論文 / 著書情報
Article / Book Information

論題(和文)	コンピュータによる音声認識のこれまでと今後の展望
Title(English)	
著者(和文)	古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会2011年春季講演論文集, , No. 1-13-1, pp. 1721-1724
Citation(English)	, , No. 1-13-1, pp. 1721-1724
発行日 / Pub. date	2011, 3

コンピュータによる音声認識のこれまでと今後の展望 *

○古井 貞熙（東工大）

1　コンピュータによる音声認識技術の変遷

音声を自動的に認識する工学的技術が世界で最初に発表されたのは、1952 年のことであった。それ以前には、1925 年頃に、声で名前を呼ぶと犬小屋から犬が飛び出してくるおもちゃが作られているが、その後につながるような技術ではなかった。1950 年代以降、半世紀以上にわたって、音声認識は様々な技術的進歩を遂げてきた。その変遷は、Fig. 1 に示すように、第 1 世代、第 2 世代、第 3 世代、第 3.5 世代に分けることができ、現在、新たな第 4 世代の枠組みの提案が期待されている [1, 2]。

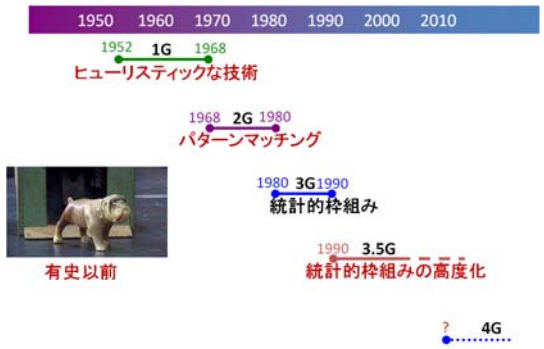

Fig. 1　音声認識技術の変遷

1.1 第 1 世代（1950 年代～1960 年代）

信号処理や計算機技術は極めて未熟であったため、米国、日本および英国を中心に、それまでの音響音声学の知識を元に、いわゆるフォルマント周波数に着目し、アナログフィルタバンクと論理回路を用いた極めて初歩的方法によって、音素、１桁数字、単音節単語などの音声認識が試みられた。

1.2 第 2 世代（1960 年代末～1970 年代）

あらかじめ蓄えておいたテンプレートと入力音声のマッチング（テンプレートマッチング）において、非線形に伸縮する音声の時間軸を合わせる効率的な技術として、動的計画法（DP）を用いる方法（DTW: dynamic time warping、日本では DP マッチングと呼ばれる）が、日本とソ連（現ロシア）の研究者によって同時に発明された。ソ連の研究は外国に発信されなかったが、日本の研究が世界的に知られ、今日まで、この技術が音声認識の基本技術の一つとして用いられている。

このテンプレートマッチング技術を用いて、孤立単語認識の研究が米国と日本を中心に盛んに行われた。ベル研究所では、マルチテンプレートを用いた不特定話者化、IBM では大語彙化に重点が置かれた。

CMU で、音素トラッキングによる連続音声認識の試みが行われ、やがて、DARPA の 1011 単語の語彙を対象とした、音声理解プロジェクトへ発展した。CMU、BBN などで種々の試みが行われ、人工知能的な方法も試みられたが、グラフサーチに基づく CMU の Harpy システムが最も優れた性能を示し、これが現在でも使われている音声認識法の基礎となった。

1.3 第 3 世代（1980 年代）

1970 年代末に、2 段 DP 法を中心とする、音素（あるいは音節）レベルと単語レベルの 2 段階のパターンマッチングに基づく連続音声認識が行われた。これをベースに、1980 年代に入ると、コンピュータの進歩に支えられて、現在の音声認識技術につながる、統計的枠組み（統計的モデル）による音声認識が広く使われるようになり、1980 年代半ばにはそれが主流となった。音響モデルには HMM（隠れマルコフモデル）、言語モデルには、N-gram が用いられるようになった。これらは、主に IBM の研究者によって先駆けて研究され、Bell 研究所の研究者による、HMM の VQ（ベクトル量子化）から GMM（混合ガウス分布）への改良や体系化によって、広く使われるようになった。特徴量としては、ケプストラムとその動的特徴（デルタ特徴）が使われるようになった。これらの枠組みは、四半世紀が過ぎた現在でも変わっていない。

---

* Past, present and future of computer-based automatic speech recognition, by FURUI, Sadaoki (Tokyo Institute of Technology)

