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.
SP2003-82 (2003-08)

波形接続型 Speech-to-Speech 音声合成のための可変長音声単位による単位選択手法

正木 敦之† 柏岡 秀紀†,†† ニック キャンベル†, ††,‡

†奈良先端科学技術大学院大学 〒630-0101 奈良県生駒市高山町 8916-5
††国際電気通信基礎技術研究所 〒619-0288「けいはんな学研都市」光台 2-2-2
‡JST/CREST 〒619-0288「けいはんな学研都市」光台 2-2-2

E-mail: †atsu-mas@is.aist-nara.ac.jp, ††,‡{nick, hideki.kashioka}@atr.co.jp

あらまし 波形接続型音声合成はその合成音の高い自然性から人気を集めているが，現時点では利用可能な場面は限られており，日常会話への適用にはいくつかの課題が残されている．我々はこれらの課題のうち，ラベリングされた大規模データベースが必要な点，パラ言語情報を考慮したターゲットの作成が困難である点に着目し，Speech-to-Speech 合成のための単位選択法を提案する．提案法では，データベース中の音声および入力の音声に対して，波形から抽出できる音響特徴量を用いて可変長の音声単位を切り出し，その音声単位をスペクトル情報・韻律情報により特徴付け，その特徴ベクトルの距離計算により単位選択を行う．本稿では大規模音声データベースから音声単位を切り出す技術，そして音響的特徴に基づいた単位選択について提案する．2 種類の音声単位切り出し手法を比較し，両手法で切り出されたコーパスを使った Speech-to-Speech 合成音について，割り当て音素ラベル列によるラベル正解精度，聴覚実験による書き取り正解精度・了解度・自然性を確かめた．

キーワード 波形接続型音声合成，単位選択，Speech-to-Speech 音声合成，可変長音声単位，音響的特徴に基づいた選択

Using variable-sized speech segments as targets for concatenative Speech-to-Speech synthesis

Atsushi MASAKI† Hideki KASHIOKA†, †† and Nick CAMPBELL†, ††,‡

†Nara Institute of Science and Technology 8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0101 Japan
††Advanced Telecommunications Research Institute International
2-2-2 Hikari-dai, "Keihanna Science City", 619-0288 Japan
‡JST/CREST 2-2-2 Hikari-dai, "Keihanna Science City", 619-0288 Japan
E-mail: †atsu-mas@is.aist-nara.ac.jp, ††,‡{nick, hideki.kashioka}@atr.co.jp

Abstract Concatenative speech synthesis is growing in popularity due to the high naturalness of its resulting voice quality, but it is still domain-specific and has not yet been tested with conversational speech. We propose a method of unit selection that will overcome some of the problems that have prevented this development. In particular, we address two problems; one is the need for an extremely large database of labelled speech, the other is the incorporation of paralinguistic information in the speech synthesis. In our proposed 'speech-to-speech' method, we use acoustic criteria to segment the database into variable-sized units, and then use an acoustic waveform as a target for the unit-selection search. In a final stage, prosodic criteria are applied to select the optimal sequence of units for the output waveform generation. In this paper, we describe the techniques for segmenting the large speech database and the acoustic criteria used for unit selection. We present results comparing two methods of speech database segmentation, and further results from accuracy based on phonetic labels and a perceptual test which confirm the intelligibility and naturalness and accuracy of dictation.

Keyword concatenative speech synthesis, unit selection, Speech-to-Speech synthesis, variable-sized speech units, acoustic-based selection.

