# マルチ言語マルチ話者音声合成システムCHATR
## －お気に入りの声をコンピュータで合成－

(株)ATR音声翻訳通信研究所
第二研究室
ニック・キャンベル、樋口　宜男

当研究所では、ある話し手の声を一定量以上（例えば30分程度）、コンピュータに記憶して音声データベース化しておき、その中に多数ある同じ種類の音韻（子音や母音）の中から、出力したい文に応じて、音韻の並び方や声の高さ、音韻の長さなどの条件が良く適合し、しかも滑らかにつながる音韻を選び出し、それらの音声波形をつなぎ合わせるという方法を開発し、短期間でその人の声を合成することを可能にしました。

### [1]　はじめに

現在の日常生活ではコンピュータから出力される音声を耳にする機会が多くなりました。例えば駅のアナウンスでは「ひかり123号が12番線に参ります。」と言ったり、高速道路の料金所では「料金は1200円です。」という声が聞こえたりします。また、文章を読み上げる機能が付いているパソコンもあり、電子メールを音声で聞くこともできます。

これらの音声出力の方法は

(1) 単語単位の音声を予めアナウンサに読んでもらい、それらをつなぎあわせて出力する方式（通常、録音編集方式と言う）と、

(2) 音韻（子音や母音）や仮名のような比較的小さな単位をつなぎあわせて出力する方式（通常、規則合成方式と言う）

の２つに大別されます。規則合成方式はどんな内容の文でも音声に変えられるという利点を持っていますが、どうしても合成音特有の響きがあるため、これまで高い品質を要求される公共サービスへの利用は録音編集方式のみに限られて来ました。

一方、録音編集方式を用いた場合、利用する単語すべてを予め録音しておかなければならないために、新幹線に「のぞみ」が登場したときや高速道路に新しいインターチェンジができたときのように新しい単語を増やす必要が生じたときに、既にあるものと同じ声の質で新しい単語を録音することがかなり難しい問題になってきます。

この点を解決するために、声の高さや音韻の長さを変える処理を最小限に抑え、自然な音声の響きをそのまま残した合成音声を出力するシステムCHATR（チャター、おしゃべりを意味するchatとATRの合成語）を開発しました。このシステムは単に自然な音声の響きを残すというだけでなく、その人の個人性やそのときの雰囲気を忠実に再現できるので、音声合成器のカスタマイズという点から大いに注目されます。

### [2]　目的文に合った音声波形データの選択

これまでの音声合成システムでは、まず出力する文の内容に従って音韻の特徴を表わすスペクトル特徴を生成し、その後で抑揚を表わす基本周波数（ピッチ周波数とも言う）を制御する方式が採られていました。具体的には、一旦音声波形をスペクトル特徴パラメータに変換してから音声波形に戻したり、音声波形を$\hat{1}$周期毎に分離してピッチ周波数を変えたりしてきました。このため、音声データにはいろいろな信号処理が施されることになり、元の音声が持っていた自然な響きが損なわれてしまい、合成音特有のこもった響きの原因となっていました。

この点を解決するために当研究所では、ある話し手の声を一定量以上（例えば30分程度）、コンピュータに記憶して音声データベース化しておき、その中に多数ある同じ種類の音韻の中から、出力したい文に応じて、音韻の並び方や声の高さ、音韻の長さなどの条件が良く適合し、しかも滑らかにつながる音韻を選び出し、それらの音声波形をつなぎ合わせるという方法を開発しました [1]。

音声単位の選び方は図１に示すように、目標とする音韻の持つべき特徴（$t_{i-1}$, $t_i$, $t_{i+1}$）とのずれを表わす$C^t$と、接続される音声単位（$u_{i-1}$, $u_i$, $u_{i+1}$）間での不連続さを表わす$C^c$との和を最小にするような音韻の波形データを音声データベースの中から選んできます。

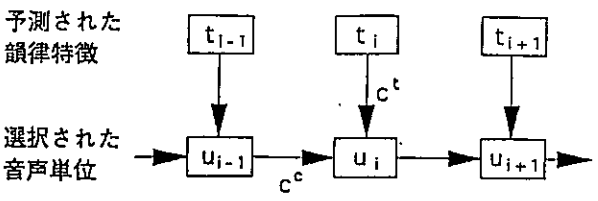


図１　音声単位の選択規準
$C^t$は目標とする音韻とのずれ
$C^c$は隣接単位間での不連続さの度合い

信号処理という観点から見ると音韻単位の波形データを相互に接続するだけですから、以前の方式に比べてずっと簡単になっていると考えられますが、最適な音声単位を音声データベースの中から探し出すという観点から見ると膨大な計算を行なっていることが分かります。例えば、30分の音声データには約15,000の音韻が含まれており、その中で特に生起頻度の高い/a/（ア）の音について見ると音声データベース中に約1,200個の波形データが含まれています。従って、その中から最適のものを選んでつなぎあわせるという作業がいかに大変かは容易に想像できると思います。この点を解決するために音声認識で用いられているViterbiアルゴリズムを利用して高速処理を行なっています。

