Vol.2012-MUS-94 No.15
Vol.2012-SLP-90 No.15
2012/2/4

HMM 歌声合成における音高正規化学習の検討

大浦 圭一郎[†1] 間瀬 絢美[†1]
南角 吉彦[†1] 徳田 恵一[†1]

隠れマルコフモデル (Hidden Markov Model; HMM) に基づく歌声合成システムは HMM テキスト音声合成システムを応用したシステムで，歌声から抽出したスペクトル，基本周波数，ビブラートを HMM でモデル化し，学習した HMM からパラメータを生成することで，任意の歌声が合成できる．しかし，HMM 歌声合成が合成可能な音高は学習データベースに強く依存するため，学習データベースの中に特定の音高が少ない場合や存在しない場合にその音高をうまく合成できない問題があった．この問題を軽減するため，音高をシフトさせたデータを用いて擬似的に学習データを増やす手法や，あらかじめデータを正規化する手法が提案されているが，疑似学習データによる学習時間の増大や，学習アルゴリズムとデータの不一致などの様々な問題があった．そこで本稿では，音符の音高を基準とした対数基本周波数系列の正規化を学習に内包する音高正規化学習手法を提案し，主観評価実験により提案手法の有効性を確認した．

HMM-based synthesis of singing voices using pitch adaptive training

Keiichiro Oura,[†1] Ayami Mase,[†1] Yoshihiko Nankaku[†1]
and Keiichi Tokuda [†1]

In Hidden Markov Model (HMM)-based singing voice synthesis approach, the spectrum, excitation, and vibrato of singing voices are simultaneously modeled with context-dependent HMMs and waveforms are generated from the HMMs themselves. HMM-based singing voice synthesis systems are heavily based on the training data in performance because these systems are “corpus-based.” Therefore, pitches hardly ever appear in the training data cannot be well-trained. A technique using pitch-shifted pseudo-data is one solution to this problem. However, there are various problems such as large computational costs. Although data-level pitch normalization has also been proposed, there are still some other problems such as the inconsistency between data and training. In this paper, we proposed “pitch adaptive training” which make it possible to normalize pitch based on musical notes in the training process. The experimental results demonstrated that the proposed technique could alleviate the data sparsness problem.

1. はじめに

近年，隠れマルコフモデル (Hiddeen Markov Model; HMM) に基づくテキスト音声合成手法[1] が盛んに研究されている．この手法では，音声から抽出したスペクトル，基本周波数に基づいて HMM のモデルパラメータが推定され，合成音声波形は推定された HMM 自体から生成される．システムに波形を保持しないため軽量である，モデルパラメータを適切に変換することで様々な声質の音声を合成できるなどの特徴を持ち，これまで話者適応[2]，話者補間[3]，固有声[4] などの手法が提案されてきた．HMM 歌声合成[5] は HMM テキスト音声合成を歌声に応用したものであり，歌声データベースから歌い手の声質や歌唱スタイルなどの特徴を自動学習し，その特徴を再現した任意の歌声を合成することができる．HMM 歌声合成は統計手法のため，その合成音声品質は学習に用いるデータベースに強く依存する．そのため，学習データベースの中に特定のコンテキストが少ない場合や存在しない場合には，そのコンテキストをうまく学習することができない．本来，必要なコンテキスト要因を全て網羅するデータベースを学習に用いることが望ましいが，調や歌詞，強弱，音符の位置，長さ，音高などの全てのコンテキストの組合せを網羅するデータベースを用意することは不可能である．ただし，音高のコンテキストは歌声の合成音声品質に強く影響するため，データベース内の音高の偏りを軽減する手法が提案されている．音高のシフトによって擬似的に学習データを増やす手法[6] を用いることでこの問題を大幅に軽減できるが，膨大な計算量などの様々な問題があった．また，学習データをあらかじめ正規化する手法[7] も提案されているが，データと学習の不一致などの問題があった．そこで本稿では，音高の正規化を学習に内包した音高正規化学習手法を提案する．

