概要

音声合成の手法の１つとして波形接続型音声合成法が提案されている. この手法は録音音声から音節単位で波形素片を取り出し, 信号処理をせずに接続することで, 自然性の高い音声が合成できる．しかし，任意の一般名詞を作成しようとするには大量の録音単語が必要である．そこで収録されている DB に対して木に基づくクラスタリング (以後クラスタリング) を行い，音響パラメータが似た音節素片をグループ化する．クラスタリングにより得られた情報を利用して波形接続型音声合成を行うことで作成可能な単語数が飛躍的に増加する．しかしクラスタリングを行う際の最適な条件については明らかにされていない.

そこで本研究ではクラスタリングに関して, 特徴パラメータとして FBANK 及び MFCC を利用し，特徴パラメータの違いを調査するために対比較実験を行った. 同時にクラスタリングで条件緩和を施すのに適した言語的な情報の調査のために対比較実験を行った. また全体に関して, 音声を合成する事で音質の劣化が懸念される為，オピニオン評価実験により音質の評価を行った.

その結果, オピニオン評価実験では, クラスタリングを利用した合成音声で 3.6, 波形接続型音声合成で 3.9, 自然音声で 4.5 というオピニオンスコアが得られた. クラスタリングを利用した合成音声は, 自然音声には少し及ばないものの, 波形接続型合成音声とあまり差がなく, 品質の高い合成音声が作成出来たことが分かった.

対比較実験において, 特徴パラメータに FBANK を用いた合成音声が 47%,MFCC を用いた合成音声が 53%となった. 両パラメータにほとんど差はないが, 若干 MFCC を用いた合成音声の方が良い結果が得られたため, クラスタリングの特徴パラメータには MFCC を用いた方が良いことが分かった.

またクラスタリングで条件緩和を行う言語的な情報としてモーラ情報を用いた合成音声が 81%, モーラ情報と前後音素環境を用いた合成音声が 19%となった. この結果よりクラスタリングで条件緩和を行う言語的な情報にはモーラ情報を用いた方が良い事が分かった.

目 次

図 目 次

表 目 次

1　はじめに

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等のように，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，システムやサービスに必要となる音声を，サービスが利用される時間・場所等を示す固有名詞のように比較的短い単語・文節 (以下，可変部) と，それ以外の比較的長い文節・文 (以下，固定部) に区別する．この可変部と固定部を別々に録音し，組み合わせを変化させることで出力音声を作成する．

例えばカーナビゲーションシステムにおいて，ユーザの入力に対し,「次の目的地は○○です」という音声が出力される．出力音声において，○○の部分が可変部，それ以外の部分が固定部に相当する．ユーザが目的地に「鳥取駅」を入力した場合，出力される音声は「次の目的地は鳥取駅です」となる．

例のように録音編集方式では可変部と固定部の組み合わせにより大量の出力音声が作成出来る．しかし録音編集方式の問題点の１つとして，可変部と固定部を接続する際に違和感を生じさせないために，同一話者の音声である必要がある．可変部と固定部を分離して録音することにより，必要となるすべての音声を録音する場合に比べて話者に対する負担は若干軽減されるが，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となる．さらに，録音環境の違いにより発話速度 F_0 周波数にばらつきが出るため，安定した品質の音声を得ることは非常に困難となる．

そこで，固定部と可変部に必要となる音声をすべて音声合成によって作成する方法が考えられる．例えば，音素や音節，CV，VCV を単位とした規則音声合成がある．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，基本的には，音声の特徴をパラメータとして抽出し，変形することによって合成音声を作成する．PSOLA 方式による音声合成については，現在も多くの研究がなされている．また，最近では HMM を用いて直接音声を合成する研究も行われている．しかし，いずれの場合においても，直接人の声を録音した音声に比べ，安定して高い品質を得ることが難しい．

一方，録音した音声波形の一部 (以下，音声素片) を用いて別の音声を合成する方法があり，一般に，波形接続型音声合成と呼ばれる．波形接続型音声合成は，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．

また波形接続型音声合成は，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，その一方で韻律の扱いが問題となる．しかし本研究で対象とする「普通名詞」を合成する場合において，アクセント型を考慮することで，明瞭性が高く，自然性の高い合成音声の作成が可能である事が示されている [1].

波形接続型音声合成には，更に音節素片選択時の条件が厳しく，作成出来る音声の数が少ないという問題がある．過去の研究では，木に基づくクラスタリングを利用する事で素片選択の条件を緩和し，作成できる音声を増加する事が可能であると報告されている [3]．しかし木に基づくクラスタリングを行う際の最適なパラメータについては明らかにされていない．そこで本研究では木に基づくクラスタリングを行う際に，特徴パラメータとしてFBANK及びMFCCを利用し，特徴パラメータの違いを調査する．同時にクラスタリングに適した言語的な情報の調査も行う．またクラスタリングにより音質の劣化が懸念される為，作成した音声の音質の評価を行う．

結果として，オピニオン評価実験において，クラスタリングを利用した合成音声が3.6という高い値を得た．これよりクラスタリングを利用した合成音声は品質の高い合成音声であることが分かった．

対比較実験において，特徴パラメータによる差を調査した結果，FBANKと比べほとんど差は無かったが，若干MFCCを用いた合成音声が良いという結果が得られた．また緩和条件による違いを調査した結果，モーラ情報の緩和を用いた合成音声が良いという結果が得られた．

以上より本研究におけるクラスタリングの最適なパラメータは，特徴パラメータとしてMFCC，緩和条件としてFBANKであることが分かった．

以降，2章，3章，4章で波形接続型音声合成，HTKを用いた音節モデルの作成，木に基づくクラスタリングの説明を行う．また5章で木に基づくクラスタリングを利用した波形接続型音声合成について説明し，6章で実験条件を述べる．7章で実験結果を報告し，その考察を8章で行う．

2　波形接続型音声合成

2.1　波形接続型音声合成の概説

本研究で用いる波形接続型音声合成では，始めに，録音された音声に対して表 1 の言語的な情報が一致する音節素片を選択する．次に言語的な情報が一致した音節候補を持つ音声の中から１つをランダムに選択する．その音声のラベリングデータを参照し，音節の開始時間と終了時間に基づいて波形データを切り出す．最後に選択された波形データを接続して合成音声を作成する．

表 1: 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

2.2 波形接続型音声合成の問題点

2.2.1 韻律の問題

波形接続型音声合成は，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，その一方で韻律の扱いが問題となる．しかし本研究で対象とする「普通名詞」を合成する場合において，アクセント型を考慮することで，明瞭性が高く，自然性の高い合成音声の作成が可能である事が示されている [1].

2.2.2 作成出来る音声数の問題

過去の研究 [1] より，波形接続型音声合成の有用性が示されたが波形接続型音声合成では音節素片選択の条件が厳しいために作成できる音声が少ないという問題がある，音声データベースとして，ATR 単語発話データベース Aset(5240 単語) を使用した場合，5240 単語中の 470 単語しか作成できない [1].

そこで任意の音声を合成可能にするために，収録されている録音音声に対して木に基づくクラスタリング [2] を行い，音響パラメータが似た音節素片をグループ化する．そのグループ化された情報を利用して波形接続型音声合成を行うことで作成可能な単語数が飛躍的に増加する [3].

2.3 波形接続型音声合成に関する捕捉

波形接続型音声合成では，接続部の違和感の発生が自然性に大きく影響する．本研究では，波形の接続位置を音素境界とする．さらに，接続部における 2 素片間の波形の位相を考慮し，接続部の振幅の差がゼロに近付くように調整を行う．具体的には，あらかじめラベル付けされた素片終了時間をもとに，振幅が負から正に変わる部分を，波形が短くなる方向 (開始時間は進む方向，終了時間は戻る方向) に探し，抽出する位置を修正する．

2.4　波形接続型音声合成の例

本研究で作成した合成音声の例を図 1 に示す．左側の音声が本研究で作成した合成音声であり，右側の音声において，太字部分で示されている音節素片の波形データを切り出し，接続する事で合成音声の作成を行う．

