Vol.2015-MUS-106 No.13
Vol.2015-EC-35 No.13
2015/3/3

HMM歌声合成における音声データの誤りに頑健なモデル化手法の検討

虫鹿 弘二[1,a)]　中村 和寛[1]　橋本 佳[1]　大浦 圭一郎[1]　南角 吉彦[1]　徳田 恵一[1]

概要：隠れマルコフモデル (HMM) に基づく歌声合成システムは，あらかじめ用意された歌声データから統計モデルを学習し，任意の歌声を合成する．HMM 歌声合成の性能は学習データに強く依存するため，高品質な歌声を合成するためには高品質な歌声データベースが必要になる．しかし，実際のデータベースには，歌い間違いやノイズなどの誤りが含まれていることが多い．特に，これからは音声合成の分野でも，インターネット上の大量のデータを学習に有効活用するという流れが加速していくと考えられ，そのような誤りを多く含むデータから高精度なモデルを学習する手法が必要である．そこで本稿では，学習データ内の誤りを局所的に除外することによる誤りに頑健なモデルの学習手法を提案し，主観評価実験により提案手法の有効性を評価する．

キーワード：HMM 音声合成，HMM 歌声合成，連結学習，コンテキストクラスタリング

A robust modeling technique against training data errors for HMM-based singing voice synthesis

Koji Mushika[1,a)]　Kazuhiro Nakamura[1]　Kei Hashimoto[1]　Keiichiro Oura[1]
Yoshihiko Nankaku[1]　Keiichi Tokuda[1]

1. はじめに

歌声合成の手法として，VOCALOID [1] を代表とする素片接続に基づく合成手法や，統計モデルの一つである隠れマルコフモデル (Hidden Markov Model; HMM) に基づく合成手法 [2] などが挙げられる．素片接続に基づく手法では，データベース内の音声波形素片を合成したい楽譜に従って選択・接続して歌声を合成する．録音した音声波形そのものを合成に用いるため，高品質な歌声が合成可能な反面，音声波形素片の接続歪みが生じやすい問題もある．

一方，HMM に基づく歌声合成では，あらかじめ用意した歌声データから，音色を表すスペクトルや音高を表す基本周波数，音の長さを表す継続長などの歌声の特徴を抽出し HMM によってモデル化する．通常，音声は音素単位でモデル化されるが，音符の高さや長さ，テンポや強弱記号などの楽譜情報を考慮することで，より精度の高いモデル化を行う．合成時には，与えられた楽譜に従ってモデルを連結し，動的特徴量を考慮してパラメータを生成することで，歌声を生成する．HMM 歌声合成は，素片接続に基づく合成手法にはない以下のような特徴を持つ．

(1) 与えられたデータに基づいてモデルを自動学習するため，声質だけでなく，プレパレーションやオーバーシュートなどの基本周波数の変化や，前のりや後のりなどの音符に対する発音タイミングの変化といった歌

[1] 名古屋工業大学
Nagoya Institute of Technology
[a)] mushika@sp.nitech.ac.jp

唱表現を再現できる．
(2) 比較的少ない量の学習データで高品質な歌声を合成できる．
(3) 学習データに含まれる波形をシステムに蓄積する必要がないため，フットプリントが小さい [3]．
(4) モデルパラメータを適切に変更することにより，様々な声質の歌声を合成できる．

特に (4) は他の手法では実現困難な特徴であり，実際に「声を真似る」話者適応手法 [4][5]，「声を混ぜる」話者補間手法 [6]，「声をつくる」固有声手法 [7] などの手法が提案されている．