以降，2 節で HMM 歌声合成システムの概要を紹介し，3 節で音高正規化学習，4 節では主観評価実験とその評価について述べる．そして最後に 5 節でむすびとし，本稿のまとめと今後の展望について述べる．

2. HMM 歌声合成システム

HMM 歌声合成システムは HMM テキスト音声合成システムを歌声に応用したものであ

†1 名古屋工業大学大学院工学研究科
Graduate School of Engineering, Nagoya Institute of Technology

© 2012 Information Processing Society of Japan

Vol.2012-MUS-94 No.15
Vol.2012-SLP-90 No.15
2012/2/4

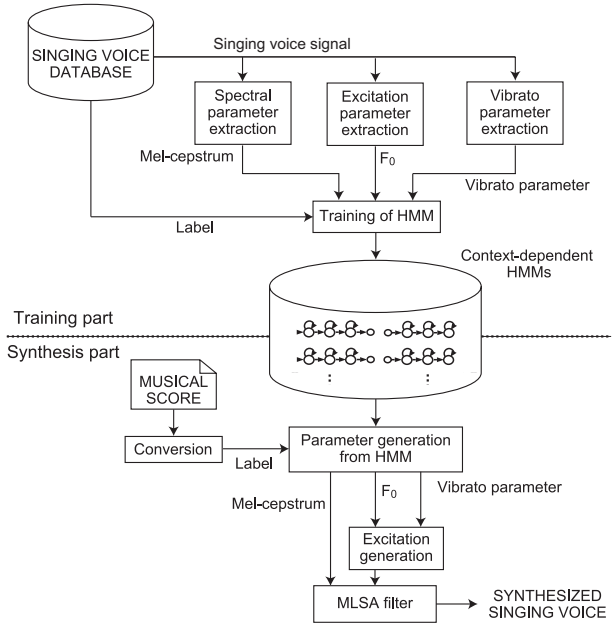

図 1 *HMM* 歌声合成システムの概要
Fig. 1 *The overview of the HMM-based singing voice synthesis system.*

る．概要を図 1 に示す．学習部と合成部に分かれており，学習部ではデータベースから抽出したスペクトル，基本周波数，ビブラートをコンテキスト依存 HMM でモデル化する．また，状態継続長に関しても同時にモデル化される．合成部ではまず任意の歌詞付き楽譜がコンテキスト依存ラベルに変換され，そのラベルに従って HMM が連結される．各音符の継続長と状態継続長モデルに基づいて状態継続長が決定された後，スペクトル，基本周波数，ビブラートのパラメータがパラメータ生成アルゴリズム[8] によって生成され，MLSA フィルタ[9] により歌声が合成される．

3. 音高正規化学習

HMM に基づく音声合成手法は統計手法のため，その合成音声の品質は学習データベースに強く依存する．そのため，学習データベースの中に特定のコンテキストが少ない場合や存在しない場合には，そのコンテキストをうまく学習することができない．この問題を解決するため，出現するコンテキストのバランスを考慮するアルゴリズムを用いた読み上げデータベースの設計[10] などが提案されている．歌声合成システムに関しても，本来，必要なコンテキスト要因を全て網羅するデータベースを学習に用いることが望ましいが，読み上げ音声に含まれるコンテキストに加えて，音高，テンポ，調，拍子，強弱などの全てのコンテキストの組合せを網羅するデータベースを用意することは不可能である．ただし，音高のコンテキストは歌声の合成音声品質に強く影響するため，データベース内の音高の偏りを軽減する手法が提案されている．

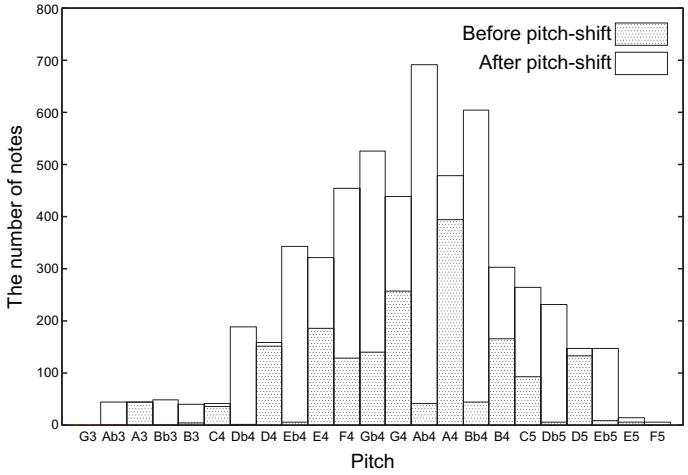

図 2 学習データ *10* 曲に含まれる音高の音符数の分布
Fig. 2 *The distribution of pitch in training data (10 songs).*