なお 図中の「_|￣」はアクセントの高低を表す．

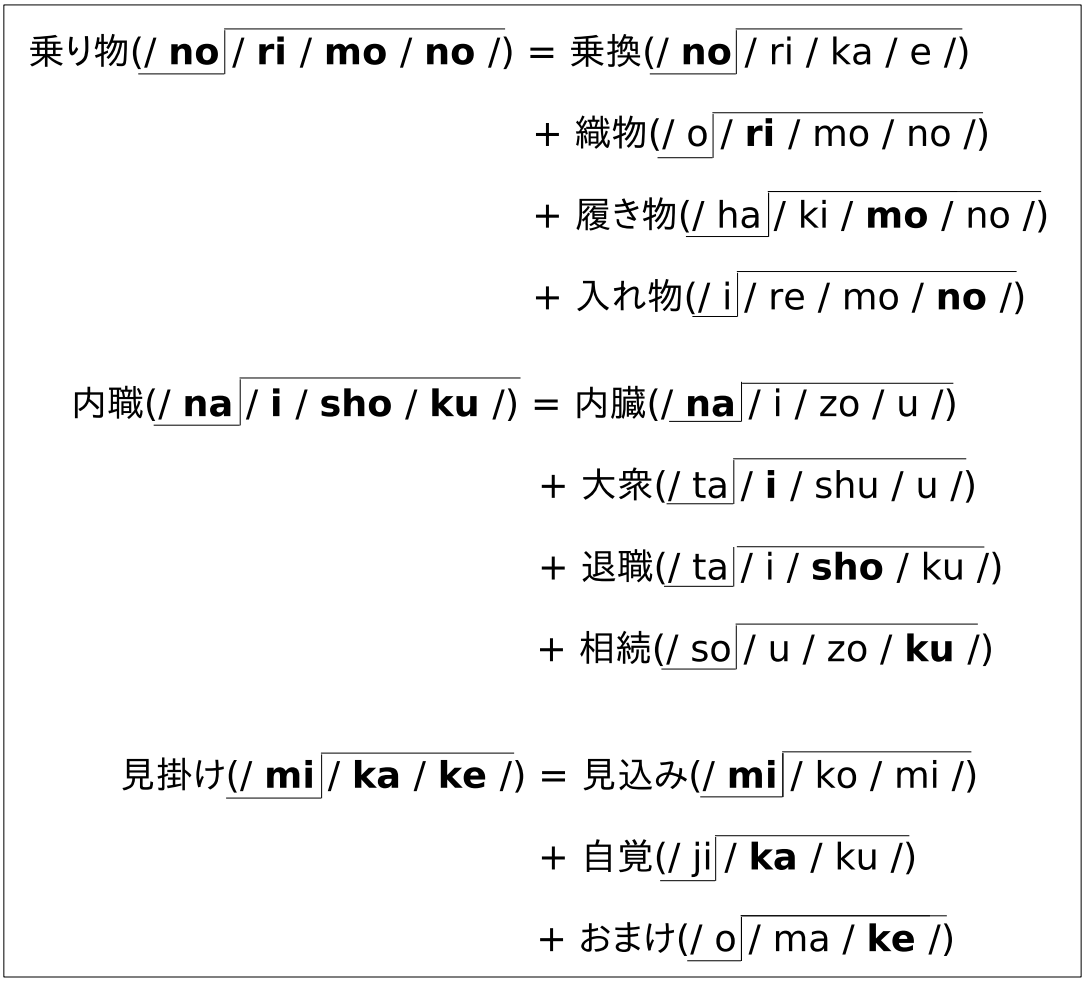

図 1: 波形接続型音声合成の例

図 1 の波形データをそれぞれ図 2〜図 7 に示す．

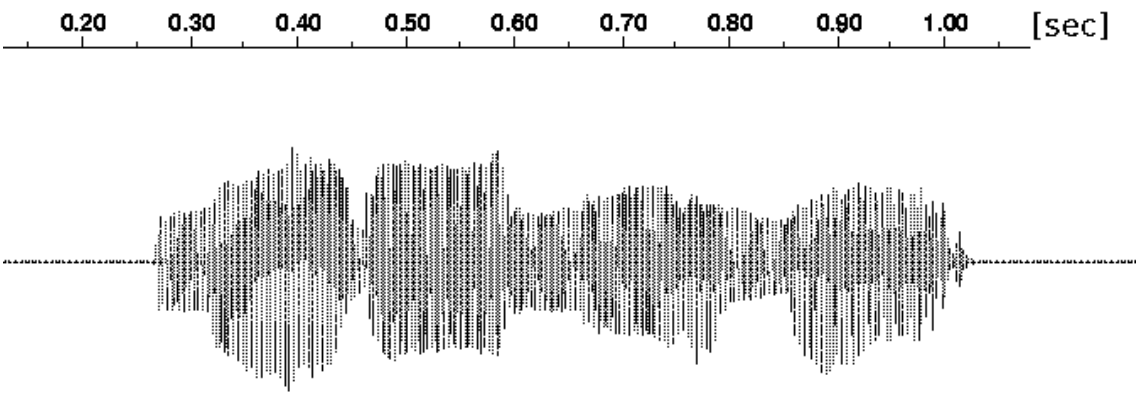

図 2: 「乗り物」の波形データ (自然音声)

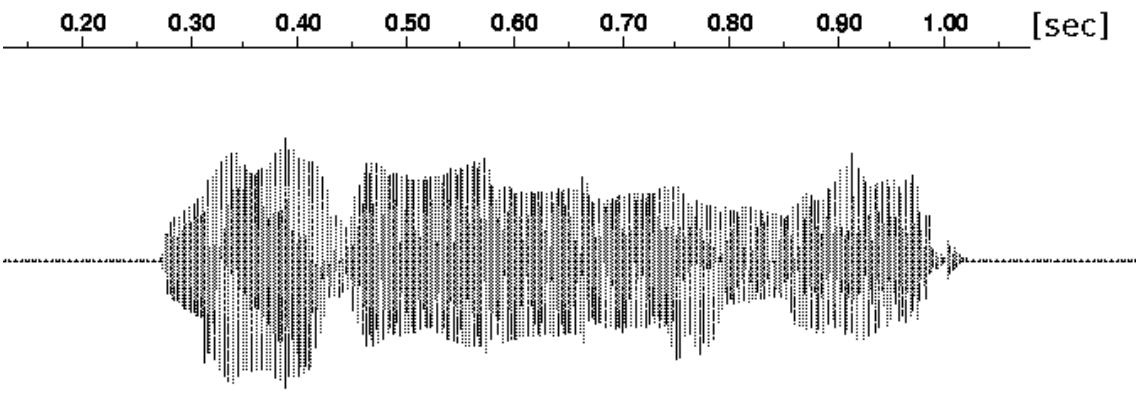

図 3: 「乗り物」の波形データ (波形接続型音声合成)

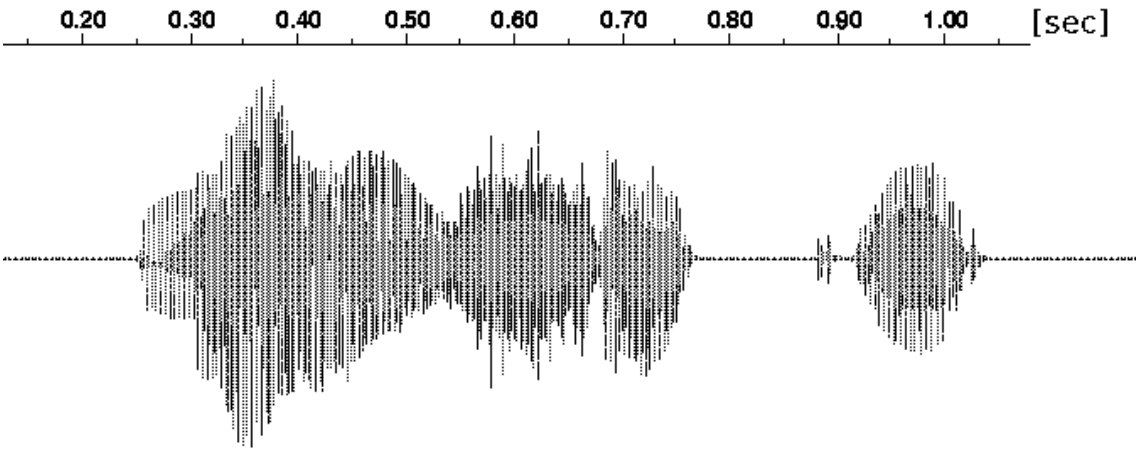

図 4: 「内職」の波形データ (自然音声)

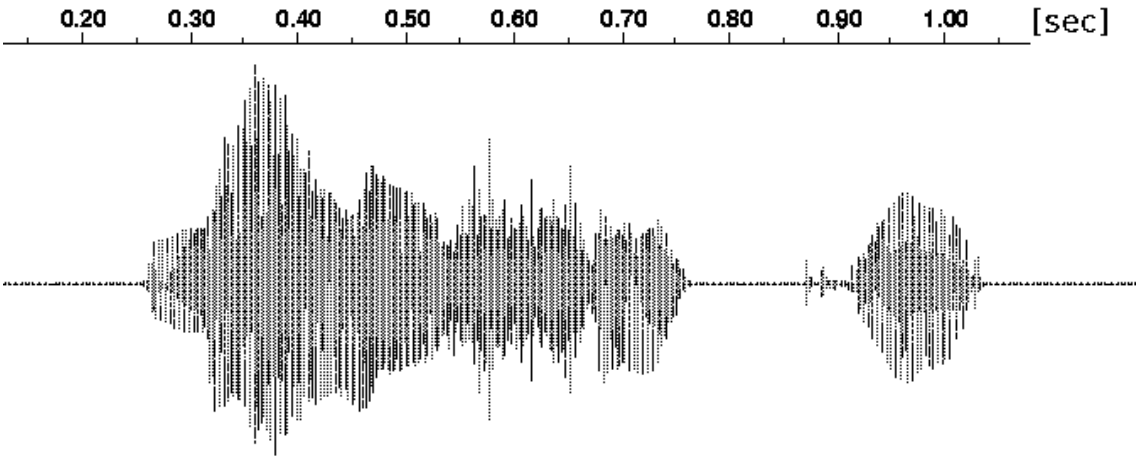

図 5: 「内職」の波形データ (波形接続型音声合成)

