# 無声子音を含む遷移区間を考慮したマルチパルス音声分析合成システム

李 時雨† 高橋 寛†

本論文では，無声子音を含む遷移区間(Transition Segment Including UnVoiced Consonant : TSIUVC)を考慮したマルチパルス音声分析合成システムを提案し，6.9 kbit/s の低ビットレートで良好な音質を得るシステムであることを明らかにしている．従来のマルチパルス音声符号化方式では，有声音の音質に比べ，無声子音で音劣質化が現れている．その音質劣化の原因に，白色雑音やマルチパルスの無声音源の算出方法，これらの無声音源と全極型の合成フィルタとの非整合性，無声子音や有声音と無声子音が混在しているフレームで全極型の合成フィルタの係数が不安定になることなどがある．従来の方式ではフレーム内の音声信号を有声音源か無声音源のどちらかで合成しているため，フレーム内に無声子音と有声音が混在している場合，音質劣化が現れる．本論文では，まず TSIUVC 区間と有声音区間が混在しないようにフレームを再作成するとともにピッチパルス間隔の変動に対応できるよう FIR-STREAK ディジタルフィルタと補間処理による個別ピッチパルスの抽出法について述べている．さらに，無声音源および全極型や極零型の合成フィルタを使わずに短文から TSIUVC 区間を探索・抽出し，区間内の周波数帯域信号を情報圧縮・近似合成する TSIUVC 近似合成法を用いた．今回の TSIUVC の抽出率は目視数と抽出数との比較を行った結果，男声で91.2%，女声で86%であった．また，TSIUVC 近似合成法を用いたシステムと用いないシステムとで合成した音声の品質を評価した結果，用いないシステムに比べ TSIUVC 近似合成法を用いたシステムの方が聴覚的に音質が改善されることを明らかにした．

# A Multi-Pulse Speech Analysis-Synthesis System Considering Transition Segment Including Unvoiced Consonant

See Woo Lee† and Yutaka Takahashi†

In this paper, we propose a method of approximate-synthesis of the transition segment including unvoiced consonants (hereafter TSIUVC Approximate-Synthesis Method). The TSIUVC is extracted by using the zero-crossing measurement and individual pitch pulses. The number of the extracted TSIUVC was compared with that of the as-observed, indicating such high extraction rates as 86% for female voice and 91.2% for male voice respectively. The TSIUVC signals in the time domain are transformed to those in the frequency domain by means of the FFT, and followingly compressed by deviding the frequency band. For their reproduction, the compressed signals are put back to the time domain using the IFFT. We evaluate the MPC system (system (A)) which does not use TSIUVC Approximate-Synthesis Method and the MPC system (system (B)) use TSIUVC Approximate-Synthesis Method. As the result, we knew that synthesis speech of the system (B) was better in speech quality than synthesis speech of the system (A).

## 1. まえがき

近年，DSP (Digital Signal Processor) の技術進歩を背景に通信網のディジタル化が急速に進められている．通信網の回線を有効に利用し，伝送コストを低減させるためには低ビットレート符号化が望ましく，通信サービスの多様化に伴い高能率でかつ高品質な音声分析合成技術が求められている．

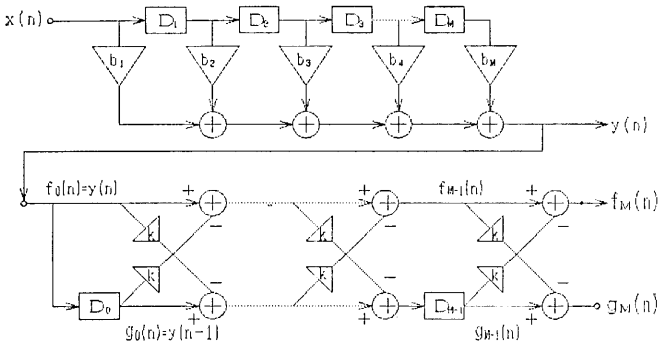
音声を能率的に符号化する LPC (Linear Prediction Coding) 方式は有声音源にインパルス列，無声音源に白色雑音という単純な音源モデルを用いるため，音質面では不十分となる．そこで，有声音の調音特性を複数のパルス，いわゆるマルチパルスで LPC 合成フィルタを駆動させるマルチパルス音声符号化方式が提案されている[1]．さらに，ピッチ情報を導入したマルチパルス音声符号化方式についても報告されている[2]．ピッチ情報は音声の有声/無声を区別する重

† 日本大学理工学部電子工学科
College of Science & Technology, Nihon University

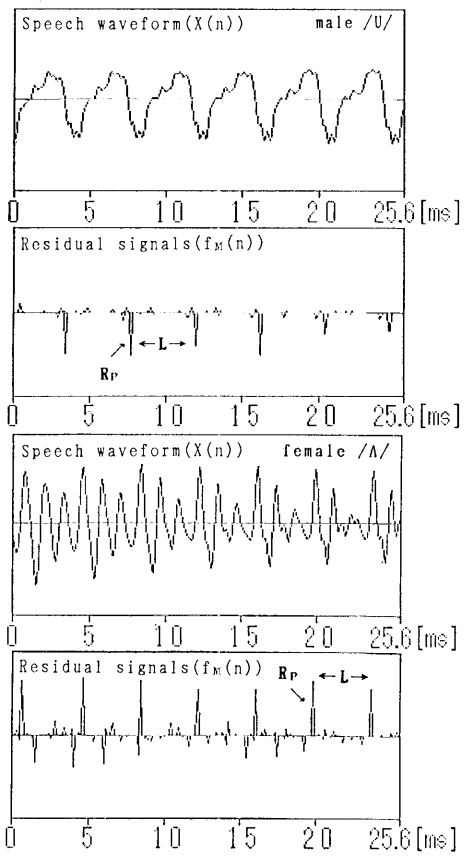
要なパラメータであり，ピッチの違いは男女の区別や話者の個人性，情緒性，音声の自然性などに有効である．一般に，広く用いられているピッチ抽出法に自己相関法やケプストラム法などがあり，平均的なピッチ間隔を算出している[3)~7)]．しかし，音声波形のメカニズムは音韻のミクロな変動や音韻の調音干渉により常に変動するのでピッチパルスの間隔も一定ではなく時間とともに常に変動する．したがって，平均的なピッチ間隔は音声のミクロな特性の変化に対応できず，無声音と有声音が混在しているフレームなどでピッチ抽出の誤りを生じ，音質低下の原因となる．

一方，音声分析合成や規則合成などの分野において，無声子音の音質を改善させるため，無声音源に圧縮残差信号，マルチパルスなどを使用したり，マルチパルスと白色雑音とを混合して用いるなど様々な工夫がなされている[1),8),9)]．しかし，無声子音の音質劣化は根本的に解決されていない．その原因に次のようなことがあげられる．第1に，相関演算に基づいて算出した無声音源では非定常的で非線形な無声子音の特性を正確に近似することは困難であると考えられる．第2に，無声子音の声道特性は極めて短時間にわたって変動するので零点の個数も時間とともに変化する．したがって，次数固定の全極・極零型の合成フィルタでは無声子音の特性を最適に推定することは困難であると考えられる．また，従来のマルチパルスの音声符号化方式では音声合成に線形モデルに基づくPARCOR合成フィルタを用いているが，無声子音や有声音と無声子音が混在しているフレームでPARCOR合成フィルタの係数 $k$ が不安定（$|k|>1$）になることが多い．このようなことを考慮すれば，従来の方法では非定常的な無声子音のメカニズムを最適に近似することは困難であるといえる．

そこで，本論文では，まず，ピッチパルス間隔の変動に対応できるよう FIR-STREAK ディジタルフィルタの正と負の残差信号から求めた個別ピッチパルス




（a） フィルタの構成
（a） Structure




（b） 音声波形と残差信号
（b） Speech waveform and residual signals

**図 1** FIR-STREAK ディジタルフィルタ
Fig. 1 FIR-STREAK digital filter.

を用いる方法について述べる．次に，無声音源および全極・極零型の合成フィルタを使わずに短文から無声子音とこれに続く遷移区間（**T**ransition **S**egment **I**ncluding **U**n**V**oiced **C**onsonant：以下 **TSIUVC** と呼ぶ）を探索・抽出し，近似合成するマルチパルス音声分析合成システムを提案する．さらに，TSIUVC 近似合成法を用いたシステムと用いないシステムとで合成した音声に対し客観・主観的な比較評価を行い，その評価結果について述べる．

