論　文

音節波形接続による単語音声合成

村上　仁一[†]　　水澤　紀子[††]　　東田　正信[†††]

Simple Word Synthesis by Concatenating Syllabic Components based on Positional Features with Mora Length

Jin'ichi MURAKAMI[†], Noriko MIZUSAWA[††], and Masanobu HIGASHIDA[†††]

あらまし

録音編集方式による音声合成において，可変部の単語件数が多い場合，必要な音声を全て同一話者の音声で録音するのは困難である．本論文では，必要な単語の一部を同一話者の声で録音し，その音声波形から切り出した音節波形を信号処理をしないで接続することで，録音していない単語音声を合成する方法を検討した．本方式は，各音節の韻律的な情報として単語のモーラ数と音節の単語内モーラ位置を使用し，ピッチ周波数やパワーの定量的な分析や予測を行わないのが特徴である．日本の地名を合成対象として必要録音件数の調査を行ったところ，1 万 7 千件の録音音声から 4，5，6 モーラ語の地名 10 万 5 千件が合成できることがわかった．また，地名を合成して聴覚実験を行ったところ，合成音声の品質も十分実用的なものであることがわかった．

キーワード

単語音声合成，録音編集方式，音節波形接続，韻律的特徴，モーラ位置，モーラ数

1. ま え が き

近年，音声を用いて情報を提供する自動応答システムが広く用いられている．このようなシステムにおいてユーザごとに提供する情報が異なる場合は，入力を促したり確認を行ったりする手続きが不可欠となるため，ユーザの入力状況にきめ細かく対応した音声ガイダンスが必要である．

このような音声ガイダンスの文章を合成する方法として，録音編集方式が広く使われている．これは，出力するガイダンス文をユーザの入力に依存する単語 (以下「可変部」と呼ぶ) とそれ以外の部分 (以下「固定部」と呼ぶ) とに分けて別々に録音しておき，出力内容に応じた可変部を固定部に挿入して出力音声を作成する方法である．現在 NTT-BJ（NTT 番号情報株式会社）がサービスを提供している自動電話番号案内システム [1], [2]「あんないジョーズ」のガイダンス文も録音編集方式を採用しており，「お調べのところは○○でよろしいですか.」などのガイダンス文 (下線部 が固定部，○○が可変部が挿入される部分) に日本全国の地名，姓名，企業名などの単語を挿入し，「お調べのところは“神奈川県” “横須賀市” でよろしいですか.」といったガイダンス文を出力する．

通常このような用途に用いる音声は，固定部，可変部間，可変部同士間に生じる違和感を軽減するため，全て同一話者が発声したものを録音する．しかし例えば前出の自動電話番号案内システムでは，可変部として日本全国の地名 18 万件，姓名 22 万件，企業名 660 万件という膨大な件数の可変部音声が必要である．このように膨大な件数を全て同一の話者が発声するのは時間的に非常に困難であり，複数の話者が手分けして発声せざるを得ない．この結果，一応答文中に複数話者の声が混在し，ユーザに応答文の違和感を感じさせる原因となる．仮に同一話者による発声が出来たとし

[†] 鳥取大学工学部知能情報工学科，鳥取市
Department of Information and Knowledge Engineering Faculty of Engineering Tottori University 4-101, Minami Koyamachou, Tottori city, 680-8552 Japan
[††] NTT 東日本東日本電信電話株式会社，東京都
Nippon Telegraph And Telephone East Corporation Iidabashi First-Build. 5th floor 5-1 Kouraku 2-Chome Bunkyouku, Tokyo 112-0004 Japan
[†††] 国際通信基礎研究所，京都府相楽郡
Corporate Strategy Planning Division, Advanced Telecommunications Research Institute International 2-2-2, Hikaridai, Seikacho, Sourakugun, Kyoto 619-0288, Japan

ても，収録が長期間にわたるため，高さや速さが均質な音声を得るのは難しく，一応答文中で声質にばらつきが生じるという問題が残る．

上記のような問題に対して，固定部まで含めて規則合成などの音声合成方法で合成するという解決方法も考えられるが，人間の自然音声に近い高品質な合成音声が得られるとは言い難い．また，別話者の発声による単語音声を話者変換技術により全て同一話者の音声に変換するという解決方法も考えられるが，現在の技術では音質の劣化は免れない．

そこで，本研究では，固有名詞の音声合成に限定して，選択した音節を信号処理をせずに単に接続して他の単語音声を合成する方法を検討する．そして波形の選択のパラメータとして単語内のモーラ位置と単語のモーラ数と音素環境を利用する．この方式は，信号処理を行わないため，話者性や自然性を残したまま他の単語音声が合成できる．録音すべき単語数が，同一話者が短期間で発声出来る数であれば，固定部と同じ話者に単語も読み上げてもらうことでシステムからの応答を全て同一話者の音声で出力でき，ユーザに与える不自然さを軽減できる．具体的には日本の地名を合成対象とし，検討を行う．

