音声情報入出力技術とその応用

中 田 和 男[†]

音声情報入出力技術すなわち音声認識と音声合成について、その技術の現状レベルを紹介するとともに、実用上の問題点について解説する。とくに放送技術との関連でその実用を考えたとき、ニーズと技術レベルがどれ位へだたっているかを明らかにし、実用化のためには体系的な事前評価が非常に大切であることを述べる。また技術の現状レベルを理解するために、現在の音声合成と音声認識の基本原則を述べ、その性能制約の原因となっている要因とくに連続化（調音結合）と不特定話者化（音声の個人性）について説明する。最後に装置の LSI の傾向をふくめて、今後の発展の方向をさぐる。

1. ま え が き

“Speak & Spell”（写真 1 参照）の実現以来¹⁾、多くの人々の関心が音声入出力装置、すなわち人間の声で機械に指令やデータを与え、また声の形で機械が情報を伝える装置、に高まるようになった。

音声合成チップ実現の背後に、新しい音声情報処理技術と LSI 化技術の結合による、音声入出力技術の新しい展開と応用を予見し期待したからである。商用 MOS・LSI 技術と線形予測理論にもとづく音声合成（音声出力）の飛躍的な小形化、低価格化実現の延長線上に、VLSI 技術と新音声認識技術の結び付きによる

音声認識（音声入力）装置の本格的な実用化を期待したわけである。

しかし問題はそう簡単ではない。音声認識は単に LSI 化というハード化技術だけで片づくものではなく、その実現には音声の本質に関する深い知識と経験を必要とするからである。

この解説では、音声入出力を主とした音声情報処理技術の現状を紹介するとともに、その制限のもとで、どのような用途に実用化の効用を期待しうるかを考えてみたい。

2. 音声入出力の効用と放送技術への応用

音声は人間にとって最も自然で使いなれた情報授受の手段であるから、音声によって機械に指令やデータを入力することができ、また機械から音声の形で情報を受けとることができれば、人間にとって便利であることはいうまでもない。

しかも文字や図形の場合にくらべて、音声の場合、その入出力のセンサー（変換器）がマイクロホンとスピーカーで非常に簡便であり、ワイヤレスマイクやイヤホンのように装着して移動することも可能であるから、機能的にも経済的にも利用価値が大きい。

さらに電話機を入出力端末として電話網と結合すれば、遠隔地での入出力も技術的のみならず経済的にも可能となる。

では放送技術の分野でどのような応用が可能かを考えてみよう。

（1）放送への直接の応用

写真 1 Texas Instruments 社の “Speak & Spell” 外観

[†] 東京農工大学 工学部
“Speech Input-Output Techniques and Their Applications”, by Kazuo Nakata (Tokyo University of Agriculture and Technology, Tokyo)

(a) テレビニュースの文字表題(テロップ)の
自動作成(音声認識とカナ漢字変換)

ニュースキャスターがニュースのタイトルを読むにつれて自動的に背後のディスプレイに表題が表示される。

あるいは放送の直前にテロップを作るにしても、音声認識によって自動作成する。

問題は固有名詞とくに人名と地名の認識と表記、同音異義語の正しい漢字表示への変換であり、当面人手の介入が必要となる。

(b) 天気図からの天気予報文の自動作成と自動
アナウンス(新しい情報とデータベースにもと
づく文章の作成と音声合成による出力)

どの程度自然な文と音声で作成、合成できて出力されるかがポイントである。

(2) 放送機器への応用

(a) カメラ操作の音声制御(音声認識による制
御)

ディレクターの思うままに複数台のスタジオテレビカメラの配置、移動、視野を制御する。問題はスタジオの実環境下でいかにS/N良くディレクターの声のみをピックアップすることができるかにあるだろう。その他、照明、音声、背景音楽、効果音などの音量制御も原理的には全く同じである。

(b) 時間調整のための音声の編集、伸縮(音声
の分析合成の応用)

放送指定時間長に合わせて映像は編集、伸縮が比較的容易である。画面にともなう音声もこれらの時間合せ処理によって、たとえば話しの途中でブツツリと切れたり、録音テープの早回し・遅回しのようにならないで時間長を伸縮できることが必要とされる。音声の分析合成技術で、ある範囲の編集、伸縮には応じられる。

(c) テープまたはフィルムの保管、取出しの音
声制御(音声認識とそれによる制御)

新しいテープやフィルムを収納するとき、そのインデックス(たとえばタイトル)を音声で登録し、格納棚の格納先のアドレスを音声で指示することによって所定の棚に格納される。取出しに当たっては、インデックスを音声で呼び出すことによって、必要なものが取出される。

