音声言語情報処理 17－12
（1997. 7. 19）

様々な音声表現を実現できる音声作成ツール
－ *Speed97* －

阿部匡伸 水野秀之 中嶌信弥
NTT ヒューマンインタフェース研究所
〒239 横須賀市光の丘 1－1

音声作成ツール(Speech editor 97)を開発した。本システムは、グラフィカルユーザインタフェース(GUI)を用いて、音声合成のパラメータが操作できるものであり、その目的は、従来の TTS では不可能であったきめ細かな制御を可能とし、多種多様な品質や表情で音声を合成することにある。操作法には、漢字かな混じり文、アクセント型等をテキストベースで修正するモードと、音声のパワー、基本周波数、継続時間をパラメータレベルで修正するモードとがある。UNIX 上と Windows95 上で動作している。*Speed97* で作成された音声は、音声信号と音素記号等との対応が明確になっているため、他のメディアとの同期が容易にとれる等のメリットがある。また、*Speed97* は、音声ガイダンスの作成等の音声メッセージの作成ばかりでなく、例えば、感情を込めてせりふを読ませるなどして演技させることも可能である。さらに、*Speed97* で作成された音声は、1kbit/s 以下の高能率音声符合化音声として利用することも考えられる。

Speed97: A speech creation tool for synthesizing various kinds of speech

Masanobu ABE, Hideyuki MIZUNO, Shin'ya NAKAJIMA
NTT Human Interface Laboratories
〒239 1-1 Hikarinooka Yokosuka-Shi

We developed a tool (Speech editor 97) to create speech messages. *Speed97* provides a graphical user interface to manipulate parameters of speech synthesis, and makes it possible to synthesize various types of speech. The manipulation is performed in text level such as to change Chinese characters and accent types, and in parameter level such as to modify speech power, fundamental frequency and duration. *Speed97* runs on UNIX and Windows95. Speech messages created by *Speed97* have several advantages. Examples include easy synchronization with other media such as moving picture, because the speech is associated with phoneme symbols, and a low bit rate; i.e., only phonetic symbols and prosodic parameters should be transmitted; approximately 1kbit/sec or less.

1. はじめに

規則による音声合成の研究は、過去20年あまりの間、テキストからの音声合成（TTS）というアプリケーションを中心に推移してきた。TTS は、実用的であるばかりか、その実現自体が魅力的な研究テーマであった。TTS の特徴は、テキスト解析から合成音声の出力までを全て自動で行う点にある。そのため、コンポーネントとしての独立性が高く、文字情報から音声情報へのメディア変換として、またはヒューマンインタフェースの道具として、様々なシステムに容易に組み込むことができる。一方、全自動であるが故に、合成された音声の品質には限界がある。

本稿で報告する音声作成ツール（Speech editor）とは、上述した TTS とは異なるものであり、音声合成の新しいアプリケーションの試みである。即ち、全自動で音声を合成するのではなく、ある程度人間が介在することを許し、このことにより、従来の TTS では不可能であったきめ細かな制御を可能とし、多種多様な品質、表情で音声を合成する。つまり、「音声を制作」する手段を与えるものである。

音声はガイダンスや情報アナウンスに使われるばかりか、マルチメディアコンテンツにも不可欠であるが、そこで利用されているのは、自然音声、または、自然音声を変形したものであり、音声を人工的に創造して用いる例は殆どない。現在のところ、自然音声は声優が発声していることが多く、その収録は時間的にも、金銭的にも高価である。また、一般のユーザが自室やオフィス環境で自然音声を録音する場合には、音声入力装置の貧弱さや、SN 比が高く取れない等の理由によって、高品質な自然音声を収録することは容易ではない。音声を人工的に創造できれば、これらの問題が解決できる。「音声を制作」することの潜在的なニーズは大きいものと考えられる。

