音 声 合 成

都 木 徹[†]

近年、身近になってきた音声合成技術に関し、その原理、研究・開発および応用の現状、より人間の発声に近い自然な合成方式の実現に向けた研究について、日本語音声合成する場合を中心に解説する。

1. ま え が き

人々は声を用いて様々な情報をやりとりしている。声はそのまま文字列に置き換えた言語としての情報以外に、個性や感情といった非言語情報も含んでおり、多様性に富んだ情報を同時に伝達できる。音声合成技術の最終目標は、これらの多様な情報をすべて、人間の発声と同等の自然性をもって表現可能にすることである。しかし当面は、言語情報をいかに正確かつ聴きやすい音質で伝えられるかが重要な課題である。

音声合成方式には、録音編集、パラメータ編集、規則合成方式などがあり、それぞれ異なる特徴がある。

音声合成の用途は、初期には録音テープの代用といった要素が強かったが、使い勝手や音質が向上した結果、現在では原稿の校正やパソコン通信の読み上げ装置として、あるいは、視覚障害者向けのサービスなどに応用されるようになってきた。音声合成は、基本的には機械から人への情報伝達手段であるが、パソコン通信の例のように人から人への情報伝達手段の一部にもなり、音声合成ならではの応用分野が広がつつある。

2. 音声合成の歴史¹⁾²⁾

人工的に音声を発生させる試みは古くからある。記録が残されている最初のものは1779年で、ロシアのKratzensteinは人の発声器官を模擬した共鳴管を作り、リードを振動させることによって母音を合成し

た。その後1791年にドイツのVon Kempelenによってさらに精巧なモデルが考案され、電気回路のブリッジで有名なWheatstoneはKempelenの記述に基づいて図1のような音声合成器を作った。右肘でふいご(肺に相当)を押して空気を送り、右手の指を使ってリード(声帯に相当)の振動を制御したり、笛から摩擦性の音を出したり、鼻孔を開閉する。同時に左手で革製の筒(口腔に相当)を変形させて様々な共鳴を起こし各種の母音・子音を合成した。しかし、このような機械的な合成器では、良好な合成音声を得るのは困難であった。

一方、1930年代から急速に発展してきた電気回路技術がこの方面にも応用された。これに最初に成功した装置は、おそらくDudleyらによるボーダ(Voder)であろう。肺から声帯に至る、いわば音源に相当する部分はバズ発振器やランダム雑音発生器で、口腔での共鳴は10個の帯域フィルタで、口からの放射はスピーカで置き換えられた。足踏みペダルでバズ発振器の周波数を制御して声の高さを変化させ、10本の指で

図1 Kempelenの記述に基づいてWheatstoneが作った音声合成器(Flanagan¹⁾による)

[†] NHK放送技術研究所 音響聴覚研究部

“Speech Synthesis” by Tohru Takagi (Auditory Science and Acoustics Research Division, NHK Science and Technical Research Laboratories, Tokyo)

フィルタ特性を制御するキーを押し母音・子音の音色を付与し、連続な音声合成した。合成装置単体での操作は熟練を要したが、その後、分析装置と組合せたチャンネルボコーダ (Channel Vocoder) (1939) へと発展し、分析合成技術の基礎を築いた。分析合成とは、実際の音声进行分析し、声帯の振動や口腔での共鳴など音声の性質を特徴づける情報を時系列の形でパラメータ化し、そのパラメータで合成装置を動作させる方式である。送信側に分析装置、受信側に合成装置を置くことで通信にも応用でき、音声の符号化伝送方式の端緒を開くことにもなった。Vocoder とは Voice Coder の略で、現在では分析合成方式全体をさす用語となっている。

その後、1950 年代の終盤から 1960 年代の初頭にかけて、スウェーデンの G. Fant らによって音声生成の音響理論³⁾ が明らかにされ、電気回路による音声合成が容易となり盛んに研究が行われた。さらに、録音技術の発達、コンピュータの出現とこれに伴うデジタル信号処理技術の発達が拍車を掛け、様々な合成手法が提案されて急速に実用化に向かった。なかでも、1960 年代終盤に行われた、板倉・斉藤や Atal・Schroeder らの研究⁴⁾⁵⁾ が発端となった、音声の線形予測符号化 (Linear Predictive Coding: LPC) が果たした役割は大きい。

