音声認識技術とその実用化の動向

坪井 宏之 (株)日本電子化辞書研究所
正員 竹林 洋一 (株)東芝 総合研究所

特 集

解 説

1. はじめに

音声認識の研究は、話し言葉をそのまま入力できる音声タイプライタや、機械との対話を可能にする音声理解システムを目標として、古くから行われている。1970年代のアメリカの ARPA 音声理解システムプロジェクトでは、Harpy 以外のシステムは未完に終わったが、多くの優れた研究成果を残した。その後、各国で実用的な研究が行われるようになり、音声認識装置が実用化されるに至った。しかし、音声認識装置を実際に応用してみると、マンマシンインタフェースとして要求される認識性能を十分に満足していないことが、数年前から明らかになってきた。

1985年代にアメリカでは DARPA の大規模なプロジェクトが再び開始され、MIT, CMU, BBN を中心として、不特定話者大語集(い)連続音声認識を目指した基礎研究が1,000万ドル/年の規模で行われている。また、HMM (Hidden Markov Model) によるリアルタイム大語集単語認識装置の開発が IBM, AT & T で行われている。更に、イギリス、フランス、西ドイツや日本でも基礎的研究が盛んになってきた。

しかし、上述したような音声認識の最新の研究成果と実際に商用システムを使用する際の認識性能との間には相当大きなギャップが存在する。一般のユーザは、マスコミで報じられる認識装置の性能から、現在の音声認識を過大評価したり、逆に、音声認識装置の利用経験のあるユーザや技術者は、音声認識の実用化に関して過度に否定的な意見をもっている場合が多い。以下、実用化の観点から音声認識技術の現状分析と今後の動向、特に、雑音の問題につき解説する。

Trends in Speech Recognition Technology and Its Applications.
By Hiroyuki Tsuboi (Japan Electronic Dictionary Research
Institute, LTD) & Yoichi Takebayashi (Toshiba Corporation,
Research and Development Center).

本内容は筆者が(株)東芝在職中に行った研究成果の一部である。
(坪井)

2. 音声認識技術と実用化の現状

2.1 音声認識技術の動向⁽¹⁾⁻⁽⁴⁾

音声認識は、表1に示すような話者登録方法、発声方法、語彙数の観点から分類できる。このうち、現在実用化されているのは、主として特定話者の離散/連続単語認識装置、小語集の不特定話者単語認識装置である。話者登録方法に関しては、不特定話者方式は音声の登録が不要であるため、装置が簡素化できるという利点があり、LSI 化に適している⁽⁶⁾。また、話者適応形は、不特定話者の標準パターンを基本とするため、特定話者形よりも少数のパターンで、各ユーザの認識用辞書パターンを登録できるので、今後、多く採用されると考えられる。発声方法に関しては、文、文節、連続単語とユーザの発声に対する制約が順に増すが、認識率は逆に向上する。また、認識対象語彙数と認識率も相反関係にある。このため、発声の容易さや語彙数と認識性能とのトレードオフが、音声認識の実

表 1 音声認識システムの分類

項 目	内 容	
ユーザの 音声登録方式	特定話者	システム準備 なし ユーザ登録 すべてを登録
	不特定話者	システム準備 多数話者より作成 ユーザ登録 不 要
	話者適応	システム準備 多数話者より作成 ユーザ登録 一部を登録
入力音声の 発声単位	音 節	「お」「ん」「せ」「い」「に」「ん」「し」「き」
	単 語	「おんせい」「にんしき」
	連続単語	「おんせいにんしき」
	文 節・文	「おんせいにんしきをする」
認識対象語彙数	少	~100 語程度
	中	~1,000 語程度
	大	1,000 語程度以上

用システム設計上の大きな問題となる。

特定話者単語認識装置は、単語全体のパターンに基づく DP マッチング法を利用している場合が多い。語彙数は高々数百語であり、時間軸方向の単語パターンの変形を、非線形な時間軸伸縮を用いて吸収している。しかし、この方法は、語彙の変更の際には新たに認識対象単語を、標準パターンとして登録しておく必要があり、また、ユーザは認識対象単語すべてを発声しなければならないという問題がある。一方、不特定話者認識は、識別関数、マルチテンプレート、複合類似度法などが、実用レベルの認識装置⁽⁵⁾や音声認識 LSI⁽⁶⁾に採用されている。特定話者認識と同様に、ここでも、大語彙化や語彙の変更および認識性能が課題になっている。