このためには新しい格納機構の開発が必要であるが、物流システムでは自動倉庫の音声制御として類似のものが実用化されつつある。

(3) 受像(信)機への応用

(a) 音声によるチャンネル切り換えや音量調整

(音声認識とそれによる制御)

デモ用としてはすでに実際に試みられている。しかし手もとにマイクロホンを持つての音声制御が、スイッチや超音波、赤外光といった他のリモコン手段に対してどれだけメリットがあるかは疑問なしとしない。むしろ番組予約などにおけるキー入力の手代りに音声入力を使う方が有効ではないだろうか。

一例として上にあげたような音声入出力の利点が真に発揮できるためには、多くの場合、人間なみの音声入出力機能をもった装置ができて初めて可能だという場合が多い。現状では技術的に可能なことと、音声入出力として望まれる性能との間のギャップが大きく、その整合をとるシステムの配慮が実用上非常に重要である。

音声入出力の機能についてのユーザーの要求には限度がない。一方、技術的に可能なことには、特に音声の認識についてはまだまだ大きな制約がある。これをシステムの対処によって実用上さしつかえないものにしようかどうか、この“事前評価”が大切であり、これをおろそかにすると全くの無駄使いになったり、無理難題ということになりかねない。

それでは現在技術的にどんなことが可能か、それには音声入出力の技術的基礎となっている音声の合成と認識の原理のおよそを理解することが必要となる。

3. 音声合成の原理

音声合成の内容は、技術的には次の2段階に分けられる。

(1) 音声として使われる個々の単位音(母音と子音あるいは仮名文字1字で表記される単音節)を、どういう仕掛けで電氣的に発生させるかという“離散的な音声合成”。これはいわば人間の口の機能の単純なシミュレーションといえる。

(2) その個々の単位音をつないで、1つの単語なり句なり文なりを作る、いわば言語音としての“連続的な音声合成”。この場合、個々の単位音の自然なつながり方とそれが正しい自然な、ひとまとまりの単語なり句、文なりとして聞えるための声の抑揚(イントネーション)やアクセントのつけ方、いいかえれば連続化のための修飾の制御が複雑になる。これはいわば人間の頭と口の連動動作の高度なシミュレーションといえる。

音声の場合、文字のように、句読点さえ正しく付けられれば、単純な仮名文字の連結がそのまま句や文として受容されるというように簡単ではない。

(1)については、現在までにかなり詳しく、また明

快に物理音響現象として定量的に説明でき実現できるようにになっている²⁾³⁾。

その基本的な考え方は、音声波形（音声を物理的に具体化したもの）の生成は、エネルギー源としての“音源の生成”（通信用語でいえばキャリアの発生）と、そのうえに音声をあらわす音としての情報をのせる“調音作用”（通信用語でいえば変調）とからなると考えることである。

具体的にいえば、音源には声帯の振動による呼気の断続と、口の中に形作られた狭い空隙（調音点という）を呼気が通過するときに生ずる乱流の発生の2つがある。

前者（有声音源という）はほぼ周期的なパルス列で、後者（無声音源という）はほぼ白色のランダム雑音で近似される。

調音作用は、舌や唇や顎の運動による口の形の変化によって実現されるから、その形作る空間（声道という）の共鳴特性によって特徴づけられる。

これを外部から物理現象として観測すれば、調音作用は音声波形のスペクトルの（マクロな）エンベロップ特性、いいかえるとエネルギーの周波数分布として観測される。また2種類の音源はスペクトルの（ミクロな）内部構造、すなわち調波関係にある粗い線スペクトル構造（有声音）か、密な分布エネルギー構造（無声音）かの違いとして観測される。

以上の理解から、音声合成の基本回路（単位音の発生回路）は図1のような構成となる。

しかし、すべての音声の特性がスペクトル特性のみで特徴づけられるわけではなく、時間的な構造によっても区別される。すなわち定常的な母音 /i/ と /a/ に対して、その間を特性が時間的に動くことによって生じる半母音 /ja/ とか、定常的なまっす子音 /s/ に対して短い過渡的なはれつ子音 /t/ があるなど。

