概 要

音声合成の手法として近年注目されている波形接続方式は，大量の録音音声から音素や音節を単位とした音声素片を取り出し，接続することによって合成音声を作成する．音節波形接続方式は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．例えば，音節を単位とする音節波形接続方式では，固有名詞と普通名詞を対象に実験をした結果，実用的な品質が得られたことが報告されてる [1][2]．また，普通名詞の場合には自然性向上のためにアクセント型が有効であることも報告されている [3]．

そこで本研究では，文節発声でゆっくりと発話された音声を使用し，音節波形接続方式を文節に適用し，有効性を調査した．

音節波形接続型音声合成では基本的に信号処理を行わない．加えて，文節は名詞に比べアクセントが複雑である．そのため，素片単位や適切な素片を選び出す方法などが非常に重要となる．

そこで，品質向上のために，従来の音節波形接続方式に音節選択の条件を 2 つ追加した．1つ目として，接続部において不自然さの軽減のために，母音と撥音が連続する部分を連続母音として扱った．また 2 つ目として，条件を満たす素片の中から録音した時間帯が近い音声を選んで音声を作成した．そして，作成した合成音声の品質を調査した．

その結果，聴覚実験における合成音声の了解度は条件を追加していない場合が 98.7 %，条件を追加した場合が 99.3 %となった．これは自然音声の 99.3 %と比べても同程度の高い値となった．また，オピニオンスコアは条件を追加しないものが 3.55 であったのに対し，条件を追加した場合は 3.83 となった．そして，対比較実験においても条件を追加した音声としない音声では，60.7 %が条件を追加した音声の方が自然だと判定され，条件を追加することが自然性の向上に有効であることが分かった．

一方，自然音声はオピニオンスコアが 4.75，条件を追加した合成音声との対比較においても 74.3 %が自然音声の方が自然だと判定されており，自然性の面では合成音声は自然音声には及ばなかったが，高い品質の合成音声が作成可能であることが分かった．

目 次

図 目 次

表 目 次

1　はじめに

現在，カーナビゲーションシステムや電車の車内アナウンスなどのように，音声ガイダンスを利用したシステムやサービスが様々な場面において利用されている．このようなシステムでは，録音編集方式が広く使われている．録音編集方式では，まず，システムやサービスに必要となる音声を，システム利用者の入力やサービスの利用される場所・時間などに依存するような比較的短い単語音声（以下,「可変部」）と，それ以外の比較的長い文節・文音声（以下,「固定部」）に区別する．そして，可変部と固定部を別々に録音しておき，必要に応じて組み合わせることで出力音声を構築する．

例えばカーナビゲーションシステムにおいて,「目的地は○○でよろしいですか.」というガイダンス音声を出力したい場合，○○の部分には，駅名や建物名などの単語音声が挿入される．ユーザーが目的地に「東京駅」を指定した場合，ガイダンス文は「目的地は“東京駅”でよろしいですか.」となる．例の場合,「東京駅」などの駅名や建物名などの単語音声が可変部,「目的地は～」という部分が固定部となる．

録音編集方式を用いた音声合成においては，可変部と固定部を接続した場合の違和感を軽減するために，一般に同一話者の音声が必要となる．可変部と固定部を分離して録音することにより，必要となるすべての音声を録音する場合に比べて話者に対する負担は若干軽減されるが，可変部に挿入する単語が増大した場合，同一話者から全ての音声を録音することは困難となる．さらに，録音環境の違いにより発話速度や F_0 周波数にばらつきが出るため，安定した品質の音声を得ることは非常に困難となる．

そこで，可変部や固定部に必要になる音声をすべて音声合成によって作成する方法が考えられる．例えば，音素や音節を単位とした規則音声合成がある．規則音声合成は，古くから TTS 音声合成において用いられてきた方法であり，基本的には，音声の特徴をパラメータとして抽出し，変形することによって合成音声を作成する．PSOLA 方式による音声合成については，現在も多くの研究がなされている．また，最近では HMM を用いて直接音声を合成する研究も行われている [4][5]. しかし，いずれの場合においても，直接人の声を録音した音声のように，高い品質を安定して得ることが難しい点が問題である [6].

一方，録音した音声波形の一部（以下「音声素片」）を用いて別の音声を合成する方法がある．その代表的な手法が CHATR[8] である．CHATR は，あらかじめ合成したい話者の音声を録音しておき，そこから部分的に切り出した音声波形を信号処理をせずに音声を合成する手法である．

これと似た手法で単語の音声合成を行うのが音節波形接続方式である．音節波形接続方式は，あらかじめ録音しておいた音声波形を，音素単位や音節単位などで分割し，接続することで合成音声を作成する方法である．この場合，波形に信号処理を行わずに接続することにより，話者性と高い自然性が保たれる特徴がある．

しかし，波形接続型音声合成においては，音声波形に信号処理を加えないため，韻律の扱いが問題となる．尤も，波形接続型音声合成に限らず，一般に音声合成

において，韻律制御は重要な課題であるが [9]，音声合成の対象として小さな単位である単語を合成する場合においては，地名などの固有名詞では F_0 周波数のばらつきが比較的小さく，アクセント型がほぼ一意に決まるため，F_0 周波数とモーラ情報の依存関係を効果的に利用することが可能である [1]．そして，固有名詞を対象とした実験では，実用的な品質が得られたことが報告されている．また，普通名詞に適用した場合も，明瞭性の高い合成音声が作成でき，さらにアクセント情報としてアクセント型を考慮することで，より自然音声に近い合成音声の作成が可能であることが示されている [2][3]．

