



HOKKAIDO UNIVERSITY

Title	音声符号化システムにおけるコンパクトARMAラティスフィルタの有効性について
Author(s)	広重, 真人; 宮永, 喜一; 栃内, 香次
Citation	北海道大學工学部研究報告, 153, 91-102
Issue Date	1990-11-29
Doc URL	https://hdl.handle.net/2115/42261
Type	departmental bulletin paper
File Information	153_91-102.pdf



音声符号化システムにおける コンパクト ARMA ラティスフィルタの有効性について

広重 真人 宮永 喜一 栃内 香次

(平成 2 年 8 月 31 日受理)

Effectiveness of a Compact ARMA Lattice Filter in a Speech Coding System

Makoto HIROSHIGE, Yoshikazu MIYANAGA and Koji TOCHINAI

(Received August 31, 1990)

Abstract

In this report, a new speech coding method using a compact ARMA lattice filter is proposed. This report mainly discusses some differences between the proposed method and a conventional coding method using PARCOR. In comparison, spectrum parameters and residual error signals are quantized at the same bit rate. First, we compared their performances of parameter coding without sophisticated coding schemes for residual signals. Next, we applied a CELP coding method to a compact ARMA lattice coding system. The results of experiments show that the proposed method is superior to the conventional method in the segmental S/N ratio and in audible experiments of synthesized speech, especially, when the lattice parameters are quantized at low bit-rate.

1. ま え が き

近年、音声情報処理の分野では、移動体通信や音声によるデータベース、ボイスメール等の需要の高まりにともなって、低ビットレートで高品質な音声符号化・合成方式の研究が盛んに行なわれている。代表的な符号化・合成方式としてはマルチパルス符号化⁽¹⁾、CELP⁽²⁾等がある。

これらの符号化方式のほとんどは、音声信号のスペクトル包絡を表わすモデルとして AR モデルを用い、モデルのパラメータは LPC⁽³⁾ を基礎とする一括処理型の分析手法を用いている。

AR モデルは音声のフォルマントのみを表わすモデルであるが、鼻子音等の音声にはアンチフォルマントも含まれており、これらの音声ではアンチフォルマントがスペクトルの傾きやフォルマントの位置に影響を及ぼすため、AR モデルでは十分にスペクトル包絡を表現できない。そこでより正確なスペクトル推定には、アンチフォルマントも表現できる ARMA モデルを用いる必要があると考えられる。

音声の符号化・合成では、スペクトル包絡を表わす予測器のパラメータと、予測誤差を符号化して伝送する。スペクトル推定の精度が不十分な場合、残差信号には音源の情報に加えて、推定

しきれなかった音声生成フィルタの情報の一部が加わることになり、残差信号の持つ情報量が増大するため、高品質の合成音声を得るためには残差信号に多くのビット数を割り当てる必要があると考えられる。一方高精度のスペクトル推定が行なわれれば、残差信号に含まれるのは音源の情報のみとなり、少ないビット数で符号化しても高品質の合成音声を得られると考えられる。言い替えると、残差信号に関する符号化ビット数が同じである場合、AR モデルによる符号化システムよりも、ARMA モデルを用いた符号化システムの方が、高品質な合成音声を得られる可能性があるといえる。

本報告では、音声符号化・合成システムにおけるスペクトル推定アルゴリズムとして、既に提案されているコンパクト ARMA ラティスフィルタ⁽⁴⁾を導入する。このフィルタは最少のパラメータを持つラティス型 ARMA モデルに基礎をおいており、パラメータは適応的アルゴリズムによって 1 観測サンプルごとに求められる⁽⁴⁾。

本報告では、コンパクト ARMA ラティスフィルタを用いた音声符号化システムを従来の PARCOR⁽⁹⁾ による符号化システムと比較することにより、音声生成モデルの違いによる符号化システムの性能の差を検討する。

2 章では、コンパクト ARMA ラティスフィルタについて簡単に紹介する。3 章では本報告で提案するシステムと従来のシステムとの違いについて、一般的な音声符号化・合成システムのブロック図を用いて説明する。4 章では単純な線形量子化を用いて、スペクトル推定フィルタの残差波形を量子化し符号化・合成実験を行う。実験結果より、特に鼻子音において、提案するコンパクト ARMA ラティス符号化システムを用いると高品質の合成音声を得られることを示す。5 章ではさらに、残差波形に対してピッチ構造を表現する長期予測フィルタと CELP 符号化方式を導入し、実用的なビットレートを持つコンパクト ARMA ラティス符号化・合成システムを構築する。いくつかのビットレートを設定して実験を行ない、コンパクト ARMA ラティス-CELP 符号化システムは PARCOR-CELP システムとほぼ同等の性能を持つことを示す。

