論　文

HMMに基づいた極低ビットレート音声符号化

広井　順[†*]　徳田　恵一[†]　益子　貴史[††]　小林　隆夫[††]
北村　正[†]

Very Low Bit Rate Speech Coding Based on HMMs

Jun HIROI[†*], Keiichi TOKUDA[†], Takashi MASUKO[††], Takao KOBAYASHI[††], and Tadashi KITAMURA[†]

あらまし　本文ではHMM（Hidden Markov Model）に基づく極低ビットレート音声符号化について述べている．符号化器側では，HMMによる音素認識を行い，復号化器に音素インデックス列，状態継続長，ピッチ情報を伝送する．復号化器側では，音素インデックス列に従い音素HMMを連結する．そして，連結したHMMから，状態継続長に従って，ゆう度最大化基準による音声パラメータ生成アルゴリズムにより，メルケプストラム列を生成する．最後に，生成されたメルケプストラムを係数としてもつMLSA（Mel Log Spectrum Approximation）フィルタを，ピッチ情報に従って励振することによって合成音声を得る．主観評価実験の結果，ピッチ情報を除いて146 bit/s（26%の無音区間を含む）の提案方式により，同じくピッチ情報を除いて400 bit/s（8 bit/frame × 50 frame/s）のベクトル量子化に基づくボコーダと同等の性能を得ることができた．

キーワード　隠れマルコフモデル，MLSAフィルタ，音声符号化，極低ビットレート

1. まえがき

100ないし数百bit/s程度のビットレートで音声を符号化するためには，音素ボコーダ，あるいはセグメントボコーダを用いるのが最も一般的な方法である[1]〜[10]．これらの符号化法では，音声を音素単位（[1]〜[6]）または音響的なセグメント単位（[7]〜[10]）などの音声単位に分割し，得られた音声単位のインデックス列と継続長を復号化器側に伝送する．復号化器側では，伝送されたインデックスと継続長に従い，音声単位を連結することにより音声を合成する．

音素ボコーダ，セグメントボコーダにおける符号化器は一種の音声認識器とみなすことができるため，近年のHMM（hidden Markov model）に基づいた音声認識器の性能向上を考慮すると，HMMに基づいて音素ボコーダ，あるいはセグメントボコーダを構築することは一つの有効な方法と考えられる．実際に，文献[2], [3], [5], [6]では，HMMに基づいた音声認識器を符号化器として用いている．一方，復号化器は，音声合成器に対応するものであり，文献[6]では，符号化器と独立なホルマント型音声合成器を，文献[3], [5]では，LSPで表現された音素（ただし音素環境依存）の代表セグメントを接続することにより，音声を合成している．

本論文では，HMMに基づいた音声合成方式[11], [12]を利用することにより，符号化器・復号化器を通してHMMに基づいた音素ボコーダを構築し，その性能を評価することを目的とする．HMMに基づいた音声合成方式[11], [12]では，音声パラメータの静的及び動的特徴に関する統計情報に基づいて音声パラメータが生成されるため，音声単位の接続の際の，接続ひずみの問題が発生しにくいという特徴がある[11], [12]．本方式以外にも，符号化器，復号化器を通してHMMに基づいて音声符号化系を構成する手法[13]があるが，この手法は，LPC-VQに基づいた離散エルゴディックHMMを一つだけ用いて得られた状態系列を伝送し，復号化器では，伝送された状態系列から，ヒューリスティックなコードワード間の遷移確率（板倉・斉藤距離で近似）を用いることより，コードワードの列を選ぶというものである．したがって，本論文で対象とす

[†] 名古屋工業大学知能情報システム学科，名古屋市
Department of Computer Science, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya-shi, 466–8555 Japan
[††] 東京工業大学大学院総合理工学研究科，横浜市
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama-shi, 226–8502 Japan
[*] 現在，ソニー（株）コーポレートラボラトリー

る音素/セグメントボコーダとは異なる手法であり，むしろベクトル量子化の一手法と位置づけられる．

提案するボコーダではメルケプストラム分析 [14] により得られたメルケプストラム係数により音声スペクトルを表現し，各音声単位のメルケプストラム係数ベクトル列を音素 HMM によってモデル化している．符号化器側では，HMM に基づいた音素認識を行い，音素インデックスと状態継続長をエントロピー符号化あるいはベクトル量子化を用いて復号化器側に伝送する．また，合わせてピッチ情報を復号化器側に伝送する．復号化器側では，伝送された音素インデックス列に従い音素 HMM を連結し，伝送された状態継続長から状態列を決める．そして，メルケプストラム係数ベクトル列を，連結した HMM と状態列のゆう度が最大となるように定める [11], [12]．最後に，得られたメルケプストラム係数から MLSA（Mel Log Spectrum Approximation）フィルタ [15], [16] によって音声信号を合成する．従来の音素ボコーダでは，音声分析，音声単位のモデル化及び入力音声の音声単位へのセグメンテーション（音声認識），音声合成（音声単位の接続）において異なる音声表現，評価基準を用いていることが多いが，提案手法では，一貫して，音声をメルケプストラムにより表現しており，メルケプストラムの静的及び動的特徴に関するゆう度最大化基準により，音声単位のモデル化，入力音声のセグメンテーション，接続されたモデルからのメルケプストラム列の生成を行っていることを特徴としている．

