社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Tree Based Clusteringを利用した音節波形接続型音声合成法に関する検討

植村　和久† 　村上　仁一† 　池原　悟†

† 〒 680-8552 鳥取県鳥取市湖山町南 4-101 鳥取大学工学部知能情報工学科
E-mail: †{s032007,murakami,ikehara}@ike.tottori-u.ac.jp

あらまし　音声合成法の手法の 1 つとして，音節波形接続型音声合成法が提案されている．この手法の問題点の 1 つとして，任意の一般名詞を作成する際に大量の録音単語が必要となることが挙げられる．そこで収録されているデータベースに対して Tree Based Clustering を行うことで，理論上は全ての音声を作成出来る．しかし，音声品質が非常に悪い音声もある．本論文では，異なる 2 つのモデルで Tree Based Clustering を行い，音声品質の改善を目指す．音声品質の評価には，オピニオン評価実験および対比較実験を用いる．オピニオン評価実験の結果，標準モデルの音声が 1.7，拡張モデルの音声は 2.5 という値が得られた．また，対比較実験の結果，標準モデルの音声が 20%，拡張モデルの音声が 80%となった．

キーワード　クラスタリング，木に基づく状態共有，音節波形接続型音声合成，音節素片

Study for Word Synthesis by Concatenating Syllabic Components using Tree Based Clustering

Kazuhisa UEMURA†, Jin'ichi MURAKAMI†, and Satoru IKEHARA†

† Faculty of Engineering, Tottori University, Minami 4-101, Koyama-cho, Tottori-shi, Tottori-ken, 680–8552 Japan
E-mail: †{s032007,murakami,ikehara}@ike.tottori-u.ac.jp

Abstract　Word synthesis by concatenating syllabic components method is proposed as a speech synthesis method. As a problem of this technique, large amount of recording words is needed when we make an arbitrary general noun. Then, all speech ideally can be made by doing tree-based clustering for collected database. However, very low quality speech is generated sometimes. In this paper, our aim is to improvement of the speech quality by doing tree-based clustering with two different models. The mean opinion score (MOS) and the ABX test are used for the evaluation of the speech quality. As a result of the MOS, the standard model was obtained 1.7 and the enhanced model was obtained 2.5. Moreover, as a result of an ABX test, the standard model was obtained 20% and the enhanced model was obtained 80%.

Key words　tree-based clustering, concatenating syllabic components, word synthesis, MFCC

1. はじめに

音声合成法の手法の 1 つとして，音節波形接続型音声合成法 [1] が提案されている．この手法は，録音した音声波形の一部（以下，音節素片）を取り出し，接続することによって合成音声を作成する．音声波形に信号処理を加えないため，自然性の高い音声が作成出来るが，音節素片選択時に，全ての条件を一致させなければならない．そのため，この手法の問題点の 1 つとして，任意の一般名詞を作成する際に大量の録音単語が必要となる．

その問題を解決するために，Tree Based Clustering（以下，クラスタリング）を利用した手法が提案されている [2]．この手法は，音節素片選択時の全ての条件を完全に一致させるのではなく，一部の条件をクラスタリングを用いて緩和することにより，理論上全ての合成音声が作成可能となる．しかし，音節素片選択時の条件を緩和したことにより，音声品質が非常に悪い音声も出来る．

そこで本研究では，文献 [3] で作成したモデル (以下，標準モデル) において，音節素片選択時の条件の緩和と音声品質の関係について調査する．

ところで，音声合成において，アクセントは非常に重要な情報である．標準モデルにおいても，音節素片選択時にアクセントに関する条件を緩和すると音声品質が悪くなることがある．そこで本研究では，あらかじめアクセントに関して分類を行い，その後クラスタリングを行うモデル (以下，拡張モデル) を提案する．

