音声言語情報処理 33－4
（2000．10．27）

対話システムにおける音声合成

山下洋一
yama@cs.ritsumei.ac.jp

立命館大学理工学部情報学科
〒525-8577 滋賀県草津市野路東 1-1-1

音声合成は、人間と機械が情報交換するマルチモーダルな対話システムを構築するための重要な技術である。これまでに機械によるテキストの読み上げを実現することを目指して、多くの音声合成システムが開発されてきた。しかし、対話システムでの利用に適した音声合成システムを容易に利用できるような状況にはなっていない。本報告では、テキスト音声合成の処理について述べ、対話システムでの音声合成を実現するための課題と問題点について述べる。
キーワード：テキスト音声合成、テキスト解析、韻律生成、対話音声合成

Speech Synthesis for Dialogue System

Yoichi Yamashita
yama@cs.ritsumei.ac.jp

Department of Computer Science, Ritsumeikan University
1-1-1 Noji-Higashi, Kusatsu-shi, Shiga, 525-8577 Japan

Speech synthesis is a key technique for multi-modal communication between man and machine. Many speech synthesis systems have been developed for reading text by machine. The speech synthesis system appropriate to speech output module in the spoken dialogue systems is not yet available. This paper describes the outline of text-to-speech (TTS) conversion and some improvements required for realizing spoken dialogue synthesis.
keywords: text-to-speech, text analysis, prosody generation, spoken dialogue synthesis

1 はじめに

機械との対話を考えたとき、利用者に情報を提示する一つの手段として、音声合成が考えられる。音声合成の研究では、音声によるテキストの流暢な読み上げを目標として精力的に研究が行われ [1]、現在では多数の高品質のテキスト音声合成システムが市販されている [2, 3]。さらに、対話音声の生成を指向した音声合成の研究も多数報告されている [4, 5, 6, 7, 8, 43]。しかし、音声合成を用いた音声対話システムを構築しようとしたときに、使い勝手の良い音声合成システムは見当たらないのが現状である。

本稿では、テキスト音声合成の処理を概観し、音声対話システムに組み込んで利用できる音声合成システムを作成するための問題点と課題について述べる。

2 テキスト音声合成

テキスト音声合成 (TTS: text to speech) は漢字仮名混じり文を合成音声に変換する技術である。図1に示すように、形態素解析から波形生成までの一連の処理が行われる。

2.1 テキスト解析

2.1.1 形態素解析

テキストで表現された漢字仮名混じり文を音声を変換するには、読みやアクセント型などの情報が必要であり、これらは辞書を参照して獲得する。形態素解析では、辞書の見出しに登録されている形態素の単位に文を分割し、形態素の読み、品詞、アクセント型などの情報を得る。

2.1.2 係り受け解析

音素の時間長やピッチ(基本周波数)パターンは、形態素の品詞やアクセント型だけでなく係り受けの仕方によっても変わってくるため、形態素解析の結果をもとに係り受け情報を抽出する[10]。

2.1.3 読み付与

形態素解析の結果、読みが一意に決定できていない場合には、読みの選択を行う。また、形態素が「株式/会社」のように短く分割されている場合には、「かぶしき」+「かいしゃ」の読みを「かぶしきがいしゃ」とするなどの連濁の処理も行う[11, 12, 13]。

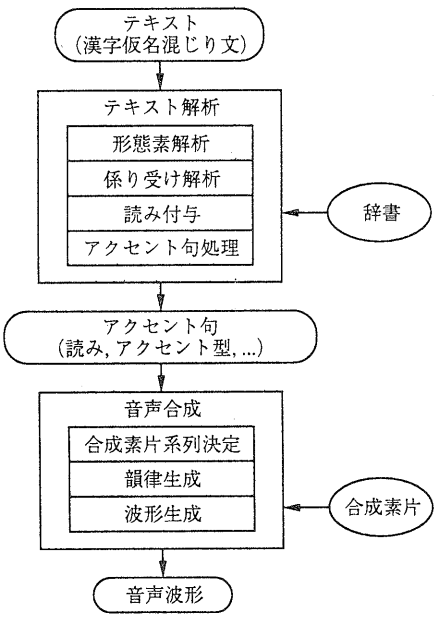

図 1: テキスト音声合成の処理の流れ

2.1.4 アクセント句処理

ピッチ(基本周波数)の時間変化パターンは、ピッチが一つの山型を示すアクセント句を単位として考えることが多い。一般に、アクセント句は複数の形態素から構成され、形態素解析から得られる品詞情報などの形態素の素姓をもとにアクセント句を決定する。

