アクセントを考慮した波形接続型単語音声合成*

○村上仁一 石田隆浩 池原悟 （鳥取大学工学部）

1 まえがき

音声合成の手法として近年注目されている波形接続型音声合成は，大量の録音音声から音素や音節を単位とした波形素片を取り出し，接続することによって合成音声を作成する．信号処理せず接続することで，話者性と高い自然性を保てる特徴がある．また，波形接続型音声合成を用いた単語音声合成において，前後音素環境やモーラ情報などの付加的な情報を用いることで，品質が向上することが知られている [1].

例えば，音節を単位とした音節波形接続方式 [1] では，地名を対象に実験した結果，実用的な品質が得られたことが報告されている．また，同様の手法を普通名詞に適用した場合も，明瞭性の高い合成音声を作成できたことが示されているが，アクセント型のばらつきによる自然性劣化も指摘されている [2].

また，波形接続型音声合成では基本的に信号処理を行わないため，素片単位や接続位置，最適な素片を選び出す方法なども非常に重要となる．

そこで本研究では，普通名詞を対象として，音素，または音節を素片単位とした場合に，アクセントを考慮することで合成音声の品質をどの程度改善できるかについて調査する．そして，波形接続型単語音声合成におけるアクセントの有効性および最適な素片単位について検証する．

2 波形接続型単語音声合成

2.1 モーラ情報，アクセントと F_0 周波数

波形接続型音声合成においては，一般に韻律の扱いが問題となる．しかし，録音音声波形の正確な F_0 周波数を直接推定することは困難である．

一方，特定話者の固有名詞発話において，単語のモーラ数とモーラ位置（合わせて「モーラ情報」）が決まれば，F_0 周波数がほぼ決定できることが知られている [1] が，特に普通名詞においては，アクセントの未考慮による自然性劣化が指摘されている [2]．アクセントがモーラ情報と F_0 周波数の関係に影響を及ぼすことも報告されている [?]．そこで本研究では，素片選択においてモーラ情報および単語のアクセントを考慮する．

2.2 波形接続型単語音声合成の概説

本研究で用いる波形接続型音声合成では，まず，以下の情報が一致する素片を選択する．単語のアクセントについては，NHK 日本語発音アクセント辞典 [6] を参考にラベルデータに対してアクセントを付加する．

- 音素または音節
- 直前の音素 (前音素環境)
- 直後の音素 (後音素環境)
- 単語中のモーラ位置
- 単語のモーラ数
- 単語のアクセント

そして，素片の開始時間と終了時間を元に波形データを切り出し，接続して合成音声を作成する．

2.3 合成音声の例

本研究で作成した合成音声「伝説」(/de/N/se/tsu/) について，音節単位で合成した例を以下に示す．なお,「_ | ¯」は音の強弱（アクセント）を表している．また，括弧内強調部は，実際に選択される部分を示している．

伝説 (/de/N/se¯/tsu¯/) = 伝票 (/de/N/pyo¯/u/)
+ 伝染 (/de/N/se¯/N/)
+ 建設 (/ke/N/se¯/tsu¯/)
+ 鋼鉄 (/ko/u¯/te¯/tsu¯/)

3 評価実験

3.1 実験環境

本研究では，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 3，4 モーラ語名詞について，以下の条件で各 100 音声（3，4 モーラ語各 50 音声）を準備する．

- 自然音声
- 音節を単位とした波形接続型合成音声（syl）
- 音素を単位とした波形接続型合成音声（tri）

また，波形接続型音声合成による合成音声では，アクセントを未考慮のもの（na）と，考慮したもの（ac）を準備する．

話者には，F_0 周波数のばらつきが比較的小さく，収録された音声にエコーの少ない，FTK と FYN の 2 話者を選ぶ．

3.2 波形接続に関する補則

波形接続型音声合成では，接続部の違和感の発生が自然性に大きく影響する．本研究では，波形の接続位置を音素境界または音素中心とし，さらに，接続部における 2 つの音節波形の位相を考慮し，接続部の振幅の差がゼロに近づくように調整を行う．音素境界と音素中心のどちらがより適しているかについて，評価実験により明らかにする．