2. 音声合成方法の検討

2.1 概　　要

本論文の合成対象である単語音声は，ガイダンス文の可変部に挿入して使用される．固定部と可変部の間の違和感を軽減するためには合成音声が固定部の話者の声質と高い自然性を持つことが望まれる．

話者性および自然性を持った音声を合成するために，合成したい話者の音声を収集し，そこから切り出した音声波形に全く処理を加えずに接続する方式が ATR の Nick Campbell らによって提案されている [3], [4]．この方式は話者性の保存という点ですぐれた成果をあげている [3]．この音声合成方法 CHATR は合成対象を文にしている．そしてピッチ周波数を TOBI モデルで推定している．そのため韻律が不自然に聞こえる場合がある．また音質がデータベースの品質に大きく依存し，しかもそのデータベースの作成方法が議論しにくい．

本論文では, [3] と同様に，収集しておいた単語の録音音声から適切な部分を切りだし，その波形に信号処理を施さずに接続する方式を採用する．ただし，本論文では合成対象を，ガイダンス文の固定部，すなわち固有名詞に限定する．そして，従来の方法 [3] と異なる波形の接続単位や韻律的な特徴を表わすパラメータを採用する．本論文で用いた韻律的特徴を表わすパラメータを表 1 に示す．これらのパラメータを採用したことにより，どのような単語を録音すれば良いのか，その単語はどのようにして選択すれば良いのかを明確に決定することができる．

表 1　本論文で用いるパラメータ
Table 1　Parameters for proposed speech synthesis

1	単語のモーラ数
2	単語内のモーラ位置
3	前後の音素環境

2.2 韻律的な特徴を表わすパラメータ（モーラ数とモーラ位置と音素環境）

本来，自然な文章の音声合成を目指すのであれば，文章全体に渡って韻律的特徴を予測し，呼気段落レベル，文節レベル，音節レベル，と徐々に細部の特徴を決定していく必要がある．しかし本論文では録音編集方式によるガイダンス文の中の可変部の音声合成を前提とし，同一話者が大量の単語を全て発声できた場合と同じ音声が得られることを目標としている．そこで固有名詞の音声を合成する場合について考える．

韻律を構成する具体的なパラメータとして，ピッチ周波数，パワー，音韻継続長などが考えられる．通常はモデルを用いてこれらのパラメータを予測し，合成音声が予測された値を持つように合成単位の波形をデータベースから選択したり，選択した波形に処理を加えたりする [3]．

本論文では，各合成単位が持つ韻律的特徴 (高さ，強さ，長さなど) を個別のパラメータで扱うのではなく，固有名詞（単語）内のモーラ位置と固有名詞（単語）のモーラ数で代表させることができると仮定する．

2.2.1 ピッチ周波数（モーラ数とモーラ位置）

単語のピッチパターンはアクセント型によらず語頭で上昇したのち語尾に向かって下降する成分とアクセントの位置に対応して上昇・下降する成分の和として近似できる [7]．日本語の単語アクセントは「低高高高低」のように各音節ごとの高低で表わされるため，各音節のピッチ周波数は単語のモーラ数とアクセント型，および音節の単語内における位置で表現できると仮定できる．したがって M モーラ語であれば理論上 $M+1$ 通りのアクセント型が存在する．

しかし，合成単語の対象が「地名」「姓名」のような

固有名詞の場合，アクセント型の出現に偏りがある．実際に 4〜6 モーラの地名を同一話者が発声した音声から 100 件ずつランダムに選んで検聴したところ，ほとんどのアクセント型は $M-2$ 型，もしくは 0 型であった．

さらに，これらの単語音声のピッチ周波数を Xwaves+ [9] を用いて調査した．図 1 に単一話者が発声した地名 4 モーラ語 1,500 件のピッチ周波数の平均値と分散を示す．同様に 5 モーラ語 2,800 件の結果を図 2 に，6 モーラ語 2,200 件の結果を図 3 に示す．

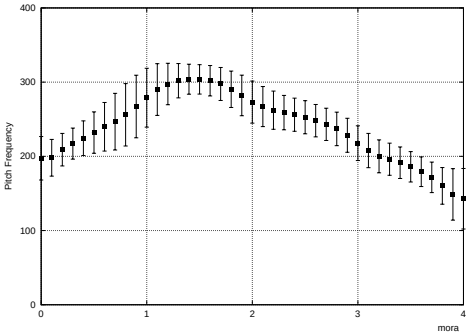

図 1　4 モーラ語 1,500 件の f_0 の平均と分散
Fig. 1　Means and deviation of f_0 for 1,500 word (4 mora)

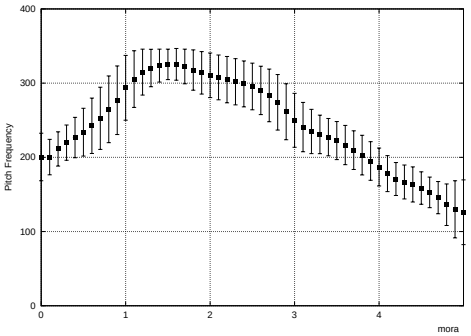

図 2　5 モーラ語 2,800 件の f_0 の平均と分散
Fig. 2　Means and deviation of f_0 for 2,800 word (5 mora)

