

HMM に基づいた極低ビットレート音声符号化における不特定話者への対応

益子 貴史[†] 小林 隆夫[†] 徳田 恵一^{††}

Very Low Bit Rate Speech Coding Based on HMM with Speaker Adaptation

Takashi MASUKO[†], Takao KOBAYASHI[†], and Keiichi TOKUDA^{††}

あらまし 本論文では、HMM を用いた音声認識、音声合成に基づく極低ビットレート音声符号化方式である HMM 認識ボコーダにおける不特定話者への対応手法について検討する。HMM 認識ボコーダでは、符号化音声の声質が復号化器において音声合成に用いる HMM のみによって決まるため、不特定の入力話者に対応するためには復号化器の HMM を入力音声に適應させる必要がある。そこで本論文では、音声認識により入力パラメータ列に対応づけられた HMM の出力分布列の平均ベクトルを、セグメントごとにパラメータ空間上で一様に並行移動させることにより、入力音声に適應する手法を提案する。この移動量をここでは移動ベクトルと呼び、符号化器においてこの移動ベクトルを求め、量子化して伝送する。主観評価実験により、提案手法で移動ベクトルを 100 bit/s 程度となるように量子化し不特定話者 HMM を適應した場合に入力話者の音声データで学習した特定話者モデルを用いた場合と同等の音質となることを示した。

キーワード 音声符号化、音素ボコーダ、HMM、不特定話者

1. ま え が き

近年、音声・オーディオ信号のデジタルデータ圧縮を行う高能率符号化法が各種、提案、標準化されている。例えば音声符号化の場合、誤り訂正を除いて 3.4 kbit/s の PSI-CELP 方式 [1] が携帯電話に利用され、更に低ビットレートの符号化方式として 2.4 kbit/s の MELP 方式 [2]、2.0 kbit/s の HVXC 方式 [3] などが標準化されている。しかし、地震やその他の災害などの非常時には、多くの人々が地域的、時間的に集中して通信を行おうとするため、通信回線が不足して固定電話、携帯電話ともに使用不能な状態に陥ってしまう。このような場合には、多少音声の品質を犠牲にしてもなるべく多くの通信回線を確保したいという要求があると考えられ、そのためには更に低ビットレートの音声符号化方式の開発が必要となる。

これまでに、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方式として、音素ボコーダあるいはセグメントボコーダが提案されている [4] ~ [14]。これらの符号化法では、符号化器において音声を音素 [4] ~ [10] や音響的なセグメント [11] ~ [14] 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と継続長を復号化器に伝送する。復号化器では、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。我々も HMM に基づく音声認識システムと我々の提案する HMM に基づく音声合成システム [15] を組み合わせた音素ボコーダ [16] (以下 HMM 認識ボコーダと呼ぶ) を提案し、ピッチ情報を除き、150 bit/s 程度のビットレートで 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が得られることを示している。

HMM 認識ボコーダの問題点の一つとして、符号化音声の声質が合成システムで用いる HMM によって決まるため、入力話者によらず一定となることが挙げられる。入力話者の集合が特定される場合には、それぞれの話者ごとにモデルセットを用意しておき、入力話者に応じて符号化、復号化に用いるモデルセットを切り換えることにより、復号化器で入力話者の声質をも

[†] 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, Yokohama-shi, 226-8502
Japan

^{††} 名古屋工業大学知能情報システム学科, 名古屋市
Faculty of Engineering, Nagoya Institute of Technology,
Nagoya-shi, 466-8555 Japan

つ音声を合成することができる。しかし、入力話者が不特定の場合には、すべての話者のモデルを用意することは不可能であるため、入力音声の声質に関する何らかの情報を復号化器へ伝送し、入力話者の話者性を反映するように復号化器の HMM を適応させる必要がある。

そこで本論文では、音声認識により入力パラメータ列に対応づけられた HMM の出力分布列の平均ベクトルを、セグメントごとにパラメータ空間上で一様に並行移動させることにより、入力音声に適応する手法を提案する。この移動量をここでは移動ベクトルと呼び、符号化器においてこの移動ベクトルを求め、量子化して伝送する。復号化器では伝送された移動ベクトルを HMM の出力分布列の平均ベクトルに加えることにより、入力音声へ適応する。移動ベクトルを求める方法として、入力パラメータ列とモデルとの最尤基準に基づく方法、入力パラメータ列と出力パラメータ列の 2 乗誤差最小基準及び最尤基準に基づく方法の 3 通りの方法について検討を行う。

2. 不特定話者対応 HMM 認識ボコーダ

図 1 に不特定話者対応 HMM 認識ボコーダのブロック図を示す。符号化器では、まず、音声信号からメルケプストラム分析 [19] によりメルケプストラムを求め、これとデルタメルケプストラムを合わせて特徴ベ

クトルとする。この特徴ベクトル列に対して HMM に基づく音素認識を行うことにより、音素列と状態継続長を求める。この際、特定話者 HMM を用いた場合には、学習に用いた話者と入力話者の組合せによって入力話者への適応の精度が影響を受けると考えられることから、不特定話者 HMM を用いる。次に、特徴ベクトル列と音声認識により得られた音素列から話者適応パラメータである移動ベクトルを求める。そして、これらとゲイン、ピッチ情報をそれぞれ符号化して復号化器に伝送する。

復号化器では、まず伝送された音素列と状態継続長から出力分布列を生成する。そして、移動ベクトルを出力分布の平均ベクトルに加えることにより分布列を入力音声へ適応し、ゆう度最大化基準に基づくパラメータ生成アルゴリズム [20] を用いてメルケプストラム列を生成する。これとゲイン、ピッチ情報を用い、MLSA フィルタ [21] により音声进行合成する。

2.1 ゆう度最大化基準に基づく HMM からのパラメータ生成

HMM 認識ボコーダの復号化器では、与えられた HMM のパラメータセット λ と、音素列と状態継続長列から定まる状態遷移列 $Q = (q_1, q_2, \dots, q_T)$ に対し、ゆう度 $P(O | Q, \lambda)$ を最大化する出力ベクトル O を生成し、音声合成に用いている。以下では、このゆう度最大化基準に基づくパラメータ生成アルゴリズム [20] について述べる。なお、簡単のため、各状態は単一ガウス出力分布をもつとする。

時刻 t ($1 \leq t \leq T$) の出力ベクトルを o_t として、 $O = [o'_1, o'_2, \dots, o'_T]'$ ($'$ は行列の転置を表す) と定義すると、 $P(O | Q, \lambda)$ の対数は、

$$\begin{aligned} \log P(O | Q, \lambda) \\ = -\frac{1}{2}(O - \mu)'U^{-1}(O - \mu) - \frac{1}{2}\log |U| \\ - \text{Const.} \end{aligned} \quad (1)$$