音高シフトによる疑似学習データを用いる手法[6] はこの問題に対する解決策の一つである．音高は対数基本周波数パラメータとして保持しているため，それを半音単位でシフトするだけで音高をシフトした疑似学習データを簡単に用意することができ，大量の歌声を新たに収録することなく音高のコンテキストを増やすことができる．図 2 に学習データ 10 曲に含まれる音高の音符数の分布を示す．音高をシフトした疑似学習データを追加することで，本来学習データに少ない音高を補完できていることが確認できる．ただし，スペクトル及びビブラートパラメータに関しては半音程度の音高の違いによる影響はほとんど受けないと考え，元の音高のパラメータをコピーして利用する． この手法により擬似的に学習データ量が 3 倍になるため，最小記述長 (Minimum Description Length; MDL) 基準[11] に基づくコン

© 2012 Information Processing Society of Japan

Vol.2012-MUS-94 No.15
Vol.2012-SLP-90 No.15
2012/2/4

テキストクラスタリングを用いて決定木を構築する際，決定木の大きさを調整する重みを適切な値にする．音高シフトによる手法を用いることで歌声の自然性は大幅に向上するものの，以下の 4 つの問題点があった．

問題点 1　異なる音高のコンテキストを混ぜてモデル化してしまうため，特定の音高または音域にのみ含まれる特徴をモデル化することができない．

問題点 2　学習データの音域外の合成がシフト量に依存する．

問題点 3　擬似学習データを用いるためにデータ量が増え，学習時間が大幅に増大する．

問題点 4　シフト量及び決定木の大きさを調整する重みを適切に調整しなければならない．

これらの問題点に対し，歌声の対数基本周波数系列と音符の音高の差分をモデル化する音高正規化手法が必要になる．

音高の正規化には，学習データをあらかじめ正規化しておいてから通常どおり学習する手法と，モデルの中に正規化を組み込む手法が考えられる．このうち，歌声の対数基本周波数系列と音符の音高の差分をそのまま HMM の学習データとして用いる手法[7] が提案されているが，

問題点 5　学習データを用意する際に正確なアライメントがなければならない．

問題点 6　学習時にアライメントを固定する場合は，スペクトル，基本周波数，ビブラート，状態継続長を同時最適化できない．学習時にアライメントを固定しない場合は，データと学習アルゴリズムの不一致により，静的特徴量に矛盾が生じる．

などの問題点があった．

そこで本稿では，これら 6 つの問題点に対し，モデルに音高の正規化を組み込んだ音高正規化学習手法を提案する．この手法は話者毎の正規化を学習に内包する話者正規化学習[12] を音高に応用したものである．学習データをあらかじめ正規化する手法と，モデルの中に正規化を組み込んだ提案法の比較を図 3 に示す． 話者正規化学習では，状態 i における各話者と平均声の差分を以下のように線形回帰で定義する．

$$\boldsymbol{\mu}_i^{(f)} = \boldsymbol{W}_i^{(f)} \boldsymbol{\xi}_i \tag{1}$$

$$\boldsymbol{W}_i^{(f)} = \left[\boldsymbol{\zeta}_i^{(f)}, \boldsymbol{\epsilon}_i^{(f)} \right] \tag{2}$$

$$\boldsymbol{\xi}_i = \left[\boldsymbol{\mu}_i^{\top}, 1 \right]^{\top} \tag{3}$$