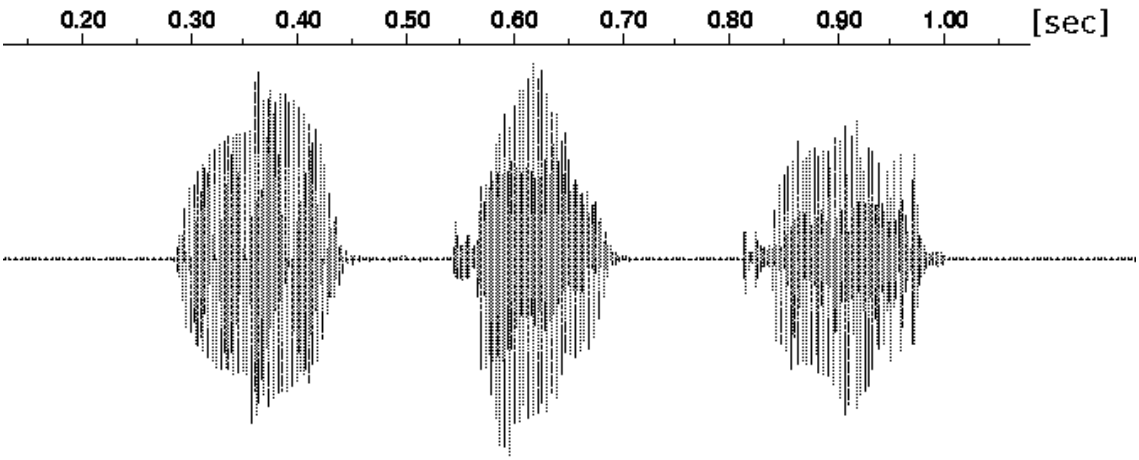

図 6: 「見掛け」の波形データ (自然音声)

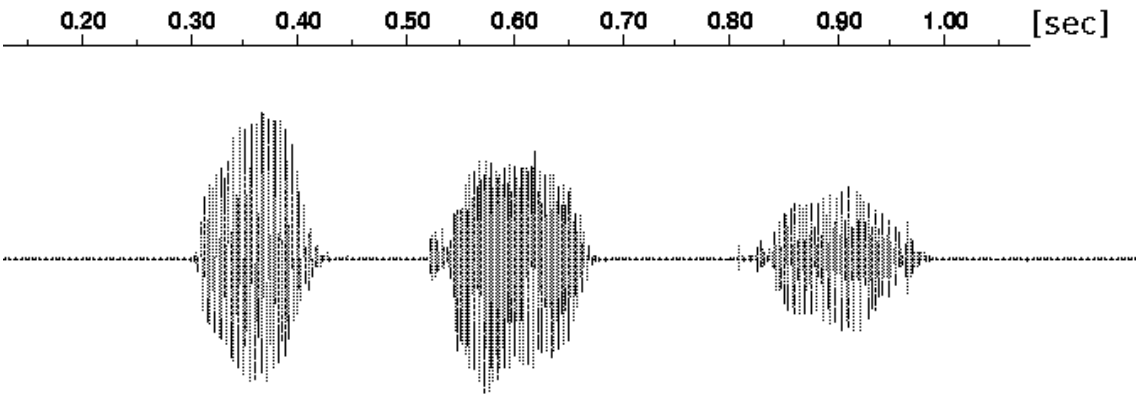

図 7: 「見掛け」の波形データ (波形接続型音声合成)

3　HTK を用いた音節モデル作成

本研究では音節モデル作成に HTK[4] を使用する．HTK とは，HMM を利用する音声ツールである．

3.1　HMM

HMM とは Hidden Markov Model(隠れマルコフモデル) の事であり，出力シンボルによって一意に状態遷移先が定まらないという非決定状態オートマトンとして定義されている．HMM には，ある状態から全ての状態に遷移出来る Ergodic モデルや，左から右へと状態遷移する left to right モデル等がある．

図 8 に HMM(left to right) の例を示す．

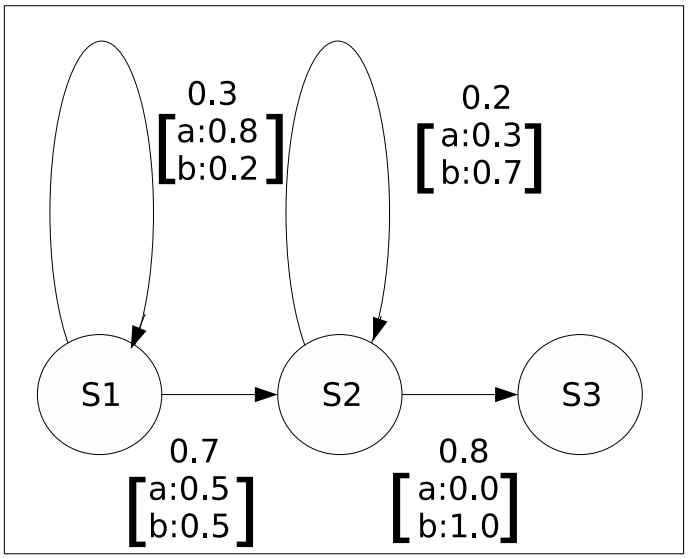

図 8: HMM(left to right) の例

図 8 の例では，$S1$，$S2$，$S3$ の 3 つの状態で構成されており，出力は有限個のシンボル a と b の 2 種類である．初期状態は $S1$，最終状態は $S3$ である．

例として出力シンボルが aab であった場合，状態遷移系は $[S1 \rightarrow S1 \rightarrow S2 \rightarrow S3]$，$[S1 \rightarrow S2 \rightarrow S2 \rightarrow S3]$，の 2 通りがある．

3.2 連続 HMM

出現するスペクトルパターンを連続値として表す分布モデルである. 出現確率を表す方法としては単一ガウス分布や混合ガウス分布が用いられる. パラメータの自由度を減らすために無相関ガウス分布を用いることが多い.

出現確率 $b_ij(o_t)$ が混合ガウス分布に従う場合は,

- M_{ij}...状態 i から状態 j の遷移における混合数
- C_{ijm}...状態 i から状態 j の遷移における混合数のときの重み
- $\mathcal{N}(;\mu,\Sigma)$...平均ベクトルμ, 共分散行列Σをもつ混合ガウス分布

とすると, 以下のように計算される.

$$b_{ij}(o_t) = \sum_{m=1}^{M_{ij}} C_{ijm}\mathcal{N}(o_t;\mu_{ijm},\Sigma_{ijm}) \tag{1}$$

$\mathcal{N}(;\mu,\Sigma)$ は

- n...観測行列の次元数
- $(O-\mu)^t$...$(O-\mu)$ の天地行列
- $\mid \Sigma \mid$ ...Σの固有値
- Σ^{-1}...Σの逆行列

とすると, 以下の式で表現される.

$$\mathcal{N}(O;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n \mid \Sigma \mid}} \exp(-\frac{1}{2}(O-\mu)^t\Sigma^{-1}(O-\mu)) \tag{2}$$

3.3 離散 HMM

出現するスペクトルパターンを有限個のシンボルの組合せで表す分布モデルである. スペクトルパターンのベクトル量子化によって, 符号ベクトルを生成し, 各符号ベクトルの出現確率の組合せによって出現確率を表す.

3.4 半連続型 HMM

半連続型 HMM は離散 HMM の出力確率値に分布を与えた HMM である. 半連続分布は, 離散 HMM の符号張の 1 つずつのベクトルに分布を与えたもので, 連続密度符号張 (continuous density codebook) とも呼ばれている. そこでは, 出力確率を連続密度符号張の分布の混合で表す. 符号張のなかの分布数を M とすると,

$$b_{ij}(x) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(x) \tag{3}$$

と混合正規分布で表す. ただし,

$$\sum_{m=1}^{M} \lambda_{ijm} = 1 \tag{4}$$

である. 平均値と共分散はすべての出力確率で同一であり, 遷移 $s_i \to s_j$ での分布の重み λ_{ijm} のみが変わる [5].

4　木に基づくクラスタリング

4.1　木に基づくクラスタリングの概説

木に基づくクラスタリング [2] は音響的特徴が類似した triphone HMM の状態集合に対して，音声の決定木に基づいて行う．本研究では始めに，2.1 で示した波形接続型音声合成において考慮する言語的な情報に対して，質問を用いることで木に基づくクラスタリングを行う．クラスタリングにより状態分けされた音節情報を基に，その音節情報を持つ単語をランダムに 1 つ選択する．最後に，選択された単語の波形データを切り出し，接続することで音声合成を行う．

なお具体的なクラスタリングの手順については以下に示す．またクラスタリングの動作例を図 9 に示す．

1. 初期状態として，全ての状態をルートノードにまとめる．
2. 1 に対して，*log* 尤度が最大になるように親ノードの状態を分割する質問を見付け，状態を分ける．
3. 全ての状態を共有したときの *log* 尤度と，状態を分割したときの *log* 尤度を比較する．
4. 3 において，*log* 尤度の増加が閾値を下回れば決定木の構築を終え，下回らなければ状態の分割を繰り返す．
5. 最終的な状態数が 500 になるように分割を行う．