アクセント句はピッチ変化と密接に関連するアクセント型を持つ。アクセント句全体のアクセント型は、アクセント句を構成する形態素のアクセント型だけで決定されるのではなく、形態素によるアクセントの結合様式の違いによっても変化する。これらの知見をもとにアクセント句のアクセント型を決定する[14, 15, 16, 17]。そのためには、アクセント結合様式など複合語のアクセント型を決定するための情報を辞書にあらかじめ登録しておく必要がある。

2.2 音声合成

2.2.1 合成素片系列の決定

規則音声合成では、音素、音節(CV), VCV, CVCといった比較的短い単位を合成単位として用いる。音声の特徴を特徴パラメータとして、あるいは時間波形として合成システムに蓄えておき、それを接続することによって任意の発声内容を合成する[18, 19]。接続部での不連続を減らし、音素の多様な変形に対応するために、同じ合成単位に対しても複数の素片を統計的手法により生成する手法[21]や、合成素片の長さを可変にして適切な素片をデータベースから選択的に利用する手法[22]が提案されている。テキスト解析の結果得られた読みの系列に対して、長音化、無声化、鼻音化などの音韻変形を考慮して合成単位の系列を決定する。

2.2.2 韻律生成

音声の韻律的特徴は、ピッチ、パワー、継続時間長の三つの要因で構成される。ピッチパターンの生成では、ピッチの時間変化を表現するモデルと、アクセント型やモーラ長などの言語情報からモデルパラメータを決定する規則の作成が問題となる。ピッチパターンは、アクセント句に対応した局所的な起伏(アクセント成分)と時間とともに緩やかに下降する大局的な変化(話調成分)の重ね合わせとして現れ、このような知見に基づいたモデルが提案され

ている [23, 24]。藤崎モデルでは、2 次力学系へのインパルス入力およびステップ入力に対応する応答として、話調成分およびアクセント成分の変化をそれぞれ表現し、言語情報をもとにヒューリスティックな規則でモデルパラメータ (力学系への入力の大きさと時刻) を決定している [24]。しかし、アクセント成分と話調成分の分離は自明な処理ではないため、これらを分離しないモデルも提案されている。阿部らのモデルでは、アクセント句の最高ピッチ周波数とアクセント区内の相対的なピッチ変化で全体の形状を表現し、数量化 I 類を用いて言語情報をモデルパラメータに対応付けている [25]。

音素の継続時間長の決定では、音素環境などの言語情報、発話速度、モーラ性などを考慮しなければならない [26, 27, 28]。また、長い文章の発声では、ポーズの適切な挿入が必要になる [29, 30, 31]。

パワーについても、研究事例は多くはないが、数量化理論を使ったモデル化が試みられている [32]。

2.2.3 波形生成

合成単位および韻律パラメータの時系列から合成波形を生成する。波形を生成する手法としては、フレームごとに声道スペクトルの特徴を音源で駆動するボコーダタイプが過去広く研究されてきた [33, 34]。近年、計算機の処理能力の向上に伴い、合成素片を時間領域で処理する波形接続による合成方式が提案され [35, 36]、明瞭性の高い合成音の生成が可能になってきた。一方で、ボコーダタイプの合成方式としても、スペクトルなどの特徴を HMM (Hidden Markov Model) で表現した合成単位から音声を生成する手法が提案され [37]、音質の向上が図られている。