一方、音声に比べて他のメディアの制作では、コンピュータを用いた創造支援が広く普及している。例えば、ワードプロセッサは、作家の文筆活動や雑誌の編集など、文書の作成を生業とする職業で使われるばかりでなく、一般の会社でも報告書の作成などで広く利用されている。電子楽器を用いた音楽は、デスクトップミュージック（DTM）とよばれ、最近の楽曲制作には不可欠といっても過言ではない。さらに、趣味として DTM を楽しむ人達も増加している。また、コンピュータグラフィックス（CG）は、映画製作やゲーム制作では、ポピュラーな手段の1つとして定着している。今後は、CGも趣味の世界へ広がっていくであろうと予想される。DTM や CG の普及の裏には、使い勝手のよいツールの存在がある。「音声を制作」する場合にも、その制作環境を整えることによって、「音声を制作」することが、広く受け入れられるようになると考える。 *Speed97* は、それを目指した第1歩である。

「音声を制作」するツールとしては、様々なレベルで、様々な機能が必要であろう。このうち、*Speed97* は、最も低レベルに位置し、音声合成のパラメータを直接操作する機能を有する。また、より上位の音声作成法として、我々は MSCL(Multi-Layered Speech/Sound Synthesis Control Languate) と呼ばれる記述方式を提案している。これは、HTML のようなマークアップランゲージに類似しており、抽象化された制御コマンドで音声現象を表現することによって音声作成を行う。詳細は、文献[1]を参照されたい。

2節では、*Speed97* の利点、その応用分野について述べる。3節では、*Speed97* の機能について説明し、4節では、そのシステム構成について述べる。

2. *Speed97* の利点と応用分野

Speed97 は TTS をベースとする音声作成ツールである。*Speed97* で作成された音声は、音声信号と音素記号等との対応が明確になっているため、自然音声にはない次のようなメリットがある。

(1) 他のメディアとの同期が容易に可。
マルチメディアタイトルの制作では、画面切り替えのタイミングなど、音声と画像の同期が重要となる。*Speed97* で作成された音声は、音素記号レベル、分析フレーム長レベルなど、詳細な時間分解能で他のメディアとの同期が容易にとれる。

(2) 音声の検索が可。
対応づけられた文字情報をたより音声信号を検索でき、必要な部分のみを聞くことなどに利用できる。

(3) 再利用可。
一度作成した音声を、文字情報および韻律情報と

共に、文単位、フレーズ単位などで辞書に蓄積しておけば、これを再利用して、新たな文を容易に作成できる。言い回しや語彙の限られたアナウンス文等の作成に有効である。

Speed97 は、従来の TTS のように、音声ガイダンスの作成や、情報アナウンスのための音声メッセージの作成[2]ばかりでなく、様々な音声表現を実現できることが特徴である。例えば、感情を込めてせりふを読ませるなどして演技させることも可能であり、アニメーションのアフレコ用の音声が作成できる。音声制作者は、演出家であり、同時に声優となる。また、機械の出力に合成音声を利用する場合には、きめ細かな音声表現が要求される[3]。*Speed97* は、このような音声を合成する場合にも有益である。文献[4]では、*Speed97* による合成音声を用いて、フレンドリーなユーザインタフェースの検討を行っているので参照されたい。

規則合成において、規則による韻律生成を行うのでなく、自然音声から抽出した韻律パラメータを利用すれば[5]、音素をセグメントに持つセグメントボコーダとみなすことができる。この場合、*Speed97* は、セグメントボコーダのエンコーダと言える。*Speed97* で作成された音声は、1kbit/s 以下の高能率音声符合化音声として利用できる。

3. *Speed97* の機能

Speed97 による音声の作成は以下のようにして行う。

(1) かな漢字混じり文のエディット
かな漢字混じり文をキーボードから入力し、エディットする。キーボードから入力する代わりに、既存のテキスト文書を読み込むことも可能である。

(2) テキスト解析
かな漢字混じり文に対して、読みを付与するとともに、文節のアクセント型と文節間の結合度合いを推定する。

(3) かな文字とアクセント型等のエディット
読み誤りの修正、および、推定されたアクセント型と文節間の結合度合いの誤りを修正する。また、後者の修正には、合成音声を視聴して参考にすることもできる。