1.4 第 3.5 世代（1990 年代～2000 年代）

尤度最大化に基づく統計的モデルをベースとしながら、その高度化として、MCE（minimum classification error）や MMI（maximum mutual information）原理に代表される、誤り率最小化に基づく識別的モデルの研究が行われるようになった。

DARPA プロジェクトでは、より自然な音声を対象とした研究が行われるようになり、ATIS（air travel information service）情報検索システムの研究、放送ニュース（BN, broadcast news）音声の文字化の研究、電話での会話を対象とした Switchboard タスクの研究等が行われた。音声認識システムを、音声の個人差や周囲雑音に対して頑健にするための研究も盛んに行われるようになり、MLLR（maximum likelihood linear regression）法、PMC（parallel model combination）法などが提案された。

2000 年代になると、話し言葉音声を対象とした研究が盛んに行われるようになった。DARPA では音声を文字化するだけでなく、要約を含む多様な情報の抽出を目指した EARS プロジェクトが行われた。最近の GALE などの DARPA プロジェクトでは、会議音声を対象とした Rich Transcription タスクが盛んに研究され、音声の文字化以外に、話題の推移、質問なのか説明なのか、誰が発言しているか、会議を主導しているのは誰か、どのような結論が得られたか、議事の要約の作成など、種々の情報の自動抽出が試みられている。我が国では「話し言葉工学」プロジェクトが行われ、世界最大かつ最も精密な話し言葉音声データベース、CSJ（corpus of spontaneous Japanese）が構築され、多様な音声認識研究に広く用いられている。

2 人の音声認識とコンピュータの音声認識

2.1 限られた情報に依存したコンピュータの音声認識

コンピュータによる音声認識は、一言でいえば、人の音声認識機能をコンピュータ上に実現することである。現在のコンピュータの能力と人の能力を厳密に比較するのは、使われる環境などによって異なるので難しいが、コンピュータによる音声認識誤りは、人の 5～10 倍くらい大きいと言われている。特に、モデルの学習環境と実際の利用環境にミスマッチがあると、コンピュータによる音声認識性能は大きく低下してしまう[3]。

非線形周波数・対数スペクトルを基礎とする特徴抽出、特徴量の多様な変動を考慮した音響的類似性、単語の頻度や接続可能性を考慮した言語的妥当性を利用している、現在の音声認識システムの基本的原理は、人の音声認識における多様な機能の単純化としては、間違っているとは思えないが、明らかに人が音声認識に利用していると思われる情報で、コンピュータが使えていない情報が沢山あるのも事実である[3]。

音声認識システムに頑健性（robustness）が乏しいのは、日本人のような英語の非母国語話者が、英語で会話や会議をする場合と似ている。母国語話者が多様な知識を臨機応変に用いて言語を理解しているのに対して、非母国語話者は使える知識が少ないため、電話を通したり、周囲に雑音があったりすると、とたんに理解能力が低下してしまう。

2.2 スペクトルの動的情報

現在の音声認識技術が取扱いを不得意とする情報に、動的な情報がある。いうまでもなく、音声は常時波形やスペクトルが変化しており、停止することがない。その結果、音素特徴も、前後の音素の影響を大きく受けて変化し（調音結合）、個々の音素のみでは精度よく認識することができない。連続音声において調音結合のために調音が不完全にしか行われない現象に対して、ある種の補正機構が人の聴覚処理系内に存在し、補正された特徴が知覚されていると考え、それを実現する工学モデルもいくつか提案されているが[4]、まだ音声認識システムに使えるモデルはない。現状では、前後の音素を考慮したトライフォン、あるいはさらに前後に広げたクインフォンという形で、変形したモデルを大量の音声データから学習してしまう方法が、認識精度を向上させる近道になっている。

日本語の単音節音声波形を、前と後ろからなめらかに切断して聴取実験を行った結果から、音節の知覚に最も重要な情報は、子音部から母音部へのスペクトルの変化が最も大きい区間に含まれていることがわかっている[5]。人の音声認識では、積極的に動特性を利用して、音素を知覚していると思われる。音声認識システムでは、トライフォン、音節、単語などの粒がひもでつながっているような形で音声をモデル化しており、それらの動的な連続関係を明示的に表現することができない。

2.3 高次言語情報

音声認識システムで使うことが困難な情報としては、他に、アクセント、イントネーションなどの超分節的特徴（韻律的特徴）がある。これも、動的情報であり、それがゆえに、モデル化が難しい。使えていない情報としては、他に、構文的情報、文中の離れた位置にある単語の関連性、文や単語の意味情報などがある。