3 対話システムでの音声合成

書き言葉の読み上げとして開発されたテキスト音声合成システムを対話システムでの出力モジュールとして利用するために解決すべき課題について考えてみる。

3.1 話し言葉の形態素解析

入力が漢字仮名混じり文で与えられる場合には、まず形態素解析を行う必要がある。形態素解析システムは主に書き言葉を対象に開発されており、「〜しちゃう」などの話し言葉特有の表現や言い回しが解析できなかったり、解析精度が落ちてしまう。

3.2 入力の表現形式

対話では、対話の状況や話者の意図によって同じ文でも話し方が多様に変化する。このような多様性を合成音で表現するには、音声合成システムへの入力を単なる漢字仮名混じり文で表現するだけでは不十分である。入力テキストにコマンドを埋め込み、音声の特徴を制御するための言語の設計も行われている [38] また、コーパスのタグ付けでは XML に基づいた記述が行われ始めており、音声合成への入力や音声対話システム内部の情報の受け渡しにそのような標準化された記述方法を用いることも一つの手段である。最近、日本電子工業振興協会の作業グループによって制定された「日本語テキスト音声合成用記号の規格:JEIDA-62」[39, 40] でも、XML に準拠した形式が採用されており、対話システムでの出力メッセージの記述方式としても普及することを期待したい。

さらに、合成システムへの入力の表現方法として、テキストではなく意味表現を利用する合成方式もいくつか検討されている [41, 42, 43]。

合成システムへの入力をどのように表現するにせよ、出力される合成音の品質は合成システムの能力だけで決まるわけではない。様々な制御を受け入れる合成システムが提供されても、そこへの入力を決めるのは対話システムにおける対話管理部や対話を遂行する問題解決部などの他のモジュールであり、対話システムの総合的な能力が問われることになる。

3.3 多様な声質

対話システムで利用するには、男声・女声だけでなく多様な声質の音声が合成できることが望まれる [44, 45]。特に、顔画像出力を伴うマルチモーダルな出力を行う場合には、顔とマッチした声質での合成音が必要になる。波形接続方式の音声合成では、基本的に声質を変えることは難しく、多数話者の素片を用意するか、後処置として声質変換の技術を使うことになる [46, 47]。ボコーダタイプの音声合成では、声質をパラメトリックに制御できる可能性もある [48]。

3.4 多様な韻律

対話での発話らしい合成音を生成するには、対話の状況や話者の意図や感情などのコンテキストに応じて韻律を制御しなければならない。これを実現するには、このようなコンテキスト情報と韻律的特徴の関係を明らかにしてパラメータを生成する手法を開発するだけでなく [4, 5, 6, 7, 8, 44, 45]、対話を通してコンテキストを獲得するメカニズムも必要になる。

3.5 外部からの制御

読み上げ音声の合成を目的として開発された音声合成システムでは、韻律的な強調や感情を表現する手段が提供されていない。このような合成システムでも、外部から任意の韻律パターンを設定することができれば、多様な韻律の合成音が出力可能であるが、多くのテキスト音声合成器では、漢字仮名混じり文や簡単な制御コードを含んだ音韻列を入力とする場合が多く、声質や韻律的特徴は合成器の内部処理で決まってしまう。対話音声合成では、様々な調子の音声出力が求められることから、声質や韻律的特徴を外部から制御できるような機能が必要となる。

3.6 他モジュールとの同期

機械とのマルチモーダルな対話では、利用者に対する出力は音声だけでなく画像や文字テキストなどと同時に提供されることになる。特に擬人化エージェントを利用して顔画像と同時に出力を行うときには、音声発話と顔画像の口唇の同期 [49] がなければ自然なマルチモーダル出力は得られない。顔画像と音声が同期した出力を生成するには、時間情報の共有や画像と音声の同時生成を行う必要がある。

3.7 出力制御

人同士の対話では、相手の話の途中で割り込んで話し始めたり、相手に割り込まれたら話を途中でやめたりすることが自然に行われている。対話システムでこのようなインタラクションを実現するには、音声合成システムが音声出力を途中で打ち切るような機能を持っていなければならない。さらに、このような真に柔軟な音声出力を実現するには、音声の打ち切り処理を単に出力ルーチンの機能の問題として捉えるのではなく、発話する文全体が決定していなくても漸次的に音声を生成できる文生成および音声生成のアーキテクチャが必要になってくる [50]。