3.3 評価方法

合成音声の評価のために，20 代男性 5 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聴かせ，了解度試験とオピニオン評価を行う．評価は，作成した単語を文に埋め込んで行うのではなく，単語音声のみで行う．

(1) 了解度試験

単語音声の明瞭性を調べるために了解度試験を行う．了解度試験では，どのように聞こえたかを仮名で書き取らせる．自分の知識を用いず，聞こえたとおりに書き取るように指示する．

(2) オピニオン評価

単語音声の自然性を調べるためにオピニオン評価を行う．オピニオン評価では，自然に聞こえた度合を 5 段階（5 が最も自然，1 が最も不自然）で評価するように指示する．

---

* “ Common Noun Word Synthesis by Concatenating Syllabic Components using Accent ” by Jin’ichi Murakami and Takahiro Isida and Satoru Ikehara (Faculty of Engineering Tottori University)

4　実験結果

実験結果を表 1，表 2 に示す．

表 1: 実験結果 (1): 接続位置＝音素境界

	了解度　正解率 (%) 評価音節数: 1,750			オピニオンスコア 評価単語数: 100		
	FTK	FYN	平均	FTK	FYN	平均
自然音声	99.7	99.9	99.8	4.8	4.9	4.9
syl(na)	99.4	99.7	99.6	3.8	4.1	4.0
syl(ac)	99.4	99.7	99.6	4.2	4.4	4.3
tri(na)	98.2	99.3	98.8	3.4	4.0	3.7
tri(ac)	99.1	99.5	99.3	4.0	4.2	4.1

表 2: 実験結果 (2): 接続位置＝音素中心

	了解度　正解率 (%) 評価音節数: 1,750			オピニオンスコア 評価単語数: 100		
	FTK	FYN	平均	FTK	FYN	平均
自然音声	99.7	99.9	99.8	4.8	4.9	4.9
syl(na)	98.4	99.6	99.0	3.4	3.8	3.6
syl(ac)	99.0	99.7	99.4	3.6	4.0	3.8
tri(na)	98.3	98.6	98.5	3.1	3.6	3.4
tri(ac)	98.2	99.1	98.7	3.4	3.8	3.6

表 1，表 2 から，了解度はアクセントの考慮，未考慮に関わらず，非常に高い正解率を得た．一方，オピニオンスコアは，アクセントを考慮することで改善することが分かった．また，音素を単位とした場合より音節を単位とした方が良くなり，音素中心で接続するより音素境界で接続した方が良い結果が得られた．しかし，いずれの条件下でも，了解度は自然音声と同程度であったが，オピニオンスコアは及ばなかった．

5　考察

5.1　自然音声と合成音声の差

実験結果から分かる通り，波形接続型音声合成による合成音声は，了解度は自然音声と同等の正解率が得られたが，自然性は依然として低いままであった．アクセント考慮時のオピニオンスコアは音節単位で 4.3，音素単位で 4.1 であり，3.3 節で述べたオピニオン評価の基準によると，音量やアクセントは正常だが，接続部の違和感が残った状態であると言える．

よって，被験者が自然音声と合成音声との差を感じる最も大きな原因は，接続部における違和感である．また，スコアが音素単位より音節単位の方が高いことからも，接続部が少ない方が違和感が発生する可能性が低くなり，自然性が向上すると考えられる．

5.2　了解度試験の解析

了解度試験において，波形接続型音声合成で作成した音声で，多くの被験者が間違えた音声について表 3 に示す．

表 3: 了解度試験における間違いの例

	正解	間違いの例
f1	便利 (べんり)	でんり
f2	無断 (むだん)	むらん
f3	売店 (ばいてん)	ばいて_
f4	来年 (らいねん)	らえねん
f5	黒板 (こくばん)	ここばん

表 3 から分かるとおり，例えば「べ」と「で」のように，発声方法の似た音，特に子音部分を間違える場合が多かった．また，最終モーラの撥音を聴き逃す被験者が多かった．母音では,「い」と「え」,「う」と「お」を間違える場合があった．

