最近の音声認識技術の進展と問題点

齋 藤 収 三*

1. ま え が き

話し言葉を聞き分ける人の機能を代行することのできるロボットを開発することは、音響科学技術者の古くからの夢であった。わが国では 1930 年代に、田中館愛橋先生が日本語音声で働く音声タイプライタの開発研究の必要性を説いておられたという。音声認識に関する学術論文として数字音声の機械認識の研究が発表されたのは 1952 年であるが、1960 年代以降デジタルコンピュータによる音声処理技術が進展して単語音声認識装置の商品化が行なわれるようになり、今日では音声認識技術の研究は連続単語や文音声を対象とする連続音声認識装置、発話者が相手に伝えたいと考える情報内容の理解を目指す対話形音声理解システム、あるいは発話者個人名の識別や照合を目標とする話者認識システムなど、幅広い応用を目標として研究が盛んである。ここでは話し言葉の意味情報の認識と発話者の個人性情報の認識との、2 種の認識技術について最近の研究に重点をおいて解説することとしたい。

2. 音声の意味情報の認識

2.1 音声認識装置の構成

音声認識装置の基本的な構成は図 1 に示すように、音響分析によって音声信号より音響的特徴量を抽出し、これを装置に内蔵された標準パターンと比較して、パターンの整合度を計量し、認識判定を行なうようになっている。ここで標準パターンは、言語情報が既知の音声信号について音響分析により予め抽出した音響的特徴量により記述されている。標準パターンに用いる音声信号には、単語、音節、あるいは音素などの音声単位が使われている。単語の標準パターンを使う場合は、単語中の音素と音素のわたり部分における調音結合によって生ずる音響的特徴量の変形が標準パターンにそのまま含まれるので、認識音声信号とのパターン整合が良くして認識性能

が一般によいが、単語の全時間長に対応した長さの標準パターンを内蔵することになるので、そのための記憶容量が大きくなる。

音素の標準パターンを使う場合は、標準パターンを音響的特徴量が定常な部分のみを用いて記述するので、それに要する記憶容量は少なくすむが、調音結合などによる音響的特徴量の変形が標準パターンに含まれていないと認識性能は劣化し易い。この対策として後述のように音素の定常部分以外に、わたり部分における差分を音素標準パターンに付け加える構成がよく用いられている。

2.2 パターン整合に用いる音響的特徴

認識音声入力と標準パターンとのパターン整合に用いる音響的特徴量として、音声信号の周波数スペクトル包絡と音帯振動周波数が使われる。音声信号の周波数スペクトル包絡を抽出するには、帯域フィルター、ケプストラム (CEPSTRUM) 分析、あるいは線形予測分析などが使われる。そして分析された周波数スペクトル包絡特性を用いて音声入力を逆フィルタリングしたり、ケプストラム分析結果をリフターで分離することにより声帯振動周波数を抽出する。

これらの音響的特徴の分析法の内容は音声情報処理に関する技術書¹⁾に記されているのでここでは省略する。

2.3 音声認識アルゴリズムの進展

音声波に含まれる意味情報と音声波の音響的特徴とは必ずしも 1 対 1 に対応しない。意味情報はカテゴリー的な性格を持つので、様々な音響的特徴の変形、例えば音素区間の時間的な伸縮や調音結合によるホルマント周波数の移動など、に対応して認識が行なわれる。このような音響的特徴の変形に対応できる認識アルゴリズムが考案されている。

図 1 音声認識装置の基本構成

原稿受付 1990 年 12 月 29 日

Recent Advance on Speech Recognition

* 工学院大学

2.3.1. 非線形時間軸マッチング

図1に示した構成による単語音声認識では、標準パターンにおける個々の音素の時間長と、認識音声入力における対応する音素の時間長との間に、かなりの差が生ずるのが普通である。同じ単語でもゆっくり発話すると1秒間に4モーラ程度の速さであるが、急げば1秒間に10モーラ以上を発話することができる。このように発話の速さが異なる音声入力では、音素の継続時間長を伸縮して能率良くパターン整合をとることが必要である。このための認識アルゴリズムを動的計画法による非線形時間軸マッチングと呼ぶ²⁾。図2はその例である。ここで横軸は音声入力、縦軸は標準パターンで、左下の単語の始端から右上の終端まで、音声入力と標準パターンとが時間長まで一致していれば対角線に沿って類似度最大のパターン整合がとれる。一般には標準パターンの時間軸を伸縮して、音声入力と標準パターンとの最適な整合がとれるようにする。ここでは時間軸を伸縮して整合をとる整合バスの方向を図に示す3方向にとっている。また整合をとる窓幅の制限を付けて、誤認識を減らすようにする。この認識アルゴリズムは単語音声認識における基本的手法として広く用いられている。

