Vol. 10 No. 46　　電波研究所季報　　January 1964
PP.27－40

調　　査

UDC 534.78

音声合成の新しい発展

情報処理研究室

内　容　目　次

ま　え　が　き

音声の研究における“合成”の重要さはつとに認められてきたところであり，当研究所においてもすでに研究，実験を行なってきた。しかし最近になって音声の研究手段としての合成の重要さはとみに増大し，analysis-by-synthesis（合成による分析法）の手法に示されるような active な研究手段として盛んに研究されるようになった。音声合成研究の重要さは，ただ単に以上のような研究的な面のみでなく，最近では言語オートマント(language automation)の一つとして実用的な重要さも認められてきた。たとえば不連続な入力（文字とか音韻記号とか）から連続的で了解度と自然性の高い音声をほぼ real time で合成できるとすれば，問い合せや案内業務のような“information retriervl” service において retrieave された情報を音声の形で表示することができ，人間への情報再生を非常に能率的にまた自然な形で行なうことができることとなる。

このような音声合成の新しい発展の背景には次のような理由が考えられる。

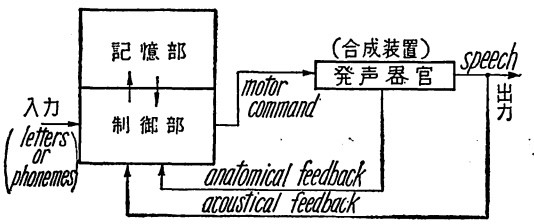
(1) G. Fant らによる音声発生理論が原理的に確立され，音響的なまたは発声運動的な段階で，operational な音声発生（合成）のモデルを作ることができるようになった。

(2) 従来の合成音声の聞き取り実験結果の解析から，音韻知覚過程のモデルとして articulatory reference theory が提唱され，さらに analysis-by-synthesis の手法が音声研究における active な研究方法として原理的に確立され，一部実験的にその有効性が実証された。

(3) ディジタルおよびアナグロ計算機の発達，実用化によって，能率的で適用性の広い音声合成装置またはその適切な制御ができるようになり，連続的に音声を合成しうる可能性がみとめられてきた。

このように研究的にまた実用的に重要性を増し，新しい発展段階をむかえた音声の合成について，現在の研究状況を総合的，系統的に展望し，その問題点を明らかにするとともに，今後のわれわれの音声合成の研究の方向を search するために調査を行なったので，その概要をここに報告する。

人間の発声機構を工学的な音声合成装置との対応に重点をおいて考えると第1図のように考えられる。このような過程を工学的に実現する場合に，記憶装置に重点を



第1図 人間の発声機構の工学的な説明

おき，合成の過程を“compilation”(編集) のみとしてしまったのが，すでに録音されている音声セグナントからの連続的な音声の合成(speech synthesis from stored segments or compiled speech)であり，合成制御装置に重点をおいてパラメターの連続的な制御という形で合成するのが合成法則による連続音声の合成 (speech synthesis by rules)である。さらにその中で合成装置を人間の発声機構のアナグロ回路とし，音声の構造的な情報を合成装置に内蔵させたものが，いわゆる模擬音声合成 (analog speech synthesis) である。

この調査報告では第1部にパラメータ制御による合成の諸問題を，第2部に録音された音声セグメントからの合成の諸問題を論ずることにする。

なお音声合成の基本的な原理に関しては次の諸論文を参照されたい。

(1) G.Fant ; “Acoustic Theory of Speech Production,” Mouton Co., 1960.

(2) E.E.David, Jr. ; “Signal Theory in Speech Transmission,” IRE. Trans., CT-3, 4, 232-244(1956).

(3) G.E.Peterson and E.Sivertsen ; “Studies on Speech Synthesis,” Rep. No.5, Speech Res. Lab., Univ. Michigan (1960).

(4) 中田和男 ; “日本語音声の合成的研究,” 電波研季報臨時特集, No.4 (1961).

第1部 パラメタ制御による音声の合成

(Speech Synthesis by Rules)

中田和男　　光岡輝義※

緒言

パラメータ制御による音声の合成法にはいろいろの原理のものがあるが大きく分けて次の三つに分類されよう。(1)ボコーダ型,[1] (2)基本信号型,[2] (3)アナグロ型。[3] 不連続な入力 (たとえば文字とか音韻記号とか) の系列から連続的な音声を合成するという点からみて，また音声の研究用という点からみても，音声の構造的な情報の多くが合成装置自体の構成に内蔵されているアナログ型のものが最も適しており，したがって最もよく研究されているので，ここではアナログ型の音声合成について論じる。**さらに不連続な入力の系列から連続的な音声を合成するためには，直接音声の合成に，あるいは間接に音声合成装置の制御に，計算機を利用する合成法が能率的でまた適用範囲の広い方法と考えられるので，ここでは計算機の使用法によって音声の合成法を分類し，その研究の現状と問題点を明らかにし，今後の研究の参考とした。

計算機を利用した音声合成法は原理的に次のように分類される (アナグロ型の合成法についてのみ考える)。

(1) Computer Synthesis (計算機を音声合成装置simulatorとして用いる方法)

(1.1) Vocal Tract Analog型

(1.2) Terminal Analog型

(1.21) 時間領域での計算法

(a) Convolution積分による方法

(b) 微分方程式による方法

(1.22) 周波数領域での計算法

(2) Computer Control (計算機を音声合成装置の制御に用いる方法)

(2.1) Vocal Tract Analog型

(2.2) Terminal Analog型

以下これらの各方法についてその研究，実験の現状と問題点，およびその検討について論ずる。

＊東京電気大学大学院

＊＊Haskins研究所の“Pattern Play Back”はボコーダ型であるが，この装置による音声の合成は例外的に非常によく研究されている。

1. Computer Synthesis

現在までのところ主としてベル電話研究所において研究，実験されているが,[4] わが国においても電気試験所の猪股氏らによる初期の研究[5]があり，最近電々公社通研の橋本氏らによっても研究，実験されている。[6] またやや異色のもとしては東大のRAAG Phonetical Research Groupによるアナログ計算を用いた合成実験がある。

1.1. Vocal Tract Analog 型

J.L.Kelly らによってベル電話研究所において研究，実験されている方法であり，分布常数的な声道(vocal tract)の音響的な特性をちえん線と反射係数で模擬したものであるが，他の computer synthesis の方法にくらべ計算機での処理に適した方法として注目すべきものと思う。

(1) 合成法の概要

分布常数的な声道の音響特性を次のようなちえん線の連結として近似する。

（1）vocal tract を区間長 $l=0.8$ cm の21区間に分割する。(全長を16.8cmに固定する。)

（2）各区間を $l/c \approx 1/40$(m.sec)のちえん線とする。(c は音速，320 m/sec と仮定)

（3）各区間の特性インピーダンス Z_0 は $\sqrt{L/C}$ であり，かつ $L=\rho/A$, $C=A/\rho c^2$ (A は区間の断面積，ρ は空気の密度，c は音速) であるから，$Z_0=\sqrt{\frac{\rho}{A}\cdot\frac{\rho c^2}{A}}=\frac{\rho c}{A}$ となり，その断面積に逆比例する。

（4）各区間のつなぎ目をつぎのような反射係数で結ぶ。

$$\alpha=\frac{Z_0^2-Z_0'}{Z_0^2+Z_0'}=\frac{1/A_2-1/A_1}{1/A_2+1/A_1}=\frac{A_1-A_2}{A_1+A_2}$$