本研究では，文節に対して音節波形接続方式を適用し，有効性の確認を行う．ただし，文節は名詞単体に比べてアクセントが複雑になるため，通常の発話の音声では音声合成が困難だと考えられる．そこで，本研究では実験に文節発声で発話速度が遅い音声を使用する．また，作成した合成音声の問題点から，音声波形の選択条件を追加し，さらに自然音声に近い合成音声の作成を目指した．

以降，2 章で音節波形接続方式を用いた音声合成について説明する．そして，3 章で評価実験に関する説明を行い，4 章で実験結果を報告する．実験により現れた問題点を 5 章で述べ，6 章で改善策を提案する．そして，改善案に沿った実験を 7 章で行い，結果について 8 章で考察する．

2　音節波形接続方式

2.1　モーラ情報とアクセント情報

波形接続型音声合成を含め，一般に音声合成においては，韻律の扱いが問題となる．韻律を扱う場合，録音音声および出力音声の F_0 周波数が必要となる．しかし，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，音声合成の対象として小さな単位である単語を合成する場合においては，地名などの固有名詞では F_0 周波数のばらつきが比較的小さく，アクセント型がほぼ一意に決まるため，F_0 周波数とモーラ情報の依存関係を効果的に利用することが可能である [1]．そして，このモーラ情報は音素ラベリング [10] や音声認識 [11][12] などの分野において効果があることが報告されている．

しかし，より一般的な普通名詞では例えば「雨」と「飴」のように同音異義語が多数現れるため，モーラ情報を考慮しただけでは不適切な音声素片が選択される場合がある．そして，普通名詞で素片選択においてモーラ情報に加えてアクセント型を考慮した研究 [3] が行われており，アクセント型が合成音声の自然性を向上するために有効であることが示されている．

そこで本研究では，文節を対象とした場合に素片選択にモーラ情報とアクセント型を考慮することで，どの程度の合成音声の品質が得られるかを調査する．

2.2 発話速度

文の発話は名詞のみの発話と比べて韻律が複雑となる．しかし，音節波形接続方式では信号処理を加えない．そのため，音声素片に様々な情報を付加することで韻律の問題を解決しなければならない．通常，文の音声合成を波形接続方式で行う場合には，CHATR 等で使用されている ToBI モデルや藤崎モデルなどの複雑な韻律モデルが使用されている．しかし，文節発声で発話速度が遅い音声を用いる場合には，文節間で区切ることでピッチが初期化される．そのため，文節においても名詞の場合と同じ ように扱うことができ，ToBI モデルや藤崎モデルのような複雑な韻律モデルを使用しなくても合成音声の作成ができると考えられる．

そこで，本研究では文節発声で発話速度が遅い音声を用いて文節の合成音声を作成する．

2.3 連続母音の扱い

母音が連続する場合には音素境界がはっきりしない場合がある．例えば「英語に (e/i/go/ni)」の「エイ」や「感想は (ka/N/so/u/wa)」の「ソウ」等である．このような音素境界がはっきりしない部分については無理に切り離さずに，連続母音として扱う．

そして，本研究では「英語に」と「感想は」は「英語に (e-i/go/ni)」,「感想は (ka/N/so-u/wa)」と扱うこととする．

2.4　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．文節のアクセントについては，NHK 日本語発音アクセント辞典 [13] を参考にラベルデータに対してアクセントを付加する．

・音節
・直前の音素 (前音素環境)
・直後の音素 (後音素環境)
・文節中のモーラ位置
・文節のモーラ数
・文節のアクセント型

次に，情報が一致する音節侯補の中から，データベースの上位の素片を選択し，音節の開始時間と終了時間から波形データを切り出し，接続して合成音声を作成する．

2.5　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とする．さらに，接続部における 2 素片間の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．具体的には，あらかじめラベル付けされた素片開始時間と素片終了時間をもとに，振幅が負から正に変わる部分を，波形が短くなる方向 (開始時間は進む方向，終了時間は戻る方向) に探し，抽出する位置を修正する．

2.6 合成音声の例

本研究で作成した合成音声「聞かれて (/ki/ka/re/te/)」および「銀行の (/gi/N/ko/u/no/)」についての例を下に示す．なお,「_|¯」は音の強弱（アクセント）を表している．
() 内強調部は，実際に選択される部分を示している．

むかって (/mu/ka/q/te/) ＝昔の (/**mu**/ka/sh**i**/no/)
＋使って (/tsu/**ka**/q/te/)
＋当たった (/a/ta/**q**/ta/)
＋終わって (/o/wa/q/**te**/)

銀行の (/gi/N/ko¯−u/no/) ＝銀行に (/**gi**/N/ko¯−u/ni/)
＋深刻に (/shi/**N**/ko/ku/ni/)
＋天候に (/te/N/**ko**¯−**u**/ni/)
＋民謡に (/mi/N/yo¯−u/**ni**/)