1. はじめに

波形接続型音声合成はその合成音の高い自然性から人気を集めているが，この手法にはいくつかの問題点が考えられる．まず，従来の波形接続型音声合成では，パラ言語情報を考慮したターゲットの作成が困難で，たとえば肯定の「うん」と否定の「うん」を合成し分けることは難しかった．また，コーパスにはあらかじめ音素などの単位に切り出した音声を格納しておく必要がある．この切り出し作業は自動化が検討されている[1]が，前述のような肯定の「うん」と否定の「う

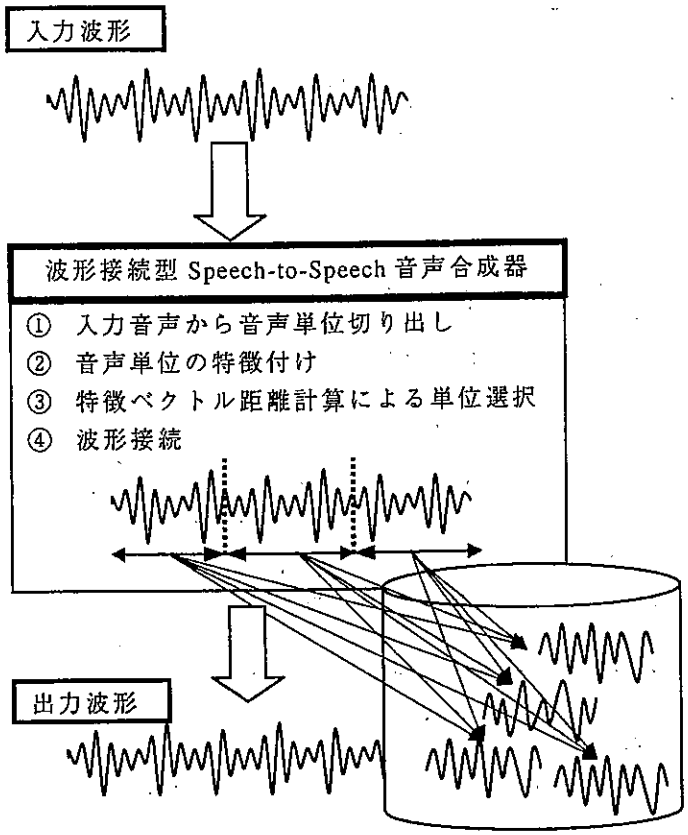

図 1 Speech-to-Speech 音声合成の流れ

ん」を合成し分けるには，日常の家族や友人との対話に表れる表現豊かな音声を収録したコーパスが必要と考えられる．しかし，このような音声には発音のなまけが含まれているため，コーパスとして使う場合に必要な音素単位への切り出しは手作業でも困難を極めると予想される．このようにパラ言語情報を考慮した音声合成を考えると，従来の音素などの音声単位を使った波形接続型の合成手法では新たなコーパスの作成や既存のコーパスの拡張において大きなコストがかかるという問題がある．

我々はパラ言語情報を考慮したターゲットの作成と，音声単位の切り出しコストの 2 点の問題を解決すべく，音声のスペクトルの流れを保存した可変長の音声単位を定義し，この音声単位を使った波形接続型 Speech-to-Speech 音声合成器を実現した．この合成手法は，入力音声に最も近い音声をデータベース中の音響情報から再構築する技術である．このとき，入力音声の韻律情報をそのままターゲットとすることでパラ言語情報を考慮したターゲット作成が可能となる．またコーパスの新規作成や拡張についても，2.1 節で説明するように音声単位を自動的に切り出す手法を採用しており，これも容易である．図 1 に提案法の概要を示す．

Speech-to-Speech 音声合成は以下のような場面で有用であると考えられる．たとえば，従来のテキスト合成器の出力を入力として使えば，音素などの単位に切り出されていない大規模コーパスを使った合成音を作成することができる．また，ささやき声あるいは非可聴つぶやき(NAM)[2,3]を入力とし，通常音声で再現する NAM-to-Speech 音声合成技術実現への第一段階と位置付けることもできる．NAM とは人に聞かせる目的ではなく口の中で自分に対してしゃべる小さな無声音のつぶやきのことで，ささやき声と同様，音源に声門付近の乱流雑音を使用している．つまり，ささやき声や NAM は声帯振動を必要としない発声方法であり，疾患による声帯摘出者へのコミュニケーション補助として NAM-to-Speech は非常に有用な技術と考えられる．ただし，波形から F_0 を抽出できないので，別の情報から音声の高さを推定する必要がある．