したがって各区間を第2図に示すような計算過程でおきかえて処理する。実際には第2図を第3図のように変換し，$1/2\tau=20$ kc として出力よみ出しの sampling 周波数と合わせて処理している。細部の処理として，

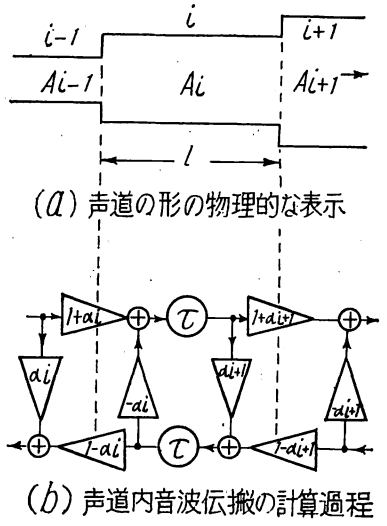
(1) Damping: 反射波(逆方向の音波)に対して適当な減衰をあたえることで dampingの効果をもたせる。

(2) Terminal impedance: glottisにおける反射係数は断面積0.2cm²の source impedanceとして計算する。lipsにおけるradiation impedanceは，半無限空間と考えて $A_l\to\infty$ とすると0となって $\alpha=-1$(完全反射)となり不合理なので適当な α の値を仮定する。

(3) Nasal tract: oral tractの単位区間長の3倍の区間長と，より大きな dampingを有する4区間で近似する。

(4) Voice excitation: 有声音源および aspiration は glottisに，無声音源は vocal tract configuration の greatest constrictionの点に加えられる。

(5) Control method: configurationは 10 samples の間(0.5m.sec) 一定に保たれ，次にこの8区間すなわち4 m.secにわたって excitation parameters と reflection coefficient を linearに interpolateしながら変え，4 m.secごとに新しい area informationをよみこむ。

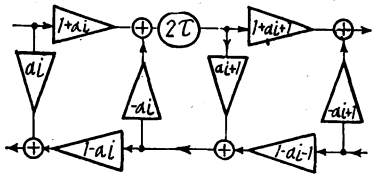


第2図 声道の形からの音声波形の計算法 (A system)

$$a_{i-1}=\frac{Z_0^i-Z_0^{i-1}}{Z_0^i+Z_0^{i-1}}=\frac{1/A_i-1/A_{i-1}}{1/A_i-1/A_{i-1}}=\frac{A_{i-1}-A_i}{A_{i-1}+A_i}$$

区間のつなぎ目における反射係数(A_i は i 番目の区間の等価断面積)

$\tau=l/c$ ちえん時間(c は音速，l は区間長)



第3図 第2図の等価変形計算法 (B system)

(2) 制御法の概要

上述の vocal tractを制御して連続的な音声を合成するためには次の26個のパラメターの値を時間の関数として与えなければならない。

vocal tract 各区間の等価断面積	21個
nasal coupling(鼻音化の程度)	1個
有声音源の強度とピッチ周波数	2個
気音源(aspiration)の強度	1個
無声音源(affrication)の強度	1個

不連続入力（この場合には音韻記号）からこのような制御出力を作り出す programを signal generatorといっている。不連続入力のカードは6個のパラメターを記入できるようになっており，第1は音韻の名称，第2は母音の stress，第3は loudness，第4は pitch，第5

は transition time, 第6は duration であるが, 第3以下の四つのパラメターは空白であってもprogramによってある法則にのっとって制御される。

制御の第1原則はすべてのパラメターはその旧の値から新しい値に向って時間的に直線的に変化するよう制御される。新しい値は入力の音韻記号に対応して stored tablesの中からえらび出される。そして入力パラメターまたは stored tables から指定される時間の間その状態をつづけ, それからまた次の状態（音韻）へと移ってゆく。

これにいくつかの経験的な例外法則(ad-hoc exception rules) を加えて実際の合成は行なわれる。

(3) 問題点およびその検討

(1) 合成法について

(a) 第2図(A systemとする) と第3図(B systemとする) の比較：音源が加えられてから最初に出力が得られるまでの時間おくれが2倍になる以外は原理的には等価である。ベルでは delay time と出力よみ出しの sample周期とを一致させるために B system をとった。計算のサンプル周期が声道の形の時間的な変化の速さにくらべてじゅうぶん速いときは, 両 system の計算時間はほぼ同じとなる。

(b) tractの長さを一定としたこと：Fant の実測によれば, ロシヤ語の母音の場合で, 19.5cm (/u/) から16.5 cm (/i/) まで変化しており, 母音の場合でも, radiation impedanceと関連して, 唇での terminationにもうひと工夫必要であると思われる。(子音についてはデーター不足)

(c) damping の与え方：区間長の長さが短いか, 減衰係数が小さい場合には逆方向の波のみに減衰を与える方法と, 両方向の波に減衰を考える方法との差は小さい（その比は $e^{\Gamma l}-qe^{-\Gamma l}/1-qe^{-\Gamma l}l$ は区間長, Γ が減衰係数, qは反射係数である)。

(d) back coupling: glottisの source impedanceを断面積0.2cm²に対応させたのは平均的に考えて無難と思われるが, さらによい近似として back couplingの時間的な変化ということを考えてもよいであろう。

(e) radiation impedanceの問題：radiation impedanceのみに限らず dampingにしても, 時間領域での取扱いなので, 周波数特性を簡単にもちこむことができないのがこの計算法の一つの欠点である。

(2) 制御法について

(a) パラメター値の直線的な変化：一つの音韻から次の音韻へ連続的に移る場合, 各制御のパラメター（主として調音パラメター）の値をその二つの音韻間で時間とともに直線的に変わるとしているが, この仮定には明確な生理音響的なまたは物理音響的な根拠はなく, 計算の便宜のためと考えられる。この点発声器管の動きの段階での analysis— by— synthesisの分析結果や, 新しいX線観測データにもとづいてより実際に近い動きをさせることが必要である。

1.2. Terminal Analog型

現在までのところわが国において実験された音声の computer synthesisはすべてこの型のものである。この型のものはその計算法によって時間領域での合成と周波数領域での合成に分けられる。以下その合成法について簡単にのべる。

(1) 時間領域での合成法

(1) convolution積分による方法

Vocal tractの impulse responseを $h(t)$, 励振音源波形を $g(t)$とすれば, 出力としての音声波形 $v(t)$は次のような convolution積分を計算することによってえられる。

$$v(t)=\int_0^t g(t)h(t-\tau)dt=\int_0^t g(t-\tau)h(t)d\tau \qquad (1-1)$$

vocal tractの impulse response はその伝達関数 $Z(p)$の Laplace変換として次式によって求める。

$$h(t)=\frac{1}{2\pi j}\oint Z(p)e^{pt}dp \qquad (1-2)$$

母音型の音声については

$$Z(p)=\prod_{i=1}^{n}\frac{p_i\cdot p_i^*}{(p-p_i)(p-p_i^*)} \qquad (1-3)$$

($p_i=\sigma_i+j\omega_i$, $p_i^*=\sigma_i-j\omega_i$でω_i, σ_iは i 番目のホルマントの周波数とその帯域幅であるから,

$$h(t)=\sum_{i=1}^{n}A_ie^{\sigma_i t}\sin(\omega_i t+\phi_i) \qquad (1-4)$$

となる。

猪股,[5] および橋本[6]の実験はいずれも$h(t)$として（1−4）式を使っており, 音源波形$g(t)$として猪股は三角波を, 橋本は気管外壁からとった有声音源波形を用いている。