HMM 歌声合成は統計的手法であるため，合成される歌声の品質は学習データに強く依存しており，高精度な歌声合成には十分な量の高品質なデータベースが必要になる．しかし，実際のデータベースには，歌詞・音高の歌い間違いやノイズなどの誤りが含まれていることが多い．特にインターネット上の大量の音声データを学習データとして利用するような場合には，データベース内に多くの誤りが含まれると考えられる．誤りを含む音声データを学習に用いた場合，誤りの周辺で適切な音素境界の推定が行われず，モデル推定精度に影響を与える可能性がある．通常，HMM テキスト音声合成 [8][9] の場合は，文章単位（数秒〜十数秒）の学習データを用いるため，誤りを含む文章を学習データから除外しても，総学習データ量への影響は少なく，合成される音声の品質に与える影響は小さいが，HMM 歌声合成の場合は，曲単位（数分）の学習データを用いることが多く，誤りを含む曲を除外することで，総学習データ量が大きく減少し，合成される歌声の品質に大きな影響を与える可能性がある．

そこで本研究では，音声データ内の誤りを局所的に除外することにより誤りによる影響を軽減し，誤りに頑健なモデルを学習する手法を提案する．音声データの誤りの有無を表す誤りフラグを楽譜情報に追加し，誤りを考慮してモデル化することにより，適切な音素境界の推定が行われ，誤り以外を表すモデルへの影響が軽減されると考えられる．合成時には，音声データの誤りのモデルを用いない事で合成音声の品質が向上することを期待する．

以下，2 章で HMM 歌声合成システムを紹介し，3 章では提案法となる音声データの誤りに頑健なモデルの学習手法について，4 章では主観評価実験について述べる．そして 5 章で全体をまとめ，今後の展望について述べる．

2. HMM 歌声合成システム

図 1 に HMM 歌声合成システムの概要を示す．本システムは学習部と合成部で構成されている．

2.1 学習部

HMM の学習のために，歌声データベースから各種特徴量を抽出する．特徴量は，歌声の音色や音高，歌唱表現を表すスペクトル，基本周波数，ビブラート [10] のパラメータから成り，スペクトルパラメータとしてメルケプストラムを，基本周波数パラメータとして対数基本周波数を，ビブラートパラメータとして対数基本周波数の揺らぎの振幅と周期を用いる．また，特徴量の時間的変動を考慮するためにこれらの Δ, Δ^2 [11] を求め，それらを結合した特徴ベクトルから HMM を学習する．基本周波数としては対数基本周波数が用いられるが，無声部で値が無いという特殊な時系列であるため，このような時系列を扱うことのできる出力確率分布 (Multi-Space Probability Distribution HMM; MSD-HMM) [12] を用い，曲単位で HMM を連結し，EM アルゴリズム [13] により HMM のパラメータの推定を繰り返す．