変数 $\boldsymbol{\mu}_i^{(f)}$，$\boldsymbol{W}_i^{(f)}$，$\boldsymbol{\xi}_i$ は，それぞれ，話者 f の平均ベクトル，話者 f と平均声の差分を表現する変換行列，拡張された平均声の平均ベクトルである．話者正規化学習におけるパラメータ推定では，尤度が最大になるように平均声の HMM のパラメータと話者毎の変換行列が推定される．一方，音高正規化学習における状態 i の対数基本周波数の静的特徴量の平均推定値 $\hat{\mu}_i$ は以下のように定義する．

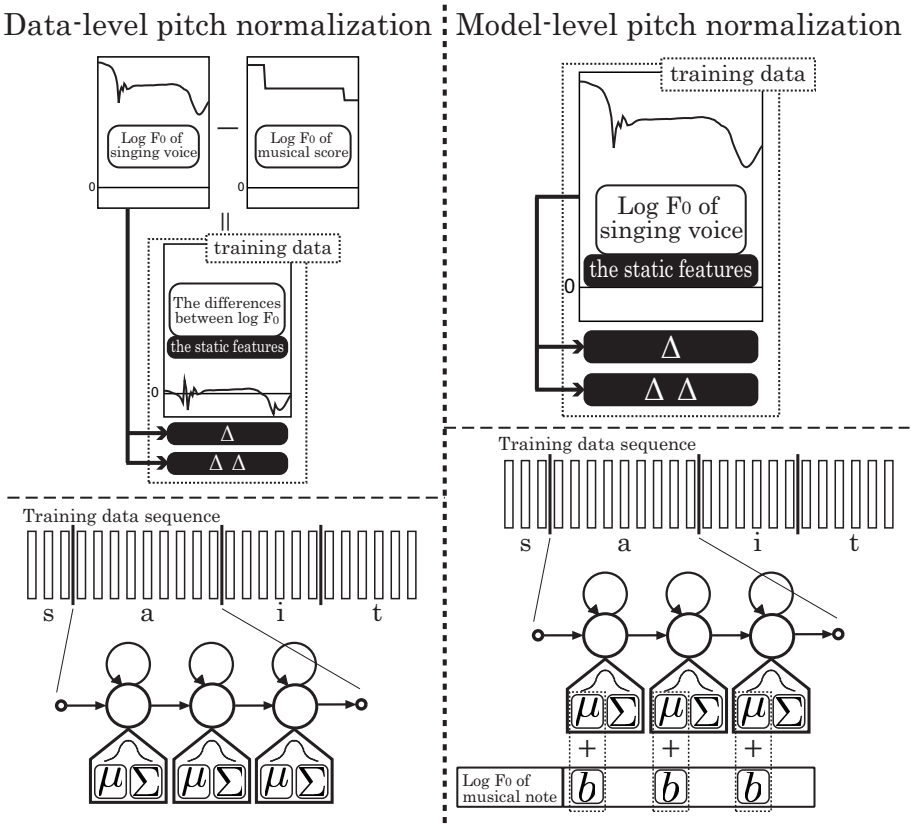

図 3　データレベルとモデルレベルの正規化手法の比較
Fig. 3　*Comparison between data-level and model-level pitch normalization.*

$$\hat{\mu}_i = \boldsymbol{W}_i \boldsymbol{\xi}_i \tag{4}$$

$$\boldsymbol{W}_i = [1, b_i] \tag{5}$$

$$\boldsymbol{\xi}_i = [\mu_i, 1]^{\top} \tag{6}$$

© 2012 Information Processing Society of Japan

Vol.2012-MUS-94 No.15
Vol.2012-SLP-90 No.15
2012/2/4

ここで，変数 μ_i は抽出された対数基本周波数と音符の音高の差分の平均，b_i は音符の対数基本周波数である．音高正規化学習では話者正規化学習と異なり，楽譜に基づいて変換行列 W が固定されるので，通常の学習と同様に HMM のパラメータだけを推定すれば良い．

