論　文

アクセントを考慮した音節接続型音声合成における普通名詞の評価

村上　仁一[†a)]　　石田　隆浩[†]　　池原　悟[†]

Evaluation of Common Noun using Concatenative Syllable Speech Synthesis with Accent

Jin'ichi MURAKAMI[†a)], Takahiro ISHIDA[†], and Satoru IKEHARA[†]

あらまし
現在，高い自然性が要求される音声ガイダンスにおいて，録音編集型音声合成が使用されている．しかし，この方式では，同一話者による単語音声が大量に必要となる．そのため，長期間にわたって収録する必要がある．その結果，収録された音声には，録音環境の違いや話者の疲労により，発話速度やピッチにばらつきが出る．このため，音声ガイダンスの品質が低下することが多い．そこで，一部の音声を音声合成によって作成することで，録音量を大幅に減らす方式が考えられる．このような音声の合成を目的として提案されている音節波形接続方式では，固有名詞に対する有効性が示されている．
本研究では，普通名詞を音節波形接続方式を用いて音声合成して，得られた合成音声を評価した．実験の結果，従来の音節素片の前後音素環境と単語のモーラ数と音節素片の単語中のモーラ位置に加えて単語のアクセントを考慮することで，自然性の高い音声が得られることを示した．
キーワード　音声合成 波形接続 モーラ長 モーラ位置 アクセント

1. はじめに

1.1 研究の背景

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．

1.2 録音編集型音声合成の問題点

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，通常，同一話者の音声が必要となる．しかし，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である．仮に同一話者で収録したとしても，収録時間が長期間になるため，録音環境の違いや話者の疲労により，発話速度やピッチにばらつきが出る．

そこで，可変部に必要になる音声を音声合成によって作成する方法が考えられる．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，多くの手法が提案されている．基本的には，音声

† 鳥取大学工学部，鳥取県
Faculty of Engineering, Tottori University, 4-101, Minami Koyamachou, Tottori city, Tottori, 680-8552 Japan
a) E-mail: murakami@ike.tottori-u.ac.jp

の特徴をパラメータとして抽出し，信号処理によって合成音声を作成する．現在商用化されている音声合成は PSOLA [4] 方式が主流のようだ．最近は，HMM を用いて直接音声を合成する研究も盛んである [3] [6]．しかし，いずれの方式も，人の声のような自然性の高い音声を安定して得ることが困難である [2]．そのため，可変部に規則音声合成を用いても，ガイダンス音声に，違和感が残ることが多い．

1.3 波形接続型音声合成

ところで，録音した音声波形の一部を用いて，信号処理を加えずに音声を合成する方法がある．これを一般に，波形接続型音声合成と呼んでいる [1] [8]．

波形接続型音声合成は，収録された大量の音声から，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．接続単位については，音素，CV，VCV，CVC など，様々な単位が提案されている [7]．いずれの場合においても，基本的に信号処理を行わず，収録された録音音声から取り出した波形をそのまま用いる．このため，話者の声の特徴（以下,「話者性」）や高い自然性を保つことが可能である．この方式の１つに CHATR [8] がある．CHATR は，合成する音声の音響パラメータをモデルを利用して予測し，もっとも近い音素素片を接続することで，合成音声を得ている．そして音素素片の選択のとき，ピッチやケプストラムなどの音響パラメータを利用している．

1.4 音節接続型音声合成

水澤らが提案している「音節接続型音声合成」[9] では，音節素片の前後音素環境・音節素片が属する単語のモーラ数・音節素片の単語中のモーラ位置の，言語パラメータのみを用いて，接続する音節素片を選択している．そして音響パラメータを一切用いない．この合成方法は，地名・人名などの固有名詞の合成音声において自然性の高い合成音声を得ている．これは，地名などの固有名詞ではアクセント型がほぼ一意に決まるため，単語のモーラ数と音節素片のモーラ位置の情報がアクセントの情報になるためと考えている．

しかし，この「音節接続型音声合成」を普通名詞に適応した場合，例えば「雨」と「飴」のように同音異義語が多数あるため，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．実際，過去の研究において，アクセントを未考慮であったために不自然な音声が作成される場合があった [11]．そこで，音節接続型音声合成を普通名詞に適用した場合，単語のアクセントを考慮する必要がある．