(2) 微分方程式による方法

音声発生過程は微分方程式によっても記述される。微分方程式の形としては, アナログ型音声合成装置から考

えてもわかるように常微分方程式と考えてじゅうぶんである。いま vocal tract の伝達関数に三つの polesと一つの zeroを仮定とすると

$$a_6\frac{d^6Y}{dt^6}+a_5\frac{d^5Y}{dt^5}+\cdots\cdots+a_1\frac{dY}{dt}+a_0Y=\frac{d}{dt}X(t) \qquad (1-5)$$

ここで$X(t)$は励振音源波形であり，解としての$Y(t)$が出力音源波形を与える。

計算算機プログラムの技術としては上式を連立の一階微分方程式になおして，適当な初期条件を与えて解けばよい。しかし連続音声を発生させるには物理音響的なまたは生理音響的な制御パラメターから係数a_iを求める過程が複雑になる。計算法としてはアナログ計算機による処理に適している。

(2) 周波数領域での合成法

(1) ディジタル計算機による方法

定常的な音声に対しては，vocal tract の周波数特性を$H(\omega)$，励振音源の周波数スペクトルを$G(\omega)$とすれば時間領域での comolution積分法に対応して次の周波数領域での Fourier変換による合成式がえられる。

$$u(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}G(\omega)\cdot H(\omega)e^{j\omega t}d\omega \qquad (1-6)$$

しかしこの方法は時間的に変化する連続音声の合成にはあまり便利でなく，またFourier変換のためにはsin x cos xのサブルーチンを使わなければならず，そのため計算所要時間が長くなる欠点をもっている。

(3) アナログ計算機による合成法

vocal tractの伝達関数は Fant によって求められているように次のようにあらわされる

直列型
$$Z(p)=Aop\cdot\frac{\prod_{i=1}^{n}(p-p_i)(p-p_i^*)}{\prod_{k=1}^{m}(p-p_k)(p-p_k^*)}$$

$$p_k=-\sigma_k+j\omega_k \qquad (1-7)$$

並列型
$$Z(p)=p\sum_{k=}^{m}\frac{A_k}{(p-p_k)(p-p_k^*)}$$

$$p_k^*=-\sigma_k-j\omega_k \qquad (1-8)$$

いずれにしてもその基本をなす共振特性は次のようにあらわされる。

$$H_k(p)=\frac{A_k\cdot p}{p^2+2\zeta\omega_k p+\omega_k^2(1+\rho^2)} \qquad (1-8)$$

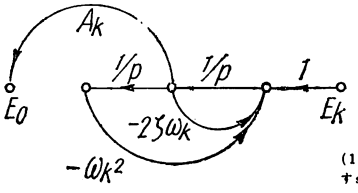
ここで$\zeta\omega_k=-\sigma_k$がこの共振の dampingを与える。(1－8) 式は次のようにかきかえられる。

$$E_k\approx\frac{1}{A_k}\cdot\frac{(p^2+2\zeta\omega_k p+\omega_k^2)}{p}E_0,\ \rho^2\ll1 \qquad (1-9)$$

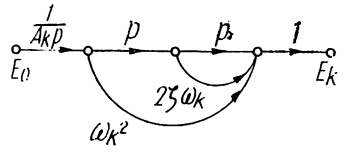
ここで(1－9)式は第4図のような signal flow-graph であらわされるからその逆関係をあらわす(1－8)式は第5図のようなflow-graphであらわされることになり，この関係は第6図のようなアナログ計算機の結線であらわされることになる。第6図で$A=A_k$でk番目のホルマントの振幅を，$B=\omega_k^2$でその周波数を，$C=2\zeta\omega_k=-2\sigma_k$がその帯域幅(damping) を与えることになる。

東大の PAAG Phonetical Research Groupではこのような原理によってアナログ計算機による第7図のような結線（並列型）で母音の合成を実験している。

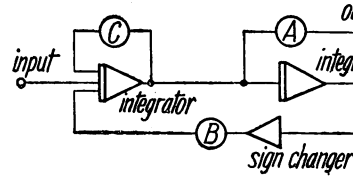
最後に computer synthesisの計算所要時間を比較してみると第1表のようになる。



(1－9)式の関係をあらわすsignal flow graph



第4図の逆で(1－8)式の関係をあらわす signal flow graph

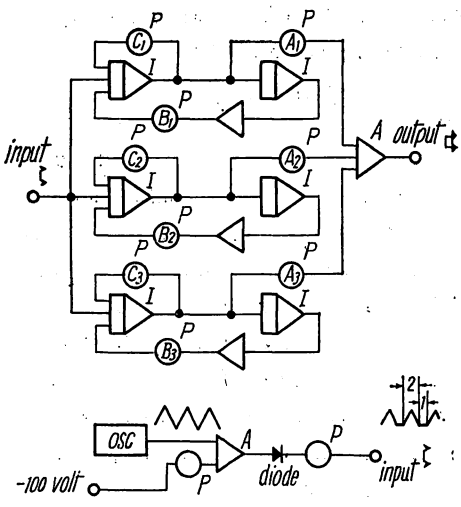


第6図 第5図のsignal flow graphをあらわすアナログ計算機結線

1.3. Vocal Tract Analog型(V型)とTerminal Analog型(T型)の比較

(1) 制御パラメターについて

V型では主要な制御パラメターがいわゆるanatomical parameters であり，articulation(発声運動) との対応が明確であるが，vocal tract configurationの適当なparameter表示を用いない限り，制御パラメターの数が多すぎ，またX線による観測データも少ないので合成のルールに現在まだ不明の点が多い。



A : adder, I : integrator, O : operational
P : potentiometer amplifier.
第7図 アナログ計算機による音声合成方式
(母音について，並列型)

第1表 Computer Synthesisによる音声合成の time scaleの比較

主要研究者	合成法	使用計算機	time scale
J. L. Kelly	Vocal Tract Analog型	IBM7090	40 : 1
猪股	Terminal Analog型，時間領域 Convolution 積分型	ETL-Mark-4A	※ 1週間：0.5秒
橋本（新）	Terminal Analog型，時間領域 Convolution 積分型	M1B	9時間：1秒
RAAG，音声研究グループ	Terminal Analog型，周波数領域アナログ計算機	Hitachi-ALS-200	1500 : 1

※ホルマント周波数からインパルスレスポンスを求める Laplace変換まで含んでいる。定常的な音声ぺすべてのサブルーチンを数表化した場合4000：1になると報告されている。

T型では制御パラメターがいわゆる acoustical parameters、であり，物理的に特徴的な事象，たとえばホルマント周波数など直接に対応しているので，従来の研究実験結果から比較的容易に合成のルールを作ることができる。

(2) 連続音声の合成能力について

V型では音韻から音韻への transision をきわめて自然に合成することができ，連続音声の合成能力は大きい。

T型では自然な移行がむずかしく，ことに子音から母音への過渡においてスペクトルの zeroのふるまいが複雑で，連続音声合成のためにはルールが複雑になるおそれが多い。

結論的にいって現在までのわれわれの音声合成の知識が主として acoustical parameters についてのもので，現状としてはT型の方が取扱いやすいが，その原理的な潜在能力としてはV型の方が，とくに連続音声の合成に対して大きいと考えられる。*