しかしいづれにしても、音声の単位音の離散的な生成は、図1の回路の比較的簡単な制御で、充分明瞭に

作り出す（合成する）ことができる。

ところが単位音（たとえば単音節）を連結して単語や句、文をあらわす連続的な音声を合成しようとするとき、問題は急にむずかしくなる。

その原因のひとつは、連続音声の発生における調音動作を、その結果である音声波形、あるいはそのスペクトルの時間的な変化として観測したとき、それらが外見上非常に複雑であって、そのあらわれ方を簡単に規則化することができないためである。これを“調音結合”と呼んでいる。いまひとつは、アクセントやイントネーションを与えるピッチ周期（声帯の振動周期）や音源強度を、単語なり句、文なりの全体にわたってどのように制御すればよいかという音源特性の連続化固有の制御法がまだ充分に解明されていないためである。

しかも人間は、こうした複雑な変化の中にもこそ、明瞭な連続した自然音声を聞きとり、反対に、単純な単位音の連結はひどく不自然に、ときにはその意味さえ不明と感じる。

現在のところ、文字表記のような単純な情報だけから実用上十分な自然さで任意の連続音声を合成することはできていない。ことに実用レベルで考えたとき、連続発声のための制御がマイコンとROM程度で簡易に実時間で行われることが必須であり、実用上はまだまだといわなければならない。

この問題に関連する研究を、純合成あるいは“規則による合成”（synthesis by rules）⁴⁾または文字からの合成（synthesis from text）と呼んでおり、現在音声合成研究のひとつの主要課題となっている。

これに対して、合成すべき音声の語彙が比較的少ない場合（たとえば 10^2 語のオーダー）には、実用上、人間によって発声された音声そのものを分析し、その情報から合成（厳密にいえば復元）することによって連続音声を作り出すことができる。

現在いわゆる音声合成装置として実用化されている

図1 音声合成の基本回路構成

図2 分析合成形音声合成装置の基本構成

ものはほとんどすべてこの方式であり、これを“分析合成方式” (analysis-synthesis system)⁵⁾と呼んでいる。

分析に当って、音声生成の機構として図1に示すような合成モデルを想定し、このモデルを制御するパラメーターの形で情報を抽出するから、大幅な情報圧縮が可能となる。この情報圧縮された分析結果をランダムアクセスメモリーに記憶しておき、必要に応じてそれを読み出し、図1の回路によって音声波形に復元する。

その基本構成を図2に示す¹⁾⁶⁾。LPC, PARCOR, LSPなどと略称される各方式(付録参照)がその代表的な実例である。

このとき必要とされる情報量は10～1 kbit/secの範囲にある。もしその量を1 kbit/secと仮定すれば(合成音の音質は原音声と直接比較すればやや劣るが実用上は十分な品質である)、1語の平均時間長を1秒として、128 kbitのROMチップ1つで128語の単語が記憶でき、それをもし16個まで接続すれば2000語以上の単語を出力することができる。

図1に示す音声生成モデルの導入によって、分析時に大幅な情報圧縮(音声波形そのままのデジタル化記憶に比べて60分の1以上)が可能となり、その記憶に半導体メモリーを使うことが量的にもコスト的にも可能になったため、記録されている中の任意の単語にテープレコーダー録音ではとても実現できない高速アクセスが可能となり、状況に応じた臨機の内容を選択出力することができるようになった。その結果、学習機器(そろばん練習機など)、機器使用に当っての操作の説明、指示(おしゃべり電子レンジなど)、警報の発生(自動車用)、キー入力の確認(音声電卓など)など広い分野への音声出力の応用が可能になった⁷⁾。

4. 音声認識の原理

音声生成についての現在の科学的な理解からいって、音声認識の最も基本的なプロセスは、図3に説明的に示すような入力未知音声波形と標準既知音声波形の“スペクトル次元での整合比較”である。

しかしスペクトルは音声波形の時間的なある1断面についての準定常的な短時間情報であるから、1つの単語に対応する音声波形全体は、こうしたスペクトル断面の系列として、たとえば図4のように記述される。

単語全体をスペクトルの系列としてとらえたとき、時間的な長さ、すなわちスペクトル系列の個数(これをフレーム数と呼ぶ)の異なる標準と未知入力の間で、

図3 スペクトルパターンの整合比較による音声認識原理の説明

図4 単語音声/9/ (キュウ) のスペクトル系列

どのように対応づけして整合を評価すればよいかという問題が起こる。

これに対して、主として3つの対策がとられている。

(1) 定フレーム数処理

1つの単語の中にとるスペクトル断面の数(フレーム数)をすべて一定(たとえば20フレーム)とし、時間の長さを規格化して処理する。フレームをどの時点にとるかにはいろいろの考え方があるが⁸⁾、最も簡単なのは等間隔にとる方法であり、等価的に線形伸縮比較となっている⁹⁾。