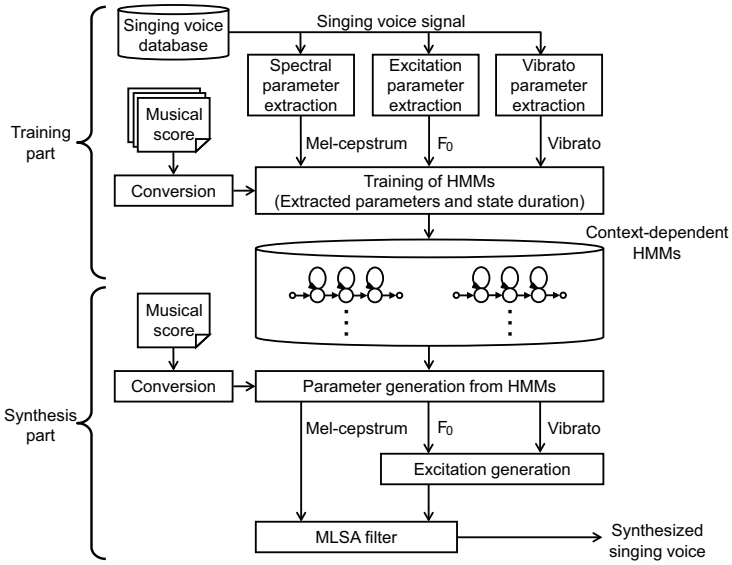

図 1 HMM 歌声合成システム

モデルの学習は音素単位で行うが，同じ音素であっても前後の音素や楽譜情報から得られる音符の高さや長さなどの組み合わせによりその特徴は大きく異なることが知られているため，コンテキストと呼ばれるこれらの変動要因を考慮したコンテキスト依存モデル [14] を用いることで，より詳細なモデル化を行う．一方で，コンテキストの組み合わせは膨大であるため，有限のデータからすべてのコンテキスト依存モデルの学習を行うことは困難である．この問題を解決するために決定木に基づくコンテキストクラスタリング [15] が用いられる．コンテキストクラスタリングによる決定木構築の例を図 2 に示し，手順を以下に示す．

(1) コンテキストに対して yes か no で答えられる質問を用意する．
(2) クラスタリングの対象となるすべての状態を統合したルートノードを作成する．
(3) すべてのリーフノードに対してすべての質問を適用し，分割を仮定した場合の尤度を計算する．

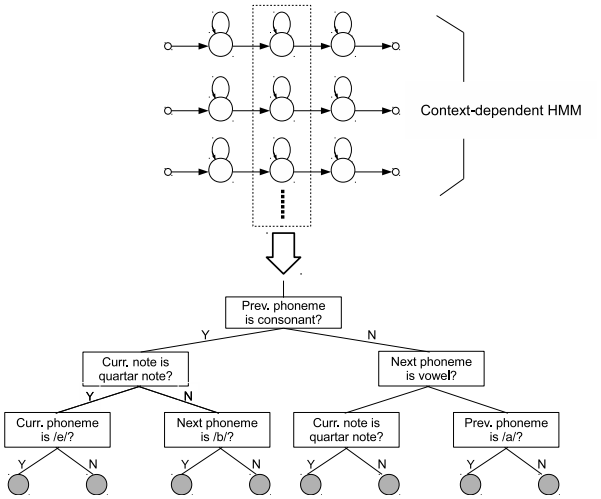

図 2 コンテキストクラスタリングによる決定木の構築

(4) 分割後の尤度が最大となるノードと質問の組を選択する．
(5) 分割前後の尤度差があらかじめ定められた閾値以下であればクラスタリングを停止し，そうでなければ選択したノードを分割して (3) へ戻る．

決定木のリーフノードには音響的に類似する状態が集まるため，それらの状態間でパラメータを共有することで，各モデルに十分な量の学習データを割り当てることが可能になる．

以上を踏まえ，学習部の流れを以下にまとめる．

(1) 歌声データを分析して特徴量を抽出し，特徴ベクトルを作成する．
(2) HMM のパラメータを初期化し，コンテキスト非依存 HMM の学習を行う．
(3) コンテキスト非依存 HMM にコンテキスト情報を与えてコンテキスト依存 HMM に変換する
(4) コンテキスト依存 HMM の学習を行う．
(5) メルケプストラム，対数基本周波数，ビブラートパラメータ，状態継続長に対してそれぞれ独立したコンテキストクラスタリングを行い，各々の決定木を構築する．
(6) パラメータが共有されたコンテキスト依存 HMM の学習を行う．

2.2 合成部

合成したい曲の楽譜から得られた歌詞や音高などのコンテキストを元にモデルを連結する．次に楽譜の音符長情報と学習した継続長モデルから各状態の継続長を求め，連結したモデルからパラメータ生成アルゴリズム [11] によりメルケプストラムと対数基本周波数，ビブラートのパラメータ系列を生成する．そして，ビブラートパラメータから計算した正弦波を対数基本周波数系列に重ね合わせてビブラートを再現し，生成されたパラメータに基づいて

