音声の高能率符号化と認識・合成

国際電信電話株式会社 研究所 樽 松 明

音声高能率符号化の最近進展が著しい方式の紹介として、適応量子化差分符号化 (ADPCM)、適応予測符号化、サブバンド符号化、線形予測符号化を重点的にとり上げ動向を述べた。音声認識では、単語認識と連続音声理解を、音声合成では、音声応答および文字記号からの音声合成について、それぞれ現状と課題を述べた。

1. ま え が き

音声信号のデジタル処理の研究は、デジタル伝送路の限られた伝送容量を有効に利用するための高能率符号化の分野と、コンピュータに対するマンマシンコミュニケーションをより便利にするための音声認識および音声合成に関して、最近盛んに行われている。

音声信号に含まれる冗長性を除去して、最低限の情報のみを抽出して伝送することは、情報理論が研究されはじめた頃からの夢であり、音声信号から特徴を抽出して記号に変換すること、記号から自然な音声を作り出すことが、それぞれ音声認識と音声合成の究極の目標である。

音声高能率符号化については、音声の性質に着目して、時間領域および周波数領域の両面からのいろいろな方式が提案されているが、許容される品質の程度に応じて、高能率化の限界に向けた挑戦が行われている。

狭帯域の符号化は、品質の劣化は伴うが、むしろ伝送路帯域の制約内でのデジタル伝送や伝送路の有効利用の観点から研究が行われ、音声とデータの統合通信の実現がひとつの目標になっている。

音声の分析合成系は、ボコーダーとよばれ、分析系で音声信号内の有意な情報をパラメータとして抽出するが、この過程は音声認識の最初の段階での処理過程にそのまま利用されうるものである。合成系はパラメータから合成器を働かせて音声を再生するもので、音声合成の波形出力の過程そのものである。

このように音声高能率符号化と音声認識合成の研究は、一体となって進められている。

本文では、音声高能率符号化方式の最近の主な研究

の紹介および動向と、音声認識と音声合成について、それぞれ現在における研究レベルと将来の実用化の動向について述べる。

2. 音声高能率符号化

デジタル信号処理技術の進展と大規模集積回路の経済性の向上によって、音声信号の高能率符号化伝送の研究が活発になっている。研究の目標は、最小のコストで、最小のチャンネル容量で、最高の可能な品質を得る音声デジタル符号化方式を樹立することである。

音声符号化伝送速度と品質とは、互いに関係があり、図1に示すように、品質の良い方から、符号化速度が200~64 kbps の範囲を放送品質、64~16 kbps の範囲を一般公衆品質、9.6~7.2 kbps の範囲を特定通信品質、4.8 kbps 以下を合成音品質と位置づけできる。

放送品質とは、音楽など広帯域な信号をひずみの少ない高品質で出力するもので、放送用音声のデジタル伝送に適している。一般公衆品質とは、現用のアナログ音声伝送品質（すなわち周波数範囲 200~3200 Hz、信号対雑音比 30 dB 以上、高調波ひずみ 2~3% 以下）に相当する品質を指す。特定通信品質とは、十分に了解性は高いが、品質の劣化が検知され、気がつく程度のひずみがあり、誰が話しているかという話者識別が多少むずかしいという品質である。合成音品質では、了解性はあるが、自然性の点で難点があり、いかにも合成音という感

図1 音声符号化伝送速度と品質の関係

“High Efficiency Speech Coding, Speech Recognition and Speech Synthesis” by Akira Kurematsu (Research and Development Laboratories, Kokusai Denshin Denwa Co. Ltd., Tokyo)

じがあり、話者識別もかなり劣り、品質の程度が話者に依存して一定でない場合がある。

音声高効率符号化の方式は、波形符号化系と分析合成系の2つに大きく分けられる。波形符号化系は、音声信号の性質を利用して高効率に符号化し、復号化して元と同じ波形を再現するものである。一方、分析合成系は、音源から音声が生産されるモデルにもとづき、音声の特徴づけるパラメーターを抽出して符号化して伝送し、受信側で、音源信号とフィルタを送られたパラメーターで制御しながら音声を合成するもので、ボコーダーとよばれている。