2. コンパクト ARMA ラティスモデル

既に提案されているコンパクト ARMA ラティスモデル⁽⁴⁾は、最少のパラメータでフォルマントとアンチフォルマントの両方を表現できるラティス型の音声生成モデルである。パラメータはラティス基本区間 1 つにつき 1 個のみ必要とされる。このモデルは文献(4)に示される適応的アルゴリズムによって同定される。

モデルが安定であるとき、各区間のラティスパラメータの絶対値は 1 より小さくなる。そのためこれらのラティスパラメータは、トランスヴァーサル型 ARMA モデルの係数よりも符号化に向いているといえる。

コンパクト ARMA ラティスモデルおよびその同定アルゴリズムの詳細は文献(4)に記されている。

3. 音声符号化システム

一般的な音声符号化システムの概略図を図 1 に示す。入力音声波形 $y(k)$ はスペクトル包絡推定フィルタ（短期予測逆フィルタ）を通過することによってスペクトルの包絡成分が除去され、残差信号 $x(k)$ となる。フィルタの係数は量子化され合成器側へ伝送される。残差信号 $x(k)$ は適当な符号化手法で効率よく符号化された後、合成器側へ伝送される。図 1 において、5 章で述べる長期予測フィルタは、残差符号化部に含めて考えるものとする。

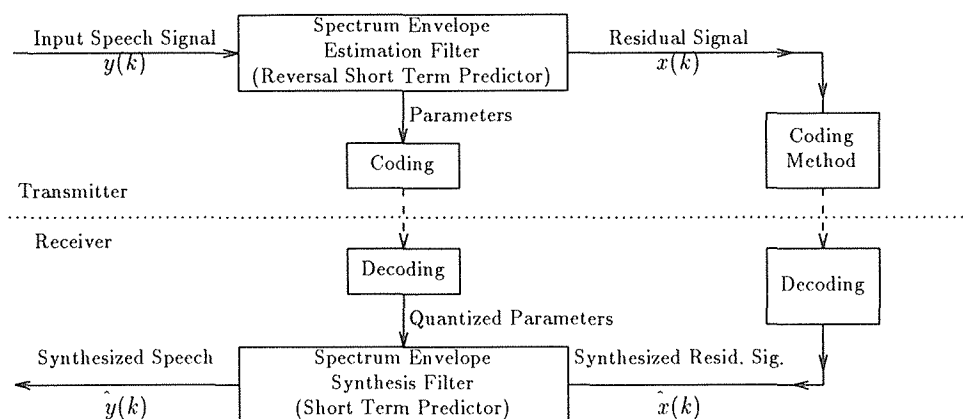


Fig. 1 Block diagram of a speech coding system.

従来の音声符号化システムのほとんどは、図1におけるスペクトル包絡推定フィルタに、PARCOR等のARモデルに基づくフィルタを用いている。本報告ではこのスペクトル推定フィルタに、ARMAモデルに基づくコンパクトARMAラティスフィルタを導入することにより、スペクトル推定の精度を向上し、再生された合成音声の品質向上をはかる。

4. 残差を単純に量子化した場合

本章では図1の残差符号化部で用いる符号化手法として、単純な線形量子化を用いた場合について考察する。コンパクトARMAラティスフィルタから得られた残差波形と従来のPARCORフィルタで求めた残差波形をそれぞれ同じビット数で線形量子化することにより、残差の持つ情報量を等しくすることができ、再生された合成音声の品質からスペクトル包絡推定アルゴリズムの性能を評価することが可能となる。

4.1. 実験システム

本章で用いた実験システムを図2に示す。実験に用いた音声サンプルは成人男性話者1名によって発声された47都道府県名で、10kHzサンプリング、12bit-ADで作成されたものである。コンパクトARMAラティス、PARCOR共にパラメータの数は8個(ARMAラティス：AR 6次，MA 1次，追加型 1次，PARCOR：AR 8次)である。PARCORのフレーム周期は120サンプル(12msec)であり、適応信号処理であるコンパクトARMAラティスの係数は、係数に関する情報量をPARCORに合わせるため120サンプルの平均を用いる。

それぞれの係数は不均一ビット割当⁽⁵⁾によって量子化される。本実験で使用したビット割当のパターンを表Iに示す。

残差信号は120サンプル毎に求められるそれぞれのフィルタの係数を用いて、逆フィルタによって計算される。本章の実験では、残差は量子化前のフィルタ係数から求められる。得られた残差信号は、短区間(360サンプル=36msec)の極大値によって正規化され、極大値は6ビット、正規化された信号は3ビットで量子化される⁽⁶⁾。