2. 音節波形接続型音声合成法

2.1 音節波形接続型音声合成の概要

音節波形接続型音声合成では，表 1 に示す言語的な情報を用いる．まず，データベース中の各音節素片に対して，単語のモーラ数，モーラ位置，アクセント型，各モーラ位置のアクセントの高低，さらに前後の音素環境のラベルを付与する．付与された音節素片の例を表 2 に示す．なお，本研究において，単語のアクセントは NHK 日本語発音アクセント辞典 [4] を利用する．

表 1　音節波形接続型音声合成における言語的な情報

1. 中心の音節
2. 直前の音素 (前音素環境)
3. 直後の音素 (後音素環境)
4. 単語のモーラ数
5. 単語のモーラ位置
6. 単語のアクセント型
7. 単語のアクセントの高低

表 2　音節素片の詳細 (例：a-ka0202001+pau)

前音素環境	a
中心音節	ka
モーラ数	02
モーラ位置	02
アクセント型	00
アクセントの高低	1（高）
後音素環境	pau

次に合成する単語内に含まれる音節素片と，言語的な情報が一致する音節素片を選択する．最後に選択した音節素片を接続して合成音声を作成する．

例えば「乗り物 (no/ri/mo/no)」という合成音声を作成する際の例を図 1 に示す．なお，図中の「‾|_」はアクセントの高低を表す．また，太文字で示している部分は，接続する音節素片である．

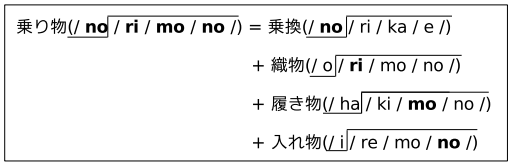

図 1　音節波形接続型音声合成法の例

2.2 合成可能な音声数の問題

過去の研究 [5] より，音節波形接続型音声合成法の有用性が示されたが，この手法では音節素片選択時の条件を全て一致させるため，作成できる音声が少ないという問題がある．例えば音声データベースとして，ATR 単語発話データベース Aset(5,240 単語) を使用した場合，5,240 単語中の 470 単語しか作成できない [5].

3. クラスタリング

3.1 クラスタリングの概要

クラスタリング [2] は，音声の決定木に基づいて，音響的特徴の類似した音節素片の状態集合を作成する．クラスタリングの質問は，データベース中に存在しない単語の音節素片を，データベース中の音節素片が含まれているクラスタに分類出来る．したがってクラスタリングを利用することで，データベース中に存在しない単語を作成できる．本研究では，表 1 で示した言語的な情報に対して，HTK [6] の HHED を用いることでクラスタリングを行う．作成した質問の例を以下に示す．また，クラスタリングの動作例を図 2 に示す．

- モーラ数は 4 であるか？
- アクセントは高いか？
- 前音素環境は撥音であるか？
- 後音素環境は鼻音であるか？

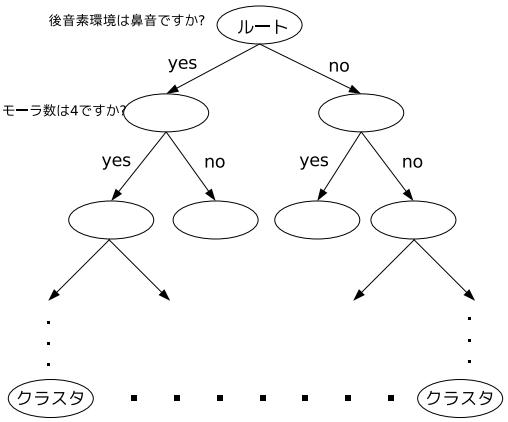

図 2　クラスタリングの動作例

図 2 のようにクラスタリングを行うと，複数のクラスタが作成出来る．各クラスタには，音響的特徴の類似した音節素片が含まれている．本研究では，クラスタ内の音節素片を同一とすることで，音節波形接続型音声合成法の条件の緩和とみなす．表 3 にクラスタの例を示す．