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，縦棒の長さは分散を示している．この結果から，モーラ数が同じ場合，ピッチ周波数の分散は非常に小さく，アクセント型を意識しなくて良いと考えた．従って，4〜6 モーラの地名の場合には，各音節のピッチ周波数は単語のモーラ数と音節のモーラ位置で表現できると仮定した．

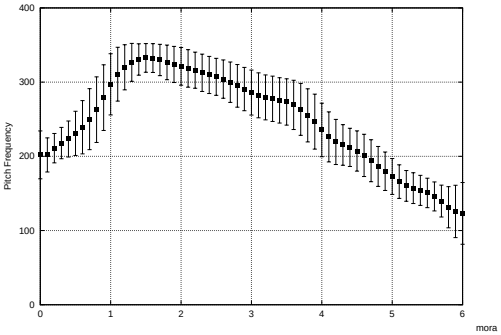

図 3　6 モーラ語 2,200 件の f_0 の平均と分散
Fig. 3　Means and deviation of f_0 for 2,200 word (6 mora)

2.2.2　パワー，音韻継続長 （音素環境）

パワーと音韻継続長は音韻種類と前後音素環境に大きく依存していると思われる．また，単語発声の場合，一般に語頭は強く発声され，語尾に向かってパワーは弱くなる．語尾の音節は語中に比べて比較的長めに発声される．そして短い単語は比較的遅く (つまり各音節は長く)，長い単語は比較的早く (各音節は短く) 発声される傾向がある．そのため，音節種類と前後音素環境に加えて各音節の単語内モーラ位置と単語のモーラ数を用いることで，各音節のパワー，音韻継続長を十分記述できると予想した．

以上のことから，本論文では，地名の音声合成の場合は各音節の持つ韻律的特徴を単語のモーラ数と各音節の単語内モーラ位置と前後音素環境で表現することができると仮定した．

なお，本論文では，モーラは拗音「ャ」「ュ」「ョ」以外の仮名 1 文字 (促音「ッ」を含む) を 1 モーラと数え，仮名 2 文字で表わされる音節のモーラ位置は 2 文字のうち最初の方のモーラ位置とする．例えば「東京都」は 5 モーラの単語であり，音節として「トー」「キョー」「ト」の 3 つに分かれ，各音節のモーラ位置は 1，3，5 となる．

2.3　波形の接続単位（音節）

音声の波形を単に接続する合成方法では，同じ音韻でも発声環境によりさまざまな特徴を持った音声波形が必要になる．そのため高品質な合成音声を得るためには大規模なデータベースが必要になる．データベース規模を小さく抑えるためには波形の接続単位は短い方が良いため，既存の研究では接続の最小単位を音素としているものが多い [3], [5]．しかし本論文では了解

度の点を考慮して，波形接続の最小単位を音節とする．

音節の知覚においてスペクトルの最大変化点 (すなわち子音から母音への遷移領域) が重要な役割を果たすことが指摘されている [6]．音素単位の接続の場合，データベースから選択された音声波形によっては接続時に必ずしも滑らかなスペクトル遷移が得られるとは限らない．この場合は音節明瞭度が悪化し，合成された単語音声の了解度が落ちると考えられる．合成音声がユーザに情報を提示するために使用される場合，特に前出の番号案内システムのように文脈が理解の補助にはならない場合には，単語了解度は重要である．従って，音節明瞭度を重視し，子音から母音への遷移領域は切り離さずに扱う方が良いと考えられる．

2.4 録音必要件数

提案手法においては，モーラ数とモーラ位置と前後音素環境が異なる音節部品を準備する必要がある．そのため，大量の単語を録音する必要がある．音節部品用に録音する数が合成対象の全件数と比較して大差ないようでは本論文の意味はあまりない．しかし，一般に自然言語には，Zipf の法則 [10] が成り立つため，全ての音節部品を必要としない．

この録音件数の問題は 4.2 節の地名の合成音声の実験において検討する．

3. 提案する合成方法

以下，本論文で提案する音声合成方法の具体的な手順を述べる．

3.1 音節データベースの作成

(a) **音節部品列の生成:**

音声出力したい単語各々を W_i, その集合を $\boldsymbol{W}=\{W_i\}$ $(i = 1 \sim L$, ただし L は単語の総数) とする．$\boldsymbol{W}$ 中の全ての単語を音節部品ラベルの列 $\{Sy(P,N)_{m,M}\}_i$ $(m = 1 \sim M_i, i = 1 \sim L)$ に直す．音節部品ラベル $Sy(P,N)_{m,M}$ は以下の情報からなる．

