論　文

HMMに基づくテキスト音声合成への混合励振源モデルとポストフィルタの導入

吉村　貴克[†]　　徳田　恵一[††]　　益子　貴史[†††]　　小林　隆夫[†††]
北村　正[††]

Incorporation of Mixed Excitation Model and Postfilter into HMM-Based Text-to-Speech Synthesis

Takayoshi YOSHIMURA[†], Keiichi TOKUDA[††], Takashi MASUKO[†††], Takao KOBAYASHI[†††], and Tadashi KITAMURA[††]

あらまし　本論文は，HMMに基づいた音声合成システムに混合励振源モデルを導入することにより，合成音声の品質向上を図ることを目的とする．我々はこれまでに，メルケプストラム，基本周波数，継続長をHMMの枠組みでモデル化し，HMMからこれらの音声パラメータを出力することによって音声を合成するテキスト音声合成システムを提案した．このシステムでは，合成フィルタ（MLSAフィルタ）を励振する際の励振源モデルとして，有声区間，無声区間でそれぞれパルス列と白色雑音を切り換える単純なモデルを用いている．このような励振源を用いる場合，有声摩擦音のように周期成分と非周期成分をともにもつ音声を合成することができず，合成音声の品質を劣化させる原因となる．そこで本論文では，パルス列と白色雑音を混合する混合励振源モデルを用いることにより高品質な音声を実現している狭帯域音声符号化手法MELPの混合励振源モデルを導入する．この混合励振源モデルは，狭帯域音声符号化だけでなく，広帯域音声符号化へも応用されていることから，音声合成においても有効性が期待される．更に，多くの音声符号化手法で用いられているポストフィルタを導入し，合成音声の品質を向上を図る．また主観評価実験により，本システムにおける混合励振源モデルとポストフィルタの有効性を示す．

キーワード　メルケプストラム分析合成，隠れマルコフモデル，テキスト音声合成，混合励振源モデル，ポストフィルタ

1. まえがき

我々はこれまでに，HMMに基づくテキスト音声合成システムを提案してきた[1]．提案するテキスト音声合成システムは，スペクトル，基本周波数などの音声パラメータを学習したHMMを，入力されたテキストに従って連結し，出力確率最大化基準により音声パラメータを推定することにより，音声を合成する．

本システムの特徴として，他の波形に基づいた単位接続型音声合成システムと比べ，

- 音声単位境界においてひずみが生じにくい．
- 合成音声の声質を変化させやすい．
- システムの自動学習がしやすい．

などの利点がある．HMMの学習時に特徴量として静的特徴量に動的特徴量を加えて学習し，パラメータ生成時に動的特徴量を考慮することにより音声単位境界においてひずみが少ない自然な音声パラメータ列が得られる[2]．また，学習したHMMに音声認識で用いられる話者適応手法を適用することにより，様々な話者の声質で音声を合成することができる[3]．単位接続型音声合成システムの音声単位データベースは，大量の音声データから音声単位を抽出し，適切な単位を選

[†]（株）豊田中央研究所第21研究領域，愛知県
Research-Domain 21, Toyota Central R&D Labs., Inc., Nagakute, Aichi-ken, 480–1192 Japan
[††]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 Inst. of Tech., 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 Inst. of Tech., 4259 Nagatsuta, Midori-ku, Yokohama-shi, 226–8502 Japan

択して構築される．この作業は合成音声の品質に直接かかわってくるため，人手により精度良く行わなければならなく，多大な時間を要する．一方，HMM に基づく音声合成システムでは，トランスクリプション付きの音声データがあれば，HMM の連結学習によりシステムを自動的に構築することができる．

システムの概要を図 1 に示す．学習部では，音声データベースからスペクトル，励振源パラメータを抽出し，それぞれ HMM でモデル化する．合成部では，ゆう度最大化基準に基づく音声パラメータ生成アルゴリズムにより HMM からスペクトル，励振源パラメータを出力し，スペクトルパラメータから構築された合成フィルタを得られた励振源により励振することにより合成音声を得る．