なお，一般に音声合成において，韻律制御は重要な課題の１つである [10]．波形接続型音声合成においては，韻律制御に藤崎モデルや ToBi モデルが利用されている．また，最近は HMM を用いる方法も提案されている [6]．

1.5 本研究の目的

本研究では,「音節接続型音声合成」を普通名詞に適用した場合の，合成音声の品質について報告する．具体的には，音節を接続単位として，音節素片の前後音素環境と，音節素片が属する単語のモーラ数と，音節素片の単語中のモーラ位置と，音節素片が属する単語のアクセントが一致する音節素片を結合して，普通名詞を合成する．そして，合成した音声を聴覚実験により評価する．

なお，本論文では，以後，音節素片が属する単語のモーラ数および音節素片の単語中のモーラ位置を，モーラ情報と呼ぶ．

2. 提案する音節波形接続方式

2.1 概　　要

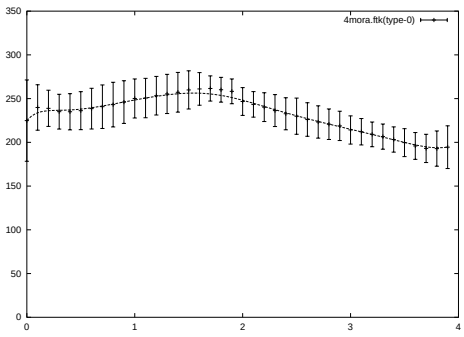
本研究では，普通名詞の音声合成に音節波形接続方式 [9] を用いる．音節波形接続方式では，まず，音節を基本単位として素片選択を行う．素片選択の条件としては，以下の条件が一致するものを選択する．

表 1　使用するパラメータ
Table 1　Parameters for Speech Synthesis

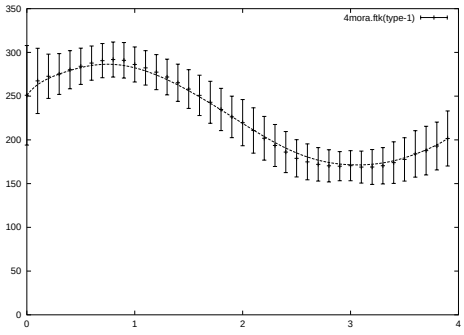
Sy:	音節素片の音節
P:	音節素片の前音素環境
N:	音節素片の後音素環境
m:	音節素片の単語中のモーラ位置
M:	音節素片が属する単語のモーラ数
ac:	音節素片が属する単語のアクセント

この方式は，信号処理を行わずに音節素片を接続するため，高い自然性と話者性を得ることが可能である．また，音響パラメータを一切考慮せずに，言語パラメータのみで波形の選択を行うことを特徴にしている．このため，録音に必要な音声を事前に決めておくことができる．また，計算量もほとんど必要としないため，リアルタイムで合成が可能である．

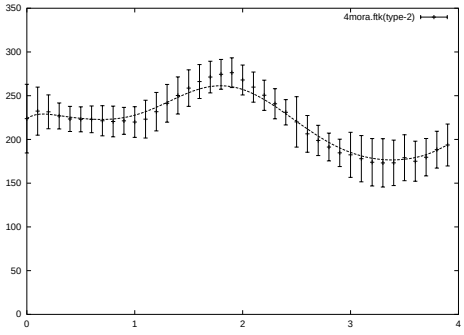
なお，従来の音節波形接続方式 [9] は，上記のパラメータのなかの音節素片が属する単語のアクセントを利用していない．音節素片の前後環境と，音節素片の単語中のモーラ位置と，音節素片が属する単語のモーラ数のみ利用している．



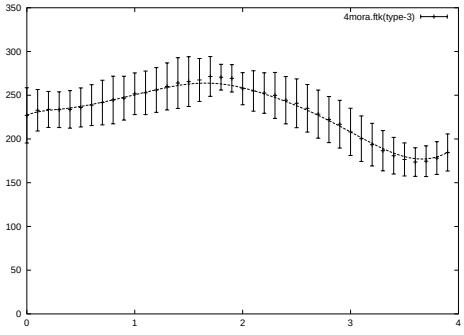
4 モーラ 0 型 (約 1,300 件の平均)



4 モーラ 1 型 (約 200 件の平均)



4 モーラ 2 型 (約 100 件の平均)