Sy：音節
P：直前の音素 (前音素環境)
N：直後の音素 (後音素環境)
m：単語中の当該音節のモーラ位置
M：単語のモーラ数

例えば「ヨコハマシ」の音節部品列は $\{$yo(, /k/)$_{1,5}$ ko(/o/, /h/)$_{2,5}$ ha(/o/, /m/)$_{3,5}$ ma(/a/, /s/)$_{4,5}$ si(/a/,)$_{5,5}\}$ となり、「ヨコハママチ」の音節部品列は $\{$yo(, /k/)$_{1,6}$ ko(/o/, /h/)$_{2,6}$ ha(/o/, /m/)$_{3,6}$ ma(/a/, /m/)$_{4,6}$ ma(/a/, /t/)$_{5,6}$ ti(/a/,)$_{6,6}\}$ となる．「ヨコハマシ」と「ヨコハママチ」の「ヨ」,「コ」,「ハ」は，単語のモーラ数が違うため異なる音節部品ラベルとなる．

(b) **録音:**

(1) (a) で生成した音節部品列に出現する音節部品ラベルを $\boldsymbol{W}$ 全体に渡って収集し，異なり音節部品ラベルの集合 $\{Sy(P,N)_{m,M}\}_{\boldsymbol{W}}$ を抽出する．

(2) その単語集合から得られる異なり音節部品ラベルが $\{Sy(P,N)_{m,M}\}_{\boldsymbol{W}}$ をカバーするような単語の集合 $\boldsymbol{W_R}$ を $\boldsymbol{W}$ から抽出する．

(3) $\boldsymbol{W_R}$ 中の単語を全て同一話者に発声してもらい，録音する．

(c) **音節部品の作成:**(図 4)

(1) 録音した音声をラベリングして音節区間を決定し，録音単語の ID と各音節波形の位置からなる，音節波形インデックスを作成する．

(2) 音節部品ラベルに対応する音声波形のインデックスを加え，音節部品とする．

(3) $\boldsymbol{W_R}$ 中の単語全てにわたって上記の操作を行い，得られた音節部品でデータベースを作成する．

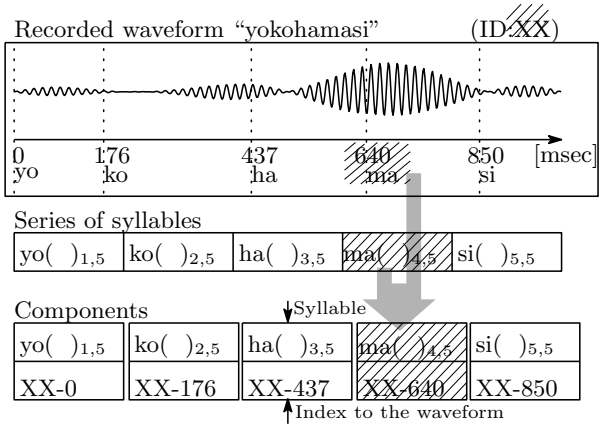

図 4 提案する音声合成方法
Fig. 4 Speech synthesis process

3.2 合成，出力

出力したい単語 W_j が $\boldsymbol{W_R}$ に含まれる場合は，その録音音声をそのまま用いる．$\boldsymbol{W_R}$ に含まれない場合は (a) で生成した音節部品列 $\{Sy(P,N)_{m,M}\}_j$ を参照して必要な音声部品を音節部品データベースから選択し，波形のインデックスを参照して録音音声から各音節波形を切りだし，接続して合成する．

例えば $\boldsymbol{W_R}$ に含まれない地名「ヨコカワシ」を出

力しようとする場合，音節部品列を参照して必要な音節部品ラベル $\mathrm{yo}(,/\mathrm{k}/)_{1,5}$ $\mathrm{ko}(/\mathrm{o}/,/\mathrm{k}/)_{2,5}$ $\mathrm{ka}(/\mathrm{o}/,/\mathrm{w}/)_{3,5}$ $\mathrm{wa}(/\mathrm{a}/,/\mathrm{s}/)_{4,5}$ $\mathrm{si}(/\mathrm{a}/,)_{5,5}$ を得る．データベースからこれらの音節部品を選択するとき，例えば同じ 5 モーラ語「ヨコハマシ」の $\mathrm{yo}(,/\mathrm{k}/)_{1,5}$，や $\mathrm{si}(/\mathrm{a}/,)_{5,5}$ は選択できるが，$\mathrm{ko}(/\mathrm{o}/,/\mathrm{h}/)_{2,5}$ は後音素環境が異なるため 2 モーラ目の「コ」であるにも関わらず選択できない．

4. 地名音声の合成実験

提案する合成方式を用いて，日本の地名の音声を合成し評価する．なお地名は，NTT の電話帳に記載されている地名（県名，市町村名，町大字名，町小字名）18 万件 [1] を利用する．