3.8 データベース

近年の音声情報処理では、大量のデータを用いて統計的にモデルを構成する手法が主流となってきている。このような手法は大規模なデータベースによって支えられており、必要なデータの収集や整備が重要な研究課題となっている。対話データに関してもデータベースの作成が進んでいるが、[51, 52, 53] まだ十分な量が確保されているとはいいがたく、今後もその拡充を期待したい [54]。

3.9 フリーソフトウェアとして

対話システムを構築し、音声出力を細かいところまで思い通りに操ろうとすれば、やはりソースコードが必要となる。現在、音声認識では、情報処理振興事業協会 (IPA) の独創的情報技術育成事業に係る「日本語ディクテーション基本ソフトウェアの開発」プロジェクトの成果として、連続音声認識システムがフリーウェアとして提供されており [55]、多くの研究機関で利用されている。一方、音声合成では、このようなフリーソフトウェアは提供されておらず、市販されている音声合成システムをブラックボックスとして利用するしかないのが現状である。フリーウェアとして提供されるシステムは、それを組み入れて対話システムなどの統合システムを構築するときに便利なだけでなく、うまくモジュール化されていれば、個々の要素技術を評価する場合にも有用なツールとなる。音声合成においてもフリーウェアとしてのシステムが提供されることが望まれる。

4 おわりに

連続音声認識の基本性能が向上し、音声入力を利用した対話システムの構築が様々なアプリケーションに対して進められることが予想される。入力だけでなく出力にも音声が利用されるようにするには、合成音の明瞭性や自然性の改善に加えて、「対話システムでの使い心地の良さ」の向上が今後音声合成システムに必要となろう。

謝辞 情報処理振興事業協会 (IPA) の独創的情報技術育成事業に係る「擬人化音声対話エージェント基本ソフトウェアの開発」プロジェクトの音声合成グループおよび読みグループの皆さんに感謝します。

参考文献

[1] 広瀬啓吉：“音声の出力に関する研究の現状と将来”, 日本音響学会誌, **52**, 11, pp.857-861 (1996).

[2] 山崎信英：“最近のテキスト音声合成とその技術”, bit, **27**, 3, pp.11-20 (1995).

[3] (社) 日本電子工業振興協会：“音声合成の製品動向”, 音声入出力方式に関する調査研究報告書, 00-標-2, pp.29-48 (2000).

[4] J. Hirschberg, “Using Discourse Context to Guide Pitch Accent Decisions in Synthetic Speech,” Proc. of ESCA Workshop on Speech Synthesis, Autrans, pp.181-184 (1990).

[5] A.I.C. Monaghan, “Intonation Accent Placement in a Concept-to-Dialogue System,” Proc. of 2nd ESCA/IEEE Workshop on Speech Synthesis, New York, pp.171-174 (1994).

[6] 白井克彦, 岩田和彦：“音声合成のための単語の強調表現の規則化”, 信学論, **J70-A**, 5, pp.816-821 (1987).

[7] 広瀬啓吉, 阪田真弓：“対話音声と朗読音声の韻律的特徴の比較”, 信学論, **J79-DII**, 12, pp.2154-2162 (1996).

[8] 山下洋一, 作田瑞, 溝口理一郎：“対話音声合成のための2段階予測に基づく韻律規則の生成と評価”, 日本音響学会誌, **53**, 2, pp.103-109 (1997).

[9] 山下洋一：“朗読から対話音声合成へ”, 日本音響学会誌, **49**, 12, pp.860-865 (1993).

[10] 鈴木和洋, 斉藤隆：“日本語テキスト音声合成のためのN文節構造解析とそれに基づく韻律制御”, 信学論, **J78-DII**, 2, pp.177-187 (1995).