4 モーラ 3 型 (約 100 件の平均)

図 1　モーラ情報およびアクセントと，ピッチ周波数の関係

Fig. 1　The Relationship between Pitch and Mora Information and Accent

2.2　普通名詞における，モーラ数・モーラ位置・アクセントとピッチ周波数の関係

普通名詞における，ピッチ周波数とモーラ数・モーラ位置・アクセントの関係を調査するために，ATR の単語発話データベース Aset を利用する．ATR の単語発話データベース Aset のうち，話者 FTK の 4 モーラ語普通名詞約 1,700 件をアクセント型に分類して，ピッチ周波数の平均と分散を調べる．ピッチの抽出には ESPS/Waves+ [12] を利用する．結果を図 1 に示す．図は，それぞれ，0 型アクセントの 1,500 件，1 型アクセントの 200 件，2 型アクセントの 100 件 3 型アクセントの 100 件，のピッチ周波数の平均値と分散を示す．これらの図において，時間軸はモーラ数で正規化したのち計算した．図中の■はピッチ周波数の平均値を，縦棒の長さはピッチ周波数の分散を示している．

図 1 より，単語のモーラ長とモーラ位置とアクセントは，ピッチ周波数と強い相関があると考えられる．そこで韻律情報として単語のモーラ長とモーラ位置とアクセントが利用できると考えられる．

2.3　音節波形接続方式による音声合成の例

音節波形接続方式では，まず，音節を単位として素片選択を行う．素片選択の条件としては，表 1 中のパラメータが一致するものを候補とする．なお，複数の候補が残った場合はランダムに選択する．

音節素片を「$Sy(P,N)_{m,M(ac)}$」と表現する．記号の意味は表 1 と同じである.「反対」(/ha/N/ta/i/) を合成する音節素片を，以下のように表現する．

- 普通名詞「反対」(/ha/N/ta/i/) のモーラ数は 4 であるため，$M=4$ となる．またアクセントとして，音の高低を 0（低）と 1 （高）で表すと，$ac=0111$ となる．
- 1 モーラ目の音節は ha であり，前音素はなく，後音素は N である．よって，$\mathtt{ha}(,\mathtt{N})_{1,4(0111)}$ と表現できる．
- 同様に，2 モーラ目の音節は N で，前音素が a，後音素が t であるので，$\mathtt{N}(\mathtt{a},\mathtt{t})_{2,4(0111)}$ と表現できる．
- 3 モーラ目，4 モーラ目もそれぞれ，$\mathtt{ta}(\mathtt{N},\mathtt{i})_{3,4(0111)}$，$\mathtt{i}(\mathtt{a},)_{4,4(0111)}$ と表現できる．
- すべての音節素片において，条件に合う音節素片を録音データベースから抽出する．
- ラベルデータから音節素片の開始時間と終了時間を取得して波形データを取り出し，接続することで合成音声を作成する．

たとえば，「反対」(/ha/N/ta/i/) の/ha/は

「反映」(/**ha**/N/e/i/) の/ha/, /N/は「簡単」(/ka/**N**/ta/N/) の/N/, /ta/は「団体」(/da/N/**ta**/i/) の/ta/, /i/は「洋裁」(/yo/u/sa/**i**/) の/i/から作成する.

概略を図 2 に示す.

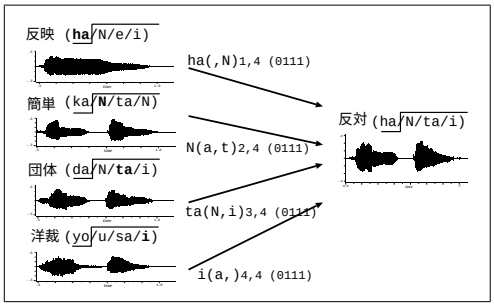


図 2 提案手法の概略
Fig. 2 Outline for Proposed Method

2.4 自然性向上のための補助的な操作について

音節素片を接続する際には，波形の位相を考慮する必要がある．位相を考慮せずに波形を取り出し，接続すると，接続部分にノイズが乗り，合成音声の品質，特に自然性を低下させる原因となる．この問題を解決するために，音素間のスペクトル遷移の始終端の時間を用いることで品質の向上をねらった研究がある [13]. 文献 [13] では，音素間で前後音素が不一致の場合を対象として効果が得られたことが報告されている．