さて，本稿の構成は以下の通りである．2 節で 提案法である Speech-to-Speech 合成のための可変長音声単位による単位選択について，3 節で評価実験について，最後に 4 節で本稿の結論と今後の課題について述べる．

2. 可変長音声単位による単位選択

提案法は大まかに，音声単位の切り出し(2.1 節)，音声単位の特徴付け(2.2 節)，単位選択(2.3 節)の 3 つの部分に分かれる．

2.1. 音声単位の切り出し

本稿では，2 種類の切り出し方法を検討した．自動的に音節単位への切り出しを試みた Mermelstein の提案法[4]を拡張したもの(2.1.1 節)と，音楽の開始点自動検出を試みた Klapuri の提案法[5]を音声へ適用したもの(2.1.2 節)である．

2.1.1. エネルギ局所的最小点での切り出し

Mermelstein の提案法は Mokhtari らによって音声単位切り出しに応用されている[6]．ここではまず Mermelstein の手法を述べた後，Mokhtari らの応用，これらを拡張した我々の提案法について述べる．

Mermelstein の手法

Mermelstein は音声からラウドネスを抽出し，その時間推移の軌跡が局所的最小点となる点を音節境界とした．一般的に，パワーの小さい点で音声単位を切り出し，接続すると，接続歪みの軽減が期待できる．つまり，ラウドネスの小さい点で区切ることは，合成への応用には有利であると考えられる．

さて，局所的最小点を見つけるためには以下のようなアルゴリズムを経る．まずラウドネスから凸型殻関数(convex-hull function)を算出する．凸型殻関数は，ラウドネスのピーク点へ向かって単調増加し，ピーク点を過ぎると単調減少する関数である．この凸型殻関数

と元のラウドネス関数との差が最大となる点を探し，その差がある閾値以上であればその点を音節の境界であるとした(図 2 に概念図を示す)．このプロセスを経て決定された境界と境界の間で再び，

1. ピークを検出
2. 凸型殻関数の算出
3. ラウドネス関数との差の最大値を検出
4. 閾値以上ならその点を境界と認定

という流れを再帰的に繰り返し，順に境界を決定していく[4]．

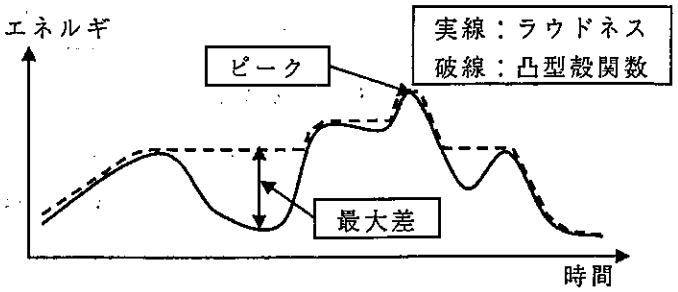

図 2　局所的最小点での切り出し

Mokhtari らの応用

Mokhtari らは，母音中心の音声単位を切り出すため，ラウドネスの代わりに SE (Sonorant Energy, 60～3400Hz のエネルギ)を参照し，このエネルギの局所的最小点で音声単位を切り出した．彼らは，より短い単位が切り出せれば音韻的・韻律的なカバー率を上げることができるという結論を得た[6]．

提案法

我々は予備的に SE を参照した Mokhtari らの手法を試した．切り出された音声単位と音素ラベル数の関係を表 1 に示す．対象としたデータは日本人男性話者 MHT による ATR 音素バランス文(503 文)読み上げデータである．このデータには手作業による音素ラベルがついており，表の値は切り出された音声単位に割り当てられた音素ラベルの数を示している．音素の内部で切り出された音声単位には，その両方の単位に音素ラベルを割り当てた．また，音声単位境界は 2.1.3 節で示す切り出し位置修正を経たものである．