2.3.2. 標準パターンの音声単位

音声認識に用いる標準パターンを認識単語毎に用意する、いわゆる単語単位の標準パターンを用いると、単語の構成単位である音素の言語環境による変形が内包されるので、誤認識を減らすことができる。しかしながら単語内の音響的特徴の時間的変動をすべて記憶する必要があるので、標準パターンの記憶容量が大きくなり、認識語彙数が増すと膨大な記憶容量を必要とする。

単語を構成する音素や音節を標準パターンに用いることにすると、それらの音声単位毎の標準パターンの他に、音声単位記号の系列で表示された単語辞書を記憶する必要があるが、全体の記憶容量は単語単位の標準パターンに比して遙かに少なく済む。しかし音素や音節は言語環境の影響による音響的特徴の変形が生ずるので、この変形を無視したり、時間平均化したりすると誤認識が増大するようになり、認識語彙数が増すとこの傾向は強くなる。このためこれまでの音声認識装置の実用的認識語彙数には数100単語までという制約がつけられている。

この制約を取り除いて大語彙の音声認識装置を実現するため、言語環境における音素スペクトルの動的特性としてLPCケプストラムの時間変化量を特徴量とする認識アルゴリズム³⁾や、認識単語内の音素のLPCケプストラム分析により導いた音素特徴ベクトルのクラスタ化アルゴリズム⁴⁾、あるいは半音節や音素結合⁵⁾を音声単位とするアルゴリズムなど種々の手法の研究が行なわれて

図2 DPによる時間軸マッチング

いる。

2.3.3. 標準パターンの学習

標準パターンの音響的特徴には発話者の個人的特徴が含まれているので、特定の発話者の音声標準パターンに用いる音声認識装置に他の発話者の音声を入力すると、多くの場合音声認識率が低下する。そのため音声認識装置の標準パターンは発話者に合わせて作り直す、いわゆる学習処理をすることが多い。標準パターンの学習をする形式を特定話者用と呼び、学習しない形式を不特定話者用と呼ぶ。

特定話者用認識装置では、標準パターンを適切に設定すれば認識語彙数が数百語以下では97~99%程度の認識率が達成できるが、標準パターンの学習に要する手間と時間が可成り大きくなる。特に標準パターンの音声単位が単語の場合には認識すべき単語をすべて少なくとも一回は発話しなければならない。標準パターンが音素や音節単位であれば標準パターンの学習を簡略にできる可能性があり、種々の研究が行なわれている。

不特定話者用認識装置のための標準パターンの作成法は現在研究中の課題である。認識語彙数が30単語以下では実用性のある認識装置ができるが、語彙数を100単語以上にするための研究が現在盛んに行なわれている。このためには音声データベースの充実が不可欠である。

2.3.4. 隠れマルコフモデル (HMM)

前述の言語環境、あるいは発話者による音素の音響的特徴の変形に対する認識手法としてHMM⁶⁾がある。これは始端と終端との間にいくつかの状態があって、状態間は一定周期毎に遷移する。この状態遷移確率は与えられている。また各々の状態が幾つかの記号を出力する確率が与えられている。この状態遷移確率と記号出力確率の2つが与えられているが、個々の状態は隠されていて、外部からは見えない。この点が通常マルコフモデルと

図3 HMMによる音声のモデル化

異なるので、隠れマルコフモデルと呼ばれる。図3はその例である。

ここでは始端と終端との間に3つの状態 S_1 , S_2 , S_3 がある。状態遷移確率は a_{ij} で与えられ、状態 s_i において記号出力が v_k である確率が $b_i(k)$ で与えられている。このような HMM を用いることにより特定の単語音声に対応する記号出力系列の生起確率を計算することができる。認識対象のすべての単語について学習用音声信号を用いて a_{ij} と $b_i(k)$ を決めると、認識単語の HMM が与えられる。このモデルは記号出力が確率分布を持つので、ある特定の単語の HMM から別の単語に相当する記号出力が得られることも少ない確率で起こり得るという点が特徴で、認識誤りを減らすのに有効である。

HMM により音声認識を実行するためには、図4に示すように音声信号を分析して音響的特徴量を抽出した後、特徴ベクトルに関するコードブックを参照してベクトル量子化した後、記号出力系列との対応を調べる。この記号出力系列に基づいて予め学習処理を済ませた HMM における記号出力の尤度を計算し、最大尤度を得る単語を認識出力とする。