## 2. ピッチ抽出法

本章では，まず，FIR (Finite Impulse Response) ディジタルフィルタと STREAK (Simplified Technique for Recursively Estimating Autocorrelation $K$-parameters) ディジタルフィルタとを組合わせた FIR-STREAK ディジタルフィルタによるスペクトル平坦化処理について述べる．次に，FIR-STREAK ディジタルフィルタにより得られた正と負の残差信号から個別ピッチパルスを求める方法について述べる．

### 2.1 スペクトル平坦化処理

音声波形は無音区間 (silence segment)，子音区間(consonant segment)，遷移区間 (transition segment)，そして母音区間 (vowel segment) により構成されている．母音区間と有声子音区間には声帯振動に伴う相似的な波形の繰り返しが存在する．その繰り返し波形の長さは声帯の開閉時間に基づいており，声帯の開閉間隔がピッチパルスの間隔となる．また，声帯の開閉時点がピッチパルスの時間位置である．したがって，ピッチパルスを抽出するために声帯の開閉時点のところに鋭いピークが現れるよう，各高調波成分の振幅を平坦化する必要がある．文献10)によれば，スペクトル平坦化は STREAK ディジタルフィルタにより達成でき，その誤差信号を低域フィルタに通すことにより，著しく高い相関のピークを得ることができるとされている．しかし，スペクトル平坦化の見地から，低域フィルタを STREAK ディジタルフィルタの後に置くより，前に置いた方がピッチ周期の境界で鋭いピークを得やすいと判断し，実験によりこれを確かめた．なお，そのピークの出現頻度は低域フィルタの次数および遮断周波数に依存している．

ここでは，**図1**(a)に示すように低域フィルタとして零位相特性を有する FIR ディジタルフィルタを用いた．また，次数 ($M_F$) および遮断周波数 ($F_F$) は，$20 \leqq M_F \leqq 120$，$400\,\mathrm{Hz} \leqq F_F \leqq 800\,\mathrm{Hz}$ の条件で，声帯の開閉位置のところに現れるピークの等間隔性より40次，800 Hz とした．そして，FIR ディジタルフィルタの係数は，直交フィルタの構成を簡単化できる利点から複素係数とした．結局，音声信号を FIR ディジ

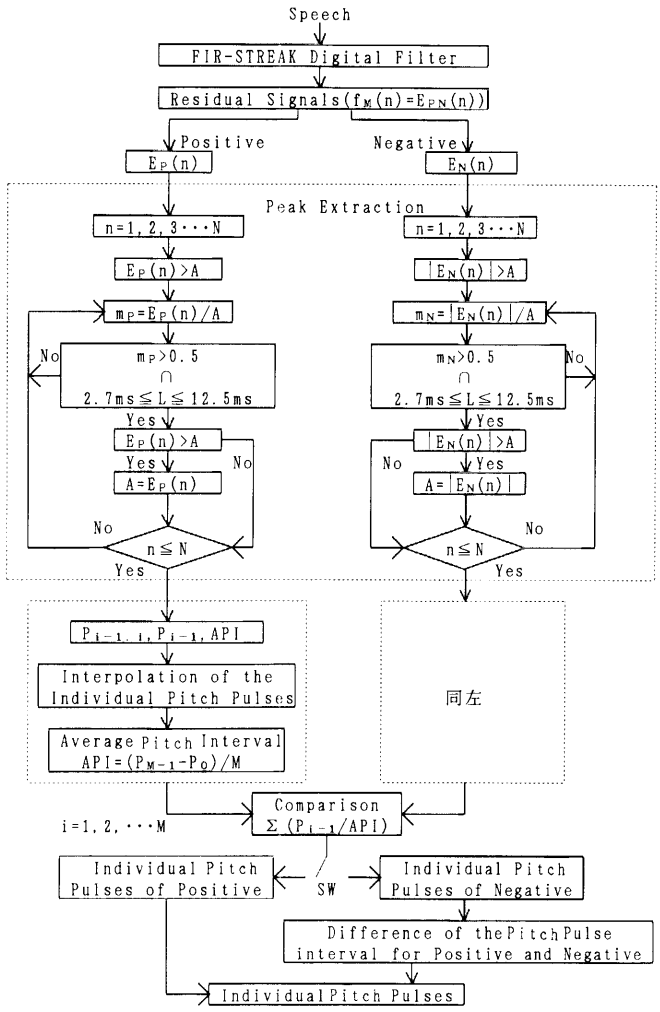



**図2** 個別ピッチパルスの抽出法
Fig. 2 Extraction method of the individual pitch pulses.

タルフィルタで帯域制限した上，STREAK ディジタルフィルタに通すことにより平坦化された残差信号（$f_M(n)$）を得る．その一例を図1（b）に示す．

### 2.2 個別ピッチパルスの抽出法

このアルゴリズムの構成を図2に示す．図1（b）のように FIR-STREAK ディジタルフィルタの出力の残差信号（$f_M(n)=E_{PN}(n)$）は時間軸に対して正の残差信号の振幅（$E_P(n)$）と負の残差信号の振幅（$E_N(n)$）で構成されている．また，$E_P(n)$ と $E_N(n)$ にはパルス性の残差信号（$R_P$）と雑音性の残差信号が含まれている．$R_P$ はピッチ構造に起因するインパルス的な信号であり，雑音性の残差信号はランダム的な信号である．したがって，個別ピッチパルスを抽出するためには雑音性の残差信号から $R_P$ を分離する必要がある．しかし，図1（b）のように音声波形の形状によって $R_P$ が時間軸に対して正のところで等間隔に現れる場合と負のところで等間隔に現れる場合とがある．そこで，個別ピッチパルスの候補となる $R_P$ を正と負の両方から抽出した方が望ましい．時間軸に対して正あるいは負のところから $R_P$ を抽出する手順は正や負とも根本的に同じであるので，本論文では正のところから $R_P$ を抽出する方法について述べる．まず，**図2**の peak extraction 部に示されているように残差信号の振幅値を規格子(A)により正規化する．つぎに，**表1**の音声試料を基に $m_P$ を算出した結果，$R_P$ 時刻における $m_P$ の値は 0.5 以上となることがわかった．そこで，$E_P(n)>A$ と $m_P>0.5$ である残差信号を $R_P$ とし，ピッチ周波数は 80～370 Hz に存在することを考慮して $R_P$ の間隔 $L$ が $2.7\,\text{ms}\leqq L\leqq 12.5\,\text{ms}$ となる $R_P$ の位置を個別ピッチパルスの位置とした．

しかし，瞬時値の緩やかな正弦波の場合，$R_P$ の欠落により個別ピッチパルスの欠落が現れる場合がある．個別ピッチパルスの欠落には，25.6 ms のフレーム（短時間分析の区間長）当たり個別ピッチパルスが1個だけ欠落する場合と連続に欠落する場合とがある．そこで，前者に対しては過去の個別ピッチパルスの間隔（$P_{i-1}$），平均ピッチ間隔（API），個別ピッチパルスの間隔偏差（$P_{i-1,i}$）で補間を行う．後者については，まず，正と負の個別ピッチパルスの間隔偏差を求める．次に，$\Sigma$（個別ピッチパルスの間隔/平均ピッチ間隔）が小さい方の個別ピッチパルスを選択するようにした．ただし，時刻0から最初の個別ピッチパルス（$P_0$）までの時間隔は個別ピッチパルスの間隔偏差の算出に含めない．

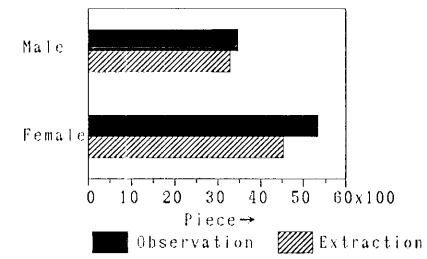
個別ピッチパルスの抽出において，①本来の個別ピッチパルスを抽出できなかった場合，②本来は存在しないにもかかわらず抽出された場合がある．この①②を抽出誤りとする個別ピッチパルスの抽出率（Automatic Extraction Rate : AER）を次式により算出した．

$$\mathrm{AER}_1=\sum_{j=1}^{N}\sum_{i=1}^{m}(a_{ij}-(b_{ij}+c_{ij}))\Big/\sum_{j=1}^{N}\sum_{i=1}^{m}a_{ij} \qquad (1)$$

