究の展開が期待される。

参考文献

1. B.-H. Juang and S.Furui：Automatic recognition and understanding of spoken language-A first step towards natural human-machine communication，IEEE Proceedings（to be published）
2. 古井貞熙：音声情報処理，森北出版（1998）
3. T.Kawahara，et al：Combining key-phrase detection and subword-based verification for flexible speech understanding，Proc.Int.Conf，Acoust，Speech，Signal Processing，pp.1159-1162（1997.4）
4. B.-H. Juang：From speech recognition to understanding：Shifting paradigm to achieve natural human-machine communication，Proc.16th ICA and 135th Meeting ASA，pp.617-618（1998.6）
5. C.Hori and S.Furui：“Improvements in automatic speech summarization and evaluation methods”，Proc.Int.Conf，Spoken Language Processing，Beijing（2000.10）
6. S. Furui，K. Maekawa，H. Isahara，T. Shinozaki and T.Ohdaira：“Toward the realization of spontaneous speech recognition - Introduction of a Japanese priority program and preliminary results -”，Proc.Int.Conf，Spoken Language Processing，Beijing（2000.10）
7. 古井貞熙：Ubiquitous/Wearable Computing 環境における音声認識の展望，信学技報，SP99-114（1999.12）

ふるいさだお
古井貞熙

１９７０年日本電信電話公社（現NTT）研究所入所。米国ベル研究所客員研究員，NTT基礎研究所第四研究室長，音声情報研究部長，古井特別研究室長を経て，１９９４年東京工業大学大学院情報理工学研究科計算工学専攻客員教授，１９９７年同教授，１９９９年からNHK放送技術研究所客員研究員。音声認識，話者認識，音声知覚，音声合成などの研究に従事。IEEE論文賞，科学技術庁長官賞など受賞。IEEEおよび米国音響学会フェロー。工学博士。