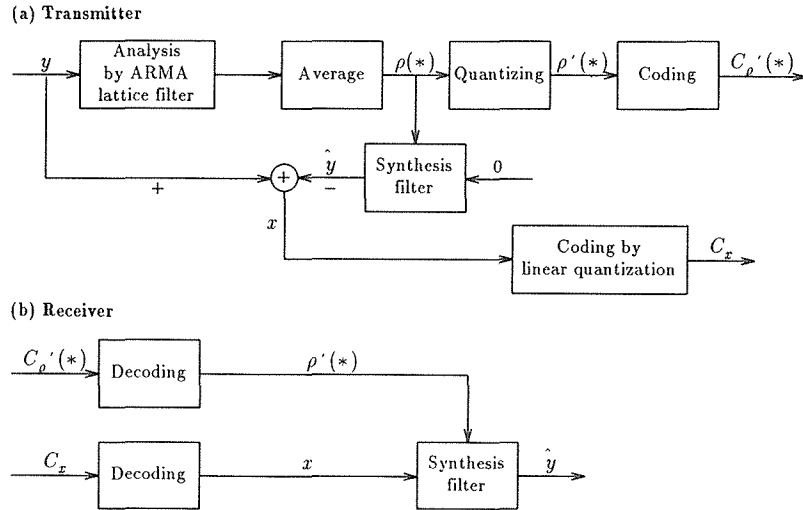


Fig. 2 Block diagram of a coding system with linear quantization.

TABLE I Bit patterns of nonuniform bit allocation

Name of bit pattern	Bit allocation								Total(bits/120 points)
bits.6.1-3	9	7	6	6	5	5	5	5	48
bits.6.1-5	7	5	4	4	3	3	3	3	32

4.2. 実験結果

実験結果の例を図3に示す。図3において、(A)はコンパクト ARMA ラティスによる合成音声、(B)はPARCORによる合成音声、(C)はセグメンタル SNR の比較グラフ、(D)はコンパクト ARMA ラティスのセグメンタル SNR 値から PARCOR のそれを引いた値で、正であればコンパクト ARMA ラティスの方が SNR の面ですぐれていることを示す。

図3によると、コンパクト ARMA ラティスフィルタを用いた符号化・合成システムは、信号パワーの極めて小さい部分を除くほとんどの部分で、PARCOR と同等、もしくは PARCOR よりすぐれたセグメンタル SNR を示している。特に、アンチフォルマントを含んでいると考えられる /N/・/m/等の鼻子音部では、1 次の MA パラメータを導入したことによりスペクトルの傾きがより正確に表現され、高いセグメンタル SNR が得られている。また、係数の量子化ビット数を少なくすると(図3(b))、PARCOR との差はさらに大きくなり、より多くの点で ARMA ラティスがすぐれた SNR を示す傾向があることがわかる。このことから、コンパクト ARMA ラティスフィルタは低ビット音声符号化に適していると考えられる。信号パワーの極めて小さい無音部分等では、わずかな雑音もセグメンタル SNR を大きく悪化させる。コンパクト ARMA ラティスによるシステムでは無音部分にわずかな信号が残る傾向があり、よい SNR が得られていないが、そのような無音部分のわずかな雑音は合成音声全体の品質にはあまり影響しないと考えられる。実際に合成音声を試聴した結果も、コンパクト ARMA ラティスによる合成音声が高い品質を示した。