表 3　クラスタの例

クラスタ名	N_1	pe_3
クラスタ内の音節素片	a-N0202001+pau	N-pe0403001+k
	a-N0303001+pau	pau-pe0201011+N
	o-N0303001+pau	pau-pe0301000+e
	u-N0202001+pau	q-pe0403001+i
	u-N0303001+pau	

3.2 クラスタリングを利用した音節波形接続型音声合成法

本研究では，まずデータベース中の音節素片に対してクラスタリングを行う．本来クラスタリングは，データベース中に存在しない単語を合成する際に利用されるが，そのような音声を作成した場合，自然音声との比較が出来ない．そこで本研究では，データベース中の単語を作成することで，クラスタリングを利用した合成音声の音声品質を評価する．次に合成したい単語の音節素片を含むクラスタの中から，音節素片を 1 つ抽出する．その後，音節波形接続型音声合成法で合成音声を作成する．本研究における音声合成の流れを図 3 に示す．また，本研究で作成した合成音声の例を図 4 に示す．

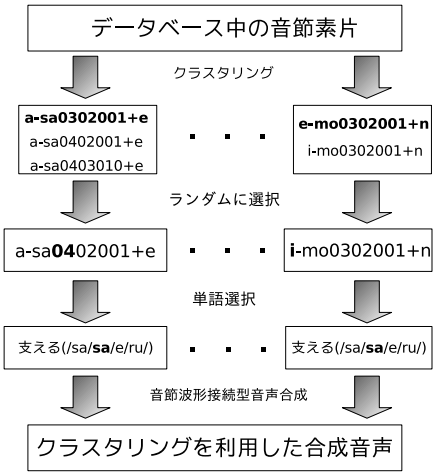

図 3　クラスタリングを利用した波形接続型音声合成の流れ図

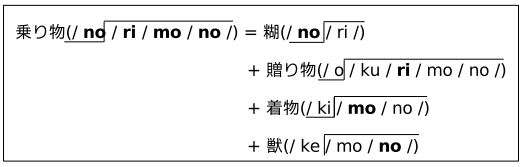

図 4　クラスタリングを利用した音節波形接続型音声合成法の例

図 4 は,「乗り物 (no/ri/mo/no)」という合成音声を作成する際の例である．2.1 節の図 1 と比較すると，合成に使用した音節素片全てが，モーラ数およびモーラ位置の異なる音節素片である．

3.3 合成音声の評価方法

本研究では，合成音声の評価方法としてオピニオン評価実験および対比較実験を行う．各実験に使用する音声は表 4 に示す３または４モーラの音声とする．また，特に断りがない場合，以後の実験の評価者は，音声研究に関わったことのある１名とする．

表 4　作成する音声

名称	説明
自然音声	ATR 単語発話データベース ASET に含まれる音声
オリジナル合成	クラスタリングを利用しない音節波形接続型音声合成
クラスタリング合成	クラスタリングを利用した音節波形接続型音声合成

4. 標準モデルを用いた実験

4.1 実験の目的

本章の実験では，データベース中の音節素片に対してクラスタリングを行うモデルにおいて，クラスタ内の音節素片をランダムに選択した場合に，得られる音声品質を調査する．本研究では，以後このようなモデルを標準モデルと呼ぶ．標準モデルの流れを図 5 に示す．

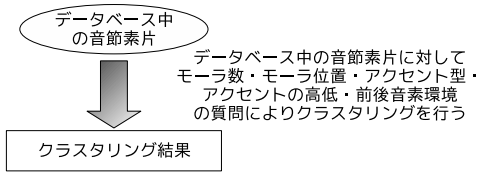

図 5　標準モデル

4.2 実 験 条 件

本章の実験では，HTK [6] の HHED を使用してクラスタリングを行う．クラスタリングの質問は，単語のモーラ数，モーラ位置，アクセント型，アクセントの高低，前後音素環境に関する質問とする．また，クラスタリングの状態数は，閾値によって決定しており，本研究では状態数が 500 程度となるようにクラスタリングを行う．クラスタリングの結果，状態数は 538 となった．学習データは，ATR 単語発話データベース Aset の 5,240 単語／話者の女性話者 1 名を用いる．なお，本研究で作成する合成音声は，自然音声との比較のため，学習データ内の単語とする．その他の実験条件を表 5 に示す．