3. 音声の生成と音声合成器の原理

人間がことばを発声する場合を模式的に考えてみると、図 2 のようになる。まず、何を話すかという概念が生まれ、それが言語(文章)になり、それを声にするために発声器官を操る神経指令が発せられ、最後に発声器官で音響信号となる。この最後の発声器官を機械に置き換えたものを、ここでは音声合成器と呼ぶことにし、以下これについて説明する。

3.1 音声の生成

図 3 は発声器官の断面図である。音声は肺から押し出される空気に種々の器官が作用して作られる。発声器官は声帯 (Vocal Cord) と声道 (Vocal Tract) に大別

図 2 言語発声の模式図

される。声道は喉頭より上の部分全体を指す。

音声は“音”の性質として、有声音 (Voiced Sound) と無声音 (Unvoiced Sound) に区別される。有声音は声帯振動を伴う音声で周期的な振動として観測される。無声音は声帯振動を伴わない音声で非周期的な雑音として観測される。通常の音声では大部分が有声音 (母音、半母音、鼻音、流音など) であり、無声音は無声子音と呼ばれる特殊な子音 (無声摩擦音/s/, 無声破裂音/p/, /t/, /k/ など) のみである。有声音の周期は声帯振動の周期で決まり、これをピッチ周期、その逆数をピッチ周波数または基本周波数と呼ぶ。これらは声の高低やイントネーションを決める重要な要因である。

ある音声を発するための固有の口の構え (例えば、/o/ は唇を丸めて舌を大きく後ろに引く) をすることを調音 (Articulation) という。調音に関係する器官が声道で、唇、舌、あご、口蓋などがある。口蓋帆は一種の弁であり、通常は閉じているが、/m/, /n/ などの鼻音を発声する際には開き気流が鼻腔に送られる。声道全体としては損失のある音響管であり共振特性を持つが、鼻腔が関与した場合には反共振特性も加わる。

有声音の周波数スペクトルは、近似的にはピッチ周波数の高調波からなり、それが声道の共振や反共振特性によってその包絡が変形される。スペクトル包絡の各ピークの周波数が声道の共振周波数で、ホルマント周波数と呼ぶ。周波数の低い方から順に第 1、第 2、…ホルマントといい、母音では第 1 と第 2 ホルマントの組合せでほぼその音響的特徴が区別できる。

一方、無声摩擦音は声帯を振動させず、舌と口蓋との間に狭い透き間を作り、そこを勢い良く気流が流れることによって作られるため、白色雑音に近く、その周波数スペクトルは連続的となる。また破裂音は、唇や舌によって口腔が閉じられ、肺からの呼気がその後

図 3 発声器官の断面図

方にたまり圧力が高まった後、急に閉鎖が開放されることによってインパルス的に発生する。

音声は唇や鼻孔から空間へ放射される際には音波の放射インピーダンス特性が加わるので、音声合成の全過程を簡単にまとめると、①有声または無声音源の生成、②声道の形による調音、③放射の3要素となる。

3.2 音声生成のモデル化

音声合成の際、前節で述べたような音声生成の過程を完全に模擬するのは困難であるので、合成器の構成を簡単にするためのモデル化を行う。Dudleyらのボーダー以来、現在まで用いられてきたのが、線形分離等価回路モデルである。これは音源の特性 $G(\omega)$ と調音（共振・反共振）の特性 $H(\omega)$ を完全に分離して相互作用がないという仮定のもとに、各々の等価回路を接続して音声波形 $S(\omega)$ とするものである。すなわち、

$$S(\omega) = G(\omega) \cdot H(\omega) \quad (1)$$

であり、一般に音源はパルス（有声音源）と白色雑音（無声音源）で近似し、声道における調音は全極型または極零型のフィルタの特性で近似する。これを図4に示す。音源をパルスや白色雑音で近似した場合、 $G(\omega)$ のスペクトル包絡は平坦であるので、音源のスペクトル包絡特性は、放射特性とともに声道フィルタ $H(\omega)$ に含めて考えることになる。

図4 線形分離等価回路

声帯の振動や声道の形は時々刻々変化するので、これに合わせて $G(\omega)$ や $H(\omega)$ の特性も更新しなければならない。ただし、20～40 ms 程度の短時間内では音声の性質は一定であると考えてよく、実際には更新の周期として、5～20 ms を採用する。

3.3 音声合成器の種類

主な合成器の種類を表1⁶⁾に示す。

アナログ合成は、音声の発生機構を電氣的に近似 (Analogue) する方式で、もとなる音声波形が必要ない。実際にはスペクトログラム（声紋の分析装置）や次節で説明する線形予測分析による結果をもとに、合成器に入力するパラメータを決定する。