(2) フレーム単位音韻認識

整合は1断面のスペクトル単位(フレーム単位)でとる。いいかえると標準パターンは各カテゴリー(音

表 1 音声認識基本手法の分類

課題	要因	分 析 (特徴抽出)			整 合 尺 度	
		フ ィ ル タ ー	特 殊	Q の 低 い 単 共 振 聴 覚 シミュレータ ー	絶 対 値 距 離 自 乗 距 離	
普 通	少 数 チャンネル 多 数 チャンネル					
線 形 予 測 分 析			全 極 形 極 零 形	対 数 尤 度 比 (LLR) cosh 尺 度 荷 重 尤 度 比 (WLR)		
時 間 長 整 合	フ レーム 間 隔			整 合 法		
	可変：定フレーム数分析 (線形伸縮整合)			識別関数 (多次元空間) 単純パターン整合		
	固 定	音 韻 判 定	書 き か え 規 則			
			(直 接)	非線形伸縮整合	音 韻 単 位 D P	
					両 端 固 定	
					端 点 自 由	
		連 続 D P				
		伸 張 D P				

韻) 毎に 1 フレームのスペクトルパターンのみからなっており、未知入力に 1 フレーム毎に、フレーム単位で音韻認識される。

その結果はフレーム毎の音韻系列となり、同じ音韻が数フレーム以上続いたり、母音と母音や、母音と子音の移り変りの部分に間違った音韻があらわれたりする。これらはある規則でまとめ、書きかえて結果をまとめるか¹⁰⁾、このフレーム毎の音韻系列の形で後に説明する非線形伸縮整合を行うかする¹¹⁾。

(3) 非線形伸縮整合

単語に対応するスペクトル系列全体として入力と標準の間で整合をとる。その際、入力と標準のフレーム間で、許されるあらゆる対応(時間軸を逆行するような対応だけは許されない)をとって、最適に整合した場合をもって 2 つの単語間の整合とする。これを“非線形伸縮整合法”と呼ぶ。この処理を効率的に実行する手段として、いわゆる DP (ダイナミックプログラミング) 処理を利用する手法が定着しており、DP マッチングと略称されている¹²⁾。

いいかえれば、音声認識の基本処理は、フレーム単位で、スペクトルをどう記述し、そのフレーム間の整合をどう評価するかということと、スペクトル系列として整合をどう評価するかの 2 点に要約される。

前者は、音声認識のための分析法(特徴抽出法)と整合尺度の問題として、後者は、単語パターン全体としての整合処理の問題として、種々研究されてきた。それらの代表的なものを表 1 に示しておく。

このような音声波形のスペクトル系列の整合を基本とする技術で、どのようなレベルの音声認識が可能なのだろうか。

音声認識の機能を論じたり比較したりするとき、次

の 3 つの次元で考えることができる。

- (1) 認識できる語の数——有限か無限か、少数か多数か
- (2) 話者の範囲——特定(登録)話者か不特定話者か
- (3) 発声のモード——離散発声か連続発声か

いまのところ、特定(登録)話者の離散発声・少数語彙のもの(たとえば 1 桁単位の数字音声といくつかの制御語)はほぼ商用のレベルに達しており、特定話者の連続発声・少数語彙(たとえば数字音についてのみ、桁なしの棒読み連続発声を許す)の認識の実用化が本格化しようとしている。

不特定話者の音声認識は、ことに電話機を使つての遠隔音声入力用として銀行用などにニーズが強いが、不特定話者と電話回線の特性変動と不確定要因が 2 重に重なっており、まだ本格的な実用化にはいまひとつ問題がある状態といえよう。いずれの場合にも実用上は音声区間をみつけること、いわゆる音声区間の切出しが一番問題である。

一般的な連続発声音声の認識については、まだ研究段階にあるといわなければならない。

5. 音声入出力の問題点

現状の音声入出力の問題点は結局次の 2 点に集約される。

- (1) 連続的な発声の処理
- (2) 音声の個人差の処理

(1) は音声の生成機構の複雑さに起因するもので、これを一括して調音結合という言葉であらわしている。

簡単にいうと、単独の個々の離散的な発声のままの単音や単音節を単純につないでも決して明瞭で自然な連続発声音とはならない。また逆に、連続発声音中の個々の音の物理特性は、離散発声のときのその音の特性とはかなり違ったものとなっており、前後の音や、語、句、文といったものの全体の構造と深くかかわっている。