音高正規化学習を用いることで，これまでに挙げた 6 つの問題点が全て解決可能である．特定の音高または音域にのみ含まれる特徴を抽出することができない問題 (問題点 1) に対して，音高正規化学習では全てのデータが正しい音高コンテキストを持つので，特定の音高または音域にのみ含まれる特徴を効果的にモデル化することができる．学習データの音域外の合成がシフト量に依存する問題 (問題点 2) に対しては，学習時に歌声の対数基本周波数系列と音符の音高の差分のみをモデル化することであらゆる音高が合成できる．また，学習時間が増大する問題 (問題点 3) に対しては，音高正規化学習では学習データ量が変化しないため学習時間はほとんど変わらない．調整すべきパラメータの問題 (問題点 4) に対しても，音高正規化学習では音高シフトによる擬似学習データが存在しないためクラスタリングの分割停止条件に MDL 基準を用いることができ，シフト量も調整しないため，人手でパラメータを調整する必要がない．学習データにあらかじめアライメントが必要な問題 (問題点 5) に対して，提案法はモデルに正規化が組み込んであるので，適切なアライメントが自動的に推定される．HMM の学習時にアライメントを固定しなければならない問題 (問題点 6) に対しても，提案法では学習時にアライメントを固定しなくとも全てのパラメータが同時最適化でき，データと学習アルゴリズムの不一致も無い．

4. 評価実験

提案法の有効性を示すため，評価実験を行った．

4.1 実験条件

学習及び主観評価実験には女性 1 名による童謡 70 曲，合計 71.8 分の歌声データベースを用いた．サンプリング周波数は 48kHz，量子化ビット数は 16bit，モノラルである．スペクトルパラメータとしては，STRAIGHT[13] によって抽出されたスペクトルに，メルケプストラム分析[14] を適用することにより得られた 49 次元のメルケプストラムパラメータとその Δ, Δ^2 を用いた．基本周波数パラメータとしては対数基本周波数とその Δ, Δ^2 を用いた．ビブラートパラメータはビブラートの振幅，周波数とそれらの Δ, Δ^2 を用いた．HMM は 5 状態の left-to-right 型 HSMM[15] を用いた．

4.2 主観評価実験

音高正規化学習による歌声合成の自然性の向上を確認するため，ベースライン，音高シフトによる疑似学習データを用いた従来法，音高正規化学習を用いた提案法の計 3 手法で合成した歌声について主観評価実験を行った．これらの手法のパラメータ数は MDL 基準に基づいて決定したが，従来法では疑似学習データを用いるため適切なパラメータ数を設定することができないため，従来法の基本周波数以外の分布に関してはベースラインとパラメータ数が同程度になるように閾値を手で調整した．なお，従来法のシフト量は上下 1 半音とした．学習データに含まれない 10 曲の楽譜を音域内セット，音域内セットを半オクターブ上に移調した楽譜を音域外セットとした．2 種類のテストセットを用いて歌声を合成し，被験者 10 名にランダムに選ばれた 15 フレーズを聞かせ，歌声の自然性についてそれぞれ 5 段階 MOS で評価させた．なお，学習データの 10 曲の音域は C4 から F5，60 曲の音域は G3 から F5，音域内セットの音域は C4 から D5，音域外セットの音域は F#4 から G#5 である．10 曲で学習し，学習データの音域に含まれるテストセットで評価した実験結果を図 4 に，10 曲で学習し，音域を半オクターブ移調したテストセットで評価した実験結果を図 5 に示す．また，60 曲で学習し，学習データの音域に含まれるテストセットで評価した実験結果を図 6 に示す． 図 4 より，学習データに存在する音域の合成に差は見られないものの，図 5 より，従来法と提案法では学習データに無い音域ではベースラインより高い自然性を得ることができたことがわかる．図 5 の従来法と提案法を比較するとわかる通り，提案法では音高の正規化により，あらゆる音高を合成することができるので，学習データに無い音域において従来法より高い自然性を達成した．また，図 4 と図 6 を比較すると，学習データを増やした場合には学習データと同じ音域でも，提案法はより適切に音高をモデル化する傾向が見られた．