分析合成では、もとなる音声波形の一定時間内の特徴を少数の特徴パラメータに変換して保存するので、元の音声波形の情報量を圧縮することができる。したがって、波形を直接扱うのに比べ、合成装置単体で用いる場合には話す内容の記憶容量を、符号化伝送の場合には伝送容量を、大幅に節約できるという利点を持つ。

3.4 音声の線形予測

音声信号処理の発展に大きく寄与した、音声の線形予測について説明する。まず、音声の離散的な時系列 $\{s_1, s_2, \dots\}$ を考える。ある時点でのサンプル値 s_n は過去の p 個のデータ $\{s_{n-1}, s_{n-2}, \dots, s_{n-p}\}$ の1次結合（線形結合）によって次のように近似されるとする。

$$\hat{s}_n = \sum_{k=1}^p a_k s_{n-k} \quad (2)$$

ここで、 $\{a_k, k=1, \dots, p\}$ は予測係数、 $\hat{s}_n$ は近似値である。このように現在の値を過去の値の線形結合で推定することから線形予測と呼ばれている。予測係数の決定は、対象とする短時間の音声区間内において(3)式に示す予測誤差 ε_n の2乗平均 $E(\varepsilon_n^2)$ が最小になる

表1 合成器の種類

	分類	合成方法	備考
分析合成	アナログ 声道アナログ ターミナルアナログ	声道の幾何学的特徴を円形音響管により模擬し、電気回路に変換する 数個の変可共振・反共振回路を用いて声道特性を模擬する	ホルマント型ともいう
	チャンネル	帯域フィルタで分割した帯域ごとの利得を調節して声道特性を模擬する	チャンネルボコーダなど
	ケプストラム	ケプストラム係数をパラメータとするディジタルフィルタにより声道特性を模擬	対数パワースペクトルの包絡をパラメータ化
	線形予測に基づく	線形予測係数により制御するディジタルフィルタにより声道特性を模擬する PARCOR LSP PARCOR 係数により制御するディジタルフィルタにより声道特性を模擬する 線スペクトル対周波数をパラメータとするディジタルフィルタにより声道特性を模擬する	最尤推定法と共分散法がある ハードウェア化に向く PARCORより補間特性が良く、圧縮率も高い

図 5 線形予測による声道フィルタ

ようにする。

$$\varepsilon_n = s_n - \hat{s}_n = s_n - \sum_{k=1}^p a_k s_{n-k} \quad (3)$$

この計算方法のアルゴリズムについては種々のものが開発されているが、ここでは省略する。

ここで、(3)式の右辺第2項を移項すると(4)式となり、

$$s_n = \varepsilon_n + \sum_{k=1}^p a_k s_{n-k} \quad (4)$$

これを物理的に解釈すると、音声 s_n は、図5に示すような線形フィルタに ε_n を入力したときの出力で与えられる。 z 変換を用いて、このフィルタの伝達関数を求めると(5)式となる。 $H(z)$ は、対象とした音声が発したときの声道特性を表している。

$$H(z) = 1 / \left(1 - \sum_{k=1}^p a_k z^{-k} \right) \quad (5)$$

ここで、 z^{-k} は“ k サンプルの遅れ”を意味し、標準化周波数を f_s とすれば、

$$z^{-k} = \exp(-jk\omega/f_s) \quad (6)$$

である。

有声区間では、 ε_n は予測が容易な位置では0に近いが、ピッチドライブの位置(声門開放時)では予測が困難であるため値が大きくなる。線形分離等価回路モデルでは、声帯の周波数特性も声道フィルタ $H(z)$ に含めているため、 ε_n は理論的にはピッチ周期に等しいパルス列となる。すなわち、図5のフィルタにパルス列 δ_n を入力すれば、いい換えれば(4)式の ε_n を δ_n に置換すれば近似的に元の音声合成(復元)できる。また、無声区間については δ_n の代わりに白色雑音を入力する。

男声に対するフィルタの次数 p は、 $f_s = 8$ kHz 程度では8~10くらい、 $f_s = 16$ kHz 程度では12~16が適当である。女声や子供ではこれらより2次くらい少ない値が適当とされている。(5)式の分母を0とする z の根 $\{z_k, k=1, \dots, p\}$ が極を与え、 p の値が最適であれば、

$$f_k = f_s \cdot \arg(z_k) / (2\pi) \quad (7)$$

がホルマント周波数を表す。

以上、音声の線形予測の原理を簡単に述べたが、この方法は音声の性質を簡潔に表現できるため、活発に研究され発展した。表1のPARCORやLSPはハードウェア化に向けた優れた方式で、盛んに用いられている。