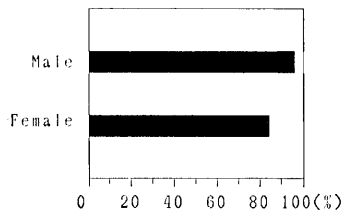
$m$：個別ピッチパルスが存在するフレーム数

$N$：音声試料数

ここで，$a_{ij}$, $b_{ij}$, $c_{ij}$ はそれぞれ1フレーム当たり真の個別ピッチパルス数と見なされる目視数，①の誤り数，②の誤り数を示す．**表1**の音声試料を用いた結




（a） 個別ピッチパルス
（a） Individual pitch pulses




（b） 抽出率
（b） Extraction rate

**図3** 個別ピッチパルスの抽出率
Fig. 3 Extraction rates of the individual pitch pulses.

**表1** 音声試料（個別ピッチパルス）
Table 1 Speech sample (Individual pitch pulses).

| 条件 | 男声 | 女声 |
|---|---|---|
| 発話者 | 4 | 4 |
| 発話時間 | 3.4秒 | 3.4秒 |
| 短文数 | 16 | 16 |
| 母音数 | 145 | 145 |
| 無声子音数 | 34 | 34 |

果，目視の個別ピッチパルスの数（式(1)の分母）は男声で 3483 個，女性で 5374 個であり，抽出した個別ピッチパルスの数（式(1)の分子）は男声で 3343 個，女声で 4566 個であった．したがって，男女声における個別ピッチパルスの抽出率は男声で 96%，女性で 85% が得られた．これらの結果を**図3**に示す．

## 3. TSIUVC 近似合成法

本章では，聴覚的な音質改善を目的とする TSIUVC 近似合成法について述べる．まず，TSIUVC 近似合成法の基本概念，無声子音の特徴分析について述べる．次に，零交差レートと2章で述べた個別ピッチパルスを用いて短文から TSIUVC を探索・抽出する方法について述べる．さらに，TSIUVC を周波数軸上で情報圧縮・近似合成する方法について述べる．

### 3.1 基本概念

音声認識分野では音節知覚に重要視される無声子音の分析について数々の論文が報告されている[11)~13)]．一方，音声分析合成分野では母音を重視する傾向から，無声子音とこれに続く遷移区間（TSIUVC）の分析・合成について報告された例は筆者らが調査した限りでは見当たらなかった．

文献8)によれば，無声音区間を白色雑音のような無声音源により励起した場合，無声音区間や有声音区間と無声音区間の接続部で音質劣化が生じると示されている．また，有声音区間と無声音区間との接続方法の必要性についても示されている．そして，無声破裂子音の知覚実験における主要な cue は，後続母音のホルマント遷移よりも破裂開始時点から声帯振動の開始時点までの区間に存在するものと考えられている[14)]．同様に，無声摩擦子音や無声破擦子音も音節知覚に重要とされる情報は TSIUVC 区間に存在するものと考えられる．

以上，上記の根拠と1章で述べた無声音源および無声子音の劣化原因などを考慮すれば，従来の無声音源および全極・極零型の合成フィルタを使わずに TSIUVC を忠実に再現することで音質を改善することができると考えられる．また，後述する3.3節の方法により定常的な有声音と非定常的な無声子音が同フレーム内に混在しないようにした上，後述する3.4節の方法により有声音区間と無声子音区間の間に存在する遷移区間の音声信号を忠実に再現することで有声音区間と無声子音区間とを低歪みに接続することができる．

### 3.2 特徴分析

単音節に含まれている男女各 25 個の無声子音を目視した結果によれば，無声子音の区間長は男性より女性のほうが短く，また単音節の場合より短文の場合の方が短い．また，短文における無声子音の平均区間長は**図4**に示すように約 10～20 ms 前後であり，無声破裂子音（p, t, k）や無声破擦子音（ts, tʃ）は無声摩擦子音（s, h）より短い傾向を示す．また，無声破裂子音/t/の区間長は/p/および/k/の区間長より長くなる傾向がみられる．ただし，後続母音の影響による区間長のばらつきは認められる．ここで使用した音声試料は表2における 195 個の無声子音である．システムのフレーム長を 25.6 ms，遷移区間を 5 ms とし，図4の各無声子音の区間長が正規分布に従うものとすれば，/s/，/h/，/t/に関してはそれぞれ 66.3%，44.8% 23.3% の全情報が伝達されない可能性があるが，他の無声子音についてはほとんど全情報が伝達可能である．

音声波形の時間的変動は零交差レートの変動で表すことができ，一般に有声音の零交差レートに比べ無声子音の零交差レートの方が高い．表2の音声試料における CV（子音＋母音）型，VCV 型および CVC 型の零交差レートの一例を**図5**に示す．図5において，縦軸の零交差レートの理論値は0～1であり，横軸は約 3.25 秒に相当する 127 フレームを示す．また，破線は後述する3.3節の方法により抽出した TSIUVC 区間の前半部（12.8 ms）の零交差レートを示す．図5に示されているように有声子音や有声音が含まれているフレームの零交差レートに比べ無声子音が含まれているフレームの零交差レートの方が顕著に高いことがわかる．また，各無声子音の零交差レートは隣接母音の種類や個人差によりばらつきはあるが，いずれの場

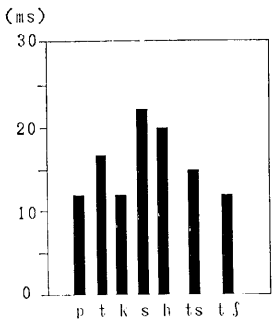



**図4** 無声子音の区間長
Fig. 4 Segment scale of the unvoiced consonants.

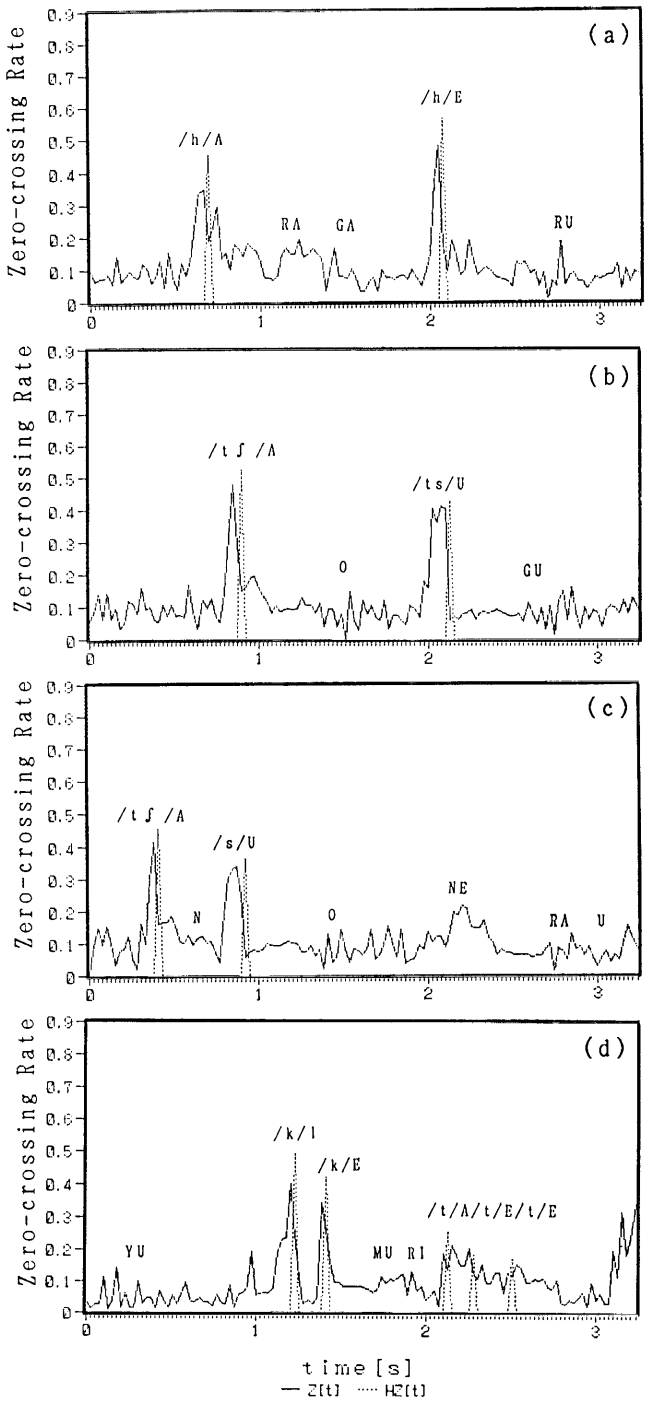



図 5 音声の零交差レート
Fig. 5 Zero-crossing rate of the speech.
(a)CV form, (b)CV form, (c)CVC form, (d)VCVC form.