学習に要した時間を表 1 に示す． 表より，従来法は擬似学習データ量が増加するために

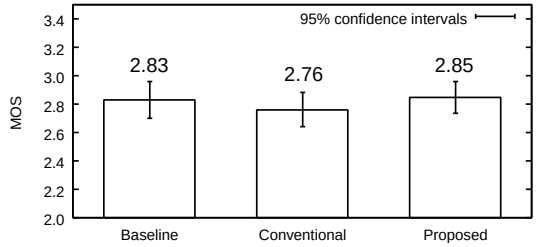

図 4 主観評価実験結果（*10* 曲で学習/音域内セット）
Fig. 4 *Subjective evaluation results: 10 songs were used for training. The pitch range of the test songs was included in the pitch range of the 10 training songs.*

© 2012 Information Processing Society of Japan

Vol.2012-MUS-94 No.15
Vol.2012-SLP-90 No.15
2012/2/4

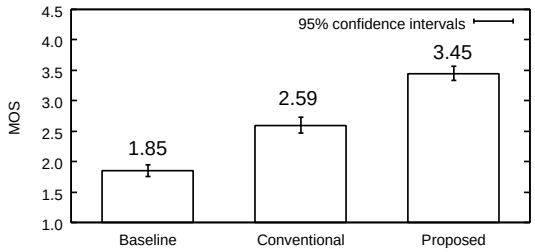

図 5 主観評価実験結果（*10* 曲で学習/音域外セット）
Fig. 5 *Subjective evaluation results: 10 songs were used for training. The key of the test songs was transposed up to a half octave.*

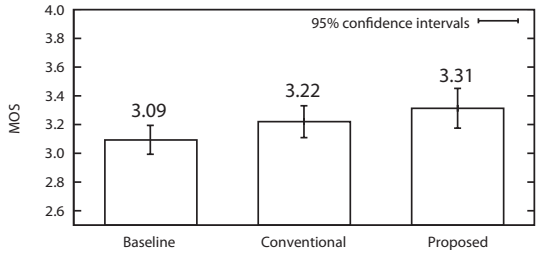

図 6 主観評価実験結果（*60* 曲で学習/音域内セット）
Fig. 6 *Subjective evaluation results: 60 songs were used for training. The pitch range of the test songs was included in the pitch range of the 60 training songs.*

ベースラインの約 3 倍の時間が必要であるのに対し，提案法はベースラインとほぼ同程度の時間で学習できることを確認した．

学習データの高域境界における，各手法の対数基本周波数のモデリングの性能を確かめるため，図 7 に A4 から G5 の 4 つの「あ」の音を合成した合成歌声の音高を示す．実線が楽譜から計算される音高，点線が各手法が生成する対数基本周波数をプロットしたものである．なお，音高を確認しやすくするためビブラートモデルを使わずに合成した．ベースラインでは学習データに含まれる最高音高 F5 よりも高い音高 G5 と，学習データに含まれているものの数が少ない音高 E5 の合成ができていない．また，従来法でも音高シフトによる擬似学習データの最高音高 F#5 よりも高い音高 G5 が合成できていないが，提案法では全ての

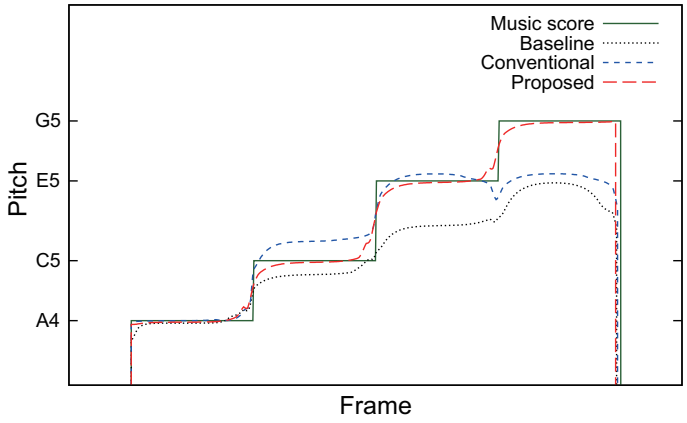

図 7 合成歌声の音高のプロット
Fig. 7 *Log F_0 sequences of synthesized singing voices.*

音高を合成可能であることが確認できる．

5. む す び

HMM 歌声合成の高精度化のため，音高正規化学習を提案した．学習データに含まれない音高をモデル化できない問題に対し，歌声の対数基本周波数と音符の音高の差分をモデル化することで，主観評価実験により学習データの音域外の合成音声の自然性の向上を確認した．今後の課題として，合成音声品質に強く影響するパラメータの調査が挙げられる．