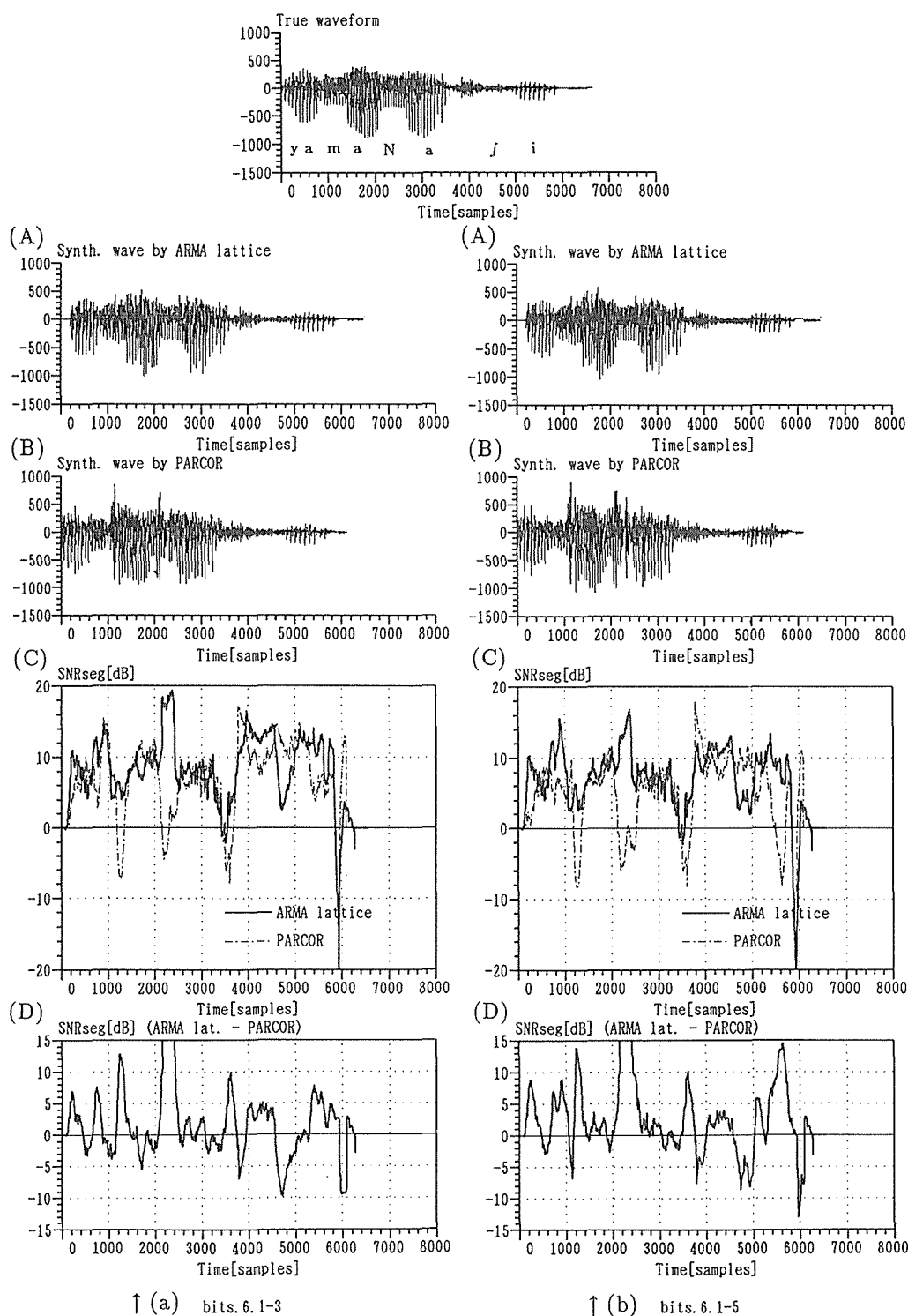


Fig. 3 Experimental results with linear quantizing method.

5. 長期予測フィルタと CELP の導入

本章ではより実用的な符号化システムとして、図 1 の残差符号化部に、長期予測フィルタと CELP 符号化方式を導入する。

5.1. 長期予測フィルタ

長期予測フィルタ⁽⁷⁾は波形の周期性を表現するフィルタであり、周波数領域では音声スペクトルの微細構造を表わしている。短期予測逆フィルタの出力である残差波形 $x(k)$ を長期予測逆フィルタに入力すると、残差波形に残っている周期性が取り除かれ、ガウス雑音に近い駆動波形 $d(k)$ が得られる。本報告で用いた長期予測フィルタの伝達関数は、

$$H_{LTP}(z) = \frac{1}{1 - bz^{-M}} \quad (1)$$

であり⁽⁸⁾、この伝達関数にはピッチ周期 M とピッチ係数 b の二つのパラメータがある。実音声(男性)の基本ピッチの範囲を考慮して、ピッチ周期 M の値を $20 \leq M \leq 200$ (10kHz サンプリング時で 50~500Hz) と制限する。誤差の 2 乗和を評価関数とし、それぞれの M の値について評価関数を最小にする b とそのときの評価関数値 V_M を求め、 V_M が最小となる M と b を長期予測フィルタ係数の推定値とする。実際の符号化・合成システムでは、 M と b は量子化されて伝送される。

