入力音声の韻律を用いた音声合成 *

◎藤澤 謙　　ニック キャンベル
(ATR 音声翻訳通信研究所)

1　はじめに

近年，音声合成技術の向上にともなってカーナビゲーションなどへの応用が検討されており，GUI を用いて合成音声の韻律を制御するシステム [1] [2] が提案されているが，録音再生方式の音声と比較した場合，音声の自然性が不十分である．しかし，録音再生方式のアプリケーションでは発声すべき単語が増えたとき同じ話者で追加録音する必要がある．録音する際のコンディションや前回の録音からの時間的な間隔などによって必ずしも以前と同じような声質の音声が録音できるとは限らず，録音話者と同じ音声を合成で実現できると便利である．

一方，我々は話者の特徴を活かした波形接続型音声合成システム CHATR [3] [4] を提案している．CHATR は，あらかじめ録音された音声データベース中の音素単位の音声波形を，信号処理を行わずに接続して連続音声として出力するため，話者の特徴を失わずに音声を合成することを特長としているが，(a) ターゲットとする基本周波数 (以下 F_0) の予測エラーや (b) 音素単位の選択エラーによる不自然なイントネーションが問題となっている．

本稿では (a) を解決するため，あらかじめ与えた文章を発声し，その韻律情報を出力話者のレンジに変換したものをターゲットとして音声を合成するシステムについて報告する．韻律情報として実際に発声した音声のものを使用することにより，合成音声の自然性の向上に加え，部分的な強調や韻律の変更など，さまざまな韻律の合成音声を簡便に作成することを目的とする．

2　システム構成

システムの実行画面を図 1に，構成を図 2に示す．システムには図 3に示すように合成する文章とその音素表記をあらかじめ与えておく．音声が入力されると与えられた音素ラベルにしたがって Viterbi alignment [5] を行い，音素境界を決定する．また，声門閉鎖点 [6] に基づいて入力音声の F_0 を抽出する．

音声合成には波形接続型音声合成システム CHATR を用いた．F_0 とパワーはそれぞれ平均，標準偏差より合成音声出力話者のレンジに変換して，音声合成セグメント情報を作成し，CHATR により合成する．

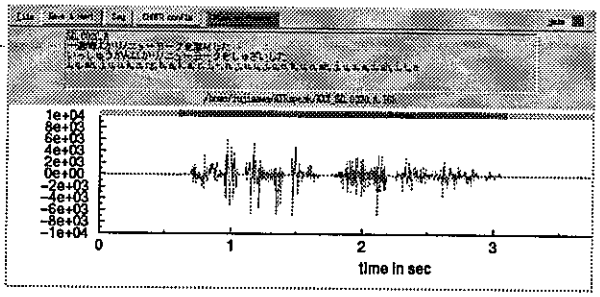

図 1: 実行画面

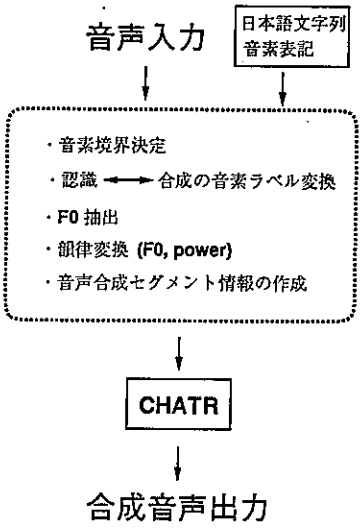

図 2: システム構成

```
SD.0010.A
あらゆる現実を／すべて自分の方へねじ曲げたのだ
あらゆるげんじつを／すべてじぶんのほうへねじまげたのだ
a,r,a,y,u,r,u,g,e,N,j,i,ts,u,o,#,s,u,b,e,t,e,j,i,b,u,N,n,o,h,o,o,e,
n,e,j,i,m,a,g,e,t,a,n,o,d,a

SD.0020.A
一週間ばかり／ニューヨークを取材した
いっしゅうかんばかり／ニューヨークをしゅざいした
i,ssh,u,u,k,a,N,b,a,k,a,r,i,#,n,y,u,u,y,o,o,k,u,o,sh,u,z,a,i,sh,i,t,a
```

図 3: 入力テキスト例

*Prosody adaptation for speech synthesis using voice input, by K. Fujisawa, and N. Campbell (ATR Interpreting Telecommunications Research Labs.)