3　評価実験

3.1　実験環境

作成した合成音声の評価のために聴覚実験を行う．

本研究では音声データベースとして，複数の電子辞書から重文複文を抽出した日英対訳の例文集 (CREST コーパス [14]) の文を使用する．この例文集は機械翻訳を目的にしたものであるが，日本語の文としては短く，本研究で使用するのに適していると考えた．そこで，この例文集に収録されている 1000 文を使用し，女性話者 (轟美穂 (プロのナレーター)) に文節発声で遅く発声してもらった音声を音声データベースとして用いる．そして，この音声データベースに含まれる 4，5，6 モーラの文節について，以下の条件で各 100 文節を準備する．

・自然音声
・音節波形接続方式で作成した合成音声

なお，実験に使用する 100 文節は，それぞれのモーラごとの作成可能な文節数の割合から表 1 のように定める．

表 1: 実験に使用する文節のモーラごとの内訳

モーラ数	文節数
4mora	17
5mora	70
6mora	13

3.2 評価方法

合成音声の評価のために，音声研究に関わったことのない 9 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，オピニオン評価，対比較実験の 3 つの実験を行う．

(1) 了解度試験

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，比較対象の文節が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

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．

了解度試験とオピニオン評価では，比較対象となる文節を文節発声された自然音声の文の中に埋め込んで行い，比較対象の文節のみについて評価をしてもらう．了解度試験とオピニオン評価で使用する文の例を下に示す．

(例) 全部員が **優勝を** 目指して 練習に 励んでいる

(3) 対比較実験

作成した音声の評価のために対比較実験を行う．対比較実験では文節を文に埋め込んで行うのではなく，文節の音声のみを聞かせる．そして，対比較実験は自然音声と合成音声の同じ内容の文節で 2 種類の音声を続けて聞かせ，どちらの音声が自然に聞こえたかを判定してもらう．

4　実験結果

4.1　了解度試験の実験結果

了解度試験の結果を下に示す.

表 2: 了解度試験の実験結果

	了解度　正解率 (%)
自然音声	99.3(894/900)
合成音声	98.7(889/900)

合成音声の了解度は 98.7 %と高い値が得られ，明瞭な音声が作成できたことが分かる．また，自然音声と比較すると，合成音声も自然音声に近い高い値が得られた.

聞き間違えた音声を見てみると，多くの場合が発音方法の似ている子音の聞き間違えであった.

4.2　オピニオン評価の実験結果

オピニオン評価の全被験者の平均を下に示す.

表 3: オピニオン評価の実験結果

	オピニオンスコア
自然音声	4.75
合成音声	3.55

合成音声のオピニオンスコアは 3.55 となり，高い品質の合成音声が作成できたことが分かる．しかし，合成音声は 3.55 であるのに対して自然音声が 4.75 となっており，自然性の面で自然音声との差があることが分かった.

素片ごとの音量の違いがオピニオンスコア低下の原因となっている場合が多かった．また，スムーズに次の音素に移行していく部分での，微妙な声質の違いによる違和感もオピニオンスコアに影響を与えていた.

4.3 対比較実験の実験結果

自然音声との対比較実験の結果を下に示す.

表 4: 対比較実験の実験結果

	自然音声 (%)	合成音声 (%)
文節数 100	82.7	17.3

対比較実験では合成音声の方が良い音声だと判定された文節が 17.3 %であった. この結果から合成音声は品質高い音声が作成されているが, 自然性の面では自然音声との間にまだ差があるということが分かる.

5　問題点

5.1　接続部の違和感

音節波形接続方式で作成した合成音声は音声素片の接続部の違和感が問題となる．特に違和感を感じるのは母音や撥音が連続する部分である．違和感を感じた例として「膨大な」の自然音声と合成音声のスペクトラムと音声波形を示す．この文節では「アイ」の部分に違和感が感じられた．波形データおよびスペクトログラムの出力には Wavesurfer[15] を使用した．

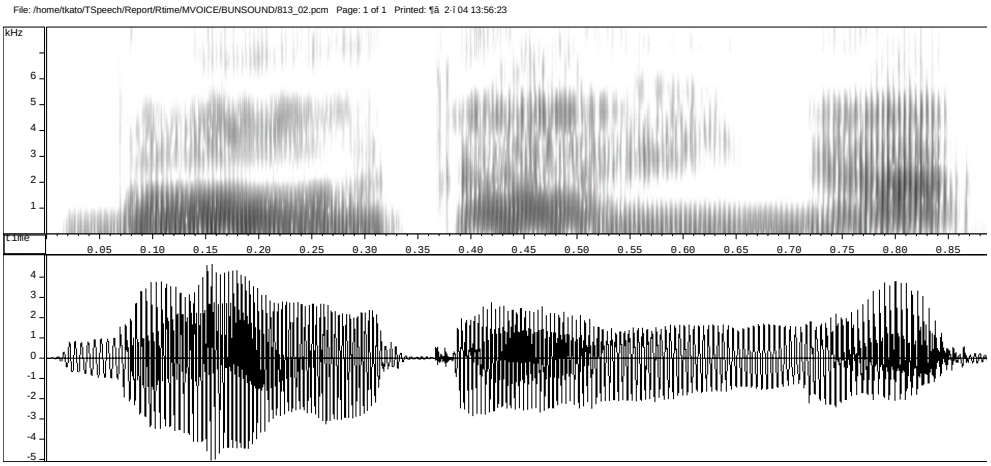

図 1:「膨大な」(自然音声)