SE を使った局所的最小点切り出しは Mokhtari らの結果と一致し，長めの音声単位が切り出された．コーパスのカバー率や特徴ベクトル(2.2 節)の説明能力を考え，より短い単位を切り出すためこのアルゴリズムの拡張を試みた．

拡張は 0～8000Hz までをメルスケールで 16 バンドに等分割し，ラウドネスの代わりに各バンドのエネルギを参照した局所的最小点切り出しを行い，その結果をマージするというものである．マージは，まず分割前の 0～8000Hz 帯のエネルギで切り出した結果と分割後の各バンドの切り出し結果を 20 msec 以内の差なら同じ境界であるとし前者に代表させる．残った境界候補と先ほどのマージ結果(分割前の 0～8000Hz 帯のエネルギで切り出した結果と同じ)とを比べ 40 msec 以上離れていれば残った候補も境界として認めることとした．理解を簡単にするためにバンド分割数 2 のときのマージ例を図 3 に示す．提案法での切り出し結果を表 1 に示す．

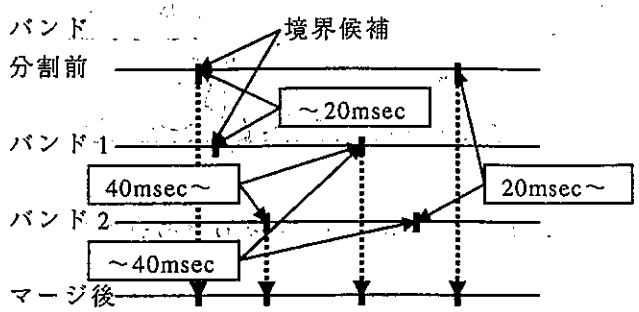

図 3　局所的最小点切り出しのマージ例

2.1.2. エネルギ変化ピーク点での切り出し

Klapuri は音楽を対象として，各楽器の開始点検出を試みた．この Klapuri の手法と，これを音声に応用した我々の提案法について述べる．

Klapuri の手法

Klapuri は複数バンドの対数エネルギ推移に対し時間微分した係数を取り，すべてのバンドの微分係数の絶対値の単純総和を取ることでエネルギ変化を際立たせ，閾値を超えたピーク点を開始点であるとした[5]．

提案法

対象を音声とするにあたり，以下のような拡張を施した．音声のエネルギ分布が変化するということは，その時点で調音動作の最中であることを意味する．この調音動作のピーク点より少し戻った点(10msec と決めた)を調音開始点であるとし，この点で切り出すこととした．この音声単位切り出しでは，使うバンド数・バンド幅は 0～8000Hz までをメルスケールで 16 バンドに等分割とし 2.1.1 節のエネルギ局所的最小点での切り出しと同じにした．図 4 に概念図を示す．また提案法での切り出し結果を表 1 に示す．