波形符号化系は高い符号化速度で品質が良い所に用いられ、分析合成系は音声を低ビットに符号化できるので、伝送帯域の大幅な経済化が図れる。

以下、波形符号化系と分析合成系の主な方式の最近の研究をとり上げて動向を述べる¹⁾。

2.1 適応量子化差分符号化 (ADPCM)

音声信号中に存在する冗長性を除いて、PCM より高効率の符号化をする方式の中で、最も代表的なものが ADPCM である²⁾。音声波形では、相隣るサンプリング点の振幅値の間に高い相関があるため、相隣るサンプリング点の差分値の分散は、元の振幅値の分散より小さくなる。分散が小さいほど符号化誤りは小さくなるので、差分値を符号化することにより信号対雑音比 (SNR) は向上し、同じ SNR を得るには少ないビット数の符号化で済むことになる。

音声信号は時間的に振幅が大きく変動するので、量子化の細かさを信号レベルに応じて適応的に可変にする方法が有効であり、これを適応量子化差分符号化方式 (ADPCM) とよんでいる。図2に ADPCM の構成を示す。

量子化のステップサイズ (時刻 $r+1$) Δ_{r+1} は、時刻 r における符号器出力の大きさ (P_r) によってきまる関数 $M(|P_r|)$ と、時刻 r でのステップサイズ Δ_r との積、

図2 ADPCM 符号器の構成

すなわち、 $\Delta_{r+1} = \Delta_r \cdot M(|P_r|)$ で与えられる。この方式では、これまでに得られなかったような広いダイナミックレンジにわたり高品質が得られ、32kbps で一般公衆通信の品質が得られており、実用に供し得る段階といえる。

2.2 適応予測符号化 (APC)

音声信号が時間的に定常でないという性質をさらに積極的に利用し、音声信号の予測符号化をよりきめ細かく行うことによって符号化レートを下げることができる³⁾。適応予測は2つの面から施すことができ、ひとつは相つづく音声サンプル間の相関にもとづくもので、他のひとつは有声音声で存在する音声の概周期性にもとづくものである。

前者の予測器は、相つづく p 個のサンプル値からの線形予測を行い、予測係数は、予測しきれなかった値をさす予測残差を、一定のフレーム長の間 (10~30 ms) にわたって最小とするようにきめられ、フレームごとに予測係数の値を更新していく。後者の予測器はサンプル値を遅延回路で遅らせピッチ長だけずらしたサンプル値間の差

図3 一般的適応予測符号化器の系統図

を最小にするように、予測器のゲインとピッチ長を定める。

さらに人間の耳で知覚されるひずみを小さくすることが最終の目標であることと考えると、量子化雑音を音声信号のスペクトルに依存する形で評価でき、より能率の良い符号化を施すことができる⁴⁾。すなわち、耳では信号のスペクトルが大きい周波数では雑音がマスクされることを利用して、予測符号器の量子化雑音のスペクトルを信号のスペクトル成分が小さい所で小さくなるように変えるフィルターをおく。

図3にこれらの適応的予測による一般化した符号化器の系統図を示す。ここで、 P_s は音声サンプル間の相関を取り除く予測器、 P_d はピッチ周期にもとづく予測器、 F は量子化雑音を知覚するひずみを小さくするためのフィルターであり、 d_n は P_s による予測残差、 d'_n は P_d による予測値、 f_n は量子化誤差をフィルター F に通した後ピーク制限をした値で、 q_n は $d_n - d'_n - f_n$ の値である。ADPCM にくらべると回路は複雑であるが、16 kbps 程度の符号化速度で、PCM と同程度の品質が得られており、さらにビットレートを低くする研究がすすめられている。

2.3 サブバンド符号化 (SBC)

音声信号は周波数スペクトルが一定ではなく、長時間スペクトル密度は、周波数が2倍になると約 6 dB の傾

斜で小さくなることがわかっている。このような周波数領域での性質を利用した高能率符号化方式として、音声信号を帯域通過フィルターによりいくつかの周波数成分に分割し、それぞれの成分を別々に符号化するというサブバンド符号化方式 (SBC) がある⁵⁾。周波数分割数は、2～8 個がとられる。