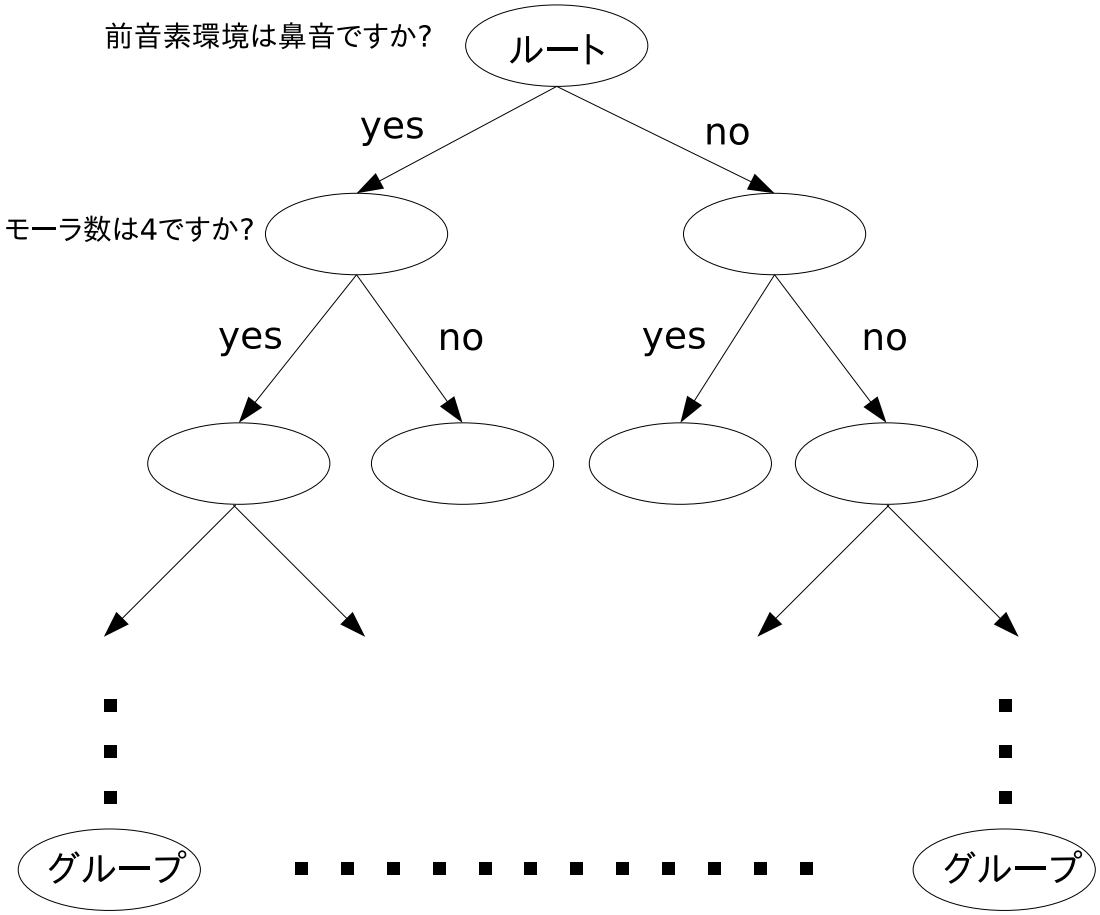

図 9: 木に基づくクラスタリングの動作例

4.2 木に基づくクラスタリングの特徴パラメータ

4.2.1 FBANK

FBANK は音声波形をフーリエ変換して得られたパワースペクトラムの周波数を使用する. パワースペクトラムを少ない次数で効率的に表現するために, メル分割されたフィルタバンクの対数パワーを使用する. またパワーケプストラムの全域に, 人間の聴覚の特性にあわせて低周波部分は細かく，高周波部分は大まかに調べるためメルスケールに沿って等間隔に配置された三角関数のフィルタをかける. この三角関数の個数がフィルタバンクのチャンネルのチャンネル数 (特徴パラメータにおける次数) を表している. 周波数メル分割の式は

$$Mel(f) = 2592 \log_{10}(1 + \frac{f}{700}) \tag{5}$$

となる. そして, フィルタバンクの出力に log 対数をとったものを FBANK として使用する.

4.2.2 MFCC

ケプストラムパラメータには, 多様な計算方法がある. その中には MFCC(Mel-Frequency Cepstrum Coefficient) がある.MFCC の計算では, スペクトラル分析は周波数軸上に三角窓を配置し, フィルタバンク分析により行う. すなわち, 窓の幅に対応する周波数帯域の信号のパワーを, 単一スペクトルチャンネルの振幅スペクトルの重みづけ和で求める. さらに, 窓はメル周波数軸上に等間隔に配置される.

最終的に, フィルタバンク分析により得られた帯域におけるパワーを離散コサイン変換することで,MFCC が求められる.

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos(\frac{\pi i}{N}(j - 0.5)) \tag{6}$$

N はフィルタバンクチャンネの数を表し,m_j は対数フィルタバンクの振幅を表す.

4.3 木に基づくクラスタリングに用いた質問の例

木に基づくクラスタリングに用いた質問の例を表 2 に示す.

表 2: クラスタリングの質問の例

番号	質問名	共有化する情報	説明
1	R_g2	*+g,*+gy	後音素環境は g，gy のどちらか?
2	L_Voiced-Stop6	b-*,d-*,g-*,gy-*	前音素環境は b，d，g，gy のどれか?
3	L_hh	h-*	前音素環境は h か，それ以外か?
4	AC15	*-*010+*,*-*011+*	アクセント型は 01 で，アクセントの高低は 0,1 のどちらか?

4.4 質問によるグループ分け

質問によってグループ分けされた例を 4.4.1，4.4.2 に示す.
またグループ分けされた音節に関する捕捉説明を以下に示す.

- ST_N*_227 のように ST で始まるのは状態名を示している.
- a-N0202001+pau のようにコンマ区切りでアルファベットと数字で構成されているのは，音節情報を示している．詳細を表 3 に示す.
- 同じ状態に記述されている音節は，お互いに代用しても良い事を示している.

表 3: 音節情報の詳細 (例：a-N0202001+pau)

前音素	中心音節	モーラ数	モーラ位置	アクセント型	アクセントの高低	後音素
a	N	02	02	00	1	pau

4.4.1 FBANK を用いた際のグループ分けの例

1. ST_N*_227

   a-N0202001+pau,a-N0303001+pau

2. ST_gya*_21

   pau-gya0201000+k,pau-gya0401000+k

3. ST_i*_226

   a-i0202010+pau,a-i0303010+pau,a-i0303020+pau,a-i0404010+pau,a-i0404020+pau,a-i0404030+pau,a-i0505030+pau,a-i0505040+pau,a-i0606050+pau,o-i0202010+pau,o-i0303010+pau,o-i0303020+pau,o-i0404030+pau,o-i0505040+pau

ST_N*_227，ST_gya*_21 にはそれぞれ 2 個，ST_i*_226 には 14 個もの音節情報が共有されている．このように状態により，共有された個数はばらばらとなっている．

4.4.2 MFCC を用いた際のグループ分けの例

1. ST_N*_23

   a-N0202001+pau,a-N0303001+pau,a-N0404001+pau

2. ST_gya*_21

   pau-gya0201000+k,pau-gya0401000+k

3. ST_i*_25

   a-i0202010+pau,a-i0303010+pau,a-i0404010+pau,a-i0404030+pau,a-i0505030+pau

ST_N*_23 には 3 個，ST_gya*_21 には 2 個，ST_i*_25 には 5 個の音節情報が共有されている．

4.4.1 のグループ分けと比較すると，FBANK の ST_N*_227 と MFCC の ST_N*_23 では，a-N0202001+pau と a-N0303001+pau が同じ状態にグループ分けされているが，MFCC では更に a-N0404001+pau も一緒の状態とされている．FBANK の ST_gya*_21 と MFCC の ST_gya*_21 では，状態名も共有されている音節情報も全く同じとなっている．FBANK の ST_i*_226 と MFCC の ST_i*_25 では，一部の音節情報のみが同じ状態に記述されているが，FBANK の方がより多くの音節情報が 1 つにまとめられている．

5　木に基づくクラスタリングを利用した波形接続型音声合成

5.1　本研究における木に基づくクラスタリング

本研究では木に基づくクラスタリングを利用するために，まず HTK[4] を用いて，前後音素環境，モーラ情報，アクセント情報を考慮した音節モデルの作成を行う．次にその音節モデルに対して，木に基づくクラスタリングにより音響的特徴が類似した音節情報に状態分けを行う．しかし音響的特徴による分類だけでは，選択出来る音節が多数存在するので，本研究では更に状態分けされた複数の音節情報に対して，言語的な情報の緩和条件を用いて音節情報を選択する．最後に選択された音節情報の中からランダムに１つを選択し，それらを接続することで合成音声を作成する．

なお本研究におけるクラスタリングの” 緩和条件” とは，指定された言語情報以外が一致する条件を指す．” 緩和条件” の例を図 10 に示す．

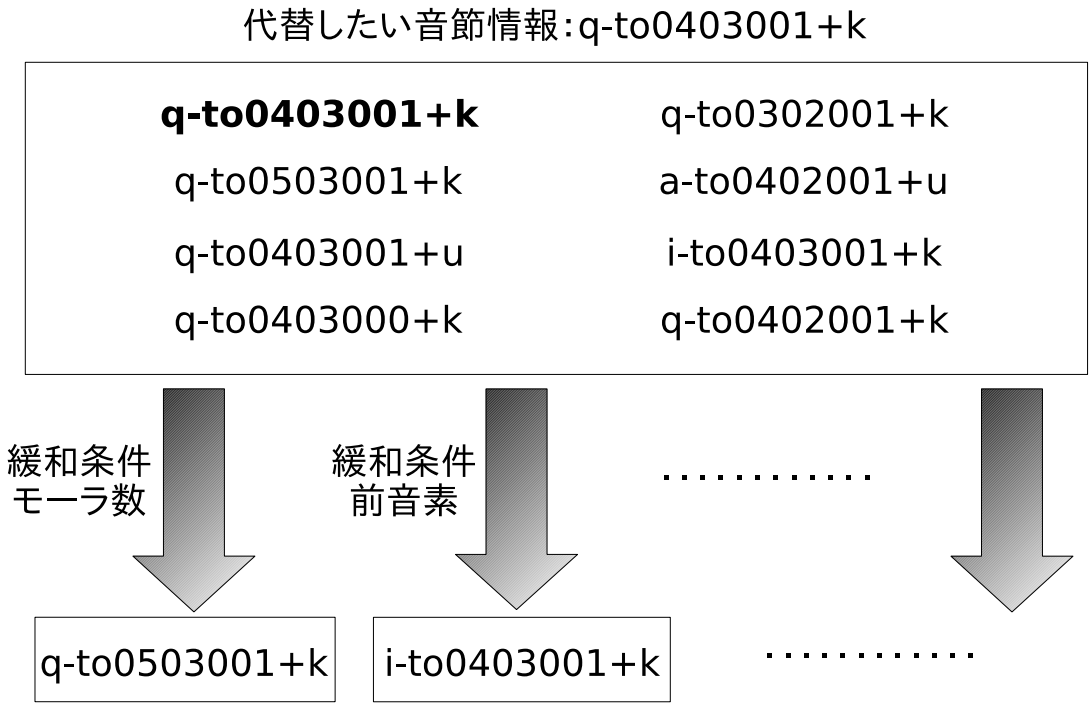

図 10: 緩和条件の例

図 10 には，音節情報 q-to0403001+k と代替可能な音節情報として，7 つの音節情報が示されている．このグループに対して，緩和条件としてモーラ数の緩和を行うと q-to0503001+k が，緩和条件として前音素の緩和を行うと i-to0403001+k がそれぞれ抽出出来る．

5.2　木に基づく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

本研究における木に基づく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を図 11 に示す.