実用化につながる、最近の技術の傾向として特筆すべき点は以下の3種類のアプローチである。

- HMM (Hidden Markov Model)
- ニューラルネット
- エキスパートシステム

HMM は、1970 年代後半から Jelinek らにより文章入力を目的として研究されてきた確率モデルに基づく音声認識の手法である。従来の DP マッチング法は数百語程度の特定話者単語認識に広く適用されたが、HMM は、不特定話者化、大語彙化、連続化に適用可能であり、各研究機関で盛んに研究されており、かなり実用化に近いレベルに達している。

一方、ニューラルネットワークは、研究室レベルでシミュレーションを行っている段階であるが、最近のブームを反映して、音素認識、雑音除去、パターン照合などに関する研究が急速に立ち上がってきた。音声認識は、研究者の層が厚く、また格好の応用分野であるため、今後の成果が期待できる。

また、エキスパートシステムは、音声の知識を整理し、それを自動認識に利用するという観点からのアプローチである。音声の知識を蓄積するという意味で意義が大きいが、知識の自動獲得、すなわち学習機構の組み込みが最大の課題である。

以上、最近の音声研究の傾向について簡単に説明した。上記の3種類の方法すべてが強力なコンピュータパワーを必要としているが、今後の半導体や計算機技術の進歩を考慮すると、有効なアルゴリズムが提案されれば、実用システムへのインプリメントは比較的容易であると思われる。

2.2 実用化の動向

前述したような種々の研究成果に比べて、音声認識装置の実用化はあまり進んでいない。音声は人間の最

も自然な通信手段であるが、反面、計算機のインタフェースとしての入力手段としては、あいまい性(誤認識)を含むという決定的な欠点を有している。一方、キーボードは、ユーザにとって従来負担の大きい入力手段であったが、ワープロの急速な普及により、一般のユーザの大部分が使いこなすようになってきた。両者を比較すると、音声認識装置の認識性能と認識対象とする語彙数がユーザの要求するレベルに至っておらず、実用化を妨げている。

現時点で、音声認識装置が利用されている分野は極めて限られている。音声の利点をいかした下記の応用分野では、それなりの成果をあげていることがわかる。

- (1) クレーン制御システム
- (2) 電話による銀行残高照会システム
- (3) 自動車電話用自動ダイヤルシステム
- (4) 自動車用話者照会システム

(1)はクレーン作業の省力化を目的とする応用で、特定話者の音声認識によりクレーンの制御を行う作業者の人数を削減するためのものである。両手のふさがっている作業に対する同様の応用として、郵便物、製品検査の仕分けシステムも実用化されている。しかし、不用意に発声した音声やノイズへの対応が今後の問題となっている。

(2)の応用は、日本の都市銀行で7～8年前から導入されている、電話による不特定話者音声認識を用いた振り込み通知や残高照会を自動化するためのシステムである。現在でも多くの銀行などで使用されており、音声認識装置の応用の中で利用者数が最も多い分野である。この応用では、電話回数に重畳するノイズと長距離伝送によるレベル変動が誤認識の原因になっている(3.2 節の(3)参照)。

(3)、(4)の応用は、エレクトロニクス化が進む自動車を対象とする応用例である。(3)の自動車電話の普及は特にアメリカで目覚ましく、音声ダイヤルは、両手をハンドルに、視線を進行方向にとられている運転者にとって非常に有用なダイヤルの入力手段である。アメリカを中心に実用化が開始されているが、車内騒音に対する対策と低価格化が今後の普及の鍵となっている。

(4)は、セキュリティ関連の音声認識の応用であり、音声の有する身体的な個人情報を利用したシステムである。自動車のキーの代用をして、ドアロックとイグニッションのオン・オフを行うためのものである。この話者照会システムには話者のデータとして音声パターンの登録が必要なので、特定話者の音声認識

機能も付加されている。自動車内のライトの点灯や、ワイパー、ラジオの操作に、この特定話者音声認識は応用可能であるが、まだ本格的に利用されるには至っていない。

上述した応用例と対照的な例が、音声入力ワードプロセッサである。これは、数万語に及ぶ語彙を入力対象とするため、認識性能の向上が困難である上に、ワープロのユーザ層の拡大により“キーボードアレルギー”のユーザの割合は減少を続けている。更に、オフィス内などでの騒音に対処するため現状では接話マイクの使用が不可欠であるという問題がある。従って、音声認識装置の飛躍的な認識率向上、および価格、速度面での十分なメリットがなければ、本格的な実用化はかなり先のことになるざるを得ない。アメリカでは、離散単語発声のワープロ（ディクテーションマシン）の実用化を目指した研究が行われている。英語の場合、単語と単語を区切って発声するのに比較的抵抗感が少なく、また、日本語と比べて接頭語、接尾語や付属語のような変化が少ないので、本格的な実用化は日本語よりも先になるかもしれない。