図4 HMMによる単語音声認識システムの構成

HMM による音声認識アルゴリズムは 2.3.1. で述べた非線形時間軸 マッチングと極めて類似した手順で、HMM の各状態からの記号出力が確率的であることを除けば、本質的には類似の手順と見なせる⁷⁾。HMM はその意味から非線形時間軸 マッチングアルゴリズムを包含した、より一般的な認識アルゴリズムといえる。

さらに HMM はここで説明した音声の音響的特徴ベクトルに基づく記号出力系列の認識処理に有用であるばかりでなく、単語系列などより高次の記号出力の認識処理にも有用であって、文音声の連続的な認識に適用されている⁸⁾。最近では連続音声認識に有用な基本的アルゴリズムとして幅広く研究が進められている。HMM の内容についてここではこれ以上立ち入らないが、確率モデルによる音声認識について優れた図書がある⁹⁾ので、より深い内容についてはそちらを参照されたい。

2.3.5. ニューラルネットワークによる音声認識

人の情報処理能力に類似した情報処理システムの実現を目指して、人の脳における神経細胞の生理学的特性に類似した素子によるネットワークモデル、いわゆるニューラルネットワークの研究が盛んで、これを音声認識に応用するいくつかの試みが行なわれている。

ニューラルネットワークモデルには種々の構造のものがあるが、音声認識への応用に用いられているのは図5に示すような入力層と出力層との間に複数個の隠れ層(中間層)を持つ階層構造のモデルである。ここで入力層には音声进行分析するのに用いた音響的特徴量の次元数に等しい数のユニットがあり、出力層には認識カテゴリの種類数に等しい数のユニットがある。中間層のユニット数は通常入力層のユニット数より少なくする。これらのユニットは階層間で網目状に接続されていて、入力層ユニットの出力に、結合係数と呼ばれる重みをかけた荷重和が中間層ユニットの入力となる。この入力と神経細胞の域値との関数の形で、中間層ユニットよりシグモイド関数形応答が出力される。これを数式的に表現すると、

$$H_i = f(\sum W_{ji} \cdot I_j - h) \quad (1)$$

ここで I_i : 入力層 i の出力

W_{ji} : 入力層 j から中間層 i への結合係数

図5 3層構造のニューラルネットワーク

h : 中間層 j の域値

H_j : 中間層 j の出力

f : シグモイド関数

中間層ユニットと出力層ユニットの関係も同様で、中間層出力の荷重和 ($\sum V_{jk} \cdot H_j - h$) のシグモイド関数形応答として出力層出力 O_k が得られる。このモデルにおいて学習パターンの教師信号と出力層出力との差を最小にするように、 V_{jk} および W_{ij} を修正する手順がバックプロパゲーションと呼ばれる学習アルゴリズムである。

この階層形ニューラルネットにより有声音 /b/, /d/, /g/ の認識実験が行なわれている¹⁰⁾。ここでは音声入力を16チャンネルの帯域に分割し、フレーム周期10ms、フレーム数15の周波数-時間パターンで表わす。従って入力層は16×15ユニットで与えられる。中間層は2層あり、第1層は8×13ユニット、第2層は3×9ユニットで構成され、出力層は/b/, /d/, /g/の3ユニットである。この認識システムを発話者毎に学習させた結果、認識率が98.5%に達し、この値はHMMによる93.7%に比べて格段によい結果であった。

認識率はきわめて良いが、学習に膨大な時間がかかり、かつ出力層に多数の音素を配置できない点が問題である。またバックプロパゲーションという学習法により音素境界が過度に強調され易い。音声認識への適用には今後より深い検討が必要と思われる。

2.4 音声認識装置の性能

2.4.1. 単語音声認識装置

区切って発話した単語音声に対する特定話者用認識装置として1972年アメリカThreshold Technology社よりVIP-100が商品化され、1978年には日本電気よりDP-100が発売されて、実用化の時代に入った。そして1981年にはアメリカ市場で7品目程度であった単語認識装置が、1988年には表1に示すように17品目¹¹⁾に拡

大した。これらの性能のすべては明かではないが、最上級の製品では語彙数が100~200単語程度であれば97~99%以上である。4桁程度の数字音声などのような連続単語音声でもほぼ同程度の認識率は達成できるようになっている。

現時点での単語音声認識装置の研究目標は主として認識語彙数の増大と不特定話者用認識装置の開発であって、2.3に述べた新しい音声認識アルゴリズムに基づく研究が進んでいる。特定話者用単語認識装置として開発中のTangoraシステム¹²⁾は5,000および2,000単語の認識誤り率がそれぞれ、2.9, 5.4%と報告されている。