先ほどの Mermelstein 法と比べ，切り出し位置決定のプロセスがシンプルなこと，音楽に関して既に実績があることが特徴として考えられる．

2.1.3. 切り出し位置修正

2.1.1 節，2.1.2 節で決定された切り出し位置は，波形接続時の歪み軽減をねらい近傍ゼロクロス位置に修正した．

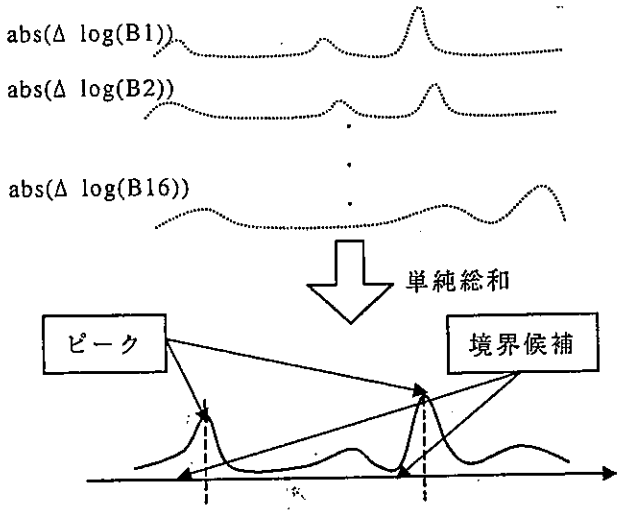

図 4　エネルギ変化ピーク点での切り出し

表 1　音声単位切り出し比較

切り出し手法	平均	標準偏差	最大長
局所的最小点 (SE のみ)	3.69	1.32	15
局所的最小点 (16 バンド)	2.44	0.83	6
変化ピーク点	2.54	1.34	14

2.2. 音声単位の特徴付け

提案する合成手法では，音声単位を音韻ラベルの参照無しで切り出し，単位選択するため，スペクトル情報を音声単位の特徴として組み込む必要がある．さらに韻律情報を組み込み音声単位の特徴ベクトルを作成した．音声単位内の音響パラメータについて 0～4 次の離散 Cosine 変換(以下 DCT)係数を特徴ベクトルに組み込む Mokhtari らの手法[6]の応用となっている．

2.2.1. スペクトル情報の特徴付け

2.1 節で示したように，切り出し位置決定にはメルスケール 16 バンドのエネルギの推移を用いている．このエネルギが，スペクトル情報を示すパラメータであるとし，各バンドのエネルギ遷移を DCT 係数 5 つに変換して特徴ベクトルに組み込んだ．16 バンドのエネルギ推移があるので，16*5 = 80 個のパラメータで音声単位のスペクトル情報を特徴付けることになる．

2.2.2. 韻律情報の特徴付け

さらに韻律情報も音声単位の特徴として組み込む．パラメータとしては F_0，継続時間長(Dur)を採用した．F_0 は先と同様 5 つの DCT 係数に変換する．その結果，音声単位は 80+5+1 = 86 次元のベクトルとして表現される．韻律情報としてよく知られるパワーは，スペクトル情報を特徴付けた 80 個のパラメータが説明していると考えている．図 5 に特徴ベクトル算出の概要を示す．

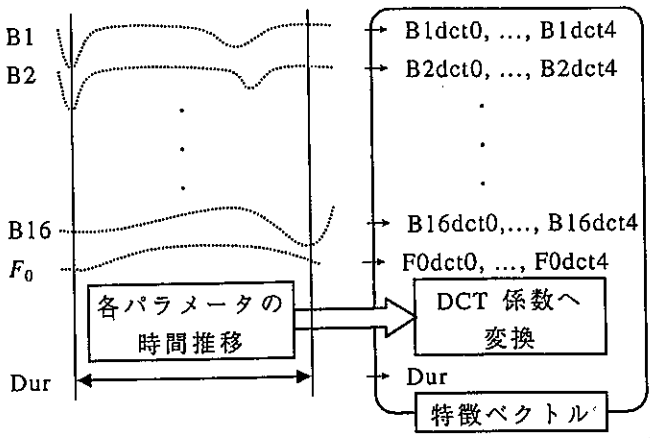

図 5　音声単位の特徴付け

2.3. 単位選択

単位選択は 2.2 節で定めた特徴ベクトルに基づいて行われる．具体的には，コスト関数を次式のように定め，その値の一番小さいもの，つまり重み付き距離の一番近いものが選ばれる．

$$C=\sum_{i=1}^{86} w_i\sqrt{(t_i-u_i)^2}$$

ここで，t はターゲットの特徴ベクトル，u はコーパスに格納された音声単位の特徴ベクトル，w は重みを意味し，重みはパラメータ間でスケールの異なることを考慮し，スペクトル情報および韻律情報のパラメータ 86 個すべてについてそれぞれの標準偏差の逆数と定めた．