前述した音声認識の応用例でも示したとおり、実用化に際して大きな問題となっているのが雑音であると考えられる。しかし、これまでのところ音声認識における雑音の影響に対して十分な検討がなされていない。以下、雑音の問題について説明する。

3. 雑音環境下での音声認識

音声認識装置が使用される環境では、通常、入力音

声以外の雑音が周囲に存在し、雑音に起因する誤認識が生じる。ここでは、その原因をまとめ、それぞれについての最近の研究を中心に説明する。

3.1 音声認識に影響する雑音の分類

図1は、音声認識に影響する種々の雑音を示す模式図である。音声認識装置にとって、騒音とは入力音声以外の不要な雑音のことであり、他人の音声も、認識装置にとっては雑音となる。ユーザの相づちや独り言なども認識装置に不要な入力であり、雑音と等価なものである。その他、電話送話器の非線形ひずみ、伝送系のひずみ、反射音も雑音と見なすことができる。現状では、図1の①に示した外来雑音に関する検討がわずかに成されているだけで、本格的な雑音の影響に関する研究は行われていない。

区分	音源の位置	例
①	ユーザの周囲	騒音 エンジン音、ドアの開閉音
		音声 背景の会話
②	ユーザ自身	騒音 呼吸音、舌打ち音
		音声 「えーと」など入力対象外の発話
③	伝達系	ひずみ 電話の送話器、長距離伝送ひずみ
		反射 部屋の壁、机による反射音

図1 音声認識における雑音の分類

項目	内容	主要な課題
音声収集	音声の収音方式	マイクの種類・個数・位置 (7)
信号処理	雑音の除去・音声の強調	雑音スペクトルの推定、音声の周期性の利用、波形の相関処理、ニューラルネットの利用 (8), (9), (10), (11), (12) 耐雑音性能の高い特徴、聴覚モデルの利用 (13), (14)
音声区間検出	音声の始終端検出	複数の始終端候補、トップダウン的情報の利用 (15), (16)
照合モデル	入力パターンと標準パターンの照合方式	距離尺度の耐雑音性能、雑音を含むスペクトルと含まないスペクトルの対応の利用、雑音モデルの組み込み (17), (18), (19)
認識システム	複数の特徴による判定の統合	全体的特徴と部分的特徴の並用 (5)
その他	カクテルパーティ効果	複数話者からの音声の分離 個人性の抽出とその分離 (20), (21)
	Lombard 効果	高雑音下の発声の変化 スペクトル形状の変形補正 (22)

図2 音声認識における雑音対策

3.2 音声認識における雑音対策

騒音や明りょう度改善のための雑音除去に関する研究は従来から広く行われている。しかし、音声認識の立場からの雑音対策に関する研究はあまり行われていない。音声認識における雑音対策を認識処理の中で項目に分けて図2に示す。これらの項目のほとんどは外来騒音に対する雑音対策である。

(1) 音声収集 音声認識のためには、ユーザの発声を S/N の高い信号として収集することが望ましい。接話形マイクロホンは S/N 改善には有効であるが、その着脱は煩わしく、実用化の障害となっている。マイクロホンを複数個並べてユーザの方向を得るような指向性をもたせる⁽⁷⁾方式などが検討されている。このマイクロホンの段階で S/N を改善できれば、続く認識処理の性能は大幅に向上する。

(2) 信号処理 従来から、音声の S/N 改善のための信号処理技術が研究されている⁽⁸⁾。1入力の信号処理の場合、雑音の音響的特徴入力から推定して入力信号から雑音成分を除去する(スペクトル減算処理)方式、音声の周波数スペクトルの調波構造を利用して周波数領域で音声成分を抽出する方式(くし形フィルタ処理)、人間の聴覚特性(例えば、音声信号の位相の変化に鈍感である)を利用して波形の相関処理を行う方式などがある。

単語音声認識システムに使用した、くし形フィルタ処理による雑音除去の流れを図3に示す⁽⁹⁾。従来の方式に加え調波構造成分の雑音バイアスの推定除去を行い、ピッチの抽出は帯域フィルタ群を用いて、相補的な処理を併用している。話者2名で67単語の実験が行われ、 S/N が12 dBで、白色雑音付加の場合で55%、室内騒音付加の場合で35%認識率が向上して