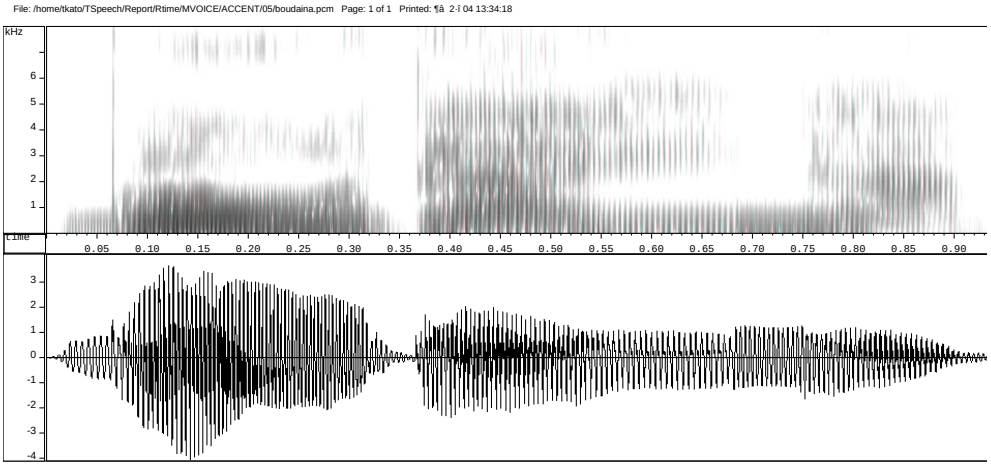

図 2:「膨大な」(合成音声)

これらの音素は前後の音が連続的に変化する部分であり，音素を切り分けることが困難である．しかし，切り分けることが可能であったため今回は別々の音素として扱った．その母音や撥音が連続する部分を別々の素片として扱うことが，自然性の低下に結びついていると考えられる．

5.2　データベースの音量のばらつき

作成した合成音声の中には素片ごとの音量のばらつきも現われた．音節波形接続方式には録音された音声が大量に必要となる．そのため，録音された音声によって音量のばらつきが出てしまう．そして，音節波形接続方式は録音した音声に信号処理を加えないため，音量のばらつきが作成した合成音声に反映する．例として今回作成した音声の中では「部長の」がある．この「部長の」という音声では，最後の「の」に使用した「不況の」の音声が他の素片よりも音量が大きく，最後の「の」が強調されて違和感が現われた．

下に「部長の」の自然音声と合成音声のスペクトラムと音声波形を示す．

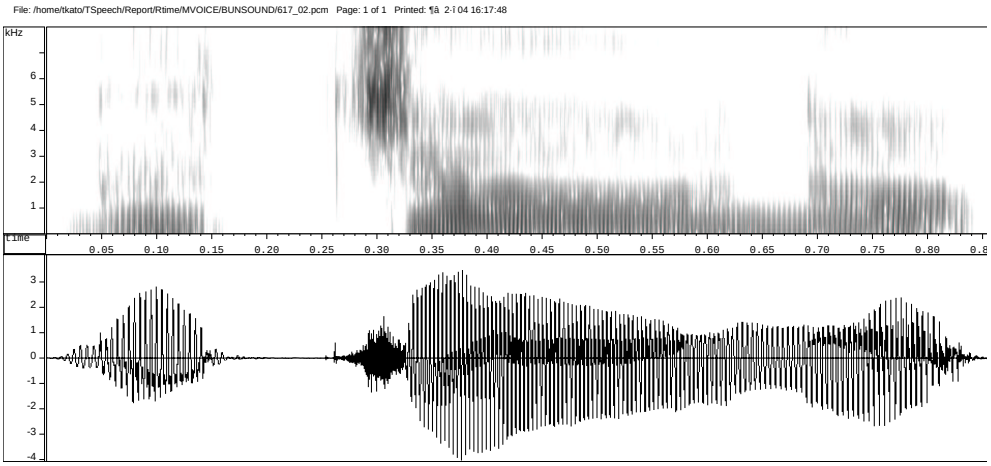

図 3:「部長の」（自然音声）

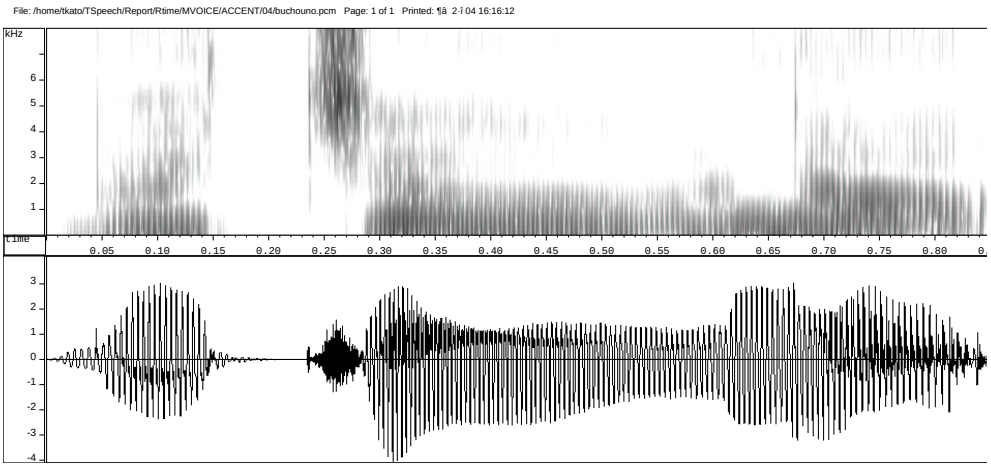

図 4:「部長の」（合成音声）

6　改善案

本研究では問題点解決のための 2 つの改善策を提案する．

6.1　接続部の違和感軽減

前述の実験では音素境界が明確でなく，音素境界で切り離すことが困難とされる「エイ」や「オウ」等の部分のみを連続母音として扱って合成音声を作成した．改善策として，母音と撥音が連続する部分を連続母音として扱う．