本研究では，振幅が負から正に変化する点 (以下，零点) を探し，音節素片の開始時間と終了時間から見て最も近い零点で波形データを切り出す．本研究では前後環境が一致した音節素片を選択しているため，このような簡易な方法を採っても合成音声の品質の低下は小さいと考えている．

3. 評価実験

3.1 音声データベース

音声データベースとして，ATR の単語発話データベース Aset(5,240 件) を用いる．音声合成では，Aset に含まれる 3 モーラおよび 4 モーラ語名詞約 3,000 件を利用する．話者には，女性話者 FTK,FYN の 2 話者を用いる．なお，ATR の単語発話データベースには，アクセントが付与されていない．そこで，NHK 日本語発音アクセント辞典 [14] を利用してラベルデータにアクセントを付加する．そのため，ATR の Aset 中に，収録の誤りのため NHK のアクセント辞書と一致しない音声データがある．

3.2 評価対象

3 モーラの普通名詞 50 個および 4 モーラの普通名詞 50 個，計 100 単語において、2.3 節に示した方法によって音節素片の選択を行い．合成音声を作成する．また，比較実験のために，自然音声、および，PSOLA 方式を用いた市販の合成機の合成音声を利用する．作成した音声の一部を表 2 に載せる．表中の太文字は，使用した音節素片を意味する．また，上付線，下付線はアクセントを意味する．

表 2 合成した普通名詞 (一部)
Table 2 Synthesis Speech (partly)

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/)
安心 (/a/N/shi/N/)
= 安静 (/a/N/se/i/)
+ 関心 (/ka/N/shi/N/)
+ 妊娠 (/ni/N/shi/N/)
+ 農民 (/no/u/mi/N/)
身分 (/mi/bu/N/)
= 身振り (/mi/bu/ri/)
+ 気分 (/ki/bu/N/)
+ 処分 (/sho/bu/N/)
威厳 (/i/ge/N/)
= 意外 (/i/ga/i/)
+ 機嫌 (/ki/ge/N/)
+ 無限 (/mu/ge/N/)

3.3 評価方法

合成音声の評価のために，了解度試験とオピニオン評価と対比較試験を行う．被験者は音声研究に関わった経験のない人 5 名である．評価は，作成した普通名詞（可変部）を自然音声の文（固定部）に埋め込み，自然音声が含まれる音声ガイダンスと合成音声が含まれる音声ガイダンスをランダムに再生して，ヘッドフォンを使って行う．

- 了解度試験

初めに，合成した普通名詞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，音声ガイダンスの固定部の発話内容と可変部の場所を予め示しておき，可変部がどのように聞こえたか仮名で書き取らせる．自分の知識などは用いず，聞こえたとおりに書き取るように指示する．

- オピニオン試験

次に，合成した普通名詞の自然性を調べるために，

オピニオン評価を行う．音声ガイダンスの全ての発話内容と可変部の場所を予め示しておき，可変部にのみ評価するように指示する．オピニオン評価では，自然に聞こえた度合を 5 段階 (1 が最も不自然，5 が最も自然) で評価する．

- 対比較試験

最後に，対比較試験を行う．音声ガイダンスの発話内容を一切提示せずに，自然音声の音声ガイダンスと合成音声を含む音声ガイダンスを聞いて，自然に聞こえる音声ガイダンスを選択する．

4. 実験結果

4.1 了解度

提案手法による合成音声の了解度試験の結果を表 3 に示す．比較のために，自然音声の普通名詞と市販の合成器による普通名詞の結果も併せて示す．この結果から，了解度では，提案手法により作成された音声は自然音声と差がほとんどないことがわかる．

表 3　実験結果　了解度
Table 3　Results of Experiments (Intelligibility)

	了解度　正解率 (%) 評価音節数: 1,750		
	FTK	FYN	平均
自然音声	99.7	99.9	99.8
提案手法	99.4	99.7	99.6
市販の音声	96.7	97.6	97.2

了解度試験において，被験者が間違えた音声の一部を表 4 に示す．なお，表中の下線部は，被験者が間違えた箇所を示す．

表 4　間違いの例
Table 4　Examples of Erros

	正解	間違いの例
1	はいせき (排斥)	か いせき
		た いせき
2	さいだい (最大)	さい が い
3	せつだん (切断)	せつ が ん
4	かいだん (階段)	かい が ん
5	ぐんかん (軍艦)	ぶ んかん
6	れんさい (連載)	め んさい
		え んさい
7	ちょっかく (直角)	しょっかく