4.1 合成音声の対象

本論文では以下の理由から，合成対象 $\boldsymbol{W}$ を 4，5，6 モーラの地名に限定する．

（1） 2.2.1 節で述べたように，これらの地名に関してはアクセント型を意識しなくて良いと考えられる．

（2） 町字などまで含めると，日本全国の地名の件数はおよそ 18 万件 [1] に及ぶが，その中で 4，5，6 モーラの地名は 105,000 件，すなわち 6 割を占めるため，録音件数の削減効果が大きい．

（3） 7 モーラ以上の地名は 55,000 件，全体のおよそ 3 割を占める．しかし，これらの地名は，例えば「綾野町+東 (アヤノチョウ+ヒガシ)」のように，ポーズを入れることで短い単語に分割できる．したがって 6 モーラ以下の単語で合成できれば，それらを組み合わせて 7 モーラ以上の地名を合成できる．

4.2 録音必要件数の調査

提案手法においては，前後音素環境に加えてモーラ数，モーラ位置ごとに異なる音節部品を予め録音しておく必要がある．そこでまず，音節部品の種類数 $|\{Sy(P,N)_{m,M}\}_{\boldsymbol{W}}|$ とそれを全てカバーするのに必要な録音地名件数 $|\boldsymbol{W_R}|$ を調査した．

理論的には，モーラ数 M の単語を合成するにはおよそ $|Sy||P||N|M$ 種類の音節部品が必要となる．(ただしここで $|Sy|$,$|P|$,$|N|$ は各々音節 Sy，前音素環境 P，後音素環境 N の種類数を表す．) 音節として長音を考えずに $|Sy|=100$, $|P|=7$, $|N|=18$ (注1)とすると，5 モーラ語の場合およそ 63,000 種類の音節部品が必要となる．音節種類として長音も考えるとこの数はさらに増える．

しかし，実際の 5 モーラの地名全て (およそ 38,500 件) について出現する音節部品の種類を調べたところ，音節種類に長音を含めてもその数はおよそ 10,700 種類であった．必要な音節部品が全て含まれるような地名のセットを選出すると，その件数は 6,000 件強(注2)，合成対象である 5 モーラ語全体の 16%程度であった (表 2 参照)．

同様に調査した 4 モーラ語，6 モーラ語の結果を表 2 に示す．合成対象 $\boldsymbol{W}$ を 4〜6 モーラの地名 ($|\boldsymbol{W}|$ =105,287 件) とした場合，必要な音節部品の種類数 $|\{Sy(P,N)_{m,M}\}_{\boldsymbol{W}}|$ はおよそ 30,000 個であり，録音が必要な件数 $|\boldsymbol{W_R}|$ は合成対象全件数の 16%，17,000 件であった．

表 2　音節部品種類数と録音地名件数
Table 2　Number of syllable parts and Number of recording words

単語モーラ数	4	5	6	4〜6 合計
必要な音節部品種類数 $\lvert\{Sy(P,N)_{m,M}\}_{\boldsymbol{W}}\rvert$	7,606	10,670	12,025	30,301
地名の全件数 $\lvert\boldsymbol{W}\rvert$	31,222	38,430	35,635	105,287
録音する件数 $\lvert\boldsymbol{W_R}\rvert$	4,560	6,066	6,262	16,888

4.3 評価実験

提案した合成方法を用いて，5 モーラ語の地名音声を作成して，ナレータが発声した音声および市販の合成音声と比較実験を行った．

4.3.1 評価データ

600 件の 5 モーラ語の地名をランダムに選び，それらを同一話者が読み上げ，録音した．これを手作業でラベリングして 2,810 個の音節部品を得た．同じ種類の音節部品が複数ある場合は，その中からランダムに一つずつ選択し，1,549 種類の音節部品を得た．これらを用いて実在する 5 モーラの地名を合成したところ，3,167 件を合成することが出来た．

合成できた 3,167 件のうち 100 件を選び，単語了解度試験とオピニオン評価を行った．ただし 100 件の地名は，単語了解度試験への影響を避けるために被験者になじみのない地名を選ぶようにし，さらに，条件の厳しい音声の評価を行うために複数の音声部品が同一発声 (同じ地名の発声) から連続して選択されているものは避けて選択した．合成した地名の一部を表 3 に示す．

(注1)：前音素環境は母音，撥音「ン」，促音「ッ」．後音素環境は母音 5 種類と各行に対応する子音，撥音．ただし子音/k/，/t/，/p/と促音はまとめて 1 種類とした．

(注2)：必ずしも最適値ではない．選出のアルゴリズム次第でさらに減る可能性がある．

表 3　合成した地名（一部）
Table 3　Examples

表記	よみ	表記	よみ
青上	アオンジョウ	瓜喰田	ウリクイダ
乙沖田	オツオキダ	川原保	カワラツポ
九合取	クゴウトリ	下居合	シモイアイ
拾貫内	チコウウチ	中瓜生	ナカウリュウ
不動岩	フドウイワ	轆轤石	ロクロイシ

