日本語文音声認識システムを利用した音声圧縮・編集方法の提案

6G－10

西村雅史　阪本正治　大嶋良明　斎藤隆　鈴木和洋
（日本アイ・ビー・エム　東京基礎研究所）

１．はじめに

近年，音声認識の分野では，音素ＨＭＭのようなサブワードを認識単位とした統計的連続音声認識手法の研究が活発である．発声単位等にまだ制約が残るものの，既に英語に対しては，ディクテーションシステムが実用化されるに至っている．一方，テキスト音声合成の分野でも，ピッチ同期波形重畳法（PSOLA）による音声合成手法が開発され，合成音の音質を，大幅に改善出来ることが示された[1],[2]．ただ，韻律については，まだ不自然な部分も多く，テキスト解析の精度向上とともに，今後の課題として残っている．

もし，音声認識システムを用いて発声内容のテキスト（読み），ならびに韻律情報を自動抽出し，これを既存のテキスト音声合成システムの言語解析部の出力と置き換えることが出来れば，情報圧縮の観点からは非常に効率のよい分析合成系を構築できる．ただ，これを一般的な音声の符号化法として用いるには，まだ，認識率が不十分であるが，ユーザーが入力内容の確認・修正を行う，ディクテーションシステムの利用を前提とした場合には，音声付きの文書データを効率良く作成でき，さらに，テキスト上で，音声の編集が可能になるという利点がある．また，同様にして，個人用の音声合成素片辞書を自動生成することも可能になるだろう．逆に，他人の素片辞書を用いることで，声質変換としての応用も期待出来る．

本論文では，このような音声認識を前提とする分析合成システムの構成について述べるとともに，文音声の韻律（音素継続長と，基本周波数）の自動抽出に関して基本的な検討を行ったので報告する．

２．システムの構成

本システムの構成を図１に示す．音声認識システムと，音声合成システムは共通のサブワード（ここでは音素）をそれぞれ認識，合成の基本単位としている．認識は，音素ＨＭＭを用いて行い，一方，合成は，PSOLAに基づいて音素波形の編集を行う．入力発声からは，音素記号列，各音素の継続時間長，強度，並びに有声音区間のピッチ周期（２点分の点ピッチ）を自動推定し，音声合成システムの入力とする．

これらの情報のデータ量は，それぞれを１バイトで表現するとすれば，１音素当たり，５バイト程度でしかない．また，発声内容のテキストと完全な対応が付くため，音声としての編集が容易である．

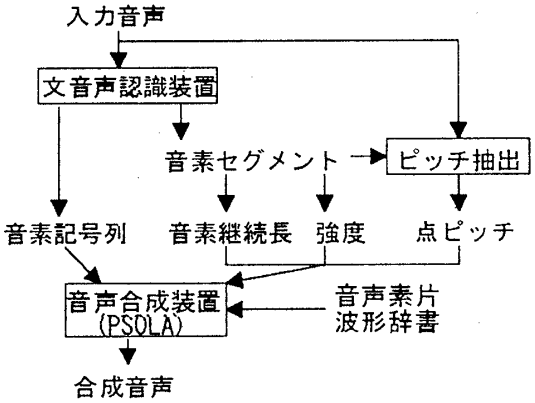

図１システム構成

２．１ 音素セグメンテーション

音声認識によって，音素セグメンテーションと同時に音素記号列を推定するが，誤認識が生じた場合はユーザーが必ず修正を行い，正解の音素列をシステムに教えるものと仮定する．このため，本稿の認識装置では，発声対象の文音声のみを認識するようにマッチングを制限し，音素セグメンテーションのみを行った．なお，このようにしても最適解が保証される探索手法を用いている限り，実際に認識を行った場合と同じセグメンテーション結果を得ることが出来る．

認識には前後音素環境を考慮したトライフォンモデルを用いた．トライフォンモデルはモデル数が膨大になるので，その対策としてデシジョントリーによる音素環境クラスタリングを行っている[3]．なお，各モデルの状態数は１で，継続時間長制限付きのモデルになっている．

２．２ ピッチ抽出

PSOLAに基づく音声合成では，音声素片波形辞書に，ピッチ開始位置の情報（ピッチマーク）を与える必要がある．ピッチ抽出法としては，辞書の自動生成にも適した方法であることが望ましい．我々は，波形の接続性を考慮して，声門閉鎖点にピッチマークを与えることとし，Kadambeらによって提案された，二進ウェーブレット変換によるピッチ抽出法を適用した[4]．また，二進ウェーブレット変換の極大点検出のためのしきい値設定

A Study on Speech Compression and Editing Using Continuous Speech Recognition System
By M.Nishimura, M.Sakamoto, Y.Ohshima, T.Saitoh and K.Suzuki
(Tokyo Research Laboratory, IBM Japan Ltd.)