表 4 からも分かるように,「は」と「か」,「だ」と「が」など，似た音韻を間違える場合が多かった．また，特に「a」の音を持つ場合に多く間違えていた．

4.2 オピニオンテスト

提案手法による合成音声のオピニオンテストの結果を表 5 に示す．比較のために，自然音声の普通名詞と市販の合成器による普通名詞の結果も併せて示す．この結果から，提案手法により作成された音声は，オピニオンスコアでは自然音声にやや劣るものの，非常に自然性の高い合成音声が得られることが示された．

表 5　実験結果 オピニオンテスト
Table 5　Results of Experiments (Opinion Sore)

	オピニオンスコア 評価単語数: 100		
	FTK	FYN	平均
自然音声	4.8	4.9	4.9
提案手法	4.2	4.4	4.3
市販の音声	2.4	2.4	2.4

オピニオン評価においては，波形の接続部分が不自然に聞こえる音声の評価が特に悪かった．例えば「最大」という音声で「祭日」の第 1 音節と「階段」の第 2 音節の接続部分が不自然に聞こえた．

4.3 対比較試験

自然音声と提案手法による合成音声の対比較試験の結果を表 6 に示す．

表 6　実験結果 対比較試験
Table 6　Results of Experiments (ABX Test)

	対比較試験 評価単語数: 100		
	FTK	FYN	平均
自然音声	87.0 %	74.2 %	79.6 %
提案手法	13.0 %	25.8 %	19.4 %

この表から，提案手法によって得られた合成音声は，自然音声には及ばないものの，合成した普通名詞の 19%が，自然音声より自然性が高いと評価された．

これらの実験結果から，アクセントを考慮した音節接続型音声合成で得られた普通名詞は，非常に明瞭度が高く自然性の良い音声であると言える．

5. 考　　察

5.1 音声データベースの音量のバラツキ

音声データベースに使用した Aset には，5,240 単語の音声の音量や音質にばらつきがある．本研究では，音節間での音量差については考慮していなかったため，音量の大きい波形と音量の小さい波形をつないだ結果，自然性が落ちることがあった．例えば今回の実験において「解散」という普通名詞で，第 2 音節に利用した

「財政」と第 3 音節に利用した「計算」の音量差による自然性の低下が見られた．

この問題に対しては，録音環境 (日時，場所など) が分かる場合，それらが近い音声を優先的に用いることで改善することができる．しかし，基本的には，音量差を考慮した選択方法を導入する必要があると思われる．

5.2 個 人 差

Aset には 10 名の単語音声のデータがある．この話者の中に残響がある音声がある．また，非常に声が高い（ピッチの高い）話者がいる．そのため，今回の実験では，比較的落ち着いた（ピッチの低い）話者である FTK と FYN を利用した．作成した FTK と FYN の合成音声を比較してみると，FYN のほうが了解度試験・オピニオンテスト・対比較試験とも，良い結果を得ている．特に対比較試験において明確に差がでている．この原因として，FYN は FTK と比較すると，原音において比較的抑揚がすくない．そのため音節が滑らかに接続され，非常に自然性の高い音声が作成できたと考えている．

5.3 アクセントの有効性

図 3 に，ATR の単語発話データベース Aset の，話者 FTK の 4 モーラ語普通名詞詞約 1,700 件のピッチ周波数の平均と分散を示す．図から，ピッチ周波数の分散は小さく，モーラ位置が決まれば，ほぼピッチ周波数が決定できることが分かる．また，図 1 と比較すると，4 モーラ 0 型に良く似ていることがわかる．

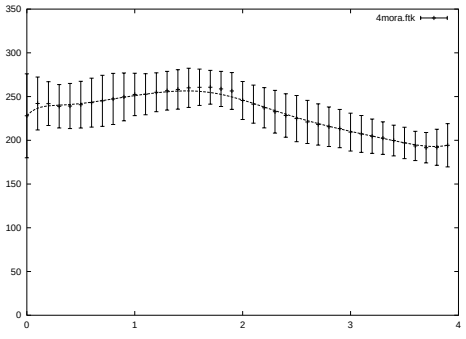


図 3 モーラ情報とピッチ周波数の関係
Fig. 3 Relationship between Mora and pitch

そこで，表 1 から音節素片が属する単語のアクセントを用いずに，音節素片の前後環境・音節素片の単語中のモーラ位置・音節素片が属する単語のモーラ数のみ利用して，普通名詞を合成した．その実験結果を表 7 に示す．