サブバンド符号器の構成は図4に示すように、帯域通過フィルターと適応 PCM (APCM) 符号器とマルチプレクサーからなる。

SBC の利点は各周波数バンドで、そのエネルギーが小さいときにはそれに応じて量子化ステップサイズが小さくするように変化させて量子化雑音の影響を小さくできることである。周波数バンドの幅は、聴覚における感度に対応するように、低い周波数では狭く、高い周波数では広くとる。また、量子化ビットの割当も、有声音のエネルギーが集中する低い周波数では多く割当て、無声音や雑音によるエネルギーが出現する高い帯域では少なく割当てる。4 個の周波数バンドで 16 kbps の SBC の品質が、ADPCM の 22 kbps の品質にほぼ等しいという結果が得られている。

SBC は、新しい CCD フィルター技術の適用が可能で、簡単なハードウェアで実現できる見通しが大きい。現在より低ビットをねらって努力が払われている。

また、SBC はニュースや音楽などの放送用の広帯域の信号を、56 kbps または 64 kbps の通常のデジタル伝送回線の速度で伝送する用途にも利用できる⁶⁾。

図5は 7 kbps の帯域幅の音声信号を低い量子化雑音で 56 または 64 kbps で符号化する場合の系統図で、0～3650 Hz の低い帯域は 8 kHz でサンプルして4ビット ADPCM で符号化し、高い方の 3.6～6.8 kHz の帯域は 7 kHz でサンプルし、56 kbps で伝送するときは、APC 符号器を3ビットで量子化し

図4 サブバンド符号化方式の系統図

図5 高品質の音声高能率符号化方式の系統図

て $32+21=53\text{ kbps}$ で信号を表わし、 64 kbps の伝送では高域を4ビットで符号化し、それぞれ同期信号などの付加情報をつけて伝送する。この方式は、商用のデジタル回線の伝送速度でハイファイの高品質な音声伝送ができるので、音声プログラムの放送用のみならず会議電話の音声伝送用にも有効である。

2.4 線形予測符号化 (LPC) ボコーダー

線形予測符号化 (LPC) ボコーダーの原理は、音声信号のある時点における値は、それ以前の連続したサンプル値 (8個~12個) から線形予測できることにもとづいている⁷⁾⁸⁾。予測係数は、予測値と真値との誤差の2乗を最小にするように決められる。線形予測分析をモデルで表わすと、音声信号が全極型のデジタルフィルターによって生成されるということになり、これは母音の音声にはフォルマントとよばれる共振周波数が4~6個あることによく合致している。また子音についても全極モデルで充分によく表わし得ており、また、発声する際、喉から口あるいは鼻の先まで音声が通過する道程を声道とよんでいるが、声道の状態を一定長の円筒管を接続した形でモデル化したときの声道の断面積比が線形予測係数から導かれるパラメーターと1対1に対応することが理論的に裏づけられている。このことから、線形予測分析は音声の物理的事象をうまくとらえた、すぐれた分析方法とされている^{9)~12)}。

LPC ボコーダーの構成は図6に示される。線形予測係数をそのまま用いる代りに、量子化に対して影響が少ない、断面積比係数とよばれるパラメーターを、音声信号の自己相関分析をした後に導出する。音源情報として、有声音無声音の判定と有声音区間におけるピッチ周期の抽出を行う。各パラメーターは、一定長 (たとえば 20 ms) を1フレームとして、フレームごとにパラメーターを抽出して符号化する。表1に $2,400\text{ bps}$ の場合の符号化の一例を示す¹³⁾。

合成側では、音源として有声音の場合にはパルス、無声音の場合には一様スペクトルの雑音を用い、ピッチ情報によってパルスの間隔をピッチにあわせ、音の強さの値に応じて音源の大きさを増幅する。フレーム間隔ごと

表1 LPC ボコーダーの符号割当の一例
($2,400\text{ bps}$, フレーム長 20 ms)