2. Computer Control

"DAVO" や "POVO"のようなすぐれた性能のアナログ音声合成装置をもっている M.I.T. において主として研究されているが，わが国ではまだ発表された研究,[8][9]実験はない。computer controlの利点は次のようである。

(1) 操作が簡単で．適応性が広く，制御が確実である。たとえば DAVOでは synthesisのための調整点が多く，入力波形および制御電圧は梯形波に限られており，すべての制御は同期的にしか行われない。

(2) 音韻記号のような不連続入力から連続的な音声を real timeで合成することができる。たとえばDAVOの現在の制御機構では2音節程度の音声しか連続的に合成することはできない。またcomputer synthesisでは real timeに近い高速処理はなかなか困難である(第1表参照)。

computer control にも制御する音声合成装置の型に応じて vocal tract analog型と terminal analog型があり，原則的にはすすでにのべたような利害得失が考えられるが，従来の hardwareによる制御機構に計算機が代わるわけだから複雑な制御を確実に行なうことができるようになり，潜在的な能力の大きい vocal tract analog型の方が有利となる。

2.1. Vocal Tract Analog Synthesizer の Computer Control

M.I.T. のJ.B.Dennisらによって，すでに G.Rosenらによって開発された vocal tract analog型のspeech synthesizer DAVO を computer TX-0 で制御すべく研究が行なわれている。

(1) 合成法の概要

DAVOによる合成の概要についてはすでによく知られているのでここでは省略する。詳細は下記の二つの論文を参照されたい。

G. Rosen ; "Dynamic Analog Speech Synthesizer," JASA., **30**, 3, pp. 201-209 (March, 1958)

Michael H.L.Hecker ; "Studies of Nasal Consonants with on Articulatory Speech Synthesizer," JASA., **34**, 2, pp. 179-188(Feb., 1962).

(2) 制御法の概要

制御法をできるだけ簡単にし，しかもできるだけ広範囲の音声を合成することができるようにするために，制御は次のような順で行なわれる。

*人間の発声による音声の構造的な制扼が，自然な形で合成装置の機能の中に内蔵されている程度のちがいによる。

Experimenter→ Event Compiler→ Control Program→ Control of DAVO
‖ ‖ ‖
Phonetic Input→ Translation(Set of Rules) → Manipulation

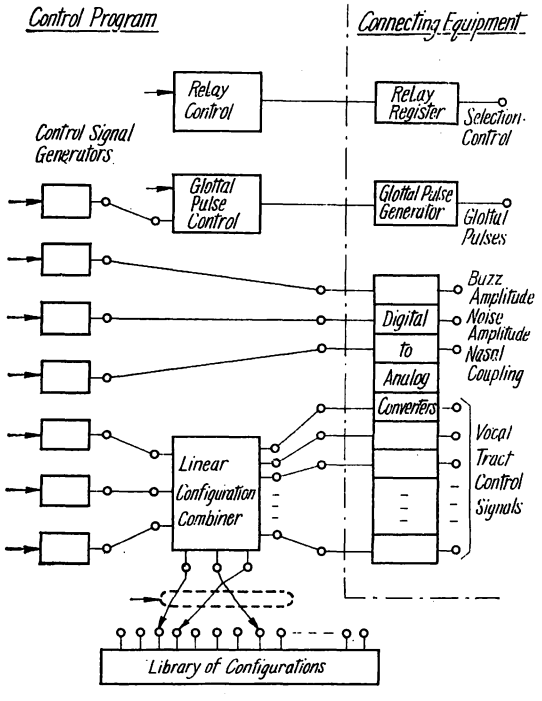
制御の主要部は，不連続な音韻入力から連続的な音声合成を行なうために必要な制御出力を作り出す translationの部分である。その中心となる合成のための“Set of rules”は，合成modelが実際の人間の発声機構に近づけば近づくほど簡単になり，結果が自然になる。

Event compiler は，入力の音韻記号からそれに対応した音声を合成するために必要な eventsの Listを簡単に準備できるようにするためのものであり，ここでその listを modefyしたりパラメターを変更したりできる。

control programは，event compiler の出力である eventの list から，実際の制御に必要な出力の形に，時間的にその情報を organizeする。

時間的には buzzの波形，noiseの強度，nasal couplingの三つの情報について，各 segmentごとにその初期値，中間値，最終値および継続時間が与えられて，その間を parabolicに内挿する。

vocal trct の configulationは libraryからえらばれたいくつかの configulationの linear combination として与えられる。しかもこの linear combination のための係数を，時間的に上にのべたように各 segment ごとに2次曲線的に制御する。このようにすれば configulationの libraryの内容を少なくして，しかも configulation の時間的な変化を詳細にあらわすことができる。



第8図 Control programの内容の説明

この制御の要点は第8図でよく理解されると思う。ここで control signal generator というのが時間的に2次曲線的に変化する制御出力を発生する。最初の三つがそれぞれ buzz amplitude, noise amplitude と nasal coupling を controlし，次の三つがconfigulation の linear combinationの係数を制御する。

この control programの入力すなわち event-comilerの出力である eventの list は，次のような事項を含む。

(a) Control signal generatorの起動：特定の control signal generator に次の segment の paraboric制御電圧を発生するように必要な四つのパラメター（初期値，中央値，最終値と継続時間）を与える。

(b) Glottal pulse entry：control programに control signal generator の出力に比例した周波数のパルスを発生するか，または定められた時間にパルスを発生するかを指令する。

(c) Relay control entry：vocal tract のどこのsectionに noise source を挿入すべきかを control programに指令する。

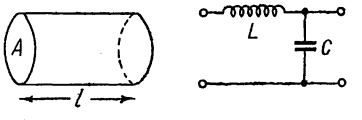
(d) Configuration entry：linear combination の入力として libraryからとりあげるべき configulationを指示する。

control programは各 time unitごとに2次曲線的な control signal を発生しつつ進み，その時間が次の event list をよむべき時間に達すると，次の event が解釈される。

(3) Computer controlに適した合成装置の改良。

実際に現存する DAVOを TX－0で制御してみたところいろいろと問題があることがわかったので，すべてをトランジスタ化し，ディジタル制御が容易にできるような形に改良しようとする試みがなされた。その一つとしてここにM.I.T.における検討の1例を紹介する。

vocal tract の基本 unitとしての1区間は第9図のような音響管で近似され，その特性は電気的に同図の右に示されるような等価回路であらわされる。



第9図
音響管の電気的等価回路

ここで $L=\frac{\rho}{A}$，$C=\frac{A}{\rho c^2}$ (1－10)

または $A=\rho c\sqrt{\frac{C}{L}},\quad l=c\sqrt{LC}$ (1−11)

Aは管の断面積，lは管の区間長，ρは空気の密度，cは音速である。

この電気的等価回路マトリックス表示を考え，その中で一番簡単な表示をとると次のHマトリックスとなる。

$$[\mathbf{H}]=\begin{vmatrix} p\frac{\beta}{\omega}\sqrt{K} & 1 \\ -1 & p\frac{\beta}{\omega\sqrt{K}} \end{vmatrix}=\begin{vmatrix} j\frac{\omega\sqrt{K}}{C} & 1 \\ -1 & j\frac{\omega}{C\sqrt{K}} \end{vmatrix}$$

$$K=\frac{L}{C},\quad \beta=\omega\sqrt{LC} \qquad (1-12)$$

このHマトリックス表示の等価回路を考えると第10図の(a)のようになり，これから同図(b)(c)に示すような安定性向上のための変形を加えてゆくと第10図(d)に示すような等価回路がえられる。これは第11図に示すようなM.I.T.での試作回路の原理を示すものである。