(4) 韻律パラメータのエディット
パワー、基本周波数、継続時間長をグラフィカルユーザインタフェース (GUI) を用いて操作する。また、音声の合成と視聴を試行錯誤的に繰り返しながら、音声作成を進めることができる。

(1)から(3)の操作は、すべてテキストベースである。これら一連の操作を行う画面を図1に示す。この画面には、「入力」「登録」「保存」「テキスト解析」「合成」「D/A」「対比較 D/A」「韻律修正」「終了」のボタンがあり、ボタンをクリックすることで処理が起動される。

「登録」とは辞書登録の機能である。すなわち、エディット中の文書の中で、再び利用しそうな部分を選択し、その部分の文字情報、アクセント情報および韻律パラメータをファイルに書き出す。「保存」もファイルに書き出す処理であるが、「登録」と異なるのは、選択した部分でなく、エディット中の文書全体が操作対象になる点である。

「D/A」では、文書全体、または、選択した部分だけを聞くことができる。「対比較 D/A」では、ある時点までに作成した合成音声の中から、任意の2つの音声を選んで聞くことができる。

「韻律修正」をクリックすると、図2に示すような画面が表示され、(4)の韻律パラメータのエディットできるようになる。画面には、音声のパワー、合成波形、基本周波数パタン、音素記号が、上から順に表示されている。マウスの操作によって、各パラメータが操作できる。即ち、音声パワーの表示領域では音素毎の平均パワーが、合成波形表示領域では音声の継続時間が、基本周波数パタン表示領域では音素毎に3つの基本周波数が設定できる。操作モードには、全体を一括して操作するモードと一点毎に操作するモードとがある。さらに、基本周波数パタンに対しては、特別に自由曲線で操作できるモードがある。これらのモードは、図2の画面右側に表示されたボタンで切り替えることができる。

図2の画面には、「合成」「D/A」「対比較 D/A」「リセット」「スケール」「保存」「終了」のボタンがあり、ボタンをクリックすることで処理が起動される。「リセット」は、それまでの韻律パラメータの修正を全て取り消し、図2の画面を開いた時点の韻律パラメータにセットしなおす。