文献 [1] では，スペクトルパラメータとしてメルケプストラム，励振源パラメータとして基本周波数を用いており，基本周波数に従って有声区間，無声区間でそれぞれパルス列と白色雑音を切り換える単純な励振源モデルを用いて合成フィルタ（MLSA フィルタ）を励振し，音声を合成した．このような励振源を用いる場合，有声摩擦音のように周期成分と非周期成分をともにもつ音声を合成することが難しく，しばしばボコーダ特有の（Buzzy などと形容される）音質が問題となる．合成音声の品質を改善するには，励振源パラメータをより高精度にモデル化する必要がある．

そこで本論文では，高精度な励振源モデルの一つとして知られている混合励振源モデルを HMM に基づくテキスト音声合成システムに導入することを考える．現在，いくつか混合励振源モデルが提案されているが，本論文では高品質な音声符号化を実現している MELP [4] の混合励振源モデルを導入する．狭帯域音声符号化手法である MELP では，パルス列と白色雑音を混合する混合励振源モデルを用いることにより高品質な音声を実現している．MELP で用いられている混合励振源モデルでは，周波数帯域を五つに分け，それぞれの帯域で有声の強度（本論文では帯域有声強度と呼ぶ）を求める．有声の強度が高い帯域はパルス列，低い帯域は白色雑音を割り当て，周波数帯域上でパルス列，白色雑音を混合する．この混合励振源モデルは，狭帯域音声符号化だけでなく，広帯域音声符号化 [5] や音声合成システム [6] へも応用されている．また，MELP では自然発声に近い合成音声を実現するため

- 残差信号のフーリエ振幅の使用
- 基本周波数の揺らぎ
- パルス拡散

など，様々な工夫を行っており，我々が提案するテキスト音声合成システムにおいても，合成音声の品質が向上することが十分期待できる．更に本論文では，多くの音声符号化手法で取り入れられているポストフィルタを導入し，合成音声品質の更なる向上を目指す．

以下，本論文は次のように構成されている．**2.**, **3.** では，本論文で導入する混合励振源モデルとポストフィルタの概要を述べ，**4.** では，HMM に基づくテキスト音声合成システムに混合励振源モデルとポストフィルタを導入する．**5.** では，混合励振源の生成，主観評価実験による混合励振源とポストフィルタの評価を行い，**6.** でまとめる．

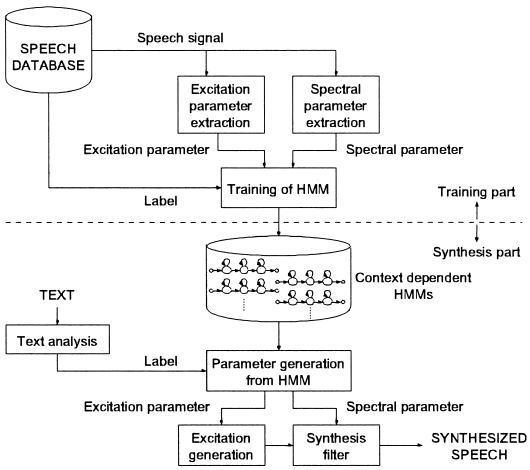


図 1　HMM に基づく音声合成
Fig. 1　HMM-based text-to-speech synthesis.