2.4.2. 連続音声認識装置

1971年アメリカで音声理解システム(SUS)の研究が開始されて¹³⁾、文音声の連続的認識処理の研究が具体化した。ここでは言語情報に基づくトップダウン処理と音響分析によるボトムアップ処理との協調を計るとともに、音声合成技術を利用する対話形音声認識システムの実現を目標として研究が進められた。その結果CMUのHarpy, Sphinx, BBNのHWIMなどの音声理解システムが試作された¹⁴⁾¹⁵⁾。SphinxはHMMによる不特定話者用連続音声認識システムで、1,000単語の認識率が94~96%と報告されている。わが国でも京都工繊大、山梨大、ATRなどで種々の音声理解システムが試作されている¹⁶⁾。

3. 発話者の個人性情報の認識

3.1 話者認識処理の形態

音声信号を受信して発話者の個人性を判定する処理を話者認識と呼ぶ。話者認識装置の構成は音声認識と同じで、予め分析した標準パターンと未知話者パターンを比較・判定する図1のようである。この個人性の判定には、その応用から2つの形態がある。第1は、ある発話者未知の音声か予め登録されている複数の発話者の内の誰であるか、あるいは登録話者の誰でもないかの判定である。第2は、登録話者の内の特定の個人であると自称する音声か、その特定個人に間違い無いかの判定である。前者は脅迫電話の犯人を電話音声から鑑定したい場合などに相当し、話者識別(speaker identification)と呼ばれる。後者はIDカードなどの携行者がカード記載の本人に間違い無いかどうかを判定したい場合に相当し、話者照合(speaker verification)と呼ばれる。一般に話者識別では、登録話者数が増えると認識誤りが増えるが、話者照合では登録話者数が増えても認識誤りは変わらないので、信頼性が高いといえる。

3.2 サウンドスペクトログラムによる話者識別

1962年にサウンドスペクトログラムによる話者識別

表 1 音声認識装置の現状

会 社 名	商 品 名	語い数	認識率 (%)	性 能	用 途	価格 (\$)
Audec	SSB-1000	144	95	特定話者 孤立単語	音声データ 入力	250
	VoDialer	48	95	同 上	音声電話	349
Dragon	VoiceScribe	1000	NA	同 上	口 述	1195
IBM	VoiceComm	64	95-98	同 上	音声データ	1195
Intel	IWS	200	NA	同 上	検 査	NA
Inter- state	Vocaline SRB	400	98	特定話者 連続単語	音声データ	395
	Vocaline CSRB	100	99	特定話者 連続音声	同 上	4000
Kurz- weil	KVS3000	1000	NA	特定話者 孤立単語	口 述	6000
NEC	SAR-10	250	98	同 上	命令・制御	599
	SR-10	128	98	同 上	同 上	600
	DP-200	150	NA	特定話者 連続単語	音声データ	7500
Speech System	Phonetic Engine	20000	90	特定話者 連続音声	口 述	5000
TI	Speech Command	1000	NA	特定話者 孤立単語	命令・制御	995
Voice Connec	Introvoice V	400	98	同 上	同 上	495
Voice Controle	VCS1000	40	98.5	不特定 孤立単語	同 上	1000
Votan	Voicecard	13	93	同 上	電話応答	1350
	Voicecard	640	98	特定話者 連続音声	音声データ 検査	1200

の研究が発表されて、声紋による話者識別とその犯罪捜査への応用の可能性が提案された¹⁷⁾。ここでは 10 種の単語を用い、5, 9, 12 名の発話者の内の 1 名を声紋で識別させた結果、誤認識率はそれぞれ 0.4, 0.7, 0.8% に過ぎなかった。

2 名の発話者による単語「高原」のサウンドスペクトログラムは図 6 のようで、2 kHz 以下の母音ホルマントはよく似ているが、2 kHz 以上の高次ホルマントの周波数、/g/ の 5 kHz 付近の周波数成分には大きな差があって、同一人の音声ではないことが判り、声紋による話者識別の可能性が理解されよう。しかしながら声紋による犯罪捜査の有効性について、アメリカで大規模な実験が行なわれた結果、サウンドスペクトログラムは (1) 標準パターンと未知パターンとの発話時期差、(2) 犯人が標準パターンの話者の 1 人であることの有無、(3) 標準パターンと未知音声の文脈差、によって認識率に差が生じ、文献 17) は最良条件の結果であるが、最悪条件

では認識誤りが約 20% 増すことが明かとなった¹⁸⁾。視覚による話者識別ではこのような問題点があるが、発話状態が類似した同一単語、乃至同一文章音声のサウンドスペクトログラムによって、かなり高い確度で話者識別を行なえることが実験的に確かめられており、最近では視覚による話者識別をより客観性を高めるべく再評価する方向にある。