しかし，この場合「病院を (byo/u/i/N/o)」という文節の場合には (byo-u-i-N-o) と 1 つの音声素片として扱われてしまうことになる．そこで，母音や撥音が連続する場合には最高で 2 音素までをまとめて１つの音素として扱うことにする．そうすることで,「病院を」は「病院を (byo-u/i-N/o)」と 3 つの音節として扱う．

6.2　録音時間帯による絞りこみ

音声は録音した時間帯が近い場合には音量や発話速度が同程度となると考えられる．そして，録音した時間帯が近い音声素片で合成音声を作成した場合，音声合成に使用する音声素片の音量や発話速度のばらつきが抑えられると考えられる．そこで，本研究で使用する音声コーパスは録音した順番が分かっているため，選択条件に一致した音声素片の侯補の中で録音した時間帯が近い素片を選択し合成音声を作成する．

本研究で使用する音声データベースは，音声が録音した順番に文番号が割り当てられている．そこで，音声素片全体の音量や発話速度を均一化する具体的な絞り込み手法として，音声素片の含まれる文の文番号の値の偏差が小さい素片の組み合せを選択し，音声を作成する．

6.3 合成音声の例

作成した合成音声「むかって (/mu/ka/q/te/)」および「銀行の (/gi/N/ko/u/no/)」について，前述した従来の音節波形接続方式を文節に適用した手法と，新しく提案した手法で作成した場合の例を下に示す．なお,「_|¯」は音の強弱（アクセント）を表している．() 内強調部は，実際に選択される部分を示している．また，() の右にある数字は文番号を表している．

従来手法の合成音声

むかって (/mu/ka/q/te/) ＝昔の (/**mu**/ka/sh**i**/no/).831
＋使って (/tsu/**ka**/q/te/) .223
＋当たった (/a/ta/**q**/ta/) .178
＋終わって (/o/wa/q/**te**/) .416

銀行の (/gi/N/ko−u/no/) ＝銀行に (/**gi**/N/ko−u/ni/) .595
＋深刻に (/shi/**N**/ko/ku/ni/).054
＋天候に (/te/N/**ko**−**u**/ni/).027
＋民謡に (/mi/N/yo−u/**ni**/) .001

提案手法の合成音声

むかって (/mu/ka/q/te/) ＝昔の (/**mu**/ka/sh**i**/no/).831
＋使った (/tsu/**ka**/q/ta/) .890
＋止まった (/to/ma/**q**/ta/) .854
＋歌って (/u/ta/q/**te**/) .791

銀行の (/gi ┴ N/ko = u/no/) ＝銀行に (/**gi** ┴ **N**/ko = u/ni/) .595
＋健康に (/ke ┴ N/**ko** = **u**/ni/) .545
＋劇場の (/ge/ki/jo − u/**no**/) .558

