21世紀に向けての音声合成の技術展望

東京大学新領域創成科学研究科
広瀬 啓吉

【音声合成の歴史】

音声を人工的に生成することへの興味は古くからあり，すでに，今から200年も前に，声道に見立てたゴム管を手で変形して母音を合成しようとする試みが行われていた．このように，当初は理学的な興味の対象であった音声合成が，ここ20～30年の半導体集積回路技術の急速な進歩に支えられて工学的な研究対象となり，最近では，パーソナル計算機のソフトウェアとして一般的になりつつある．任意のテキストを音声化するテキスト音声合成の試みは1960年代の後半から始まり，1970年代の後半には実用的な英語音声合成システムが開発された．その後，日本語や各言語についてテキスト音声合成システムが開発されているが，その合成手法は，現在では，当初のものとは大きく変わってきている．すなわち，当初は人間の音声生成過程を模擬する規則合成が目標とされ種々の研究が行われたが，現在では，自然音声波形データをもとに，これを統計的処理によって編集して音声波形を合成するコーパスベースの音声合成が主流となっている．

ここでは，音声を合成するために必要な処理を解説するとともに，これらの音声合成手法について概説する．さらに，コーパスベース音声合成の利点とともに問題点をみることによって，音声合成の将来を展望する．

【音声合成の対象と手法】

音声合成の応用として，携帯端末で電子メールの内容を読み上げたり，盲人のために文書を読み上げるといったことが行われている．これは，テキストを入力として音声を生成する技術であり，テキスト音声変換（Text-to-Speech Conversion）と呼ばれる．パーソナル計算機用の音声合成ソフトウェアはもっぱらこれであり，実用化が進んでいる．一方，最近，音声によって応答する対話システムの開発が盛んであるが，このようなシステムでは，ユーザに伝えたい内容（概念）を文書化するとともに音声として出力することが求められる．これは，概念音声変換（Concept-to-Speech Conversion）と呼ばれる．概念の表現（意味表現）形式としては，概念依存表現形式☆1のような言語の種類によらない抽象度の高いものの方が，文生成，音声生成の自由度の観点からは望ましいが，取扱いは困難となる．したがって，現状の音声対話システムでは，抽象度のない文（表層文）自体，あるいはそれに近い表現を利用することが多い．その場合，

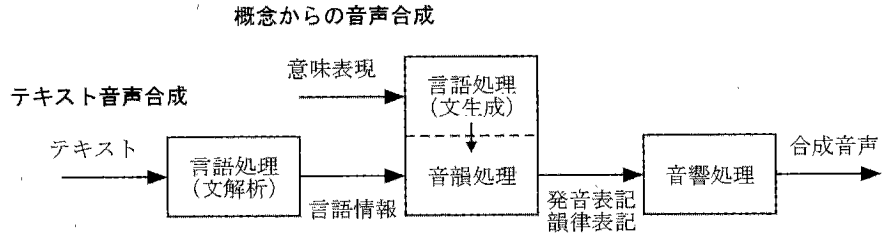

図-1　テキストからの音声合成と概念からの音声合成の処理の流れ

市販のテキスト音声合成ソフトウェアを利用して音声生成することがよく行われているが，高品質音声合成のためには概念から一貫して音声を生成するプロセスが望まれる．テキストあるいは概念を入力として音声を合成する場合，図-1に示すような言語処理，音韻処理，音響処理といった各処理が必要となるが，言語処理の主要プロセスは，前者では文解析，後者では文生成であって大きく異なる．具体的な処理内容は個々の合成システムにより細かい違いがあり，コーパスベースの音声合成では，段階の区分が明確でない場合もある．

高品質の音声を合成するためには，もちろん，各段階の処理の高度化が必要であるが，ここでは，音韻処理の結果得られる読みあるいは韻律の情報から音声を生成する音響処理について説明する．一般に，音声の特徴は，個々の音を表現する分節的特徴と抑揚やリズムを表す韻律的特徴とに大別されるが，両者では処理の様子が異なる．ここでは，主に前者を念頭において話しを進め，後者については次にまとめて述べる．