従来の多くの音素/セグメントボコーダが音声単位の継続長を伝送しているのに対し，状態継続長を伝送している点も提案方式の特徴の一つである．主観評価実験により本符号化系の性能を評価しているが，特に，状態継続長の表現精度と主観性能の関係について明らかにするため，ピッチの量子化なしで実験を行う．

以下，**2.** では提案符号化法で用いられる HMM に基づいた音声合成法について簡単に述べ，**3.** では提案符号化方式について述べている．**4.** では提案符号化系に対する主観評価実験の結果を示した上で，**5.** において，セグメントボコーダとの関係などについて考察を加える．

2. HMM に基づく音声合成

図 1 に，提案する音声符号化方式のブロック図を示す．符号化器は HMM に基づく音素認識器に相当し，復号化器は，メルケプストラムに基づく分析合成

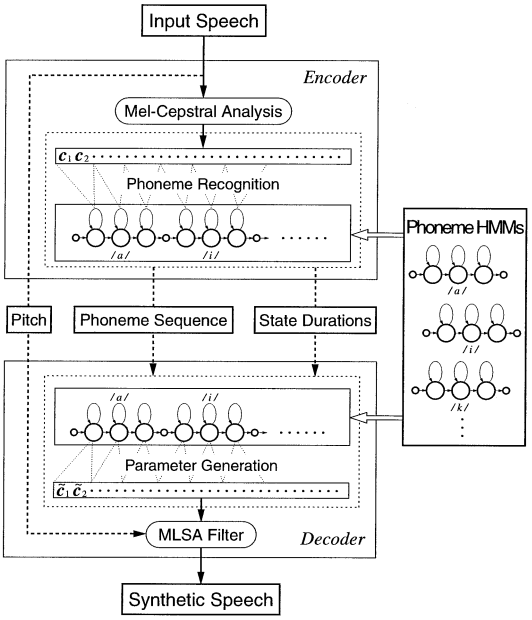

図 1　HMM に基づく極低ビットレート音声符号化システム
Fig. 1　Very low bit rate speech coding system based on HMMs.

法 [14]〜[16] と HMM に基づく音声パラメータ生成アルゴリズム [11] との二つの手法からなる音声合成方式 [12] を用い，符号化器と逆の操作を行う．本章では，これら二つの手法について簡単に述べる．

2.1 メルケプストラムに基づく音声分析合成

$M+1$ 次のメルケプストラム係数，つまり周波数変換されたケプストラム係数により，音声スペクトルをモデル化する．このとき，最小位相合成フィルタは次式で与えられる．

$$D(z) = \exp \sum_{m=0}^{M} c(m)\, \tilde{z}^{-m} \tag{1}$$

ここで，$\tilde{z}^{-1}$ は次式で定義されるオールパス伝達関数である．

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1 \tag{2}$$

標本化周波数が 10 kHz のとき，$\alpha = 0.35$ とすることにより，オールパス伝達関数の位相特性はメル周波数軸をよく近似する．メルケプストラム係数は，式 (1) に対数スペクトルの不偏推定のための評価関数 [17] を適用することにより決定される [14]．この手法は，入

力音声列 $\boldsymbol{x} = [x(0), x(1), \ldots, x(N-1)]'$ を平均 0 のガウス過程としたときには，ゆう度 $P(\boldsymbol{x} \,|\, \boldsymbol{c})$ を最大とするメルケプストラム係数 $\boldsymbol{c} = [c(0), c(1), \ldots, c(M)]'$ を求めることに等価である．ここで $\cdot'$ は転置を示す．$P(\boldsymbol{x} \,|\, \boldsymbol{c})$ は $\boldsymbol{c}$ に関して凸であることが示されるので，この最大化問題は [14] で述べられている反復法により効果的に解くことができる．また，$P(\boldsymbol{x} \,|\, \boldsymbol{c})$ を $\boldsymbol{c}$ に関して最大化することは，$D(z)$ 及び $1/D(z)$ のインパルス応答の時刻 0 での値が 1 であるという制約のもとで逆フィルタ出力のエネルギーを最小化することと等価である．