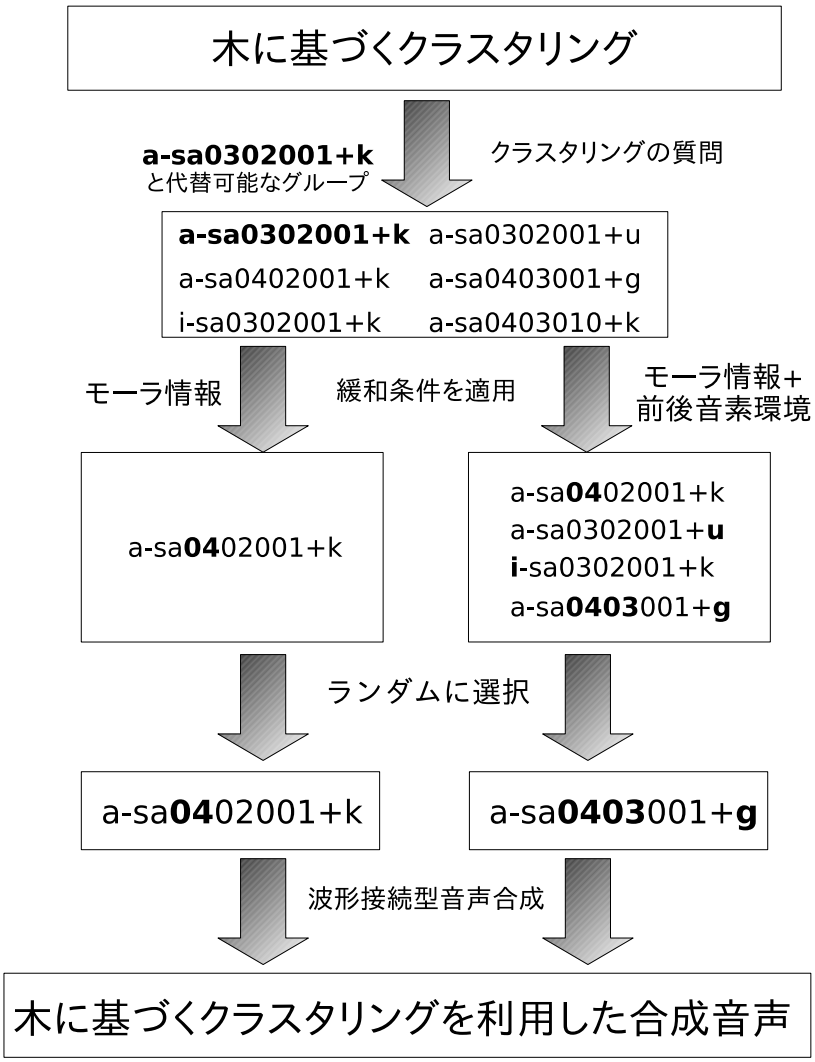

図 11: 木に基づくクラスタリングを利用した波形接続型音声合成の流れ図

まず木に基づくクラスタリングの質問により，音響的特徴が類似した音節素片がグループ化される．図 11 では a-sa0302001+k と代替可能なグループとして 6 つの音節情報が 1 つのグループに共有されている．次にこのグループに対して緩和条件としてモーラ情報を用いると a-sa0402001+k が，モーラ情報+前後音素環境を用いると図のように 4 つの音節情報が抽出される．最後にこれらの音節情報において，ランダムに 1 つずつ選択し，選択された音節情報を持つ単語に対して，波形接続型音声合成を行うことで，木に基づくクラスタリングを利用した合成音声の作成を行う.

5.3 木に基づくクラスタリングを利用した波形接続型音声合成の例

5.3.1 特徴パラメータ：FBANK

特徴パラメータとしてFBANKを使用した際の合成音声の例を図12に示す.

図 12: 木に基づくクラスタリング (FBANK) を利用した波形接続型音声合成の例

図12において，左側の音声が本研究で作成した合成音声であり，右側の音声において，太字部分で示されている音節素片の波形データを切り出し，接続する事で合成音声の作成を行う．なお図中の「_|¯」はアクセントの高低を表す.

また図12の波形データを，図14，17，20に示す.

5.3.2 特徴パラメータ:MFCC

特徴パラメータとして MFCC を使用した際の合成音声の例を図 13 に示す.

図 13: 木に基づくクラスタリング (MFCC) を利用した波形接続型音声合成の例

図 13 において，左側の音声が本研究で作成した合成音声であり，右側の音声において，太字部分で示されている音節素片の波形データを切り出し，接続する事で合成音声の作成を行う．なお図中の「_|¯」はアクセントの高低を表す.

また図 13 の波形データを，図 15，18，21 に示す．更に図 12 と図 13 に関する詳細な情報を，それぞれ図 16，図 19，図 22 に示す．なお図中の網掛けされた部分は，クラスタリングにより緩和された言語的な情報を示している.

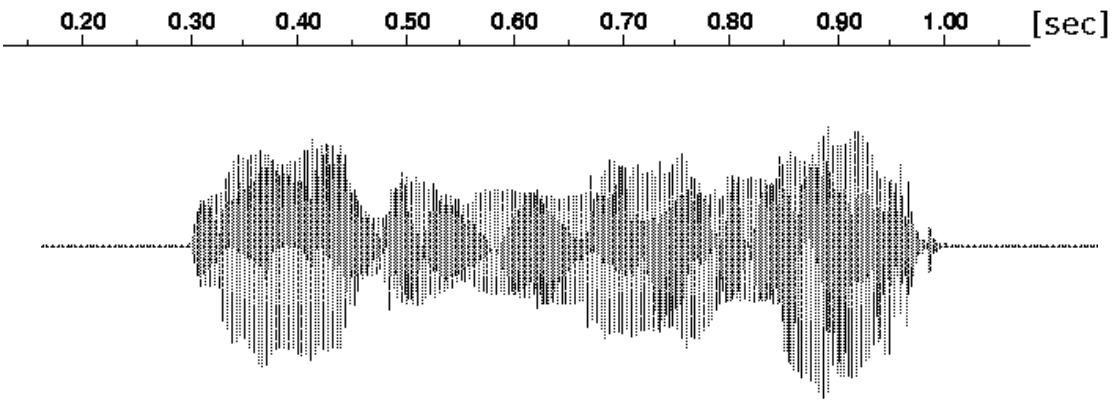

図 14: 「乗り物」の波形データ (FBANK)

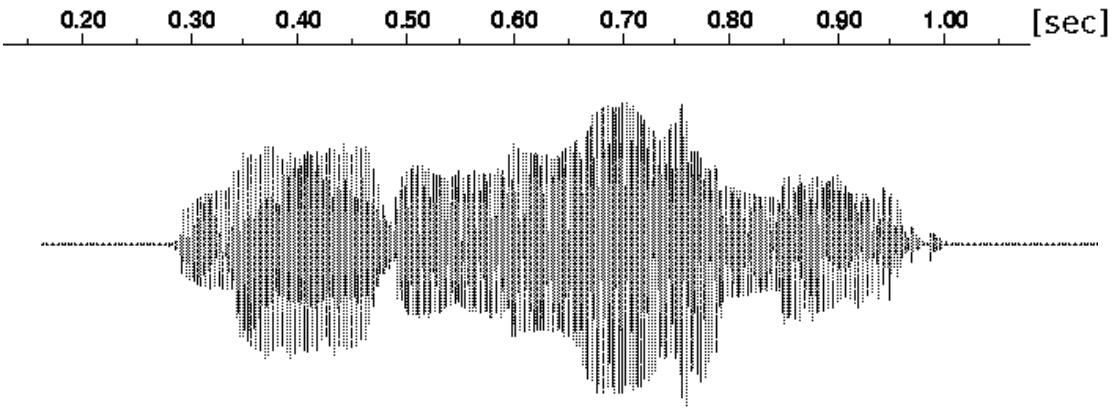

図 15: 「乗り物」の波形データ (MFCC)

音声「乗り物」についての詳細を図 16 に示す.