本章の実験では，まず女性話者 1 名に関して，表 4 の音声を各 100 単語ずつ作成する．次に，作成した合成音声の評価方法として，オピニオン評価実験を行う．

表 5　実 験 条 件

基本周波数	16kHz
分析窓	Hamming 窓
分析窓長	25ms
フレーム周期	10ms
音響モデル	3 ループ 4 状態・半連続分布型
stream 数	3
共分散行列	Diagonal-covariance
MFCC 特徴ベクトル	12 次 MFCC+Δ 12 次 MFCC+ 対数パワー+Δ 対数パワー (計 26 次)

4.3 実験結果

オピニオン評価実験の結果を表 6 に示す．表 6 より，クラスタ内の音節素片をランダムに選択した場合には，自然音声やオリジナル合成と比べて，音声品質は悪いことが分かった．

表 6　オピニオン評価実験結果 (総評価音節数：300)

音声の種類	オピニオンスコア
自然音声	4.9
オリジナル合成	4.0
クラスタリング合成	1.7

5. 標準モデルにおけるクラスタ内の選択

5.1 実験の目的

4. 章では，標準モデルにおいて，クラスタ内の音節素片をランダムに選択した結果，クラスタリング合成の音声品質が悪いと分かった．しかし，クラスタ内の音節素片は複数あるため，さらにこれらの音節素片を選択することで音声品質が向上する可能性がある．

そこで本章の実験では，標準モデルにおいて，クラスタ内で最良と考えられる音節素片を選択した場合に，得られるクラスタリング合成の音声品質を求める．

5.2 実験条件

基本的な実験条件は 4.2 節と同様である．予備実験の結果，モーラ数およびモーラ位置の異なる音節素片を選択した場合に，良い音声品質が得られた．しかし，アクセント型およびアクセントの高低の異なる音節素片を選択した場合には，音声品質が悪くなることが多かった．よって以下の番号順に，選択を行う．

（1） モーラ数およびモーラ位置の異なる音節素片
（2） 前後音素環境の異なる音節素片
（3） ランダムに選択

本章の実験では，まず女性話者 1 名に関して，表 4 の音声を各 50 単語ずつ作成する．クラスタリング合成において，11 単語はモーラ数およびモーラ位置の異なる音節素片のみで作成し，残り 39 単語はモーラ数およびモーラ位置が異なり，かつ，前後音素環境の異なる音節素片で作成する．次に，作成した合成音声の評価方法として，オピニオン評価実験を行う．ただし本章の実験に限り，オピニオン評価実験の評価者は，音声研究に関わったことのない 5 名とする．

5.3 実験結果

オピニオン評価実験結果を表 7 に示す．

表 7　オピニオン評価実験結果 (総評価音節数：750)

音声の種類	オピニオンスコア
自然音声	4.5
オリジナル合成	3.9
クラスタリング合成	3.6

表 7 より，クラスタリング合成は，オピニオンスコアで 3.6 という値が得られた．また，自然音声と合成音声の比較において，オピニオンスコアに多少差が見られるが，オリジナル合成とクラスタリング合成にはあまり差が見られない．以上よりクラスタ内で最良と考えられる音節素片を選択した場合，クラスタリング合成は音声品質の高い合成音声であると言える．

6. アクセント型の影響

6.1 実験の目的

表 6 の結果から，クラスタ内の音節素片をランダムに選択した場合に，クラスタリング合成の音声品質は悪いことが分かった．しかし，表 7 の結果から，クラスタ内で最良と考えられる音節素片を選択した場合には，非常に良い音声品質が得られた．これは同一クラスタ内に，音声品質を落とさない音節素片と，音声品質を落とす音節素片が混在している事を示している．音声品質を落とす音節素片にも様々な種類があると考えられるが，予備実験からアクセントの高低が異なる音節素片を選択した場合に，音声品質の低下を確認している．しかし，アクセント型に関しては未調査である．そこでアクセント型（6 種類）が音声品質に与える影響を調査する．

