モーラ数，モーラ位置に基づいた音節波形接続による単語音声合成*

○水澤紀子 (NTT PF 研)　村上仁一 (鳥取大・工)　東田正信 (ATR-I)

1　はじめに

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]「あんないジョーズ」では応答音声の合成に録音編集方式を採用している．　これは，ガイダンスの文章に日本全国の地名，姓名，企業名などの単語を挿入し，例えば「お調べのところは“神奈川県”“横須賀市”でよろしいですか.」(下線部がガイダンス文章，“ ”内が挿入単語) などの応答音声を出力する方式である．

ここで使われる単語音声は日本全国の地名 18 万件，姓名 22 万件という膨大な件数にのぼる．そのため，複数の話者が手分けして発声しており，しかも収録が長期間にわたっている．その結果，各単語の音質は不均質であり，一応答文中に様々な声質が混在して聞きにくいという問題が生じる．

そこで，一部の単語のみを同一話者が発声し，その録音音声を用いて必要な単語音声全てを合成する方法を提案し，日本の 4,5,6 モーラの地名を合成対象として検討を行った [2]．本稿ではさらに，録音件数を減らすために前後音素環境に基づいた音節部品種類の削減も検討した．

2　合成方法

2.1　基本的な合成方法 (合成方法 A)

我々が提案する合成方法は，単語の録音音声から音節単位の波形を切りだし，これらを信号処理せず選択・接続することにより新たな単語音声を合成する方法である [2]．合成時の音節波形選択に用いるパラメータは音節，前音素，後音素，当該音節のモーラ位置，単語のモーラ数の情報のみであり，ピッチ周波数や音韻継続長等の分析・予測は行わないことが提案方法の特徴である．

本合成方法の具体的な手順を簡単に述べる．

まず合成対象の単語全てを音節に分割し，必要な音節部品を調べる．ここで音節部品とは合成の単位となる音節波形を指し，音節・前音素・後音素・当該音節のモーラ位置・単語のモーラ数の情報から成る音節部品ラベルが付与されている．例えば「ヨコハマシ」の「ハ」の音節部品ラベルは ha$(/o/, /m/)_{3,5}$ (5 モーラ語の 3 モーラ目，前音素が/o/，後音素が/m/の ha) のように表記できる．

次に，出現する音節部品ラベル全種類が含まれるような合成対象単語の部分集合を抽出し，それらの単語を録音する．録音した音声をラベリングして音節部品を得る．

録音していない単語を合成する場合は，音節部品ラベルが合成対象単語を構成する音節と一致するような音節部品を選択し，接続する．

2.2　録音件数の削減を目指した合成方法 (合成方法 B)

提案手法においては，モーラ数，モーラ位置により異なる音節部品を準備することから録音件数が多くなりがちであり，用途によっては録音件数すなわち音節部品種類の削減が必要になる．

このような場合，音節内の先行子音によっては前音素の影響を受けにくいためこれを無視する，後音素の調音様式が同じ場合は受ける影響も似ているので同じ環境として扱う，など，音節内の先行子音・後音素環境の子音によって音節部品種類を削減出来ると考えられる．

上記の考察に基づいて，前後音素環境が完全には合わない部品を意図的に選択してサンプルを作成し，検聴を繰り返しながら，了解度を落とさない音節部品種類の削減範囲を検討した．

その結果，以下の削減範囲では単語了解度に影響を与えなかった．1. 音節内の先行子音が破裂音，摩擦音，鼻音，半母音の場合は前音素環境を無視する．2. 撥音以外の全音節について，破裂音，摩擦音の後音素環境は同一視する．

3　録音件数の調査

本論文で合成対象とする語は 4〜6 モーラの日本の地名に限った．これは，以下の 2 点の理由からである．1. 予備調査の結果，これらの地名についてはモーラ数が決まればアクセント型がほぼ決まるのがわかったこと．2. これらの地名だけでも全地名件数 18 万件の 6 割である 105,287 件を占め，大きな録音件数削減効果が期待できること．

---

* Word Synthesis by Concatenating Syllabic Components.
By Noriko MIZUSAWA (NTT Information Sharing Platform Labs), Jin'ichi MURAKAMI (Faculty of Engineering, Tottori University) and Masanobu HIGASHIDA(ATR International)

表 1: 音節部品種類数と録音地名件数

	単語モーラ数 地名の全件数	4 31,222	5 38,430	6 35,635	4~6 合計 105,287
合成方法 A	音節部品種類数	7,606	10,670	12,025	30,301
	録音件数 (割合)	4,560 (14.6%)	6,066 (15.8%)	6,262 (17.6%)	16,888 (16.0%)
合成方法 B	音節部品種類数	2,749	3,821	4,491	11,061
	録音件数 (割合)	1,633 (5.2%)	2,147 (5.6%)	2,373 (6.7%)	6,153 (5.8%)

合成方法 A,B について音節部品の種類数と録音が必要な地名件数を調査した. 結果を表 1 に示す. 表中の録音件数の割合は, 各モーラ数の地名全件数に対する割合を示す.

合成方法 A では録音が必要な件数は約 17,000 件 [2], 合成方法 B では約 6,000 件であった.

4 評価試験

合成音声の評価試験を行った. 100 件の 5 モーラ語の地名について自然音声, 合成方法 A,B による合成音声を用意した. 話者の声質により違いが出る場合を考慮して 2 人の話者の録音音声で合成音声を作成した. (ただし 100 件の地名の内容は話者間で異なる.) これらの音声について,10 名の被験者による単語了解度試験 (聞き取り試験), オピニオン評価 (1 から 5 の 5 段階評価, 1 が最も不自然,5 が最も自然) を行った. 表 2 に結果を示す.

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	
話者	a	b	a	b
自然音声	90.4	97.9	4.6	4.6
合成音声 A	91.5	98.0	3.5	3.4
合成音声 B	90.6	96.1	2.9	2.6

方式 A,B いずれの合成音声も自然音声と同程度の単語了解度が得られており, 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.

5 考察

今回合成したサンプルでは, 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オピニオン評価値を下げたと思われる例が多かった.

音質を重視するような用途に対しては, より多くのサンプルについて実験が必要であるが, 以下のような方法を検討する必要がある.

1. 合成方式 A において, 音節部品選択時に複数候補があった場合にパラメータとしてピッチ周波数や LPC 係数などを加え, そのつながりの良いものを選ぶ.

2. 合成方式 B を用いる際, 録音コストの要求が厳しくない場合には削減する音節部品種類をより狭い範囲に限定し, 半母音, 鼻音など有声音の音素環境が関係する音節部品については録音時から準備する. 録音件数を増やしたくない場合は, 上記 1 の方法のように部品選択時のパラメータの一つとして音素環境を反映する.

6 まとめ, 今後の課題

単語の録音音声から音節種類, 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, 信号処理せずに接続する単語音声合成方法を提案した. さらに, 前後音素環境に基づいた音節部品種類の削減を試みた.

4 ~ 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わかった. また, 音節内の先行子音が破裂音, 摩擦音, 鼻音, 半母音の場合は前音素環境を無視, 撥音以外の全音節について, 破裂音, 摩擦音の後音素環境を同一視することにより音節部品数を削減したところ, 録音件数は 6,000 件で済むことが分かった. 5 モーラ語の合成音声について評価試験を行ったところ, いずれの合成方法でも比較的良好な了解度が得られた.

参考文献

[1] 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)

[2] 水澤, 村上, 東田, “音節波形接続による単語音声合成,” 信学技報,SP99-2,pp9-16(1999.5).

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

[5] 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).