メルケプストラム係数から音声を合成するために，非有理式である式 (1) の伝達関数を実現しなければならない．幸い，MLSA フィルタ [15], [16] により十分な精度で $D(z)$ を近似することができる．MLSA フィルタは [15], [16] で示される特殊な構造をもつ IIR フィルタであり，通常の音声スペクトルを近似する際には安定となるよう実現することができる．また，MLSA フィルタの係数は，メルケプストラム係数から M 回の積和演算により得ることができる．したがって，MLSA フィルタにより，容易にメルケプストラム係数から音声を合成することができる．

2.2 HMM からの音声パラメータ生成

フレーム t でのメルケプストラム係数ベクトルを $\boldsymbol{c}_t$ とする．このとき，フレーム t の動的特徴 $\Delta \boldsymbol{c}_t$ と $\Delta^2 \boldsymbol{c}_t$，つまりデルタ，デルタデルタメルケプストラムは，それぞれ次のように計算される．

$$\Delta \boldsymbol{c}_t = \sum_{\tau=-L_1}^{L_1} w_1(\tau) \boldsymbol{c}_{t+\tau} \tag{3}$$

$$\Delta^2 \boldsymbol{c}_t = \sum_{\tau=-L_2}^{L_2} w_2(\tau) \boldsymbol{c}_{t+\tau} \tag{4}$$

$\Delta \boldsymbol{c}_t$，$\Delta^2 \boldsymbol{c}_t$ は，それぞれ，$\boldsymbol{c}_t$，$\Delta \boldsymbol{c}_t$ の時間方向に関する変化率を表すパラメータであり，ここでは，$L_1 = 1$，$L_2 = 2$ として，$w_1(-1) = -1/2$，$w_1(0) = 0$，$w_1(1) = 1/2$ 及び $w_2(-2) = w_2(2) = -1/4$，$w_2(-1) = w_2(1) = 0$，$w_2(0) = 1/2$ とした．フレーム t での音声パラメータベクトル $\boldsymbol{o}_t$ は，$\boldsymbol{o}_t = [\boldsymbol{c}_t', \Delta \boldsymbol{c}_t', \Delta^2 \boldsymbol{c}_t']'$ のように静的，動的特徴ベクトルからなる．

連続 HMM λ と状態列 $\boldsymbol{Q} = \{q_1, q_2, \ldots, q_T\}$ が与えられたとき，メルケプストラム係数ベクトル列 $\boldsymbol{C} = [\boldsymbol{c}_1', \boldsymbol{c}_2', \ldots, \boldsymbol{c}_T']'$ は，音声パラメータ列 $\boldsymbol{O} = [\boldsymbol{o}_1', \boldsymbol{o}_2', \ldots, \boldsymbol{o}_T']'$ に対する λ，$\boldsymbol{Q}$ のゆう度 $P(\boldsymbol{O} \,|\, \boldsymbol{Q}, \lambda)$ を，式 (3)，(4) の制約下で，$\boldsymbol{O}$ に関して最大化することにより得られる．各状態の出力分布を単一ガウス分布と仮定すれば，$P(\boldsymbol{O} \,|\, \boldsymbol{Q}, \lambda)$ の対数は次のように書くことができる．

$$\begin{aligned} &\log P(\boldsymbol{O} \,|\, \boldsymbol{Q}, \lambda) \\ &\quad = -\frac{1}{2}(\boldsymbol{O} - \boldsymbol{M})' \boldsymbol{U}^{-1} (\boldsymbol{O} - \boldsymbol{M}) - \frac{1}{2} \log |\boldsymbol{U}| \\ &\qquad + \mathrm{Const} \end{aligned} \tag{5}$$

ここで，

$$\boldsymbol{M} = [\boldsymbol{\mu}_{q_1}', \boldsymbol{\mu}_{q_2}', \ldots, \boldsymbol{\mu}_{q_T}']' \tag{6}$$

$$\boldsymbol{U} = \mathrm{diag}[\boldsymbol{U}_{q_1}, \boldsymbol{U}_{q_2}, \ldots, \boldsymbol{U}_{q_T}] \tag{7}$$

であり，$\boldsymbol{\mu}_{q_t}$ と $\boldsymbol{U}_{q_t}$ はそれぞれ，状態 q_t の平均ベクトルと共分散行列である．式 (3)，(4) の制約を考えないとき，$P(\boldsymbol{O} \,|\, \boldsymbol{Q}, \lambda)$ は $\boldsymbol{O} = \boldsymbol{M}$ のときに最大化されることは明らかである．これは，メルケプストラム係数ベクトルが共分散行列 $\boldsymbol{U}$ とは無関係に，平均ベクトルによって与えられることを意味する．