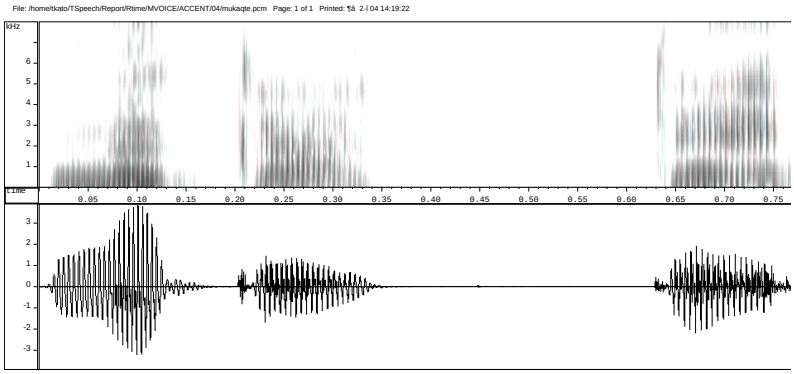

図 5: 「むかって」（従来手法）

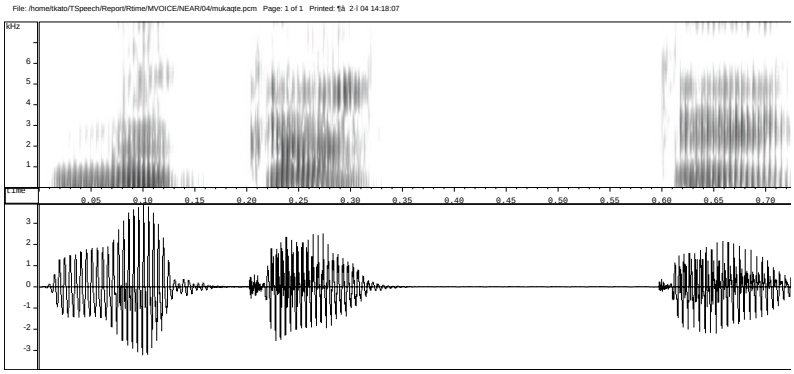

図 6: 「むかって」（提案手法）

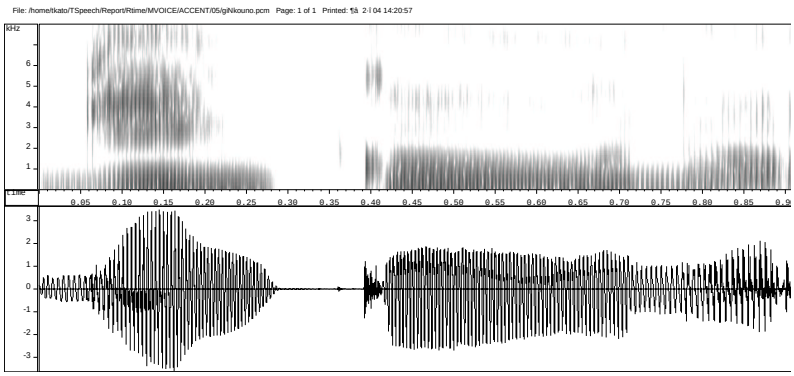

図 7: 「銀行の」（従来手法）

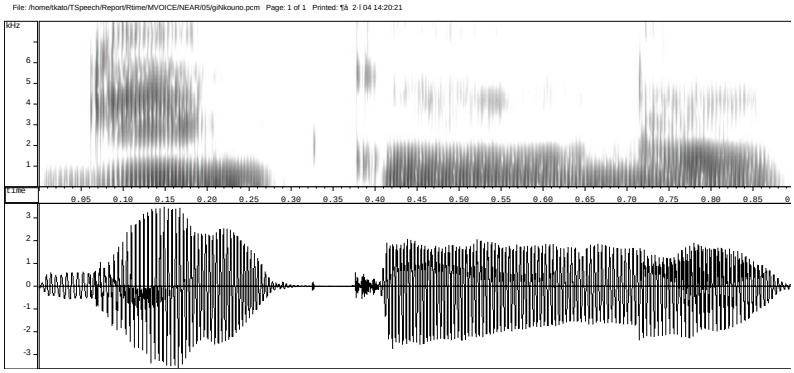

図 8: 「銀行の」（提案手法）

7　評価実験

作成した音声の評価のため，聴覚実験を行う．聴覚実験は前述した従来手法の合成音声の評価の時に行った実験と同じ条件で行う．また，対比較実験は自然音声と提案手法の合成音声，従来手法の合成音声と提案手法の合成音声の組み合せで行う．

7.1　了解度試験の実験結果

了解度試験の実験結果を下に示す．

表 5: 了解度試験の実験結果

	了解度　正解率 (%)
自然音声	99.3(894/900)
提案手法	99.3(894/900)
従来手法	98.7(889/900)

提案手法の了解度は 99.3 %であり従来手法の 98.7 %に比べると提案手法の方が高い値が得られた．自然音声と比較してみても，提案手法は自然音声と同程度の高い値が得られた．

7.2　オピニオン評価の実験結果

オピニオン評価の全被験者の平均を下に示す．

表 6: オピニオン評価の実験結果

	オピニオンスコア
自然音声	4.75
提案手法	3.83
従来手法	3.55

提案手法のオピニオンスコアは 3.83 となり，自然音声には及ばないものの，高い品質の音声が作成されていることが確認できた．また，従来手法の 3.55 より高い値が得られ，素片選択の条件を追加することで品質が向上していることが分かる．

7.3 対比較実験の結果

7.3.1 自然音声との対比較

従来手法と提案手法のそれぞれの自然音声との対比較実験の結果を下に示す.

表 7: 自然音声との対比較の結果

	自然音声 (%)	提案手法 (%)
文節数 100	74.3	25.7

	自然音声 (%)	従来手法 (%)
文節数 100	82.7	17.3

この結果から提案手法は自然性の面では自然音声との間にまだ差があるが，品質の高い音声が作成されていることが分かる．また，従来手法は 17.3 %の文節が自然音声よりも良い音声だと判定されているのに対し，提案手法は 25.7 %となっており，音節波形の選択条件を追加した効果が現われている.

7.3.2 従来手法と提案手法との対比較

従来手法の合成音声と提案手法の合成音声の対比較実験の結果を示す.

表 8: 従来手法の合成音声との対比較

	従来手法 (%)	提案手法 (%)
文節数 95	39.3	60.7

従来手法との対比較では，提案手法の方が良いと判定された文節が 60.7 %となった．このことから，音節の選択条件を追加することで自然性が向上したことが分かる.

8　考察

8.1　了解度試験の解析

了解度においては自然音声も提案手法の合成音声に関しても同程度の高い値となった．了解度試験において，合成音声で多くの被験者が間違えた音声を示す．

表 9: 聞き間違えた音素

原因		提案手法	従来手法	合計
最終モーラの欠落		1	3	4
母音部の誤り	[i] → [-]	1		1
	[a] → [u]		1	1
撥音部の誤り	[N] → [bu]		1	1
促音部の誤り	[q] → [-]	1		1
子音部の誤り	[g] → [r]		1	1
	[k] → [h]		2	2
	[k] → [t]	1	1	2
	[ky] → [ry]		1	1
	[n] → [r]	1		1
	[sh] → [j]	1		1
	[t] → [k]		1	1