となる。ただし、

$$\mu = [\mu'_{q_1}, \mu'_{q_2}, \dots, \mu'_{q_T}]' \quad (2)$$

$$U = \text{diag}[U_{q_1}, U_{q_2}, \dots, U_{q_T}] \quad (3)$$

であり、 μ_{q_t} 及び U_{q_t} はそれぞれ状態 q_t の出力分布の平均ベクトル及び共分散行列である。また、Const. は出力分布の正規化係数の対数であり、 O 、 μ 、 U と独立な定数項である。

ここで、出力ベクトル o_t が、静的特徴量 $c_t =$

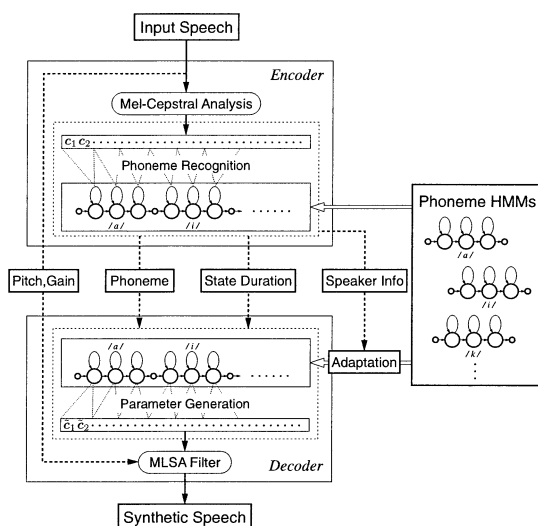


図 1 不特定話者 HMM 認識ボコーダ

Fig. 1 Speaker independent phonetic vocoder based on HMM.

$[c_t(1), c_t(2), \dots, c_t(M)]'$ (例えばメルケプストラム係数, ただし M は次数) と, 前後の複数フレームの静的特徴量から定まる動的特徴量 Δc_t からなる, すなわち, $\mathbf{o}_t = [c'_t, \Delta c'_t]'$ とする. $\Delta^{(0)} c_t = c_t$, $\Delta^{(1)} c_t = \Delta c_t$ とおけば,

$$\Delta^{(n)} c_t = \sum_{i=-L_-^{(n)}}^{L_+^{(n)}} w^{(n)}(i) c_{t+i}, \quad n = 0, 1 \quad (4)$$

と定義することができる. ただし, $L_-^{(0)} = L_+^{(0)} = 0$, $w^{(0)}(0) = 1$ とする. $w^{(1)}(i)$ は, 例えば

$$w^{(1)}(i) = \begin{cases} 1/2, & i = 0 \\ -1/2, & i = -1 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

と定義される.

$$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_T]' \quad (6)$$

$$\mathbf{w}_t = [\mathbf{w}_t^{(0)}, \mathbf{w}_t^{(1)}] \quad (7)$$

$$\begin{aligned} \mathbf{w}_t^{(n)} = & [\mathbf{0}_{M \times M}, \dots, \mathbf{0}_{M \times M}, w^{(n)}(-L_-^{(n)}) \mathbf{I}_{M \times M}, \\ & \quad \quad \quad (t-L_-^{(n)})\text{-th} \\ & \dots, w^{(n)}(0) \mathbf{I}_{M \times M}, \dots, w^{(n)}(L_+^{(n)}) \mathbf{I}_{M \times M}, \\ & \quad \quad \quad (t+L_+^{(n)})\text{-th} \\ & \mathbf{0}_{M \times M}, \dots, \mathbf{0}_{M \times M}]', \quad n = 0, 1 \quad (8) \end{aligned}$$

とすると, \mathbf{O} は \mathbf{W} と静的特徴量 c_t ($1 \leq t \leq T$) からなるベクトル $\mathbf{C} = [c'_1, c'_2, \dots, c'_T]'$ との積として表すことができる.

$$\mathbf{O} = \mathbf{W} \mathbf{C} \quad (9)$$

これを用いて式 (1) を \mathbf{C} について表すと,

$$\begin{aligned} \log P(\mathbf{O} | \mathbf{Q}, \lambda) \\ = -\frac{1}{2} (\mathbf{W} \mathbf{C} - \boldsymbol{\mu})' \mathbf{U}^{-1} (\mathbf{W} \mathbf{C} - \boldsymbol{\mu}) - \frac{1}{2} \log |\mathbf{U}| \\ - \text{Const.} \end{aligned} \quad (10)$$

となる.

このとき, $P(\mathbf{O} | \mathbf{Q}, \lambda)$ を最大化する $\overline{\mathbf{C}}$ は, $\mathbf{0}_{TM}$ を TM 次の零ベクトルとして,

$$\frac{\partial}{\partial \overline{\mathbf{C}}} \log P(\mathbf{O} | \mathbf{Q}, \lambda) = \mathbf{0}_{TM} \quad (11)$$

により得られる連立方程式

$$\mathbf{R} \overline{\mathbf{C}} = \mathbf{r} \quad (12)$$

ただし,

$$\mathbf{R} = \mathbf{W}' \mathbf{U}^{-1} \mathbf{W} \quad (13)$$

$$\mathbf{r} = \mathbf{W}' \mathbf{U}^{-1} \boldsymbol{\mu} \quad (14)$$

を解くことにより得られる [20].

3. 入力話者への適応

入力音声のフレーム数を T , フレーム t においてスペクトル分析によって得られた静的特徴量を x_t , 式 (4) を用いて求めた動的特徴量を Δx_t とし, フレーム t における特徴パラメータを $z_t = [x'_t, \Delta x'_t]'$, 音声認識により入力音声のパラメータ列 (z_1, \dots, z_T) に対応づけられた状態遷移列を $\mathbf{Q} = (q_1, q_2, \dots, q_T)$ とする. 復号化器では 2.1 で述べたアルゴリズムに従って \mathbf{Q} からスペクトルパラメータ列を生成するが, このとき, 入力パラメータ列とモデルとのミスマッチが大きいくほど入力パラメータ列と生成パラメータ列との誤差が大きくなり, 符号化音声の話者性や音質の劣化の原因となる. 特定話者モデルを用いた場合にはこのミスマッチはそれほど大きくはなくあまり問題とはならないが, 不特定話者モデルを用いた場合にはミスマッチが大きいため, 何らかの補正を行い, モデルを入力音声へ適応する必要がある.