3　音声動特性の知覚とモデル化

3.1 動特性知覚の階層構造

上でも述べたように、音声現象と聴知覚は、本質的に動的現象である。動特性の知覚単位（時定数の違い）は、階層構造を持っていると考えられており、それぞれの階層によって、感じる音響的特徴が違っている[6]。その階層構造は厳密にはわかっていないが、いろいろな実験結果から、筆者は、基本的に｛数 ms、数十 ms、数百 ms、数 s｝の 4 階層から成っていると考えている[7]。前の 2 階層と、後の 2 階層とは、聴覚による動的現象のとらえ方が、本質的に異なっている。

前の 2 階層（数 ms と数十 ms の単位）は、「知覚的現在（perceptual present）」（50ms～100ms）の中に入るため、知覚的な意味での情報は、動きとして感じるものではなく、時間形式を失って音色として感じている。知覚的現在の中では、時間・強さの相補関係が成り立ち、刺激の時間的加算あるいは積分が行われていると考えられている。これまでの実験から、2ms 程度あれば、その中の 1 個と 2 個のクリックの音色の違いを感じることができ、聴覚の「時間感度」と言われている。また、20ms 程度あれば、その中の音刺激 AB または BA の時間的順序の違い（逐次感）を、（順序はわからないが）音色の違いとして感じることができる。この長さは「時間の粒」あるいは「知覚瞬間（perceptual moment）」と言われ、聴覚情報処理の最小単位と考えられている。コンピュータによる音声認識でも、20ms～50ms の窓でスペクトル分析や、デルタ特徴を抽出して用いており、聴覚的な意味での「瞬間」とほぼ一致している。

後の 2 階層（数百 ms と数 s）は、時間の流れ（動き）として感じることができる長さである。逐次感を超えて、明確に分離した 2 音の感じを得るためには、2 つの知覚的現在にわたっている必要があり、少なくとも 50～100ms 以上の間隔がある必要がある。200ms 程度で、高次神経レベルでの時間的加算が起こっていると考えられており、能動的聴取機構が動作するのに必要な時間と考えられている。また、2s 程度になると、知覚の範囲を超えて、記憶による処理が行われていると考えられる。

音素や音節（分節的特徴）の移り変わりや、アクセント（超分節的特徴）は、基本的に数百 ms の長さの現象であり、イントネーション（超分節的特徴）は数百 ms から数 s に及ぶ。

なお、動特性の階層構造は、｛1～20ms、20～100ms、100ms 以上｝に大別されるという説もある。最初の範囲は、「感覚・末梢過程（sensory process）」（音色感覚）、中央の範囲は、「知覚過程（perceptual process）」（順序感覚）、最後の範囲は、「認知過程（cognitive process）」（言語反応）に対応すると考えられている。

3.2 動特性のモデル化

これらの動特性は、分節的特徴として用いているケプストラムの動的特徴を除くと、現在の音声認識の枠組みでは効果的に用いることができていない。数百 ms およびそれよりも長い動的特徴は、音声認識に効果的に使うモデルがない。数十 ms 単位の動特性も、本来は、分節的特徴として時間の粒の中に押し込んで、粒と粒の単なる接続としてモデル化するのでなく、連続的な動的特徴として取り扱うモデルを構築することが必要と思われる。

4　多様な知識の利用

人が簡単に音声を聴き取っているように見えても、その背景には、その人が生まれてからの多様な経験と膨大な知識があり、人には優れた汎化能力がある。人の情報処理は、並列的、分散制御的、自律的、協調的、学習的、適応的（柔軟性）、自己修復的（再生的）な特性で、特徴づけることができる。

人は音声の音響的特徴、雑音や部屋の残響に関する知識、構文的知識、意味的知識、話題やトピックスの移り変わりに関する知識、さらに人の感情表現などの知識を、並列的かつ適切に組み合わせて、音声を認識・理解している[3]。コンピュータによる音声認識でも、Fig. 2 に示すような枠組みで、大規模な知識の収集・体系化と、問題解決のモデル化を実現することが必要と思われる[8]。そのためには、構造を持った大規模で効率的な統計的知識モデルが必要となる。

大規模知識システムの構築には、多様な条件の組み合わせによる大量な音声コーパスが必要に

なる。不可欠な基本技術の一つとして、現在の音声認識のコミュニティーが持っている話し言葉音声のコーパスに比べて桁違いに大きいコーパスを、膨大な手作業を伴わずに構築し、「教師なし(unsupervised)」あるいは「軽い教師あり(lightly supervised)」学習で使うことができる技術を開発することが重要と思われる。