一方，式 (3)，(4) の制約下では，$\log P(\boldsymbol{O} \,|\, \boldsymbol{Q}, \lambda)$ を最大にするメルケプストラム係数ベクトル列 $\boldsymbol{C}$ は線形方程式 $\partial \log P(\boldsymbol{O} \,|\, \boldsymbol{Q}, \lambda) / \partial \boldsymbol{C} = \boldsymbol{0}$ によって定められる．この方程式は，文献 [11] の高速アルゴリズムにより容易に解くことができる．得られたメルケプストラム係数ベクトルは，静的及び動的特徴ベクトルの平均ベクトルだけでなく，それらの共分散行列によって定められることになる．

3. HMM に基づく極低ビットレート音声符号化

本章では，音声符号化システムの各部について簡単に述べる．

3.1 音声認識

今回の実験では対象話者は 1 名であり，音素 HMM の学習には ATR 日本語音声データベースの男性話者 MHT による音韻バランス 503 文を用いた．サンプリング周波数 10 kHz，フレーム周期 5 ms とし，長さ 25.6 ms の Blackman 窓を用い，メルケプストラム分析によりメルケプストラム係数を得た．特徴ベクトルは 0 次係数を含む 13 個のメルケプストラム係数とそれらの Δ，Δ^2 係数からなる．HMM は，スキップのない 3 状態 left-to-right 型のトライフォンモデルである．各状態は対角共分散をもつ単一ガウス分布によりモデル化される．音素ラベルには表 1 に示した 34 種類の

表 1　音素ラベルの種類
Table 1　Phoneme labels used in the system.

母音
a, i, u, e, o
子音
N, b, by, ch, d, f, g, gy, h, hy, j, k, ky, m, my, n, ny, p, py, r, ry, s, sh, t, ts, w, y, z
促音
cl
無音
sil

音素と無音を用い，各音素モデルを用意した．トライフォンモデルには決定木に基づくモデルクラスタリング法 [18] を適用し，その総分布数を約 1800 とした．

符号化器の音声認識器は日本語音素配列の制約（子音が二つ以上続かないなど）のみを利用した音声タイプライタ型の音素認識を行う．主観評価（**4.**）で用いたテストデータに対する音素認識率は 73.7%（挿入誤りを無視した場合は 88.7%）であった．テストデータ中の音素は 9.5 音素/s であるが，テストデータの認識結果では 11.7 音素/s となっている．

3.2　音素情報の符号化

音素認識により得られた音素列はエントロピー符号化を用いて伝送される．音素及び音素対の出現頻度は学習データを認識した結果を用いて計算した．音素の出現確率分布に基づいたハフマン符号化を使ったとき，テストデータにおける音素情報のビットレートは 54 bit/s となった．音素対の出現確率分布（音素バイグラム確率）を用いた場合には，46 bit/s のビットレートを得ることができた．

3.3　状態継続長の符号化

状態継続長を伝送するために，状態継続長をエントロピー符号化する方法（方法 1），音素継続長をエントロピー符号化し，復号化器で状態継続長に分割する方法（方法 2），状態継続長をベクトル量子化する方法（方法 3）の 3 通りを試みた．以下に，それぞれの方法について説明する．

［方法 1］

状態継続長をハフマン符号により伝送する．一つの音素モデルは，三つの状態継続長をもつため，一つの音素に対して，三つのハフマン符号語が伝送される．状態継続長は音素ごとに異なる分布をもつため，ハフマン符号表は音素ごとに用意し，切り換えて用いた（ただし，同一音素モデル内の三つの状態は区別せず，同じハフマン符号表を用いた）．

各音素のハフマン符号表は，学習データを **3.1** で述べた音声認識器で認識することにより得られた状態継続長の出現頻度に基づいて作成した．

［方法 2］

方法 1 では，状態継続長を伝送したのに対し，方法 2 では，音素継続長のみを伝送し，復号化器側でこれを三つの状態継続長に分割する．その際，音素継続長は，音素ごとに用意されたハフマン符号により伝送される．また，音素継続長の状態継続長への分解は，あらかじめ学習しておいた状態継続長に関する確率分布に基づいて行われる．つまり，状態継続長の分布をガウス分布で近似しておき，ゆう度最大化基準による方法 [11] で

$$d_k = m_k + \rho\sigma_k^2 \tag{8}$$

$$\rho = \left(T - \sum_{k=1}^{3} m_k\right) \Big/ \sum_{k=1}^{3} \sigma_k^2 \tag{9}$$