表 7 アクセントを考慮しない合成音声
Table 7 Speech Synthesis without Accent

	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
アクセント利用	99.4	99.7	99.6	4.2	4.4	4.3
アクセント未使用	99.4	99.7	99.6	3.8	4.1	4.0

この表から，アクセントを利用せずに音声合成をおこなった場合と，アクセントを使用して音声合成をおこなった場合を比較すると，了解度には差がないが，オピニオン評価において低下していることがわかる．

この理由として以下の理由を考えている．通常，普通名詞は，0 型の名詞が多い．そのため，多くの単語が 0 型として作成される．このときは自然性の高い音声が作成される．しかし 2 型や 3 型の合成において，0 型の名詞を使用して合成した場合，自然性が低下する．[9].

5.4 自動ラベリング

波形接続型音声合成においては，録音音声に対してラベル（音素境界位置）が付与されたデータが必要となる．今回使用した ATR の Aset では、人手によるラベルデータがある．しかし，人手によるラベリングには多くのコストがかかるため，自動ラベルが好ましい．

そこで，ラベルデータを自動的に作成し，自動ラベルと手動ラベルによる音声合成の明瞭度および自然性の差を調査する．自動音素ラベリングには様々な方法があるが，本研究では，HMM による基本的な方法を用いる．ツールとしては HTK [17] を利用する．使用したパラメータを表 8 に示す．

自動音素ラベリングでは，まず Aset の 5,240 件を奇数番と偶数番に分ける．次に奇数番データで HMM を学習する．最後に Viterbi アルゴリズムを利用して偶数番データに自動ラベルを行う．また，偶数番データで HMM を学習して奇数番データに自動ラベルを行う．

表 9 に結果を示す．ただし，実験は 4 モーラ単語 50 個についてのみ行った．

この結果をみると，オピニオンスコアにおいて，手動ラベルは自動ラベルと比較してやや良いが，あま

表 8　自動ラベリングにおいて使用するパラメータ
Table 8　Parameters for Automatic Labeling

標本周波数	16kHz
音響モデル	4 状態 3 ループ Diagonal
mixture	3

表 9　手動ラベルと自動ラベルの差
Table 9　Difference Hand Labeling and Automatic Labeling

	了解度　正解率 (%) 評価音節数: 1,000			オピニオンスコア 評価単語数: 50		
	FTK	FYN	平均	FTK	FYN	平均
手動ラベル	94.4	98.4	96.4	4.3	4.3	4.3
自動ラベル	94.0	96.4	95.2	4.1	4.2	4.15

り大きな差がない．また，了解度では，ほとんど差がない．

なお，文献 [16] では，本実験の結果と同様に，波形接続型音声合成において，手動ラベルと自動ラベルで得られる合成音声に，品質に差があまりないことが紹介されている．

5.5　素片の選択単位の違いによる品質の差（音節単位と音素単位）

音節接続型音声合成では，必要とされる音声素片をすべて録音する必要がある．そのため実際に使用するばあい，大量の録音音声が必要になる．この大量の録音音声を削減するため，音節素片ではなく，音素素片で音声合成を試みる．前後音素環境・モーラ数・モーラ位置・アクセントは，両者とも同様に扱う．結果を表 10 に示す．

表 10　音節素片と音素素片の違い
Table 10　Difference between Syllable Unit and Phone Unit

	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
音節素片	99.4	99.7	99.6	4.2	4.4	4.3
音素素片	99.1	99.5	99.3	4.0	4.2	4.1

表 10 から，オピニオンスコアでは，音節単位で作成するほうが，音素単位で作成するより，評価が高いことがわかる．

5.6　音節素片の接続位置

自然性向上のための補助的な操作として，音節素片の接続位置を，音節の中心とすべきか，境界とすべきかの問題がある．手動ラベルにおいて，この問題を調査した．結果を表 11 に示す．