音韻処理で得られた読みから対応する音声の特徴パラメータ時系列を生成する手法として，まず音声生成過程に基づく規則合成が考えられる．この手法は，発話速度や声質などの柔軟な制御が可能，必要なメモリ量が小さい等の点からは理想的であるが，このような生成過程に関する知見は，母音連鎖についてはある程度得られているものの，子音が含まれた連続音声に関しては未知の点が多く，高品質音声合成の実現は現状では困難であるといわざるを得ない．これに対し，特に音声の分節的特徴が音素や音節程度の小さな単位で取り扱うことが可能であることに着目し，自然音声の分析結果をもとに作成した音節等の特徴パラメータパターン（あるいは直接に音声波形）をデータベースに蓄積しておき，合成に際しては，読みに従って必要なものを取り出して接続することが一般的に行われている．

合成単位の接続は合成音声の品質低下の大きな要因となり，この観点からは，なるべく大きな単位での合成が望ましいが，反面，蓄積単位数の著しい増加を招く．蓄積単位の作成は，自然音声を切り出すことによって行うが，その発声単位は，調音結合☆2の影響を考慮した場合，母音-子音-母音の連鎖以上が必要となる．話者の発声のしやすさからは単語程度が望ましい．単語音声を利用する場合，合成対象と長い区間で一致する音素系列が単語音声データ中に得られる可能性があり，そのような音素系列を利用する方が画一的な単位を利用するよりも効率的であり，音質もよい合成音声が得られると考えられる．このような観点から，複合音声単位の合成手法が提案されている[1)]．

音声波形を合成する方式としては，音声波形をそのまま時間領域で処理する波形編集方式，周波数領域で処理する分析合成方式と，音声の生成過程を周波数領域で模擬するターミナルアナログ方式，直接，調音器官の運動として模擬する声道アナログ方式がある．このうち，ターミナルアナログ方式は，英語を対象とした最初のテキスト音声合成システムに採用され，欧州の各言語や日本語などを対象としたシステムも作成された．規則合成という観点から有望視されたが，現状では音質の向上に限界がある．データベースを活用する統計的手法では，自動的な処理に適した前2者が有利であり，特に，音質の点から最近では波形編集方式が音声合成の主流となっている．

波形編集方式では，まず，自然音声波形から，合成単位の音声波形を，前後の音素環境，韻律的情報などの情報とともに切り出し，波形辞書（合成の素材）として蓄積する．次に，合成時には，音韻環境がテキストの音韻処理結果と最も合致する波形を選択して接続する．自然音声波形そのものを用いるために個々の合成単位の品質は高いが，韻律の制御に難点があり，従来は，駅の案内や時報の案内のように，数種の単語音声を定形文の中に

☆1 Conceptual Dependency表現形式．動詞を少数の基本的な行動の組合せで表現した上で，各格の関係として概念を記述する．

☆2 個々の音を連続的に発声した場合，その音響的特徴に隣接の音が影響すること．これによる音響的特徴の変形は相当に大きく，音声認識を困難にしている要因ともなっている．

埋め込んで出力するといった限られた用途に用いられていた．これに対し，波形に付与したピッチマークを基準とし，基本周期単位で波形窓掛けを行い，合成時の基本周期に従って再配置するTime Domain Pitch Synchronous Overlap and Add法（TD-PSOLA法）が開発され[2]，高い品質が得られたことから，合成に広く用いられるようになった．

TD-PSOLA法によって基本周波数の変更が可能となったが，波形操作に際してスペクトルの連続性が保証されないため，品質上から許される操作の範囲は狭い．このような観点から，波形操作を周波数領域で行うPSOLA法として，FD（Frequency Domain）-PSOLA法がある．これは分析合成方式と似ているが，分析パラメータレベルで接続してから合成しないため，パラメータの連続性が大きな問題とならない．したがって，線形予測分析など，音声の生成モデルを前提とした分析の他，音声のスペクトル包絡を現象的に再現する種々の分析手法の利用が考えられる．実際，スペクトル時系列を正弦波成分の変化として表す正弦波モデリングなどでよい結果が得られている．

TD-PSOLAにしてもFD-PSOLAにしても，合成単位を加工・変形することが合成音声の品質劣化に結びつくとの観点から，波形の加工を基本的に行わないCHATRと呼ばれる合成方式が開発され，品質の高い合成音声が得られている[3]．基本周波数を変更しないかわりに，韻律の観点からバラエティーに富む文音声データから合成単位を選択している．この方法は，ピッチマークの付与などの誤差要因が少なく，自動化に適している．後述するHMMによるセグメンテーションと単位選択の客観的評価の導入により，音声データベースさえあれば，各言語をはじめ，種々の合成音声が比較的簡単に得られるようになっている．