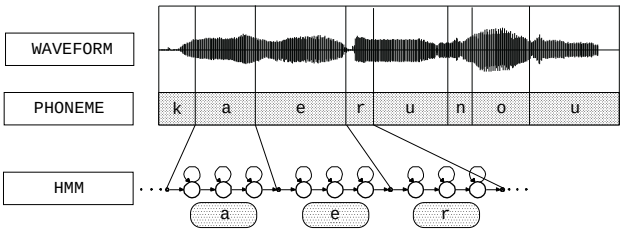

図 3 連結学習

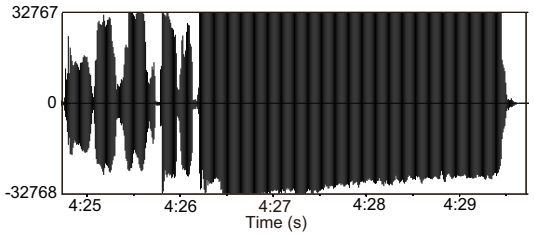

図 4 音声データ誤り（クリッピング）の例

MLSA (Mel Log Spectrum Approximation) フィルタ [16] を励振させることで歌声を合成する．

3. 音声データの誤りに頑健なモデルの学習手法

3.1 連結学習における音声データの誤りの影響

HMM 歌声合成では，コンテキスト依存モデルを楽譜情報を元に曲単位で連結し，EM アルゴリズム [13] を用いた連結学習を行う．連結学習の例を 図 3 に示す．まず，与えられた楽譜の歌詞情報から得られる音素列 “k a e r u n o u” と音符の高さや長さといったコンテキスト情報を元にして，対応する音素単位の HMM を連結する．次に，連結した HMM を用いて，学習データの尤度がより高くなるように，各モデルの出力確率分布と遷移確率の推定を繰り返す．

合成される歌声の品質は学習データの品質に強く依存するため，モデルの学習には高品質な学習データを用いることが望ましい．しかし，実際に収録された歌声データベースにはモデル化に適さない以下のような歌声データが含まれていることが多い．

- 歌詞，音符長，音高などの歌い間違い
- 曲中で用いられる他言語の歌詞
- 咳などの文字で表記できない音声
- クリッピング，反響音，雑音

本論文では，モデル化に適さないこれらのデータを音声データ誤りと呼ぶ．歌詞の歌い間違いのように，音声に合わせて楽譜を修正することで正しい学習データとして用いることができる誤りもあるが，図 4 に示すクリッピングや咳のように楽譜を修正するだけでは対処できない誤りも存在する．音声データ誤りを含むデータを連結学習に用いた場合，音声データ誤りの周辺で適切な音素境界の推定が行われず，周辺音素のモデルの推定にも影響を与える可能性

