技術解説

音声の合成と認識について

角　所　　収*

1．はしがき

このテーマに関しては，やや堅い学問的な見地から，すでに2，3回論じたことがある[1)~4)]ので，今回は少し記述法を変えて，必ずしも学問分野的に近い位置におられない方のために，音声研究の歴史的発展を述べながら，可能な限り平易に音声研究の問題点を説明してみよう．

2．音声研究の開始

世界で最初の音声合成器が試作され，音声の研究が開始されたのは今から約200年前である（1779年のKratzensteinの研究，1771年のvon Kempelenの研究が有名である）[5)]．日本では昭和17年頃，東京外国語学校（現在の東京外国語大学）の千葉・梶山両氏による先駆的研究が行われた．その後，昭和19年頃阪大（牧田）[6)]で音声合成の研究が行われ，ついで，昭和30年頃から日本音響学会とか電気通信学会を発表の場として電気通信大学（藤村），郵政省電波研究所（中田，鈴木），東北大学（大泉，馬渕），山梨大学（重永），電々公社電気通信研究所（三浦，斉藤），京都大学（坂井，堂下），大阪大学（加藤，北村），日本電気（永田，加藤）が今日でいう音声合成，認識の研究を開始した．

3．帯域フィルタとゼロ交叉波—
——観測装置上の制約

この頃までの研究は帯域濾波器が音声の合成と認識の何れの研究においても主役を果していた．当時，音の分析といっても，測定用のツールとしては1/3オクターブ・フィルタ等の帯域濾波器位しかなく，その出力を整流してペンレコーダで記録するのが唯一といってもよい位の分析法であった．音声の合成や認識もこの程度の分析の結果を基礎にして行うのであるから，合成の場合には10個程度の帯域濾波器を並列に接続し，それを弛張発振器等で励振し，各帯域濾波器出力の相対的振幅を制御することにより音声を合成した．また認識は同様な帯域濾波器の相対的出力を比較することにより，認識論理を構成するのが普通であった．このことは，研究の各時点で利用可能な観測装置の能力が，その研究の方法にいかに大きな制約となって現れるかを示す好例である．

その後音声波形をそのまま記録できる応答速度の速い各種計測器の発達に支えられて，音声波形を仔細に観測したり，また今日でも使用されているソナグラムにより音声の周波数スペクトルの三次元的表示（時間，周波数，振幅）をすることが可能になり（昭32～35年），色々な音声（単音節程度）の短時間周波数スペクトルの面からみた特性がかなり明らかになった．ほぼこの頃（昭31年），ゼロ交叉波が2値的な波形情報しかもっていないのにも拘らず，文章のような音声の分析合成には充分な明瞭度を保存しているという研究結果（1950年 Licklider）に着目して，ゼロ交叉波を処理対象とした音声認識の研究が行われるようになった．これは情報処理装置の能力が不充分（NEAC 2203が出始めた頃）であったので，その弱点をカバーするためにある程度データ量を低減させて処理を行う必要があったためである．実際には音声波そのものをゼロ交叉波にするのではなく，数個の帯域に分割濾波しその出力を処理する方法がとられた．このゼロ交叉波処理システムは比較的簡単な論理回路による処理を組み合せれば，音節程度の認識に威力を発揮し，音声認識システムの構成までも行われた（坂井，堂下）．