このように，ある話者の音声データを用いることにより，その話者らしい音声を合成することが行えるようになった．しかしながら新しい話者に対しては，再度音声データを取り直す必要がある．そこで，取直しを行わずに，新しい話者の少量の音声データをもとに，声質の変換を行う声質変換技術が研究されている．また，最近では，画像のモーフィングにヒントを得て，話者Aから話者Bへの変換を連続的に行う音声モーフィングという考え方が登場してきた[4]．

話者Aと話者Bの音声の特徴空間内での対応をもとに話者の変換（適応）をすることは音声認識において従来から行われていたが，これを利用して話者A，B間の声質変換が可能である．声質を決定する要因としては，このような声道特性（分節的特徴）に起因するものの他に，音源特性（韻律的特徴）に起因するものがあり，後者に関する処理も必要である．音声モーフィングは中間的な特徴の音声を生成する必要があり，たとえば，スペクトル包絡について，話者A，B間の対応をとり，連続的に変化させることで実現できる．

声質変換にしても音声モーフィングにしても，音声の特徴を大きく変形するために，現状では得られる音声の品質に問題が多い．今後の研究成果が待たれる点である．

種　類		内　容
言語情報	語義情報	アクセント型，声調型
	統語情報	係り受け
	意味情報	疑問文
	談話情報	話題・焦点，段落
パラ言語情報		態度，意図
非言語情報		感情，個人性

表-1　韻律によって伝達される情報

【音声合成と韻律】

分節的特徴とは異なり，音声の韻律的特徴は，基本的にはテキストには明示されていない．しかしながら，それが情報伝達に果たす役割は表-1のように多岐にわたっている．したがって，人間の発声らしい合成音声を実現する上で，韻律の適切な制御は重要な課題となっている．

韻律的特徴はテキストに明記されていないので，テキスト解析によってそれを導き出す必要がある．しかしながら，そのためには辞書引きによる単語のアクセント型の決定だけでは不足であり，高品質の合成音声実現のためには，文の統語構造や文章の談話構造の利用が不可欠である．連続的な文音声を発声した場合，その韻律的特徴は構成単語のそれとは大きく異なる．たとえば，連続音声中では，単語の基本周波数パターンのアクセント成分が，隣接韻律語のアクセント型，文の統語構造，文章の談話構造の影響を受けて融合，増大，減少，消滅などの複雑な変化を示す．また，文のイントネーションに対応する基本周波数パターンのフレーズ成分も統語構造と深くかかわっている．このような言語情報と韻律的特徴との関係を規則として表現するためには，両者を明確に結びつけ得るモデルが必要である．基本周波数パターンをフレーズ成分とアクセント成分との対数周波数軸上での和として表現する生成過程の重畳モデルが提案され，これを用いて韻律規則が作成され，アクセント・イントネーションの観点から優れた品質の合成音声が得られている[5]．図-2に合成された基本周波数パターンの一例を示す．

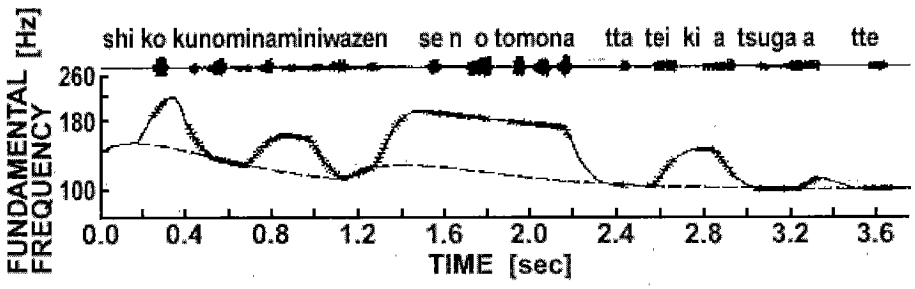

図-2　重畳モデルに基づく基本周波数パターンの合成例

この韻律規則は，音声の分析結果をもとに，人間が発見的に構築しており，たとえば，朗読以外の発話スタイルについては，新たに分析を行って規則を作成するための手間が大きい．また，高次の言語情報との関係を記述する規則であるため，分野を限定しない一般のテキスト音声合成では利用しにくいという問題がある．このような観点から，数量化手法，さらには音声認識で一般的な隠れマルコフモデル（HMM）やニューラルネットワーク（NN）を用いて音声データから自動的に韻律規則を導出することが行われている．このような統計的手法では，テキスト解析で比較的容易に得られる言語情報（たとえば品詞情報）に限定して韻律との関係を記述することができ，テキスト音声合成に適した規則が作成される．特に，音素の持続時間に関しては，種々の要因が関係しているために発見的手法では定式化が困難で，数量化手法が有効であり，早くから研究が進んでいる．数量化手法としては数量化I類による線形回帰モデル，要因の組合せによってデータを逐次的に分割する回帰木によるモデルなどがあるが，匂坂によりよくまとまられているので，そちらを参照されたい[6]．