断面積比係数								振幅	ピッチ と有 無 声	同期	計
g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8				
7	6	5	4	4	4	3	3	5	6	1	48

に伝送されるパラメーター値を適当な間隔で線形補間して合成フィルターに加えて、元の音声に近い出力波形を得る。

LPC ボコーダーの音声品質は、了解性は充分あるが、伝送する情報を圧縮しているために自然性がいく分そこなわれている。ボコーダーは伝送帯域が制限されている移動通信や、限定された専用回線の範囲で音声の内容が伝達できれば充分で、むしろ通信回線の有効利用の必要性が高いという用途に使われはじめている。LPC ボコーダーの今後の課題は、不自然性の改善と、周囲雑音が大きいときの品質劣化に対する対策で、これらを解決すべく研究がつけられている¹⁴⁾¹⁵⁾。

3. 音 声 認 識

音声自動認識は、わが国通産省のパターン情報処理大型プロジェクトや ARPA* の音声理解システムプロジェクトが刺激となって最近の技術的進展が著しい。音声認識装置は主としてコンピュータの音声入力として用いられ、音声人間にとって最も自然な入力手段で特別な訓練を要せずに容易に使用でき、また荷物の仕分けなど手作業に関連したデータ入力に音声により容易に行える特徴を生かして、実用的に使用できる装置が出現しはじめている。

音声認識は、認識する対象のちがいに、単語音声の認識と連続会話音声の認識に分けられる。また、音声認識は話者に依存して性能が変わることがあるので、特定話者用か不特定話者用かの区別もつけて研究が行われている。

3.1 単語音声認識

* ARPA: アメリカ合衆国防省の研究開発機関, Advanced Research Projects Agency の略。

図6 線形予測符号化ボコーダーの構成

単語音声の認識は、数字、地名、品名あるいはコマンド用語などを単語を単位として認識するもので、単語を1つのまとまったパターンとして扱っている。認識できる単語の数は次第に増加して、現在 100~500 語を認識できるシステムが発表されている^{16)~18)}。この種のシステムでは、話者によって標準パターンを学習しなおす方式をとっており、使用者が認識対象の全部の単語を発声して標準パターンを作成し、これを登録しておき、認識の際は、発声した単語の音声パターンと標準パターンとのマッチングを行うことによって認識する。入力音声の処理は、音声を十数個の帯域フィルター群に通し、その出力から特徴を抽出するものが多い。

入力音声パターンは発声の工合によって発声速度が異なるので、標準パターンとのマッチングは、動的計画法によるマッチング (DP マッチング) を用いて、長さが異なるパターンのマッチングを最適に行うようにしている。さらに DP マッチングの範囲を広げて、認識対象の単語を連続して発声しても認識できるようになってきている。

一方、不特定話者を対象とした単語音声認識の研究も進められており、日本電気による日本語の 100 語程度の語彙を対象としたシステム¹⁹⁾や、ベル研における英語の数字とアルファベットの認識実験²⁰⁾が報告されている。いずれも入力音声の前処理部で、個人性によるパターンの相異を除去するために適応等化を行った後、自己相関分析をしている。

日本電気のシステムでは 12ms のフレーム周期で 20 個の自己相関係数を求めた後、K-L 変換 (Karhunen-Loève 変換) により 7 次元に圧縮し、これに振幅を加えて 8 次元の分析パラメーターを得ている。その後、単語の種類や個人差によって語長が異なることを解消するために、各単語について重要な特徴の抽出手順を単語テーブルに書いておき、これを参照しながら、比較的安定な抽出点として各単語当り 15 点を選んでいる。この結果、入力音声は $8 \times 15 = 120$ 次元のベクトルに変換される。次に識別部では、区分的線形識別関数を用意して、識別関数の値が最大となる単語を識別結果とする。

