# 音節波形接続による単語音声合成

水澤 紀子[†]　　村上 仁一[‡]　　東田 正信[†]

[†]NTT 情報流通プラットフォーム研究所
〒 239-0847 横須賀市 光の丘 1-1
0468-59-5858
{mizusawa,higasida}@isl.ntt.co.jp

[‡]鳥取大学 工学部
〒 680-8552 鳥取市 湖山町南 4-101
0857-31-6788
murakami@ike.tottori-u.ac.jp

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には必要な音声を全て同一話者の音声で準備するのが困難である. 本研究では必要な単語の一部のみ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ほぼそのまま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特徴を単語のモーラ数と音節の単語内モーラ位置で代表させており, ピッチ周波数やパワーの定量的な分析, 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および合成実験を行ったところ,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, 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.

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

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

Noriko Mizusawa[†]　　Jin'ichi Murakami[‡]　　Masanobu Higashida[†]

[†]NTT Information Sharing Platform Laboratories
1-1, Hikarinooka, Yokosuka-shi
Kanagawa 239-0847, Japan
+81-468-59-5858
{mizusawa,higasida}@isl.ntt.co.jp

[‡]Faculty of Engineering Tottori University
4-101, Minami Koyama-cho, Tottori-shi
Tottori 680-8552, Japan
+81-857-31-6788
murakami@ike.tottori-u.ac.jp

This paper proposes a method that synthesizes a huge number of words by concatenating syllabic waveforms obtained from recorded words selected carefully. This method assumes that just the mora position and the mora length represent enough of the prosodic features to realize synthesis at the word level.

An experiment study to synthesize Japanese city and town names with five moras by the proposed method was carried out. Competitive opinion scores were obtained by the synthesized voices compared with those obtained by actual voices. By our estimation, only 17,000 names need to be recorded to cover the remaining names with 3, 4, and 5 mora, which totals about 105,000.

word synthesis, slot filling method, syllable, prosodic features, mora position, mora length

# 1　はじめに

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るようになってきている. 提供する情報がユーザごとに異なる場合は, 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が, このためにはユーザの入力状況にきめ細かく対応した応答音声が必要である.

このような応答音声の合成方法として, 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章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 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」の応答音声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.」などのガイダンスの文章 (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.」といった応答音声を出力する.

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ても, 収録が長期間にわたると高さ, 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, ガイダンス文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, 波形を大きく操作するために, 音質の劣化は免れない.

そこで, 必要な単語の一部のみを同一話者の音声で録音し, その音声波形を利用して, 話者性を残したまま他の単語音声を合成する方法を検討した. 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, ガイダンス文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.

# 2　音声合成方法の検討

## 2.1　概要

本研究の合成対象である単語音声は自然音声である固定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を持つことが望まれる.

話者性, 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が, 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や十分性が議論できない.

本研究でも, 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, その波形になるべく信号処理を施さずに接続する方式を採用する. ただし, 波形の接続単位や韻律的な特徴を表わすパラメータが従来の方法と異なる. 特に韻律的特徴を表わすパラメータとしてモーラ位置とモーラ数という簡単なもの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決定することができる.

## 2.2 波形の接続単位

2.1 節で述べたような波形に対して信号処理を行なわずに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].

しかし本研究では以下の 2 点を考慮して, 波形接続の最小単位を音節とする.

### a. 了解度

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, 一つのものとして扱う方が良い.

### b. データベースサイズ

本研究の合成対象は単語の読み上げ音声であり, 聞き取りやすいようにゆっくり, 明瞭に発声されるべきものである. 従って, 自然文や感情を伴う音声を合成する場合に比べて抑揚が抑えられ, 各音韻ごとに用意しておくべき韻律のバリエーションが少なくて済むと考えられる. また, 合成対象単語は「地名」「姓名」など特定の分野のものなので, 一般的な単語全てを合成する場合に比べて, 用意すべき音韻の種類や韻律のバリエーションは限られる.

従って, 音節を接続単位として選び, 各音節について必要な韻律のバリエーションを用意してもデータベースを現実的なサイズに抑えることができると考えられる.

主に以上 2 点から波形接続の最小単位として音節を選択したが, 録音単語の選び方や波形の選択の方法によっては 2 音節以上分の波形が連続して選択される場合もある.

音節の種類については, 基本的に仮名一文字で表わされるもの (撥音「ン」, 促音「ッ」も単独で一音節とする) とし, 拗音「ャ」「ュ」「ョ」, 長音「ー」はその前の音節と合わせて一音節とする. また, 今回は母音の無声化は考えない.

また, 前後の音素との調音結合は無視できないと考えて, 各音節について, 前後音素環境に応じた複数の音節波形を用意する.

## 2.3 韻律を表すパラメータ

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かなくてはならない. しかし本研究では録音編集方式による音声合成を前提としており, 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ので, 孤立単語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.