文字にたとえていえば、音声は活字のような構造ではなくて、手書きの、しかも草書体のような構造をもっているといえよう。

草書は一見でたために書かれているように見えるが、そのくずし方とつけ方には規則があり、それを修得した人にはごく自然に読み書きすることができる。

これと同じように、連続発声の音声もまた、人間にとっては、しごく明瞭であり自然であり、了解も発生

も容易である。ただその規則を陽にアルゴリズムとして抽出しえない現在の我々の技術にとっては大へん難解なのである。“音声発生機構の解明”が、調音と音源を含めて、さらに必要なゆえんである。

(2)は音声認識における不特定話者の取り扱い、音声合成における声の自然さ、さらには特定の人の声らしさの再現にかかわる。

音声の個人性は、主にその人の音声発生器官の生理的な条件に起因する静的な長時間平均スペクトル的な変動と、その人の発話習慣の学習条件に起因する動的な時間構造的な変動の2つから成っていると考えられるが、現実にはこの2つを分離することがむずかしく、有効で定量的な正規化手段がないまま結局それらを“こみ”にした単語単位の“マルチテンプレート”(複数標準パターン)で対処するしかないのが現状である¹³⁾。

電話音声となると、そのうえにさらに回線特性の変動あるいは送話機の非線形特性とその変動、周囲雑音の重量といった問題が加わり、問題はいつそう解決が困難になる。

6. 音声入出力実用化の検討

さて、それでは現状の音声入出力装置は全く使いものにならないのかといえ、決してそうではなく、多くの実用例をあげることができる。そこで音声入出力とくに音声入力を実用化したいと考えたとき、どういう事前評価をすればよいかを考えてみたい。

- (1) 音声入出力導入のメリット——単なるサービスの向上、マン・マシン性の向上あるいは省力化だけなのか、あるいはその導入によって従来不可能であったことが可能になり、あるいはニーズの強い全く新しい機能が実現されるのか、後者であるほど望ましいことはいうまでもない。
- (2) 他の入出力手段とのコスト比較——キー入力やディスプレイ出力などとの比較をすることが必要である。
- (3) 導入しようとするシステム全体からみて、その効率、機能に音声入出力の実現がどう影響するかを全体として評価する。
- (4) システム的に誤り対策、訂正手段、補助手段は充分可能か——音声入出力では100%完璧ということを保証することはできない。そこで誤ったときの対策が必要である。
- (5) システム的に要求機能をはっきりと定式化でき、それを技術的に可能な範囲に限定することができるか——ただ便利だからという安易な態度は

危険である。語彙数、話者条件、話者の態度(協力的か否か)、発声モード、周囲騒音、などをすでに述べた技術的に可能な範囲におさえることができるかどうか、(4)とも関連して導入メリットがあると評価される範囲で要求される音声認識率はいくらか、それは音声認識にとって過大な要求となっていないかどうか。

- (6) 現状技術でニーズに応えるために、システム的な考慮(主として各種条件の限定、入出力手順の構成など)を十分に払ったか——音声出力での誘導指示や確認を使うと、かなり確実有利に条件を設定できることが多い。
- (7) 人間の学習効果を期待できるか——専用オペレーターによる使用では、この効果が大きいことが多くの実用例で立証されている¹⁴⁾。しかし、そのためには数週間にわたる導入試用期間が必要である。また長期的にみたオペレーターの心理、疲労などへの考慮も必要である。

このような事前評価がなされたうえで導入された時の音声入出力の効用は大きく、満足すべき結果がえられている。

その音声入力実用化の一例は、物流システムにおける仕分けの省力化であり¹⁴⁾、他の例は生産の検査ラインにおける能率の向上である¹⁵⁾。前者は機器の音声制御の適例であり、後者は音声による遠隔分散端末からのデータ入力の好例である。

常識的に達成可能な音声認識率は、特定話者の離散発声、少数語彙で97~99%、同じく連続発声で95~97%(離散発声単位で)位である。

7. 音声入出力装置のLSI化

音声出力(合成)装置のLSI化は、Speak & Spellの実現以来順調にすすんでおり、方式的にも、PARCORからLSPへと発展し、別にADPCM(適応差分PCM)やAΔM(適応デルタ変調)のものも作られている。PARCOR方式についていえば、CMOS1チップのものが作られ、消費電力の面でPMOSやNMOS版にくらべていちじるしい改善が行われた¹⁶⁾。その仕様の一例を表2にあげておく。