このシステムでは、女性を含む 40 名が静かな無響室で発声したデータを用い、35 名分のデータで標準パター

ンを作り、残り 5 名のデータを認識した結果、99.3% の正答率が得られたと発表されている。

一方、ベル研の実験では、15ms のフレーム間隔ごとに自己相関分析した後、線形予測分析 (LPC) を施して、8 次までのパラメーターを求め、あらかじめ記憶されている標準パターンとの間の距離を、LPC パラメーター間の距離尺度を用いて求める。不特定話者の認識を行うには、標準パターンを各単語ごとに複数個用意しておく必要があり、ベル研では 100 人の話し方のパターンをその共通性によってグループ分けを行い、各単語 12 個ずつの標準パターンにまとめている。

入力パターンの識別は、入力パターンと各単語の全標準パターンとの距離を求め、各単語の最も近い k 個 (たとえば 4 個) の標準パターンとの距離の平均の値が最小となる単語を識別結果とする方法を用いている。この方法は k -最近隣法と呼ばれるパターン識別の一手法である。図 7 にベル研における不特定話者認識実験の構成を示す。

この実験では電話機からの入力を可能としており、数字単語の認識で、以前に聞いたことのない人の声で 98.2% の正答率が得られている。また、アルファベット入力による電話番号案内の実験も行っており、姓の最初の 5 文字と名前の頭文字を言って番号簿の名前と一致するまで検索を行い、一致すると対応する番号を回答する実験で 97% の適中率を得ている。認識結果の出力まで約 5 秒を要しているという。

限定単語の認識で特定話者用のシステムでは、標準パターンの登録に時間と労力がかかるため、語彙数をそれほど多くはとれない。また語数の増大に伴う認識率の低下もある程度避けられなくなる。したがって、利用の範囲はこの条件が許容される分野に限定されよう。しかし、信頼度が高い認識が実現されれば、人間にとって最も楽なデータ入力方法となるので、コンピューターを相手としたタイプライターに代わるコマンドや情報入力として利用が拡大されていくであろう。

単語数の制限を解決するには、アルファベットまたは五十音を単位として認識する方法をとることになり、その試みも行われている。しかし、単音だけの認識は、まぎらわしい音たとえば B と D、P と T などの識別がな

図 7 不特定話者単語認識の構成

かなかわずかしいので、より精密な音声分析により高度な特徴を抽出していく必要がある。

不特定話者の認識は、標準パターンを学習することが不要となるので、利用者にとって使いやすい。標準パターンの数が増えると認識結果を得るまでに時間がかかるので、並列処理などにより実用上許される時間内に認識結果を出す必要がある。

3.2 連続会話音声認識

自然言語文の会話音声認識する最終目標は、入力音声と言語符号を1対1に対応させるという音声タイプライターの姿であるが、この場合は認識対象とする語彙数が著しく多くなり、非常にむずかしい。そこで実際にはシステムが入力として扱う対象（これをタスクとよぶ）を、たとえば座席予約やニュースなどの情報検索に限定し、タスクの中で出現する語彙数を数百程度におさえておき、会話的なやりとりを通して発話内容の理解を目標とする研究が、音声理解システムと称して進められている^{21)~23)}。

音声理解システムでは、連続的に発声される音声に対して、①音響分析、②セグメンテーションと音韻認識、③単語認識、④文の理解、という段階で処理がすすめられる。図8に典型的な音声理解システムの構成を示す。

まず音響分析では、たとえば線形予測分析(LPC)を行ってパラメータを抽出し、音韻の標準パターンとの間の距離を計算して、音韻単位でセグメンテーションと音韻認識をして、音韻記号系列に変換する。音韻認識の際、候補がいくつかあって決定が困難な場合には、可能性のある音韻名を併記しておいて、次の単語同定の段階で最も確からしいものを推定することを行う。また、連続音声で発声速度が速い場合などには、音韻は必ずしも正しく認識されないことがあるが、音韻の前後関係を見て正しい音韻に変形するための音韻変形規則も用意して参照することがある。