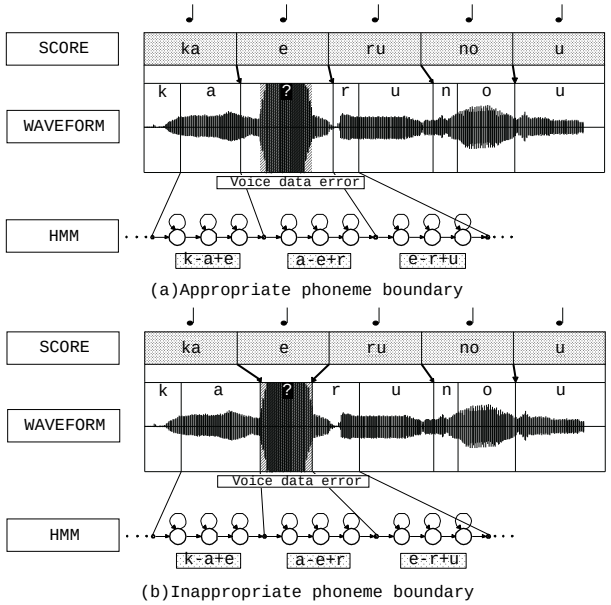

図 5 音声データ誤りが音素境界の推定に与える影響

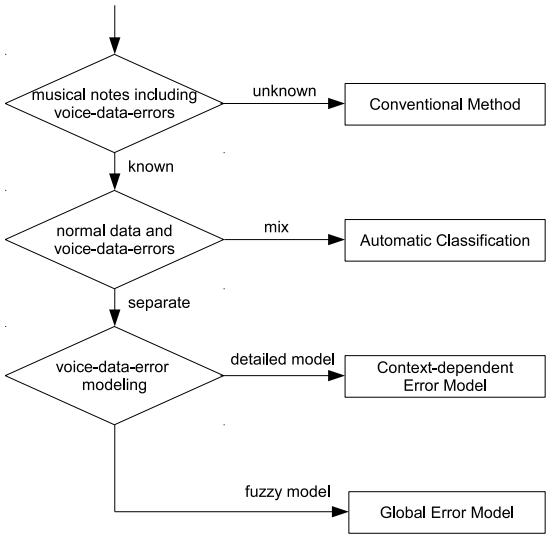

図 6 モデル化手法の関係

がある．音声データ誤りが音素境界の推定に与える影響の例を図 5 に示す．“k a e r u n o u ” の “e” にクリッピングが存在する場合，図 5 (a) のような正しい音素境界が推定されず，図 5 (b) のように誤った音素境界が推定され，“e” のモデルだけでなく，周辺音素のモデルの推定精度も低下する可能性がある．

通常，HMM テキスト音声合成の場合は，文章単位（数秒〜十数秒）の学習データを用いるため，誤りを含む文章を学習データから除外しても，総学習データ量への影響は少なく，合成音声の品質に与える影響は小さいが，HMM 歌声合成では曲単位（数分）の学習データを用いるため，誤りを含む曲を除外することで，総学習データ量が大きく減少し，合成歌声の品質に大きな影響を与える可能性がある．また，話し声の収録と比較して，歌声の収録は曲ごとに歌唱の練習が必要であるため，収録にかかるコストが高い．このため，音声データ誤りを含む曲を除外せずに有効活用できる方法が必要である．

そこで，音声データ誤りに該当する音符に音声データ誤りの有無を表すコンテキストを付与し，音声データ誤りを考慮したモデル化を行う，音声データ誤りに頑健な学習手法を提案する．“k a e r u n o u” の “e” にクリッピングが存在している場合でも，音声データ誤りのモデルを割り当てて連結学習を行うことで，図 5 (a) に示すような正しい音素境界が推定されると考えられる．

3.2 音声データ誤りのモデル化手法

音声データ誤りのモデルを学習するために，音声データ誤りの有無を表す誤りフラグをコンテキストに導入する．あらかじめ，人手で音声データ誤りを判別し，音声データ誤りに該当する音符のコンテキストに誤りフラグを付与する．合成時は，誤りフラグを用いない事で音声データ誤りの影響が軽減されると考えられる．ただし，音声データ誤りには様々なパターンが考えられ，その出現箇所や頻度も未知であることから，あらゆる音声データ誤りを適切にモデル化することは難しい．そこで，音声データ誤りのモデル化に関して以下の 3 手法を提案する．各手法の関係を 図 6 に示す．

Automatic Classification 音声データ誤りのモデルと音声データ誤り以外のモデルを分類するかどうかを自動で決定する．

Context-dependent Error Model 音声データ誤りのモデルと音声データ誤り以外のモデルをあらかじめ分類し，音声データ誤りを詳細にモデル化する．

Global Error Model 音声データ誤りのモデルと音声データ誤り以外のモデルをあらかじめ分類し，音声データ誤りを曖昧にモデル化する．

これらの手法の実現のためにコンテキストクラスタリングを用いてモデル化を行う．各手法のコンテキストクラスタリングの例を 図 7 に示す．

Automatic Classification では，音声データ誤りの有無がコンテキストとして有効であると仮定して，音声データ誤りのモデルと音声データ誤り以外のモデルが混在する決定木を構築する．決定木を構築する際の質問には誤りフラグに関する質問も含まれており，音声データ誤りのモデルと音声データ誤り以外のモデルを分類するかどうかは尤度最大化基準に基づき自動的に決定される．また，音声データ誤りの音響的特徴が楽譜情報と何らかの関係性を持つかどうかを確認するために，以下の 2 手法を試みる．どちらの手法においても，あらかじめ誤りフラグの情報を用いて，音声データ誤りのモデルと音声データ誤り以外のモデルを分類する．**Context-dependent Error Model** では，音声データ誤りと楽譜情報との間に何らかの関係性