4.3.2　評価方法（単語了解度試験とオピニオン評価）

合成した音を評価するために，単語了解度試験およびオピニオン評価を行った．評価方法の具体的な方法を以下に示す．

（1） 提案方法による合成音声と自然音声と市販の合成音声

合成音声 100 件と同一話者で同一発話内容の自然音声 100 件を収録した．また，市販の合成音声と比較するため F 社の linux で動作する音声合成ライブラリを使用して合成音声を作成した．なお，この合成音声を作成するとき，音声品質がもっとも良くなるように，各種パラメータを操作した．

（2） 評価用ガイダンス文

提案方法による合成音声 100 件と自然音声 100 件と市販の合成音声 100 件を，同一のガイダンス文に埋め込んで評価を行った．具体的には表 4 のガイダンス文を各々100 件作成した．

表 4　評価ガイダンス文
Table 4　Evaluation sentence

「入力された住所は」（地名の合成音声もしくは自然音声）「で，よろしいですか.」

（3） 評価者

単語了解度試験，オピニオン評価とも，被験者には音声研究に携わった経験のない人 6 名を選び，ヘッドフォンからガイダンス文を提示した．

（4） 単語了解度試験

提案方法による合成音声と，自然音声と，市販の合成音声，各 100 件を同一のガイダンス文に埋め込んで評価を行った．自然音声と合成音声は，ランダムな順序で提示し，聞こえた地名の音声を仮名で書き取るよう指示した．自分の知識や，前に聞いた地名のことは考えず，そのときに聞こえた音を書き取るよう指示した．

（5） オピニオン評価

提案方法による合成音声と，自然音声と，市販の合成音声，各 100 件を同一のガイダンス文に埋め込んで評価を行った．自然音声と提案した合成音声と市販の合成音声は，ランダムな順序で提示し，各々の音声の自然さについて 1 から 5 の間の 5 段階 (1 が最も不自然，5 が最も自然) で評価するよう指示した．同じような地名が何度か出てきても，それ以前に聞いた音と比較せずに評価するよう指示した．

4.4　評価試験結果

表 5 に評価試験の結果を示す．なお，ナレータの声質により合成音声の品質が変わる場合を考慮して，ナレータを変えて同様の方法で評価試験を行った．ただし，両者の合成音声の発話内容 (地名) は，異なる．これらの結果を表 5 に示す．

表 5　評 価 結 果
Table 5　The result of evaluation

	単語了解度試験 正解率 (%)			オピニオン評価 不自然 1:→ 5:自然		
	ナレータ A	ナレータ B	平均	ナレータ A	ナレータ B	平均
本手法による合成音声	97.9	99.1	98.5	4.13	4.03	4.08
自然音声	97.9	99.6	98.8	4.86	4.91	4.89
市販の合成音声	90.9	94.3	92.6	1.76	1.72	1.74

（1） 単語了解度

評価実験より，本論文で提案した合成音声の単語了解度は，平均 98.5%であった．一方自然音声の単語了解度は，平均 98.8%であった．したがって，本論文で提案した合成音声は，自然音声とほぼ同程度の単語了解度が得られることが示された．一方，市販の合成音声の単語了解度は平均 92.6%であった．したがって，誤り率からみると，本論文で提案した方法は，市販の合成方法と比較すると，誤り率が 76.7% (100.0-(2.13/9.13)*100) も減少していることがわかる．一方，自然音声と提案した方法を比較すると，誤り率は 2.9% (100.0-(2.07/2.13)*100) しか低下しないことが示された．

（2） オピニオン評価

オピニオン評価では，自然音声が平均 4.89 であったのに対し，提案手法が平均 4.07 と若干低くなった．しかし，市販の合成音声が平均 1.74 であったことを考慮すると十分に高い自然性が得られることが示された．

以上のことから，本論文で提案して得た合成音声は，自然音声に近い単語了解度と自然性を持っていることが示された。

5. 考　　察

（1） 地名の音声合成

提案した合成方法が有効であった理由として，以下の点を考えている．

（a） 固有名詞

本論文の合成対象は「地名」「姓名」などの固有名詞であるため，必要な音韻の種類や韻律のバリエーションは限定されている．

（b） アクセント型

ナレータ（発話者）になじみのない地名は，同一のアクセント型で発声されることが多い．そのため，多くの固有名詞が同一のアクセント型で発話されている．

（c） 録音した音声品質

合成に利用した音声は，聞き取りやすいように比較的，低速で明瞭で均一に発声されている．

これらを考慮すると，本論文で提案した合成方法は，特に固有名詞の音声合成に, 有効である可能性がある．

（2） 普通名詞の合成実験

本論文で提案した手法の有効性を普通名詞で評価するために，ATR の A セット 5240 単語を用いて実験を行った．話者には A セットの女性話者 10 名の中から，比較的ピッチの低い FTK と FYN を利用した．合成方法や評価方法は 4. 章と同様である．ただし，評価した単語は 4 モーラ語の 50 単語である．合成したリストの一部を表 6 に示す．