さて、実際に音声合成器を操るには煩雑な作業が必要で、Kempelenの装置でも、現代のデジタル装置でも変わらない。操作を簡単にするには、図2において発声器官よりも前の部分を機械化する必要がある。次章では、これを実現する音声合成方式について説明する。

4. 音声合成方式²⁾⁶⁾

現在、一般に行われている音声合成方式は、

- ① 録音編集方式：あらかじめ人が発声した音声を波形のまま、あるいは波形符号化して記録し、発声内容に応じて適宜接続して合成。
- ② パラメータ編集方式：分析合成を応用。実音声から特徴量を分析・抽出してパラメータの形で記録し、このパラメータをそのまま、または若干変更して発声内容に応じた時系列とし、合成器を制御。
- ③ 規則合成方式：音素記号列を基に音声学的・言語学的規則によって合成器を制御。さらに発展した、日常の文章(日本語では漢字かな交じり文)からの合成は、テキスト合成(Text to Speech)と呼ばれ、現在の主流。

の3種類に大別される。これらの特徴を表2²⁾にまとめた。技術的には規則合成方式が最も高度であり便利に使えるが、現状では音質や自然さの点で他の2方式が優っており、実際には各方式の特徴を考慮し、使用目的にあった方式を選択することが必要である。

4.1 録音編集方式

最も早くから実用化されており、合成器は用いず、アナログまたはデジタルで蓄えられた音声波形をそのままつなぎあわせる。デジタルの場合ADPCM(Adaptive Differential PCM: 適応差分PCM)などの波形符号化を用いて記憶容量を削減することが多い。

ほぼ元の波形が再生されるので、音質は合成器を用いる方式に比べてかなり良いが、単位と単位との間の聴感的な接続のスムーズさが問題となる。声の高さ、大きさ、発声速度、響きなどの連続性が重要で、これらを考慮しないと、ぎくしゃくした聴き取りにくい音声

表 2 3種類の音声合成方式の特徴の比較

		録音編集方式	パラメータ編集方式	規則合成方式
品質	了解性 自然性	高 高	高 中	中 低
語い数		小 (500 語以下)	大 (数千語)	無限
記録方式または合成器		アナログまたは PCM 録音, ADPCM	チャンネルボコーダ, PARCOR, LSP など	ターミナルアナログ, ケプストラム, PARCOR, LSP など
情報量	伝送容量 1Mbでの 音声長	24~64 kb/s 15~40 s	2.4~9.6 kb/s 100 s~7 min	50~75 b/s 無限
合成単位		単語, 文節, 文章	単語, 文節, 文章	音素, 文節, 形体素など
装 置		簡単	やや複雑	複雑

となる。合成単位を文のように長くすれば問題はないが、記憶容量が膨大になる。逆に、合成単位を音節のように短くとると合成できる語いは原理的に無限となるが、接続をスムーズに行うのは極めて困難となる。

日本語において実用化されているものは単語以上の長い単位を採用し、定型文の中の単語を入れ換えて合成する形式がほとんどである。この場合、スムーズな接続を確保するには、声の高さ等がある範囲を越えないように、各単位の発声の仕方に注意を払う必要がある。さらに厳密に行うには、文中の位置などによって同じ単語であっても声の高さやアクセントが異なることを考慮し、何種類かの異なる発声で録音しておく必要がある。また、ポーズの長さも適当に制御する。

4.2 パラメータ編集方式

録音編集方式の場合と同様に、単語、文節などを単位とするが、分析合成方式を用いるため情報量の圧縮ができ、録音編集方式に比べて合成音声の音質は低下するものの、装置の小型化・低価格化が可能である。

3.2 節で述べたように 5~20 ms の周期で音声进行分析し、ピッチ周波数、振幅、スペクトル包絡を特徴づける係数(線形予測係数、ケプストラム係数等)などのパラメータを時系列の形で記録する。合成時には、適宜接続したパラメータ時系列を用いて、音源パルスの周期や振幅、合成フィルタの特性などを制御する。