音声入力(認識)装置については、さすがにまだその全機能を1チップにまとめるというところまではいっていないが、部分的なLSI化は行われている。

最も機能的にすすんだものは電電公社武蔵野電気通信研究所の試作品¹⁷⁾で、標準パターンの登録機能のものを除けばほとんどの音声認識機能が1チップのうえにもられている。その仕様を表3に示しておく。

表2 音声合成 LSI の性能

項目	形名	HB3880B	HD61885
システム構成		3チップ	1チップ*
方式		PARCOR	PARCOR
発売		1979年	1981年
デバイス技術		PMOS	CMOS
電源電圧		-10V±10% -5V±5%	3.6~5.5V
消費電力	発声	70mA	25mW
	休止	1mA	2.5mW
クロック		640kHz	800kHz
ピン数		28	28
チップサイズ		6.0×5.5mm	5.5×5.5mm
データレート (kbit/sec)		2.4/4.8/9.6	2.5/5.0/9.9
フレーム周期		10/20ms	10/20ms
サンプリング		8kHz	10kHz
接続可能 ROM		128k×16	128k×16*
内蔵 ROM		なし	32kbit
発声時間		50~100s	5~20s**
デジタル フィルター	精度	10×15bit	10×15bit
	段数	8/10	10
音源	有聲	パルス/sin ² ωt	積分音源
	無聲	M系列符号	M系列符号
D/A		±8bit	±9bit
パラメータ補間		2.5ms	2.5ms
発声速度可変範囲		0.6~1.4	0.75~1.25
応用製品		そろばん練習機 自動車用警報	時計 電卓、家電品

* ROM 外付拡張可

** 1チップ構成のとき

表3 1チップ音声認識 LSI 仕様

項目	仕様
認識アルゴリズム	音韻尤度にもとづく離散単語 音声認識、特定話者
特徴パラメーター	0~10次自己相関係数
時間軸整合	DP マッチング
サンプリング	8 kHz
フレーム周期	16 ms
認識単語長	0.3~1.6 sec
認識単語数	最大 200 語
認識所要時間	発声終了後 350 ms以内
素子数	39,800
論理ゲート数	約10kゲート
チップサイズ	5.7×6.4 mm
クロック	4 MHz 1相
電源電圧	5 V単一
消費電力	1.3 W
パッケージ	64ピン DIP
LSI 技術	2ミクロン E/D・MOS
	poly-Si, Al 2層配線

特徴抽出のための分析部も、いわゆる信号処理プロセッサ (DSP) を使って LSI 化されつつあり¹⁸⁾、非線形伸縮整合部についても日米両国で LSI 化がすすめられている¹⁹⁾²⁰⁾。近い将来、機能別に LSI 化され、1ボードにのせられた本格的な音声認識装置が市販されるようになるものと期待される²¹⁾。

VLSI による実現については、VLSI の技術そのものが現在開発中だから、まだ直接の試みは発表されていない。しかし LSI 技術そのものが最近ますますシステム指向を強くし、system on a chip という考え方の設計でその利点をフルに活かそうとしている。音声認識や音声の分析合成はその恰好の目標とされており、ここ数年の間に、続々と新しい LSI 化が発表発売されるようになるものと予想される。

8. むすび——今後の方向

さて、音声入出力技術は今後どのように発展していくであろうか？ この難問にあえて著者の個人的な見解(それは個人の好みに左右されている危険性がある)を示して、この解説のむすびとしたい。

(1) 音声出力 (合成)

(a) 合成音声の品質改善 PARCOR や LSP 方式の音質を ADPCM なみに向上させる。そのひとつの試みとして残差波形を音源波形として利用することが検討されている²²⁾。さらに今一步すすめて、音源とくに有聲音源である声帯振動パルスの発生を、単なるパルス発生ではなくて、より実体的な声帯振動機構のシミュレーションに拡張する、いわゆる実体音源の導入も検討に値しよう²³⁾。

(b) 音声実時間分析の実現 PARCOR や LSP 方式の音声合成の応用がいまひとつ広がらないのは、合成のための分析がやっかいなことに原因のひとつがある。音声の実時間分析チップの開発によってこの問題を解決すれば、ADPCM なみの手軽さで PARCOR や LSP が使われるようになり、もっと応用分野が広がるであろうと期待される。