合でも零交差レートが0.1以上でほとんどの場合は約 0.2〜0.6 であることを表3の音声試料から確認した．特に，図5(c)，(d)のように CVC 型あるいは VCVC 型では先行無声子音の零交差レートより後続無声子音の零交差レートの方が低くなる傾向を示す．また，舌の位置変動に対しては先行母音より後続母音の方が強く影響する[13]ことを考慮すれば，零交差レートの変動も先行母音より後続母音の方が強く影響するものと考えられる．

### 3.3 探索・抽出

有声音と TSIUVC とを判別させるパラメータに個別ピッチパルス，零交差レート，音声波形から求める音声エネルギーなどが考えられる．しかし，発話者が低い音圧レベルで発声した場合の有声音部や音声の立ち下がり部のような音声エネルギーの低いところが無声子音と判別される恐れがある．一方，個別ピッチパルスや零交差レートは音圧の高低に影響されにくく，1) 無声子音の零交差レートは高いが，有声音の零交差レートは低い，2) TSIUVC 区間には個別ピッチパルスが存在しないが，有声音区間には個別ピッチパルスが存在するといった特徴がある．このようなことを考慮すれば，実音声の音声波形から有声音と TSIUVC とを判別させるパラメータに個別ピッチパルスおよび零交差レートが有効であると考えられる．そして，上記の1)，2)のような特徴と個別ピッチパルスの位置が声帯振動の開始位置であることに着目すれば，無声子音と有声音が混在しているフレームから TSIUVC を容易に抽出することができる．この考え方に基づき，著者らは短文から TSIUVC を探索・抽出し，近似合成する方法の研究を進めてきた[15]．そのアルゴリズムを**図6**に示す．このアルゴリズムの処理手順は，まず短文から切出したフレームが有声音(V)か無音(S)かをピッチ係数 ($PC[t]$) により判別する．ピッチ係数はフレーム内に個別ピッチパルスが1個でも存在すれば1，そうでなければ0とする．すなわち，ピッチ係数が1なら有声

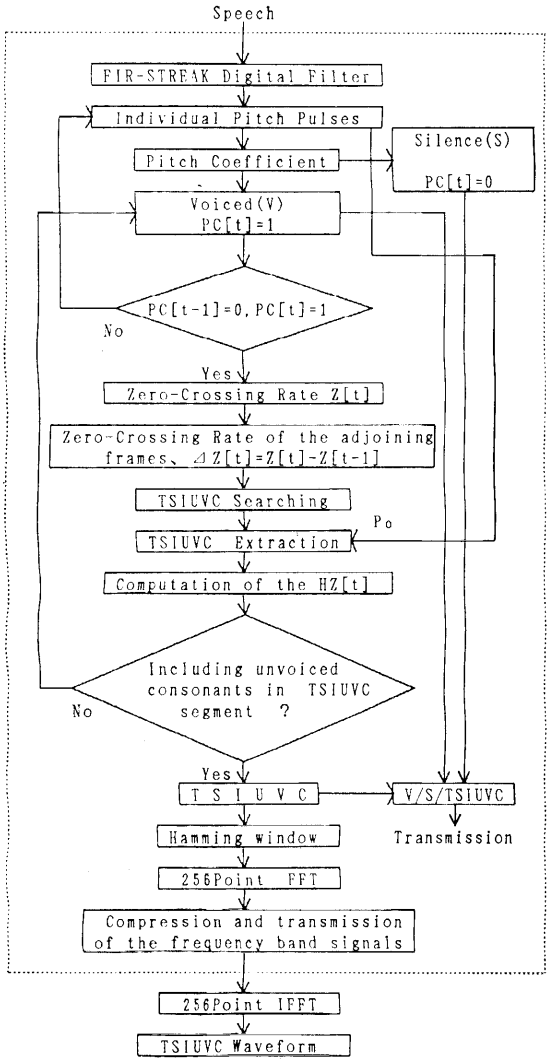



図 6 無声子音を含む遷移区間 (TSIUVC) の近似合成法
Fig. 6 Approximate-synthesis method of the TSIUVC.

音，0なら無音とする．次に，フレームが有声音と判別された場合，2.2 節のアルゴリズムにより求めた複数の個別ピッチパルスから最初の個別ピッチパルス ($P_0$) を声帯振動の開始点，すなわち TSIUVC 区間の終点とし，この終点から 25.6 ms 前の時点を TSIUVC 区間の開始点とする仮区間を求める．さらに，この仮区間に対し，零交差レート $Z[t]$ および隣接フレーム間の零交差レートの差 $\Delta Z[t]$ を求め，$\Delta Z[t]$ が負でかつ $Z[t-1] \geqq 0.1$ の条件を満たしているかを判別する．さらに，上記の仮区間の前半部 (12.8 ms) について零交差レート $HZ[t]$ を求め，$HZ[t] \geqq 0.128$ を満

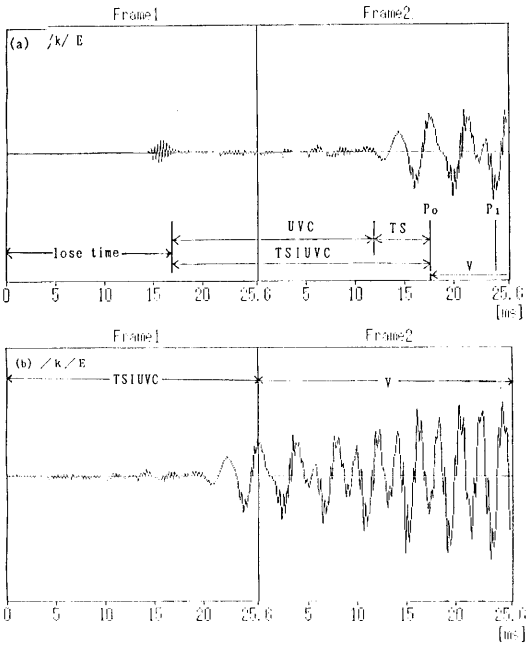



図 7 フレームの再作成
(a) 原フレーム，(b) 再成フレーム
Fig. 7 Reform of frames.
(a) Original frames, (b) Reformed frames.