重畳モデルは声帯の物理的な特性を考慮した優れたモデルであるが，観測される基本周波数パターンから解析的に抽出することができず，逐次近似法によっている．このため，統計的手法の基本となるデータベースの整備が進んでいない．HMMやNNを用いた場合，重畳モデルのような生成モデルを明に必要としないため，多量のデータを用意することが容易という利点がある．HMMで基本周波数パターンを表現する場合，基本周波数のない無声区間，無音区間の取扱いが問題となる．これに対し，有声区間の基本周波数と無声区間のピッチなしの情報を，異なる空間での事象と考え，それを確率分布で表現するHMMの枠組みが提案されている．基本周波数に影響を与える諸要因を決定木により整理した上でクラスタリングを行い，また，HMMの特徴パラメータに基本周波数の変化量を導入することによって，任意の文章に対して，適切な基本周波数パターンが生成されている[7]．

韻律を記述する指標として単語境界情報，抑揚情報等を表記するTones and Break Indices（ToBI）が英語について開発され，それをもとに各言語について韻律データベースの作成が行われている[8]．これも人手をまったく介さないというものではないが，重畳モデル分析よりもデータベースの規模の拡大は容易であり，これをもとに韻律の合成規則を作成し，得られた指標を用いて，直接あるいは重畳モデルを経由して基本周波数パターンを合成することが行われている．

【音声合成と統計的手法】

計算機の性能向上によって多量のデータを短時間に取り扱うことが可能となり，これが，音声合成における統計的手法の本格的な導入を可能とした．このようなコーパスベース音声合成では，合成音声の物理的に定義可能な客観的評価基準のもとに，処理の自動化が進められている．韻律的特徴の統計的処理による生成に関しては，前章ですでに概説したが，ここでは，音声認識で成功したHMMを用いて合成を行う手法を中心に説明する．

合成に用いる音声データには音韻情報があらかじめラベリングされている必要がある．当初は，これは人手で行われていたが，非常に労力のかかる作業である．発話内容が表記されていれば，HMMを用いて音声のセグメンテーションとラベリングを自動的に行うこと（強制アライメント）が可能である．まず，文脈独立の音素HMMを用いて，モデルの学習とセグメンテーションを交互に行って，HMMとセグメンテーションを最適化した後，次に文脈依存の音素HMM（Tri-Phone Model）の学習を行う．この際，クラスタリング手法によりモデル数の削減を図る．合成に際しては，音素HMMあるいはその各状態ごとに平均値に近い特徴の波形データを選択して用意しておき，PSOLA法により接続して合成する．

人間がラベリングした場合と比較して遜色のない品質の音声が得られている[9]．波形の接続によらず，音素HMMを接続してパラメータ時系列を生成し，音声合成をすることも行われている．メルケプストラム係数とそのΔパラメータ，Δ^2パラメータを用いた合成システムが報告されている[7]．この合成方式は，前章のHMMによる基本周波数のパターンの表現手法と組み合わされ，HMMによる一貫した合成の枠組みに発展している．パラメータ時系列を生成することは，分析合成方式ということであり，現状では，得られる音質に一定の制約があるが，音声認識で確立している手法がそのまま適応可能という利点がある．たとえば，話者適応手法によってHMMを変更することにより，声質変換が容易である．

多量の音声データベースを用いた場合，そこから合成単位を適切に選択し得るか否かが，音質の良否に大きく影響する．一般には，合成の際の音韻環境との一致度，隣接単位間のスペクトルの不連続の程度，合成の際の基本周波数との一致度などを指標として，合成単位を自動的に選択する．指標間の重みについては，自然音声との歪み最小を基本とした学習が行われる．