ここまでに得られた音韻記号系列は発音された音韻名で、いわば発音記号の系列のようなものである。音韻記号の系列から規則的に単語を割り当てることは、発音の

し方が前後の音韻によって変ることや、連続に発声して変形していることが多いため、非常にむずかしい。そこで、可能性のある単語あるいは文節を仮定して予測をし、音韻記号の系列を、可能な全部の変形を含めて用意し、入力記号系列とマッチングをとり、最も確からしい単語を同定する。単語を予測する際に、自然言語としての文法的な制約を表わす構文規則と意味情報およびタスクの場面における制約を示すプラグマティクスを援用して、候補単語の種類をしぼって性能を高めている。

単語同定の段階では、多くの可能な記号系列の探索を行うので、音韻記号をノードとしたネットワークを作り、ネットワークの状態を順次遷移して目標に到達するような制御をすることもある。

最後の段階は、内容の理解で、単語系列から文章を組み立て、内容を理解して結果を出力する。

音声理解システムの研究に用いられたタスクは、数百語を用いた情報検索などの質問に回答する形式のものが多く、表2にカーネギーメロン大学のHARPYというシステムの成果を示す。1011の語彙をもつニュース検索のタスクで性能は、5名の話者の184発話文に対して、文の意味理解率95%、文認識率91%、単語認識率97%であった。現在の段階では、まだ実用に供し得る性能をもつ音声理解システムは実現がむずかしいが、タスクの範囲を小さくし、話者登録をして音響レベルでの精度を上げ、かつ質問に対する応答を用意して、たとえ認識誤りが起こっても再度発音しなおすことを許すなどをしていけば、実時間で動作するシステムが可能であると思われる。

表2 HARPY (カーネギーメロン大)の音声理解システムの成果

対象タスク	ニュース検索
連続音声	可
単語数	1011
認識時間	実時間の80倍
学習訓練	話者ごとに20個の訓練文章
入力	接話マイクロホン

図8 音声理解システムの構成

音声タイプライターのように、発音した音声を文字に変換するには、まぎらわしい音節は単語と組合せて発音するなど、発声側に制約を設け、かつ音韻系列から文字の正書法への変換に精力的に力を注ぐ研究の積み重ねが必要である²⁴⁾。

4. 音 声 合 成

音声合成は、コンピューターなど装置から音声を人工的に作り出すことで、電話機を通して情報の通知や応答が手軽に行える利点に注目して利用が拡大し、研究がかなり進んでいる²⁵⁾²⁶⁾。

音声合成の方法には、3つの方式がある。すなわち、①あらかじめ出力音声の内容を記憶しておき、これをそのままつなぎあわせて出力する録音編集方式、②音声を情報圧縮されたパラメーターの形で記憶しておき、音声合成器を用いて音声を出力するパラメーター合成方式、③規則により制御情報を発生し、任意の音声を合成する規則合成方式である。

録音編集方式は、比較的少ない語彙数の組合せですむ音声の出力用に、すでに多く実用に用いられている。ディスク装置あるいは光学記憶装置などに、文章、文節、単語などを0.5秒程度の長さを1つのユニットとして記憶し、制御部で選択して出力する。現在では、2,000語以上の語彙を768回線に多重出力する大規模なものから、フロッピーディスクを用いて300～1,500語、16～32回線用のものなどの製品化が進み、電話座席予約システムや電話勝馬投票システムなどに用いられている。この方式では出力音声の内容を変更することが厄介であるので、固定的な用途に用いられる。出力する語彙を変更する必要がある用途の場合は、パラメーター編集方式によらなければならない。

パラメーター編集方式は、パラメーターの形で情報圧縮して記憶するので、出力語彙数を飛躍的に増大できる。情報圧縮の方法としては、線形予測パラメーターの形が品質の点ですぐれており、よく使われている²⁷⁾。また、線形予測音声合成器は、最近LSIチップが製造され安価に使用できるようになったので、この方式の普及が広まると予想される。大量の音声情報をファイルにもって

表 3 音声ファイル装置の仕様

音声応答方式	PARCOR 型編集合成方式
音声出力単位	最大3.4秒の節・文など
音声パラメーターの符号化条件	情報容量: 9.6kb/s フレーム周期: 10ms PARCOR 係数: 10段、計81ビット 駆動音源情報: 計15ビット
出力回線数	最大32回線