(c) 高能率伝送方式の開発 ADPCM の下限界が 24 kbit/sec、PARCOR、LSP の上限界が 10 kbit/sec であり、16 kbit/sec を中心とする高能率伝送方式が欠けている。この程度まで高能率化して品質は ADPCM なみということになれば、voice mail (音声メッセージの蓄積交換) や自動車電話への応用など新しい展開が期待できる。ベル研の SBC (Sub/Split Band Coding)²⁴⁾ や ATC (Adaptive Transform Coding)²⁵⁾、電電公社通研の APC-AB²⁶⁾ など、そこを狙った研究開発といえよう。

(d) 規則合成の実用化 音声合成による音声出力が本当のマン・マシンインターフェイスとして使われるためには、なんといっても規則合成の実用化が必要である。マイコンと ROM による規則合成の実用化、それは音声合成研究のひとつの夢である。

(2) 音声入力 (認識)

(a) 不特定話者の音声認識 音声認識の広い実用化のためには、たとえ少数語彙であっても、不特定話者の音声認識、すなわち誰の声でも認識できるといった音声入力技術の開発が急務である。そのとき、語彙が少なければ少ないほど、それを更新する手続きが簡単であることが実用上必要となる。

どんなに認識率が高くても、語彙が変わるたびに何百人何千人の人間の声を集めて分析しなければ標準パターンや識別関数が作れないというのでは問題である。

そのコストにたえられるほどの大きなニーズがいくつあるだろうか。“音声の個人性の解明”が必要な理由はこれだけでも明らかである。

(b) 装置のLSI化、ワンチップ化 これについてはすでに説明した。

(c) 雑音対策 音声認識の技術が研究室から現場に出て一番問題になるのは、雑音を含めて音声区間を正しくピックアップし、非音声区間の入力で誤動作しないことである。ことに背景の人の話し声とか、息や咳、舌うち、無意味な音声（アーとかエーとかなど）の除去などは大へんむずかしい。マイクロホンの近接音場に対する感度が良くなればなるほど、これらの妨害がクローズアップされてくる。しかもこれらは偶発的なもので、シミュレーションではなかなか対策しにくい問題である。また S/N の悪い音声の認識も実用上重要な課題である。

世人の音声認識実用化への期待にこたえるためには、現状の技術であらゆる工夫をこらして、1%、2%の認識率の向上を目ざして努力するという研究も必要である。しかしこの解説でもその一端を示したように、現在の手法は多分に便宜的なものにすぎない以上、一般的な（語彙の大きい）連続音声の認識や不特定話者の音声認識のためには、もっと音声の本質を明らかにするような地道な研究が必要と思われる。

それを怠っては、この音声ブームも一夜の花火と消えてしまうのではないかと恐れている。

最後に、音声入出力についての日本語の参考図書をあげておく。

(1) 音声情報処理の基礎について

(a) 手軽な入門書：音声の合成と認識，中田和男，総合電子出版

(b) 本格的な参考書：音声情報処理の基礎，斉藤収三，中田和男，オーム社

(2) 音声認識の専門書：音声認識，新美康永，共立出版

(3) 音声合成の専門書：音声の線形予測，鈴木久喜訳，コロナ版または(1)の(b)

(4) 音声一般について

(a) 音声，中田和男，コロナ社

(b) 聴覚と音声，三浦種敏ほか，電子通信学会

(c) 音声科学，藤村靖ほか，東大出版会

(昭和57年4月26日受付)

[参 考 文 献]