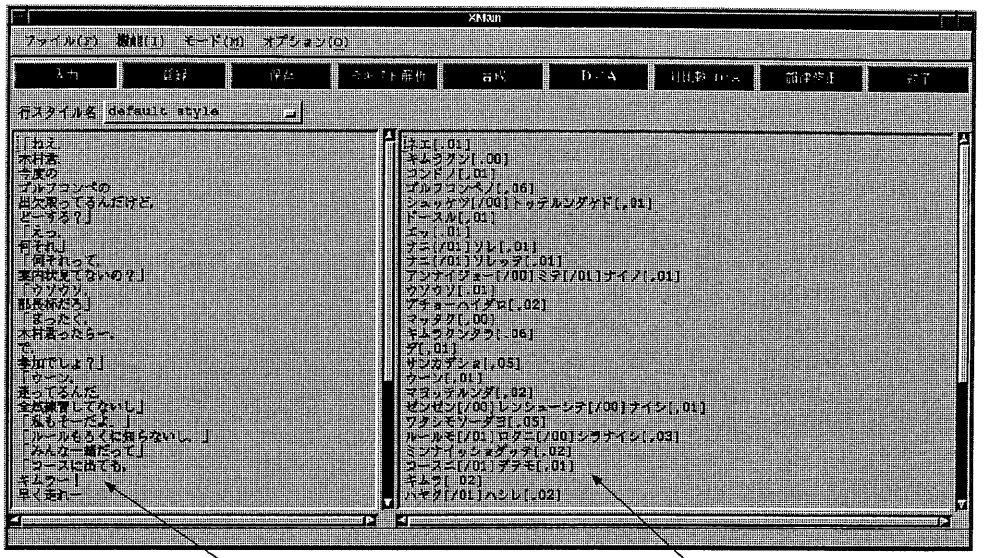

図１ テキストレベルでの修正画面

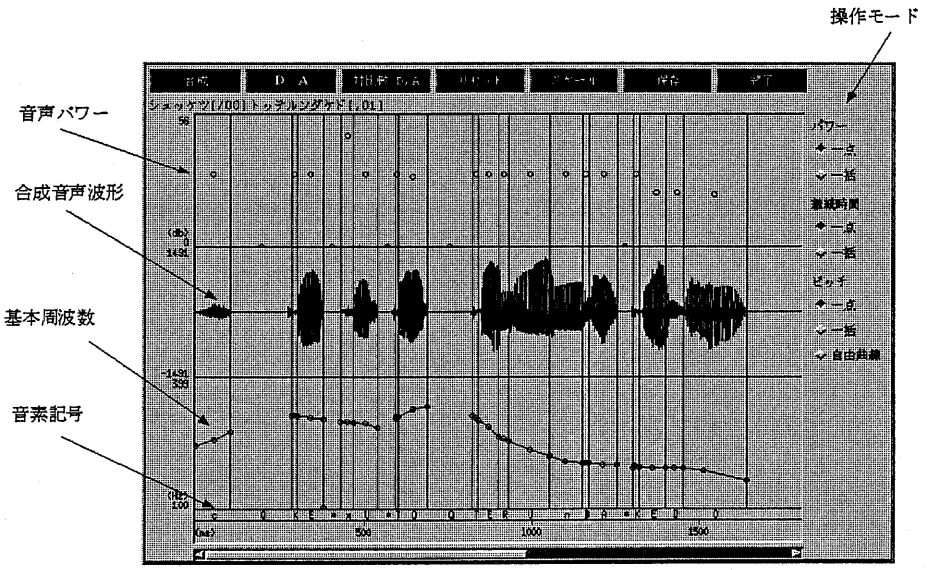

図２ 韻律パラメータ修正画面

通常、図2の画面に表示されるパラメータの目盛りは、パラメータの大きさによって自動的に設定される。「スケール」は、この画面に表示されるパラメータの目盛りを任意に変更するためのものである。

以上の「韻律修正」は行単位に行う。即ち、ある行の韻律パラメータが修正したら、このモードを終了し、再び図1の画面に戻り、次の「韻律修正」する行を指定する。

以上のように(1)(2)(3)(4)の操作によって、音声を合成することができる。その他に、図1の画面の「オプション」で「スタイル」を選択すると、音声の声質を設定することができる。図3は、スタイル設定画面である。スタイルは、男女声の別、発話速度、音量、基本周波数の平均値とダイナミックレンジのパラメータがある。スタイルは、名前を付けて登録することができ、図1の画面の行毎にスタイルが設定できる。各行にカーソルを持っていくと、その行のスタイル名が、図1の画面の「行スタイル」のところに表示される。

4. *Speed97* のシステム構成

図4にシステム構成を示す。テキスト解析部、韻律パラメータ生成部、音声合成部は、TTS での機能と同じである。基本となる TTS システムは、文献[6]のものである。注意すべき点は、テキストエディタ、カナアクセントエディタ、韻律パラメータエディタへのデータの流れが一方向であることである。下位での修正を上位レベルに反映させることは、不可能では無いが、煩雑な処理となるため本システムでは実装していない。

本システムは、UNIX と Windows95 とに実装されている。両者の機能は、全く同じであるが、ウインドウシステムが異なるため、見た目は多少異なる。

5. まとめ

本報告では、規則合成音声のアプリケーションである音声作成ツール「*Speed97*」について述べた。 数名の被験者で評価実験を行ったところ、 *Speed97* では、韻律パラメータの再表示や音声合成にかかる時間は極めて短時間であり、スムーズに音声作成できることが確認されている。音声の作成手順から分かるように、*Speed97* で音声を作成するためには、ユーザが音声の基礎知識を持つことが前提となる。また、パラメータの変更と音声合成を繰り返しながら、試行錯誤的に音声作成するためには、ユーザがノウハウを獲得する必要もある。音声作成を作成するという観点から、必要な