3. 評価実験

提案法の性能を評価するため次の実験を行った．以下，まず実験に使った音声データについて触れ，発話内容の書き取りと了解度・自然性を評価する主観評価実験，音韻ラベルに基づく客観評価実験について述べる．

3.1. 音声データと実験試料

実験には，表 1 で示したものと同じ日本人男性話者 MHT が発声した ATR 音素バランス文(503 文)の読み上げ音声を利用した．このデータには手作業による音素ラベルが付与されている．

まず，このデータから 2.1 節 の 2 つの手法で音声単位を切り出し，次にこの音声単位を 2.2 節 の手法で特徴ベクトルに変換することで，2 種類のコーパスを作成した．そしてコーパス作成に利用した音声からターゲットとなる音声をひとつずつ選び，2.3 節の手法で単位選択し，その後波形接続した．なお，503 文の

1 文目を合成する時は残りの 502 文から選択するようにし，単位選択時にはターゲットとなる音声はコーパスから削除した．こうして，それぞれのコーパスで 503 文すべてを合成し，計 1006 文の合成音を得た．

図 6 に「あらゆる現実を」を合成した例を示す．紙面の都合上，局所的最小点切り出し(2.1.1 節)で構築したコーパスの元の音声と合成結果のスペクトログラムのみを示す．

3.2. 主観評価実験

3.2.1. 実験条件

主観評価実験として次の 2 つの実験を試みた．被験者は日本人成人男女 12 名である．

実験(1) 書き取りによるかな正解精度

今回の合成手法が音韻ラベルを使用しない手法であることから，合成のターゲットとした音声の発話内容を伏せた上で合成音を聞き内容を書き取ってもらった．書き取る際には，聞き取れた通りではなく各被験者の許容範囲内で正しい日本語としてかなで書き取ってもらった．

この実験で用いる合成音のターゲットとして 503 文から 20 文選んだ．前半 10 文を局所的最小点切り出し(2.1.1 節)で合成し，後半 10 文を変化ピーク点切り出し(2.1.2 節)で合成した実験試料セットとその逆の組み合わせの試料セットを作成し，被験者の半分には前者を，もう半分は後者を聞かせることでターゲットの文章と切り出し手法の違いを相殺した．

そして正解と被験者の回答とでかな単位のアラインメントを取り，かな正解精度 (音声認識の認識率計算に使われる単語正解精度に準じたもの) を算出した．

実験(2) 了解度・自然性

発話内容を示した(1)とは別の合成音を聞かせ，その文章を読んだものとして了解できるかを 5 段階評価(1:了解度が低い～5:了解度が高い)してもらうと同時に，自然性を 5 段階評価(1:自然性が低い～5:自然性が高い)してもらった．自然性は接続の歪み感や韻律など総合的に判断してもらう．

この実験に用いる音声試料のターゲットには 503 文から 10 文を採用した．切り出し手法別に同じ 10 文を合成し，被験者は計 20 文の了解度・自然性を評価する．その評価値平均と標準偏差を算出した．

なお，合成音を聞く際，両実験とも何度でも試料を聞いて良いこととした．

3.2.2. 実験結果

実験(1),(2)の結果を表 2 に示す．(1)はかな正解精度を百分率で，(2)は評価値平均(上段)と標準偏差(括弧内下段)を示す．

書き取りによるかな正解精度は局所的最小点切り出しで 75.7%と高い数字を得た．変化ピーク点切り出しでは 60%弱と低い数字にとどまった．了解度・自然性は両手法とも平均値は 3 を下回る低い値であるが，いずれも局所的最小点切り出しが変化ピーク点切り出しを上回っている．

3.3. 客観評価実験

3.3.1. 実験条件

2 つの手法で切り出した音声単位に音素ラベルを割り当て，ターゲット音声単位の割り当て音素ラベル列と選択された音声単位の割り当て音素ラベル列の間で，かな正解精度と同じく音素ラベル単位のアラインメントを取り，音素ラベル正解精度を算出した．