2. 混合励振源モデル

音声符号化手法 MELP の混合励振源の生成過程を図 2 に示す．混合励振源モデルでは，図 2 のように周波数をいくつかの帯域に分割し，それぞれの帯域に対し有声の強さ（帯域有声強度）を求める．求めた帯域有声強度があるしきい値以上であれば，その帯域を有声，しきい値以下であれば無声と判定する．有声と判定された帯域はパルス列，無声と判定された帯域は白色雑音を割り当て，それぞれ帯域通過フィルタに通し，両者を混合することにより，混合励振源を得る．

帯域有声強度は，分割された周波数帯域ごとの自己相関法に基づいて定義される．つまり，時刻 t での帯

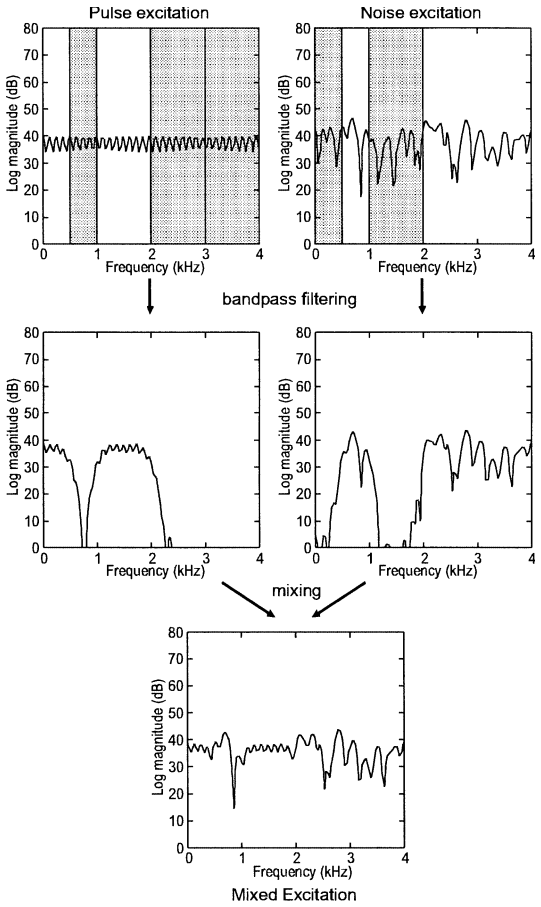


図 2　混合励振源の生成過程
Fig. 2　A generation process of mixed excitation.

域有声強度 c_t は，以下のように求められる．

$$c_t = \frac{\sum_{n=0}^{N-1} s_n s_{n+t}}{\sum_{n=0}^{N-1} s_n s_n \sum_{n=0}^{N-1} s_{n+t} s_{n+t}} \tag{1}$$

ここで，s_n はサンプル n における音声信号，N は基本周波数の分析窓のサイズである．

また本論文では合成音声の自然性を向上させるため，MELP と同様に，

（1） パルス列を生成する際，残差信号(注1)から得られたハーモニクスのフーリエ振幅（以下，フーリエ振幅と呼ぶ）を用いる．

（2） 低域の帯域有声強度に従い基本周波数に揺らぎを加える．

（3） パルス拡散フィルタを使用する．

の処理を施す．

3. ポストフィルタ

MELP を含む多くの音声符号化方式で用いられているポストフィルタを用いてフォルマント強調を行うことにより，合成音声の明瞭性を向上させる．ここでは，スペクトルがメルケプストラムにより表現されていることから，文献 [7] と同様，メルケプストラムに基づいたポストフィルタリングを行う．

メルケプストラム $c(m)$ で表されたスペクトルのモデルは，

$$\begin{aligned} D(z) &= \exp \sum_{m=0}^{M} c(m)\tilde{z}^{-m} \\ &= \exp \sum_{m=0}^{M} b(m)\Phi_m(z) \end{aligned} \tag{2}$$

$$b(m) = \begin{cases} c(m) & m = M \\ c(m) - \alpha b(m+1) & 0 \le m < M \end{cases} \tag{3}$$

のように表される．ここで $\Phi_m(z)$ は，

$$\Phi_m(z) = \begin{cases} 1 & m = 0 \\ \dfrac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}} \tilde{z}^{-(m-1)} & m \ge 1 \end{cases} \tag{4}$$