たすかを判別する．これらの条件をすべて満たせば，上記の仮区間を TSIUVC 区間と決定するが，そうでなければフレーム全体を有声音 (V) と決定する．ここで，$t$ はフレーム番号である．そして，本手法では TSIUVC を探索・抽出するために 25.6 ms の遅延を施しており，**図 7** に示すように有声音と TSIUVC が混在しないようフレームを再作成している．すなわち，処理中のフレーム内に TSIUVC が存在すれば，TSIUVC 区間を1つのフレームとし，最初の個別ピッチパルス ($P_0$) の時点を次のフレームの始種り時点とする．この手法により TSIUVC 区間と有声音区間とを適切な方法により分析合成することができ，また伝送すべきパラメータの混在を防ぐことができる．

TSIUVC の抽出において，①本来の TSIUVC を抽出できなかった場合と，②本来は存在しないにもかかわらず抽出された場合とを抽出誤りと判定し，次式により TSIUVC の抽出率 ($AER_2$) を算出した．

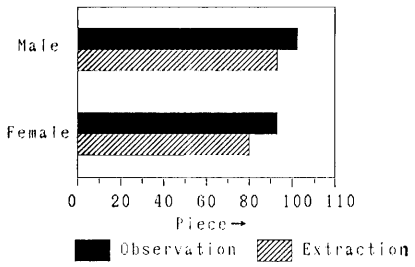
$$AER_2 = \sum_{j=1}^{N} (a_j - (|b_j| + |c_j|)) \Big/ \sum_{j=1}^{N} a_j \qquad (2)$$

ここで，$a_j, b_j, c_j$ はそれぞれ1文当たり真の TSIUVC 数と見なされる目視数，①の誤り数，②の誤り数を示す．**表 2** の音声試料を用いた結果，目視の TSIUVC

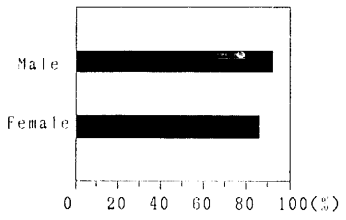
の数（2式の分母）は男声で102個，女声で93個であり，抽出したTSIUVCの数（2式の分子）は男声で93個，女声で80個であった．無声破裂/摩擦/破擦子音別にみると，目視によるTSIUVCの数は男声でそれぞれ50/39/13個，女声でそれぞれ46/34/13個であった．また，抽出したTSIUVCの数は男声でそれぞれ44/37/12個，女声でそれぞれ39/30/11個であった．したがって，無声破裂/摩擦/破擦子音を含むTSIUVCの抽出率は男声でそれぞれ88%，94.9%，92.3%，女声でそれぞれ84.8%，88.2%，84.6%であった．結局，男女声におけるTSIUVCの抽出率は男声で91.2%，女声で86%が得られた．これらの結果を図8に示す．

**表2** 音声試料 (TSIUVC)
Table 2 Speech sample (TSIUVC).

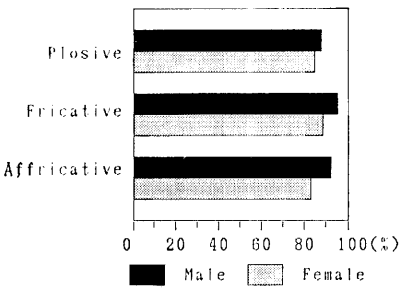
| 条件 | 男声 | 女声 |
|---|---|---|
| 発話者 | 9 | 9 |
| 発話時間 | 3.4秒 | 3.4秒 |
| 短文数 | 39 | 34 |
| 母音数 | 317 | 292 |
| TSIUVC数 | 102 | 93 |




(a) TSIUVCの数
(a) Number of TSIUVC




(b) 抽出率
(b) Extraction rate




(c) 抽出率
(c) Extraction rate

**図8** 無声子音を含む遷移区間 (TSIUVC) の抽出率
Fig. 8 Extraction rates of the TSIUVC.

### 3.4 情報圧縮・近似合成

TSIUVCは図7のように無声子音区間 (UVC) 遷移区間 (TS) とで構成されている．これらの信号は低ビットレート化のために情報圧縮する必要があり，そうすれば当然ながら波形に歪みが生じる．したがって，TSIUVCの情報圧縮と波形歪みとの関係を考慮しなければならない．そこで，無声破裂/摩擦/破擦子音を含むTSIUVCの信号を低歪みに再生するに当たって，どの周波数帯域が有効であるかについて簡単な実験を行った．実験に用いた音声は3.4 kHzのLow-Passフィルタで帯域制限し，10 kHzでサンプリング，12 bitで量子化したものである．まず，3.3節の方法により抽出した時系列のTSIUVC信号をフリーエ変換し，周波数分解能が39.0625 Hzである周波数

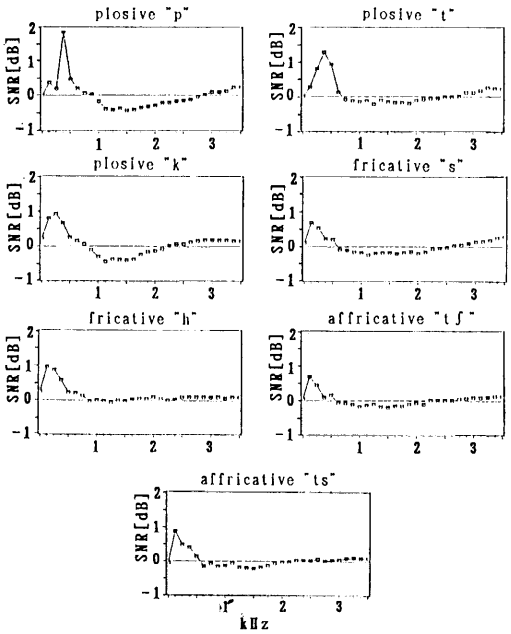



**図9** 周波数帯域におけるTSIUVCのSNR
Fig. 9 SNR of the TSIUVC for the frequency band.

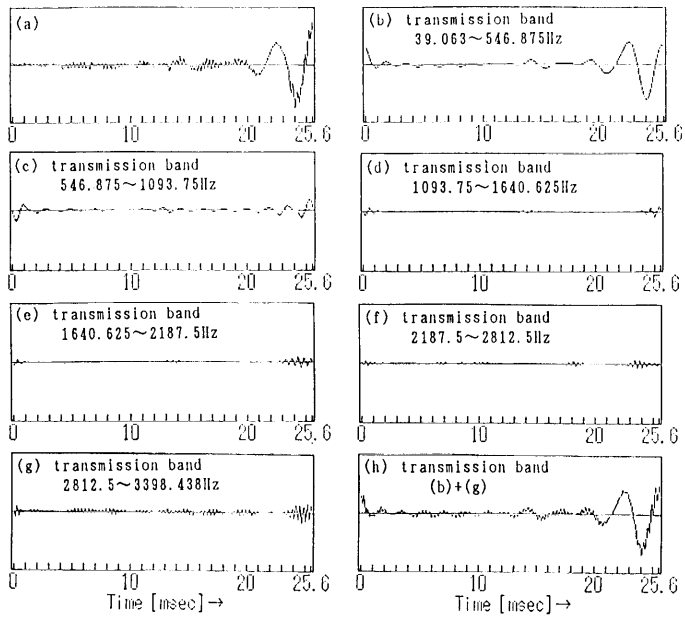



**図 10** 無声子音を含む遷移区間 (TSIUVC) の近似合成波形
Fig. 10 Approximate-synthesis waveforms of the TSIUVC. (a) Automatic extraction TSIUVC, (b)～(h) Approximate-synthesis waveforms of the TSIUVC.