「**か**たった」を「**た**たった」や「あった**が**」が「あった**ら**」等の発音方法が似ている子音を間違える場合がほとんどであった．そして，音量や音声の継続時間の違いから現われる最終モーラの聞き逃しが，従来手法では 3 つであるが，録音時間帯を考慮した提案手法では 1 つと減少した．

8.2 オピニオン評価の解析

了解度では自然音声も合成音声も同程度の値が得られたが，オピニオンスコアでは自然音声は 4.75 の対して従来の手法では 3.55，そして今回提案した手法でも 3.83 となり，自然音声には及ばなかった．

オピニオンスコアの評価が悪かった合成音声の多くが接続部での違和感が原因となっていた．それは，音素ごとの微妙な音量の違いから，接続部が強調されてしまい不自然に聞こえたものである．

また，オピニオン評価の音声の種類による比較を下に示す．

表 10: オピニオンスコアの平均の種類による比較

	自然音声	同じ値	従来手法
文節数 100	96	2	2

	自然音声	同じ値	提案手法
文節数 100	87	6	7

	従来手法	同じ値	提案手法
文節数 95	20	5	70

この結果を対比較実験と比較すると自然音声の自然性の高さが際立っている．文節の音声は文の音声の中に埋め込んでこそ意味を持つものであると考え，本研究ではオピニオン評価は比較対象となる文節を自然音声の文の中に埋め込んで行った．そのため，自然音声の文と比較対象の文節との音量差から，合成音声が不自然に聞こえてしまうことがあった．そして，合成音声は自然性が低く評価されてしまい，自然音声のオピニオンスコアの値が際立った結果になったと考えられる．

8.3 発話速度に関する考察

通常の速度で文節発声している ATR の単語発話データベースの中の DSB を用いて実験を行った．使用した音節選択条件は下に示す．

・音節
・直前の音素 (前音素環境)
・直後の音素 (後音素環境)
・文節中のモーラ位置
・文節のモーラ数
・文節のアクセント位置

使用したデータベースは文を文節ごとに区切って発話されているが，区切る時間が短く，普通の発話に近い音声となっている．また，文が 115 文と少なかったため，作成できた 12 文節で評価を行った．評価は 8 人の被験者について，了解度試験とオピニオン評価を行い，本研究と同様に評価対象の文節は自然音声の文の中に埋め込んで実験を行った．その結果を下に示す．

表 11: 通常の発話速度を用いた実験

	了解度　正解率 (%)			オピニオンスコア		
	FTK	FYN	平均	FTK	FYN	平均
自然音声	96	99	98	4.5	4.7	4.6
合成音声	96	98	97	3.2	3.0	3.1

本研究で作成した合成音声の了解度は 99.3，オピニオンスコアは 3.83 であり，この実験の音声は本研究で得られた音声に品質では及ばなかった．これは通常の発話速度で発話した音声では，文節間での区切りの時間が短いため，ピッチが初期化しきれず，アクセントのばらつきがあったためだと考えられる．したがって，今回のような波形選択を通常の発話速度の音声に適用することは困難だと考えている．

8.4　アクセントに関する考察

文節は普通名詞よりもアクセントの多様性があり，複雑だと考えられるが，本研究ではアクセント情報としてアクセント型のみを使用した．アクセント型に加えてアクセント核の情報も波形選択に使用してみたところ，今回作成したほとんどの文節において同じ波形が選択された．

この結果から，波形接続方式を文節に適用した場合にも，普通名詞の場合と同様にアクセント型のみを考慮すれば十分であると思われる．

8.5　不自然な音声の解析

本研究では母音や撥音が連続する部分は音が連続的に変化している場合が多く，切り離すことで自然性が低下すると考え，最大で 2 音素をまとめて連続母音として扱い，音声を作成した．その結果，波形の接続部での違和感が軽減でき，品質が向上した．

しかし，今回は最大で 2 音素までという制限を加えたため，例えば『遺体を (i/ta/i/o)』という文節の場合には『遺体を (i/ta-i/o)』となってしまい，最後の『い』から『を』へのつながりが不自然になっているということがあった．

また，文節の場合には，助詞との接続の部分も音が切れずスムーズにつながることが多く，その部分での音声の質の違いが自然性の低下に結びついている場合もあった．

本研究では，方式の評価のために最大で 2 音素までという制限を加えて合成音声を作成したが，同じ音声から使用する音声の制限をなくすことで，さらに自然性の低下を抑えられる可能性があると考えられる．

8.6　録音した時間帯が近い素片を選択した効果

本研究は音量や発話速度を揃えることを目的に音声の録音した時間帯が近い素片の組み合せを選択し，音声を作成した．そして，融合ラベルを用いずに録音した時間帯による制御のみを使用した音声の対比較の結果を示す．

表 12: 時間帯による制御の効果

	品質向上	品質低下
文節数 50	59.2(266/450)	40.8(184/450)

表より，録音した時間帯による制御が合成音声の品質向上に大きく影響していることがわかる．ただ，音量や音声の継続時間を直接制御しているわけではないので，不適切な音声素片が選択され，品質が低下する場合もあった．

8.7　データベースの音量のばらつき

従来手法では音量のばらつきが問題となり，自然性が損なわれることがあった．今回の実験では音量のばらつきを抑えるために，録音した時間帯が近い音声を選んで音声を作成するようにした．その結果，音量のばらつきは少なくなり品質の高い音声の作成ができた．

