道しるべ：音声合成研究への招待 ―自由な合成の実現に向けて―

広瀬啓吉 hirose@gavo.t.u-tokyo.ac.jp
東京大学　新領域創成科学研究科

音声合成の歴史

人工的に音声を合成することの試みは，大変古くからあったが，任意のテキストを音声化することは，半導体集積回路技術，計算機技術の進展と歩調を合わせ，1960年代の後半から始まった．1970年代の後半には本格的な英語テキスト音声合成システムがMITで開発され，故D. Klatt博士によるKlattalkで実用の域に達した．これは，人間の音声生成過程を模擬する規則合成を目標としたものであったが，日本では自前のPARCORなどの優れたVocoder技術を土台として素片接続型の音声合成がNTTを中心に研究された．音節の数が英語と比べて少ない日本語では，音節を合成の基本単位としたシステムの構築が容易という事情もある．その後，合成音声の品質を自然音声に近づける努力が多くの研究者によってなされ，多量の自然音声波形の素片から最適なものを選択して接続する，波形編集コーパスベース音声合成が開発され，現在の主流となっている．米国での音声合成初期の歴史についてはKlatt博士が詳しくまとめている[1]．また，S. Lemmettyが修士論文で音声合成をまとめており，〈URL:http://www.acoustics.hut.fi/~slemmett/dippa/contents.html〉でKlattalk他の合成音声を聞くことができる．

本稿では，まず音声合成が最も広く利用されているテキストからの音声合成で，テキストにどのような処理が行われて音声が出力されるのかを概説し，その後，実際に音声を合成する手法について，若干詳細に説明する．さらに，研究の動向と今後の研究の方向を述べ，これから音声合成の研究を始めようと考えている読者の一助としたい．

テキスト音声合成

仮名漢字混じりの日本語文章を入力として音声を合成するテキスト音声合成（Text-to-Speech Conversion）は，パソコン用のソフトウェアとしてすでに一般的であるが，実際にどのような処理が必要なのかみてみよう．これは，我々が文書を読み上げる過程を考えてみるとよく分かる．我々は，テキストの内容を理解し，実際の読み方を考え，声に出すという複雑な作業を無意識のうちに瞬時に行っている．これを機械（計算機）で実現するとなると，言語処理，音韻処理，音響処理といった，それぞれ内容の異なる処理を一括して行うことになる．このため1人でテキスト音声合成の研究を進めようとすると結構大変である．現状では，各処理とも人間のように正確に行うことはできず，得られる合成音声の品質を抑える結果となっている．

まず，言語処理であるが，単語の境界が明示されていない日本語では，形態素解析によるテキストの単語単位への区分が重要である．これは，入力文と辞書項目との照合（最長一致など）を基本とした手法によりかなりの精度が得られるが，100％の精度を得るためには人間がやっているような知識を用いた意味解析が必要となる．たとえば，「東海上」が「東＋海上」であるか「東海＋上」であるかは，いまどこの話をしているかの知識がないと決められない．形態素解析での誤りは，まったく異なった読みになることが多く，合成音声の品質への影響が大きい．また，一般に任意のテ

キストを対象とするため，辞書にない語（未登録語）の問題も生じる．人間でも人の名前や略号などをどう読んでいいのか困って，ちょっと曖昧な発音をしたり，周りの人に読み方を聞く等，適当にごまかしたりしているが，テキスト音声合成ではこのような融通が利かない．市販のテキスト音声合成ソフトウェアなどを用いて合成した音を注意して聞くと，時々，おかしな読み方をしていることに気づくであろう．

高品質の音声合成には，統語解析，談話解析といった高次の言語処理も必要である．人間は，句の先頭でピッチを上げるなどの統語構造に対応したイントネーションをつけて発声する．イントネーションは人間の発声の現象であるので，必ずしも深い統語境界で大きなピッチ上昇があるわけではないが，統合構造を無視して音声合成を行うことはできない．品詞程度の情報でも，ある程度の性能は得られる．

形態素解析された段階で，テキストの読みは辞書記述レベルでは得られているが，これは実際の発音を示すものではなく，音韻処理によって実際の発音に対応した表記にする．辞書の読みの記述にもよるが，たとえば，「目覚まし時計」は“と”の部分が濁音化して“めざましどけい”となる．数詞と助数詞の組合せでは，「1本，2本，3本」のように読みの変化は多様で，もちろん，数詞＋助数詞として辞書に記述しておけばよいが，作成の手間がかかりメモリを多く使うことになる．このレベルの表記は音素表記であって，まだ実際の発音と正確には対応しない．前後の音素環境を考慮して素片の選択をするコーパスベース音声合成では，ほぼ自動的に対処されるが，/g/の濁音化や母音の無声化など注意を要する部分も多い．