いる。

ニューラルネットを使った波形処理による雑音除去も検討されている⁽¹⁰⁾。ネットワークは3層からなり、各層は60ユニットをもち、層の間のすべてのユニットが結合されている。音声波形に雑音を付加する前と付加後の60サンプルを1フレームとしてバックプロパゲーションで学習を行った。スペクトル減算処理とニューラルネット処理の合成音のプレファランステストを行い、スペクトル減算処理に比べて良い結果を得ている。

入力が複数の場合には適応フィルタの考え方に基づく処理方式が提案されている⁽¹¹⁾。雑音源から主マイクロホンまでの伝達特性をトランスバースフィルタで近似させ、主マイクロホンでの雑音を補助マイクロホンの入力から推定することにより、雑音を打ち消そうとするものである。この方式では、トランスバースフィルタの係数の適応アルゴリズムは収束速度が実用上問題となるが、実時間で動作する装置が試作されている⁽¹²⁾。

音声の周波数と時間に関する二次元ケプストラム分析による耐雑音性の向上方式が提案されている⁽¹³⁾。更には、ピッチ周波数に同期して、フィルタバンク出力の各チャンネルの自己相関係数を分離スペクトルとして抽出することにより S/N が高い音声スペクトル情報を得るモデルが提案されている⁽¹⁴⁾。

(3) 音声区間検出 音声認識は一般に入力音声の始末端を検出した後、認識処理を行うが、高雑音では始末端の検出が困難となる。そこで始末端を一意に決定せずに認識と音声区間決定を同時に行う方式が有効である⁽¹⁵⁾。また、非定常な騒音と音声の区別を行うために入力対象となる語の音響的、時間的特徴や入力対象語の特徴などをトップダウン的に利用して検出する方式⁽¹⁶⁾も提案されている。この区間検出は、音声認識などに決定的な影響を与えるので、今後も更に研究を進める必要がある。

(4) 照合モデル 照合モデルは、音声のパターン認識の中核をなす部分である。ベクトル量子化における雑音を考慮したスペクトルマッピング⁽¹⁷⁾、HMMにおける雑音モデルの組み込み⁽¹⁸⁾などが提案されている。

また、LPC分析とフィルタバンク分析に基づく特定話者単語認識における各種距離尺度の評価に関する研究があり⁽¹⁹⁾、LPC分析では、白色雑音に対する耐雑音性能は周波数の低域に重みをつけた尺度(WLR)が高く、また、フィルタバンク分析に基づく特徴パラメータは疑似的にピークに重みを置いた尺度が耐雑音

図3 くし形フィルタ処理による雑音除去⁽⁹⁾

図4 ハイブリッド認識方式の構成図⁽⁵⁾ 図5 ニューラルネットによるカクテルパーティ効果のモデル⁽²¹⁾

性能が高かったと報告されている。

(5) 認識システム 前述したような個別の雑音対策がそれぞれ適用できるが、複数の認識方式を統合して耐雑音性能の高い音声認識システムを構築すると効果的である⁽⁵⁾。図4に、電話音声認識の実用システムに適用されたハイブリッド構造マッチング法による単語全体のパターン照合と単語の内部構造を解析する遷移ネットワークを併用することにより、電話回線のノイズやレベル変動、また、送話器の非線形ひずみに対処し、不用意な発声のリジェクトを可能とした。

(6) その他 雑踏中の人声から相手の声を聴き分けられる現象をカクテルパーティ効果という。これは両耳の方向性ではなく、音声の個人性による区別と音声内の言語的情報によるものである。個人性の一つとしてピッチ周波数の時間的連続性を利用して、一人のピッチ周波数を抽出し、くし形フィルタにより一人分の音声を抽出することが行われている⁽²⁰⁾。また、ニューラルネットを用いて異なる特徴をもつ入力信号を区別するモデルが提案されている⁽²¹⁾(図5)。興奮性セルと抑制性セルという2種類のセルが図5(a)に示すような結合で結ばれており、異なる二種の刺激に

対して興奮性セルのブロック化と抑制性セルの働きにより一方の入力のみに興奮するというものである。その活動度の例を図5(b)に示す。

高雑音下では発声された音声の発話そのものが変形し、エネルギー、ピッチ周波数、スペクトル形状、継続時間長などの変形が複合された雑音と等価なものとなり、認識に悪影響を及ぼす。これを、Lombard 効果と呼び、その対処としてスペクトル形状の変化の補正が検討されている⁽²²⁾。

4. おわりに

音声認識技術と実用化の現状を概観し、雑音環境下での問題点について述べた。現在、音声認識の研究は音素認識、言語処理などの不特定話者大語彙連続認識を目指した研究が盛んに行われている。しかし、実用化の際に大切な雑音環境を考慮した音声認識技術の地道な研究はあまり行われていない。このため、現状の音声認識システムは極めて脆弱であり信頼性が低く、接話マイクや防音室の使用が必要である場合が多い。