・自然音声:乗り物(/no/ri/mo/no/)の詳細

前音素	中心音節	モーラ数	モーラ位置	アクセント位置	中心音節の高低	後音素
pau	no	04	01	00	0	r
o	ri	04	02	00	1	m
i	mo	04	03	00	1	n
o	no	04	04	00	1	pau

FBANKでの組み合わせ

1. 乗り場(/ **no** / ri / ba /)
2.ガソリン(/ ga / so / **ri** / N /)
3. 贈り物(/ o / ku / ri / **mo** / no /)
4.物(/ mo / **no** /)

MFCCでの組み合わせ

1.糊(/ **no** / ri /)
2.贈り物(/ o / ku / **ri** / mo / no /)
3.着物(/ ki / **mo** / no /)
4. 獣(/ ke / mo / **no** /)

・クラスタリング音声(FBANK):乗り物(/no/ri/mo/no/)の詳細

前音素	中心音節	モーラ数	モーラ位置	アクセント位置	中心音節の高低	後音素
pau	no	03	01	00	0	r
o	ri	04	03	00	1	N
i	mo	05	04	00	1	n
o	no	02	02	00	1	pau

・クラスタリング音声(MFCC):乗り物(/no/ri/mo/no/)の詳細

前音素	中心音節	モーラ数	モーラ位置	アクセント位置	中心音節の高低	後音素
pau	no	02	01	00	0	r
u	ri	05	03	00	1	m
i	mo	03	02	00	1	n
o	no	03	03	00	1	pau

図 16: 音声「乗り物」についての詳細

図において網掛け部分は自然音声と比べて，言語情報が緩和されている部分を示している．またクラスタリング音声の詳細における音節情報は，特徴パラメータごとの組み合わせの太字で表されている部分の音節情報を示している．

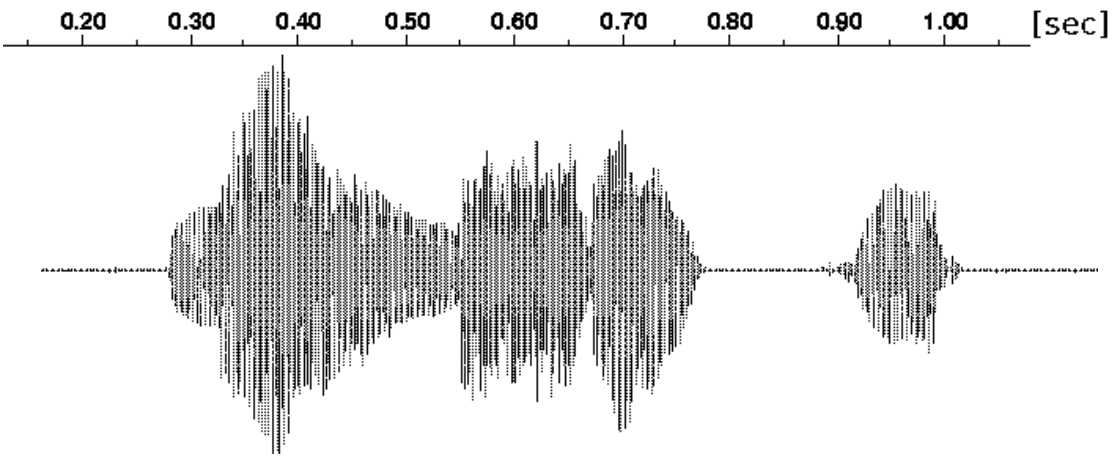

図 17: 「内職」の波形データ (FBANK)

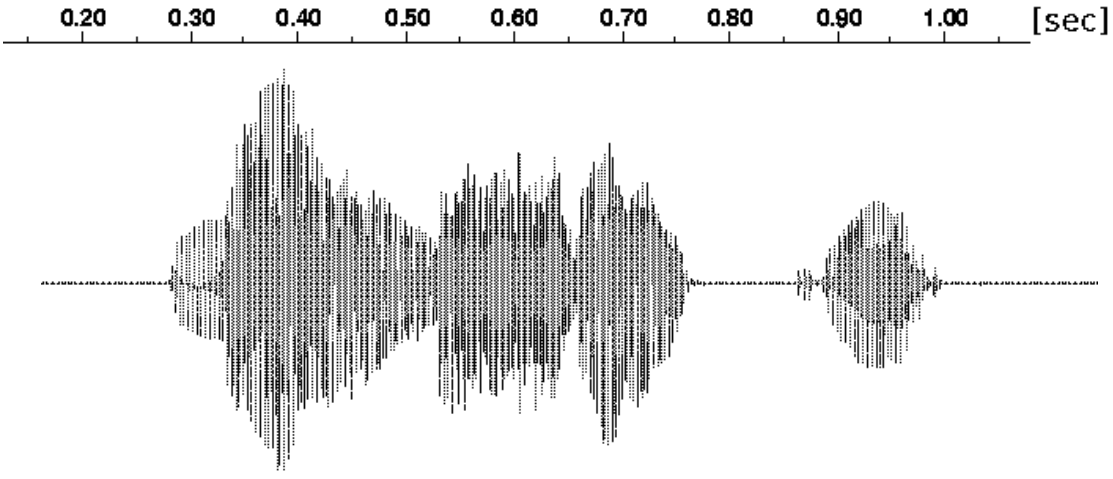

図 18: 「内職」の波形データ (MFCC)

音声「内職」についての詳細を図 19 に示す.

・自然音声:内職(/na/i/sho/ku/)の詳細

前音素	中心音節	モーラ数	モーラ位置	アクセント位置	中心音節の高低	後音素
pau	na	04	01	00	0	i
a	i	04	02	00	1	sh
i	sho	04	03	00	1	k
o	ku	04	04	00	1	pau

FBANKでの組み合わせ

1.内緒(/ **na** / i / sho /)
2.開始(/ ka / **i** / shi /)
3.辞職(/ ji / **sho** / ku /)
4.置く(/ o / **ku** /)

MFCCでの組み合わせ

1.内緒(/ **na** / i / sho /)
2.会社(/ ka / **i** / sha /)
3.辞職(/ ji / **sho** / ku /)
4.汚職(/ o / sho / **ku** /)

・クラスタリング音声(FBANK):内職(/na/i/sho/ku/)の詳細

前音素	中心音節	モーラ数	モーラ位置	アクセント位置	中心音節の高低	後音素
pau	na	03	01	00	0	i
a	i	03	02	00	1	sh
i	sho	03	02	00	1	k
o	ku	02	02	00	1	pau

・クラスタリング音声(MFCC):内職(/na/i/sho/ku/)の詳細

前音素	中心音節	モーラ数	モーラ位置	アクセント位置	中心音節の高低	後音素
pau	na	03	01	00	0	i
a	i	03	02	00	1	sh
i	sho	03	02	00	1	k
o	ku	03	03	00	1	pau

図 19: 音声「内職」についての詳細

図において網掛け部分は自然音声と比べて，言語情報が緩和されている部分を示している．またクラスタリング音声の詳細における音節情報は，特徴パラメータごとの組み合わせの太字で表されている部分の音節情報を示している．

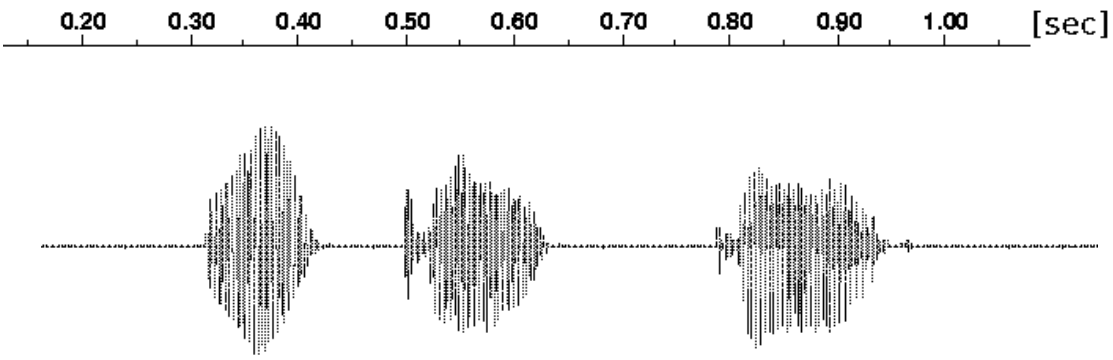

図 20: 「見掛け」の波形データ (FBANK)

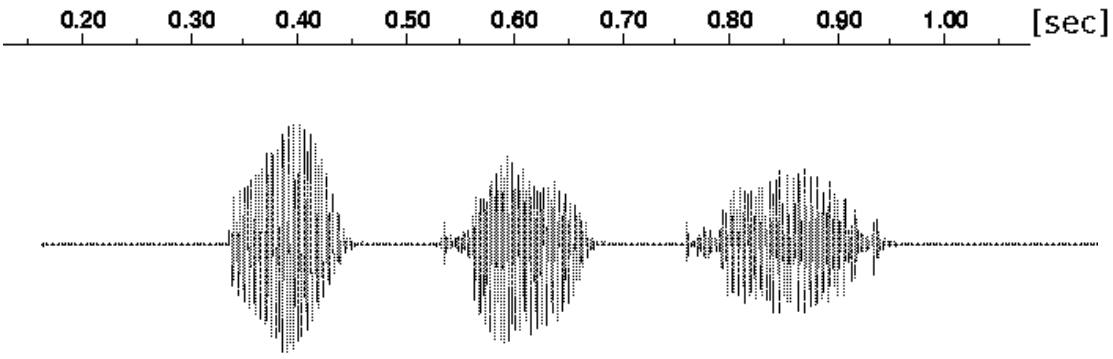

図 21: 「見掛け」の波形データ (MFCC)

音声「見掛け」についての詳細を図 22 に示す.