しかし，完全に音量の音量の統一はできず，不自然さが残る音声も作成された．

音量についてさらに品質を上げるには，接続部分の音量の同程度の素片を使用するなどという手法が考えられる．この手法により音量が原因となる接続部で感じる違和感が緩和できると考えられる．しかし，自然性に関わる要素は音量だけではないため，発話速度や継続時間等の制御も必要となり，素片選択の条件が複雑になるという問題がある．

8.8 素片単位に関する考察

今回は母音や撥音が連続した場合に連続母音として扱って音声を作成し，作成した合成音声の品質は向上した．しかし，連続母音の数が多くなるため，作成可能な文節数が減少する．本研究では日英対訳の例文集に含まれる 1000 文を使用し，従来手法に母音や撥音が連続した場合は連続母音として扱うという条件を加えた．そして，母音や撥音が連続した場合に連続母音として扱わずに作成できる 4，5，6 モーラの文節数が 382 文節だったのに対し，連続母音として作成した場合は 323 文節へと減少した．品質向上のために制御を増やすと，作成可能な文節数はさらに減少してしまう．

この問題については，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考えている．なお，環境などは異なるが，短い刺激音声について，代替により自然性劣化に関する評価がなされている [16].

9　まとめ

本研究では文節発声で発話速度が遅い音声を用いた場合の音節波形接続方式の文節における有効性を調査した．従来の単語合成に用いられている手法で作成した合成音声は，了解度が 98.7 %でオピニオンスコアも 3.55 が得られ，音節波形接続方式が文節に対しても有効であることが分かった．

また，従来手法の問題点の解決のために，音節選択に２つの条件を追加して行った実験では，了解度は 99.3 %でオピニオンスコアが 3.83 となり従来手法の値を上まわった．対比較実験でも 60.7 %の文節が従来手法よりも品質の高い音声と判定され，追加した２つの条件が素片選択に有効であることが分かった．

一方，自然音声は了解度が 99.3 %，オピニオンスコアは 4.75 となった．また，提案手法との対比較では 74.3 %の文節が自然音声の方が品質が高いと判定された．了解度では合成音声も自然音声と同程度の値であった．また，自然性の面では自然音声と合成音声の間にまだ差があるが，合成音声も高い品質が得られたことが確認できた．

今後は，接続部の違和感を軽減など様々な制御を導入し，音声の品質を高めることが必要である．また，それと同時に，音声コーパスをより有効に利用できるように，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] 村上仁一, 水澤紀子, 東田正信: 音節波形接続方式による単語音声合成, 電子情報通信学会論文誌 D-II, Vol. J85-D-II, No. 7, pp. 1157–1165 (2002).

[2] 石田隆浩, 村上仁一, 池原悟: 音節波形接続型音声合成の普通名詞への応用, 電子情報通信学会技術研究報告, SP2002-25, pp. 7–12 (2002).

[3] 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への応用, 日本音響学会 2003 年春期研究発表会, 2-Q-18, pp. 1-409,410 (2003).

[4] 益子貴史, 徳田恵一, 小林隆夫, 今井聖: 動的特徴を用いた HMM に基づく音声合成, 電子情報通信学会論文誌 D-II, Vol. J79-D-II, No. 12, pp. 2184–2190 (1996).

[5] 徳田恵一: HMM による音声合成の基礎, 電子情報通信学会技術研究報告, SP2000-74, pp. 43–50 (2000).

[6] Jan P.H. van Santen, Richard W. Sproat, Joseph P. Olive and Julia Hirschberg: Progress in Speech Synthesis, Springer, ISBN 0-387-94701-9 (1996).

[7] 戸田智基, 河井恒, 津崎実, 鹿野清宏: 素片接続型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く素片選択, 電子情報通信学会論文誌 D-II, Vol. J85-D-II, No. 12, pp.1760–1770 (2002).

[8] Nich Campbell and Alan W.Black: CHATR:自然音声波形接続型任意音声合成システム, 電子情報通信学会技術研究報告, SP96-7, pp. 45–52 (1996).

[9] 石川泰: 音声合成のための韻律制御の基礎, 電子情報通信学会技術研究報告, SP2000-72, pp. 27–34 (2000).

[10] 村上仁一, 前田智広, 池原悟: モーラ情報を用いた音素ラベリング方式の検討, 電子情報通信学会技術研究報告，SP2003-137，pp. 145–150 (2003).

[11] 妹尾貴宏, 村上仁一, 池原悟: モーラ情報を用いた単語音声認識の検討, 電子情報通信学会技術研究報告, SP2002-130, pp. 55–61 (2002).

[12] 谷口勝則, 村上仁一, 池原悟: モーラ情報を用いたフィルタバンクによる孤立単語認識, 電子情報通信学会技術研究報告, SP2002-131, pp. 63–68 (2002).

[13] NHK 出版: NHK 日本語発音アクセント辞典 新版, ISBN 4-14-011112-7 (1998).

[14] 村上仁一，池原悟，徳久雅人，”日本語英語の文対応の対訳データベースの作成”,「言語，認識，表現」第 7 回年次大会，(2002-12)

[15] Kåre Sjölander and Jonas Beskow: Wavesurfer,
http://www.speech.kth.se/wavesurfer/

[16] 河井恒, 津崎実, 舛田剛, 岩澤秀紀: 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, 電子情報通信学会技術研究報告, SP2001-22, pp. 51–57 (2001).