く音声ファイル装置の例を表3に示す²⁸⁾。また、14,000語、256回線の大容量音声応答装置も実現している。

規則による合成方式は、語彙数に制限なく、入力としてアルファベットまたはカナによるテキストを入れると、連続音声を出力するものである。合成の単位として、線形予測符号化した音韻連鎖を用いる方式が内外で研究されている^{29)～31)}。

音韻連鎖とは、1つの音韻の中央から次の音韻の中央までの音声を構成する基本要素で、子音+母音、語中の安定母音、語のはじめの母音、子音の連続などで2,000個位の音韻連鎖を用いている例がある。音の高さやアクセントや継続時間などの韻律的情報は、与えられた文字列から自動的に作り出す規則を有している。図9に、音韻連鎖による音声合成の系統図を示す。

まず入力した文字系列から、辞書を参照して最適な音韻連鎖の系列に変換する。韻律特徴制御にうつり音韻連鎖の時間を伸縮させ、音韻連鎖のつなぎで振幅とスペクトルの不連続が起こらないようにパラメーターの平滑化を行い、さらにピッチの情報を文章全体で自然に聞こえるように制御する。合成器にはスペクトルパラメーター(たとえば線形予測パラメーター)とピッチと振幅の情報を与えて合成音声を出力する。

この方式の音声品質は、了解性は良いが、音韻連鎖接続に伴う雑音的な不連続感が残るので、品質改善が今後の課題である。音韻連鎖によらないで、すべて規則により連続音声を合成する試みもなされているが³²⁾、細かい音のつながりの部分でまだ品質が充分でなく、今後の研究にまたれている。

図 9 音韻連鎖による音声合成の系統図

5. む す び

音声信号のデジタル処理に関連して、音声高効率符号化と音声認識と音声合成について、最近の研究の動向を述べた。

音声高効率符号化伝送は、今後、デジタル信号処理技術およびLSIの進歩に伴って、経済性のメリットが拡大していく状況にあり、高品質の放送用から、狭帯域まで、ネットワークのデジタル化と歩調を合わせて目的に応じて広い範囲で利用されていくであろう。研究の方向は、品質の向上と、より高効率化に向けられており、集積回路技術が使用できるという背景の下に、アルゴリズムは複雑化しても性能が良い方式が指向されている。

音声認識は、便利なデータ入力方法として、精力的に研究されているが、まだ用途が限定されており、個人性の除去や特徴抽出の精密化など、基礎的研究の積み重ねを必要としている。しかし、発声者が注意深く、1語ずつ区切るように明瞭に発声し、語彙数もさほど多くないという条件の下では、実用に耐える音声認識装置が出来る。特に電話機を通してコンピューターシステムに情報を入力したり、問い合わせをする用途が増えるので、音声合成装置と組合せた質問応答システムとして自動化システムが出現する期待がもたれる。さらに、オフィスシステムにおけるメッセージやデータの音声入力は、潜在需要が大きいので、このような目的にあった音声認識装置が、非専門者が使える簡便な入力装置として製品化される日も遠くないと思われる。

一方、音声合成に関しては、音声応答装置や音声ファイルシステムとして、すでにコンピューターシステムから比較的自然性が高い音声が出力されるようになってきているが、地名や品名など多くの語彙を任意に音声出力するような場合には、文字から規則による合成をする必要があり、音韻連鎖などによる合成法の研究の進展が望まれる。

(昭和54年6月1日受付)

(参 考 文 献)