6.2 実験条件

本章の実験では，合成音声の作成に関して，音節波形接続型音声合成法を利用しない．表 1 において，自然音声の音節素片と，アクセント型のみが異なる音節素片を入れ換えることにより，合成音声を作成する．また，アクセント型の影響を明確にするために，入れ換える音節素片は 1 つだけとする．本章の実験では，まず女性話者 2 名に関して，各組合せ 2 単語ずつ計 34 単語作成する．次に，作成した合成音声の評価方法として，オピニオン評価実験を行う．図 6 に作成した音声の例を示す．

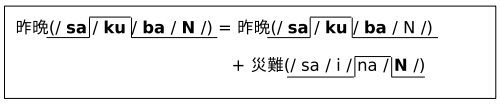

図 6　アクセント型が異なる音節素片の入れ換え例

図 6 は,「昨晩 (sa/ku/ba/N)」という合成音声を作成する際の例である．昨晩の “N” の音節素片が “a-N0404020+pau” であるのに対し，災難の “N” の音節素片は “a-N0404030+pau” である．

6.3 実験結果

オピニオン評価実験の結果を表 8 に示す．

表 8　オピニオン評価実験結果 (総評価音節数：68)

音声の種類	オピニオンスコア
作成した全ての音声 (68 音声)	4.4
中心音節 N を入れ換えた音声 (8 音声)	2.6

表 8 より，中心音節が N である場合，アクセント型を考慮しないと音声品質が悪くなる．しかし，その他の中心音節の場合，音声品質に影響がない．

7. 拡張モデルを用いた実験

7.1 実験の目的

4. 章,5. 章,6. 章で，標準モデルには，同一クラスタ内に品質

を落とす音節素片が含まれているという問題があった．そこで音声品質を落とす音節素片の条件として考えられるアクセント型の調査を行った結果，中心音節が N の音節素片を入れ換えた場合に，音声品質が悪くなることが分かった．

本章の実験ではこの問題を解決するために，あらかじめ全ての音節素片をアクセント型およびアクセントの高低で分類し，その後クラスタリングを行う拡張モデルを提案する．拡張モデルでは，図 7 のように事前に全ての音節素片をアクセント型およびアクセントの高低で分類する．本研究ではアクセント型を 6 種類，アクセントの高低を 2 種類用いているため，全 12 種類の集合となった．最後に各集合に対してクラスタリングを行う．

本来ならば，中心音節が N の音節素片のみをアクセント型およびアクセントの高低で分類し，その他の音節素片は，アクセントの高低で分類を行うべきである．しかしクラスタリングを容易に行うために，本章の実験では，全ての音節素片をアクセント型およびアクセントの高低で分類する．

本章の実験では，この拡張モデルにおいて，クラスタ内の音節素片をランダムに選択した場合に，得られる音声品質を調査する．また，標準モデルと拡張モデルの音声品質の比較も同時に行う．

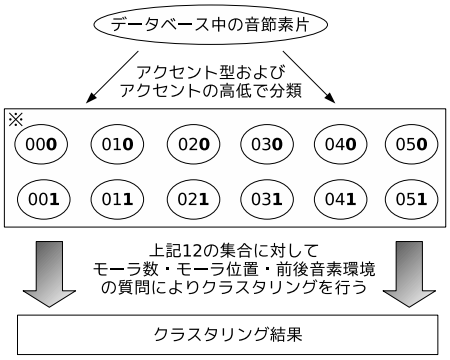

図 7　拡張モデル

図 7 中※で記述してある 3 桁の数字は，最初の 2 桁がアクセント型を示し，最後の 1 桁がアクセントの高低を示している．例えば “000” の場合，アクセント型が 00 型，アクセントの高低が 0 である．拡張モデルのクラスタの例を表 9 に示す．