入力音声への適応を行う際, 一般には入力音声の発話内容は未知であるため, この場合のモデルの適応は教師なし話者適応となる. 音声認識の分野では様々な教師なし話者適応手法が提案されているが, 認識誤りの影響により, 必ずしも十分な性能が得られるとは限らない. そこで本論文では, モデル全体を一括して適応するのではなく, 状態遷移列 \mathbf{Q} に対応する出力分布列を時間的なセグメントに分割し, セグメントごとに適応を行うこととする. また, 復号化器へ伝送することを考えると, 適応のための情報は量子化しやすいことが望ましい. そこで, 各セグメントの適応には, 当該セグメントに含まれるすべての出力分布をパラメータ空間上で一様に並行移動させることにより, 入力音声へ適応する手法を用いる. ここではこの移動量を移動ベクトルと呼び, この移動ベクトルの計算手法及び量子化コードブックの作成手法を以下に述べる.

3.1 モデル最ゅう基準

フレーム区間 $k \leq t \leq k+l-1$ において, 入力パラメータ列に対する HMM の出力分布列のフレームごとのゆう度の積が最大となるように移動ベクトルを求めることを考える [17]. すなわち, μ_{q_t} に移動ベクトル

ル m を加えることにより得られた新たな平均ベクトルを $\hat{\mu}_{q_t}$ としたとき、対数ゆう度

$$\begin{aligned} & \log P(z_k, \dots, z_{k+l-1} | q_k, \dots, q_{k+l-1}, \lambda, m) \\ &= \sum_{t=k}^{k+l-1} \log \mathcal{N}(z_t, \hat{\mu}_{q_t}, U_{q_t}) \\ &= -\frac{1}{2} \sum_{t=k}^{k+l-1} (z_t - (\mu_{q_t} + m))' U_{q_t}^{-1} \\ & \quad \cdot (z_t - (\mu_{q_t} + m)) \\ & \quad - \frac{1}{2} \sum_{t=k}^{k+l-1} \log |U_{q_t}| - lM \log(2\pi) \end{aligned} \quad (15)$$

を最大化するように m を求める。式 (15) を m で微分して 0_{2M} とおくことにより、

$$\begin{aligned} m &= \left(\sum_{t=k}^{k+l-1} U_{q_t}^{-1} \right)^{-1} \\ & \quad \cdot \left(\sum_{t=k}^{k+l-1} U_{q_t}^{-1} (z_t - \mu_{q_t}) \right) \end{aligned} \quad (16)$$

と求められる。この手法では、フレーム区間 $k \leq t \leq k+l-1$ における入力音声に対するモデルのゆう度を最大化していることから、以下ではモデル最ゆう基準と呼ぶことにする。

また、 $\mu_{q_t} = [\mu_{q_t}^{(0)'}; \mu_{q_t}^{(1)'}]'$, $U_{q_t} = \text{diag}[U_{q_t}^{(0,0)}, U_{q_t}^{(1,1)}]$ (ただし $\mu_{q_t}^{(0)}$, $U_{q_t}^{(0,0)}$ 及び $\mu_{q_t}^{(1)}$, $U_{q_t}^{(1,1)}$ はそれぞれ静的特徴量及び動的特徴量に対する平均ベクトルと共分散行列) とし、動的特徴量の適応を行わず、静的特徴量の平均ベクトルのみを移動させる場合、すなわち $\hat{\mu}_{q_t} = [(\mu_{q_t}^{(0)} + m^{(0)})'; \mu_{q_t}^{(1)'}]'$ としたとき、ゆう度を最大化する $m^{(0)}$ は、

$$\begin{aligned} m^{(0)} &= \left(\sum_{t=k}^{k+l-1} U_{q_t}^{(0,0)-1} \right)^{-1} \\ & \quad \cdot \left(\sum_{t=k}^{k+l-1} U_{q_t}^{(0,0)-1} (x_t - \mu_{q_t}^{(0)}) \right) \end{aligned} \quad (17)$$

により求められる。

3.2 2乗誤差最小基準

モデル最ゆう基準では、移動ベクトルを求める際の評価関数がモデルと入力パラメータ列とで定義されており、復号化器で生成されるパラメータ列と入力パラメータ列との直接的な評価関数とはなっていない。そ

こで、生成パラメータ列と入力パラメータ列との間で2乗誤差最小基準または最ゆう基準に基づく評価関数を定義し、この評価関数に基づいて移動ベクトルを求めることを考える。

入力音声から得られた静的特徴量の列からなるベクトルを $X = [x'_1, \dots, x'_T]'$ 、音声認識により対応づけられた状態遷移列 Q に従って生成されたパラメータ列を \bar{C} 、フレーム区間 $k \leq t \leq k+l-1$ に含まれる状態 q_t の出力分布の平均ベクトル μ_{q_t} に移動ベクトル m を加えて生成されたパラメータ列を \hat{C} とする。このとき、

$$E_{mse} = (X - \hat{C})'(X - \hat{C}) \quad (18)$$

を m に関して最小化することを考える。

ここで、

$$\begin{aligned} \hat{\mu} &= [\mu'_{q_1}, \dots, \mu'_{q_{k-1}}, \\ & \quad (\mu_k + m)', \dots, (\mu_{k+l-1} + m)', \\ & \quad \mu'_{k+l}, \dots, \mu'_{q_T}]' \end{aligned} \quad (19)$$

とすると、 \hat{C} は

$$R\hat{C} = \hat{r} \quad (20)$$

ただし、

$$\hat{r} = W'U^{-1}\hat{\mu} \quad (21)$$

である。式 (21) を式 (14) と比較すると、

$$\hat{r} = r + \left(\sum_{t=k}^{k+l-1} w_t U_{q_t}^{-1} \right) m \quad (22)$$

となる。よって、

$$\begin{aligned} \hat{C} &= R^{-1}\hat{r} \\ &= R^{-1} \left(r + \left(\sum_{t=k}^{k+l-1} w_t U_{q_t}^{-1} \right) m \right) \\ &= \bar{C} + R^{-1} \left(\sum_{t=k}^{k+l-1} w_t U_{q_t}^{-1} \right) m \end{aligned} \quad (23)$$

となる。

$$V = \sum_{t=k}^{k+l-1} w_t U_{q_t}^{-1} \quad (24)$$

$$P = R^{-1} \quad (25)$$

として式 (18) に代入すると,

$$E_{mse} = (\mathbf{X} - (\overline{\mathbf{C}} + \mathbf{P}\mathbf{V}\mathbf{m}))' \cdot (\mathbf{X} - (\overline{\mathbf{C}} + \mathbf{P}\mathbf{V}\mathbf{m})) \quad (26)$$

$\partial E_{mse} / \partial \mathbf{m} = \mathbf{0}_{2M}$ より

$$\mathbf{V}'\mathbf{P}'\mathbf{P}\mathbf{V}\mathbf{m} - \mathbf{V}'\mathbf{P}'(\mathbf{X} - \overline{\mathbf{C}}) = \mathbf{0}_{2M} \quad (27)$$

が得られ,これを解くことにより E_{mse} を最小化する \mathbf{m} を求めることができる.