は統計量に基づいて自動決定するように改良を加えた．

なお，合成時に必要なピッチ情報は，音素毎に開始位置と，中央部の２点分だけで，これを合成装置側で線形補間して用いている．

２．３ 音声合成装置

音素記号列，各音素の継続長，点ピッチ（２点分），強度の情報に基づいて，音声素片波形辞書を探索し，適切な素片をPSOLAの手法に基づいて，変形，接続して合成音声を得る．なお，辞書には，音声波形とともに，セグメント情報，音素環境，ピッチマーク等の情報が付与されている．

３．評価実験

本システムの実現可能性について，特に，韻律情報の自動抽出精度に着目して検討を行った．

３．１ 韻律の自動抽出結果

ATRの音声データベースのうち，男性話者１名（MTK）の３００文の自由発声を用いた．この内，２５０文で音素ＨＭＭの学習を行い，残り５０文で，音素セグメンテーション精度，およびピッチ抽出精度について調査した．なお，モデルの訓練は視察のセグメントデータに基づき，音素単位で行っている．結果を表１および，表２に示す．

表１　継続時間長の推定結果

	平均長	平均絶対誤差
子音	64.3msec	13.2msec
母音	109.1msec	15.4msec

表２　ピッチ抽出結果

ピッチ総数	脱落誤り	挿入誤り
20,103	415(2.1%)	155(0.8%)

３．２ 聴取実験結果

訓練に用いなかった５０文の中から，無作為に１０文を選び，韻律の自動抽出と音声合成を行い，自然性に関する主観評価実験を行った．ただし，音声素片波形辞書は，対象文の発声から作成している．つまり，音素環境を満たし，韻律の変更が最小限で済む素片が波形辞書にあったとして，PSOLA化と，点ピッチの影響，及び韻律自動抽出の影響について調べた．また，この際，辞書のピッチマークはすべて視察で修正した．

聴取対象は１）原音声，２）韻律情報を視察によって修正した合成音声，３）自動で抽出された韻律情報を使った合成音声の３種類である．被験者は，男女各５名で，声質，韻律等，全てを含めた自然性に対して，表３に示す５段階の尺度で評価を行った．結果を表４に示す．視察と自動抽出の差は大変小さく，韻律の自動抽出は，必要十分な精度が得られていると判断出来る．一方，PSOLA化と点ピッチによる劣化も，自然性を大きく損なうほどのものではなかった．

また，実際に合成に用いた点ピッチに関して，自動抽出値と視察値との関係を図２に示す．これを見る限り，ピッチの推定は安定していると言えるが，一部の値が視察値とは大きく異なっている．この多くは信号の強度が小さく，継続時間も短い/r/等の子音で，視察によっても，確実なピッチを付与するのは難しかった．なお，聞こえにはそれほど大きな影響を与えていない．

表３ 主観評価尺度

5	非常に自然
4	自然
3	少し不自然
2	不自然
1	非常に不自然

表４ 主観評価実験結果

原音声	４．７
視察で付けた韻律	３．５
自動抽出した韻律	３．３

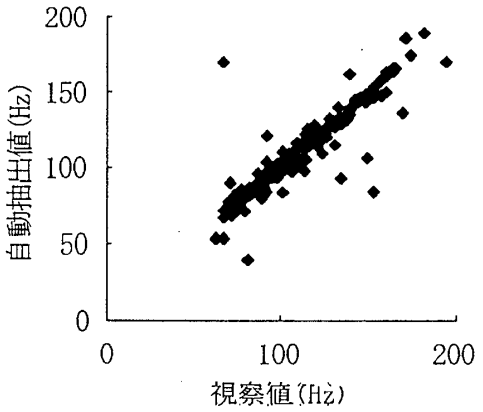

図２　ピッチの視察値と自動抽出値の分布

４．おわりに

今回の報告では，合成用音声素片データの選択方法については検討を行うことが出来なかったが，韻律の自動抽出という点に関しては，原音声との比較においても十分な自然性を持つ音声を合成出来ることが分かった．自然発声のディクテーションシステムが実用化されるようになれば，音声メール等への応用が可能だと思われるが，それ以前にも，規則合成では実現できないような自然な韻律を簡単に与えることが出来るという利点を生かし，録音に特定の話者を必要とするようなアプリケーションの省力化に役立つのではないかと考えている．また，合成用音声素片データの自動抽出等へ，積極的に応用していく予定である．

参考文献

[1] F.Charpentier, et al., Eurospeech89, 26-1 [2] 伊藤他，信学技報 SP93-121 [3] 大嶋他，情処第49回大会,6G-1 [4] S.Kadambe, et al.,IEEE Trans. Information theory, vol.38,No.2,pp.917-924,1992