この方式のもうひとつの利点は、合成時にピッチやスペクトルのパラメータを時系列として操作できることで、発声速度・振幅やアクセント・文全体のイントネーションといったいわゆる韻律の制御や、接続部でのピッチやスペクトル変化の平滑化などが行える。したがって、短い合成単位での緻密な制御が必要な規則合成方式に応用できる。さらに、音声の生成や知覚のメカニズムを明らかにする研究に寄与したり、音声の変換装置としての応用があり、個人性判断⁹⁾や男声・

女声変換⁹⁾など多くの研究がある。最近では分析合成のこのような応用の仕方を声質変換と呼ぶことが多い。

4.3 規則合成方式

合成単位に音素や音節のような小さいものを採用して任意の文章を合成可能とする。この場合の問題点は、①合成単位間のスペクトルの連続性を保つ方法、②文全体の韻律を与える方法、の2点である(録音編集やパラメータ編集では、少なくとも単語や文節内のスペクトル変化や韻律は自然のものが使えた)。①は合成単位の選び方と密接に関連し、②は入力データとして与えるか、装置内部で機械的に付与する必要がある。したがって、出力音声の品質は使用する合成器の種類のみならず、スペクトルや韻律の制御規則に大きく依存する。

初期の規則合成装置は、入力として発音記号列と各種の韻律情報を与えるもので、その作業は煩しいものであった。装置内部での制御規則が高度化するにつれ、例えば、日本語では仮名文字列と各単語のアクセントなどを入力すればよくなり、さらに近年、日常使用する漢字かな交じり文のみを入力すれば、あとはすべて規則により制御するテキスト合成方式が実用化に至った。

テキスト合成方式の構成要素は、図6に示すように、①文章解析部、②制御情報生成部、③音声合成器である。③は既出なので、①、②について説明する。

(1) 文章解析部⁹⁾

入力文章を辞書を基に解析し、各種の情報を得る。

日本語の漢字かな交じり文は、英語などのように分かち書きされないので、まず単語境界を検出しこれを切り分ける。5~10万語の辞書をメモリーやディスク上に用意し、単語の文法的接続関係を調べながら、文の頭から順々に辞書単語と照合をとり、単語の系列

図 6 テキスト合成方式の構成

図 7 文のピッチパターンの生成

に区分化する。辞書には読み仮名、文法情報、アクセントなどが蓄えられており、これらの情報も付加される。しかし、日本語特有の難しさもある。以下に幾つか例をあげるが、これらは規則で制御可能なものも多いが、例外も存在し、解析アルゴリズムを複雑にしている。

(a) 複数の区分化ができる：

「米国産業界」は「米国/産業・界」とも「米/国産/業界」とも区分できる。

(b) 表記が同じでも読み仮名が異なる：

「最高値」は<さいこうち>とも<さいたかね>ともなる。

「1本」は<いっぽん>、「2本」は<にほん>となる。

(c) 読み仮名と実際の発音が異なる：

長音化「空港<くうこう>」は/クーコー/と読む。
鼻音化「学校が…」の後の/ガ/は鼻濁音となる。
無声化(調音は母音のままで声帯は振動しない)
「北です」の<き>の/i/や<す>の/u/は無声化する。

(d) 同じ単語でもアクセントが変化する：

高く発音する部分に「」をつけると、「熱帯」は/ネッター/, 「亜熱帯」は/アネッター/, 「熱帯圏」は/ネッターイケン/。

(2) 制御情報生成部

入力情報と規則により、合成器を制御するスペクトルや韻律のパラメータを時系列の形で出力する。

発音記号列に対応させ、あらかじめ実音声から分析されて蓄えられている合成単位を接続していく。韻律は入力情報と規則により決まるので、原理的にはスペクトルに関するパラメータのみを蓄えておけばよい。

実際の合成単位としては日本語の場合、仮名文字に対応する 100 音節 (CV 単位ともいう；C:consonant 子音, V:vowel 母音) が用いられることが多い。より良好な合成音を得るため CVC を単位とする方式¹⁰⁾も検討されているが、組合せが膨大なため高頻度の約 1000 種類の CVC と CV, VC を併用する。VCV 単位を用いる方式¹¹⁾もあり、この場合は単位数は 700~800 となる。例えば、/サクラ/という単語の合成では、CV 単位: /sa/+/ku/+/ra/, CVC 単位: /sak/+/kur/+/ra/, VCV 単位: /sa/+/aku/+/ura/となる。CVC 単位の場合には子音の部分で、VCV 単位の場合には母音定常部で、各単位の接続が行われることになり、それぞれ異なる特徴がある。