また,動的特徴量の適応を行わず,静的特徴量の平均ベクトルのみを移動させる場合は,

$$\mathbf{V}^{(0)'}\mathbf{P}'\mathbf{P}\mathbf{V}^{(0)}\mathbf{m}^{(0)} - \mathbf{V}^{(0)'}\mathbf{P}'(\mathbf{X} - \overline{\mathbf{C}}) = \mathbf{0}_M \quad (28)$$

ただし

$$\mathbf{V}^{(0)} = \sum_{t=k}^{k+l-1} \mathbf{w}_t^{(0)} \mathbf{U}_{q_t}^{(0,0)-1} \quad (29)$$

により求められる.

3.3 最ゆう基準

$P(\mathbf{O}|\mathbf{Q},\lambda)$ を式 (9) の制約のもとで \mathbf{C} について整理すると,

$$\begin{aligned} P(\mathbf{O}|\mathbf{Q},\lambda) &= \frac{1}{\sqrt{(2\pi)^{MT}|\mathbf{P}|}} \\ &\cdot \exp\left(-\frac{1}{2}(\mathbf{C} - \overline{\mathbf{C}})' \mathbf{P}^{-1}(\mathbf{C} - \overline{\mathbf{C}})\right) \\ &\cdot \exp\left(-\frac{1}{2}\boldsymbol{\mu}'[\mathbf{U}^{-1} - \mathbf{U}^{-1}\mathbf{W}\mathbf{P}^{-1}\mathbf{W}'\mathbf{U}^{-1}]\boldsymbol{\mu}\right) \\ &\cdot \frac{\sqrt{(2\pi)^{MT}|\mathbf{P}|}}{\sqrt{(2\pi)^{2MT}|\mathbf{U}|}} \end{aligned} \quad (30)$$

となる.すなわち,平均 $\overline{\mathbf{C}}$,共分散 \mathbf{P} の正規分布 $\mathcal{N}(\mathbf{C},\overline{\mathbf{C}},\mathbf{P})$ を定数倍した形となる.この分布 $\mathcal{N}(\mathbf{C},\overline{\mathbf{C}},\mathbf{P})$ に対する \mathbf{X} の対数ゆう度 $\log \mathcal{N}(\mathbf{X},\overline{\mathbf{C}},\mathbf{P})$ を最大化するように,フレーム区間 $k \leq t \leq k+l-1$ の状態 q_t の出力分布の平均ベクトル $\boldsymbol{\mu}_{q_t}$ に移動ベクトル \mathbf{m} を加えることを考える.

3.2 と同様,移動ベクトル \mathbf{m} を加えることにより平均ベクトルが $\hat{\boldsymbol{\mu}}$ となったときに生成されたパラメータ列を $\hat{\mathbf{C}}$ とする.このとき,分布 $\mathcal{N}(\mathbf{C},\hat{\mathbf{C}},\mathbf{P})$ に対する \mathbf{X} の対数ゆう度は,

$$L = \log \mathcal{N}(\mathbf{X},\hat{\mathbf{C}},\mathbf{P})$$

$$= -\frac{1}{2}(\mathbf{X} - \hat{\mathbf{C}})' \mathbf{P}^{-1}(\mathbf{X} - \hat{\mathbf{C}}) + \text{Const.} \quad (31)$$

となる.ただし Const. は \mathbf{X}, \mathbf{m} と独立な項である.よって L の最大化は

$$E_{ml} = (\mathbf{X} - \hat{\mathbf{C}})' \mathbf{P}^{-1}(\mathbf{X} - \hat{\mathbf{C}}) \quad (32)$$

の最小化と等価となる.式 (32) に式 (23) ~ (25) を代入し, $\partial E_{ml} / \partial \mathbf{m} = \mathbf{0}_{2M}$ とおくことにより

$$\mathbf{V}'\mathbf{P}\mathbf{V}\mathbf{m} - \mathbf{V}'(\mathbf{X} - \overline{\mathbf{C}}) = \mathbf{0}_{2M} \quad (33)$$

が得られ,これを解くことにより式 (31) を最大化する \mathbf{m} を求めることができる.

式 (32) は式 (18) に共分散の逆行列 \mathbf{P}^{-1} で重み付けした形となっている.この評価関数を用いることで,共分散の大きい部分での誤差よりも共分散の小さい部分での誤差がより小さくなるように移動ベクトルを求めていることになるため,仮に音素認識の結果が正しいとすると,スペクトルが変動しやすい(出力分布の共分散が大きい)音素よりも変動しにくい(共分散が小さい)音素の部分でより入力音声に近づくようになると考えられる.

静的特徴量に関する平均ベクトルのみを移動させる場合は,

$$\mathbf{V}^{(0)'}\mathbf{P}\mathbf{V}^{(0)}\mathbf{m}^{(0)} - \mathbf{V}^{(0)'}(\mathbf{X} - \overline{\mathbf{C}}) = \mathbf{0}_M \quad (34)$$

により,式 (31) を最大化する $\mathbf{m}^{(0)}$ を求めることができる.

3.4 HMM 認識ボコーダへの適用

2 乗誤差最小基準または最ゆう基準の場合,移動ベクトルを求めるために符号化器でパラメータ生成を行う必要がある.発話全体の認識結果を用いてパラメータ生成を行い,一発話ごとに最適な移動ベクトル(列)を求めて伝送することも考えられるが,符号化に用いる場合には逐次的に実行される方が望ましい.そこで,認識により得られた音素列を (s_1, \dots, s_n, \dots) として,以下のような繰返しにより音素単位で移動ベクトルを求めて伝送する.

- (1) $n = 1$ とする.
- (2) 音素 s_n を確定.
- (3) s_1 から s_n までの区間のパラメータを生成(ただし, s_{n-1} までは適応済とする).
- (4) s_n の継続区間に対して移動ベクトル \mathbf{m}_n を計算.
- (5) 量子化した移動ベクトル $\overline{\mathbf{m}}_n$ を伝送.