*角所収（Osamu KAKUSHO），大阪大学産業科学研究所，電子科学研究部教授，工学博士，情報工学専攻

4．ダイナミック・アナログ型合成器

音声は，声帯により発生された音源信号が声門から唇に至る分布定数回路を通ることにより生成されるものであることは容易に理解できる（声帯が音源とならない無声音の場合は乱流による雑音が音源となる）．事実この様な着想により，電気的分布定数回路をパルス状音源とか雑音により励振して音声を合成する研究はMIT（Rosen）で行われていた（ダイナミカル・アナログ型合成器，略してDAVOといわれた）．この場合L，Cの値の制御は，電圧，電流と回路定数との非線型関係を利用していたので，その正確な制御は困難であった．丁度この頃，制御関係等の研究にアナログ計算機が導入され，威力を発揮し始めていた．アナコンを利用すれば回路定数の正確かつ高度な制御の問題はかなり解決されることになる．問題は基礎データとなる各音韻毎の声道の形とその時間的変化に関するデータである．これらのデータはX線撮影以外には適当な方法がみつからず，制御は正確に出来ても，設定すべきパラメータに関するデータの不足が大問題であった．その基礎データを明らかにするために，プラスチック平板を多数重ね合せて複雑に断面積が変化する声道を模擬し，声帯にはトランペットホーンのユニットを利用した機械的な合成器による研究が行われた．このプラスチック合成器は，既存の合成器と比較して極めて良好な音声を合成できた（寺西）．ついで，パラメータの時間的に線型な変化を仮定して，合成用のパラメータの制御のための基礎データが作られた．これに基づいてパラメータの時間変化をディジタル計算機により計算し，アナコンを制御して合成音声を発声するハイブリッド型音声合成装置は世界一長い文章の音声合成に成功した（童話「桃太郎」，“Sleeping Beauty”他，昭43年頃）．これは電子装置の進歩を巧みに捉え，かつ既存の音声生成に関する知識をフルに利用した研究であった．

5．電子計算機万能の時代

その後電子計算機の処理能力の向上に伴って合成も認識も計算機万能の方向に向かってゆくことになる．昭和35年頃はディジタル計算機のみでは1秒の声を合成するのに何時間もかかった（猪股）．また，合成された単母音も，そういわれれば何とかそう聞こえるという程度の代物であった．一方，前述のソナグラム分析から音声は短時間周波数スペクトル上の高々4つのピークと1～2個の谷で近似できることがわかってきていた．この研究結果を利用して，単一共振回路，反共振回路の縦続接続により声道の特性を近似し，有声音用音源としての声帯波には非対称な三角波を，また囁き声とか無声子音の場合には，電子管雑音を励振源として音声を合成する研究が行われた．この方式（ターミナル・アナログ方式という）による声は極めて明瞭度が良く，種々の単音節の合成パラメータの効き具合を定量的に明らかにするのに大きな貢献をした（中田，鈴木）．その後この方式は，方式自体は保存しながらハードによる合成システムの構成から計算機シミュレーションへと変ってゆくことになる．ある程度自由に，短い音声を合成する手段を手に入れた研究者達は，それぞれの合成器を駆使して合成用パラメータと明瞭度との関係を明らかにするための地道な基礎的研究を行う様になり，この様な研究の進め方が主流となった．

6．パターン認識の研究の影響
——特徴パラメータの抽出とその精度

この間に，関連分野としてのパターン認識と学習に関する研究の発展は音声の研究に大きな影響を与え，音声認識もパターン認識の一分野であるという認識が定着し，パターン認識システムにおける観測，特徴パラメータの抽出，識別という各ブロックの機能の検討が音声認識の分野においても行われる様になった．文字認識の分野においてもそうであったが，音声の研究においても，一番の問題点は特徴パラメータの選択とその計測精度の問題であった．ある特徴パラメータの組を用いて音声を合成すれば，まずまず特定の音節らしく聞こえる声が出来たとして，次にその同じ組のパラメータを用いて音声の認識を行う場合，次の2点が問題になる．

(1) その特徴パラメータの実時間抽出のためのアルゴリズムの確立

(2) どの程度の精度でそのパラメータを抽出すべきか

(1)の問題の解決のためには音声分析法の研究が重要となり，その方面の研究成果が音声研究全般に大きい影響を与えることになる．その一つの例として，高速フーリエ変換法（FFT）の出現（Cooley Tukey, 1965年）は今までディジタルで実時間周波数分析を可能にする具体的方法を持たなかった音声分析の研究に大きなインパクトを与え，この適用により，極めて正確に短時間周波数スペクトルを計算することを可能にした．さらにまた音声の重要な特徴パラメータであるピッチ情報の抽出のためのケプストラム法と呼ばれる新しい方法をも可能にした．音声波中のピッチ情報は音声の質に大きな影響を与えることは音声研究の初期から判っていたが，その抽出方法に信頼度の高いものがなく音声研究進展上の一つの大きな障害となっていた．高速フーリエ変換を2度行うことにより得られるケプストラムの開発は，一つの解決法を与えることになった．この外にもPARCOR方式の開発があるが，時期的には少し後になるので後述する．