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{5}$$

であり，M はメルケプストラムの分析次数，α はメル尺度近似のためのパラメータである．本論文では，式 (3) の $c(1)$ を 0 に置き換え，$c(2) \sim c(M)$ を β 倍したものをポストフィルタとする．ここで β は，ポストフィルタリングの強度を表すパラメータであり，$\beta = 0$ がポストフィルタリングを行わない場合にあたり，$\beta > 0$ とすることによりスペクトルが強調される．$c(1)$ を 0 にするのは，スペクトルの大局的な傾きが強調されるのを防ぐためである．また，ポストフィルタのゲイン，つまりインパルス応答の時刻 0 の値を 1 となるように規格化する．この操作は，$b(0)$ を 0 に置き換えることにより行うことができる．これらの処理を $b(m)$ の領域で表現すると次のようになる．

(注1)：本論文では MLSA フィルタを用いて得られた残差信号を用いる．

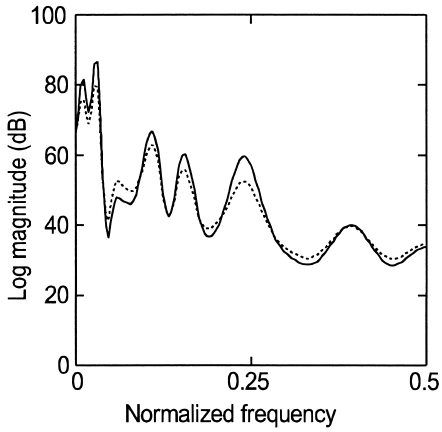


図 3 ポストフィルタリングの効果（点線：ポストフィルタリング前 $D(z)$，実線：ポストフィルタリング後 $D(z) \cdot \bar{D}^{\beta}(z), \beta = 0.5$）

Fig. 3 Effect of postfiltering (doted line: before postfiltering $D(z)$, solid line: after postfiltering $D(z) \cdot \bar{D}^{\beta}(z), \beta = 0.5$).

$$\bar{D}^{\beta}(z) = \exp \sum_{m=1}^{M} \beta \bar{b}(m) \Phi_m(z) \tag{6}$$

$$\bar{b}(m) = \begin{cases} b(m) & 2 \leq m \leq M \\ -\alpha b(2) & m = 1 \end{cases} \tag{7}$$

ポストフィルタリングの様子を図 3 に示す．図より，ポストフィルタリング後，フォルマントが強調されている様子が観察できる．

4. HMM に基づくテキスト音声合成システムへの混合励振源モデルの導入

4.1 特 徴 量

混合励振源モデルを導入するために，混合励振源の生成に必要なパラメータ

- 対数基本周波数 $\log F_0(F)$
- 帯域有声強度 ($\boldsymbol{Vbp}$)
- フーリエ振幅 ($\boldsymbol{M}$)

を，HMM の学習の際の特徴量に含める．

特徴量の構成を図 4 に示す．特徴量は，大きくスペクトルパラメータと励振源パラメータからなり，スペクトルパラメータはメルケプストラム係数，励振源パラメータは基本周波数，帯域有声強度，フーリエ振幅からなっており，それぞれのパラメータは動的特徴量であるデルタ，デルタデルタパラメータを含んでいる．

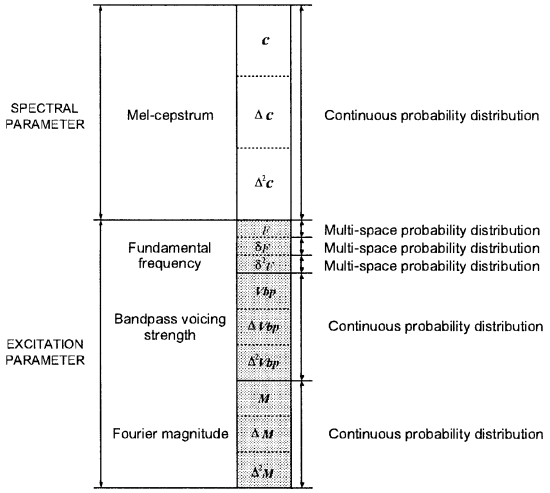


図 4 特徴ベクトル

Fig. 4 Structure of feature vector.