第10図(d)の回路の伝達関数は

I_1からV_1へは $$\frac{-\left(\frac{1}{j\omega}\right)C\sqrt{K}}{1-\left(\frac{1}{j\omega}\right)^2C^2}\simeq j\omega\frac{\sqrt{K}}{C} \qquad (1-13)$$

I_1からI_2へは $$\frac{\left(\frac{1}{j\omega}\right)^2C^2}{1-\left(\frac{1}{j\omega}\right)^2C^2}\simeq -1 \qquad (1-14)$$

であり，$C^2/\omega^2\ll 1$が近似成立のための条件を与える。

このような近似の成り立つ周波数範囲では，第11図からわかるようにその等価的なLとCの値は次のようになる。 $$C=C_iR_i\frac{C'}{A} \qquad L=C_iR_i\frac{L'}{A} \qquad (1-15)$$

M.I.T.の試作装置では$1/C'$，$1/L'$がdigital signalによって直接制御されるstep atte nuatorであり，1 dbから64 dbまで1 db stepで変えられる。このような等価回路でえられる範囲での等価的な断面積の変化範囲は，15～0.01cm²であり，ほぼ完全な閉鎖を行なうことができる。

またCRの積分回路の不完全さからdampingを生じるが，等価的なQの値で20程度までは実現可能である。

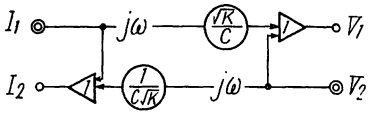
しかしこの回路もまだ完全に実用化されたわけではなく，このような近似回路を何個も直列に結合したときの誤差，唇の幅射インピダンスと声門の音源インピダンスをどう表現するかなど問題が残っている。

2．2．Terminal Analog Synthesizer のComputer Control

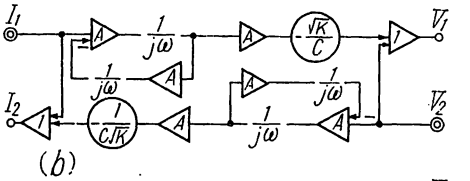
M.I.T.においてPOVOのTX−0によるcontrolが研究されている。合成装置としてのPOVOはすでによく知られているように可変共振回路3段の直列接続であり，control signalは三つのホルマント周波数，音源の振幅，ピッチ周波数などである。

制御の方法としては各制御パラメターごとにtime segmentを行ない，そのsegmentの継続時間，segment中での制御信号の初期値，中央値，最終値，他の制御パラメターとの間の相互関係などを与え，制御信号は，そのsegment中の三つの値をむすぶ連続2次曲線として出力を発生する。

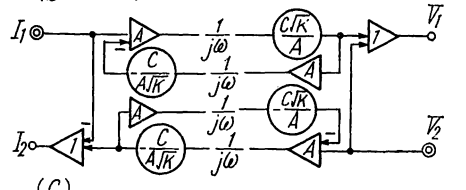
time segmentの時間幅は一定ではなく，制御パラメターの変化速度に応じてとられ，能率よく制御信号を発生する。



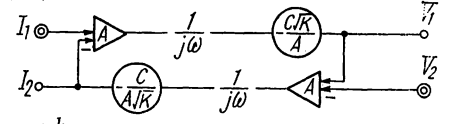
(a) (1-12)式のHマトリックスの等価回路



(b)

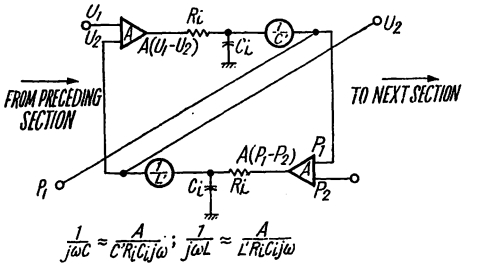


(C)



(d) 安定度向上のための変形をへた等価回路

第10図 単位長音響管のHマトリックス表示の等価回路



第11図 Computer control用に改良されたvocal tract analog synthesizerの基本回路

3. Computer Synthesis と Computer Control の比較

Computer synthesis と computer control の比較はある意味では digital simulation と analog simulation との比較ということができる。すなわち computer synthesisでは合成の可能性（合成法と制御法の flexibility）は大きいが，real timeに近い高速の合成はむずかしく，computer control では合成の能力はそのアナログ合成装置の能力によってきめられているが，real time に近い高速の合成が可能である。そこでむしろその利用の目的によって次のように考えるのがよいと思う。

(1) 合成法の原理的な研究，実験用としては，computer synthesisことに vocal tract anolog型の simulation

(2) 実用的な意味での音声合成の実験，研究用としては computer control.

(3) Computer control の対象としての合成装置としては，現在での合成能力はむしろ terminal analog型の方が大きい（すでにじゅうぶんに開発されているから）と考えられるが，潜存的な可能性としては，ことに自然な連続音声という点では vocal tract analog の方が大きいと考えられる。

(4) 将来の問題としての motor commandによる computer control を考えた場合にも vocal tract analog 型の方が有利であると考えられる。

4. 連続音声の合成法則

パラメターの連続的な制御という形での連続音声の合成のための合成法則一般については，主としてHaskins研究所において研究され，合成のための "minimum rules"として発表されている。ここで minimum rules といっているのは，合成された音声の了解度と自然性は当然その合成法則の複雑さに関係するが，法則の数ができるだけ少なく，その構造もできるだけ簡単で，しかも不連続な音韻の系列をじゅうぶんな了解度と適当な自然性をもった連続音声に通常の会話速度で変換することのできるような合成法則という意味である。

Haskins研究所での連続音声合成のための minimum rulesの研究は acoustic parameters（主としてホルマント周波数）の制御に関して行なわれたものであるが，その要点を列挙すると。

(1) 法則は subphonemic rules として，音韻発生の manner of articulationと place of articulationについて記述し，それらを同時的に，網目的に組み合せるのが能率的である。

(2) 語中の音韻の位置によって positional variationを与える。

(3) prosodic featuresの一つとして stress を考慮する。stressはピッチ周波数，強度（母音の）継続時間として制御される。(実際には unstressed syllable中の母音は定常部分としては存在しない程度にする。)

(4) articulationと soundの間の対応が1対1でないことによる variation を加える（たとえば/ g /のF_2 locusの後続母音の変化による不連続な変化）