(2)の問題について，人間も区別できない程度の精度で音声パラメータを抽出することの無意味さは，少なくとも人間を音声合成・認識システムの一つの具現とみる立場からしばしば指摘されてきていた．有用な特徴パラメータと考えられるものについてはその弁別限（DL, difference limen, 人間が知覚できる特定パラメータの最小の変化）を分析精度の基準とすべきであろうと考えられる．DLは合成器により多数の合成音を作りこれを人間が聴取するという心理実験的方法により求められるのが普通である．

7．音声合成上の問題点と認識率の評価

ここで良い声を作る時の難しい点を一寸振り返ってみよう．

8つ程の帯域濾波器を並列に接続したものに非安定マルチバイブレータを接続し，各フィルタ間の相対的利得を肉声母音の分析結果に似るように設定すれば，母音らしく聞こえる音を電気的に合成できることは前述した．この様な音はある特定の母音として聞こえる十分条件を満たしていると考えられるが，この種の音を多数作ってランダムに配列し，5母音の中のどの音韻に聞こえるかというテストをすると，最初に予期した程の認識率は得られない．音声の韻質（認識率の高さ）を正確に早く評価できる方法はまだ確立されていない．合成音の何処が悪いから別の母音に聞こえてしまうのかを明確に定量的に説明するのは難しい問題である．これらの音を何度も聞いて認識率を測定することが通常行われるが，この様な方法では合成音を聞く順序を終始ランダムにしないと，あれは“あ”の音だという先入観が入って来て認識率のスコアが固定してしまう．人間の認識系は本質が全くといってよい程未解明なすぐれた学習能力を有するが，この能力は認識率の測定実験においては，誠に具合が悪いのである．同様なことは子音についてもいえる．特定の分析方法を用いて音声を分析し（特徴パラメータを求め），その分析結果と同じパラメータの値を生ずるように，分析と逆の過程を使って音声を合成すれば，もとの声と似た声を発声させることは出来る．合成は分析の逆過程と考えられるが，逆過程の構成し易い分析方法もあればそれが難しい方法もある．たとえば，前者の例に帯域濾波器が，後者の例にゼロ交叉波がある．ターミナルアナログ合成器は，線形予測分析（後述）が登場するまで最もよく利用された合成用機器であった．音節程度の合成に関しては，現在ではアナログ時代にターミナル・アナログ合成器により得られた特徴パラメータとその時間的変化に関する知識を利用して，計算機によるディジタル・シミュレーションで合成が行われているが，アナログ合成時代に得られた結果を現在追試してみても，必ずしも良い声（韻質，声質，明瞭度，自然性）が生成できないことが多い．一寸したパラメータの設定条件の違いが大きく影響して，追試が追試になっていないことがあり，パラメータの値の設定に関する検討を最初からやり直さなければならない場合も多い．

前節までに述べた単音節程度の合成は，昭和40年代の始め頃まで続いた．しかしその頃の声は何とか特定の音節として聴き取れるが，やはり機械的な声だという程度のものから，これは /ga/ ですよといわれれば /ga/ と聞こえるといった程度，さらにもっと悪くなると，合成した当人は /ma/ のつもりでも聞いた人はどうしても /na/ にしか聞こえない程度のものまで色々あった．このような事情もあって，当時は研究会の会場で，合成した声を披露するのが1つの慣習のようになっていた．昭和45年に開催された万博に，比較的簡単な文章音声の合成器も出品されたが，長い間聞いていると気分が悪くなるという人も多く，人間の声は普段意識してはいないが，たとえダミ声といわれる人の声でも何と美的な感じのする芸術品であろうかと今更ながら感じ入ったものであった．

8．文章音声の合成

文章音声の合成は通常2つの場合に大別できる．

(1) 入力した声とそっくりの声，または若干それからは変歪してはいるが，特定の目的のために充分な基準を満足する音声を合成する場合（分析―合成といわれる）．

(2) 利用できる自然音声（肉声）をデータとして利用し，それを分析することにより特定の音韻，単音節，単語，句，節，文章の合成に有用な一般的法則を求めておき，この法則だけに基づいて離散的入力（たとえば仮名文）から所望の目的に合致した声（明瞭度，自然性，個人性の中のいくつか，場合によりそのすべてを満足するような）を合成する場合（法則合成といわれる）．