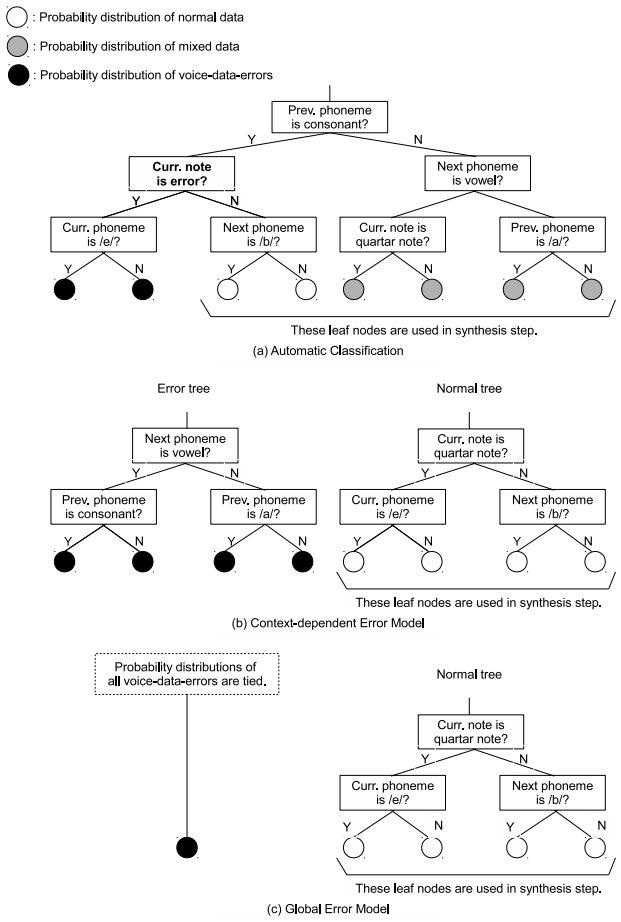

図 7 誤りフラグを用いたコンテキストクラスタリング

があると仮定し，音声データ誤りのモデルの決定木と音声データ誤り以外のモデルの決定木の 2 つの決定木を構築することで，音声データ誤りのモデルを楽譜の情報に基づくコンテキストに従って詳細にモデル化する．**Global Error Model** では，音声データ誤りには様々なパターンがあるためモデル化することが困難であると仮定し，音声データ誤りのモデルを一つにまとめることで音声データ誤りの曖昧なモデル化を行う．

4. 主観評価実験

4.1 実験条件

提案法の有効性を評価するため，主観評価実験を行った．学習には女性 1 名による J-POP 楽曲 30 曲, 約 150 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．STRAIGHT [17] 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 [18] 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係数，対数基本周波数，ビブラートの揺らぎの振幅と周期，またそれらの Δ, Δ^2 を特徴量として用いた．モデルは 5 状態の left-to-right 型 HSMM [19] とした．学習データの音素境界情報の初期値は，確定的アニーリング EM (Deterministic Annealing EM; DAEM) アルゴリズム [20] により求めた．コンテキストクラスタリングの分割停止基準に最小記述長 (Minimum Description Length; MDL) 基準 [21] を用いることでモデルのパラメータ数を決定した．また，学習データ内の音高の偏りを吸収し，学習データに含まれない音高を合成するために，音高正規化学習 [22] を行った．音声データ誤りの有無を表す誤りフラグは，音声データに対して音符単位で人手により誤りかどうかの判別を行い，コンテキスト情報として付与した．誤りがあると判別された音符は学習データ全体の 1.4%であった．なお本稿では単一言語の歌声合成システムを仮定しており，曲中で用いられる他言語の歌詞も音声データ誤りとしている．**Global Error Model** の音声データ誤りを一つにまとめたモデルに関して，音声データ誤りの確率分布をより曖昧にした場合の影響を確かめるため，音声データ誤りの確率分布の平均と分散を割り当てられた学習データから推定 **(estimate)** したもの，音声データ誤りの確率分布を学習データ全体の平均と分散で固定 **(fix)** したものの 2 種類を用いた．比較手法は，**Conventional Method**, **Automatic Classification**, **Context-dependent Error Model**, **Global Error Model (estimate)**, **Global Error Model (fix)** の計 5 手法とした．評価には童謡 10 曲の 38 フレーズを用い，被験者 10 名に被験者毎にランダムに選択した 10 フレーズを聞かせ，歌声の自然性について 5 段階 MOS で評価させた．被験者は防音室においてヘッドフォンを装着した状態で受聴し，被験者が聴きやすいレベルになるよう被験者本人に音圧を調整させた．