表 6　合成した普通名詞 （一部 ATR A セットより）
Table 6　Example of common noun

表記	よみ	表記	よみ
体積	たいせき	団結	だんけつ
人格	じんかく	間接	かんせつ
バイ菌	ばいきん	段階	だんかい
媒介	ばいかい	面接	めんせつ
栽培	さいばい	間隔	かんかく
催促	さいそく	国立	こくりつ

また，評価実験の被験者は 5 名で行ない，評価音声は文ではなく単語とした．つまり，合成した単語の前後にガイダンスをつけずに単語了解度試験およびオピニオン評価を行った．得られた実験結果を表 7 に示す．

今回の合成した単語は普通名詞である．そのため地名の実験と比較すると，類推ができるため，単語了解度はかなり高くなった．しかし，全体の傾向は地名の実験と大きな差はなかった．本論文で提案した手法では 99.9%となり，自然音声の 100%には及ばないが，

表 7　評価結果 ATR Aset
Table 7　The result of evaluation ATR A set

	単語了解度試験 正解率 (%)			オピニオン評価 不自然 1:→ 5:自然		
	FTK	FYN	平均	FTK	FYN	平均
本手法による合成音声	99.8	100.0	99.9	2.96	3.26	3.11
自然音声	100.0	100.0	100.0	4.15	4.12	4.14
市販の合成音声	99.6	99.8	99.7	2.13	2.28	2.21

市販の合成音声 99.6%より高い値を得た．また，オピニオン評価も本論文で提案した手法は 3.11 が得られた．自然音声の 4.14 と比較すると低いが，市販の合成音声 2.21 と比較すると高い自然性をもっていることが示された．

この結果から，本論文で提案した手法の有効性が普通名詞でも確認できたと考えている．

（3） アクセント型が異なる単語

本論文では，単語のアクセント型を考慮せずに，単語のモーラ数，単語内のモーラ位置，前後の音素環境をパラメータとして使用することで，高い単語了解度と自然性を持つ合成音声を得た．この結果は，固有名詞の多くは同一のアクセント型を持っているためと思われる．

しかしながら，可能性は低いが，アクセント型が大幅に異なる単語を合成する場合や，録音リストにアクセント型が大幅に異なる単語を含む場合がある．このときの解決方法を以下に考察する．

（a） アクセント型が異なる単語を合成する場合

本論文で対象にしたのは，録音編集型の固定部の合成である．したがって，合成した音声の品質が低い場合，自然音声にすることで問題が解決できる．特に 2 モーラの単語は，アクセントによって意味が異なるため，本論文で提案した方法は適用範囲外で録音するのが better であると考えている．

（b） アクセント型が異なる単語が録音リストにある場合

本論文では，音節部品は種類ごとに 1 つずつ準備した．したがってアクセント型が大幅に異なる単語が録音リストにある場合，合成した音声の品質が低下する可能性がある．

しかし，通常，音節部品の候補は複数得られる．このような場合，ピッチ周波数やケプストラム係数やパワーなどのパラメータを考慮して，つながりの良い音声部品を選択する，などの方法がある．この解決方

法により，アクセント型が異なる単語が録音リストにあっても，問題となる音節部品が使用されなくなると考えている．

（4） 音節部品の選択（有声音の接続部分）

評価実験においてオピニオン評価が悪いサンプルを調べてみると，母音ー母音，母音ー半母音など有声音の接続部分における不連続感が悪影響を及ぼしていることがわかった．このようなサンプルを減らして全体的に音質を上げるために，以下のことを検討している．

本論文では，音節部品は種類ごとに1つずつ準備した．しかし，同一種類の音節部品の候補は複数得られる．この場合，ピッチ周波数やケプストラム係数やパワーなどのパラメータを考慮して，よりつながりの良いものを選択することで，より高品質な合成音声が得られると考えられる．

6. ま と め

本論文では，大量の固有名詞を同一話者の音声で出力する方法として，音声出力が必要な単語の一部のみを録音し，録音した単語音声から以下のパラメータに基づいて音節波形を切りだして接続することで，録音していない単語の音声を合成する方法を提案した．

1 単語のモーラ数
2 単語内のモーラ位置
3 前後の音素環境

録音すべき地名の件数を調査したところ， 4 ～ 6 モーラの日本の地名 105,000 件を合成対象とした場合，約 17,000 件の録音で合成対象全体をカバーできることが分かった．

また，5 モーラ語においてナレータ 2 名で音声を合成して評価試験を行った．その結果，平均の単語了解度は 98.5%，オピニオン評価は 4.08 が得られた．一方，自然音声では，単語了解度は 98.8%，オピニオン評価は 4.89 であった．また，市販の合成音声は，単語了解度は 92.6%，オピニオン評価は 1.74 であった．