と分割を行う．ここで T は音素継続長であり，m_k，σ_k^2 はそれぞれ音素 HMM の第 k 状態における継続長分布の平均と分散である．

各音素の音素継続長に関するハフマン符号表は，学習データを **3.1** で述べた音声認識器で認識することにより得られた音素継続長の出現頻度に基づいて作成した．各状態継続長分布は，単一ガウス分布によりモデル化し，更に，各トライフォン HMM 内の三つの状態継続長分布を合わせて 3 次元ガウス分布とみなし，決定木に基づくモデルクラスタリング法 [18] を適用した．結果として，状態継続長モデルは約 1600 個の 3 次元ガウス分布で表現されている．

［方法 3］

各音素の三つの状態継続長を 3 次元のベクトルとみなし，ベクトル量子化する．VQ インデックスは，ハフマン符号化により伝送される．

VQ コードブックの学習は，学習データを **3.1** で述べた音声認識器で認識することにより得られた状態継続長に基づいて行った．**4.** の実験ではコードブックのサイズを 8（3 bit），32（5 bit），1024（10 bit）の 3 通りとした．

3.4　音 声 合 成

復号化器では，伝送された音素情報に対応するトライフォン HMM を連結し，得られた HMM からメルケプストラム係数ベクトル列を **2.2** で述べたアルゴリズムを用い生成する．ピッチ情報に従って生成したパルス列あるいは白色雑音で MLSA フィルタを励振す

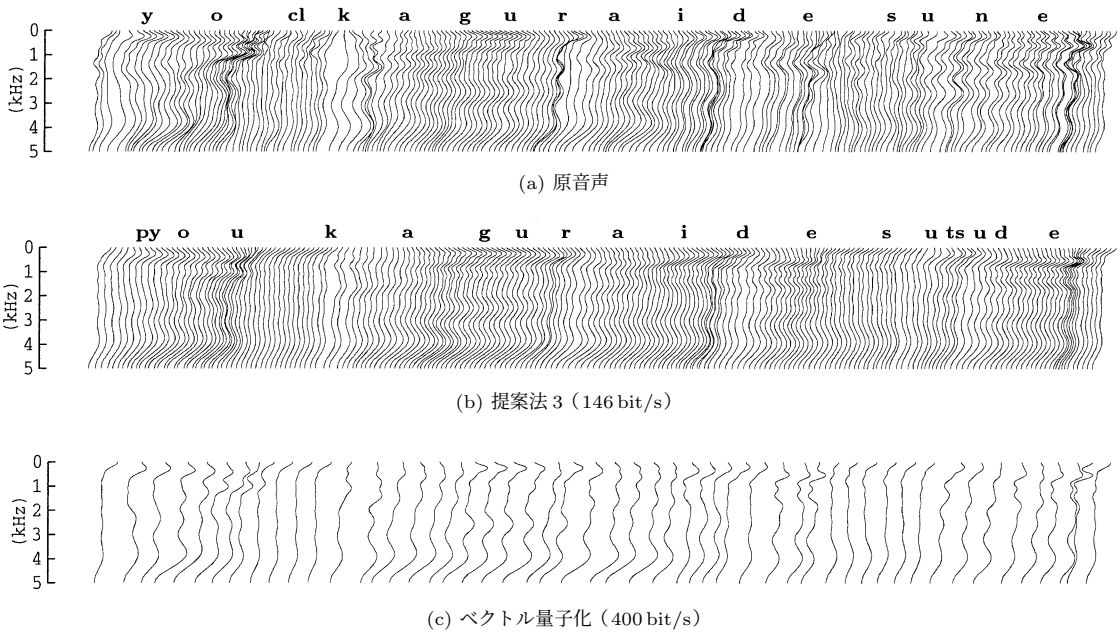

図 2　合成音声のスペクトル「4 日ぐらいですね」
Fig. 2　Spectra of synthesized speech.

ることにより，メルケプストラム係数から音声を合成する．

4. 実　　験

提案した音声符号化法の音声品質を評価するために，DCR（degradation category rating）試験 [19] を行った．DCR 試験では，被験者は，原音声に引続き，符号・復号化された音声を受聴し，原音声からの符号・復号化された音声の劣化の度合を 5 段階評価（1〜5 点で，点数が小さいほど劣化の度合が大きい）する．ある符号化方式に対して与えられた点数の平均が，その符号化方式の評価値 DMOS（degradation mean opinion score）となる．試験には，学習データに含まれていない 8 文章を用いた．被験者は 8 人の成人男性である．この実験では，ピッチは量子化せず，入力音声のピッチを音素 HMM に基づく Viterbi アライメントにより，時間的な位置合せを行った上で用いた．また，MLSA フィルタに与えるメルケプストラム係数は，従来法，提案法とも，フレーム間の直線補間により，0.1 ms（音声波形の標本化周期）ごとに更新した．