5.3　オピニオン評価の解析

FTK，FYN 共に評価の悪かった音声のうち，接続部の違和感やアクセントの不自然さ以外の原因としては，特に，音量に違和感のある音声が挙げられる．これは，音声コーパスの録音音声の音量にばらつきがあることが影響していた．例えば「会話」という音声では,「対話」と「内輪」の音量差により，接続時に第 2，3 モーラ間において音量が極端に変わり，違和感が発生した．しかしこの問題は，音声コーパスの作成時にあらかじめ音量を揃えておくことで解決可能であると考える．

また，各音素，または音節の継続時間長の問題により，違和感を感じる場合があった．例えば「結局」という音声において，第 2 モーラの促音の継続時間が極端に短くなっていた．本研究では継続時間長の制御は行わなかったが，今後，さらなる自然性の向上のためには，継続時間長の制御が重要である．

5.4　素片単位について

実験結果から，音素単位より音節単位の方が評価が高くなることが分かった．しかし，音節を単位とした場合，前後音素環境を考慮した場合の素片種類数は，一般に音素を基本単位とした場合より多くなる．そのため，音声コーパスの大きさが決定されている場合，作成できる音声は少なくなってしまう．今回の実験で作成できた音声は，音素単位で約 650 単語，音節単位で約 470 単語であり，約 180 個もの差があった．

この問題に対しては，特に後音素環境において似た子音をグループ化し，音素環境を代替して素片の種類数を少なくすることで解決していくことが可能であると思われる．なお，環境などは異なるが，単語音声より短い刺激音声について，代替による自然性劣化に関する評価がなされている [3].

5.5　素片接続位置について

素片の接続については，本研究では音素中心より音素境界で行った方が良い結果が得られた．しかし，特に母音連続の場合において，接続歪みの頻度分布は音素中心と音素境界ではほぼ同じであると報告されている [4]．実験条件等は異なるが，今後さらなる比較検討が必要であろう．

6　まとめ

本研究では，波形接続型単語音声合成におけるアクセントの有効性を検証した．聴覚実験における合成音声の単語了解度は，音素単位の場合で 99.3%，音節単位の場合で 99.6%が得られた．また，オピニオンスコアはそれぞれ 4.1，4.3 であった．またアクセントを考慮することで，オピニオンスコアはそれぞれ 0.4，0.3 向上した．これらのことから，普通名詞において波形接続型音声合成を適用した場合，アクセントが，自然性の向上のために有効であることが分かった．一方，自然音声の単語了解度は 99.8%，オピニオンスコアは 4.9 であり，波形接続型単語音声合成は，自然音声には及ばなかった．

今後の研究課題としては，さらなる自然性の向上を目指して，様々なパラメータを導入する必要がある．また，音声コーパスが同一でも，作成できる単語数を増加させるために，パラメータのグループ化，特に後音素環境における子音のグループ化を検討していくことが重要であると思われる．

参考文献

[1] 村上，水澤，東田，“音節波形接続による単語音声合成，” 信学論 D-II Vol.J85-D-II No.7 pp.1157–1165 (2002-7).

[2] 石田，村上，池原，“音節波形接続型音声合成の普通名詞への応用，” 信学技報，SP2002-25，pp.7–12 (2002-5).

[3] 河井，津崎，舛田，“波形素片接続時の音素環境代替による自然性劣化の知覚的評価，” 信学技報，SP2001-22，pp.51–57 (2001-5).

[4] 戸田，河井，津崎，鹿野，“日本語テキスト音声合成における音素単位とダイフォン単位に基づいた単位選択，” 信学論 D-II Vol.J85-D-II No.12 pp.1760–1770 (2002-12).

[5] 石田，村上，池原，“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への応用，” 日本音響学会 2003 年春季研究発表会，2-Q-18，pp.1-409,410 (2003-03).

[6] “NHK 日本語発音アクセント辞典 新版”，NHK 出版，ISBN4-14-011112-7 (1998).