このような考え方によって作り上げられた彼らの mimimum rulesによる acoustic parametersによる合成の例を第12図に示す。

結　　言

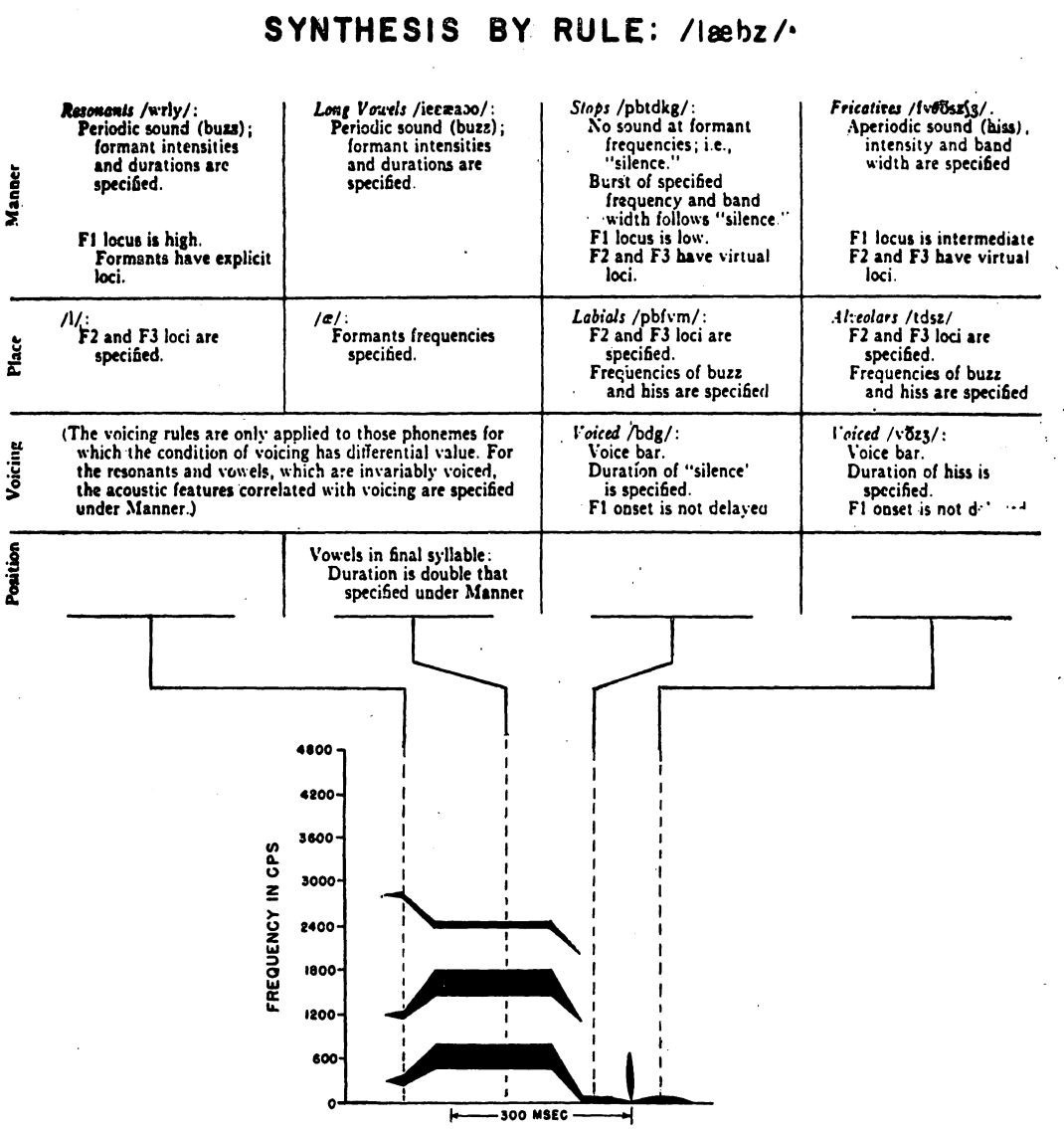
パラメターの制御という形で，音声合成装置を通じて，不連続な入力の系列から連続的な音声を合成する場合，一番問題になるのは調音結合による相互作用をいかに円滑に実現するかということであろう。

一般的にいって，音声合成装置の機能が，人間の発声機構のアナログとしての程度が高くなればなるほど，人間の発声によるために音声に課せられる制扼と構造的な情報が自然と合成装置の機能の中に内蔵され，自然でより円滑な連続音声を合成することができるようになり，さらにその制御（合成）の法則もより本質的，簡単で合理的なもの（例外則や変形の少ない）となると期待することができる。

このような考え方をおしすすめて，Haskins研究所の F. S. Cooperや A. M. Libermanらは motor theory of speech perceptionに対応して motor command によるアナログ音成合成装置の制御の可能性と有効性を主張している。[11]

結論的にいって合成手段としては，実用的な点からいえば vocal tract analog型または terminal analog型合成装置の computer control が，研究的な面からいえば vocal tract analog 型の computer simulation が最も有用な方法であろう。

合成法則の問題としては，phonemic featuresによる制御法則はほとんど知られており，あとはこれを連続的な音声としじゅうぶんの了解度と必要な自然性をもたせるための linguistic featuresによる変形（positional variation）と prosodic featuresによる修飾をいかに考慮するかが問題である。



第12図 "Synthesis by Rules"による連続音声合成の例

ただ articulatory configurationによる制御では，各音韻に対応する target configulationの間を実際にどのような configulationの変化をとってたどってゆくか，という点について実際的な情報が不足しており，X線映画撮影などによる分析的な研究と，合成実験による"best guess" configulation trojectoryの解明が大いに必要である。

第1部 参照文献

(1) 例えば (a) L.P.Schoene et al ; "Design and Development of a Digital Voice Date processing System," Rep. No. AFCRL－62－314(1962).

(b) 藤村靖 ; "音声合成の一実験," 音響学会研究発表講演論文, 2－1－14 (38年10月)。

(2) 例えば (a) W. Lawrance ; "The synthesis of Speech from Signals which have a Low Information Rate," Communication Theory (Ed. by W. Jackson) (1953).

(b) H.J.Manley ; "Analysis-Synthesis of Connected Speech in Terms of Orthogonalized Exponentially Damped Sinusoids," JASA., **35**, 464 (1963).

(3) 例えば (a) G.Rosen ; "A Dynamic Analog Speech Synthesizer," ASA., **30**, 200－209 (1958).

(b) G.Fant et al ; "Recent Progress in Formant

Synthesis of Connected Speech," JASA., 33, 834–5 (A), (1961).
(4) J.L.Kelly,Jr. and L.J.Gerstman ; "Digital Computer Synthesizes Human Speech," Bell Lab. Rec., 40, 216–218 (1962).
J.L.Kelly,Jr. and C.Lockbaum ; "Speech Synthesis," Speech Communication Seminar, Stockholm(1962).
(5) 猪股修二 ; "電子計算機による音声の発生について," 音響学会誌, 17, 93–102 (1961).
(6) 橋本新一郎ほか ; "電子計算機による母音の合成," 音響学会研究発表講演論文, 2–1–13 (38年10月)。
(7) The RAAG Phonetical Group; "Some Preliminary Experiments on the Verification of Priciple of the Composition and Decomposition of Phonemes," RAAG Memoirs,111, H, 693–713 (1962).
(8) J.B.Dennis ; "Computer Control of an Analog Vocal Tract," Speech Communication Seminar, Stockholm (1962).
(9) W.L.Henke ; "Computer Control of a Terminal Analog Speech Synthesizer," QRP., MIT., No.68, 167–169 (1963).
(10) E.C.Whitman ; "A Transistorized Articulatory Speech Synthesizer," QRP., MIT., No.68, 164–167 (1963).
(11) F.S.Cooper et al ; "Speech Synthesis by Rules," Speech Communication Seminar, Stockholm (1963).
(12) A.M.Libemran et al ; "Minimal Rules for Synthesizing Speech," JASA., 31, 1490–1499 (1959).
(13) 最近, T型の compnter synthesisについて非常に巧妙な方法が発表されたのでこゝに追記する。
J.L.Flanagan et al ; "Digital Computer Simulation of a Formant-Vocoder Speech Synthesizer" 15th Anual Meeting of Audio Eng. Soci. No. 307, (Oct. 1963)