表 9　拡張モデルにおけるクラスタの例

クラスタ名	hi_1	e_1
クラスタ内の音節素片	pau-hi0201011+f	a-e0504030+r
	pau-hi0201011+t	a-e0504030+s
	pau-hi0201011+y	e-e0604030+t
	pau-hi0301011+g	pau-e0601030+r
	pau-hi0301011+y	a-e0404030+pau

7.2　実験条件

基本的な実験条件は，4.2 節と同様である．ただしクラスタリングの質問は，4.2 節からアクセント型およびアクセントの高低を除いた質問とする．なお，アクセント型およびアクセントの高低で事前に分類を行ったため，標準モデルと同一の停止基準でクラスタリングを行った結果，最終的な状態数は 2,472 となった．また，本章の実験では，女性話者 1 名に関して，表 4 の音声を各 100 単語ずつ作成する．作成した合成音声の評価として，オピニオン評価実験および対比較実験を行う．

7.3　実験結果

オピニオン評価実験の結果を表 10 に，対比較実験の結果を表 11 に示す．

表 10　オピニオン評価実験結果 (総評価音節数：300)

音声の種類	オピニオンスコア
自然音声	4.9
オリジナル合成	4.0
クラスタリング合成	2.5

表 11　対比較実験結果 (総評価音節数:200)

	標準モデル	拡張モデル
対比較実験	20%	80%

表 10 より，拡張モデルのクラスタリング合成は，自然音声やオリジナル合成より低いオピニオンスコアとなった．しかし表 6 より，拡張モデルのクラスタリング合成は，標準モデルのクラスタリング合成よりも良い音声品質であることが分かった．

また，表 11 より，拡張モデルは 80%と高い値を得た．以上より拡張モデルの有効性を確認した．

8.　後音素環境に関する実験

8.1　実験の目的

表 6，表 10 より，クラスタ内の音節素片をランダムに選択した場合，両モデルともにクラスタリング合成の音声品質は悪いことが分かった．また，表 11 より，拡張モデルは標準モデルよりも，音声品質の点において優れていた．しかし拡張モデルにおいても，音声品質を落とす音節素片が存在している．この音声品質を落とす音節素片の条件として考えられるのが，前後音素環境である．そこで子音と母音の調査を行える後音素環境が，音声品質に与える影響を調査する．

8.2　実験条件

基本的な実験条件は，6.2 節と同様である．ただし自然音声の音節素片と入れ換えるのは，後音素環境のみが異なる音節素片とする．本章の実験では，女性話者 1 名に関して，200 単語作成する．また，作成した合成音声の評価方法として，オピニオン評価実験を行う．本章の実験で作成した音声の例を図 8 に示す．

図 8 は，「野蛮 (/ya/ba/N)」，「打診 (/da/shi/N/)」，「覆う (/o/o/u/)」という合成音声を作成する際の例である．野蛮の “ya” の音節素片が “pau-ya0301000+b” であるのに対し，野球の “ya” の音節素片は “pau-ya0301000+ky” となっている．また，打診の “da” の音節素片が “pau-da0301000+sh” であるのに対し，大事の “da” の音節素片は “pau-da0301000+i” となっている．同様に，覆うの “o” の音節素片が “pau-o0301000+o” であるのに対し，夫の “o” の音節素片は “pau-da0301000+q” であり，後音素環境だけが異なっている．

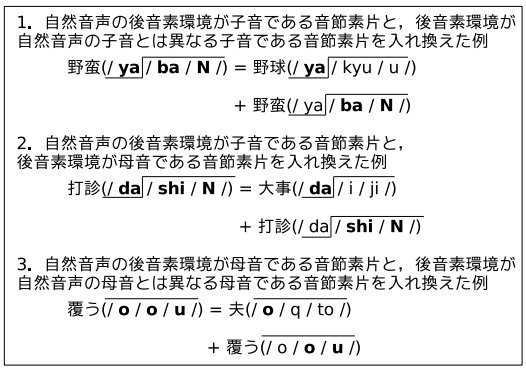

図 8　後音素環境が異なる音節素片の入れ換え例

8.3 実験結果

オピニオン評価実験の結果を表 12 に示す．なお，表 12 は，自然音声の音節素片と，後音素環境のみが異なる音節素片の入れ換えに関する結果である．

表 12　オピニオン評価実験結果 (総評価音節数：200)

