# 韻律を用いた音声合成の単位選択 *

ニック キャンベル
ATR 音声翻訳通信研究所

## 1 はじめに

従来の音声合成の研究においては、韻律的特徴と調音的特徴は分離して考えられる場合が多かった。しかし、韻律的特徴を考慮せずに調音的特徴のみに着目して音声単位を選択した後、韻律の変換を行なうという方法はスペクトルに歪みを起こす原因となっており、韻律の変換を最小限にとどめることが望ましいと考えられる。また、幾つかの研究で、韻律的特徴、例えば音韻継続時間や基本周波数、パワーなどが調音的特徴に影響を与えることが指摘されている [1][2][3]。音声合成の品質向上に伴い、以前には問題とされなかったレベルでの韻律的特徴と調音的特徴の相互関係の詳細な検討が必要となっており、韻律的特徴に基づく単位選択方法を用いて音声合成を試みた。

音声合成用のデータベースは音韻ラベルのみに依存するのではなく、韻律ラベルを考慮することによって、より適切な単位選択が可能となる。しかし、自然発話による音声データの韻律には変動が大きいために、音声単位の接続によって音声合成を行なう方式では、多くの音声資料を必要とする。データベースの量が増加すれば合成音声は質的に向上するが、記憶容量の点で問題があり、データベースの適切な予備選択が不可欠である。そこでデータベース分類時の単位選択規準を確立することによって記憶容量を縮小することを試みる。

本稿では韻律的特徴に基づく単位選択の方法および評価について述べる共に、単位選択に適した音声データベースの構築方法について述べる。

## 2 データベース中の音声単位の選択

単独音韻の種類のみに着目した場合、音声データベース中の音声単位の種類（types）の数は少ないが、各種類に含まれる実際のデータ数（tokens）はさまざまである。例えば、 ATR 503 文データベースには「ア」の数が 6000 個（tokens）以上あるのに対し、「ピャ」は 3 個のみである。そこで、文章中での出現頻度を考慮し、高頻度で使われる音韻列を新たな音声単位として登録することにより、サイズ削減の影響を受けにくい音声データベースの構築を行なった。以下にその手順を述べる。

- 登録済みの音声単位毎に音声データベース中でのトークン数を求め、その値が最大の音声単位の中から頻度最大の音韻環境を求め、その音韻環境を含めて新たな音声単位として登録する。（この過程の繰り返しにより、タイプは増加し、トークンは減少する。その結果、タイプ／トークン配列は長方形配列となる。）
- 目的とする音声データベースのサイズにより、タイプ数から深さ $n$ を決定する。
- 各タイプのトークンを韻律次元で $n$ クラスターに分割し、各クラスターのセントロイドからの最近接トークンを保存する。

以上の手法により（図 1）、function word のように出現頻度の高い基本的音声ユニットはデータベースから自動的に浮上する。また、韻律分割の結果、典型的なトークンはすべて保存され、数量的に希少なトークンについても除外されることがない。他のデータベース抽出方法と比べ、この手法の選択ユニットはデータサイズの $n$ にかかわらず、常にデータベースの代表をすべて含むことが保証される。適切なデータベースサイズ $n$ の決定はデータベースサイズと出力音声のバランスに依存する。

## 3 音声合成時の単位選択

次にこのように作成した音声データベースを音声合成に用いる方法について述べる。まず、合成する音韻列に使用可能なトークンすべてを選び出す。次に単位の長さおよび音韻環境適合度と韻律適合度により評価し、各音韻毎に最高値を示すトークンのみを選ぶ。この時、各適合度は同じ重みを持つものとする。このようにして選ばれたトークンは時間的に重なり合うことがあり、この場合には韻律適合度を優先して、最適なトークンを選択する。

ここで用いた音韻環境適合度は前後の音韻の適合の度合（完全に一致、音韻クラスが一致、その他とし、先行音韻を優先した 9 段階の値を付与、音韻クラス数は 10）であり、韻律適合度は正規化音韻時間長、正規化基本周波数、正規化パワー（以上の各特徴量を音韻別に正規化した Z-score）である。

## 4 評価実験

まず、ATR503 文データベースを用いて本手法に基づき音声データベースを構築する。次に評価実験のために 503 文の中から 1 文をランダムに選択し、選択された文章から抽出された音声データが用いられないように音声単位を選択することにより合成音声を作成し

*Prosody and the Selection of Units for Concatenation Synthesis by Nick Campbell, ATR Interpreting Telecomm.