(6) s_n の平均ベクトルに \overline{m}_n を加えて適応．

(7) $n := n + 1$ として (2) へ．

なお、移動ベクトルは必ずしも音素単位で求める必要はなく、一定フレームごとに求めて伝送してもよい．

3.5 移動ベクトルの量子化

移動ベクトルを伝送する際、全体で一つのコードブックを用いて量子化するよりも、音素ごとに別々のコードブックを作成して量子化した方がよいと考えられる．この際、各音素の出現確率や平均継続長を考慮し、学習データ全体に対する量子化誤差が最も小さくなるように各音素のコードブックサイズを決定することにより、全音素のコードブックサイズを一定にした場合より効率的に量子化することができる．実際、文献[18]では、一定周期で移動ベクトルを求め一つのコードブックで量子化する場合より、音素ごとに移動ベクトルを計算、量子化する場合の方が良い評価が得られることが示されている．そこで本論文では、以下のようにして各音素のコードブックサイズを決定する．

音素数を N 、音素 n の出現確率を p_n 、平均継続長を d_n 、コードブックサイズを b_n (bit)、コードブックのビット数を $b_n - 1$ から b_n へ増加したときの平均 2 乗誤差の減少量を $\delta_n^{b_n}$ とする．このとき、1 音素当りの平均継続長 \bar{d} は

$$\bar{d} = \sum_{n=1}^N p_n d_n \quad (35)$$

1 音素当りの平均ビット数 \bar{b} は

$$\bar{b} = \sum_{n=1}^N p_n b_n \quad (36)$$

となることから、移動ベクトルの平均ビットレートは

$$\bar{R} = \frac{\bar{b}}{\bar{d}} \quad (37)$$

となる．また、全音素で移動ベクトルを 0 bit で量子化した場合（各音素での平均値を用いた場合）からの 1 音素当りの平均 2 乗誤差の減少量 $\bar{\delta}$ は、

$$\bar{\delta} = \sum_{n=1}^N p_n \sum_{i=1}^{b_n} \delta_n^i \quad (38)$$

となる．学習データに対する移動ベクトル量子化時の平均ビットレート目標値 R (以下「目標ビットレート」と呼ぶ) に対して $\bar{R} \leq R$ となる条件のもとで全体での量子化誤差を最小にするためには、ビット当り

の量子化誤差の減少量 $\bar{\delta}/\bar{b}$ が最大となるように各音素にビットを割り当てればよい．しかし、これを解析的に求めることは容易ではないことから、以下のような繰返しにより準最適なビット割当を決定する．

(1) $i = 0, \bar{b}^{(0)} = 0, \bar{\delta}^{(0)} = 0, b_n = 0 (1 \leq n \leq N)$ とする．

(2) $i := i + 1$

(3) コードブックサイズを 1 bit 増加させたときに、全体でのビットレートが目標ビットレートを超えない音素の集合 S を求める．

$$S = \left\{ n \mid \frac{\bar{b}^{(i-1)} + p_n}{\bar{d}} \leq R \right\} \quad (39)$$

S が空集合なら終了する．

(4) S の中で、1 bit 増加させたときのビット当りの誤差の減少量を最大にする音素 \hat{n} を選択．

$$\hat{n} = \operatorname{argmax}_n \frac{\bar{\delta}^{(i-1)} + p_n d_n^{b_n+1}}{\bar{b}^{(i-1)} + p_n} \quad (40)$$

(5) $b_{\hat{n}} := b_{\hat{n}} + 1, \bar{b}^{(i)} := \bar{b}^{(i-1)} + p_{\hat{n}}, \bar{\delta}^{(i)} := \bar{\delta}^{(i-1)} + p_{\hat{n}} \delta_{\hat{n}}^{b_{\hat{n}}}$ とし (2) へ戻る．

なお、各音素のコードブックは LBG アルゴリズムにより作成する．

4. 実 験

4.1 実験条件

音声データには ATR 日本語音声データベースを用い、データベースに含まれるラベルデータに基づいて 30 種類の音素（無音、ポーズを含む）でラベル付けした．男性話者 10 名による 1,500 文章（各話者 150 文章）を用いて不特定話者モデルを学習し、学習データの話者に含まれない男性話者 2 名（MHT, MYI）を入力話者とした．また、入力話者 2 名による音韻バランス文（各話者 450 文章）を用い、各入力話者それぞれに対して特定話者モデルを学習した．

サンプリング周波数は 8 kHz とした．フレーム長 32 ms、フレーム周期 5 ms のブラックマン窓を用い、メルケプストラム分析 [19] により 0 次から 12 次までのメルケプストラムを求めた．更に式 (41) によりデルタメルケプストラムを計算し、1 次から 12 次までのメルケプストラム及び 0 次から 12 次までのデルタメルケプストラムを特徴ベクトルとした．

$$\Delta c_t = \frac{c_t - c_{t-1}}{2} \quad (41)$$

HMM は 3 状態 left-to-right モデルで、それぞれの

状態の出力分布は単一の対角共分散ガウス分布とした．MDL 基準を用いた決定木に基づくコンテキストクラスタリングにより，triphone HMM の各状態を状態別にクラスタリングして共有化し，tied triphone HMM のセットを生成した．クラスタリング後の総状態数は，不特定話者モデルでは 2127，話者 MHT のモデルでは 1401，話者 MYI のモデルでは 1226 となった．HMM の学習及び入力音声の認識には HTK (Hidden Markov Model Toolkit) [22] を用い，認識時には日本語の音韻連接規則に基づく音素ネットワークを作成して用いた．

移動ベクトルの計算には，モデル最ゆう基準に基づく手法 (MML)，2 乗誤差最小基準に基づく手法 (MSE)，最ゆう基準に基づく手法 (ML)，及び各基準で静的特徴量のみを適応させる手法 (それぞれ MML(s)，MSE(s)，ML(s)) の 6 通りの手法を用い，特定話者モデルを用いた場合 (SD) 及び不特定話者モデルで適応を行わない場合 (SI) と比較した．また，移動ベクトルのコードブックは学習データに対して音素認識を行った結果に基づいて作成した．

音素情報の符号化には音素 bigram 確率に基づく Huffman 符号化を用い，状態継続長は認識時に Viterbi アラインメントにより得られた値をそのまま各状態ごとに状態継続長のヒストグラムに基づいて Huffman 符号化して伝送する．なお，音素 bigram 確率及び各状態の継続長のヒストグラムは，学習データに対して音素認識を行った結果から求めた．

主観評価は DCR テストにより行った．メルケプストラム分析合成系 [19], [21] による分析合成音をリファレンスとし，符号化音声の音質を「5: ほぼ同等」から「1: 非常に劣化している」までの 5 段階で評価した．評価用データは学習データとは異なる 53 文章とした．被験者は 9 名である．評価用データ 53 文章から被験者ごとに重複のないようランダムに 4 文章選択し，各入力話者に対し 2 文章ずつ，1 文章につき順序をランダムに入れ換えて 2 回ずつの評価を行った．