連続音声の分析――合成に関しては，後述のPARCOR 方式及びそれを発展させた LSP 方式により，ほぼ実用化の目処が立つ所まで到達した感があるが，法則合成に関してはまだまだそこまで至っていない．音声分析方法の優劣は仮定するモデルの精粗さに依存するといわれている．従って分析の逆の過程としての合成も同じく仮定するモデルに依存する．そして仮定されたモデルの欠点を修正する方法に関しては分析とか合成の研究結果は充分な情報を与えてはくれない．この本質的な問題点が，今後研究を推進していかなければならない法則合成における最も困難な問題の一つである．

9．PARCOR 方式の発明

昭和42年の春の音響学会で音声の分析―合成系の研究に1つの大きなインパクトを与える研究結果が発表された．“統計的手法による音声スペクトル密度とホルマント周波数の推定”という題目の研究で，これは1968年の ICA（国際音響学会）や Speech Symposium Kyoto, 1968でも発表された（通研，板倉）．この研究が BTL（ベル研）の Schroeder 博士の目に止った．これが，今日 PARCOR 方式として知られる音声の分析―合成方式へと発展してゆくことになる．当時 PCM，DPCM のような音声情報圧縮の研究において，線形予測という考え方は既に存在し，ベル研ではこの方面の研究を発展させて線形予測分析法（LPC 分析）を確立し，音声の重要な特徴パラメータであるフォルマント周波数を周波数分析という過程を通らないで求め得ることを明らかにした．それまではフォルマントは音声の周波数スペクトル包絡を表現するための有用なパラメータとして認められていたが，周波数分析の過程（帯域濾波器または高速フーリエ変換を使用）を経て求める他はなかった．ただこれ等の方法は分析過程と合成過程が表裏一体的に対応していないため，分析の結果得られたパラメータを用いて合成するという操作が簡単には行えず，入力音声と酷似した音声を合成するには問題があった．ところが，板倉式の推定も，LPC 分析も(1)音声という時系列の p（10～14程度）次遅れまでの相関係数を求めておき，(2)これを係数とする連立一次方程式（正規方程式あるいは Yule-Walker 方程式とも呼ばれる）を解くことにより線形予測係数（αパラメータといわれる）を求め，(3)この α を係数とする1元 p 次方程式を解くことによりフォルマント周波数を求め得る道を開いた（その後の発展の詳細は文献(7)を参照）．この方法の開発により，かなり原肉声に似た声を再合成することが出来るようになった．

そして，この方式による合成音は「この声なら肉声標本の採取という難点はあるが，何とか商売になる」という認識を企業にもたせるのに充分であった．ここで感慨を新たにするのは，一つの学問的分野における発展は不連続に起り，それは既成の概念に捉われない斬新かつ大胆な着想によることが多いという事である．また，特定の学問分野において慣習的に使われている手法を他の分野に移植することも大切だということを示している．この PARCOR 方式は元来分析合成方式として，入力の分析，パラメータの低 bit 伝送そして受信側における高忠実度再生を目的として開発されたものであるが，その声の質の優秀さのため，現時点では音声による応答，指令等に使用する音声合成のためにも広く使われるようになっている．

10. マン・マシン・システムの構成

最近は音声による応答，指令といっても，大して人目をひかない段階にまで一般化したようであるが，それが LSI 製造技術の進歩に支えられていることを忘れてはならないであろう．PARCOR 方式の他にも，特定の使用目的だけを考えれば，もっと安直な色々の音声の合成方式が開発されている．PARCOR 方式の出現と期を同じくして，音声の認識に関する研究に拍車がかかり，音節程度の認識から単語，さらには文章の認識（ないし理解）の研究は加速度的に進展してきた．音声の認識は合成とペアになって始めて有効な場合が多いため，企業ベースに乗る音声合成装置の出現に伴って，音声認識装置の企業化が熱心に考えられるようになった．この場合，音声の認識が学問的にはもう一ランク上のパターン認識の一つの分野として捉えられるという考え方や，音声情報をその受信者が人間だけに留まらず，計算機等の広義の情報処理システムへの入力手段に使用する“man-machine system の構成”という考え方が大きい影響を与えている．パターン認識の立場からすれば，観測，特徴抽出，識別に関する一般的な考え方が音声分析の分野においても当然通用するし，また実際それらに関する研究成果は充分利用されている．しかし，その他に音声に特有な問題としては次のようなものがある．