3 アルゴリズム

3.1 音素ラベル変換

本システムに用いた音声認識と合成システムでは 表1に示す音素ラベルが一対一に対応しない問題があったため，音素ラベルは合成のラベルを与えた．また，有声／無声は入力音声の F_0 より判定し，合成時に音素ラベルの変換を行った．

表1: 音素ラベルの違い

合成の音素ラベル	認識の音素ラベル
/h/(ハ行)，/f/(ファ行)	/h/
/i/(有声'い')，/I/(無声'い')	/i/
/u/(有声'う')，/U/(無声'う')	/u/

3.2 韻律変換

入力が男性の声で女性の合成音声を出力するように，入力音声と合成音声出力話者でピッチレンジが異なる場合，合成器に与える F_0 を出力話者のものに変換しなければならない．本システムでは F_0 を以下の式で変換したものを音声合成時のターゲットとして用いた．時刻 i の入力音声の F_0 を x_i，出力音声のターゲット F_0 を y_i としたとき，

$$y_i = \overline{y} + w_{f0}\sigma_y x_i^z \quad (1)$$

$$x_i^z = \frac{x_i - \overline{x}}{\sigma_x} \quad (2)$$

ここで，σ_x は入力音声の F_0 の標準偏差，$\overline{x}$ は入力音声の F_0 の平均値を表わす．σ_y は音声出力話者の F_0 の標準偏差，$\overline{y}$ は音声出力話者の F_0 の平均値を表わし，オフラインであらかじめ出力話者のデータベースより計算しておく．w_{f0} は変換率を表わし，この値を変えることにより合成音声の抑揚の大きさを制御する．通常は 1.0 として使用する．

パワーについては F_0 と同様の変換を行い，音韻継続長は音素境界より得られるものを用いた．

発話内容 "あらゆる現実を全て自分の方へねじ曲げたのだ" を男性の声を入力し，女性の話者 (FMP) の F_0 に変換した例を図4(a) に示す．$w_{f0} = 1.0$ として変換した．入力音声の F_0 の平均は 110 [Hz]，標準偏差は 13 [Hz] であり，FMP の F_0 の平均は 228 [Hz]，標準偏差は 44 [Hz] である．FMP の波形データベース中の同発話の F_0 パターンを同図 (b) に示す．

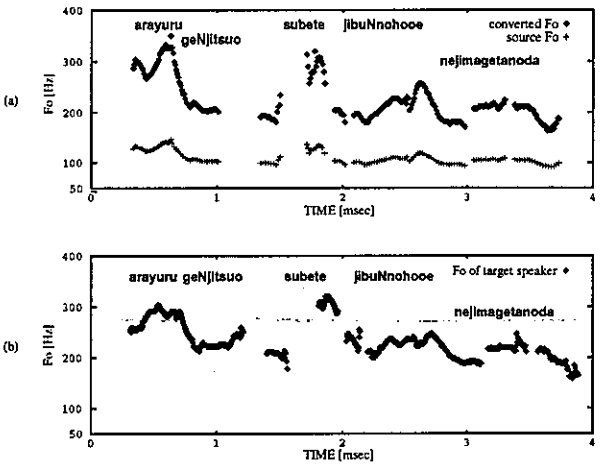

図 4: F_0 変換例

4 むすび

あらかじめ与えた文章を発声しその韻律情報を抽出，目的話者の韻律に変換して音声を合成するシステムを作成した．将来音声認識技術が発達すれば，入力を発話音声のみにして合成音声による chat やいたずら電話対策などへの応用も考えられる．

謝辞 音素ラベル変換の実装に加え，有益な議論をして頂いたハラルド シンガー研究員をはじめとする ATR 音声翻訳通信研究所の皆様に感謝致します。

参考文献

[1] Kazuo Hakoda, Tomohisa Hirokawa, and Kenzo Itoh. Speech editor based on enhanced user-system interaction for high quality text-to-speech synthesis. In *Proc. ICSLP*, pp. 1775–1778, 1994.

[2] 近藤玲史, 滝澤卓也, 若尾淳, 稲垣敬子, 吉田和永, 三留幸夫. 合成音声をデザインするシステム – スピーチデザイナの試作 –. 音響学講論, pp. 223–224, Mar. 1997.

[3] N. Campbell and A. 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, SP96–07, May 1996.

[4] A. Black and N. Campbell. Optimising selection of units from speech databases for concatenative synthesis. In *Proc. Eurospeech95*, pp. 581–584, Apr. 1995.

[5] Biing-Hwang Juang Lawrence Rabiner. 音声認識の基礎(下). NTT アドバンステクノロジ株式会社, 1995.

[6] Ashok K. Krishnamurthy and Donald G. Childers. Two-channel speech analysis. *IEEE Trans. on Acoustics, Speech and Signal Proc.*, Vol. ASSP-34, No. 4, pp. 730–743, Aug. 1986.