4.2 学 習

HMM は 5 状態の left-to-right モデルとし，音素単位で用意する．メルケプストラム，帯域有声強度，フーリエ振幅はそれぞれ連続分布 HMM，基本周波数は多空間分布 HMM (MSD-HMM) [8], [9]，継続長は多次元ガウス分布でモデル化する．

本論文で使用する HMM は，以下に示す音素環境，品詞，アクセントなどのコンテクストを考慮したコンテクスト依存モデルとする．

- 文の長さ
- { 先行，当該，後続 } 呼気段落の長さ
- 当該アクセント句の位置，前後のポーズの有無
- { 先行，当該，後続 } アクセント句の長さ，アクセント型
- { 先行，当該，後続 } の品詞（23 種類），活用形（7 種類），活用型（7 種類）
- 当該音素のアクセント句内でのモーラ位置
- { 先行，当該，後続 } 音素（42 種類）

作成したコンテクスト依存 HMM は，MDL 基準を用いた決定木に基づくコンテクストクラスタリングによりメルケプストラム，基本周波数，帯域有声強度，フーリエ振幅，継続長をそれぞれ別々に状態クラスタリングする [1]．

4.3 テキスト音声合成

音声合成システムのブロック図を図 5 に示す．まず，入力テキストに基づき連結された文 HMM からメルケプストラム，基本周波数，帯域有声強度，フーリエ振

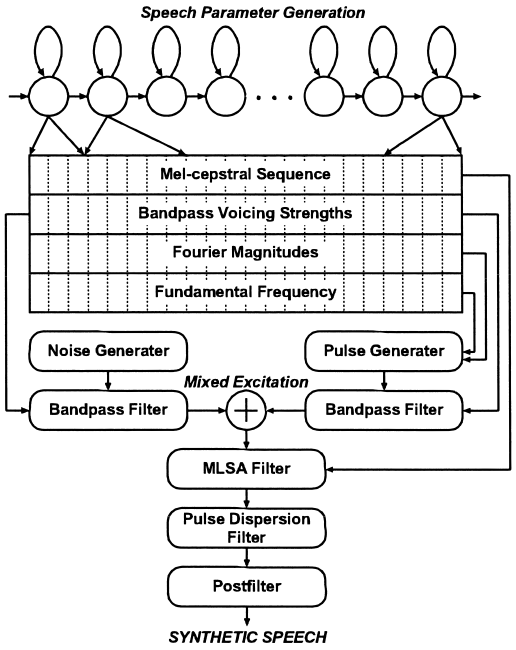


図 5　音声合成システム
Fig. 5　Block diagram of text-to-speech synthesis system.

幅を生成する．次に，生成した基本周波数とフーリエ振幅に基づいてパルス列を出力する．このとき，低域の帯域有声強度に従い基本周波数に揺らぎを加えるようにする．帯域有声強度から五つの帯域それぞれについて，しきい値により有声/無声を判定する．有声の場合はパルス列，無声の場合は白色雑音を各帯域に割り当て，周波数帯域上で混合する．混合された信号はパルス拡散フィルタに通される(注2)．本論文で用いるパルス拡散フィルタは，MELP と同様に三角パルスのスペクトルを白色化することにより得られる 130 次の FIR フィルタを使用する．メルケプストラムに基づき構成された MLSA フィルタを，得られた混合励振源により励振し，合成音声を得る．最後に，得られた合成音声はポストフィルタに通される．