表 11　音節素片の接続位置
Table 11　Concatenation Position for Syllable Unit

	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
音節境界	99.4	99.7	99.6	4.2	4.4	4.3
音節中心	98.4	99.6	99.0	3.4	3.8	3.6

表 11 から，提案手法では，音節境界で合成したほうが，オピニオンスコアの良い音が作成できることが示された．

なお，PSOLA の音声合成においては，接続位置を音節素片の中心としたほうが，自然性の高い音が合成できるようだ [5]．

5.7　接続部分の違和感の軽減について

本論文では，音節素片の接続部分の違和感を軽減するために，2.4 節で述べたような方法を用いた．これは，まったく考慮しない場合に比べて違和感を軽減することができたが，違和感が残る音声もある．それらの音声のスペクトログラムを見たところ，音節間の不連続感が比較的大きく残っていた．したがって，データベースの拡充やセグメンテーション精度の向上と同時に，接続部でのスペクトル変動が滑らかになるような音節素片を選択することによって，合成音声の品質の向上が期待できる．

また，音節素片の接続部分にクロスフェード [15] をかけることで違和感が軽減できることが知られている．しかし予備的な実験をおこなったが，単純に接続した場合と大きな違いはなかった．

5.8　録音データ量

提案手法では，合成に必要な音節素片の，前後音素環境と，音節素片が属する単語のモーラ数と，音節素片の単語中のモーラ位置と，音節素片が属する単語のアクセントごとに異なる音節素片を，予め録音しておく必要がある．そこでまず，日本語の普通名詞における音節素片の種類数 $|\{Sy(P,N)_{m,M,ac}\}_W|$ とそれを全てカバーするのに必要な録音件数 $|W_R|$ を調査した．

- NHK 日本語発音アクセント辞典には，101,700 単語が掲載されている．この単語を全て提案手法で合成すると仮定する．
- この 101,700 単語において出現する音節素片の種類を調べたところ，384,845 個であった．

- これらの音節素片が全て含まれるように，NHK 日本語発音アクセント辞典の単語を選出すると，43,000 単語の発話が必要であった．
- したがって NHK 日本語発音アクセント辞典に掲載されている単語の 42.2% を収録する必要がある．

なお，43,000 単語は，単語の選択方法により，より少なくできる可能性が高い．

現在，規則音声合成では，録音した録音データに対して決定木をもちいたクラスタリングを行うことで，存在しない音素素片をデータベース上に存在する音素素片に割り当てている [18]．これと同様な手法を音節接続型音声合成において利用できる．この場合，すべての音声を合成することが可能であるが，合成音声の品質は低下する．今後，決定木をもちいた場合の録音音声のデータ量と得られる合成音声の品質の関係を調べていきたい．

5.9 音節の選択方法に関する今後の見通し

今回の実験では，音節素片の選択において複数の候補が残った場合，ランダムに選択する方法を採った．しかし，コーパスサイズが大きくなるにつれ，安定した品質を得ることは困難となる．特に音量のバラツキがでてくる．これらを抑制するため，録音環境 (日時，場所など) が分かる場合，それらが近い音声を優先的に用いることで，ある程度改善することができると考えている．しかし，音響的な物理的尺度もしくは知覚的尺度の導入が必要になると考えている．

6. おわりに

本研究では，音節波形接続方式を普通名詞に応用したときの，合成音声の品質を調査した．音節波形接続方式は，音節素片を，音節素片が属する単語のモーラ数と音節素片の単語中のモーラ位置と音節素片が属する単語のアクセントの言語パラメータのみを用いて選択し，音響的なパラメータを用いていない．

聴覚実験における合成音声の単語了解度は，手動ラベルの場合で 99.6%が得られた．また，オピニオンスコアはそれぞれ 4.3 が得られた．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 であった．また，対比較試験において，合成音声の平均 19.4%が自然音声より自然であるとの結果を得た．

これらの結果から，アクセントを考慮した音節接続型音声合成で得られた普通名詞は，非常に自然性の高い合成音声であることが示された．

今後は，決定木を用いたクラスタリングを用いて，任意の音声を合成できるようにしたい．また，波形候補が複数残った場合の絞り込み手法や，考察で述べた手法の検討を行い，さらに自然音声に近い合成音声の作成を目指したい．

文　献

[1] 広川智久, "波形辞書を用いた規則合成法", 電子情報通信学会技術研究報告, SP88-9, pp.65–72, 1988.