韻律情報のうち最も重要なのはピッチ周波数である。日本語の単語・文節のアクセントには感覚的に高低の 2 レベルがある。その物理的主要原因はピッチ周波数で、モーラ(拍: 仮名文字単位に相当)ごとに高低どちらかのレベルを与えれば充分である。また、一息で発声する区間すなわち呼吸段落において、ピッチ周波数は声の始めでは高く次第に低くなる。この特性を話調成分またはフレーズ成分と呼び、その上に単語・文節固有のアクセント成分が重畳し、文全体のピッチ周波数パターンが定まる。この様子を図 7²⁾に示す。

表 3 製品化（含予定）されている日本語テキスト合成装置

メーカー 形式名	三洋電気 VSS-300	NTT しゃべりん坊	マイクロニクス 音次郎/Pro	大井電気 TALK-J100/200	富士通 FMVS-101	明電舎 MEITALK	リコー VC-1(N)	キャノン —
声の種類 話速制御 音量制御	男 プログラム プログラム	男・女 プログラム プログラム	男・女 プログラム つまみ	男・女 プログラム プログラム	男・女 プログラム つまみ	男・女 つまみ つまみ	男 つまみ つまみ	男・女 プログラム プログラム
インタフェース 形態	セントロ 箱型	バス直結 ボード	セントロ ラッコ人形/箱型	RS-232C/バス直結 箱型/ボード	セントロ 箱型	RS-232C 箱型	RS-232C 箱型	RS-232C/バス直結 箱型/ボード
合成器 合成単位	LSP CV, V	LSP CV, VC	PARCOR —	ホルマント 音素	PARCOR CV, V	声道アナログ CV(V,C,V)	LSP CV	ケプストラム CV, V

（注）合成音声がどのようなものか、テレホンサービスで聞ける 03-3235-7252

フレーズ成分にインパルス状、アクセント成分にステップ状の指令を対応させ、準連続的なピッチパターンを生成するモデル¹²⁾も提案され実用に供されている。

また各音素の継続時間の制御規則も合成音声に自然さを与える上で重要で、検討¹³⁾が行われている。

4.4 現時点での音声合成装置の応用分野¹⁴⁾

録音編集方式による装置はすでに、電話サービスの音声応答、駅のアナウンスなど、広く実用化されている。従来、装置が大規模になる傾向があったが、ADPCM等の波形符号化の応用と記憶装置の小型・低価格化により、個別ユーザ向けの商品も開発されている。

パラメータ編集方式では、合成 LSI の開発により、小型・安価という利点を活かし、時計や電子レンジなどの家電製品に多く応用されている。

規則合成、特にテキスト合成装置の応用分野は、現在その市場性と共に注目を集めている。最近製品化されている日本語テキスト合成装置を表 3⁹⁾に示した。これらの特徴は、パソコンとの接続が可能で使い勝手が良く、数万～35 万円という低価格であるという点で、一般のオフィスや個人ユーザでも充分使用可能なものとなっている。膨大な語いが合成でき、語い追加に際しても辞書登録を行うだけでよく、録音編集やパラメータ編集のように同じ発声者を確保する苦勞がない。

新聞の校閲支援システム、電話を用いた各種自動応答サービス、ワープロ文書や電子メールの読み上げ、監視制御の警報、車の運転など目が離せない人への情報伝達、OCR と組合せた読書器などへの応用が考えられる。また、発声器官や視覚に障害がある人の口や目の代替装置としての応用もすでに行われている⁹⁾。

現状でのテキスト合成装置は、聴き取りにくさや誤った読みやアクセントがないわけではないが、その情

報を必要とする人に伝えるには十分な品質になってきており、効果的な応用分野の拡大が期待される。

5. 将来の音声合成方式に向けて

現在の音声合成方式の品質を向上させる努力が様々な角度から続けられている一方、将来へ向けての新しい音声合成の研究も進められている。

5.1 より高度な音声合成

(1) 概念表現からの音声合成¹⁵⁾

図 2 のすべてを機械化する方式である。エキスパートシステムなどの知識ベースから概念表現を生成し、それにより音声合成する。より自然に韻律を制御するための情報(係受け関係や強調される語句等)などは、テキスト合成の文章解析部では正確に抽出するのが難しいが、概念表現として記述できれば面倒な構文解析は必要なくなる。