5.2. CELP 符号化方式

長期および短期予測フィルタを縦続接続した合成フィルタに、前節までで得られた駆動波形を入力すると音声波形が再生されるが、合成フィルタのパラメータが量子化されている場合、符号化前の真の駆動波形を合成フィルタに入力しても真の音声波形は得られない。CELP 符号化方式⁽²⁾では、合成フィルタへの入力駆動波形を、ガウス雑音から作られたコードブックによって全く新しく生成する。この新生駆動波形で合成フィルタを駆動し、真の音声波形に最も近い合成音声を生成するコードワードを最適駆動波形として、対応するコード番号を伝送する。この方式によって、真の駆動波形を用いて合成した合成音声よりも、より真音に近い合成音声を得ることができる。

雑音コードブックは、形状(雑音コード番号) k と利得 g の 2 つのパラメータを持つ。これらのパラメータの推定は、まずコードワード長に等しい長さのフレームを考え、フレーム内の観測音声信号 y から過去のフレームの影響を取り除き y_{cur} を得る。利得 g を 1 とし、コード番号 k の雑音コードワードを駆動波形として合成フィルタを駆動して、合成波形 $\hat{y}_{k0}$ を得る。観測信号 y と合成信号 $\hat{y}_{k0}$ との誤差の 2 乗和を評価関数とし、 y_{cur} と $\hat{y}_{k0}$ から、評価関数を最小とする g と、そのときの評価関数の値 V_k が計算される。 g と V_k をすべての k について求め、 V_k が最も小さくなる k とそのときの g を最適パラメータとする⁽⁹⁾。実際の符号化・合成システムでは、 g は量子化されて伝送される。

5.3. ラティス係数から ARMA 係数への変換

上述の CELP 符号化においては、長期および短期予測フィルタの計算がコードブックサイズ分の回数だけ繰り返される。そのため短期予測フィルタに、計算回数の多いラティス型フィルタを用いるのは計算コストの面で不利である。そこで量子化したラティス係数を用いたラティスフィルタと同じスペクトル包絡を持つ ARMA 係数を回帰計算により求め、雑音コード探索計算は ARMA

係数を用いたトランスヴァーサル型フィルタを用いて行なう。

5.4. 聴覚重み付けフィルタ

聴覚重み付けフィルタ⁽⁷⁾はノイズシェイピングを行なう。人間の聴覚のマスクング現象のため、音声の周波数成分のうち、エネルギーが集中して大きなパワーのある部分には比較的大きな雑音が存在しても知覚されにくく、逆にパワーの小さい周波数では小さな雑音でもよく知覚される。観測実音声と合成音声との誤差の2乗和を単純に最小化することは、白色の誤差を最小化することと等価であるが、上述の知覚特性を考慮すると望ましくない。そこで、原音声のスペクトルと同様のスペクトルをもつ逆フィルタに誤差信号を通過させ、その出力信号の2乗和を最小にすることによって、原音声と同様のスペクトルを持った誤差信号を最小化することができる。このフィルタが聴覚重み付けフィルタであり、原音声と合成音声との誤差パワーが比較的大きくても聴覚的に高品質な合成音声を得ることができる。

聴覚重み付けフィルタの周波数特性は、原音声のスペクトル包絡をそのまま用いるよりも、極の帯域を広げて鈍らせたものが効果的であることが知られている⁽⁷⁾。

5.4.1. AR モデルの場合

原音声のスペクトル包絡がARモデルで表現されている場合、聴覚重み付けフィルタの伝達関数は、

$$H_{PWF\cdot AR}(z) = \frac{A(z)}{A(z/\alpha)} = \frac{1 + \sum_{i=1}^n \hat{a}_i z^{-i}}{1 + \sum_{i=1}^n \hat{a}_i \alpha^i z^{-i}} \quad (2)$$

で表わされる⁽⁷⁾。ただし $\hat{a}_i (i=1, \dots, n)$ は推定AR係数、 α は極の広帯域化の度合を表わす定数 ($\alpha \leq 1.0$) である^{(7), (9)}。

5.4.2 ARMA モデルの場合

本節ではARMAモデルの聴覚重み付けフィルタを提案する。前節のAR聴覚重み付けフィルタが、AR逆フィルタと、極を広帯域化したAR順フィルタとの縦続接続になっていることに注目し、ARMA聴覚重み付けフィルタを、ARMA逆フィルタと、極および零点を広帯域化したARMA順フィルタとの縦続接続とする。すなわち、

$$H_{PWF\cdot ARMA}(z) = \frac{A(z)}{B(z)} \cdot \frac{B(z/\alpha)}{A(z/\alpha)} = \frac{1 + \sum_{i=1}^n \hat{a}_i z^{-i}}{1 + \sum_{j=1}^m \hat{b}_j z^{-j}} \cdot \frac{1 + \sum_{j=1}^m \hat{b}_j \alpha^j z^{-j}}{1 + \sum_{i=1}^n \hat{a}_i \alpha^i z^{-i}} \quad (3)$$