したがって，自然音声と比較すると若干音質は落ちるものの，市販の合成音声と比較すると，極めて了解度が良く自然性が高い合成音声が得られることが示された．

7. 今後の課題

今後の課題を以下に述べる．

（1） 録音件数の削減

音節部品を統合することにより，更に録音件数を削減することを考慮する必要がある．

具体的には語頭用の $Sy(P,N)_{1,M}$ のように地名のモーラ数 M が異なっても韻律の特徴があまり変わらない音節部品や，$Sy(P,/\mathrm{b}/)_{m,M}$ と $Sy(P,/\mathrm{d}/)_{m,M}$ のように音響的特徴が似ている音節部品は一つの種類にまとめることで，録音件数はさらに削減できると考えている．

（2） 音節部品の選択

本論文では，音節部品は種類ごとに1つずつ準備した．しかし，通常，同一種類の音節部品の候補は複数得られる．今後，複数の候補の中から音節部品を選択する方法を検討する必要がある．例えば，音節部品の接続部分において，ピッチ周波数やケプストラム係数やパワーをパラメータに加えることで，音質の悪いサンプルを減らし，より高品質の音声が得られると考えている．

（3） 他の分野への応用

本論文において述べた音声合成方法は，地名などの固有名詞において特に有効であると思われる．また，普通名詞においても，ATR の Aset のデータを使用することで，有効性が認められた．今後は，この方法が，どこまで汎用性があるのか実験を行って行きたい．

なお，日本語の名詞において，アクセント型が異なると，意味が異なる単語の多くは，2 モーラ単語である．（例えは「雨」と「飴」,「橋」と「箸」.）また，6 モーラ以上の名詞は，名詞連続複合語である場合が多い．これらは，名詞の間に，ある程度長いポーズを間に入れると，アクセント核の移動 [11] を考慮しなくてすむため，短い単語に分割できる．以上のことから，本論文で用いた方法は，基本的に 3 モーラ以上の名詞の合成音声に利用できる可能性があると考えている．

謝辞

本研究を行うにあたり，国際通信基礎研究所 (ATR) 音声通信研究所の Nick Campbell 氏と，KDD 研究所の樋口宜男氏には，実験の方法に関してコメントを頂きました．NTT サイバースペース研究所の浅野久子氏と水野秀之氏には，合成方法に対して討論して頂きました．鳥取大学大学院工学研究科（博士前期）知能情報工学専攻 2 年の前田智広氏と鳥取大学工学部知能情報学科 4 年の石田隆浩氏には実験を手伝ってもらいました．これらの方に深く感謝致します．

また，作成したサンプルの合成音声は，以下の URL (注3) に置いてあります．

文　献

[1] M. Higashida, “A Fully Automated Directory Assistance Service that Accommodates Degenerated Keyword Input Via Telephones,” PTC97,pp167-174(1997)

[2] 東田 正信, 村上 仁一, 奥 雅博, “オペレータレス自動電話番号検索システムの開発,” 情処研報, 98-NL-123, pp25-32 (1998.1)

[3] N.Campbell and A.Black, “CHATR:自然音声波形接続型任意音声合成システム,” 信学技報, SP96-7, pp45-52(1996.5).

[4] ATR Interpreting Telecommunications Reseach Labs, “CHATR (Generic Speech Synthesis System),” http://www.itl.atr. co.jp/chatr

[5] 広川 智久, “波形辞書を用いた規則合成法,” 信学技報, SP88-9, pp65-72 (1988).

[6] 古井 貞熙, “音声知覚におけるおけるスペクトルの動的特徴の役割について,” 音講論, pp183-184 (1984.10).

[7] 藤崎 博也, 須藤 寛, “日本語単語アクセントの基本周波数パタンとその生成機構のモデル,” 音響誌, 27, 9, pp445-453 (1971.9).

[8] 水澤 紀子，村上 仁一，東田 正信, “音節波形接続による単語音声合成,” 信学技報, SP99-7, pp45-52 (1999.5).

[9] “waves+ Manual,” Entropic Research Laboratory, Inc. (March 19, 1996).

[10] 長尾 真, “自然言語処理,” 岩波講座 ソフトウエア科学, p23 (1996).

[11] 宮崎正弘, “日本文音声出力のための言語処理に関する研究”, 博士論文, 東京工業大学, (1986).

（平成 x 年 xx 月 xx 日受付）

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修了．1986 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．

水澤　紀子　（正員）

1995 年 東京工業大学 電気・電子工学科卒．1997 年 東京工業大学 物理情報工学専攻 修士課程修了．同年 NTT 入社 情報通信研究所において音声合成の研究に従事．2000 年 NTT 東日本法人営業本部システムサービス部へ異動．現在に至る．

東田　正信　（正員）

1975 年 東京大学工学系大学院物理工学修士課程修了．同年 日本電信電話公社 (現 NTT) 入社．1996 年 情報通信処理研究所，研究企画部長．1999 年より国際電気通信基礎技術研究所 (ATR) 取締役, 経営企画部長．現在に至る

主に，計算機のアーキテクチャ，自然言語処理，音声対話処理の研究に従事．

情報処理学会，AAAI，IEEE 各会員．

(注3)：http://unicorn.ike.tottori-u.ac.jp/murakami/paper/STUDY/SP_1999_5/sound/sound.html