・自然音声:見掛け(/mi/ka/ke/)の詳細

前音素	中心音節	モーラ数	モーラ位置	アクセント位置	中心音節の高低	後音素
pau	mi	03	01	00	0	k
i	ka	03	02	00	1	k
a	ke	03	03	00	1	pau

FBANKでの組み合わせ

1.見掛ける(/ **mi** / ka / ke / ru /)
2.明確(/ me / i / **ka** / ku /)
3.酒(/ sa / **ke** /)

MFCCでの組み合わせ

1.見掛ける(/ **mi** / ka / ke / ru /)
2.対格(/ ta / i / **ka** / ku /)
3.きっかけ(/ ki / q / ka / **ke** /)

・クラスタリング音声(FBANK):見掛け(/ni/ka/ke/)の詳細

前音素	中心音節	モーラ数	モーラ位置	アクセント位置	中心音節の高低	後音素
pau	mi	04	01	00	0	k
i	ka	04	03	00	1	k
a	ke	02	02	00	1	pau

・クラスタリング音声(MFCC):見掛け(/mi/ka/ke/)の詳細

前音素	中心音節	モーラ数	モーラ位置	アクセント位置	中心音節の高低	後音素
pau	mi	04	01	00	0	k
i	ka	04	03	00	1	k
a	ke	04	04	00	1	pau

図 22: 音声「見掛け」についての詳細

図において網掛け部分は自然音声と比べて，言語情報が緩和されている部分を示している．またクラスタリング音声の詳細における音節情報は，特徴パラメータごとの組み合わせの太字で表されている部分の音節情報を示している．

5.4　木に基づくクラスタリングを利用した波形接続型音声合成の問題点

木に基づくクラスタリングを利用した合成音声は状態分けされた音節素片を基にして音声を作成するため, 1 つの単語を合成する際に合成出来る組み合わせが多数存在する. 組み合わせによって，作成出来る合成音声の音質に大きな差が生じる事から，本研究では 5 つの音声に対して各 20 通りの組み合わせを考慮し，それぞれの音質の調査を行う. またその際に，クラスタリングにより共有された音節情報の中でも，モーラ情報及び前後音素環境に対して，音節情報の緩和を行う．組み合わせの例としてモーラ情報のみを緩和した際の例を 5.4.1 に，モーラ情報に加えて前後音素環境も緩和した際の例を 5.4.2 に示す.

5.4.1 モーラ情報を緩和した際の組み合わせ

モーラ情報のみを緩和した際の合成音声の組み合わせの例を図 23 に示す.

図 23: 合成音声の組み合わせの例 (緩和条件：モーラ情報)

図 23 において，左側の音声が本研究で作成した合成音声であり，右側の音声において，太字部分で示されている音節素片の波形データを切り出し，接続する事で合成音声の作成を行う．なお図中の「_|￣」はアクセントの高低を表す.

5.4.2 モーラ情報及び前後音素環境を緩和した際の組み合わせ

モーラ情報に加えて前後音素環境を緩和した際の合成音声の組み合わせの例を図 24 に示す.

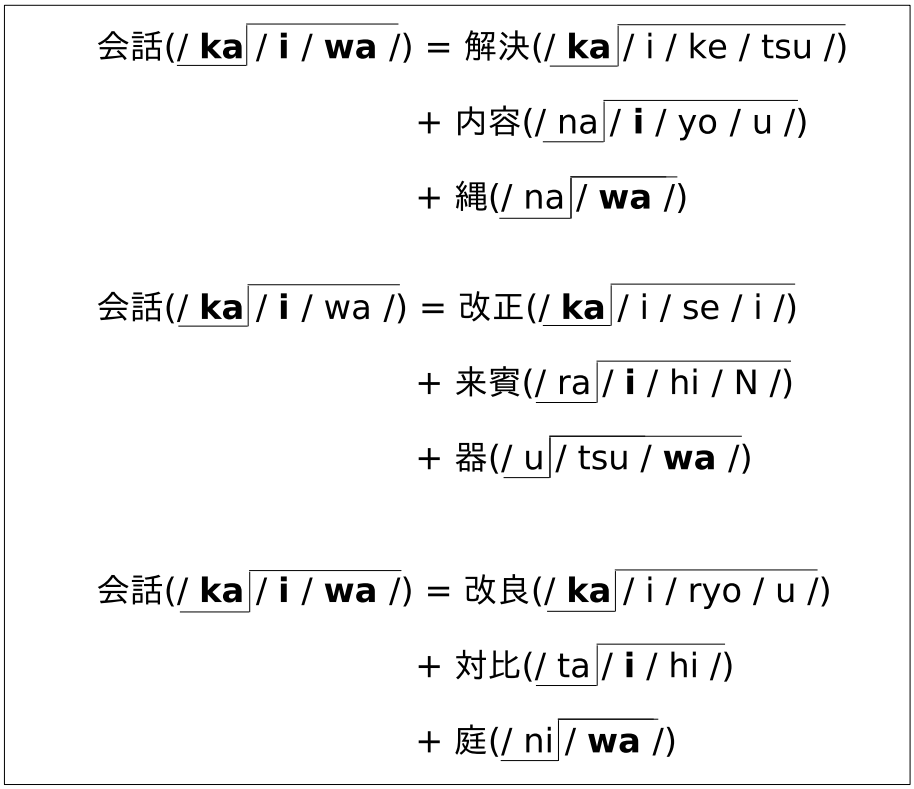

図 24: 合成音声の組み合わせの例 (緩和条件：モーラ情報+前後音素環境)

図 24 において，左側の音声が本研究で作成した合成音声であり，右側の音声において，太字部分で示されている音節素片の波形データを切り出し，接続する事で合成音声の作成を行う．なお図中の「_|￣」はアクセントの高低を表す.

6　実験条件

6.1　実験の目的

本実験では、木に基づくクラスタリングを行う際の最適なパラメータを明らかにする．本研究で調査するパラメータを以下に示す．

- 特徴パラメータ
- 緩和条件

まず特徴パラメータの違いを調査するために，FBANK と MFCC の 2 種類を比較する．またクラスタリングに適した言語的な情報の調査のために，以下の実験を行う．

1. クラスタリングの質問により，音節情報を状態分けする．
2. 状態分けされた音節情報に対して，緩和条件として「モーラ情報」と「モーラ情報及び前後音素環境」の 2 種類に分類し比較する．

6.2 実験方法

本実験では木に基づくクラスタリングを利用するために，アクセント情報，モーラ情報，前後音素環境を考慮した音節モデル (以下，音節環境モデル) を提案する．音節環境モデルの作成には，HTK[4] を使用する．

音節環境モデルの作成手順を図 25 に示す．最初に音節を中心とした初期モデルを作成する (図中 1)．次に半連続型の基本モデルを作成する (図中 2)．最後に学習を繰り返す事で，前後音素環境 (図中 3)，モーラ情報及びアクセント情報 (図中 4) を考慮し，音節環境モデル (図中 5) を作成する．

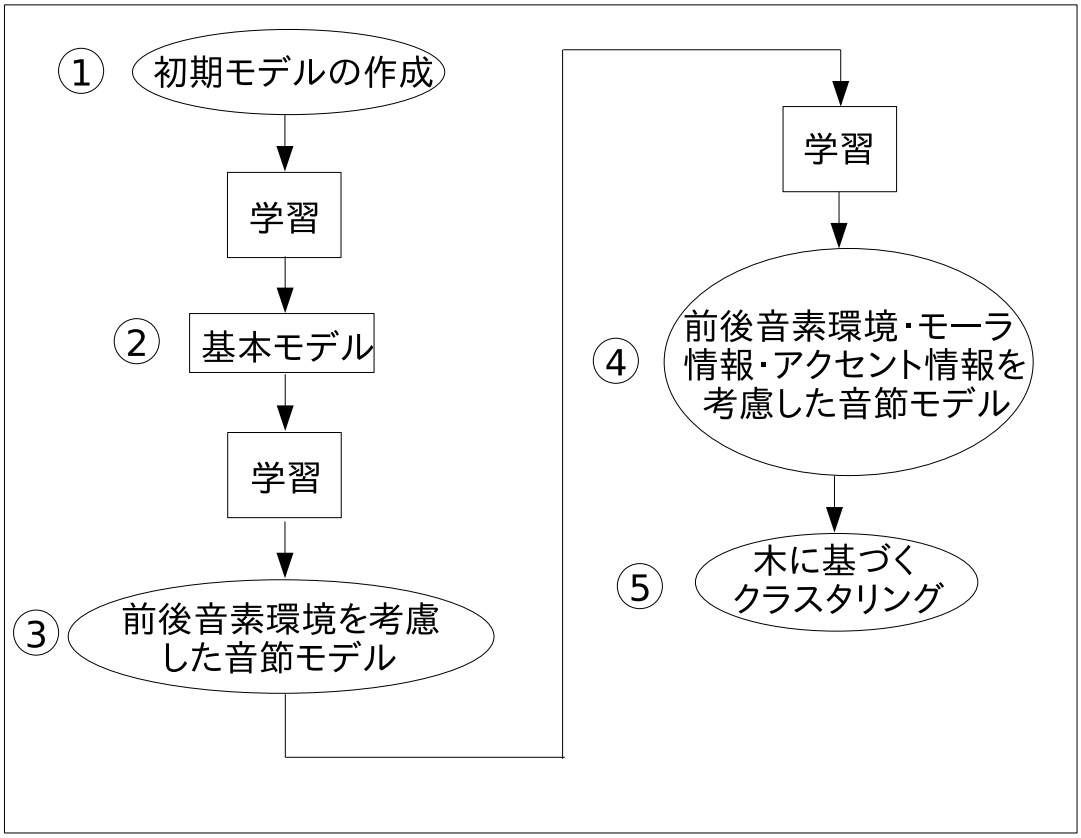

図 25: 実験手順

本実験ではまず音節環境モデルを作成する．次に作成した音節環境モデルに対して，木に基づくクラスタリングを行う．最後にクラスタリングにより得られた情報を基に合成音声を作成する．

6.3 実験環境

本実験の実験条件及び用いたデータベースをそれぞれ表 4，5 に示す．
またクラスタリングの緩和条件の詳細を表 6 に示す．

表 4: 実験条件

基本周波数	16kHz
分析窓	Hamming 窓
1 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	1
FBANK 特徴ベクトル	24 次 FBANK+Δ 24 次 FBANK+ 対数パワー+Δ 対数パワー (計 50 次)
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)
共分散行列	Diagonal-covariance

表 5: 実験に用いたデータベース

データベース	ATR 単語発話データベース ASET(5,240 単語)
話者	女性話者 1 名 (fyn)
学習データ & 合成に用いたデータ	5,240 単語／ fyn

表 6: クラスタリングの緩和条件

緩和条件	説明
モーラ情報	モーラ数及びモーラ位置の緩和
前後音素環境	前音素環境及び後音素環境の緩和

6.4 評価方法

合成音声の評価のために，音声研究に関わった経験のない 5 名を対象に，オピニオン評価実験および対比較実験を行う．対比較実験では 2 種類の同じ音声を続けて流し，どちらの音声が自然に聞こえるかを判定する．またオピニオン評価実験では，自然に聞こえた度合を 5 段階 (5 が最も自然，1 が最も不自然) で評価する．表 7 に聴覚実験に使用した音声を示す.

