波形接続型音声合成における位相情報を利用した音節境界位置決定方法*

☆橋本浩志，村上仁一 (鳥取大)

1　はじめに

音声合成の手法の 1 つとして音節波形接続型音声合成 [1] が提案されている．この手法は録音音声から音声波形を音節単位で分割し，接続することで合成音声を作成する手法である．信号処理を用いずに言語的な情報のみを用いて音節素片を選択するという特徴をもつ．

音声波形の音節開始位置・終了位置は，音節境界位置が記載されたラベルを利用している．ラベルは人手で作成されているが，波形接続用ではないため音節境界位置の精度が低い．そのため，音声合成時に波形の接続点が不連続になり，音声品質が劣化する．そこで波形が滑らかに接続するように波形接続時に人手で音節開始位置・終了位置に修正を加えている．しかし，この修正作業にはコストがかかる [2]．

そこで，昨年度の研究 [3] では音節の精密な開始位置・終了位置を自動的に決める方法を提案した．具体的には，音節素片のパワーが最大となる周波数を求め，その初期位相が “-$\frac{\pi}{2}$” となる時間を音節開始位置にする．しかし，離散フーリェ変換の窓長を音節素片の音節開始位置における 1 周期の整数倍にしなければ，初期位相に誤差が生じることがわかった．

本研究では，離散フーリェ変換の窓長を音節素片の開始時から 1 周期の整数倍にするために，誤差のある窓長の前後 0.5ms(約 0.0625ms 間隔，計 17) に対して離散フーリェ変換を行い，振幅が最も 0 に近い音節開始位置を選択する．そして，提案方法を用いて合成音声を作成し，音声品質を調査する．

2　音節波形接続型音声合成

2.1　音節波形接続方式における音声合成方法

音節波形接続型音声合成における具体的な音声合成方法 [1] を以下に示す．

1. 作成したい単語の音節素片選択条件 (Table1 に示す．) を求める．
2. 1. で指定した条件と一致する音節素片を，ATR の単語発話データベース中から選択する．
3. 選択した音節素片の開始位置・終了位置を ATR ラベルに記載されている音節境界位置から得る．
4. 接続部が滑らかに繋がるように音節開始位置・終了位置を人手で修正する．
5. 修正した音節開始位置・終了位置を用いて音節素片を切り出し，音声合成を行う．

Table 1　音節素片の選択条件

1. 中心の音節
2. 直前の音素 (前音素環境)
3. 直後の音素 (後音素環境)
4. 単語のモーラ数
5. 単語のモーラ位置
6. 単語のアクセント型
7. 単語のアクセントの高低

「乗り物 (no/ri/mo/no)」という合成音声を作成する際の例を Fig. 1 に示す．Fig. 1 の「‾|‾」はアクセントの高低を表す．また，太文字で示している部分は，接続する音節素片である．

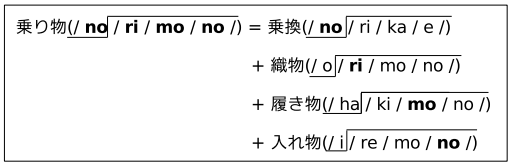

Fig. 1　音節波形接続型音声合成の音声合成例

2.2　音節境界位置の問題点

音節波形接続型音声合成では，人手で作成されたラベルを利用して各音節の開始位置・終了位置を決める．本研究では，ATR から提供されているラベル (以下，ATR ラベル) を使用する．しかし，ATR ラベルの音節開始位置・終了位置は波形接続型音声用に正確に記載されていない．そのため，音声合成時に接続点に不連続が生じる．そこで音節波形接続型音声合成では人手で，合成時に接続部が滑らかに繋がるように音節開始位置・終了位置を修正している [4]．

ATR ラベルを用いた合成例を Fig. 2 に示す．Fig. 2 は合成音声「形見 /ka/ta/mi/」の “ta/mi” の音声波形である．また，図中の縦線部は “ta” と “mi” の接続部を示している．接続部に注目すると，滑らかに接続されていないことがわかる．