である。ただし $\hat{a}_i (i=1, \dots, n)$ は推定AR係数、 $\hat{b}_j (j=1, \dots, m)$ は推定MA係数、 α は極および零点の広帯域化の度合を表わす定数 ($\alpha \leq 1.0$) である。

5.5. 実験システム

本章で用いた実験システムを図4に示す。計算コスト減少のため、聴覚重み付けフィルタは最適雑音コード探索ループの外側になるように移動されている⁽⁹⁾。本章の実験システムでは、短期および長期予測フィルタの残差信号は量子化後のパラメータを用いて計算される。

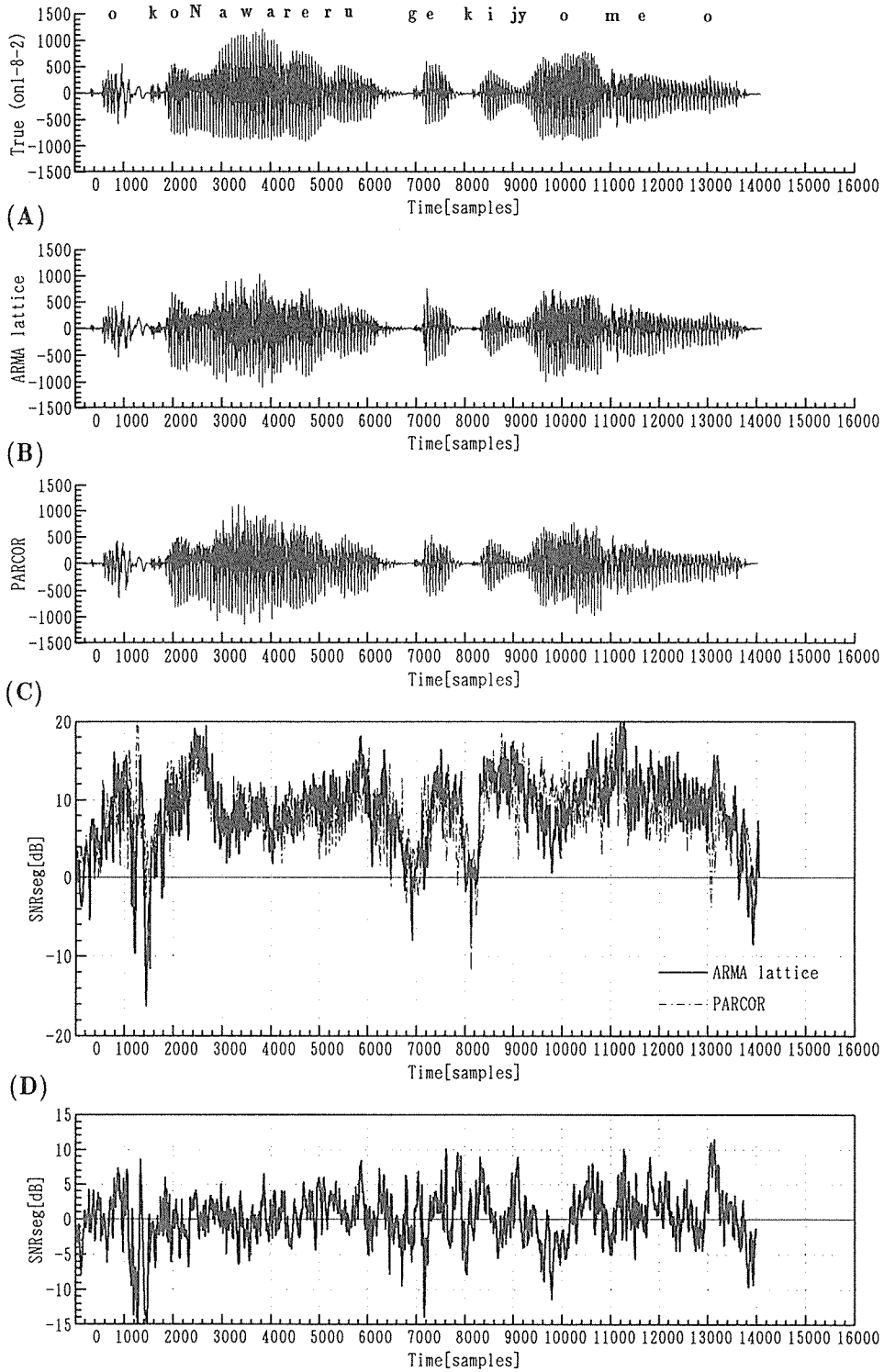


Fig. 5 Experimental result with CELP (Conditions: TABLE II (a))