(2) 音声による自動翻訳への応用

英語音声認識し、日本語文章へ機械翻訳し、日本語合成音声を出力する。あるいはその逆を行う。このようなシステムの音声合成部の機能としては、明瞭さはもちろん、元の話者の個性や感情なども表現できることが望ましい。ATRをはじめ各研究機関で、これらの要素技術の研究¹⁶⁾¹⁷⁾が始まっている。

5.2 合成方式（合成器）の高品質化

線形分離等価回路モデルでは、音源をパルスや白色雑音で近似(例えば、(4)式の ϵ_n を δ_n で)し、情報量の圧縮をはかる反面、自然さの情報を欠落させ合成音声の音質を悪くしていた。これは、スペクトルパラメータの抽出が不完全なことに起因しており、これを補完する特性を音源に持たせたマルチパルス方式¹⁸⁾や、(4)式の ϵ_n を部分的に保存して用いる方式¹⁹⁾が提案されている。また、これらとは異なる観点から高品質化を目指した新方式も各種^{20)~23)}提案されている。

当 NHK 放送技術研究所においても、放送で使用

図 8 NHK 技研で開発した声質変換方式

可能な高品質な合成方式²⁴⁾(声質変換方式)の開発を行っている。この方式では情報量の圧縮は考えず、自然さの情報を極力保存する立場をとっており、録音編集とパラメータ編集の長所を併せ持つ。無声区間は元の波形をそのまま使用するが、有声区間のピッチ周波数を高品質に制御する点に工夫があり、その原理を図8に示す。

まず、分析部で入力音声の有声区間に対して各ピッチ区間を定め、ピッチ区間毎に線形予測分析によるスペクトル包絡 $H(\omega)$ を抽出する。

次にピッチ制御部において、韻律に変化を与えるため、入力波形に対し時間軸上でピッチ区間長を変更する。

ピッチ制御部の出力波形に対し、再びスペクトル包絡 $H'(\omega)$ を求めると、この包絡はピッチ変更の影響で変化しており、入力音声のスペクトル包絡 $H(\omega)$ とは異なる。この2つのスペクトル包絡の差をピッチ変更によるスペクトル的ひずみと考える。

スペクトル制御部では、上記のピッチ変更によるスペクトル的ひずみの修正を行うと共に、個性や明瞭さを制御するため、入力スペクトル包絡に現れている極特性に変更を与える。これらのスペクトル包絡の変更や修正は、FFT パワースペクトルの各周波数成分の絶対値を制御することによって行う。

この方法により、元の音声のピッチ周波数を ± 1 オクターブ程度変化させてもかなり自然な音質が保てる。今後、放送の音声を各個人が聴きやすい話し方に変換する装置や規則合成方式などに応用していきたい。

5.3 声質変換による音声の感情制御²⁵⁾

5.2 節で述べた合成(声質変換)方式を用いて、感情表現に寄与する物理パラメータを調べた実験の一部を紹介する。女性が無感情に発声した「あーそうですか」という音声の、①スペクトル包絡、②時間構造(各母音や無音区間の時間長)、③ピッチ周波数(振幅も同時に制御)の3種の物理パラメータを、本人または他人

表 4 無感情に発声された音声の物理パラメータを“悲哀”のものに変換した音声に対する“悲哀”と判定された割合
S: スペクトル包絡 T: 時間構造 P: ピッチと振幅
T': 他人の時間構造 P': 他人のピッチと振幅

変換したパラメータ	S	T	P	S+T	S+P	T+P	S+T+P	S+T'+P'
判定率(%)	29	19	55	50	94	91	100	93

が同一文を感情を伴って発声した音声のものに変換した。“悲哀”の感情についての、主観評価結果を表4に示す。数字は、評定者が合成音を聴いて、“悲哀”の感情であると判断した割合(%)である。

ピッチ周波数単独の変換では半分程度の効果があるが、スペクトル包絡や時間構造の単独変換ではあまり効果がない。ピッチ周波数と時間構造の組合せ(韻律)ではかなり効果があり、すべてを組合せると完全となり、各パラメータの相乗効果がある。また、韻律については他人のものに変換しても十分な効果がある。規則合成方式に応用するには、今後、各パラメータの制御規則について検討していかなければならない。

6. む す び

音声合成技術について、その歴史から将来の展望まで、かいつまんで解説したが、音声の分析、生成、知覚、デジタル信号処理、音響、言語、通信、集積回路など広い分野にわたる技術であることがおわかり頂けたかと思う。今後も多くの人々の努力によって、さらに生活に身近な技術に発展していくことを期待したい。