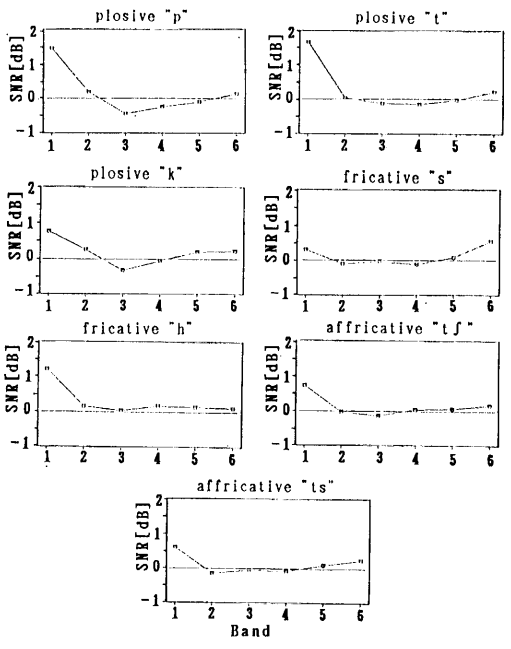



**図 11** 送信周波数帯域における TSIUVC の SNR
Fig. 11 SNR of the TSIUVC for the transmission frequency bands.

帯域信号を 117.1875 Hz ずつ細かく分割（例えば，39.0625～117.1875 Hz, 117.1875～234.375 Hz, …, 3281.25～3398.4375 Hz）した．次に，各周波数帯域信号を個別に逆フリーエ変換して時系列の TSIUVC 信号を再生した．最後に，原 TSIUVC の信号と再生した TSIUVC の信号との SNR の算出結果を**図 9** に示す．図 9 において，SNR 値は高周波数の領域に比べ低周波数の領域の方が高く，低歪みの TSIUVC を再生するには 0.586 kHz 以下と 2.813 kHz 以上の周波数帯域信号が有効とされる．その例に，0.039～0.547 kHz と 2.813～3.398 kHz とを用いて近似合成した波形を**図 10** の(b)と(g)に，これらの周波数帯域信号を用いた場合の近似合成波形を図 10 の(h)に示す．図 10 の(a)は図 7 (a)から抽出した TSIUVC の波形である．また，他の周波数帯域を用いた場合の近似合成波形と比較のため 0.547 kHz の帯域幅を高周波数の方にシフトさせて近似合成した波形をそれぞれ図 10 の(c)～(e)に示す．ただし，図 10 の(f)は(b)～(e)と(g)の余り帯域として帯域幅は 0.625 kHz である．図 10 において，0.039～0.547 kHz の周波数帯域信号を伝送した場合は遷移区間の信号が，2.813～3.398 kHz の周波数帯域信号を伝送した場合は無声子音区間の信号がよく近似されており，これらの周波数帯域信号を用いた場合が原波形によく近似されていることがわかる．この結果から，主な周波数情報として無声子音は高周波数帯域に，遷移区間の信号は低周波数帯域に分布していることがわかる．それから，図 10 の(b)～(g)における周波数帯域信号を別々に用いて再生した TSIUVC 信号と原 TSIUVC 信号との SNR 値を**図 11** に示す．この結果から，無声摩擦子音/h/を除く Band 1 (0.039～0.547 kHz) と Band 6 (2.813～3.398 kHz) で高い SNR を示している．ここで，使用した音声試料は表 2 から得た 34 種類の TSIUVC 信号 (p：5 個，t：5個，k：5 個，s：5 個，h：5 個，tʃ：5 個，ts：4 個) である．以上の実験結果により，TSIUVC 区間内の音声信号を低歪みに近似合成するためには 0.039～0.547 kHz および 2.813～3.398 kHz の周波数帯域

信号が有効であることを明らかにした．これらの周波数帯域信号は後述する4.2節の符号化条件により符号化し，合成側に伝送する．

## 4. システム

本章では，TSIUVC 近似合成法を用いたシステムと用いないシステムとの構成例および符号化条件を示す．音声は 3.4 kHz の Low-Pass フィルタで帯域制

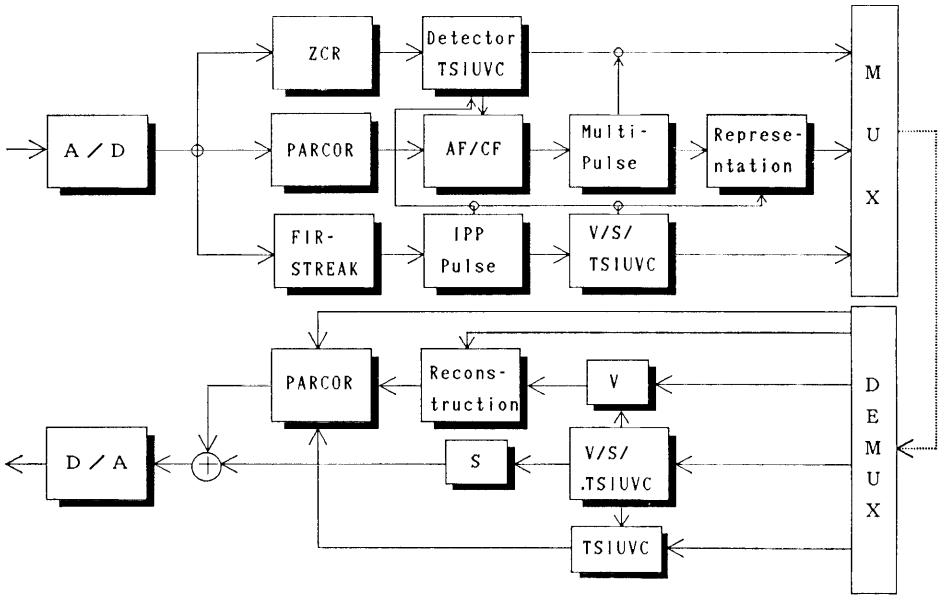



図 12 TSIUVC 近似合成法を用いない System(A)
Fig. 12 System(A) not using TSIUVC approximate-synthesis method.
ZCR: Zero-Crossing Rate, AF: Autocorrelation Function, CF: Cross-correlation Function.

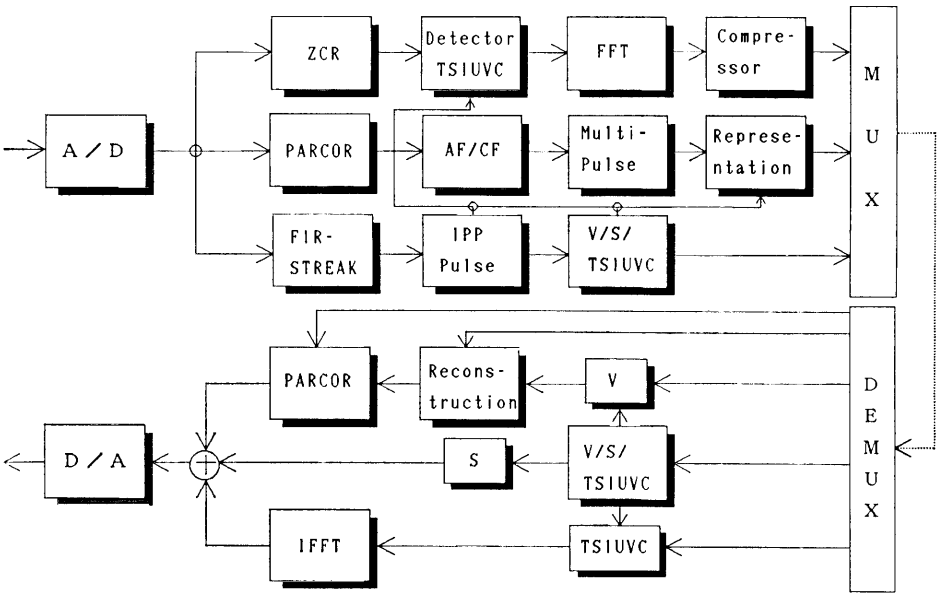



図 13 TSIUVC 近似合成法を用いた System(B)
Fig. 13 System(B) used TSIUVC approximate-synthesis method.

限した後，10 kHz でサンプリング，12 bit で量子化した．本システムにおいて，分析合成フィルタは PARCOR ディジタルフィルタを用いた．

### 4.1 構　成

TSIUVC 信号の再生に無声音源および PARCOR 合成フィルタを用いたシステム (System (A)) と TSIUVC 近似合成法を用いたシステム (System (B)) とをそれぞれ**図12**と**図13**に示す．これらのシステムは3.3節の方法により決定した有声音(V)/無音(S)/TSIUVC の判別情報（有声音：1，無音：0，TSIUVC：2）を2ビットの符号で合成側に伝送し，これらの判別情報により有声音源や無声音源，TSIUVC 近似合成法が切替わる．判別情報がVの場合は System (A), (B) とも PARCOR 合成フィルタの励起音源にマルチパルスの有声音源を用いる．また，判別情報が TSIUVC の場合，System (A) では PARCOR 合成フィルタの励起音源にマルチパルスの無声音源を，System (B) では PARCOR 合成フィルタを使わずに3章の TSIUVC 近似合成法を用いる．そして，判別情報がSの場合は System (A), (B) とも音源および合成フィルタは使用せずに V/S/TSIUVC 判別情報だけを合成側に伝送して 25.6 ms の時間シフトを施す．