(1) 音声は時系列であり，同一の対象（たとえば単語音声）でも発声時間長は不同で，これを救済する方法（例えばダイナミック・プログラミング）が必要である．

(2) 特徴パラメータとして，声道（声帯から唇までにわたる空洞）の形のような物理量と直接に結びついたパラメータを使用すべきかもしれない．

(3) あるシンボルと対応した声の要素的部分（音韻といわれる）を連続させて発声すると，その接続関係により，要素的部分が孤立して発声される場合とはその形態が特徴パラメータの時間的変化からみて変歪して来る調音結合といわれる）．

個々の音韻をどの程度の自信をもって認識しておいて，次段の単語等の認識部へ譲り渡せばよいかは難しい問題である．それは未習熟な外国語のヒアリング能力を改善するにはどのような問題を解決しなければならないかを考えればある程度理解ができる．習熟してしまった言語では意識しないが，単語の認識を例にとっても，ある時間的微少部分を単位として認識し，それらの結果を関連的かつ総合的に活用して，まとまった単語として認識するに至る過程は，認識に関与するサブシステム間に複雑にフィードバックのかかった，そして何種類もの知識工学的過程を階層的に利用した総合システムと考えられる．このような総合システムとしての捉え方は，一つには計算機の顕著な能力アップに支えられて始めて可能になってきつつある．

11. オフィスオートメーションから人間—計算機共同体へ

現在流行中のＯＡの中核はワープロとマイコンであるといわれている．これらの機器に代表されるＯＡは昨今圧倒的な社会的ニーズとなりつつある．このような状況の下で誰もが高度な情報機器を操作して仕事をするということになると，人間と機械の間の情報伝達の重要性が増してくる．その場合，人間の側からみて最も抵抗が少なく，かつ機械系に対する高能率な情報

媒体は音声以外にはない．これからさらに進んで，計算機の高度な計算能力と，計算機側での実現が難しいいわゆる heuristics を得意とする人間の共同体（man-computer symbiosis）の実現のためには，自由に話し声が出せる音声合成，不特定話者の音声を自由に認識し理解できる音声認識の問題が解決されなければならない．

12. むすび

音声研究の発展を経時的に述べ，それらに関連して研究上の問題点のいくつかを説明するつもりであったが，実際書いてみると，研究の発展を記述する場合の時間的密度にかなりの差が出来てしまい，ある所は詳しく，ある所は粗っぽすぎる結果になって誠に感心しない出来ばえになってしまい恐縮している．若い頃に取りついた研究テーマが停年になるまで種切れにならず，かつ晩年になって花が咲けば一応満足せねばならないといわれている．この観点からすれば当方も一応満足せねばならない気もする．実際に音声の研究を手掛けてみると，なかなか奥行きは深く音声を媒介とする人間と機械との情報交換，マンマシンシステムの構成，さらには man-machine symbiosis の構成という目的のためには，まだまだ解決すべき多くの問題が残っており，考え方によれば，音声の研究は現在でもまだその緒についたばかりということができるかも知れない．

参考文献

1）角所：“音声研究の現状と問題点，システムと制御”，Vol. 33，No. 2，pp. 76～84（昭54—02）．
2）角所：“音声の合成，認識関連システム開発上の問題点”，情報処理学会関西支部第 9 回予稿集，pp. 19～33（昭55—11）．
3）角所：“音声の合成と認識”，社団法人大阪工業会技術講演会予稿，pp. 1～17（昭56—11）．
4）角所：“音声合成の基礎技術の現状と課題”音声認識と合成装置の実用化とその応用技術 pp. 37～83，日刊工業（昭56—12）．
5）J. L. Flanagan：“Speech analysis, synthesis and perception”，Springer（1972）．
6）牧田：“電気的合成法による母音および子音の研究”，日本音響学会誌，Vol. 5，p. 1(昭19—05)
7）中田：“音声”，コロナ社（昭52）．