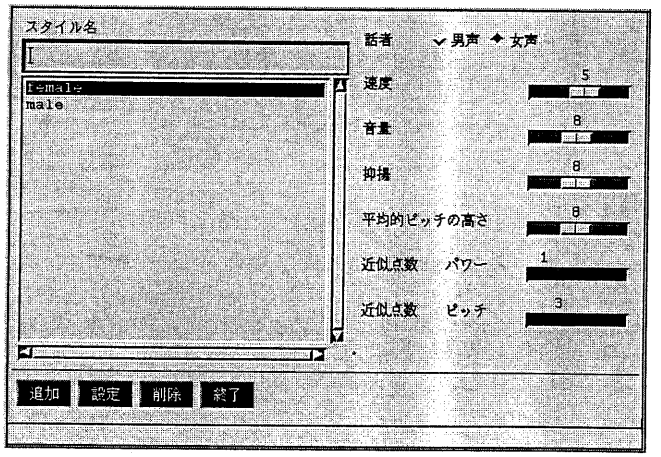

図3 スタイル設定画面

な基礎知識、ノウハウなどを整理することが重要であると考えている。今後は、試行錯誤を行う上で適切なアドバイスを与える機能や、既存の韻律パターンを再利用する機能などを付け加えて、より簡単に音声を制作できるツールへと発展させていく予定である。

参考文献

[1]水野，中嶌，“表現力豊かな合成音声を生成するための記述言語 MSCL の構想，” 本予稿集.

[2] T. Hirokawa, K. Itoh, K. Hakoda,” Speech editor based on enhanced user-system interaction,” AVIOS93, pp.39-45.

[3]阿部,”誤り指示音声の特徴分析と音声出力への適用，” 電子情報通信学会論文誌 D-2 Vol. J79-D-2 No. 12 pp.2191-2198.

[4]篠崎，阿部，“規則合成音声で躍動感を実現する方略について，” 本予稿集.

[5]B. Coile, L. Tichelen, A. Vorstermans, J. Jang, M. Staessen, “PROTRAN: A prosody transplantation toll for text-to-speech application,” ICSLP94, pp. 423-426.

[6] K. Hakoda, T. Hirokawa, H. Tsukada, Y. Yoshida, H. Mizuno, “Japanese text-to-speech software based on waveform concatenation method,” AVIOS95, pp.45-54.

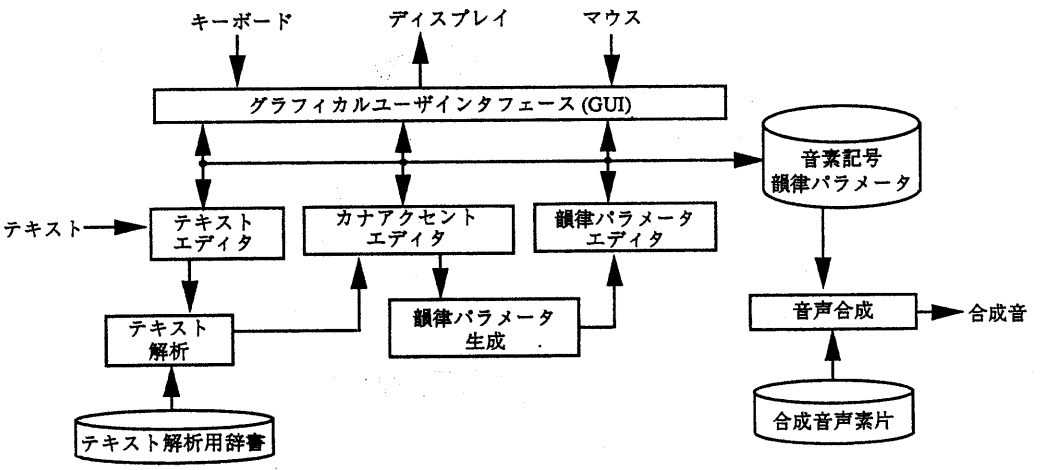

図 4　Speed97 のシステム構成