本解説を書くにあたり、工学院大学教授 斎藤収三博士、東京農工大学教授 中田和男博士、NTT 古井貞照博士らの著書²⁶⁾を多くの部分で参考にさせて頂いた。音声信号処理の入門には最適の書であり、音声合成に興味を持たれた読者は一読をお奨めする。

(1991年11月8日受付)

[参 考 文 献]

- 1) J. L. Flanagan: "Speech Analysis Synthesis and Perception", 2nd. Edition, pp. 204-232, Springer-Verlag (1972)
- 2) 古井: "ディジタル音声処理", 東海大学出版会(1985)
- 3) G. Fant: "Acoustic Theory of Speech Production", Mouton & Co. (1960)
- 4) 板倉, 斎藤: "最尤スペクトル推定法による音声分解合成伝送方式", 音響学会講演論文集, p. 231 (Nov. 1967)
- 5) B. S. Atal, M. R. Schroeder: "Predictive Coding of Speech Signals", Reports of 6th Int. Conf. Acoust., C-5-4 (1968)
- 6) 斎藤, 中田: "音声情報処理の基礎", オーム社(1981)
- 7) 桑原, 大串: "ホルマント周波数・バンド幅の独立制御と個性判断", 信学論, **J69-A**, 4, pp. 509-517 (Apr. 1986)
- 8) 安広, 尾関: "男声・女声変換実験", 音響誌, **32**, 6, pp. 362-368 (June 1976)
- 9) 国分: "もう一つのメディア音声", 情報の科学と技術, **41**, 1, pp. 27-35 (Jan. 1991)
- 10) 佐藤: "CVCと音源要素に基づく(SYMPLE)音声合成", 音響学会音声研資, S83-69 (Jan. 1984)
- 11) 佐藤: "PARCOR-VCV 音韻連鎖を用いた音声合成方式", 信学論, **J61-D**, 11, pp. 858-865 (Nov. 1978)
- 12) H. Fujisaki, K. Hirose: "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese", J. Acoust. Soc. Japan, **5**, 4, pp. 233-242 (Oct. 1984)
- 13) 匂坂, 東倉: "規則による音声合成のための音韻時間長制御", 信学論, **J67-A**, 7, pp. 629-636 (July 1984)
- 14) 中津: "音声認識・合成技術の製品化および市場動向", 信学技報, SP89-103 (Jan. 1990)
- 15) 山下, 水谷, 溝口: "合成音出力における概念表現の利用", 信学技報, SP89-115 (Jan. 1990)
- 16) 阿部: "言語間にわたる声質変換", 信学技報, SP89-123 (Feb. 1990)
- 17) 北原, 東倉: "音声の韻律情報と感情表現", 信学技報, SP88-158 (Mar. 1989)
- 18) 小澤, 小野, 荒関: "マルチパルス駆動形音声符号化法の品質改善", 音響学会音声研資, S83-78 (Jan. 1984)
- 19) 広川: "残差を用いた規則合成用駆動音源", 音響学会講演論文集, pp. 165-166 (Mar. 1986)
- 20) 中島, 鈴木: "音声スペクトルの微細構造に表われる非定常状態情報について—新しいPSE分析法による検討", 信学技報, SP87-44 (July 1987)
- 21) 佐藤, 橋本, 白井: "ピッチに同期する正弦波発振器による音声合成", 信学技報, SP89-122 (Feb. 1990)
- 22) 広川: "波形辞書を用いた規則合成法", 信学技報, SP88-9, pp. 65-72 (May 1988)
- 23) 阿部, 田村, 桑原: "FFT スペクトルからの Signal Reconstruction による音声変換手法", 信学技報, SP88-48, pp. 25-32 (Sep. 1988)
- 24) 都木, 梅田: "ピッチ変更時のひずみをスペクトル領域で修正する声質変換方式とその品質の心理評価", 信学論, **J73-A**, 3, pp. 387-396 (Mar. 1990)
- 25) 都木, 桑原, 梅田: "声質変換による, 感情表現制御のための物理的要因の解析", 音響学会講演論文集, pp. 275-276 (Oct. 1989)

たかぎ とおる
都木 徹 昭和56年, 電気通信大学通信工学専攻修士課程修了。同年, NHKに入局。昭和59年より, 同放送技術研究所音響聴覚研究部に勤務。声質変換, 音声知覚, 音声合成などの研究に従事。