図 2 には，原音声のスペクトル，提案符号化法により復元されたスペクトル，メルケプストラム係数のベクトル量子化（400 bit/s = 8 bit/frame × 50 frame/s）によるスペクトルを例示している．図より，提案符号化法では，音素の接続部においても滑らかに変化するスペクトル列が得られていることがわかる．また，提案符号化法では認識誤りが含まれているにもかかわらず原音声に似たスペクトルを再現できていることがわかる．

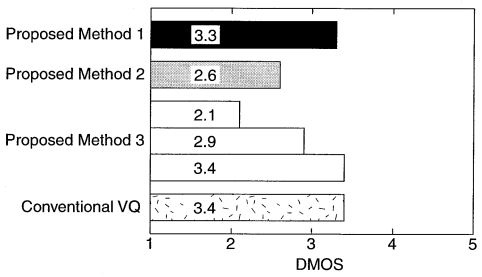

図 3　DMOS 値による主観評価
Fig. 3　Subjective performance measured by DMOS.

図 3 に，提案符号化法及びメルケプストラム係数のベクトル量子化によるメルケプストラムボコーダの DMOS 値を示す．提案法の方法 3 では状態継続長量子化のためのコードブックサイズを図中上から 3 bit，5 bit，10 bit とした．音素 HMM の学習に使われた音

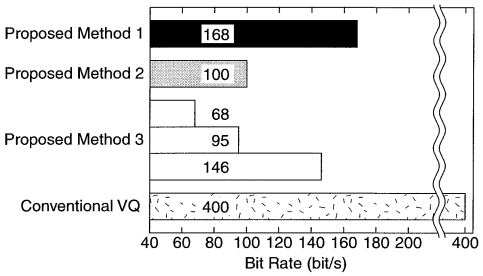

図 4　各手法のビットレート
Fig. 4　Bit rate for each method.

声データベースを VQ に基づくボコーダの VQ コードブックを学習するためにも用いた．それぞれの符号化法のビットレートを図 4 に示す．今回の実験では，いずれの符号化法ともピッチの量子化を行っていないため，提案法は，音素情報を伝送するために 46 bit/s を，残りを状態若しくは音素継続長を伝送するために使っていることになる．図 3，図 4 より，提案法においてはビットレートが高いものほど高い音声品質を示す傾向にあることから，提案法では，状態継続長の不正確な復元が符号化音声の品質を低下させていることがわかる．また，図より，146 bit/s の提案法の品質は，400 bit/s（= 8 bit/frame × 50 frame/s）のベクトル量子化に基づいたボコーダの品質と同等であることがわかる．なお，68 bit/s の提案符号化法も，音声品質は劣るものの符号化音声の了解性を保っている．

提案法の符号化遅延は以下のようにまとめられる．まず，符号化器で起こる遅延は音素認識器の探索戦略に依存し，一般に 100 ms のオーダで起こる．一方，復号化器においては，一つのトライフォン HMM を選択するために次の音素情報が必要なため，一つの音素継続長に対応する平均約 100 ms の遅延が起こる．また，復号化器での音声パラメータ生成アルゴリズムで，約 100 ms の遅延が生じる．したがって，本符号化系の符号化遅延は数 100 ms と見積もることができる．

5. 考　　察

前章の実験結果より，認識時に得られた状態継続長を，復号時にできるだけ精度良く再現することが音声品質の向上につながることが明らかとなった．また，提案法の方法 2 の性能があまりよくないことから，音素継続長のみを正確に再現しても，音素内の状態継続長を正しく再現しなければ，良い性能は得られないことがわかる．このこと以外にも，予備実験により次のような知見が得られた．

（1） 状態継続長，音素継続長，いずれも伝送せず，復号化器で各音素 HMM に付随した状態継続長分布をもとに各音素の状態継続長を決定した場合，認識誤りは，符号化音声の主観品質に重要な影響を与えるだけでなく，了解性を失わせる．

（2） 状態継続長を伝送したときには，認識誤りがあったことはほとんど知覚できない．ただし，主観品質は認識精度に比例する傾向にある．

認識誤りがあった場合，それは，正しい音素モデルより，誤った音素モデルの方が入力音声に対するゆう度が高かったことを意味しており，ゆう度の高さが聴感上の音声品質に対応しているとすれば，誤った音素モデルを用いて音声を合成した方が高い音声品質が得られることになる．つまり，認識誤りと聴感上の音声品質劣化との間には，必ずしも直接的な関連はないことになる．同様の議論は，文献 [2], [5] においても見ることができる．認識誤りが多くなるにつれ，音声品質が劣化するのは，認識誤り自体に原因があるのではなく，入力音声に対するゆう度が全体的に低くなっている，つまり入力音声とモデルとの適合の度合が低くなっていることに原因があると考えられる．ただし，以上は，復号器側で，認識時と同じ状態継続長を用いた場合の議論である．復号化器側で，それぞれの音素固有の状態継続長を用いた場合には，認識誤りも含めて認識された音素列どおりに音声が合成されることになり，発話の了解性が失われるのは当然である．