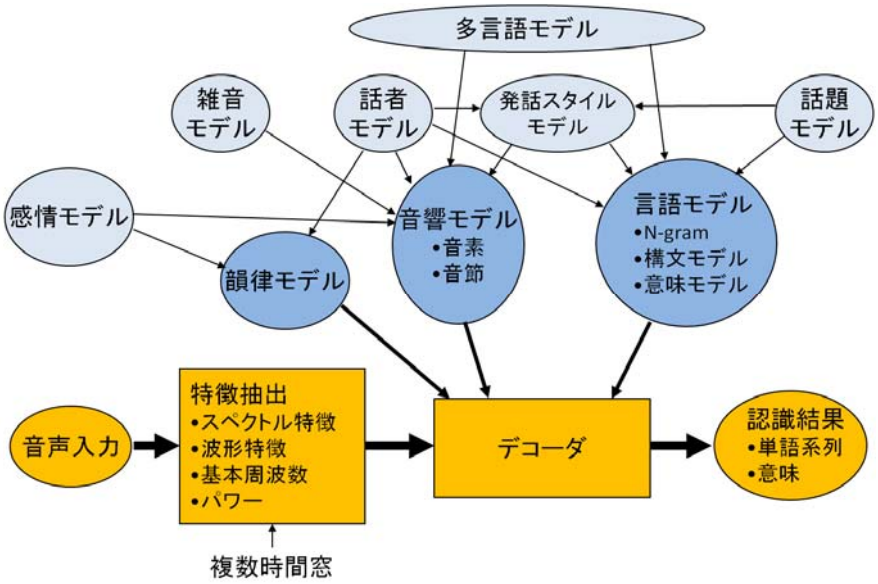

Fig. 2　多様な知識を総合的に用いる音声認識

5　今後の展望

音声認識の研究が始まってから 60 年、現在の統計的枠組みが使われるようになってから 30 年近くが経過し、アプリケーションも広がっているが、人間並みのシステムを実現するという性能目標に比べると、まだ道半ばと言える[1, 7]。音声認識の性能を基本的に上げるためには、人が音声を知覚・認識・理解している過程をもっと明らかにする必要がある。それをその通り実現する必要はもちろんないが、工学的に実現する方法を開発する必要がある。

近年の音声認識の研究は、実用化を目指している企業の開発者と、基礎研究を行っている大学などの研究者との、研究開発対象のギャップが広がりすぎているように思える。現状の限られた音声認識性能で、ヒューマンインタフェースの手段として、あるいは音声の文字化および検索手段として、どのような新たなアプリケーションが可能か、その実用化への障害をできるだけ早く取り除くにはどうしたらよいか、このような観点からの研究開発を、産学連携して積極的に進める必要がある[8]。また、基礎研究者は、研究目標をいたずらに難しい方向へ進めることによって、顕著な進歩が見えない言い訳をするのではなく、人の音声知覚・認識・理解のメカニズムを一歩ずつ明らかにしながら、基本的な課題を解いていく必要がある。

このような両輪の研究開発の過程から、新たな発想によって、第 4 世代と言える大きなブレークスルーが生まれてくることを期待したい。

参考文献

[1] 古井貞熙、何かが欠けている音声認識研究、信学技報、SP2009-80 (2009)

[2] S. Furui, History and development of speech recognition, in F. Chen, K. Jokinen (eds.): Speech Technology, Springer Science and Business Media, LLC (2010)

[3] O. Scharenborg, Reaching over the gap: A view of efforts to link human and automatic speech recognition research, Speech Communication, 49, pp. 336-347 (2007)

[4] 赤木正人、古井貞熙、音声知覚における母音ターゲット予測機構のモデル化、電子通信学会論文誌、J69-A、10、pp. 1277-1285 (1986)

[5] S. Furui, On the role of spectral transition for speech perception, J. Acoust. Soc. Am., 80, 4, pp. 1016-1025 (1986)

[6] 寺西立年、聴覚の時間的側面、難波精一郎編：聴覚ハンドブック、pp. 276-319 (1984)

[7] 古井貞熙、人と対話するコンピュータを創っています～音声認識の最前線～、角川学芸出版 (2009)

[8] 古井貞熙、音声認識技術の実用化への取り組み、情報処理、51、11、pp. 1387-1393 (2010)