System (A), (B) に用いるマルチパルス音源の算出において，マルチパルスの振幅と位置を同時に決定するには膨大な演算量が必要となる．その対策として，Atal らは原信号と合成信号との誤差電力を最小化するようにマルチパルスの振幅と位置を1つずつ決定する準最適な AbS (Analysis by Synthesis) 法を提案している[1]．そして，小澤らはマルチパルスの振幅および位置を相関演算に基づいて算出する有効な方法も提案している[16]．前者の方法より後者の方法の方が演算量の面で有利であることから，マルチパルスの振幅および位置は後者の方法により算出した．また，有声音区間には声帯振動に伴う相似的な波形の繰返しが存在しているので，System (A), (B) とも有声音源に代表個別ピッチパルス区間内のマルチパルス列を用いた．この手法により効率的な符号化を実現している．また，System (A) の無声音源には TSIUVC 区間内の音声信号から算出した全マルチパルス列を用いた．ここで，マルチパルスの数は有声音区間のピットレートと同等になるよう制御してある．結局，System (A), (B) とも合成側に伝送するパラメータは PARCOR 係数，個別ピッチパルス，マルチパルスの振幅および位置，V/S/TSIUVC の判別情報である．

**表 3** System(A), (B)の符号化条件
Table 3 Coding condition of the system(A), (B) bit/frame.

| 条　　件 | System(A) | System(B) |
|---|---|---|
| フレーム長 (ms) | 25.6 | 25.6 |
| (TSIUVC 区間) | | |
| Switch | V/S/TSUVC | V/S/TSUVC |
| ビット割当 (bit) | 2 | 2 |
| 1周波当りの振幅値 (bit) | | |
| (実数部 & 虚数部) | | |
| 13周波 低域の12周波 | | 3 |
| 13周波 最大振幅の1周波 | | 7 |
| 高域の15周波 | | 3 |
| PARCOR 係数 | 7, 6, 5, 5, 4, | |
| $k_i$ ($i$=1～10) | 3, 3, 3, 2, 2 | |
| Multi-Pulse | | |
| 振幅，位置 | 3, 5 | |
| パルス数 | 17 | |
| 総ビット数 (bit) | 178 | 178 |
| kbit/s | 約 6.9 | 約 6.9 |
| (有声音区間) | | |
| Switch | V/S/TSUVC | V/S/TSUVC |
| ビット割当 (bit) | 2 | 2 |
| PACOR 係数 | 7, 6, 5, 5, 4, | 7, 6, 5, 5, 4, |
| $k_i$ ($i$=1～10) | 3, 3, 3, 3, 3 | 3, 3, 3, 3, 3 |
| Multi-Pulse | | |
| 最大振幅値 | 6 | 6 |
| 振幅，位置 | 4, 5 | 4, 5 |
| パルス数 | 10 | 10 |
| [個別ピッチパルス抽出法] | | |
| $P_0$, $P_1$ | 7, 7 | 7, 7 |
| $P_{i-1,i}$ ($i$=3～10) | 24(3×8) | 24(3×8) |
| 総ビット数 (bit) | 178 | 178 |
| kbit/s | 約 6.9 | 約 6.9 |

### 4.2 符号化条件

System (A), (B) の符号化条件を**表3**に示す．TSIUVC 区間と有声音区間のフレーム長は 25.6 ms，ピットレートは 6.9 kbit/s，PARCOR 分析合成フィルタの次数は 10 次とした．PARCOR 係数の変化がスペクトルの変化に及ぼす影響は低次の係数ほど，その変動によるスペクトルへの影響が大きく，高次になるにつれて変動の影響が小さい[17]．そこで，PARCOR 係数へのピット割当ては高次ほど少なくなるピット配分をした．そして，System (A) と System (B) の品質を公平に比較させるために有声音区間と TSIUVC 区間における総ビット数は同等にしてある．

System (A) の *TSIUVC* 区間において，使用したマルチパルスの数は 17 個であり，PARCOR 係数，

マルチパルス，V/S/TSIUVC の判別情報に割当てたビット数を集計すると有声音区間と TSIUVC 区間における総ビットは 178 bit となる．

System (B) の TSIUVC 区間において，低域周波数は高振幅の生起確率が高く，高域周波数は低振幅の生起確率が高い．したがって，低域周波数の振幅値と高域周波数の振幅値とを同一ビット数で符号化するには低域周波数の振幅値を圧縮する必要がある．そこで，低域周波数の振幅値のみ 1/3 の圧伸符号を行った．結局，符号・伝送すべき実数部および虚数部の周波数帯域信号は 3.4 節の方法により選定した 39.063～546.875 Hz における 13 周波（以下，低域の 13 周波），2812.5～3398.438 Hz における 15 周波（以下，高域の 15 周波）である．これらの周波数帯域信号に割当てた符号ビット数は，まず実数部および虚数部における低域の 13 周波からそれぞれ最大振幅値の周波数帯域信号（以下，最大振幅の 1 周波）を検出し，その周波数帯域信号にそれぞれ 7 bit を割当てた．また，実数部および虚数部における残りの 12 周波にはそれぞれ 3 bit を割当てた．次に，実数部および虚数部における高域の 15 周波にはそれぞれ 3 bit を割当てた．これらのビット数と V/S/TSIUVC の判別情報 (2 bit) とを合わせると総ビット数は 178 bit となる．

System (A), (B) の有声音区間において，個別ピッチパルスへのビット割当ては $P_{i-1,i}$ に 3 bit, $P_0$ と $P_1$ の個別ピッチパルスに 7 bit を割当てた．その他の個別ピッチパルスは $P_0$ および $P_1$ の間隔と $P_{i-1,i}$ から求める．ここで，ピッチ周波数の存在範囲は 80～370 Hz であることから時間間隔は $2.7\,\text{ms} \leqq L \leqq 12.5\,\text{ms}$ となる．したがって，25.6 ms のフレーム内に存在しうる個別ピッチパルスの数は最大 10 個とした．また，使用したマルチパルスの数は 10 個であり，マルチパルスの最大振幅値に 6 bit，その他のマルチパルスの振幅に 4 bit，マルチパルスの位置に 5 bit を割当てた．これらのビット数と PARCOR 係数，V/S/TSIUVC の判別情報に割当てたビット数を集計すると，有声音区間の総ビット数は 178 bit となる．

## 5. 通信音質の評価

本章では，System (A), (B) で合成した音声に対して客観・主観評価を行い，その結果について述べる．

### 5.1 客観評価

客観評価は次式に示す SNRseg を用いた．

$$\text{SNRseg}=\frac{1}{m}\sum_{i=1}^{m}(\text{SNR})_i \qquad (3)$$

$m$：個別ピッチパルスが存在するフレーム数

音声試料は男女各 2 名による 24 種類の短文（合計約 81.6 秒）を用いて客観評価を行った結果，**表 4** に示すように System (A) の SNRseg は男声で 13.8 dB，女声で 13.1 dB が得られた．また，System (B) では，男声で 14.1 dB，女声で 13.6 dB が得られた．したがって，System (A) に比べ System (B) の方が，男声で 0.3 dB，女声で 0.5 dB が改善された．しかし，TSIUVC 近似合成法による SNRseg の大幅な改善値は得られなかった．その理由に，

Ⅰ）短文において，有声音のフレームが占める割合に比べ TSIUVC のフレームが占める割合は極端に少ない．

Ⅱ）SNRseg は各フレームの SNR 値を統計的な平均値で表す．

ことがあげられる．すなわち，短文に含まれている少数フレームの波形歪みを抑制しても大幅な SNRseg の改善値は期待できない．言換えれば，短文に多数の無声子音が含まれるほど，従来の方法に比べ TSIUVC 近似合成法による SNRseg の改善は期待できる．

そして，32 bit Personal Computer を用いて System (A), (B) の処理時間を測定した結果，それぞれ 326 秒，329 秒であった．また，有声音 (V)/無音 (S)/TSIUVC におけるフレーム当たりの計算時間についてはそれぞれ約 2.7 秒/0.6 秒/3 秒であった．この計算時間は表 1 の 32 種類の短文における平均時間である．そして，System (A), (B) のプログラムの Step 数はそれぞれ 1205 Step, 1383 Step であり，使用した言語は MS-C (Microsoft C) である．結局，System (A) に比べ step 数で 14.77%，計算時間で 0.92% の増である．System (B) において，原理的な時間遅れとして考えられることは TSIUVC を探索するための 25.6 ms の遅延や図 7 における lose time が挙げられる．最後に，本システムにおける計算時間については各アルゴリズムの最適化ならびに並列処理によりさらに減少できると考えられる．