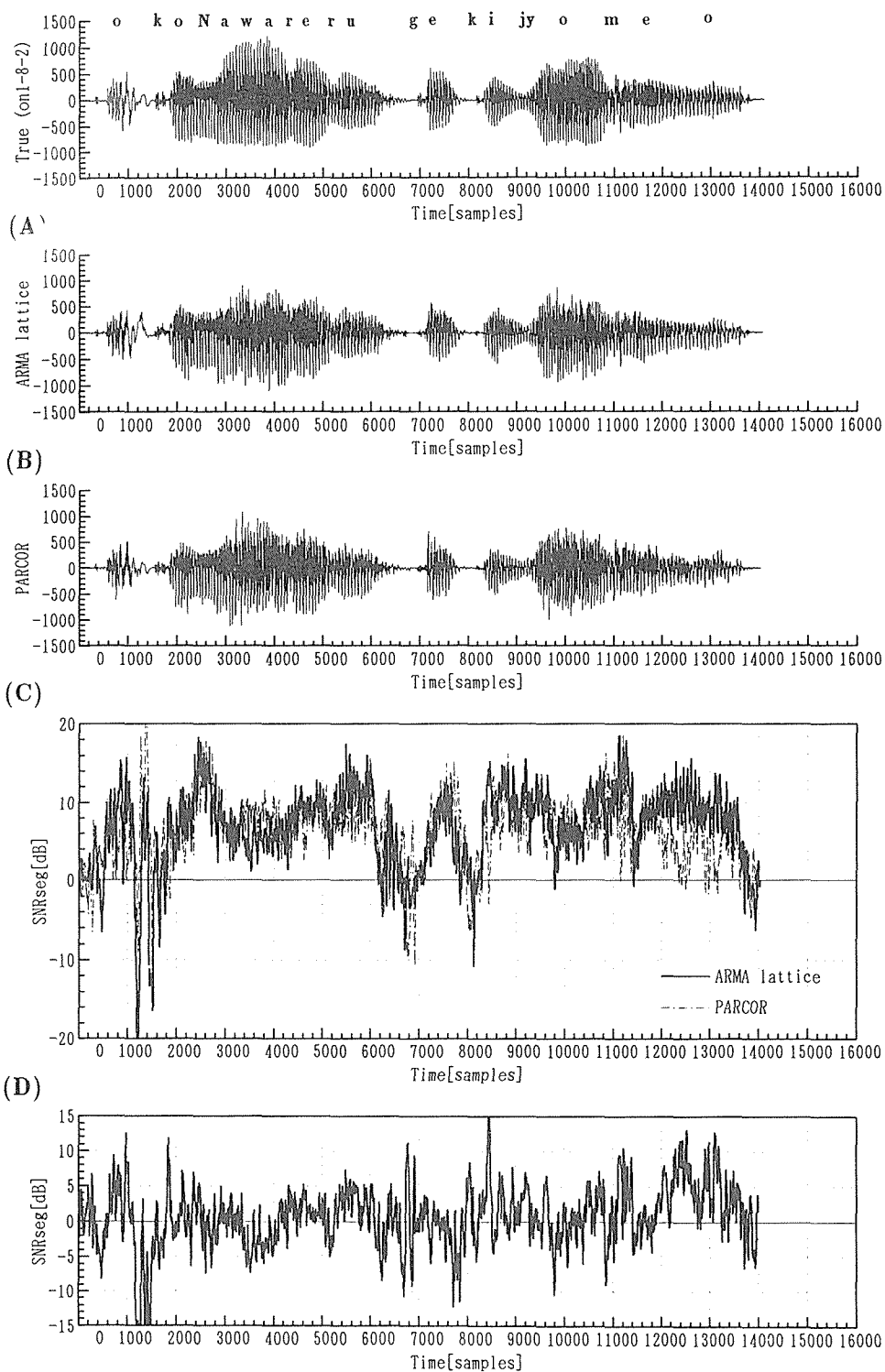


Fig. 6 Experimental result with CELP (Conditions: TABLE II (b))