第2部 録音された音声セグメントによる音声の合成

Compiled Speech or Speech Synthesis from Stored Segments

中田和男 光岡輝義*

緒言

不連続な入力の系列から連続的な音声を合成する方法として, たとえば電文を音読するように, 一つ一つの不連続入力に対応してすでに録音されている音声セグメント (現在のところ人間による発声が主として用いられているが, 必ずしも人間の発声に限るわけではない) を一つ一つ取り出して適当に時間的に連結すればよいではないかということは, 原理的には簡単に誰れでも考えつくことであり, 事実すでに1953年に C.M.Harris が音声の building block 合成方式として提案している。[1][2]

しかしこれを簡単に実験してみようと思って, たとえば録音テープの切り継ぎによって, ある単位での音声の寄せ集めから連続的な音声を編集してみると, 全然不自然で多くの場合その意味すら理解しえないという困難にぶつかり, 改めてこの問題の本質について反省し考察するということになる。[3]

このような方法 (予め録音されている音声セグメントの編集) による音声の合成が可能だと考える背後には, 連続的な音声も, 適当な単位をとれば, 時間的にその単位ごとに区切ることができるという segmentation の可能性を仮定している。しかも音韻識別におけるような音韻情報の面からだけではなくて, 自然性も保持するためには prosodic features* (duration, stress, pitch, intonation, vocal quality など) についても適当な時間分割ができなければならないから, 問題は一段と複雑にまた困難になる。

しかしこのような方法 (compiled speech) による連続音声の合成が全く不可能だというわけではなくて, 実験的に注意深く作られた posion** と prosodic features によるいくつかの変形を含んでいるような録音から編集すれば, intonation や stress がやや不自然に感じられるところはあっても, じゅうぶん了解度の高い連続音声を合成できることが実証されており, 実用的な意味からいってもじゅうぶん研究価値のある方法である。そこで以下にその問題点を検討し, 今後の研究の参考とする。

*東京電機大学大学院

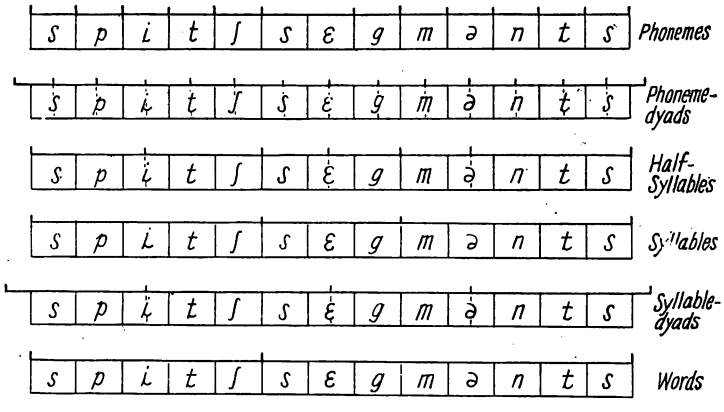
*主として音声の不自然さ, 人間の声らしさをあらわす特徴 (情報要素) と考えてよい。
**一つの単語中の音韻 (音節) の位置, 一つの章, 句の中の単語の位置などを意味する。

1．録音単位(segmentation unit)と必要な記憶内容(inventory)

録音音声の編集による音声合成の最も重要な問題は，予め録音しておくセグメント単位の大きさとその性質およびその単位で連続的な音声を合成するために必要な記憶容量とその内容(inventory)であり，まずこの点について考察する。

Petersonと Sivertsenの考察にしたがえば，予め録音しておく音声セグメントの単位として次のようなものが考えられる。(第13図参照)

(1) Phonemes(単独に分離された母音および子音)：連続音声中の音韻には調音結合による相互作用が強く，単独の phoneme 単位での連続音声の編集はほとんど不可能に近い。



第13図 録音セグメントの種類の説明図
例は/spiʧ segments/〔speech segments〕

(2) Dyads(一つの phonemeの target position (安定点) から次の phonemeの target positionまでの区間)：Petersonと Wangらによって研究されたもので，単独な phonemeによるものよりも円滑な編集ができるが，必要な inventoryは phonemeの場合に較べて急激に増大する。

(3) Half-Syllables(一つの syllableの始まりからその syllableの核の中心部まで，または中心部からその syllableの終りまで)：dyadよりも大きな単位であり，それだけ編集音声の了解度と自然性はますが，inventry さらに大きくなる。

(4) Syllables

(5) Syllable Dyas (phonemeに対する dyadの考えを syllableにまで拡大したもので，一つの syllableの核の中心から次の syllableの核の中心にいたるまでの区間)：syllableの始めも終りも自然な形で含まれるが，inventoryの大きさはさらに大きくなる。

(6) Words ：一般的にいって録音セグメントの単位が大きくなるほど調音結合による相互作用は自然な形でその単位の中に含まれることになり，それだけ了解度と自然性の高い音声が編集できることになるが，positionと prosodic featuresによる変形を含んだ inventoryの規模は大きくなり，編集(compilation)に時間を要するようになってくる。この二つの相反する要求からいって，word がおそら

第2表 各種のセグメント単位を利用したときの必要記憶量 (G.E.Petersonと E.Sivertsenの研究による)

Segment Type	Phoneme Sequences	Segments
Phoneme	37	155
Phoneme dyad	1218	8460
Half-syllable	1647	11529
Syllable :		
Dewey, total	4400	30800
Dewey, most frequent	1370	9590
Syllable dyad :	858458	40173336
Word :		
Dewey ; total	10119	
Dewey, most frequent	1027	
French, total	2822	
French, most frequent	737	
Black and Ausherman	6826	

注 (1) Phoneme Sequences の欄は phonemic に必要な segment の最小数。
(2) Segments の欄は prosodic conditions による変形まで考慮したときの必要な inventoryの最小量。
(3) Dewey, French, Black and Ausherman は推定の基礎とった言語統計の種類を示す。

く最適な単位ではないかと考えられる。

参考として petersonと Siverten の考察による録音単位の選び方とその場合に必要な inventoryの大きさの推定値を第2表に示す。[7]この表で特に注目すべき点はdyadの inventoryの大きさが他に比して過大となっていることで，セグメントの単位が音声の音韻的または言語的な単位と一致しないと inventoryの大きさが過大となって非能率的であるということを物語っている。

2. Spellingの問題

録音音声の編集による一般的な連続音声合成の難問の一つは spellingの問題である。spellingの問題というのはその inventoryにない入力の処理法として，より基本的な（音韻または音節のような）小さな単位の録音からその出力を編集することである。* inventoryの大きさが有限である以上 spellingの問題を完全に避けることはできないので，その起る確率が問題になるが，それはひいては inventroyの内容を決定するために用いられた言語統計の内容に関連してくる。

いま wordを単位とする場合について spellingの必要性に出くわす確率を推定すると，[4] inventoryの大きさが7000語のとき約5％（20語につき1語の割合），20000語のとき約1％(100語につき1語の割合）となるが，この推定ではすべての固有名詞（人名，地名など）を除いてあるから，一般的な音声の編集においては spelling の問題というのは実際的に相当重大な問題であるということがわかる。その一つの解決法は，英語の場合/$-s$/, /$-d$/, /$-t$/ のような allophonic** な共通の語尾を別に加えて inventoryの内容を能率的にすることである。

3. その他の問題点

(1) 文字や音韻記号のような音韻的（phonemic）な情報だけから wordを構成し，その wordの言語的な働きからそれが連続音声としてはどのように発音されるかをきめ，それに相当した適当な変形をほどこした wordの録音を準備し，それを抽出してくる法則，いわゆる“language problem”[9]の解決法。