**表 4** System(A), (B)の SNRseg
Table 4 SNRseg of the system(A), (B).

| | System(A) | | System(B) | |
|---|---|---|---|---|
| | 男声 | 女声 | 男声 | 女声 |
| SNRseg (dB) | 13.8 | 13.1 | 14.1 | 13.6 |

表 5 MOS 評価の尺度
Table 5 Measure of MOS estimation.

| 評価語 | 評点 |
|---|---|
| 非常に良い | 4 |
| 良い | 3 |
| 普通 | 2 |
| 悪い | 1 |
| 非常に悪い | 0 |

表 6 MOS 評価の結果
Table 6 Result of the MOS estimation.

| 方式 | MOS | |
|---|---|---|
| | 男声 | 女声 |
| System(A) | 1.61 | 1.53 |
| System(B) | 1.99 | 1.75 |
| 6 bit logPCM | 2.88 | 2.9 |
| 5 bit logPCM | 1.82 | 1.83 |
| 4 bit logPCM | 1.08 | 1.09 |

### 5.2 主観評価

System (A), (B) の主観評価には表5のような5段階の MOS (Mean Opinion Score) 尺度を用いた．受聴実験は被験者にヘッドホンから System (A), System (B), 4 bit, 5 bit, 6 bit logPCM の再生音をそれぞれ2回ずつ提示し，音声の品質を評価するようにした．被験者は音声を専門としない男性8名であり，学習効果の影響を避けるために合成音の聞き慣れていない者のみ選んだ．音声試料は客観評価に用いた試料のうち男女各2名の4種類の短文を使用した．表6に評価の結果を示す．System (A) の MOS 値は男声と女声でそれぞれ 1.61 と 1.53, System (B) の平均評点は男声と女声でそれぞれ 1.99 と 1.75 が得られた．従って，System (A) に比べ System (B) の方が音質改善されており，System (B) における男声は 5.1 ビット logPCM の音質に，女声は 4.9 ビット logPCM の音質に相当するものである．

主観評価によれば，男声および女声とも System (A) の音質より System (B) の音質の方がよいと評価されており，System (B) では女声より男声の方の音質がよい結果を得ている．音節知覚に主要とされる情報は，無声子音の開始時点から声帯振動の開始時点までの区間に存在する[14]ことを考慮すれば，System (B) は TSIUVC 区間内の音声信号を忠実に再現することで音声の品質を高めるシステムであるといえる．

## 6. まとめ

本論文では，TSIUVC 近似合成法を用いた新しいマルチパルス音声分析合成システムを提案した．

まず，FIR-STREAK ディジタルフィルタを用いた個別ピッチパルスの抽出法は男声と女声でそれぞれ 96%，85% の高い抽出率が得られることを示した．

次に，TSIUVC の探索・抽出に個別ピッチパルスおよび零交差レートを用いた結果，男声で 92.1%，女声で 86% の高い抽出率が得られた．TSIUVC 近似合成法は TSIUVC 区間内の音声信号を忠実に再現でき，これにより音質が改善されることを明らかにした．

さらに，System (A) と System (B) で合成した音声の品質を SNRseg と MOS 値によって評価した結果，SNRseg 評価では System (A) に比べ System (B) の方が男声で 0.3 dB, 女声で 0.5 dB 程度の改善を得た．また，MOS 評価では System (A) に比べ System (B) の平均評点が男声で 0.38, 女声で 0.22 の改善値を得ており，System (B) における男声は 5.1 ビット logPCM に，女声は 4.9 ビット logPCM に相当する音質であった．これらの結果から，TSIUVC 近似合成法が聴覚的な音質改善に有効であることを明らかにした．

今後は，女声における個別ピッチパルスの抽出率や TSIUVC の抽出率の向上，システムの処理何間の短縮などについて検討を進める予定である．

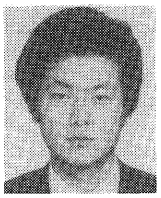
**謝辞** 本研究に対して有益な御助言を頂いた関根好文教授に感謝する．

## 参考文献

1) Atal, B. S. and Remde, J. R.: A New Model of LPC Excitation for Producing Natural-sounding Speech at Low Bit Rates, ICASSP, pp. 614-617 (1982).
2) 小澤一範，荒関 卓：ピッチ情報を用いる 9.6～4.8 kbit/s マルチパルス音声符号化方式，信学誌，Vol. J 72-D-2, No. 8, pp. 1125-1132 (1989).
3) Lawrence, R., Rabiner, L. R., Michael, J. Cheng, M. J., Rosenberg, A., McGonegal, C. A.: A Comparative Performance Study of Several Pitch Detection Algorithms, *IEEE*, Vol. ASSP-24 (1976).
4) Hodgson, L., Jernigan, M. E., Wills, B. L.: Nonlinear Multiplicative Cepstral Analysis for Pitch Extraction in Speech, *IEEE*, Vol. S 4 b. 11 (1990).
5) Hedelin, P. and Huber, D.: Pitch Period

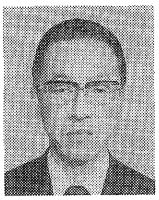
Determination of Aperiodic Speech Signals, *IEEE*, Vol. S6 b. 4 (1990).
6) Wise, J. D., Caprio, J. R. and Parks, T. W.: Maximum Likelihood Pitch Estimation, *IEEE*, Vol. ASSP-24, No. 5 (1976).
7) 藤井健作：自己相関関数法による電話帯域音声のピッチ抽出法，信学技報，SP 87-65 (1987).
8) 真野 淳，小沢慎治：LPC 有声音残差のピッチ同期メル LSP 分析合成方式，信学誌，Vol. J71-A, No. 3, pp. 634-641 (1988).
9) 武田昌一，浅川吉章ほか：残差音源利用分析合成方式とマルチパルス法の基本特性の比較検討，信学誌，Vol. J73-A, No. 11, pp. 1735-1742 (1990).
10) Markel, J. D. and Gray, A. H.: Linear Prediction.
11) 森川博由，藤崎博也：極・零モデルに基づく無声破裂音の分析と特徴抽出，音響学会誌，Vol. 38, No. 9, pp. 550-557 (1982).
12) 田中和世：日本語無声摩擦子音の分析と自動識別，音響学会誌，Vol. 38, No. 6, pp. 330-338 (1982).
13) 桐谷 滋：日本語母音，子音調音の隣接音の影響による変動，音響学会誌，Vol. 34, No. 3, pp. 132-139 (1978).
14) LaRiviere, C., Winitz, H. and Herriman, E.: Vocalic Transitions in the Perception of Voiceless Initial Stops, *J. Acoust. Soc. Am.*, Vol. 57, pp. 470-475 (1975).
15) 李 時雨，高橋 寛：無声子音を含む遷移区間の探索/抽出/近似法について，1991 信学秋季全大 A-102 (1991).
16) Ozawa, K., Ono, S. and Araseki, T.: A Study on Pulse Search Algorithms for Multipulse Excited Speech Codenr Realiiozat, *IEEE, Journal on Selected Areas in Communication*, Vol. SAC-4, No. 1 (1986).
17) 北脇信彦，板倉文忠ほか：PARCOR 形音声分析合成系における最適符号構成，信学誌，Vol. J61-A, No. 2, pp. 119-126 (1978).

(平成5年2月12日受付)
(平成6年3月17日採録)



**李 時雨**（正会員）

1987 年韓国東国大学校・工・電子卒業．1994年日本大学大学院理工学博士後期課程電子修了．博士（工学）．ディジタル信号処理，音声符号化方式，音声認識などの研究に従事．電子情報通信学会，システム制御情報学会，IEEE 各会員．



**高橋 寛**

昭和 34 年日本大学理工学部電気卒業．大学院を経て昭和 39 年日本大学理工学部勤務．現在，電子工学科教授．工学博士．列車運行システム，フェイルセイフシステム，交通情報システムなどの研究に従事．電子情報通信学会，電気学会，応用磁気学会各会員．