- 1) R. Wiggins and L. Brantingham: Three-chip system synthesizes human speech, Electronics, Aug. 31 (1978)
- 2) G. Fant: Acoustical Theory of Speech Production, Mouton's-Gravenhage (1960)
- 3) J. Flanagan: Speech Analysis, Synthesis and Perception, Springer-Verlag (1972)
- 4) 藤崎博也，広瀬啓吉：規則による音声合成，日本音響学誌，**37**, 5 (1981) 204
- 5) 板倉文忠：スペクトル符号化に基づく音声分析合成，日本音響学誌，**37**, 5 (1981) 197
- 6) 佐藤，宮原，中田ほか：PARCOR 形音声合成装置のLSI化，信学技報，**SSD 79**-122 (1980)
- 7) 中田和男：音声合成とその応用，電学誌，**101**, 2 (1979) 117
- 8) 千葉成美，旦理誠夫ほか：音声認識システム，大形プロジェクト・パターン情報処理論文集 (1980) 157
- 9) T. B. Martin: Practical application of voice input to machines, Proc. IEEE, **64**, 4 (1976) 487
- 10) R. De Mori: A descriptive technique for automatic speech recognition, IEEE Trans., **AU-21**, 2 (1973) 89
- 11) 長島，中津：音韻単位の標準パターンを用いた実時間単語音声認識装置，音声研究会資料，**S 78**, 22 (1978)
- 12) 迫江博昭，千葉成美：動的計画法を利用した音声の時間正規化にもとづく連続音声認識，日本音響学誌，**27**, 9 (1971) 483
- 13) L. Rabiner, S. Levinson, et al: Speaker-independent recognition of isolated words using clustering techniques, Trans. IEE, **ASSP-27** (1979) 336
- 14) 中田和男：音声の合成と認識（後編第1章），総合電子出版 (1979) 79
- 15) 姫路，笠井，竹田：音声入力応用官能的検査情報処理システム“VOCAL”の開発と適用，日立評論，**63**, 12 (1981) 27
- 16) H. Hara, S. Torii, Y. Toba and K. Nomiya: High-Performance Process Fits Speech System on a Chip, Electronics, Oct., 6 (1981)
- 17) 宮原，矢野，浜口，江原：1チップ単語音声認識LSI，信学総全大，**S 7-8** (1982)
- 18) 中田，宮本，稲田：PARCOR 実時間分析の検討，音響学会研究発表会 (1982/3) 3-4-14
- 19) 北爪，大平，市川ほか：連続音声認識装置認識論理部の構成と機能，音響学会研究発表会 (1981/12) 2-1-15
- 20) N. Weste, D. Burr and B. Ackland: A Systolic Processing Element for Speech Recognition, ISCC 82, FAM 19-6 (1982)
- 21) 佐藤，大城，木原ほか：SR-100 音声入力ターミナル，信学総全大 (1982) 1397
- 22) 新居，古屋：切出残差駆動波を用いたPARCOR 合成音声の明瞭度試験，音響学会研究発表会 (1982/3) 3-4-10
- 23) 石坂謙三，J. Flanagan: 声帯音源の自励振動モデル，日本音響学誌，**34**, 2 (1978) 122
- 24) R. E. Crochier, et al: Digital coding of speech in sub-band, Bell Syst. Tech. J., **55** (1976) 1069
- 25) R. Zelinski and P. Noll: Adaptive transform coding of speech signals, IEEE, Trans., **ASSP-25** (1977) 299
- 26) 菅田雅彰，板倉文忠：音声の適応ビット割当て予測符号化，音声研究会資料，**S 80-2** (1980)

〈付 録〉 LPC, PARCOR, LSP の説明

現在の音声分析、合成の主流技術は、いわゆる線形予測理論にもとづくものである⁵⁾。

音声波形のサンプル値 $\{y_n\}$, $n=0, \dots, p$ の間に,

$$y_n = \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + \dots + \alpha_p y_{n-p} + \varepsilon_n \quad (1)$$

という線形予測の関係が成り立つと仮定する。ここで ε_n はその予測誤差である。

この仮定は (1) 式の z -変換をとることによって,

$$\left. \begin{aligned} Y(z) &= H(z) \cdot E(z) \\ H(z) &= 1/A(z) \\ A(z) &= 1 - \alpha_1 z^{-1} - \alpha_2 z^{-2} - \dots - \alpha_p z^{-p} \end{aligned} \right\} \quad (2)$$

と等価であることは容易にわかる。

そこで予測誤差 ε_n の 2 乗平均値 $\overline{\varepsilon_n^2}$ を最小にするように、線形予測係数 $\{\alpha_n\}$ を求めるのが線形予測分析の基本である。

(1) 式から、付図 1 に示すような回路構成で、入力 ε_n と線形予測係数 $\{\alpha_i\}$, $i=1, \dots, p$ を与えれば、音声波形 y_n を求めることができる。このような原理にもとづく音声合成方式を LPC (linear predictive coding: 線形予測符号化) という。

この方式を、より実用的に、またより冗長さを少なく改良したのが PARCOR (partial auto-correlation: 偏(自己)相関) 方式である。T. I. 社の Speak & Spell

付図 1 線形予測による音声合成の原理

をはじめとして国内各社の音声合成 LSI はほとんどこの方式によっている。

さらに、低ビット制御 (2.4 kbit/sec 以下) のときの音質を改善すべく改良されたのが LSP (line spectrum pair: 線スペクトル対) 方式である。

これらはすべて、日本の電電公社武蔵野電気通信研究所の斉藤、板倉の両氏をリーダーとする音声研究グループによって発明考案されたもので、世界に誇るべき研究成果である。

なお、上記諸方式の詳細については、本文中にあげた参考書、とくに (1) の (b), (3) を参照されたい。