このシステムで実際に音声を出力するときには、図2に示すように音声データベース中の波形データを逐次読み出していきます。この方式では音声波形を変形したり新たに作り出したりせず、音声データベースの中から最適なものを選び出すという処理をしていますので、目標に近い音韻が音声データベース中にない場合には隣り合った波形データ間で不連続感が生じたり、抑揚の不自然さが生じたりすることがあります。この点は音声データベースの規模を拡大すれば解決しますが、小規模でも効率良く多様な音韻を含むような音声データベースをどのようにして設計するかは今後の課題となっています。

## 3 マルチ言語・マルチ話者

CHATRで音声を出力する場合、ピッチ周波数やそれぞれの音韻の時間長を文の内容に従ってどう決めるかについては言語に対する依存性がありますが、それらが何らかの方法で決められたとすると、音声データベースの中から最適な波形データを選択する過程は言語や話し手に依存しません。このため、CHATRを多言語化することは比較的容易であり、図3に模式的に示したように、現在日本語・英語・韓国語・ドイツ語の4ヶ国語の合成が可能となっています。また、いろいろな話し手の声を利用することも極めて容易であり、例えば日本語については現在16名の声を出し分けられます。

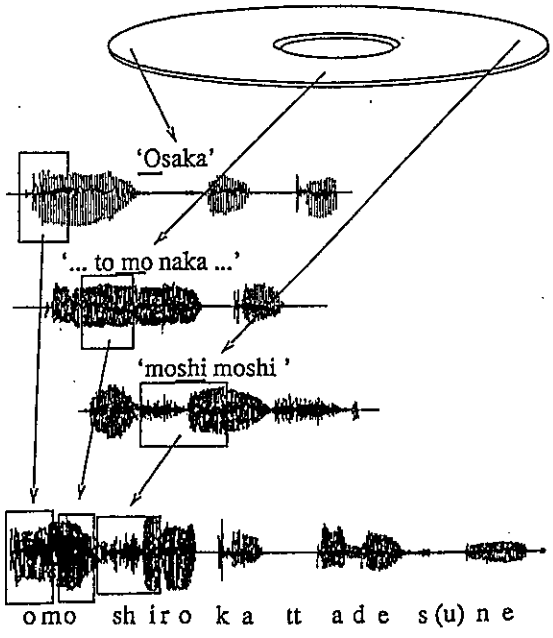

図2 CHATRによる合成音声の生成

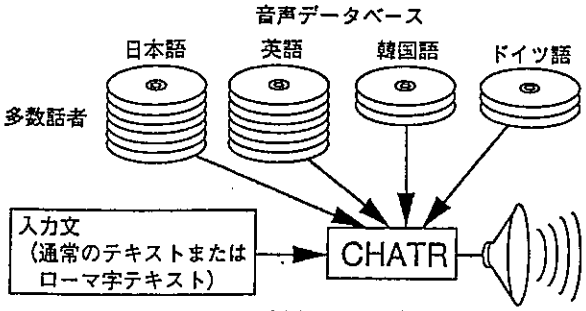

図3 マルチ言語化・マルチ話者化

新しい話し手の音声データベースを作る場合には、大きく分けて2つの方法があります。すなわち、

(1) 録音されている音声データの内容を書き起こし、それに従って音韻ラベルを付ける方法と、
(2) 決められた文を話し手に読んでもらい、既存の音韻ラベルを自動的に割り振っていく方法

です。

前者の場合、市販されているカセット・テープを利用することも可能であり、その一例として黒柳徹子さんが2つの小説を読んでいる新潮カセットブックの約1時間の音声を用いて黒柳さんの声に非常に近い合成音声が出せることを確認しています。（ただし、黒柳さんご本人からは声は非常に似ていますが、もう少し黒柳さんのアクセントを研究して欲しいというコメントを頂いております。）

一方、後者の方法を用いた場合、ほぼ自動処理が可能であり、録音から約3時間という極めて短い時間で合成音声を出力することができます。

## 4 おわりに

従来の音声合成システムと異なり、音声データベース中から、出力したい文に応じて、音韻の並び方や声の高さ、音韻の長さなどの条件が良く適合し、しかも滑らかにつながる音韻を選び出し、それらの音声波形をつなぎ合わせるという方法を用いて高品質の音声合成システムを実現しました。これらの合成音声はすべてhttp://www.itl.atr.co.jp/chatr/で聞いて頂けますので、お試し頂ければ幸いです。


参考文献

[1] N. Campbell・A. Black："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，SP 96 - 7 (1996. 5).