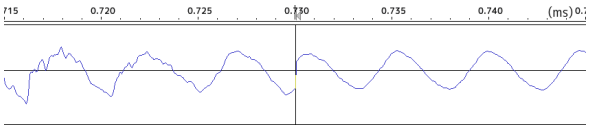

Fig. 2　ATR ラベルの音節境界位置を用いて「形見 /ka/ta/mi/」を音声合成した例

2.3　昨年度の研究

2.2 節の問題点を解決するために，昨年度の研究 [3] では，音節開始位置を音節素片のパワーが最大となる周波数の初期位相が “-$\frac{\pi}{2}$” となる位置に決めた．初期位相 “-$\frac{\pi}{2}$” は音声波形の振幅が “-” から “+” に変わる点である．音節素片は開始位置と終了位置で周波数が異なるため，音節開始位置から 4ms(約 1 周期) を窓長として，離散フーリェ変換を行う．しかし，音節素片の音節開始位置における 1 周期の整数倍の周期に対してフーリェ変換を行わなければ，推定したパワー最大の周波数に必ず誤差が生じ，初期位相に誤差が生じる．

そこで本研究では，離散フーリェ変換の窓長を音節素片の開始位置における 1 周期の整数倍にするため，2.4 節で示す方法を行う．

2.4　位相を用いた音節開始位置変更方法

本研究の提案手法を具体的に以下に示す．

1. ATR のラベルから得られる音節境界位置を用いて音節開始位置・終了位置を得る．
2. 音節素片全体の窓長に対して FFT(離散高速フーリェ変換) を行う．

*Syllable Boundary Position using DFT for Corpus Based Speech Synthesis. by HASHIMOTO Hirosh and MURAKAMI Jin'ichi(Tottori University)

3. FFT によって得られるパワーが最大時の周波数 F_{max} における 1 周期の波長 (ms) を求める．
4. 求めた波長の前後 0.5ms(約 0.0625ms 間隔，計 17) の窓長で，音節素片の音節開始位置から DFT(離散フーリェ変換) を行う．
5. それぞれの窓長に対する DFT の結果から得られるパワーが最大時の周波数 F_{max} における初期位相を求める．
6. それぞれの初期位相 ϕ を “-$\frac{\pi}{2}$” にするように音節開始位置を変更する．変更する時間幅 T は $T = \frac{\frac{1}{F_{max}} * P_0}{2\pi}$ になる．また，初期位相 ϕ は連続値で扱うので，サンプリング周波数 16kHz の音声で音節開始位置を変更するため，音節開始位置の変更値は約 0.0625ms 毎の離散値になる．
7. 求めた音節開始位置の振幅の中から最も 0 に近い音節開始位置を選択する．

2.5 音節終了位置変更方法

音節の終了位置は，次の音節の開始位置を 2.4 節の手順 1. から手順 7. を行うことによって決める．

2.6 音節開始位置・終了位置変更後の波形例

上記の方法を用いて音節開始位置・終了位置の変更を行い，作成した合成音声の音声波形例を Fig. 3 に示す．Fig. 3 は合成音声「形見 /ka/ta/mi/」の “ta/mi” 間の接続部の音声波形を示す．図中の縦線部は接続部を示している．接続部に注目すると，Fig.2 と比較して滑らかに接続されていることがわかる．

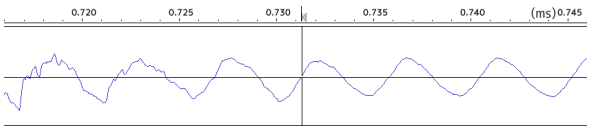

Fig. 3 提案方法を用いて「形見 /ka/ta/mi/」を音声合成した例

3 聴覚実験

3.1 実験データ

本実験では，ATR 単語発話 DB Aset(5,240 単語) 内の女性話者 2 名 (FYN, FTK) を利用する．聴覚実験には，以下の 3 種類の合成音声を用いる．

- 提案方法を用いて作成した合成音声 (以後，提案合成音声).
- ATR ラベルから得られる音節境界位置を用いて作成した合成音声 (以後，ATR 合成音声).
- 波形を滑らかに接続するように人手で作成した合成音声 [4] (以後，人手合成音声).