音韻処理は個々の発音のみならず，合成音声に適切なアクセントやイントネーションを付与することも行う．アクセント，イントネーションは，テキストには明示されていないので，統語構造などの言語情報から生成する．個々の単語（自立語）には固有のアクセント型があるが，たとえば「音声合成」のように「音声(起伏型)」「合成（平板型)」が1つの平板型として発音されるといったアクセント型の変形がある．また，自立語に付属語が付与された文節がどのようなアクセント型で発声されるかは，結構複雑である．さらに，個々の文節アクセントの大きさが前後の文節や統語構造等によって変化するという現象もある．このような現象を記述することによって，音声合成を行う準備が整う．

音韻処理で得られた情報に従って音声波形を生成する過程は音響処理と呼ばれる．最も音声合成らしい部分であり，手法も多岐に渡るので，次章にまとめる．

音声合成技術

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して音声合成を行うが，この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あげられる．この手法は，人間の発話をできるだけ模擬しようとしたもので，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期待されているが，実際は，生成過程に関する知見が限られており，高品質音声合成の実現は現状では困難である．これに対し，特徴パラメータで表した音声を，音素や音節程度の小さな単位でデータベースに蓄積しておき，必要なものを取り出して接続するコーパスベースの手法が一般的であり，比較的高品質が得られている．

パラメータから音声波形を生成する手法は，舌位置などの人間の調音器官に関連したパラメータを用いる声道アナログ方式，極・零点として声道伝達特性を表現するターミナルアナログ方式，LPC等の音声圧縮技術を利用する分析合成方式がある．このうち，前2者は規則合成に適した手法であるが，合成音声の品質は分析合成方式に劣る．前出のKlattalkはターミナルアナログ方式の最も成功した例である．分析合成方式は自動的な処理に適しており，コーパスベース音声合成に用いられる．

分析合成方式の利点は，基本周波数等の音響パラメータを比較的自由に変更できる点にあるが，得られる合成音声の自然性に問題が残る．そこで，パラメータレベルでなく波形レベルで蓄積した素片を接続する波形編集方式が多くの音声合成システムで採用されるようになった．波形編集方式自体は古い技術であるが，波形レベルでピッチを変更するPitch Synchronous OverLap-Add （PSOLA）という技術の成功によって，テキスト音声合成に利用されるようになった[2)]．

音声の特徴は，基本的に音源にかかわるアクセント・イントネーション等の韻律的特徴と，声道伝達特性にかかわる個々の音の分節的特徴に大別される．蓄積された素片パターンの接続は分節的特徴に対応したものであり，韻律的特徴の生成も必要である．韻律が情報伝達に果たす役割は**表**-1のように多岐に渡っており，人間の発声らしい合成音声を実現する上で，韻律の適切な制御が重要な課題となる．

すでに述べたように，文音声の韻律的特徴は，構成単語のそれとは大きく異なり，統語情報などの言語情報との関係を規則として表現することが必要である．

種類		内容
言語情報	語義情報	アクセント型
	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

このためには韻律のモデル化が有用である．音声の基本周波数の時間変化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[3)]．

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．このような観点から，数量化手法等の統計的枠組みで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．

研究の動向

ここでは，音声合成の研究動向の概略を紹介する．まず，現在の主流となっている波形編集方式についてみてみると，高い音質の合成が可能なデータベースの作成と蓄積波形選択が1つの関心事となっている．音声波形は人間の発声を収録して得るが，いつも同じように発声することを要求することには無理がある．異なりの大きい発声は合成音声の品質に悪影響があるのでデータベースから削除することが望ましい．また，データベースには多くの素片の候補があり，そこから最適なものを選択することが重要である．これらを自動的に行うためには，人間の主観的な評価に対応した物理的な距離の尺度が必要であるが，現在は，経験的に定義しているに過ぎず，今後，一層の研究が求められる．

PSOLA方式では，ピッチマークという波形操作の基準点を付与するなど，データベースの用意に手間がかかる．また，波形を操作するためにどうしても品質の劣化が避けられない．このようなことから，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，高品質が得られている[4)]．基本周波数を変更しないため，韻律のバラエティーに富む文音声データから合成素片を用意するが，どうしてもイントネーション・アクセントの自然さに問題が残る．PSOLAの波形加工による音質の劣化に対し，対照的な考え方として，周波数領域での操作を導入することも行われている．

HMM（隠れマルコフモデル）で音声の特徴を記述する手法が音声認識と関連して大変進んでいる．音節等の単位で用意したHMMを接続し，その尤度最大のパスを求めることによって音声合成を行うことができる．音声データを分析して得られた特徴パラメータさえあれば，学習によってHMMを自動的に構築できるため，特徴パラメータに韻律に関するものも含めておけば，分節的特徴と韻律的特徴の合成を1つの枠組みで行うことも可能である．また，ある個人の発声について構築したHMMを他の話者のものに変更（適応）する技術も音声認識で確立しているため，いったんHMMの合成システムを構築しておけば，ある話者の少数の発話のみを用いてその話者の音声に声質変換することが容易である．話者AとBの中間の音声といった音声モーフィングも実現できる．さらに，HMMを基本としているため，音声合成の経験がない研究者でも比較的取り組みやすいなどの特徴もあり，最近，注目を集めている．東工大，名工大のグループで一貫した研究が進められている[5)]．合成音声を〈URL:http://ktlab.ics.nitech.ac.jp/~yossie/TTS/demo〉〈URL:http://sp-www.ip.titech.ac.jp/research/demo-j.html〉で聞くことができる．IPAの独創的情報技術育成事業「擬人化音声対話エージェント基本ソフトウェアの開発」（代表：嵯峨山茂樹・東大）でHMM音声合成に基づいたテキスト音声合成フリーソフトウェアの開発が進んでおり，近々公開予定である．