これをすべての単位選択結果について計算し，切り出し手法別の平均値を算出する．

3.3.2. 実験結果

音素ラベル正解精度は 503 文全体と 3.2.1 節の実験 (1) で使用した試料，実験 (2) で使用した試料の 3 種について表 3 に示す．

503 文全体で見ると 7 割を超えており，まずまずの数値を得た．また 3 種すべてにおいて，変化ピーク点切り出しの方が高い値を出しており，主観評価実験の時とは逆転現象が起きている．

ここで，表 1 を見てみると音声単位の長さは局所的最小点切り出しのものより変化ピーク点切り出しの方が長くばらつきのあることがわかる．表 4, 5 に，主観評価実験で用いた音声試料のターゲットについて，表 1 と同様の表を示した．コーパス全体と同じ傾向があらわれている．長い音声単位と短い音声単位を比較すると，前者でひとつ音素ラベルを間違えることと，後者でひとつ音素ラベルを間違えることの間には音素ラベル列正解精度を計算するときの重みが変わるので，このような結果が出たのではないかと推測する．評価尺度を変えるなどして更なる分析が必要である．

表 2　主観評価実験の結果

切り出し手法	(1)書き取り	(2)了解度	(2)自然性
局所的最小点	75.7 %	2.81 (0.511)	2.43 (0.527)
変化ピーク点	59.2 %	2.38 (0.813)	2.39 (0.646)

表 3　客観評価実験の結果

切り出し手法	503 文全体	(1)試料	(2)試料
局所的最小点	71.0%	72.0%	69.3%
変化ピーク点	75.8%	77.0%	77.1%

表 4　実験(1)試料ターゲットの音素ラベル割り当て数

切り出し手法	平均	標準偏差	最大長
局所的最小点	2.36	0.84	5
変化ピーク点	2.61	1.40	11

表5　実験(2)試料ターゲットの音素ラベル割り当て数

切り出し手法	平均	標準偏差	最大長
局所的最小点	2.47	0.86	6
変化ピーク点	2.62	1.35	10

4. おわりに

本稿では，波形接続型テキスト音声合成のコーパスの新規作成や拡張には大きなコストがかかり，またパラ言語情報を考慮したターゲット作成が難しいという問題に触れ，複数のバンドに分けたエネルギを参照して自動的に決定することが可能な可変長音声単位を用いた波形接続型 Speech-to-Speech 音声合成法を提案した．可変長音声単位は，複数のバンドの局所的最小点を音声単位境界候補としマージする手法(2.1.1 節)と，エネルギの時間変化のピーク点から少し戻った点を境界とする手法(2.1.2 節)を提案した．さらにこの 2 つの方法で切り出された音声単位をスペクトル情報・韻律情報を使って特徴付ける手法(2.2 節)と単位選択法(2.3 節)について提案した．

また提案法の性能を評価するため，ATR 音素バランス文(503 文)をコーパスとした合成音を作成し，書き取り正解精度と了解度･自然性を測る聴覚実験を行い，また客観評価として音声単位に割り当てられた音素ラベルのラベル正解精度を測った．結果，局所的最小点切り出しでは 76.8% の書き取り正解精度を示し，音素ラベル列正解精度も 70% を超える値を示した．一方で主観評価の了解度・自然性ともに局所的最小点切り出し，変化ピーク点切り出しのいずれも 3 を下回るにとどまった．

今後の課題として，了解度・自然性の向上のために音韻環境などを考慮した単位選択をすることが考えられる．また，音素などの単位に切り出しにくい発音のなまけを含んだ音声，パラ言語情報を豊富に含んだ話し言葉を収録したコーパスを使って合成音を作成し，提案法の評価を行う必要がある．

謝辞

本研究を援助していただいた JST/CREST に感謝致します．

文　献

[1] 河井 恒，戸田智基，“波形接続型音声合成のための自動音素セグメンテーションの評価”，信学技報，SP2002-170, pp.5-10, January, 2003.