なお，合成時のピッチはデータベースに付属するものを量子化せずに用いた．また，ゲインに対応するメルケプストラム係数の 0 次項も分析により得られた値を量子化せずにそのまま合成に用い，パラメータ生成及び移動ベクトルの計算，量子化は 0 次項を除いて 1 次から 12 次までの係数のみを対象とした．

4.2 移動ベクトルの量子化目標ビットレートとメルケプストラム距離

図 2，図 3 に入力話者 MHT，MYI それぞれに対する入力音声と符号化音声の平均メルケプストラム距離を示す．図の横軸は移動ベクトルの量子化コードブック作成時の目標ビットレート R ，縦軸はメルケプストラム距離，破線は特定話者モデルを用いた場合，1 点鎖線は不特定話者モデルを適応せずに用いた場合，実線は静的特徴量と動的特徴量の両方の平均ベクトルを適応した場合，点線は静的特徴量の平均ベクトルのみを適応した場合で，“x”，“△”，“○” はそれぞれモデル最ゆう基準，2 乗誤差最小基準，最ゆう基準を用いた場合である．また，平均メルケプストラム距離は，学習データに付属するラベルに基づき，発話の前後の

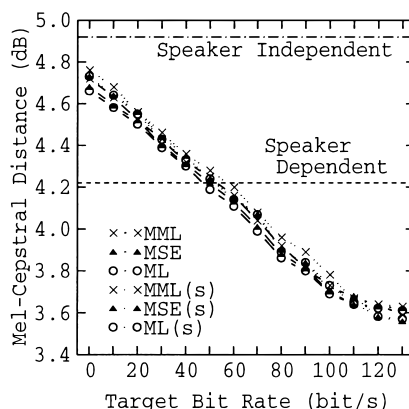


図 2 メルケプストラム距離 (話者 MHT)
Fig. 2 Mel-cepstral distance (speaker MHT).

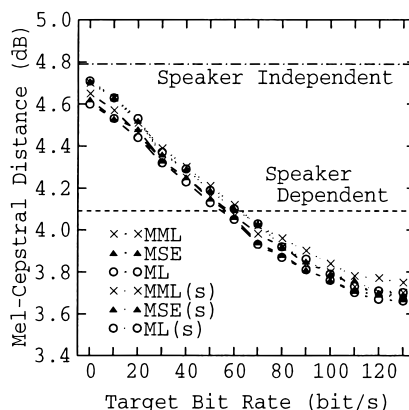


図 3 メルケプストラム距離 (話者 MYI)
Fig. 3 Mel-cepstral distance (speaker MYI).

無音区間を除いて求めている．特定話者モデルを用いた場合の平均メルケプストラム距離は入力話者 MHT 及び MYI に対してそれぞれ 4.22 dB 及び 4.09 dB，不特定話者モデルを用いた場合はそれぞれ 4.92 dB 及び 4.79 dB である．

図 2，図 3 より，各手法による差はあまりなく，どちらの話者の場合でも，目標ビットレートが 60～70 bit/s 以上の場合に特定話者モデルを用いた場合よりも平均メルケプストラム距離が小さくなり，目標ビットレートが上昇するに従って平均メルケプストラム距離が更に減少することがわかる．このことから，移動ベクトルを用いることにより，話者性のみならず，発話ごとの変動等も含めて入力音声に適応していると考えられる．また，目標ビットレートが 0 bit/s の場合，つまり各音素での移動ベクトルの平均値を用いた場合に，不特定話者モデルを適応せずに用いた場合よりもメルケプストラム距離が小さくなっていることがわかる．これは，不特定話者モデルの学習時には学習データに対する正解音素列が与えられているために符号化の際の認識誤りによる影響が考慮されていないのに対し，移動ベクトルの平均値を求める際には学習データに対して音素認識を行っているため，移動ベクトルの平均値を各音素モデルの出力分布の平均ベクトルに加えることにより，認識誤りの影響が不特定話者モデルに反映されるためであると考えられる．ただし，学習データの話者と入力話者との認識誤りの傾向が大きく異なる場合には，不特定話者モデルをそのまま用いた場合よりも移動ベクトルの平均値を加えた場合の方がメルケプストラム距離が大きくなることも考えられる．なお，平均メルケプストラム距離は必ずしも 2 乗誤差最小基準に基づく手法を用いた場合に最小とはなっていないが，これは移動ベクトルの量子化のためのコードブックの作成及び実際の量子化の際に量子化誤差のみを考慮しており，量子化された移動ベクトルが各基準に対して必ずしも最適とはなっていないためである．

4.3 主観評価結果

まず，目標ビットレート R と符号化音声の品質の関係性を調べるため，移動ベクトルの計算に最よう基準に基づく手法 (ML) を用い，目標ビットレートを 60, 80, 100, 120 bit/s とし主観評価実験を行った．図 4 に主観評価により得られた DMOS 値と 95% の信頼区間を示す．

図 4 より，目標ビットレートが上昇するに従って DMOS 値も上昇し，100 bit/s 程度で特定話者モデル

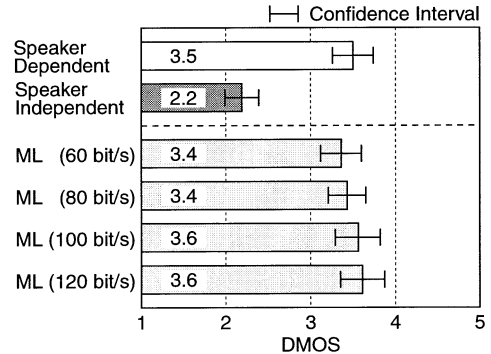


図 4 目標ビットレートと符号化音声の品質
Fig. 4 DMOS scores vs. target bit rates.

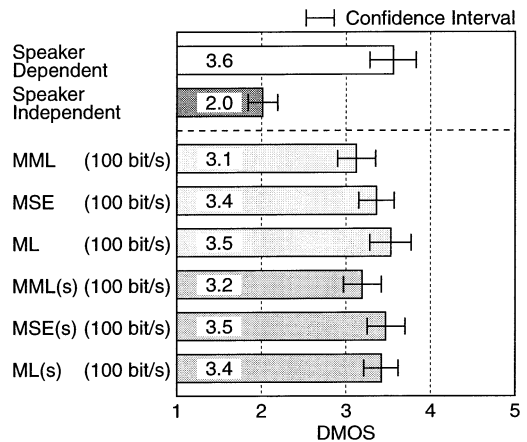


図 5 移動ベクトルの各計算法の比較
Fig. 5 Comparison of criteria for calculation of transfer vectors.