- 1) J. L. Flanagan, et al.: Speech Coding, IEEE Trans., **COM-27**, 4 (1979) 710-737
- 2) N. S. Jayant(ed.): Waveform Quantization and Coding, IEEE Press, New York (1976)
- 3) J. Makhoul, M. Berouti: Adaptive Noise Spectral Shaping and Entropy Coding in Predictive Coding of Speech, IEEE Trans., **ASSP-27**, 1 (1979) 63-73
- 4) B. S. Atal, M. R. Schroeder: Optimizing Predictive Coders for Minimum Audible Noise, ICASSP (1979) 453-455
- 5) R. E. Crochiere: On the Design of Sub-band Coders for Low-bit-rate Speech Communications, BSTJ, **56** (1977) 747-770
- 6) J. D. Johnston, D. J. Goodman: Digital Transmission of Commentary-Grade(7 kHz) Audio at 56 or 64 kbs, ICASSP (1979) 442-444
- 7) B. S. Atal, S. L. Hanauer: Speech Analysis and Synthesis by Linear Prediction of the Speech Wave, JASA, **50**, 2 (1971) 637-655
- 8) F. Itakura, S. Saito: Analysis-synthesis Telephony Based upon the Maximum-Likelihood Method, Proc. 6th ICA (1970) C17-20
- 9) J. D. Markel, A. H. Gray Jr.: Linear Prediction of Speech, Springer-Verlag, New York (1976)
- 10) B. Gold: Digital Speech Network, Proc. of IEEE, **65**, 12 (1977) 1636-1658
- 11) 板倉, 東倉: 音声の特徴抽出と情報圧縮, 情報処理, **19**, 7 (1978) 644-656
- 12) D. Y. Wong, J. D. Markel: An Intelligibility Evaluation of Several Linear Prediction Vocoder Modifications, IEEE Trans., **ASSP-26**, 5 (1978) 424-435
- 13) A. Kurematsu, et al.: A Linear Predictive Vocoder with New Pitch Extraction and Exciting Source, ICASSP (1979) 69-72
- 14) B. S. Atal: On Synthesizing Natural-Sounding Speech by Linear Prediction, ICASSP (1979) 44-47
- 15) R. D. Preuss: A Frequency Domain Noise Cancelling Preprocessor for Narrowband Speech Communication Systems, ICASSP (1979) 212-215
- 16) T. B. Martin: Practical Application of Voice Input to Machine, Proc. of IEEE, **64**, 4 (1976) 487-501
- 17) D. R. Reddy: Speech Recognition by Machine: A Review, Proc. of IEEE, **64**, 4 (1976) 501-531
- 18) 千葉: 単語音声の認識, 情報処理, **19**, 7 (1978) 667-674
- 19) 千葉, 亘理, 渡辺: 不特定話者を対象とした単語音声認識システム, 昭53電学大全, S5, (1978) 23-27
- 20) L. R. Rabiner, et al.: Speaker Independent Recognition of Isolated Words Using Clustering Techniques, ICASSP (1979) 574-577
- 21) D. H. Klatt: Review of the ARPA Speech Understanding Project, JASA, **62**, 6 (1977) 1345-1366
- 22) 中川: 機械による会話音声の認識, 理解研究の動向, 情報処理, **19**, 7 (1978) 675-685
- 23) R. De Mori: Recent Advances in Automatic Speech Recognition, 4th IJCPR (1979) 106-124
- 24) J. M. Baker: Performance Statistics of the HEAR Acoustic Processor, ICASSP (1979) 262-265
- 25) 市川: 音声の合成と出力, 情報処理, **19**, 7 (1978) 657-666
- 26) L. R. Rabiner, R. W. Schafer: Digital Techniques for Computer Voice Response: Implementations and Applications, Proc. of IEEE, **64**, 4 (1976) 416-433
- 27) 小池ほか: PARCOR 形音声応答装置, 通研実報, **23**, 10 (1974) 2107-2136
- 28) 石井ほか: PARCOR 形画面別音声ファイル装置, 通研実報, **28**, 2 (1979) 175-192
- 29) J. Allen: Synthesis of Speech from Unrestricted Text, Proc. of IEEE, **64**, 4 (1976) 433-442
- 30) R. Schwartz, et al.: Diphone Synthesis for Phonetic Vocoding, ICASSP (1979) 891-894
- 31) 佐藤: PARCOR-VCV 連鎖を用いた音声合成方式, 信学論, **J61-D**, 11 (1978) 858-865
- 32) C. H. Coker, et al.: Automatic Synthesis for Ordinary English Text, IEEE Trans., **AU-21** (1973) 293-297