[11] 窪薗晴夫：“日本語の音声”, 現代言語学入門2, 岩波書店 (1999).

[12] 佐藤大和：“複合語におけるアクセント規則と連濁規則”, 日本語の音声・音韻 (上) (杉藤美代子 編), 講座「日本語と日本語教育」第2巻, 明治書院 (1989).

[13] 佐藤大和：“連濁の分析と規則化の検討”, 日本音響学会秋季講演論文集, 1-2-10, pp.61-62 (1983).

[14] 匂坂芳典, 佐藤大和：“日本語単語連鎖のアクセント規則”, 信学論, **J66-D**, 7, pp.849-856 (1983).

[15] 宮崎正弘：“単語間の意味的結合関係を用いた複合語アクセント句の自動抽出法”, 信学論, **J68-D**, 1, pp.25-32 (1985).

[16] 野村典正：“単語の分類を用いた複合語のアクセント句分割とアクセント付与”, 信学論, **J75-DII**, 9, pp.1479-1488 (1992).

[17] NHK 放送文化研究所 編：“NHK 日本語発音アクセント辞典 新版”, 日本放送出版協会 (1998).

[18] 阿部芳春, 今井聖：“CV 音節のケプストラムパラメータからの音声合成”, 信学論, **J64-D**, 9, pp.861-868 (1981).

[19] 佐藤大和：“PARCOR-VCV 連鎖を用いた音声合成方式”, 信学論, **61-D**, pp.858-865 (1978).

[20] 佐藤大和：“CVC と音源要素に基づく音声合成”, 日本音響学会音声研資, S83-69, pp.541-546 (1984).

[21] 中嶌信弥, 浜田洋：“音韻環境に基づくクラスタリングによる規則合成法”, 信学論, **J72-DII**, 8, pp.1174-1179 (1989).

[22] 武田一哉, 安部勝雄, 匂坂芳典：“選択的に合成単位を用いる規則音声合成”, 信学論, **J73-DII**, 12, pp.1945-1951 (1990).

[23] 箱田和雄, 佐藤大和：“文音声合成における音調規則”, 信学論, **J63-D**, 9, pp.715-722 (1980).

[24] 広瀬啓吉, 藤崎博也, 河井恒, 山口幹雄：“基本周波数パターン生成過程モデルに基づく文章音声の合成”, 信学論, **J72-A**, 1, pp.32-40 (1989).

[25] 阿部匡伸, 佐藤大和, “音節区分化モデルに基づく基本周波数の2階層制御方式,” 日本音響学会誌, **49**, pp.682-690 (1993).

[26] 匂坂芳典, 東倉洋一：“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 (1984).

[27] 海木延佳, 武田一哉, 匂坂芳典：“言語情報を利用した母音継続時間長の制御,” 信学論, **J75-A**, pp.467-473 (1992).

[28] 石川 泰：“規則合成のための2モーラを単位とする音韻継続時間長規則”, 信学技報, SP97-93, pp.41-48 (1998).

[29] 海木延佳, 匂坂芳典：“局所的な句構造によるポーズ挿入規則化の検討”, 信学論, **J79-DII**, 9, pp.1455-1463 (1996).

[30] 藤尾茂, 匂坂芳典, 樋口宜男 : "確率文脈自由文法を用いた韻律句境界とポーズ位置の予測", 信学論, **J80-DII**, 1, pp.18-25 (1997).

[31] 海老原充, 石川泰 : "音声合成におけるネットワークモデルによるポーズ位置予測", 信学技報, SP96-133, pp.45-50 (1997).

[32] 三村克彦, 海木延佳, 匂坂芳典 : "統計的手法を用いた音声パワーの分析と制御", 日本音響学会誌, **49**, 7, pp.253-259 (1993).

[33] D.H. Klatt : "Software for a cascade/parallel formant synthesizer", J. Acoust. Soc. Am., **67**, pp.971-995 (1980).

[34] 河井恒, 樋口宜男, 清水徹, 山本誠一 : "隣接音素の統計的性質に基づくホルマント型音声合成方式", 日本音響学会誌, **50**, 2, pp.117-125 (1994).