なお，接続部に注目して音声品質の違いを調査するため，3 種類の合成音声にはすべて同じ単語を使用する．

3.2 評価方法

合成音声の音声品質を評価するために，音声研究に関わった経験のない 5 名を対象に，オピニオン評価実験及び対比較実験を行う．評価単語数は各 100 単語とする．また，オピニオン評価は自然音声を含めて行う．

4 実験結果

4.1 オピニオン評価実験

提案合成音声の音声品質を調査するために，オピニオン評価実験を行う．結果を Table2 に示す．

Table 2 オピニオン評価の結果 (各評価単語数 : 100)

音声の種類	スコア (fyn)	スコア (ftk)
提案合成音声	4.00	3.82
人手合成音声	3.92	3.81
ATR 合成音声	3.22	3.32
自然音声	4.61	4.59

4.2 対比較実験

提案合成音声と従来合成音声の音声品質を比較するために，次の 2 種類の対比較実験を行う．
(1) 提案合成音声と ATR 合成音声
(2) 提案合成音声と人手合成音声
結果を Table3 に示す．

Table 3 対比較実験の結果 (各評価単語対 : 100)

	比較対象 1	比較対象 2
fyn	提案合成音声 : 77%	ATR 合成音声 : 23%
	提案合成音声 : 50%	人手合成音声 : 50%
ftk	提案合成音声 : 75%	ATR 合成音声 : 25%
	提案合成音声 : 51%	人手合成音声 : 49%

4.3 実験結果のまとめ

オピニオン評価実験と対比較実験の双方の結果により，本研究の有効性が証明された．Table2，Table3 の結果より，提案合成音声と人手合成音声の音声品質にはほとんど差がないことがわかる．

5 考察

提案合成音声の音声品質が，ATR 合成音声と比較して劣化した音声波形を調査した．Fig. 4 に例を示す．Fig. 4 は「市街/shi/ga/i/」の “ga/i” の提案合成音声である．

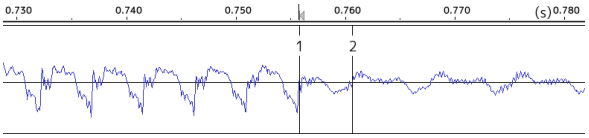

Fig. 4 音声品質が劣化した音声波形 (合成音声「市街 /shi/ga/i/」) の例

Fig. 4 の縦線部 1 は，提案合成音声の接続部を示している．しかし，音節の 1 周期の開始位置で接続されていない．Fig. 4 の縦線部 2 は，ATR 音声合成の接続部を示している．音節の 1 周期の開始位置で接続されている．したがってこの場合，提案合成音声の音声品質が ATR 合成音声の音声品質と比べて低下していることがわかる．

6 おわりに

本研究では，音節開始位置・終了位置を波形接続方式用に正確に決める方法を提案した．その結果，提案合成音声は人手合成音声と同等の音声品質を得ることができた．今後は，ピッチ同期の方法を用いて音節開始位置・終了位置を決めることで，音声品質の比較調査を行いたいと考えている．

謝辞 本論文を執筆するにあたり，聴覚実験に協力してくださった研究室の方々に深く感謝いたします．

参考文献

[1] 村上他: “音節波形接続方式による単語音声合成”, 信学論 D-II, Vol.J85-D-II, No.7, pp.1157-1165，2002.
[2] 居村他: “波形接続型音声合成のフレーズへの適用”, 言語処理学会第 14 回年次大会, pp.965-968, 2008.
[3] 橋本他: “波形接続型音声合成における位相情報を利用した音節境界位置の決定方法”, 日本音響学会 2009 年秋季研究発表会，pp.411-412, 2009.
[4] 石田: “アクセントを考慮した波形接続型単語音声合成”, 鳥取大学大学院工学研究科修士論文，2004.