を用いた場合と同等の評価が得られていることがわかる．また，目標ビットレートが 60 bit/s の場合でも，特定話者モデルを用いた場合との評価の差は有意ではなく，入力話者の話者性が再現されていると考えられる．

非公式な受聴により，入力音声への適応を行った場合は適応を行わない場合と比べ，話者性のみではなくめいりょう性も向上していることが確認された．これは，適応を行う際に話者性と音韻性を分離して考えているわけではなく，入力音声のスペクトル列そのものの再現性が良くなるように移動ベクトルを求めているためである．

次に目標ビットレート R を 100 bit/s とし，各移動ベクトルの計算手法の比較を行った．図 5 に主観評価

により得られた DMOS 値と 95% の信頼区間を示す。

静的、動的特徴量を適応する場合、静的特徴量のみを適応する場合ともに、モデル最ゆう基準を用いた場合 (MML) には 2 乗誤差最小基準を用いた場合 (MSE) や最ゆう基準を用いた場合 (ML) と比べて符号化音声の品質が劣化していることがわかる。ただしモデル最ゆう基準を用いた場合でも、適応しない場合と比較すると明らかに品質が改善されている。また、2 乗誤差最小基準及び最ゆう基準の場合には、特定話者モデルを用いた場合とほぼ同等の評価が得られている。

非公式な受聴により、最ゆう基準を用いた場合、他の手法と比べて主観的な品質に影響が大きいと考えられる母音などの部分での音質の改善が認められた。これは、3.3 でも述べたとおり、母音の定常部など、比較的分散の小さいと考えられる部分での誤差を減らすように移動ベクトルを求めているためであると考えられる。

4.4 ビットレート

表 1 に、目標ビットレート R を 100 bit/s とし、最ゆう基準を用いて静的、動的特徴量の両方を適応する場合の、テストデータに対するスペクトル情報の伝送に必要なビットレートを示す。ビットレートは、音素情報、状態継続長は 4.1 で述べた手法により符号化し、移動ベクトルは 3.5 で述べた手法により作成したコードブックを用いて符号化した場合のテストデータ全体 (文頭、文末の無音区間は除く) に対する平均値である。また、表 2 に認識率、認識精度、正解音素数

(N)、置換 (S)、脱落 (D)、挿入 (I) のそれぞれの誤りの数、及び認識結果の全音素数 ($N + S + I$) を示す。ここで、認識率、認識精度は以下の式で求めている。

$$\text{認識率 (\%)} = \frac{N}{N + S + D} \times 100 \quad (42)$$

$$\text{認識精度 (\%)} = \frac{N - I}{N + S + D} \times 100 \quad (43)$$

表 1 より、特定話者モデルを用いた場合と比べて不特定話者モデルを用いた場合には、音素で 10 bit/s 程度、状態継続長で 3 ~ 18 bit/s 程度ビットレートが上昇していることがわかる。表 2 より、特定話者モデルより不特定話者モデルの方が認識率、認識精度が低下しており^(注1)、特に挿入誤りが大幅に上昇していることがビットレート上昇の主な原因であると考えられる。

また、移動ベクトルのビットレートは、目標ビットレートである 100 bit/s を話者 MHT では約 16 bits、話者 MYI では約 27 bit/s 上回っていることがわかる。これは、学習データに対する認識結果の 1 秒当りの平均音素数が 14.1 であったのに対し、テストデータに対する認識結果の 1 秒当りの平均音素数が MHT で 16.0、MYI で 18.0 となっており、学習データよりもテストデータの方が単位時間当りの音素数が多かったことが主な原因であると考えられる。

更に、音素情報の符号化及び移動ベクトルの量子化コードブック作成時に学習データから求められた音素の bigram 確率や出現確率を考慮しているが、一般には学習データとテストデータではこれらの確率は異なっている。このことも、音素や移動ベクトルのビットレート上昇の一因となっていると考えられる。

4.5 MA 予測を用いた 2 段ベクトル量子化との比較

DCR テストによる主観評価により、提案手法において最ゆう基準に基づいて静的、動的特徴量を適応させる場合 (ML) 及び静的特徴量のみを適応させる場合 (ML(s)) と MA 予測を用いた 2 段ベクトル量子化 (以下 MSVQ と呼ぶ) との比較を行った。MSVQ におけるスペクトル分析条件、学習データ、テストデータは 4.1 と同じとし、1 次から 12 次までのメルケプストラム係数のみを量子化した。delayed decision は行わず、MA 予測の予測次数は 4、予測係数の切換は

(注 1): 認識誤りは正解音素よりも音響的なゆが度が高い音声単位が選ばれていることを示しており、認識精度の低下が必ずしも符号化音声の品質の低下につながるとは限らない [16]。

表 1 ビットレート (bit/s)
Table 1 Bit rate for spectral parameters (bit/s).

モデル	MHT	MYI	不特定話者	
入力話者	MHT	MYI	MHT	MYI
音素	45.3	53.7	56.8	64.9
状態継続長	139.8	157.6	157.3	160.0
移動ベクトル	—	—	115.8	127.3
合計	185.1	211.3	329.9	352.2

表 2 音素認識結果
Table 2 Results of phoneme recognition.

モデル	MHT	MYI	不特定話者	
入力話者	MHT	MYI	MHT	MYI
認識率 (%)	95.0	89.8	84.1	75.5
認識精度 (%)	85.1	75.8	56.2	47.8
正解音素数	2547	2366	2254	1989
置換誤り	115	214	382	539
脱落誤り	17	54	43	106
挿入誤り	267	367	746	729
全音素数	2929	2947	3382	3257

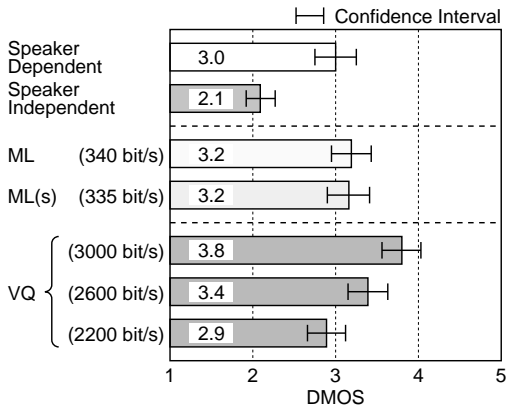


図 6 MA 予測を用いた 2 段ベクトル量子化との比較
Fig. 6 Comparison with 2-stage vector quantization with MA prediction.