【これからの研究】

コーパスベース音声合成の品質向上のためには，人間の主観的評価に対応した客観的評価の枠組みの確立が求められるが，この観点からの研究は進んでいない．最近，音素の持続時間の数量化モデルに関し，モデルと観測データの誤差評価に人間の聴知覚特性を導入することが行われ，良好な結果が得られたという報告がある．今後，このような観点からの研究が合成単位選択の指標についても重要となろう[1]．

現在のテキスト音声合成では，朗読調の高品質音声を合成することに研究の主眼が置かれているが，将来的に音声対話システムに合成音声を利用することを考えると，対話調をはじめとしてさまざまなスタイルの音声を概念から合成する技術の開発が求められる．このようなスタイルの違いは主に韻律的特徴によるものである．また，概念合成では韻律の生成に高次の言語情報を利用することが可能である．このような観点から，今後，韻律に関する重要度が増すと考えられる．これに対応して，韻律データベースの整備が切に求められる．

現在，音声規則合成の研究は，積極的には行われていないが，将来的には，声質変換など柔軟な音声合成実現のために不可欠と考えられる．新しい音声分析法の開発とそれに基づく音声生成モデルの構築といった地道な努力が求められる．なお，従来から，平静な音声をもとに感情を付与した音声を生成するなどの研究が行われているが，韻律的特徴の操作のみでは限界があり，分節的特徴の制御が求められる．これは，音声規則合成と関係するところが多い．

将来の音声合成の発展を考えると，その応用範囲の拡大が望まれる．現在の主要な応用はテキストの読上げであるが，品質がさらに向上することによって，アナウンサーの部分的な代行が可能となろう．また，種々の音声応用システムでの利用が考えられる．すでに，音声翻訳システムの出力として利用されているが，種々の音声応答システムやロボットでの利用が拡大すると考えられる．また，アニメーションへの音声付与，映画の吹替え等への利用なども近い将来可能となろう．特に，アニメーションは，不完全な合成技術でも積極的な利用が期待される有望な応用分野である．今まで，音声合成は人間の音声になるべく近い音声の実現という目標に向かって研究や開発の努力がされてきたが，それがかえって合成音声の粗を際立たせる結果となっていた．アニメーションでは，必ずしも人間に近い音声であることは要求されず，音声規則合成の利用も可能となり，合成技術の発展に結びつくと期待される．一方，吹替えには，品質の一層の向上が求められるものの，声優の特殊技能に支えられている現状を考慮すると，今後が有望な分野である．画像処理により補足した口の動きに対応して音声を合成することが求められるが，このような音声と口の動きの同期に関して，カリフォルニア大学サンタクルーズ校などで研究が進んでいる．

以上，音声合成技術について解説と展望を行ったが，技術的な詳細については，解説論文をあげておくので参照されたい[10]．

参考文献

1) 匂坂芳典: コーパス・ベース音声合成, 日本音響学会講演論文集, 1-3-10, pp.197-200 (Sep. 1999).
2) Moulines, E. and Charpentier, F.: Pitch Synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones, Speech Communication, Vol.9, pp.453-467 (1990).
3) ニック・キャンベル, アラン・ブラック: CHATR: 自然音声波形接続型任意音声合成システム, 電子情報通信学会技術研究報告, SP96-7, pp.45-53 (1996). (http://www.itl.atr.co.jp/chatr).
4) 阿部匡伸: 基本周波数とスペクトルの漸次変形による音声モーフィング, 電子情報通信学会技術研究報告, SP96-40, pp.25-32 (1996).
5) Hirose, K. and Fujisaki, H.: A System for the Synthesis of High-quality Speech from Texts on General Weather Conditions, IEICE Trans. Fundamentals, Vol.E76-A, pp.1971-1980 (1993).
6) 匂坂芳典: コーパスベース音声合成, 信号処理, Vol.2, No.6, pp.407-414 (1998).
7) 徳田恵一: 隠れマルコフモデルの音声合成への応用, 電子情報通信学会技術研究報告, SP99-61, pp.47-54 (1999).
8) ニック・キャンベル: Tones and Break Indices (ToBI) システムと日本語への適用, 日本音響学会誌, Vol.53, No.3, pp.223-229 (1997).
9) Donovan, R. and Woodland, P.: Automatic Speech Synthesizer Parameter Estimation Using HMMs, Proc. ICASSP 95, pp.640-643 (1995).
10) 広瀬啓吉: 音声合成技術, 情報処理, Vol.38, No.11, pp.984-991 (Nov. 1997).

(平成12年1月5日受付)