これらの考察から，本方式を含む音素ボコーダは，音素モデルに基づいて構築されているにもかかわらず，実際上，各モデルは音響的なセグメントモデルとして機能していることが理解される．ただし，音素環境依存の音素モデルとして学習を行っているため，セグメントボコーダにはない次のような特徴をもっている．

（1） 本方式では，音素環境依存の音素モデル（トライフォンモデル）を用いており，その数は約 15,000 種類（日本語音素配列の制約を考慮しても数千種類）となる．これらすべてをそのままセグメントを表すモデルと考えたときには，非常に多くのセグメントインデックスを用いることとなり，一つのモデルを指定するための情報量はかなり大きなものとなる．それに対して，本方式では，音素を表すインデックス（数十種類）の列を伝送するのみで，各セグメントに対応するトライフォンモデルを決定することができるという特

徴がある．これは，「トライフォンの列は，一つの音素の列を矛盾なく表現しなければならない」という形でモデル間の遷移に制約を設けたことと等価である．

（２） 認識時に日本語音素配列の制約を用いているため，音素インデックスを効率的に伝送することができる．例えば，子音の後には必ず母音が続くため，このような場合には，5 母音の中から一つを指定する情報量のみで音素を指定することができる（実際には，加えて音素バイグラム確率を利用することにより，伝送すべき情報を更に減少させている）．

（３） 論理的には約 15,000 種類（日本語音素配列の制約を考慮した場合，数千種類）あるトライフォンモデルを，そのまま学習することは，学習データ量の問題から不可能であるが，コンテクストクラスタリングの手法により，モデルの統計量は適切にタイイングされ，有限の学習データからモデルを構築することができる．

これらの特徴が本方式の性能向上に貢献していることが予想される．また，これらの特徴を残したまま，音響的なセグメントモデルとして再学習することにより，更に性能が向上することも期待できる．なお，文献 [7] のセグメントボコーダにおいては，ビットレートを下げるため，セグメント間の遷移に制約を設けており，この意味では類似した手法となっている．

6. むすび

HMM に基づいた音声認識・合成を用いた極低ビットレート音声符号化方式の構成について述べた．HMM からの音声合成は，メルケプストラムに基づいた音声合成フィルタと HMM に基づいた音声パラメータ生成アルゴリズムとからなっている．主観品質評価により，ピッチ情報を除いた場合，146 bit/s（無音区間 26%を含む）の提案法は，400 bit/s（＝ 8 bit/frame × 50 frame/s）のベクトル量子化に基づくボコーダと同等の品質をもつことを示した．また，68 bit/s の提案法でも符号化音声の了解性を保つことができた．音素認識器の性能改善により，音声品質が更に改善されることが期待される．

今後の課題として，HMM に基づいた話者適応化手法を利用した声質変換法 [20], [21] を用い，不特定話者用の符号化系を構築することが挙げられる [22]．また，考察で述べたように，音響セグメントモデルとして再学習することにより（例えば [23] と同様の学習法を用いる），性能向上を図ることも検討している．

謝辞 本研究の一部は文部省科学研究費補助金奨励研究（A）課題番号 08780333，09750399 による．

文　献

[1] R. Schwartz, J. Klovstad, J. Makhoul, and J. Sorensen, “A preliminary design of a phonetic vocoder based on a diphone model,” Proc. ICASSP-80, pp.32–35, 1980.

[2] J. Picone and G.R. Doddington, “A phonetic vocoder,” Proc. ICASSP-89, pp.580–583, May 1989.

[3] F.K. Soong, “A phonetically labeled acoustic segment (PLAS) approach to speech analysis-synthesis,” Proc. ICASSP-89, pp.584–587, May 1989.

[4] Y. Hirata and S. Nakagawa, “A 100 bit/s speech coding using a speech recognition technique,” Proc. EUROSPEECH-89, pp.290–293, Sept. 1989.

[5] C.M. Ribeiro and I.M. Trancoso, “Phonetic vocoding with speaker adaption,” Proc. EUROSPEECH-97, pp.1291–1294, Sept. 1997.

[6] M. Ismail and K. Ponting, “Between recognition and synthesis—300 bits/second speech coding,” Proc. EUROSPEECH-97, pp.441–444, Sept. 1997.

[7] S. Roucos, R.M. Schwartz, and J. Makhoul, “A segment vocoder at 150 b/s,” Proc. ICASSP-83, pp.61–64, 1983.

[8] Y. Shiraki and M. Honda, “LPC speech coding based on variable-length segment quantization,” IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-36, no.9, pp.1437–1444, Sept. 1989.