音声に関する書籍では，残念ながら音声合成については，部分的に取り扱われているに過ぎない．情報処理学会，日本音響学会，電子情報通信学会などの関連学会の解説記事等で研究動向を知ることができる．筆者も「情報処理」の Vol.38，No.11，pp.984-991やVol.41，No.3，pp.277-281に解説記事を書いているので参照されたい．なお，日本音響学会の企画で，「音声合成の基礎」講習会がNTTで開発されたVisual Speech Creatorというテキスト音声合成ソフトウェアを利用して2000年と2001年に行われたが，今後の予定は未定である．

音声研究を行うものにとって有用な電子メールの組織がある．onsei-mail-ctl@m.aist.go.jpにコンタクトをとることによって利用できるが，有志の善意によって運用されているもので，主旨に沿った利用をお願いする．前述の「擬人化音声対話エージェント」に関する情報もここに流される予定である．

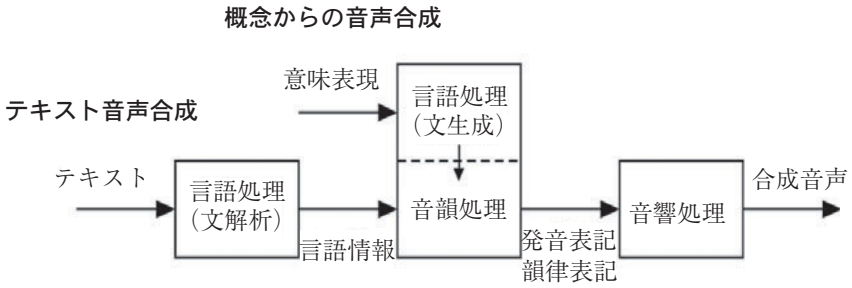

図-1　テキスト音声合成と概念音声合成

合成音声を聞いてみたい読者に有用なWebページとして，〈URL:http://www.ldc.upenn.edu/ltts/〉がある．各国の合成音声を聞くことができる．また，英語であるが，〈URL:http://www.naturalvoices.com/demos/index.html〉(AT&T) や〈URL:http://www.bell-labs.com/project/tts/ 〉(Bell Labs.) では，適当な文章を入力して合成音声を聞くことができる．

今後の展開

人間の音声言語活動は，テキストの読み上げより，自発発話が基本である．この後者に対応するものとして，概念音声合成（Concept-to-Speech Conversion）がある．この研究はあまり進んでいないが，ユーザとのやり取りを音声によって行う音声対話システムでは重要な技術となる．図-1に示すように，テキスト音声合成では言語処理として文解析が行われるのに対し，概念音声合成では文生成が行われる．高次の言語情報が文生成時に得られるため，それに対応した韻律の制御規則を構築することによって高品質の音声合成が可能である．

人間は，状況によってさまざまな調子で話す．これに対し，音声合成では，現在までもっぱら朗読調音声を合成することに研究の主眼がおかれてきた．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，対話調を始めとしてさまざまなスタイルの音声を合成する技術の開発が求められる．また，意図，態度，感情を合成音声で表現することも最近の関心事となっている．

コーパスベース音声合成では，適切なかつ十分なデータさえ与えれば高品質の音声が得られる．この観点から，音素，韻律の情報（ラベル）を記した音声データベースの構築が重要となっている．種々の調子の音声合成も音声データが用意されれば比較的容易に実現できる．ここで問題となるのがどのように音声データを用意するかである．特に韻律の観点からは手法が確立しておらず，今後の研究が重要である．現在，規則合成の研究は，高品質化が困難なため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．困難な道のりではあるが，音声合成の究極の目標といえる．コーパス音声合成では，音声データが不可欠であるが，規則合成では希望の声質，調子の音声が自由に得られることになる．若い人には意欲をもってこれにも取り組んでいただきたい．

参考文献

1） Klatt , D.: Review of Text-to-Speech Conversion for English, J. Acoust. Soc. Am., Vol.62, pp.737-793 （1987）.
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） 広瀬啓吉，藤崎博也: 音声合成とアクセント・イントネーション，電子情報通信学会誌，Vol.70, pp.378-385 （1987）.
4） 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）.
5） 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）.

（平成14年2月20日受付）