5. 実　　験

5.1 実験条件

学習データとして 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文 450 文章を用いた．サンプリング周波数は 16 kHz，分析周期は 5 ms とした．スペクトルパラメータとしては，25 ms 長ブラックマン窓を用いて 24 次メルケプストラム分析により得られた $c(0)$〜$c(24)$ のメルケプストラム係数を用いた．α は 0.42 とした．帯域有声強度は，文献 [5] に習い，五つのサブバンド 0〜1 kHz，1〜2 kHz，2〜4 kHz，4〜6 kHz，6〜8 kHz でそれぞれ求め，5 次元のベクトルとし，フーリエ振幅は，残差信号から求められたハーモニクスの上位 10 個のフーリエ振幅を 10 次元のベクトルとして用いた．特徴ベクトルは，それぞれのパラメータのデルタ，デルタデルタパラメータを含め，全 123 次元のベクトルとなった．

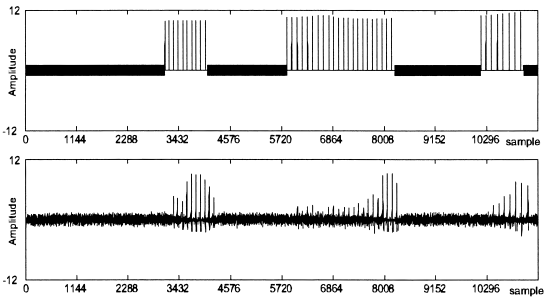


図 6　生成された励振源「少しずつ」(上：従来の励振源，下：混合励振源)
Fig. 6　Examples of generated excitation for Japanese phrase “sukoshizutu” (top: conventional excitation, bottom: mixed excitation).

HMM の学習の結果，MDL 規準を用いたコンテクストクラスタリングにより，HMM の総分布数はメルケプストラム，基本周波数，帯域有声強度，フーリエ振幅，継続長モデルそれぞれ 933, 1054, 1652, 3771, 1012 個となった．

5.2 混合励振源の生成

HMM から生成した励振源の例を図 6 に示す．この図から，生成された混合励振源が有声摩擦音/z/で周期性と非周期性をともにもつ様子が観察できる．

5.3 主観評価実験

混合励振源モデルとポストフィルタの有効性を示すため，主観評価試験を行った．受聴試験に用いた文章は，53 文章の中から被験者ごとにランダムに 8 文章を選んだ．受聴試験のサンプルとして以下の 4 種類のサンプルを用いた．

- 従来の励振源モデル+ポストフィルタなし (NORMAL)
- 従来の励振源モデル+ポストフィルタあり

(注2)：パルス拡散フィルタには遅延があるため，先にパルス拡散を行うと，励振源情報と合成フィルタ情報との間に時間差が生じる．これを避けるため，MELP 方式と同様，合成フィルタとパルス拡散フィルタの処理の順を逆にしている．

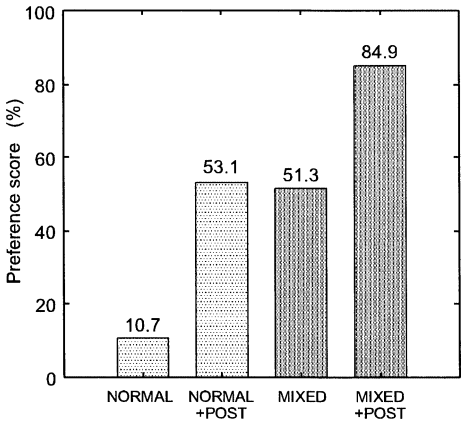


図 7 混合励振源モデルとポストフィルタの効果
Fig. 7 Effect of mixed excitation and postfiltering.