本文での雑音についての議論は、要素技術的なものに集中したが、音声認識の本質を考えた場合、音声の有する高次の言語的内容、意思・感情などの情報伝達が重要である。すなわち、雑音中の音声認識のためには、低次信号レベルの雑音処理技術に止まらず、音声の各種知識源を活用した認識方式確立のための本格的基礎研究が必要である。今後、連続化、大語彙化、不特定話者化のほかに、雑音に関する研究に注力すれば、音声認識装置の広範な実用化の道が開けるものと思われる。今後の研究により“聴こうと意図したものだけを聴きとる”、高次のノイズフィルタリング機能を備えた音声認識装置を開発すること、自然な音声入力を可能とし、広範に音声認識技術が実用化されることを望む。

(昭和63年9月27日受付)

文 献

- (1) “特集 Man-Machine Speech Communication” *Proc. IEEE* 73, 11 (1985)
- (2) 「音声処理特集」信学誌 70, 4 (昭 62)
- (3) 坪井・浮田:「音声認識・理解研究の動向」情報処理 29, 1, 30 (昭 63)

- (4) 「特集：音声の知的処理」人工知能学会誌 3, 4 (昭 63)
- (5) Y. Takebayashi, H. Shinoda, H. Asada, T. Nitta, S. Hirai, & S. Watanabe: "Telephone Speech Recognition Using a Hybrid Method" *Proc. 7th Int. Conf. Pattern Reco.g* p. 1292 (1984)
- (6) J. Tsunoda, H. Chimoto, N. Sugi, & K. Sato: "Speaker-Independent Isolated Word Recognition LSI Set" *Toshiba Reveiw* No. 159, p. 34 (1987)
- (7) H. F. Silverman: "Some Analysis of Microphone Arrays for Speech Data Acquisition" *IEEE Trans. Acoust., Speech, Signal Process.* ASSP-35, 12, 1699 (1987)
- (8) 尾関: 「音声信号の雑音除去技術の現代」NHK 技研月報 4 p. 410 (昭 36)
- (9) 長瀬: 「雑音環境下の単語音声認識の性能改善」信学論 J71-A, 5, 1100 (昭 63)
- (10) S. Tamura & A. Waibel: "Noise Reduction Using Connectionist Models" *Proc. Int. Conf. Acoust., Speech, Signal Process.* S12, 7, p. 553 (1988)
- (11) B. Widrow, et al.: "Adaptive Noise Cancelling; Principles and Applications" *Proc. IEEE* 63, 12, 1692 (1975)
- (12) 梅田・尾関・安広: 「高速マイクロプロセッサを用いた雑音除去」音声研資 S 81-81 (昭 56)
- (13) 有木・水田・坂井: 「不特定話者の発声した低 SN 比単語音声の高速認識法」信学論誌 J 71-D, 5, 790 (昭 63)
- (14) 亀井・河原田: 「高雑音下の単語音声認識」同上 J 67-D, 12, 1403 (昭 63)
- (15) 新田, 他: 「単語境界仮説と複合 LPC 類似尺度を用いた電話音声認識」信学技報 SP 86-21 (昭 61)
- (16) J. G. Wilpon, L. R. Rabiner, & T. Martin: "An Improved Word-Detection Algorithm for Telephone-Quality Speech Incorporating Both Syntactic and Semantic Constraints" *AT&T Bell Technical J.* 63, 3, 479 (1984)
- (17) B. H. Juang & L. R. Rabiner: "Signal Restoration by Spectral Mapping" *Proc. Int. Conf. Acoust., Speech, Signal Process.* 6.6, p. 2368 (1987)
- (18) A. Nadas, D. Nahamoo, & M. A. Picheny: "Speech Recognition Using Noise-Adaptive Prototypes" *ibid.* S11. 8. p. 517 (1988)
- (19) H. Matsumoto and H. Imai: "Comparative Study of Various Spectrum Matching Measures on Noise Robustness" *ibid.* 14. 19, p. 769 (1986)
- (20) T. W. Parsons: "Separation of Speech from interfering Speech by Means of Harmonic Selection" *J. Acoust. Soc. Am.* 60, 4, 911 (1976)
- (21) Ch. von der Malsburg and W. Schneider: "A Neural Cocktail-Party Processor" *Biol. Cybern.* 54, 29 (1986)
- (22) 服部, 他: 「雑音環境下で発声した音声の認識方式」信学技報 SP 88-11 (昭 63)