本手法では, 各合成単位, すなわち音節が持つ韻律的特徴 (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, 単語内のモーラ位置と単語のモーラ数で代表させることができると仮定した.

**a.** ピッチ周波数

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した.

$M$ 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. しかし本研究のように合成対象単語が「地名」「姓名」のように限られている場合, 全てのアクセント型が均等に現れるわけではない. 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調査した. 図 $1$ に単一話

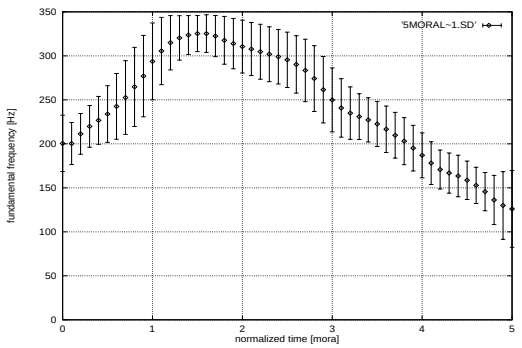


図 1: $5$ モーラ語 $2{,}800$ 件の $f_0$ 平均, 分散

者が発声した $5$ モーラ語 $2{,}800$ 件のピッチ周波数平均値と分散を示す. 時間軸はモーラ数 $5$ で正規化したのち計算した. 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. $4$,$6$ モーラ語も同様の傾向を示し, 分散も $5$ モーラ語と同程度であった.

従って, 少なくとも $4$〜$6$ モーラの地名の場合には, 各音節のピッチ周波数を単語のモーラ数と音節のモーラ位置で表現できると考えた.

**b.** パワー, 音韻継続長

パワーと音韻継続長は音韻種類, 前後音素環境に大きく依存していると思われるが, さらに単語の長さ, 単語中の位置に関して以下のような特徴がある.

孤立単語発声の場合, 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. また, 短い単語は遅く (つまり各音節は長く) 長い単語は早く (各音節は短く) 発声される傾向がある.

つまり, 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.

以上のことから, 本研究の場合は各音節の持つ韻律的特徴を前後音素環境と単語のモーラ数, 各音節の単語内モーラ位置という簡単な数値で表現することができると考えた.

モーラは, 拗音「ャ」「ュ」「ョ」以外の仮名一文字 (促音「ッ」を含む) を一モーラと数え, 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した. 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 $3$ つに別れ, 各音節のモーラ位置は $1$,$3$,$5$ となる.

## 2.4 提案する合成方法

以下, 提案する方法の具体的な手順を述べる.

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 $\{\mathrm{yo}(,/\mathrm{k}/)_{1,5}\ \ \mathrm{ko}(/\mathrm{o}/,/\mathrm{h}/)_{2,5}\ \ \mathrm{ha}(/\mathrm{o}/,/\mathrm{m}/)_{3,5}\ \mathrm{ma}(/\mathrm{a}/,/\mathrm{s}/)_{4,5}\ \ \mathrm{si}(/\mathrm{a}/,)_{5,5}\}$ となり、「ヨコハママチ」の音節部品列は $\{\mathrm{yo}(,/\mathrm{k}/)_{1,6}\ \mathrm{ko}(/\mathrm{o}/,/\mathrm{h}/)_{2,6}\ \mathrm{ha}(/\mathrm{o}/,/\mathrm{m}/)_{3,6}\ \mathrm{ma}(/\mathrm{a}/,/\mathrm{m}/)_{4,6}\ \mathrm{ma}(/\mathrm{a}/,/\mathrm{t}/)_{5,6}\ \mathrm{ti}(/\mathrm{a}/,)_{6,6}\}$ 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:

- (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.
- 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.
- $\boldsymbol{W_R}$ 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 2)

- 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.
- 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.
- $\boldsymbol{W_R}$ 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.

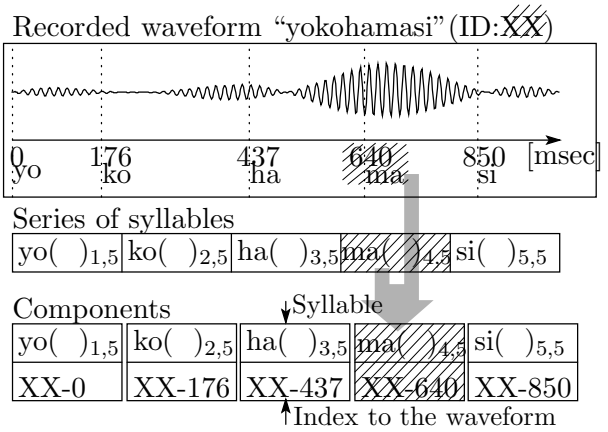


図 2: 音節部品の作成

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.

## 3 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を合成する実験を行った. ここでは以下の理由から, 合成対象 $\boldsymbol{W}$ を 4,5,6 モーラの地名に限った.