[9] P.A. Chou and T. Lookabaugh, “Variable dimension vector quantization of linear predictive coefficients of speech,” Proc. ICASSP-94, pp.505–508, April 1994.

[10] G. Baudoin, J. Cernocký, and G. Chollet, “Quantization of spectral sequences using variable length spectral segments for speech coding at very low bit rate,” Proc. EUROSPEECH-97, pp.1295–1298, Sept. 1997.

[11] 徳田恵一，益子貴史，小林隆夫，今井　聖，“動的特徴を用いた HMM からの音声パラメータ生成アルゴリズム，” 日本音響学会誌，vol.53, no.3, pp.192–200, March 1997.

[12] 益子貴史，徳田恵一，小林隆夫，今井　聖，“動的特徴を用いた HMM に基づく音声合成，” 信学論（D-II），vol.J79-D-II, no.12, pp.2184–2190, Dec. 1996.

[13] E.P. Farges and M.A. Clements, “Hidden Markov models applied to very low rate speech coding,” Proc. ICASSP-86, pp.433–436, 1986.

[14] 徳田恵一，小林隆夫，深田俊明，斎藤博徳，今井　聖，“メルケプストラムをパラメータとする音声のスペクトル推定，” 信学論（A），vol.J74-A, no.8, pp.1240–1248, Aug. 1991.

[15] 今井　聖，住田一男，古市千枝子，“音声合成のためのメル対数スペクトル近似（MLSA）フィルタ，” 信学論（A），vol.J66-A, no.2, pp.122–129, Feb. 1983.

[16] 徳田恵一，小林隆夫，深田俊明，今井　聖，“音声の適応メルケプストラム分析，” 信学論（A），vol.J74-A, no.8,

pp.1249–1256, Aug. 1991.
[17] 今井　聖，古市千枝子，“対数スペクトルの不偏推定，” 信学論（A），vol.J70-A, no.3, pp.471–480, March 1987.
[18] J.J. Odell, “The use of context in large vocabulary speech recognition,” Ph.D. Dissertation, Cambridge University, 1995.
[19] P. Kroon, “Evaluation of speech coders,” in Speech Coding and Synthesis, eds. W.B. Kleijn and K.K. Paliwal, Elsevier, Amsterdam, 1995.
[20] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, “Voice characteristics conversion for HMM-based speech synthesis system,” Proc. ICASSP-97, vol.3, pp.1611–1614, April 1997.
[21] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “Speaker adaptation for HMM-based speech synthesis system using MLLR,” Proc. Speech Synthesis Workshop, pp.273–276, Nov. 1998.
[22] T. Masuko, T. Kobayashi, and K. Tokuda, “A very low bit rate speech coder using HMM with speaker adaptation,” Proc. ICSLP-98, vol.2, pp.507–510, Nov. 1998.
[23] K.K. Paliwal, “Lexicon-building methods for an acoustic sub-word based speech recognizer,” Proc. ICASSP-90, pp.729–732, April 1990.

（平成 11 年 1 月 4 日受付，4 月 20 日再受付）

広井　順

平 9 名工大知能情報システム卒．平 11 同大大学院博士前期課程了（電気情報工学専攻）．在学中，音声符号化の研究に従事．現在，ソニー（株）コーポレートラボラトリー勤務．日本音響学会会員．

徳田　恵一　（正員）

昭 59 名工大・工・電子卒．平 1 東工大大学院博士課程了．同年東工大電気電子工学科助手．平 8 名工大知能情報システム学科助教授．工博．音声分析，音声合成・符号化，音声認識，ディジタル信号処理の研究に従事．日本音響学会，人工知能学会，IEEE 各会員．

益子　貴史　（正員）

平 5 東工大・工・情工卒．平 7 同大大学院博士前期課程了（知能科学専攻）．同年同大大学院総合理工学研究科物理情報システム創造専攻助手．音声の分析・合成，音声認識の研究に従事．日本音響学会，IEEE，ESCA 各会員．

小林　隆夫　（正員）

昭 52 東工大・工・電気卒．昭 57 同大大学院博士課程了．同年東工大精密工学研究所助手．同助教授を経て，現在同大大学院大学院総合理工学研究科物理情報システム創造専攻教授．工博．ディジタルフィルタ，音声の分析・合成，音声認識の研究に従事．日本音響学会，IEEE，ESCA 各会員．

北村　　正　（正員）

昭 48 名工大・工・電子卒．昭 53 東工大大学院博士課程了．同年東工大精密工学研究所助手．昭 58 名工大・工・電子工学科講師．昭 59 同助教授．平 7 名工大知能情報システム学科教授．工博．音声情報処理，マルチメディア情報処理の研究に従事．日本音響学会，IEEE，ESCA 各会員．