表II (a)のパラメータによる実験結果を図5, 表II (b)によるものを図6に示す。図5・6の(A)～(D)の各グラフは4章の図3と同様である。

図5・6によると, 提案するコンパクトARMAラティス-CELP符号化システムがすぐれている点と, PARCOR-CELPシステムがすぐれている点が一貫して入れ替わり, 優劣が目だつた規則性はあまりない。全体を平均的にみると, ARMAラティスとPARCORはほぼ同等の性能を示しているといえる。この傾向は異なるデータサンプル間, 異なるビットレート間でも保たれる。ARMAラティスのシステムでは4章の場合と同様に無音部にわずかな信号が残る傾向があり, SNRを悪化させているが, これは4章と同様に全体の品質にはさほど影響しないと考えられる。合成音声を試聴した結果も, 2つのシステムで品質はほぼ同等である。

4章における実験でARMAラティスはPARCORよりもかなりよい性能を示していたにもかかわらず, CELPを用いた本章の実験でほとんど差が認められない理由としては, 次のことが考えられる。まず, 長期および短期予測フィルタの係数が量子化されているため, 総当たりのCELPL符号化方式を用いると真の駆動波形を用いるよりもよい品質の合成波形を作り出すことができる。このことから, CELPが最適な駆動波形を新しく生成することによってARモデルとARMAモデルとの差を埋めてしまったと考えられる。さらに, SNRに最も貢献するといわれる長期予測フィルタが, 2つのシステムで同じであることも理由としてあげられる。最後に, ARMA型の音声生成フィルタはその性質上AR型のものより多くのパラメータを必要とするが, 本報告の符号化・合成システムでは, ビットレートの関係から, むやみに分析次数を上げてパラメータの数を増やすことができず, 特長である高いスペクトル推定精度を充分実現できないことも理由の一つである。

6. む す び

本報告では, コンパクトARMAラティスフィルタを用いた音声符号化システムを構築し, 従来のPARCORによる符号化システムと比較することによって, 音声生成モデルの違いによる符号化システムの性能の差を検討した。まず, 単純な線形量子化を用いて, スペクトル推定フィルタの残差波形を量子化し, 残差波形の情報量をそろえることによって, ARMAラティスフィルタと従来のPARCORフィルタとの性能比較を行なった。実験結果より, 特に鼻子音において, 提案するコンパクトARMAラティスフィルタがすぐれていることが示された。さらに長期予測フィルタとCELP符号化方式を導入し, 実用的なビットレートを持つコンパクトARMAラティス符号化・合成システムを構築した。本システムを用いた実験結果によると, コンパクトARMAラティスを用いた符号化システムは従来のPARCORによるシステムとほぼ同等の性能を持つことが示された。今後の課題としては, ビットレートの低減, 分析次数の問題, 長期予測フィルタの性能向上, さらに適応信号処理の特性を生かした符号化方式の検討等があげられる。

謝辞

本研究を進めるにあたり, 御協力頂いた電子機器工学講座の森助手, 須戸助手, 斎川技官に感謝いたします。また, 御討論・御協力頂いた電子機器工学講座の院生諸氏に感謝いたします。

文 献

- (1) B. S. Atal and J. R. Remde: "A New Model of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates", IEEE Proc. ICASSP-82, pp. 614-617 (Apr. 1982).
- (2) M. R. Schroeder and B. S. Atal: "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates", IEEE Proc. ICASSP-85, pp. 937-940 (1985).
- (3) 斎藤収三, 中田和男: "音声情報処理の基礎", オーム社 (1981).
- (4) 宮永喜一, 栃内香次: "ARMAラティスマデルの等価回路記述について", 信学論(A), J73-A, 6, pp. 1048-1057 (1990-06).
- (5) 北脇信彦, 板倉文忠, 斎藤収三: "PARCOR形音声分析合成系における最適符号構成", 信学論(A), J61-A, 2, pp. 119-126 (昭53-02).
- (6) 王剛, 三木信弘, 永井信夫: "ARMAモデルを用いた音声の一合成法に関する音質評価", 信学論(A), J71-A, 12, pp. 2111-2119 (1988-12).
- (7) B. S. Atal: "Predictive Coding of Speech at Low Bit Rates", IEEE Trans. Commun., COM-30, 4, pp. 600-614 (Apr. 1982).
- (8) 熊野聡, 森島繁生, 原島博: "CELP音声符号化の改良の一検討", 第4回デジタル信号処理シンポジウム講演論文集, pp. 297-302 (1989-12).
- (9) I. M. Trancoso and B. S. Atal: "Efficient Search Procedures for Selecting the Optimum Innovation in Stochastic Coders", IEEE Trans. Acoust., Speech & Signal Process., 38, 3, pp. 385-396 (Mar. 1990).