012058 木に基づく状態共有を利用した波形接続型音声合成法の検討

計算機工学講座C　山形 亮

1 はじめに

音声合成の手法の一つとして波形接続型音声合成がある．この手法は録音音声から音節単位で波形素片を取り出し，信号処理をせずに接続することで，自然性の高い音声が合成できる．しかし，任意の一般名詞を作成しようとするには大量の録音単語が必要である．音声データベースとして，ATR 単語発話データベース Aset(5,240 単語) を使用した場合，5,240 件中の 470 単語しか作成できない [1]．そこで本研究では，収録されている DB に対して木に基づくクラスタリング [3] を行い，音響パラメータが似た音節素片をグループ化する．クラスタリングにより得られた情報を利用して波形接続型音声合成を行うことで作成可能な単語数が飛躍的に増加する，一方で音質の劣化が懸念される為，オピニオン評価実験及び対比較実験を行い音質の評価を行う．

2 波形接続型単語音声合成

波形接続型音声合成 [1] では，以下の条件の音節素片を接続して音声を合成する．特徴として言語的なパラメータだけを用いて音節素片を選択し音響的なパラメータは使用しない．そのため，自然性の高い音を合成することができる．

- 中心の音節
- 直前の音素 (前音素環境)
- 直後の音素 (後音素環境)
- 単語のモーラ数
- 単語のモーラ位置
- 単語のアクセント型

3 木に基づく状態共有

木に基づくクラスタリング [3] は，音声認識で学習データを効率良く使うためによく使われている．音響的特徴が類似した triphone HMM の状態集合に対して音声の決定木に基づいてクラスタリングを行う．グループ化された音節素片の情報を使うことで任意の音を合成することが可能となる [2].

4 評価実験

収録された DB に対して木に基づくクラスタリングを行い，音響パラメータ的に似た音節素片をグループ化し，これを用いて波形接続型音声合成で合成音声を作成する．合成した音声の品質を調べるために，同一発話の自然音声と波形接続型音声合成で作成した音声を用意し聴覚実験を行う．

4.1 実験データ

本実験では音声データベースとして，ATR 単語発話データベース Aset(5,240 件) を使用する．そして，Aset に含まれる 3 または 4 モーラ語の各 50 音声 (計 100 音声) を利用する．そして以下の条件の音声を準備する．

1) 自然音声
2) オリジナル合成：木に基づくクラスタリングを利用しない波形接続型音声合成で作成した合成音声
3) クラスタリング合成：木に基づくクラスタリングを利用した波形接続型音声合成で作成した合成音声

4.2 評価方法

合成音声の評価のために，音声研究に関わった経験のない 4 名を対象に，自然音声と合成音声をランダムにヘッドフォンから被験者に聞かせ，オピニオン評価実験および対比較実験を行う．

5 実験結果

5.1 オピニオン評価の実験結果

実験結果を表 1 に示す．オピニオン評価で 3.7 というそれなりの値が得られた (評価音節数：100).

表 1 オピニオン評価の結果

音声の種類	オピニオンスコア
自然音声	4.9
オリジナル合成	4.3
クラスタリング合成	3.7

5.2 対比較実験の結果

次の三通りの対比較実験を行った．(1) 自然音声とオリジナル合成，(2) 自然音声とクラスタリング合成，(3) オリジナル合成とクラスタリング合成，その結果を表 2 に示す

表 2 対比較実験の結果

	比較対象 1	比較対象 2
(1)	自然音声：79.25%	オリジナル合成：20.75%
(2)	自然音声：90.75%	クラスタリング合成：9.25%
(3)	オリジナル合成：75.75%	クラスタリング合成：24.25%

オリジナル合成とクラスタリング合成の差は自然音声とオリジナル合成との差ぐらいである，したがってクラスタリング合成は自然性の高い音声であることがわかる．

6 考察

波形接続型音声接続では，人手でラベリングされた情報を利用して音節素片を切り出す．しかし，ラベルに誤りがある場合がある．このとき余分な前後の音声が入ることがある．

木に基づくクラスタリングを利用しない波形接続型音声合成では前後の音素を考慮している為に多少前後の音があっても違和感無く音声が合成できる．しかし，木に基づくクラスタリングを利用した合成音声では，グループ化された音素からランダムに音を選択している為に前後環境は考慮されていない．その結果，合成音声に違和感が生じ品質の低下に繋がったと考えている．

7 まとめ

本実験では，波形接続型音声合成で任意の合成音声を作成するために音声が収録されているＤＢに木に基づくクラスタリングを行い，クラスタリングされた音節素片を使用した時の合成音声の音質を調査した．

聴覚実験の結果，オピニオンスコアで 3.7 が得られた．クラスタリングにより条件を緩和したことで波形接続型合成音声よりも多少オピニオンスコアが下がったものの自然性の高い音声が作成でき，任意の音が作成可能であることがわかった．

今後は，木に基づくクラスタリングにアクセント位置を考慮に入れるなどの検討を行い，より自然性の高い合成音声の作成を目指したい．

参考文献

[1] 石田 隆浩，村上 仁一，池原 悟．“モーラ情報とアクセント情報を用いた波形接続型音声合成の普通名詞句への応用”，2-Q-18, pp.409-410(2003).

[2] 堀田 波星夫，村上 仁一，池原 悟．“不特定話者における同音異義語音声認識”，NO. 2-1-11(2006).

[3] S.J. Young, J.J. Odell, and P.C. Woodland. Tree-based state trying for high accuracy acoustic modelling. Proc. ICASSP, pp.307-312(1994).