- 前述のように, 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.
- 字などまで含めると, 日本全国の地名の件数は

表 1: 音節部品種類数と録音地名件数

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

およそ 18 万件に及ぶが, 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.

- 7 モーラ以上の地名は 55,000 件, 全体のおよそ 3 割を占めるが, 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, 短い単語に分割できるものが多い. このように複合語的な長い地名の場合は, 提案手法の音節単位の接続よりも, 構成単語ごとの接続の方が録音数, 音質の面で有利になる可能性がある. 提案手法を用いる場合でも, 韻律的特徴に関して詳細な検討を行った上で音節部品種類が持つべき属性を決定する必要がある.

### 3.1 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準備しなくてはならない. 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 $|Sy| = 100$, $|P| = 7$, $|N| = 18$ [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[2] , 合成対象である 5 モーラ語全体の 16%程度であった (表 1 参照).

同様に調査した 4 モーラ語,6 モーラ語の結果を表 1 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.

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, 録音地名件数はさらに削減できると考えられる.

### 3.2 合成実験

実際に 5 モーラ語の地名音声を合成する実験を行った.

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 個の音節部品が得られた. 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 種類の音節部品が得られた. 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

[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.

[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.

することが出来た.

### 3.3 評価試験

合成できた 3,167 件のうち 100 件を選び, 下記のような単語了解度試験, オピニオン評価を行った. ただし 100 件の地名は, 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.

1. 100 件それぞれの地名について, 音声部品作成時と同じ話者が発声した自然音声を用意した。
2. 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 10 名を選び, ヘッドフォンから音声を提示した.
3. 了解度試験**:** 自然音声と合成音声 100 件ずつを各々ランダムな順序で提示し, 聞こえた地名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.
4. オピニオン評価**:** 自然音声と合成音声の計 200 件を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.

### 3.4 評価試験結果

表 2 に結果を示す. なお, 話者の声質により合成音声の品質が変わる場合を考慮して, 話者を変えて同様の方法で合成実験および評価試験を行った. この結果も ( ) 内に合わせて示す.

合成音声は, 自然音声と同程度の単語了解度が得られている. 地名が耳なれないものであることから, 自然音声の単語了解度すら 100% ではない. このことを考慮すると, 合成音声は, 電話番号案内システムのようにユーザーの入力を反復して確認を求めるような用途では十分利用できると思われる.

表 2: 評価結果

| | 了解度試験 正解率 (%) | オピニオン スコア平均値 |
|---|---|---|
| 自然音声 | 90.4 (97.9) | 4.6 (4.6) |
| 本手法による合成音声 | 91.5 (98.0) | 3.5 (3.4) |

オピニオンスコアが悪いサンプルを調べてみると, 母音-母音, 母音-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, 音節部品選択時に複数候補があった場合はパラメータとしてピッチ周波数などを加え, そのつながりの良いものを選択する, などの方法が考えられる.

また, 件数は少なかったが, 音節部品用に発声した地名が「豊田 1 」(トヨタ+イチ) のように複合語になっていたためにイントネーションが不自然になった場合もあった. このような地名はあらかじめ合成対象から除いておく必要があるが, 接尾辞 (数字,「甲」,「乙」,「丙」など) に着目し, 漢字表記と読みの情報を用いて半自動的に除くことが出来る.

## 4 まとめ, 今後の課題

大量の単語音声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音節種類, 前後音素環境, モーラ位置, モーラ数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.

$4 \sim 6$ モーラの日本の地名 105,000 件を合成対象として録音すべき地名の件数を調査したところ, 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, 特に 5 モーラ語につい

て合成実験および評価試験を行った. 自然音声と比較すると若干音質は落ちるものの了解度の高い合成音声が得られた.

今後の課題として, 以下のことが挙げられる.

- 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. 具体的には前後音素環境に基づいて音響的特徴が似ているものを統合する, 語頭の音節などモーラ数が異なっても韻律的特徴が似ているものを統合する, などの方法が考えられる.
- 音節部品選択時に候補が複数あった場合の選択パラメータを検討すること. 特に有声音の接続部分において, ピッチ周波数や LPC 係数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すことができると思われる.

また将来的には,7 モーラ以上の単語の合成方法の検討や, ガイダンス文に合成音声を埋め込んで違和感が軽減されたかどうか確認する試験が必要である.

本研究の目的のように, 大量の単語音声をガイダンス文に埋め込んで情報提供サービスに利用する領域では, 番号案内の他にも, 住所, 氏名, 製品名等を確認する必要のあるカスタマ · ケア · サービス, 交差点名, 道路名等を用いたカーナビゲーションなど様々な応用が考えられ, 本方式の適用分野は大きいと考えられる.

## 参考文献


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

[2] 東田, 村上, 奥, "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, "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, "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, 須藤, "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).