(2) 入力信号に応じて inventoryの記憶番地を探して能率的な編集（compilation）をする方法。

(3) 大容量で random accessの記憶装置とその高速制御法。

4. 実 験 例

Compiled spechの実験例として Haskins 研究所の Cooperがその論文の中で引用しているのは，同研究所の試作的な装置による wordを単位とした実験で，[4]内容は Bertrand Russellの論文であると報告されており，じゅうぶんに了解度の高いものであったらしい。

またわが国における実験例としては，先ごろ電々公社通研において電話番号の問合せに対する自動解答装置への応用を目的として，数字音声を単位とする実験が行なわれた。その結果は学会において録音テープによって再生されたが，かなり良好でじゅうぶん了解度の高いものであった。[6]ただ random accessの大容量記憶装置として磁気ドラムを使用したがそのアナログ信号録音特性が悪く S/N比が悪い欠点がみとめられた。

結 言

以上の考察からわかるように，予め録音された音声セグメントからの編集による compiled speech の問題点は，次の3点に要約される。[9]

(1) セグメント間の継ぎ目をできるだけ円滑に連結する。具体的にいうとホルマントの周波数と強度の変化が連続的に連結されるだけでなく，ピッチ周波数の急変をさけるためにその高調波成分までも連続的に変わることが望ましい。

(2) 必要な inventoryを過大にしない。そのためにはすでに明らかにされたようにセグメントの単位と音声の言語的な単位とが一致しなければならない。

(3) Inventory中の記憶番地を能率的に探す。そのためには入力信号自体が addressとして動作するような単位が望ましい。

これらの考察からして wordがおそらく最適なセグメントの単位であると考えられる。

日本語の場合についての猪股の考察によれば，単語を単位としたとき，約1万語の inventoryでほぼじゅうぶんではないかといわれている。

Language problemの解決法については，音声的な言語学による研究が必要であり，本質的にはアナログ合成装置による synthesis by rulesの問題と全く同じであり，今後の重要な研究問題である。

---

*小さな単位から編集された（spelling）音声の了解度と自然性は当然劣っている。

**英語の場合の語尾の/s/, /d/(or/t/) のように文法的に意味をもつ最小の音声単位を allophoneという。

あとがき

以上新しい発展段階を迎えた音声合成の研究の現状について概観し，その問題点について検討した。

多くの音声合成法の可能性が考えられたが，その一つの typical な例が“synthesis by rules”であり，記憶容量は少なくてよいが，合成のための ruleが簡単であればあるほど人間の発声機構とのアナログの程度の高い高級な音声合成装置と複雑なロジックを必要とする。しかし spellingの問題はおこらない。これと対照的な例が compiled speechであり，アナログ的な音声合成装置は必要としないが，大容量の random accessの記憶装置を必要とし，spellingの問題はさけられない。

結局記憶容量の大きさと合成，制御装置の高級，複雑さとがお互に取り引きされているようなものである。*

この二つの typical な方法の中間にいくつかの中間的な hybrid systemが考えられる。その一つは入力の文字を音韻的な記述に変換するのに，発音に必要な情報まで書きこんだ dictionaryをさがし，その結果から rules によって合成装置を制御するという方法である。他の一つは入力によって音声合成装置を制御する制御電圧値を stored tableから直接よみだすという方法である。この四つの代表的な方法の相互関係を第14図に示す。

いまこの第14図に示された四つの方法について，その必要とする記憶容量を推定してみると，[4]すべての方法の能力を同じと仮定して（20×10^3 wordsの vocabylary であり，spellingの確率を1％として），

(1) Compiled speech ：約 400×10^6 bits

(2) Hybrid method
- (a) control voltage型 ：約 10×10^6 bits
- (b) phonemic dictionary ：約 1.5×10^6 bits

(3) Synthesis by rules ：約 50×10^3 bits

このような考察からして，装置（記憶，制御，合成すべてを含めて）が最も経済的でしかも最も高品質の連続音声を合成しうる可能性が多いことからいって，hybrid methodの中の phonemic dictionary look upと synthsis by rules を組み合せたものが最も有利な方法と結論される。第1部においてのべたベル電話研究所や M.I.T.における最近の研究，実験はすべてこの線に沿ったものといえる。われわれもまたこの方向に向って研究を進めるべきであると考える。しかし実用的な意味で，使用する vocabularyに制限をつけうるような場合には，compiled speechの有用性もじゅうぶん考慮に値するものと考える。

最後にこのような調査とその発表の機会を与えられた河野次長，尾方室長に感謝するとともに，電機大学の関係各位にも深謝する次第です。

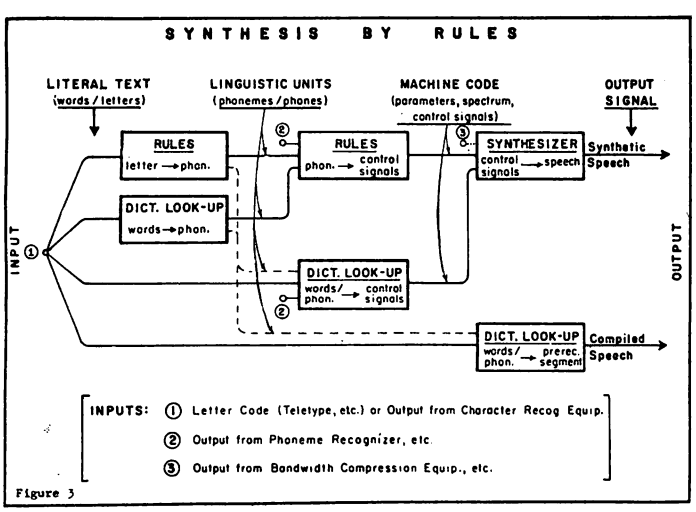


Figure 3

第14図 不連続入力から連続音声を合成するための方法の比較

*合成装置の機能の中に音声の構造的な情報が内蔵されていると考えられる。

第2部 参照文献

(1) C.M.Harris ; “A Study of the Building Blocks of Speech,” JASA., **25**, 962-969 (1953).
(2) S.Inomata ; “A New Scheme for Speech Generation, s.s.s,” 電気試験報, **24**, 1, 47－57 (1960).
(3) A.N.Stow and D.B.Hampton ; “Speech Synthesis with Prerecorded Syllables and Words,” JASA., **33**, 6, 810 (1961).
(4) F.S.Cooper ; “Speech from Stored Data,” IEEE. Conveniton Paper, 53.2 (Mar., 1963).
(5) S.E.Esteo et al ; “Speech Synthesis from Stored Data,” JASA., **34**, 2003(A) (1962).
(6) 関口茂，橋本清，三浦種敏 ; “音声編集制御方式,” 通信学会全国大会講演論文, 37 (昭37).
(7) E.Sivertsen and G.E.Peterson ; “Studies on Speech Synthesis,” Rep. No. 5, Speech Res. Lab., Univ. Michigan (1960).
(8) G.E.Peterson and W.S－Y. Wang ; “Segmentation Techniques in Speech Synthesis,” “A Segment Inventory for Speech Synthesis,” Rep. No.1, Speech Res. Lab., Univ. Michigan (1958).
(9) F.S.Cooper et al ; “Speech Synthesis by Rules,” Speech Communication Seminar, Stockholm (1963).