音声の種類	オピニオンスコア
1. 子音と子音 (144 単語)	4.7
2. 子音と母音 (51 単語)	3.6
3. 母音と母音 (5 単語)	3

表 12 より，自然音声の後音素環境が子音である音節素片と，後音素環境が自然音声の子音とは異なる子音である音節素片を入れ換えた結果，音声品質は良かった．また，自然音声の後音素環境が母音である音節素片と，後音素環境が自然音声の母音とは異なる母音である音節素片を入れ換えた結果，音声品質が悪くなった．同様に，自然音声の後音素環境が母音である音節素片と，後音素環境が子音である音節素片を入れ換えた場合，もしくはその逆の場合に，音声品質が悪くなることが分かった．

8.4 考　　察

表 12 から，自然音声の後音素環境が母音である音節素片を，後音素環境の異なる音節素片と入れ換えた場合，もしくは自然音声の後音素環境が子音である音節素片を，後音素環境が母音である音節素片と入れ換えた場合に，音声品質が悪くなることが分かった．これを考慮して，あらかじめ前後音素環境を分類するモデルを作成すれば，音声品質はさらに向上すると考えている．しかし，アクセント型およびアクセントの高低に加えて前後音素環境までも事前に分類してしまうと，音響的に分類するクラスタリングを利用する価値がほとんど見出せなくなる．以上のことから，クラスタリングによる音響的な分類を利用するよりも，データベース中の音節素片全てを言語的に分類するモデルが良い可能性がある．

9. おわりに

本研究では，クラスタリングを利用した音節波形接続型音声合成法に関する検討を行った．

異なる 2 つのモデルを構築し，各モデルに対して聴覚実験を行った．オピニオン評価実験の結果，標準モデルの音声が 1.7，拡張モデルの音声は 2.5 という値が得られた．また，対比較実験の結果，標準モデルの音声が 20%，拡張モデルの音声が 80%となった．以上より拡張モデルを用いたクラスタリング合成は，標準モデルよりも音声品質が向上していることが分かった．しかし，自然音声やオリジナル合成と比較すると音声品質の悪い音声であったため，今後は前後音素環境を事前に分類するといった改善策が必要と考えている．

また，標準モデルでは，クラスタの状態数は 538 であったのに対し，拡張モデルでは 2,472 であった．これはクラスタリングの停止基準として，同一の値を使用したためである．クラスタの状態数とクラスタ内の音節素片の平均個数は反比例するため，同程度の状態数で，今後実験を行う必要があると考えている．

謝辞　本論文を執筆するにあたり，参考にさせて頂いた論文，聴覚実験に協力してくださった鳥取大学工学部知能情報工学科池原研究室の田村 元秀氏，神田悦司氏，黒住 亜紀子氏，前田浩佑氏，水田 理夫氏に深く感謝いたします．

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] S. J. Young, J. J. Odell, and P. C. Woodland. “Tree-based state trying for high accuracy acoustic modelling. Proc”, ICASSP, pp.307-312(1994)
[3] 山形 亮, 堀田 波星夫, 村上 仁一, 池原 悟. “木に基づく状態共有を利用した波形接続型音声合成法の検討”, 1-Q-21 pp.375-376(2006)
[4] NHK 放送文化研究所, “NHK 日本語発音アクセント辞典新版”, NHK 出版 (1998)
[5] 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句への応用”, 音響全体, 2-Q-18, pp.1-409, 410(2003)
[6] “HTK Ver3.2 reference manual”, Cambridge University(2002)