[2] 中島淑貴，柏岡秀紀，鹿野清宏，ニック・キャンベル，“微弱体内伝導音抽出による無音声認識”，日本音響学会講演論文集，3-Q-12，pp.175-176, Mar. 2003.

[3] Y. Nakajima, H. Kashioka, K. Shikano, N. Campbell, "RECOGNITION INPUT INTERFACE USING STETHOSCOPIC MICROPHONE ATTACHED TO THE SKIN", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2003, MMEDIA-L3.1, April 2003.

[4] P. Mermelstein, "Automatic segmentation of speech into syllabic units", J. Acoust. Soc. Am. 58(4), pp.880-883, 1975.

[5] A. Klapuri, "Sound Onset Detection by Applying Psychoacoustic Knowledge", IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 1999.

[6] P. Mokhtari, N. Campbell, "Automatic characterisation of quasi-syllabic units for speech synthesis based on acoustic parameter trajectories: a proposal and first results", in Proceedings of the Autumn2002 Meeting of the Acoustical Society of Japan, Akita, Paper 1-10-5, pp.233-234.

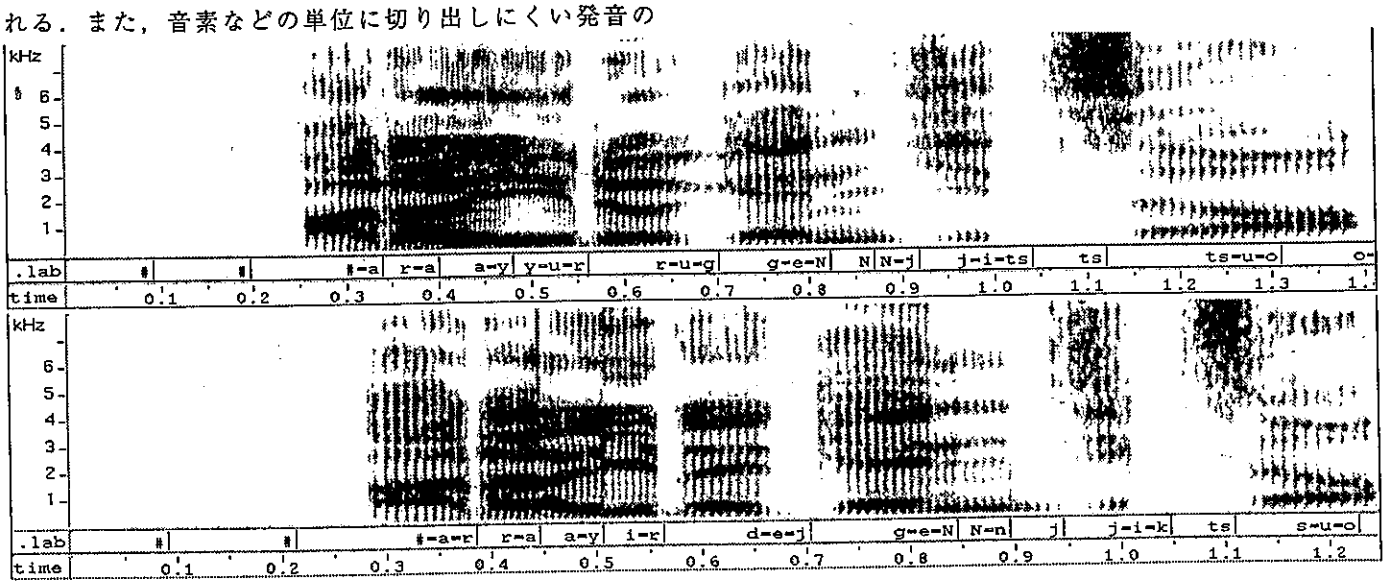

図 6　「あらゆる現実を」の切り出し例(上)と，これをターゲットとした合成例(下)