3.3 音響分析エキスパートによる話者認識

より信頼性の高い話者認識を実現するために、音声の音響的特徴量をその動的特徴を含めて精密に抽出することにより認識率を高める研究¹⁹⁾や、声紋による話者認識に見られた視覚による音声パターン認識の長所をコンピュータによるエキスパートシステムの形態で取り込む研究が行なわれるようになった。今後の進展が期待される研究分野である。

図6 2名の発話者による「高原」のサウンドスペクトログラム

4. おわりに

音声認識技術の最近の進展について、認識アルゴリズムの展開に重点をおいて説明した。紙数の関係で舌足らずになった箇所が幾つかあるが、これまでの文献などをご参照頂きたい。多数の研究者の協力と努力の結実として、より人間に近い能力を持った音声認識システムが実現することを期待している。

参考文献

- 1) 齋藤收三, 中田和男, “音声情報処理の基礎” オーム社, 1984, など
- 2) 迫江博昭, 千葉成美, “動的計画法を利用した時間正規化に基づく連続音声認識”, 音学誌, 27, pp.483-500, 1971
- 3) S. Furui, “A VQ-based preprocessor using cepstral dynamic features for speaker-independent large vocabulary word recognition”, Trans. IEEE, Acous. Speech, Signal Processing, Vol. ASSP-36, pp.980-987, 1988
- 4) 小松雄一, 好田正紀, “LVQ 法を用いた音素標準パターンクラスタ化の検討”, 信学会技報, 音声, SP 90-62, pp.41-46, 1990
- 5) 池戸丈太郎, 溜淵一博, 齋藤收三, “調音結合パターンを利用した単語音声認識”, 信学技報, 音声, SP 90-59, pp.17-23, 1990
- 6) F. Jelinek, “The development of an experimental discrete dictation recognizer”, P. IEEE, Vol.73, pp.1616-1624, 1985
- 7) J. Picone, “Continuous speech recognition using Hidden Markov Models”, IEEE ASSP Magazine, Vol.7, pp.26-41, 1990
- 8) L. R. Bahl, F. Jelinek and R. L. Mercer, “A maximum likelihood approach to continuous speech recognition”, Trans. IEEE, Pattern Anal. & Mach. Intell., Vol. PAMI-5, pp.179-190, 1983
- 9) 中川聖一, “確率モデルによる音声認識”, 電子情報通信学会, 1987
- 10) A. Weibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang, “Phoneme recognition using time-delay neural networks”, Trans. IEEE, Acous., Speech, Signal Processing, Vol.37, pp.329-339, 1989
- 11) P. Wallich, “Putting speech recognizers to work”, IEEE Spectrum, Vol.24, pp.55-57, April, 1987
- 12) A. Averbuch, L. Bahl, et al., “An IBM-PC based large-vocabulary isolated utterance speech recognizer”, Proc. IEEE ICASSP, pp.53-56, April, 1986
- 13) A. Newell, et al., “Speech understanding systems: final report of a study group”, Computer Science Dept., Carnegie-Mellon Univ., 1971
- 14) W. Lea, “Trends in speech recognition”, Prentice Hall Inc., 1980
- 15) K. F. Lee, H. W. Hon and R. Reddy, “An overview of the Sphinx speech recognition system”, Trans. IEEE, Acous., Speech, Signal Processing, Vol.38, pp.35-45, 1990
- 16) 小林 豊, 遠藤栄憲, 表 雅則, 新美康永, “音声インターフェイスをもつ情報検索システム”, 信学技報, 音声, SP 90-41, pp.15-22, 1990
- 17) L. Kersta, “Voiceprint identification”, Nature, Vol.196, pp.1253-1257, 1962
- 18) O. Tosi et al., “Experiment voice identification”, J. A. S. A., Vol.51, pp.2030-2040, 1972
- 19) 松井知子, 古井貞照, “音源・声道特徴を用いたテキスト独立形話者認識”, 信学技報, 音声, SP 90-26, pp.55-62, 1990

齋藤收三 (Shuzo SAITO)

1924年1月12日生, 昭和23年3月名大・工・電気卒, 昭和28年3月名大大学院特別研究生5年修了, 同年電電公社電気通信研究所入所, 同所調査役, 音響研究室長, 基礎研究部第4研究室長, 齋藤特別研究室長として通話品質, 音声情報処理の研究に従事, 昭和54年1月東京大学医学部音声言語医学研究施設教授, 音声科学部門担当, 昭和59年4月工学院大学教授, 工学博士, 視聴覚研究会(AVIRG)名誉会員, アメリカ音響学会フェロー会員。