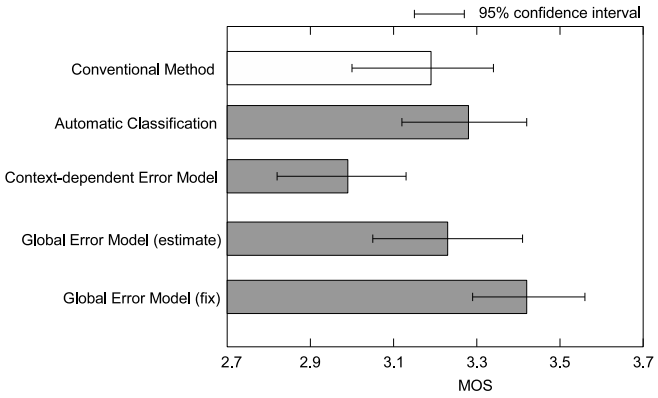

図 8 主観評価実験結果

4.2 実験結果

図 8 に主観評価実験結果を示す．**Automatic Classification** が **Conventional Method** より高い MOS を示しているのは，コンテキストクラスタリングにおいて誤りフラグに関する質問が選択され，音声データ誤りのモデルが分離されたことにより，合成時に音声データ誤りのモデルが用いられなかったためだと考えられる．また，**Context-dependent Error Model** は **Conventional Method** よ

りも低い MOS を得たが，この原因は，音声データ誤りの音響的特徴と楽譜情報のコンテキストの間に明確な関係が見い出せず，誤りのパターンが分類されなかったためだと考えられる．これに対し，**Global Error Model (estimate)** 及び **Global Error Model (fix)** は高い MOS を示しており，音声データ誤りの確率分布を曖昧にすることで，連結学習時の音声データ誤りが周辺のモデルに与える影響が軽減されたと考えられる．また **Global Error Model (estimate)** に比べて **Global Error Model (fix)** が高い自然性を得ていることから，様々なパターンを含む音声データ誤りを適切にモデル化することは困難であり，音声データ誤りの確率分布を曖昧にするほど連結学習における周辺のモデルへの影響が軽減されると考えられる．

5. むすび

本論文では，HMM 歌声合成における，音声データの誤りに頑健なモデルの学習手法を提案した．音声データ誤りがモデルの学習に与える影響を軽減するために，音声データ誤りの有無を表す誤りフラグをコンテキスト情報として付与し，音声データ誤りを 3 種類の手法でモデル化して比較した．主観評価実験の結果，あらゆる種類の音声データ誤りをすべて適切にモデル化することが困難であることから，音声データ誤りを曖昧にモデル化することで，連結学習時に周辺のモデルへの影響が軽減され，合成された歌声の自然性が向上することを確認した．今後の課題としては，より多くの音声データ誤りを含む音声データベースを用いた歌声合成実験，別の歌唱者での実験などが挙げられる．