(NORMAL+POST)

- 混合励振源モデル+ポストフィルタなし (MIXED)
- 混合励振源モデル+ポストフィルタあり (MIXED+POST)

被験者 8 名に 4 種類のサンプルを対比較試験により評価させた．なお，ポストフィルタリングの強度を表すパラメータ β は，予備実験により $\beta = 0.4$ とした．

プリファレンススコアを図 7 に示す．結果から，混合励振源モデルを用いた場合に品質が向上し，ポストフィルタを適用することにより，更に品質が改善されていることが分かる(注3)．また，従来の励振源モデルにおいてポストフィルタを適用するだけでも品質が向上しているのが分かる．ただし，本評価結果は男性話者 1 名のみの結果であるため，今後，他の話者における本手法の有効性も評価する必要がある．

6. むすび

本論文では，HMM に基づくテキスト音声合成システムに混合励振源モデルを導入し，実験において合成音声の品質が改善することを示した．また，混合励振源モデルとポストフィルタを併用することにより，合成音声の品質が大幅に向上することを示した．

今後の課題としては，混合励振源の帯域有声強度を求める際の周波数帯域の分割の仕方，励振源パラメータの学習法などの検討，他の話者の音声合成における本手法の有効性の評価が挙げられる．また，本論文では MELP の励振源モデルを用いたが，パラメータを HMM でモデル化できれば，他の励振源モデルを取り入れることもできるため，これについても今後検討する．

(注3)：提案手法によって合成された音声のサンプルは，http://kt-lab.ics.nitech.ac.jp/~yossie/TTS/index.html にて聴くことができる．

文　献

[1] 吉村貴克，徳田恵一，益子貴史，小林隆夫，北村　正，“HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化，”信学論（D-II），vol.J83-D-II, no.11, pp.2099–2107, Nov. 2000.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” Proc. ICASSP, vol.3, pp.1315–1318, June 2000.

[3] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “Text-to-speech synthesis with arbitrary speaker’s voice from average voice,” Proc. EUROSPEECH, pp.345–348, Sept. 2001.

[4] A.V. McCree and T.P. Barnwell III, “A mixed excitation LPC vocoder model for low bit rate speech coding,” IEEE Trans. Speech Audio Process., vol.3, no.4, pp.242–250, July 1995.

[5] W. Lin, S.N. Koh, and X. Lin, “Mixed excitation linear prediction coding of wideband speech at 8 kbps” Proc. ICASSP, vol.2, pp.1137–1140, June 2000.

[6] N. Aoki, K. Takaya, Y. Aoki, and T. Yamamoto, “Development of a rule-based speech synthesis system for the Japanese language using a MELP vocoder,” IEEE Int. Sympo. on Intelligent Signal Processing and Communication Systems, pp.702–705, Nov. 2000.

[7] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai, “CELP coding based on mel-cepstral analysis,” Proc. ICASSP, vol.1, pp.33–36, May 1995.

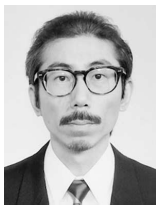
[8] 徳田恵一，益子貴史，宮崎　昇，小林隆夫，“多空間上の確率分布に基づいた HMM，”信学論（D-II），vol.J83-D-II, no.7, pp.1579–1589, July 2000.

[9] 益子貴史，徳田恵一，宮崎　昇，小林隆夫，“多空間確率分布 HMM によるピッチパタン生成，”信学論（D-II），vol.J83-D-II, no.7, pp.1600–1609, July 2000.

（平成 15 年 6 月 24 日受付，11 月 26 日再受付）

吉村　貴克

平 9 名工大・工・知能情報システム卒．平 14 同大大学院博士後期課程了．現在，(株) 豊田中央研究所客員研究員．工博．在学中，音声合成の研究に従事．日本音響学会会員．



徳田　恵一　（正員）

昭 59 名工大・工・電子卒．平元東工大大学院博士課程了．同年同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．音声言語情報処理，マルチモーダル情報処理，統計的学習理論の研究に従事．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，人工知能学会，情報処理学会，IEEE, ISCA 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．同年同大精密工学研究所助手．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を経て平 16 より（株）東芝研究開発センター勤務．博士（工学）．音声の分析・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 同大大学院博士課程了．同年同大精密工学研究所助手．同助教授を経て現在同大大学院総合理工学研究科教授．工博．信号処理，音声の分析・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．

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．現在，同大大学院教授．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，情報処理学会，IEEE, ISCA 各会員．