謝辞 本稿で述べた研究開発の一部は SCOPE による．実験に使用した歌声データベースは名古屋工業大学酒向慎司氏が中心となって収録したものである．

表 1 学習所要時間
Table 1 *Computation time for the HMM training.*

手法	時間 (10 曲)	時間 (60 曲)
Baseline	9 時間 15 分	85 時間 30 分
Conventional	25 時間 17 分	273 時間 58 分
Proposed	9 時間 27 分	93 時間 00 分

© 2012 Information Processing Society of Japan

Vol.2012-MUS-94 No.15
Vol.2012-SLP-90 No.15
2012/2/4

参 考 文 献

1) T.Yoshimura, K.Tokuda, T.Masuko, T.Kobayashi, and T.Kitamura, “Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-based Speech Synthesis,” Proc.of Eurospeech, pp.2347–2350, 1999.
2) J. Yamagishi, “Average-Voice-based Speech Synthesis,” Ph. D. thesis, Tokyo Institute of Technology, 2006.
3) T.Yoshimura, K.Tokuda, T.Masuko, T.Kobayashi, and T.Kitamura, “Speaker Interpolation in HMM-based Speech Synthesis System,” Proc.of Eurospeech, pp.2523–2526, 1997.
4) K.Shichiri, A.Sawabe, K.Tokuda, T.Masuko, T.Kobayashi, and T.Kitamura, “Eignvoices for HMM-based Speech Synthesis,” Proc.of ICSLP, pp.1269–1272, 2002.
5) K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, “Recent Development of the HMM-based Singing Voice Synthesis System — Sinsy,” Proc.of SSW7, pp.211–216, 2010.
6) A.Mase, K.Oura, Y.Nankaku, and K.Tokuda, “HMM-based Singing Voice Synthesis System using Pitch-Shifted Pseudo Training Data,” Proc.of Interspeech, pp.845–848, 2010.
7) K.Saino, M.Tachibana, and H.Kenmochi, “An HMM-based singing style modeling system for singing voice synthesizers,” Proc.of SSW7, pp.252–257, 2010.
8) K.Tokuda, T.Yoshimura, T.Masuko, T.Kobayashi, and T.Kitamura, “Speech Parameter Generation Algorithms for HMM-based Speech Synthesis,” Proc. of ICASSP, pp. 1315–1318, 2000.
9) S.Imai, “Cepstral Analysis Synthesis on the Mel Frequency Scale,” Proc.of ICASSP, pp.93–96, 1983.
10) A. Kuramatsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kawabara, and K. Shikano, “ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis,” Speech Communication, vol.9, pp.357–363, 1990.
11) K. Shinoda and T. Watanabe, “MDL-based Context-Dependent Subword Modeling for Speech Recognition,” J.Acoust.Soc.Jpn.(E), vol.21, no.2, pp.79–86, 2000.
12) J.Yamagishi and T.Kobayashi, “Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training,” Proc.of IEICE Trans., vol.E-90D, no.2, pp.533–543, 2007.
13) H.Kawahara, M.K.Ikuyo, and A.Cheneigne, “Restructuring Speech Representations using a Pitch-Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-based F_0 Extraction: Possible Role of a Repetitive Structure in Sounds,” Proc.of Speech Communication, vol.27, pp.187–207, 1999.
14) K. Tokuda, T. Kobayashi, T. Chiba, and S. Imai, “Spectral Estimation of Speech by Mel-Generalized Cepstral Analysis,” Proc.of IEICE Trans., vol.75-A, no.7, pp.1124–1134, 1992.
15) H. Zen, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, “A Hidden Semi-Markov Model-Based Speech Synthesis System,” Proc.of IEICE Trans., vol.90-D, no.5, pp.825–834, 2007.

© 2012 Information Processing Society of Japan