謝辞 本研究の一部は，科学技術振興財団「JST」の戦略的基礎研究推進事業「CREST」による支援を受けた．

参考文献

[1] H. Kenmochi and H. Ohshita, “VOCALOID - Commercial Singing Synthesizer based on Sample Concatenation,” in Proc. Interspeech, Special session, 2007.
[2] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, “Recent Development of the HMM-based Singing Voice Synthesis System — Sinsy,” in Proc. Speech Synthesis Workshop, pp. 211–216, 2010.
[3] 森岡祐介, 片岡俊介, 全炳河, 南角吉彦, 徳田恵一, 北村正, “HMM 音声合成器の小型化に関する検討, ” 日本音響学会秋期研究発表会, pp. 325–326，2014.
[4] J. Yamagishi, “Average-Voice-based Speech Synthesis,” Ph. D. thesis, Tokyo Institute of Technology, 2006.
[5] 大浦圭一郎, 間瀬絢美, 山田知彦, 徳田恵一, 後藤真孝, “Sinsy: 「あの人に歌ってほしい」をかなえる HMM 歌声合成システム, ” 情報処理学会研究報告, vol. 2010-MUS-86, no. 1, pp. 1–8, 2010.
[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Speaker Interpolation in HMM-based Speech Synthesis System,” in Proc. Eurospeech, pp. 2523–2526, 1997.
[7] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Eignvoices for HMM-based Speech Synthesis,” in Proc. ICSLP, pp. 1269–1272, 2002.
[8] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, “Speech Synthesis based on Hidden Markov Models,” IEEE, vol. 101, no. 5, pp. 1234–1252, 2013.
[9] 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, vol. J83-D-II, no. 11, pp. 2099–2107, 2000.
[10] 山田知彦, 武藤聡, 南角吉彦, 酒向慎司, 徳田恵一, “HMM に基づく歌声合成のための ビブラートモデル化, ” 情報処理学会研究報告, vol. 2009-MUS-80, no. 5, pp. 1–6, 2009.
[11] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, “Speech synthesis from HMMs using dynamic features,” in Proc. ICASSP, pp. 389–392, 1996.
[12] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Hidden markov models based on multi-space probability distribution for pitch pattern modeling,” in Proc. ICASSP, vol. 1, pp. 229–232, 1999.
[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum-likelihood from incomplete data via the EM algorithm,” J. Royal Statist. Coc. Ser. B (methodological), vol. 39，pp. 1–38，1977.
[14] K. F. Lee, “Context-dependent phonetic Hidden Markov Models for Speaker-independent Continuous Speech Recognition,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 38, no. 4, pp. 599–609, 1990.
[15] S. Young, J. J. Odell, and P. Woodland, “Tree-based state tying for high accuracy acoustic modeling,” in Proc. ARPA Workshop on Human Language Technology, pp. 307–312, 1994.
[16] 今井聖, 住田一男, 古市千恵子, “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, 1983.
[17] 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,” Speech Communication, vol. 27, pp. 187–207, 1999.
[18] 徳田恵一, 小林隆夫, 千葉健司, 今井聖, “メル一般化ケプストラム分析による音声のスペクトル推定, ” 信学論 (A), vol. 75-A, no. 7, pp. 1124–1134, 1992.
[19] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “A hidden semi-Markov model-based speech synthesis system,” IEICE Trans. Inf. & Sys., vol. 90-D, no. 5, pp. 825–834, 2007.
[20] N. Ueda, R. Nakano, “Deterministic annealing EM algorithm,” Neural Networks, vol. 11, pp. 271–282, 1998.
[21] 篠田浩一, 渡辺隆夫, “情報量基準を用いた状態クラスタリングによる音響モデルの作成, ” 信学技報, vol. SP96-79, pp. 9–16, 1996.
[22] K. Oura, A. Mase, Y. Nankaku, and K. Tokuda, “Pitch adaptive training for HMM-based singing voice synthesis,” in Proc. ICASSP, pp. 5377–5380, 2012.