[2] 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.

[3] 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.

[4] Mitsuaki ISOGAI, Kimihito TANAKA, Satoshi TAKANO, Hideyuki MIZUNO, Masanobu ABE, and Sin 'ya NAKAJIMA, "A NEW JAPANESE TTS SYSTEM BASED ON SPEECH-PROSODY DATABASE AND SPEECH MODIFICATION", ICSLP 2000, 2000.

[5] Chang K. Suh, Takehiko Kagoshima, Masahiro Morita, Shigenobu Seto, and Masami Akamine, "TOSHIBA ENGLISH TEXT-TO-SPEECH SYNTHESIZER (TESS)", Euro Speech 1999, S10.PO2.14,Volume 5, Page 2111-2114, 1999.

[6] 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.

[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] 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.

[10] 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.

[11] 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.

[12] Entropic Signal Processing System (ESPS), http://www.entropic.com/esps.html.

[13] 平井俊男, 天白成一, 鹿野清宏, "音素間のスペクトル遷移の始終端情報を用いた音声合成", 電子情報通信学会技術研究報告, SP2001-119, pp.39–44, 2002.

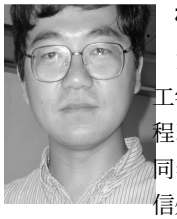
[14] NHK 出版, "NHK 日本語発音アクセント辞典 新版", ISBN 4-14-011112-7, 1998.

[15] 剣持 秀紀, 大下隼人, "コーラス音声の合成", 日本音響学会 2006 年春, 1-Q-23, pp.390-381, 2006.

[16] 河井恒, "波形接続型音声合成のための自動音素セグメンテーションの評価", 電子情報通信学会技術研究報告, SP2002-170, pp.5–10, 2003.

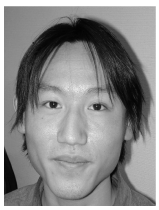
[17] Hidden Markov Model Toolkit (HTK),

http://htk.eng.cam.ac.uk/.
[18] Alistair Conkie Mark C. Beutnagel Ann K. Syrdal Philip E. Brown, ”PRESELECTION OF CANDIDATE UNITS IN A UNIT SELECTION-BASED TEXT-TO-SPEECH SYNTHESIS SYSTEM”, ICSLP 2006, 2006.



村上　仁一　（正員）

1984 年筑波大学第 3 学群基礎工学類卒．1986 年筑波大学修士課程理工学研究科理工学専攻終了．同年 NTT に入社．NTT 情報通信処理研究所に勤務．1991 年国際通信基礎研究所 (ATR) 自動翻訳電話研究所に出向．1997 年鳥取大学工学部知能情報工学科に転職．現在に至る．主に音声認識のための言語処理の研究に従事．電子通信情報処理学会，日本音響学会，言語処理学会各会員．



石田　隆浩

2002 年鳥取大学工学部知能情報工学科卒業．2004 年鳥取大学大学院工学研究科知能情報工学専攻修士課程修了．2005 年鳥取三洋電気入社．現在に至る．

池原　　悟　（正員）

1967 年大阪大学基礎工学部電気工学科卒業．1969 年同大学院修士課程修了．同年日本電信電話公社に入社．数式処理，トラフィック理論，自然言語処理の研究に従事．1996 年スタンフォード大学客員教授．現在，鳥取大学工学部教授．工学博士．1982 年情報処理学会論文賞，1993 年同研究賞，1995 年日本科学技術情報センター賞（学術賞），同年人工知能学会論文賞，2002 年電気通信普及財団賞（テレコム・システム技術賞）受賞．電子情報通信学会，人工知能学会，言語処理学会，機械翻訳協会各会員．

Abstract For speech synthesis using a slot filling method, it is necessary to record a lot of speech data with the same speaker. To avoid this high speaker cost, "concatenative syllable speech synthesis" has been proposed. This synthesis is kind of corpus based and makes it possible to retain the speaker's features and high naturalness for proper nouns. We studied the quality of common nouns produced by this synthesis and obtained high quality speech synthesis by altering mora length and positions and pre and after syllable conditions and accents.

Key words speech synthesis, corpus based, mora length, mora position, accent