1 bit/frame, VQ の各段のビット数は同じとし、各段 5~7 bit/frame, 計 2200~3000 bit/s (200 frame/s) の場合について比較を行った。また、提案手法では ML, ML(s) ともに目標ビットレート R は 100 bit/s とした。被験者は 8 名、その他の実験条件は 4.1 と同じである。

結果を図 6 に示す。図中、“VQ” は MSVQ の結果を表している。また、“ML” 及び “ML(s)” におけるビットレートは、図 4、図 5 とは異なり、入力話者 MHT 及び MYI の全テストデータに対する音素、状態継続長、移動ベクトルを含めたスペクトル情報全体のビットレートである。

図より、提案手法の性能は 2200 bit/s の MSVQ よりは良く、2600 bit/s の MSVQ よりは若干劣ることがわかる。ただし、本実験では HMM 認識ボコーダの分析条件に合わせるため MSVQ のフレーム周期を 5 ms としたが、一般にはフレーム周期は 10 ms 程度が用いられることが多く、この場合 MSVQ のビットレートは本実験の半分程度となる。

5. む す び

本論文では、HMM を用いた音声認識、音声合成に基づく極低ビットレート音声符号化方式である不特定話者 HMM 認識ボコーダにおける入力話者への適応手法について検討した。主観評価実験の結果、提案手法により復号化器で用いる不特定話者モデルを入力音声へ適応させることにより、特定話者モデルを用いた場合と同等の評価が得られた。提案手法では、音素情報

を約 60 bit/s、状態継続長を約 160 bit/s、入力音声へ適応するための情報である移動ベクトルを約 120 bit/s で符号化している。音声全体を符号化するためにはこれに加えて基本周波数とゲインを伝送する必要があるが、文献 [10] によれば基本周波数については約 120 bit/s、ゲインについては約 100 bit/s で符号化できることが示されており、全体で 550~600 bit/s 程度で符号化できると考えられる。

しかし、移動ベクトルの計算時の基準と量子化時の基準が異なっており、量子化時にも計算時の基準を適用することで更に効率的に移動ベクトルの量子化を行うことができると考えられることから、今後、量子化法について検討する必要がある。また、本論文では、復号化器側のモデルのみを適応しているが、符号化器で認識に用いるモデルも同時に逐次的に適応するなど、他の適応の枠組みについての検討も今後の課題となる。更に、基本周波数とゲインも符号化した場合の符号化音声の品質の検討も今後の課題である。

謝辞 本研究の一部は文部省科学研究費補助金（課題番号 09750399）によった。

文 献

- [1] 三樹 聡, 守谷健弘, 間野一則, 大室 伸, “ピッチ同期雑音励振源をもつ CELP 符号化 (PSI-CELP),” 信学論 (A), vol.J77-A, no.3, pp.314-324, March 1994.
- [2] A.V. McCree and T.P. Barnwell III, “A mixed excitation LPC vocoder model for low bit rate speech coding,” IEEE Trans. Speech and Audio Processing, vol.3, no.4, pp.242-250, July 1995.
- [3] M. Nishiguchi, K. Iijima, and J. Matsumoto, “Harmonic vector excitation coding of speech at 2.0 kps,” Proc. IEEE Workshop on Speech Coding, pp.39-40, Sept. 1997.
- [4] 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, April 1980.
- [5] J. Picone and G.R. Doddington, “A phonetic vocoder,” Proc. ICASSP-89, pp.580-583, May 1989.
- [6] F.K. Soong, “A phonetically labeled acoustic segment (PLAS) approach to speech analysis-synthesis,” Proc. ICASSP-89, pp.584-587, May 1989.
- [7] Y. Hirata and S. Nakagawa, “A 100 bit/s speech coding using a speech recognition technique,” Proc. EUROSPEECH-89, pp.290-293, Sept. 1989.
- [8] C.M. Ribeiro and I.M. Trancoso, “Phonetic vocoding with speaker adaption,” Proc. EUROSPEECH-97, pp.1291-1294, Sept. 1997.
- [9] M. Ismail and K. Ponting, “Between recognition and synthesis—300 bits/second speech coding,” Proc. EUROSPEECH-97, pp.441-444, Sept. 1997.

- [10] K.S. Lee and R.V. Cox, "TTS based very low bit rate speech coder," Proc. ICASSP-99, pp.181-184, May 1999.
- [11] S. Roucos, R.M. Schwartz, and J. Makhoul, "A segment vocoder at 150 b/s," Proc. ICASSP-83, pp.61-64, 1983.
- [12] Y. Shiraki and M. Honda, "LPC speech coding based on variable-length segment quantization," IEEE Trans. Acoust., Speech, & Signal Process., vol.36, no.9, pp.1437-1444, Sept. 1989.
- [13] P.A. Chou and T. Lookabaugh, "Variable dimension vector quantization of linear predictive coefficients of speech," Proc. ICASSP-94, pp.505-508, April 1994.
- [14] 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.
- [15] 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.
- [16] 広井 順, 徳田恵一, 益子貴史, 小林隆夫, 北村 正, "HMM に基づいた極低ビットレート音声符号化," 信学論 (D-II), vol.J82-D-II, no.11, pp.1857-1864, Nov. 1999.
- [17] 益子貴史, 小林隆夫, 徳田恵一, "HMM 認識ボコーダの不特定話者化に関する検討," 音講論集, 3-2-12, pp.271-272, Sept. 1998.
- [18] 高橋史生, 益子貴史, 徳田恵一, 小林隆夫, "不特定話者 HMM 認識ボコーダの音質向上に関する検討," 音講論集, 2-P-23, pp.313-314, March 1999.
- [19] 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.
- [20] 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," 音響誌, vol.53, no.3, pp.192-200, March 1997.
- [21] 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.
- [22] <http://htk.eng.cam.ac.uk/>
(平成 13 年 12 月 27 日受付, 14 年 6 月 10 日再受付)



益子 貴史 (正員)

平 5 東工大・工・情工卒・平 7 同大学院博士前期課程了(知能科学専攻)。同年東工大精密工学研究所助手。現在東工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。音声分析・合成・認識, マルチモーダルインタフェースの研究に従事。平 13 本会論文賞, 猪瀬賞各受賞。日本音響学会, IEEE, ISCA 各会員。



小林 隆夫 (正員)

昭 52 東工大・工・電気卒。昭 57 同大学院博士課程了。同年東工大精密工学研究所助手。同助教授を経て現在東工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。デジタルフィルタ, 音声分析・合成・符号化・認識, マルチモーダルインタフェースの研究に従事。平 13 電気通信普及財団賞, 平 13 本会論文賞, 猪瀬賞各受賞。日本音響学会, 情報処理学会, IEEE, ISCA 各会員。



徳田 恵一 (正員)

昭 59 名工大・工・電子卒。平 1 東工大学院博士課程了。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・合成・符号化・認識, デジタル信号処理, マルチモーダルインタフェースの研究に従事。平 13 電気通信普及財団賞, 平 13 本会論文賞, 猪瀬賞各受賞。日本音響学会, 情報処理学会, 人工知能学会, IEEE, ISCA 各会員。