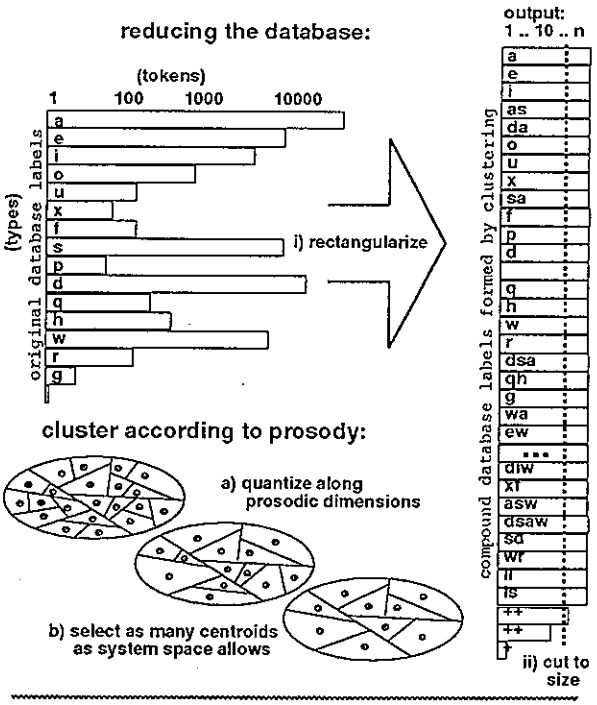


Fig. 1 Database operations

た。このような評価実験を100文について行なった。音韻時間長、基本周波数、パワーは、選択された文章の値を目標値として音声単位の選択を行なった。

高頻度の音韻列を考慮する以前の原データベース中の音声単位総数は25,264で、タイプ数は70であった。2で述べた音声単位の選択を行なった結果、635の音声単位（non-uniform units, 長さは1から7まで分布）ができた。深さ（$n$）は35とした。

ここでは評価のために以下の合成音声を作成した。

a) 音韻環境適合度のみを考慮し、音韻時間長、基本周波数、パワーを修正せずに音声データベースから選ばれた単位音声波形をそのまま接続したもの。

b) 音韻環境適合度と韻律適合度を考慮し、音韻時間長、基本周波数、パワーを修正せずに音声データベースから選ばれた単位音声波形をそのまま接続したもの。

c) (a)の波形をPSOLAを用いて原音声（ATR503文からランダムに選ばれた目的文）の音韻時間長、基本周波数、パワーに合わせたもの。

d) (b)の波形をPSOLAを用いて原音声（ATR503文からランダムに選ばれた目的文）の音韻時間長、基本周波数、パワーに合わせたもの。

これらの4種類の合成音声について原音声（ATR503文からランダムに選ばれた目的文）と合成音声との10msec毎の12次メルケプストラムのユークリッド距離を計算した。(a)および(b)の場合には各音韻の時間長が異なるため、音韻中心を一致させ、各フレーム間の距離を求めて、音韻のスペクトル距離とした。

表1に示す通り、韻律適合度を考慮して選択した場合、ケプストラム距離が0.86から0.43に減少しており（中央値）、PSOLAを用いない場合でも比較的高品質な音声が得られることが予備的な聴覚試験から確認されている。この結果からここで提案した手法の有効性が示されたものと考えられる。

韻律情報を使った場合、音韻環境情報のみを使用した場合よりケプストラム距離が減少しており（*(a) vs. (b)*: $t = 4.484$, $df = 6474$ ）、PSOLAを使うと、さらに減少することがわかった（*(b) vs (d)*: $t = 8.312, df = 6474. p < 0.001$）。

## 5 おわりに

上記の手法では、最適音韻列を選択するために韻律情報と音韻環境情報の両者を同等に活用した。音声単位の接続により得られた合成音声と同一話者が発話した自然音声とのスペクトル距離を比較した場合、韻律情報を用いた場合には韻律情報を用いない場合に比べて約半分に距離が減少していることが明らかになっており、本方式の有効性が示された。

多言語合成を目指して、この手法はこれまで日本語ならびに英語の音声合成に適用されており、今後さらに韓国語においても適用を試みる予定にある。

### 参考文献

[1] Traunmüller, H.: "Functions and limits of the F1:F0 covariation in speech", pp. 125-130 in PERILUS XIV, Department of Phonetics, Stockholm University, 1991.

[2] Di Benedetto, M.-G., "Acoustic and perceptual evidence of a complex relation between F0 and F1 in determining vowel height", pp 205-224, Journal of Phonetics, 22, 3, July 1994.

[3] Slijter, A. M. C., & van Heuven, V. J., "Perceptual cues of linguistic stress: intensity revisited", pp 246-249 in Proc. ESCA workshop on Prosody, Lund University, 1993.

表1：手法評価

Euclidean cepstral distance (quartiles)

| | min | 25% | median | 75% | max |
|---|---|---|---|---|---|
| a: | 0.007 | 0.397 | 0.858 | 1.722 | 8.855 |
| b: | 0.007 | 0.317 | 0.623 | 1.439 | 10.375 |
| c: | 0.013 | 0.241 | 0.584 | 1.189 | 9.119 |
| d: | 0.010 | 0.222 | 0.433 | 1.050 | 7.212 |