[35] E. Moulines and F. Charpentier : "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones ", Speech Communication, 9, pp.453-467 (1990).

[36] Nick Campbell, Alan Black : "CHATR: 自然音声波形接続型任意音声合成システム", 信学技報, SP96-7, pp.45-52 (1996).

[37] 益子貴史, 徳田恵一, 小林隆夫, 今井聖 : "動的特徴を用いたHMMに基づく音声合成", 信学論, **J79-DII**, 12, pp.2184-2190 (1996).

[38] 水野 理, 中嶌 信弥 : "合成音声制御のための階層型記述言語 MSCL", 人工知能研資, SIG-SLUD-9703-2 (1997).

[39] 蓑輪利光, 赤羽誠, 板橋秀一 : "JEIDA 日本語テキスト音声合成用記号", 日本音響学会秋季講演論文集, 2-1-5, pp.183-184 (2000).

[40] (社) 日本電子工業振興協会 : 日本語テキスト音声合成用記号の規格, JEIDA-62-2000 (2000).

[41] S.J. Young and F. Fallside : "Speech synthesis from concept: A method for speech output from information systems", J.Acoust.Soc.Am., **66**, 3, pp.685-695 (1979).

[42] 藤崎博也, 広瀬啓吉, 浅野康治 : "知識表現からの文章音声合成システム", 日本音響学会秋季講演論文集, 2-6-6 (1990).

[43] 山下洋一, 水谷直樹, 角所収, 溝口理一郎 : "汎用音声出力インタフェースにおける概念表現からの音声合成", 信学論, **J76-D-II**, 3, pp.415-426 (1993).

[44] 宮武正典, 匂坂芳典 : "種々の発話様式に見られる韻律特徴とその制御", 信学論, **J73-DII**, 12, pp.1929-1935 (1990).

[45] 阿部匡伸 : "異なる発話様式の特徴分析とその制御", 日本音響学会誌, **51**, 12, pp.929-937 (1995).

[46] 阿部匡伸, 嵯峨山茂樹 : "音素セグメントを単位とする声質変換", 信学技報, SP90-88, pp.25-32 (1990).

[47] 橋本誠, 樋口宜男 : "話者選択と移動ベクトル場平滑化を用いた声質変換における写像元話者の選択方法" 信学論, **J81-DII**, 2, pp.249-256 (1998).

[48] 小山晃俊, 徳田恵一, 北村正, 小林隆夫 : "固有声 (eigenvoice) に基づいた音声合成", 日本音響学会秋季講演論文集, 1-3-18, pp.219-220 (1999).

[49] 山本英理, 中村哲, 鹿野清宏 : "EMアルゴリズムを用いたAudio-Visual HMMによる音声からの画像パラメータ推定", 信学技報, SP98-65, pp.51-56 (1998).

[50] 岡田美智男, 栗原聡, 大塚裕子 : "最小指定の枠組みに基づく自然な発話の生成機構のモデル化", 人工知能研資, SIG-SLUD-9302-8 (1993).

[51] 板橋 秀一, 山本 幹雄, 河原 達也 : "重点領域研究「音声対話」における音声コーパス" 人工知能学会研究会資料 SIG-SLUD-9701-5, pp.25-30 (1997).

[52] 田中和世, 速水悟, 山下洋一, 鹿野清宏, 板橋秀一, 岡隆一 : "RWC計画における音声対話データベースの構築", 情報処理学会音声言語処理研究会報告, SLP-11-7, pp.37-43 (1996).

[53] 山本幹雄 : "音声対話データベース構築の現状" 日本音響学会誌, **54**, 11, pp.797-802 (1998).

[54] "論文特集「談話タグ:コーパスとタグ付けの支援技法」", 人工知能学会誌, **14**, 2, pp.230-295 (1999).

[55] 河原達也, 他 : "日本語ディクテーション基本ソフトウェア (98 年度版)", 日本音響学会誌, **56**, 4, pp.255-259 (2000).