表 7: 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	木に基づくクラスタリングを利用しない波形接続型音声合成
クラスタリング合成	木に基づくクラスタリングを利用した波形接続型音声合成

6.5 作成単語

本実験で作成した合成音声を表 8 に示す.

表 8: 作成単語一覧

番号	単語	ローマ字	フラグ		番号	単語	ローマ字	フラグ
1	宴会	eNkai	○		26	手順	tejuN	○
2	改善	kaizeN	○		27	伝説	deNsetsu	○
3	買い物	kaimono	○		28	突破	toqpa	○
4	会話	kaiwa	◎		29	独立	dokuritsu	○
5	角度	kakudo	○		30	内職	naishoku	◎
6	瓦	kawara	◎		31	乗り物	norimono	○
7	代わり	kawari	◎		32	発車	haqsha	○
8	学期	gaqki	○		33	発売	hatsubai	○
9	願書	gaNsho	○		34	反射	haNsha	○
10	コップ	koqpu	○		35	歪み	hizumi	○
11	祭日	saijitsu	○		36	日向	hinata	○
12	最善	saizeN	○		37	便利	beNri	○
13	市街	shigai	○		38	見掛け	mikake	◎
14	次第	shidai	○		39	身分	mibuN	○
15	尺度	shakudo	○		40	見本	mihoN	○
16	折角	seqkaku	○		41	無断	mudaN	◎
17	説得	seqtoku	○		42	免許	meNkyo	○
18	設立	setsuritsu	○		43	やかん	yakaN	◎
19	選挙	seNkyo	○		44	野菜	yasai	○
20	全体	zeNtai	○		45	来年	raineN	○
21	退職	taishoku	◎		46	利息	risoku	○
22	対比	taihi	○		47	会費	kaihi	○
23	対話	taiwa	◎		48	形見	katami	◎
24	ダンス	daNsu	○		49	盛り	sakari	◎
25	直角	choqkaku	○		50	追加	tsuika	○

表 8 のフラグは，モーラ情報の緩和のみで作成出来た音声には◎ (11 単語) を，モーラ情報と前後音素環境の緩和で作成出来た音声には○ (39 単語) を付与している.

またクラスタリングの緩和条件の調査のために行う対比較実験に使用する音声として，◎の音声である「会話」「瓦」「対話」「内職」「見掛け」について，接続する音声の組み合わせをランダムに選択することで，各 20 組作成する.

7　実験結果

7.1　オピニオン評価実験結果

得られた合成音声の音質を調査するために，オピニオン評価実験を行う．クラスタリング合成を作成する際には，考えられる複数の組み合わせの音声の中で最良と思われる音声を使用した．なお評価音節数は，音声の種類に対して，各 50 単語ずつ計 150 単語である．実験結果を表 9 に示す．

表 9: オピニオン評価実験の結果 (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 (MFCC)	3.6

表 9 よりクラスタリング合成はオピニオン評価で 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．なお過去の研究 [1] より，オリジナル合成の品質の高さは示されている．以上よりクラスタリング合成は品質の高い合成音声であると言える．

7.2　対比較実験結果

クラスタリングの特徴パラメータの違いによる合成音声の品質を調査するために，対比較実験を行う．なお評価音節数は，同じ発話内容の音声 50 対である．実験結果を表 10 に示す．

表 10: 対比較実験の結果 (総評価音節数:250)

クラスタリング合成 (FBANK)	クラスタリング合成 (MFCC)
47%	53%

表 10 よりクラスタリングの特徴パラメータによる合成音声の音質にほとんど差はないが，若干 MFCC の方が良いことが分かる．

7.3 クラスタリングの緩和条件

7.3.1 オピニオン評価実験結果

クラスタリングで緩和条件に適した言語的な情報を調査する．表 9 で示したクラスタリング合成の内訳を表 11 に示す (計 50 単語).

表 11: オピニオン評価実験の結果 (総評価音節数:250)

緩和条件	クラスタリング合成 (MFCC)
モーラ情報	3.7(11 単語平均)
モーラ情報+前後音素環境	3.6(39 単語平均)

表 11 よりクラスタリング合成の緩和条件による違いは見られなかった.

7.3.2 対比較実験結果

クラスタリング合成の緩和条件による違いを明確にするために，対比較実験を行う．クラスタリング合成を作成する際には，考えられる複数の組み合わせの音声において，ランダムに選択した音声を使用した．なお評価音節数は音声の種類に対して，同じ発声内容の音声 100 対 (5 単語各 20 組) である．実験結果を表 12 に示す.

表 12: 対比較実験の結果 (総評価音節数：500)

緩和条件	クラスタリング合成 (MFCC)
モーラ情報	81%
モーラ情報+前後音素環境	19%

表 12 よりモーラ情報の緩和が約 80%と大きな差が出た.

8　考察

8.1　特徴パラメータ

本研究において，クラスタリングの特徴パラメータによる合成音声の音質にほとんど差は見られなかった．

これはクラスタリングの特徴パラメータにおいて，特徴パラメータによって共有される音節素片の情報が変化する．しかし，実際に合成に用いる音節素片の情報を両者が共有していることが多いため，本実験では差が小さかったと考えられる．

8.2　緩和条件

クラスタリング合成の緩和条件において，表 11 の実験では，クラスタリング合成の中でも最良と思われる音声を使用した．その結果，両者の間のオピニオンスコアに差が見られなかった．しかし表 12 の実験では，ランダムに選択した音声を使用した結果，モーラ情報の緩和が良いという結果が得られた．クラスタリング合成は組み合わせにより，音質に大きな差があるが，任意に組み合わせを選択する事で品質の高い合成音声を作成する事が可能であると考えられる．

以上の結果からクラスタリングの緩和条件には改良の余地があると言える．

9　おわりに

本実験では，クラスタリング合成の最適なパラメータについて調査した．

木に基づくクラスタリングにおいて，特徴パラメータの違いを調査するために，対比較実験を行った．同時にクラスタリングに適した言語的な情報の調査のために，対比較実験を行った．また全体に関して，音声を合成する事で音質の劣化が懸念される為，オピニオン評価実験を行った．

聴覚実験の結果，オピニオン評価実験では，クラスタリングを利用した合成音声で 3.6，波形接続型合音声合成で 3.9，自然音声で 4.5 というオピニオンスコアが得られた．クラスタリングを利用した合成音声は，自然音声には及ばないものの，波形接続型音声合成とあまり差がなく品質の高い合成音声が作成できたことが分かった．

対比較実験では，特徴パラメータに FBANK を用いた合成音声が 47%，MFCC を用いた合成音声が 53%となった．両パラメータにほとんど差はないが，若干 MFCC を用いた合成音声の方が良いという結果が得られたため，クラスタリングの特徴パラメータには MFCC を用いた方が良いことが分かった．

またクラスタリングで条件緩和を行う言語的な情報を調査するために行った対比較実験では，モーラ情報を用いた合成音声が 81%，モーラ情報と前後音素環境を用いた合成音声が 19%となった．この結果よりクラスタリングで条件緩和を行う言語的な情報にはモーラ情報を用いた方が良いことが分かった．

以上より本研究のクラスタリング合成の最適なパラメータは，特徴パラメータとして MFCC，緩和条件としてモーラ情報の緩和である事が分かった．

今後は，木に基づくクラスタリングの質問を改良するなどの検討を行い，クラスタリングのより最適なパラメータについて調査していく予定である．

謝辞

本研究・本論文作成に際して，多大なる検討と種々の御助言をしていただきました鳥取大学工学部知能情報工学科計算機C工学講座池原研究室の池原悟教授，村上仁一助教授に心からお礼を申し上げます．また, 論文を執筆にあたり, 助言を頂いた徳久助手にお礼を申し上げます．さらに聴覚実験の被験者には，計算機 C 工学講座学部 4 年生および計算機 C 工学講座博士前期課程 1 年の田村元秀さんに協力して頂きました．心より御礼申し上げます.

参考文献

[1] 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(2003-3).

[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(2006-03)

[4] ”HTK Ver3.2 reference manual”,2002 Cambridge Universit

[5] 鹿野 清宏, 中村 哲, 伊勢 史朗:“音声・音情報のディジタル信号処理”, 株式会社 昭昇堂,(1997)