Vol.2010-MUS-86 No.1
2010/7/28

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

大浦 圭一郎[†1] 間瀬 絢美[†1] 山田 知彦[†1]
徳田 恵一[†1] 後藤 真孝[†2]

近年，コンピュータによる歌声合成が注目を集めている．中でも隠れマルコフモデル (hidden Markov model; HMM) に基づく歌声合成では，歌い手の特徴を歌声データと対応する楽譜から自動的に学習することができる．2009 年 12 月，無料のオンラインサービス「HMM 歌声合成システム: Sinsy」を開始した．ユーザーは楽譜をウェブサイトにアップロードすることで，任意の楽譜に対応した歌声を合成することができる．但し，Sinsy の歌声モデルには 70 曲で学習した特定話者モデルを用いており，新しい歌い手の歌声モデル追加の際の収録コストが高くなる問題があった．本稿では Sinsy のシステム構成について述べるとともに，話者適応手法により少量のデータから所望の歌い手の特徴を再現した歌声を合成することを検討する．

Sinsy — An HMM-based singing voice synthesis system which can realize your wish “I want this person to sing my song”

Keiichiro Oura,[†1] Ayami Mase,[†1] Tomohiko Yamada,[†1]
Keiichi Tokuda[†1] and Masataka Goto [†2]

A statistical parametric approach to singing voice synthesis based on hidden Markov models (HMMs) has been grown over the last few years. In this approach, spectrum, excitation, and duration of singing voices are simultaneously modeled by context-dependent HMMs, and waveforms are generated from HMMs themselves. Since December 2009, we started a free on-line service named “Sinsy.” By uploading musical scores represented by MusicXML to the Sinsy website, users can obtain synthesized singing voices. However, a high recording cost may be required to train new singer’s model because a speaker-dependent model trained by using 70 songs is used in Sinsy. The present paper describes the recent developments of Sinsy and a speaker adaptation technique to generate waveforms from a small amount of adaptation data.

1. はじめに

近年，隠れマルコフモデル (hidden Markov model; HMM) に基づくテキスト音声合成[1] の研究が盛んに行われるようになってきた．この手法では，音声データベースに基づいて 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フットプリントが小さい，統計モデルに基づく手法であるためモデルパラメータを適切に変換することで多様な音声が合成できるなどの特徴を持ち，これまでに話者適応[2]，話者補間[3]，固有声[4] などが提案されてきた．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特徴を自動学習し，その特徴を再現した歌声を合成することができる．2009 年 12 月，我々は無料のオンラインサービス「HMM 歌声合成システム: Sinsy」[6] を開始した．Sinsy は HTS[7] や hts_engine API[8]，SPTK[9]，STRAIGHT[10]，CrestMuseXML Toolkit[11] といったオープンソースのソフトウェアを用いて構築されており，ユーザーは MusicXML[12] で記述された歌詞付き楽譜をウェブサイトにアップロードすることで任意の楽曲に対応する歌声を合成することができる．

Sinsy の歌声モデルには女性 1 名の歌い手による歌声 70 曲で学習した特定話者モデルを用いている．一般ユーザーにとって同様の歌声モデルを追加するための歌声収録は簡単ではなく，十分な量の歌声データを用意することは難しい可能性がある．そこで本稿では，Sinsy のシステム構成について述べるとともに，制約付き最尤線形回帰 (Constrained Maximum-Likelihood Linear Regression; CMLLR)[13] に基づく話者適応手法を導入した結果について紹介する．

以降，2 節で HMM 歌声合成システム Sinsy の概要を紹介し，3 節で Sinsy に導入した手法の詳細，4 節で本稿で導入する話者適応手法，5 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 6 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. Sinsy

本節では HMM 歌声合成システム Sinsy の概要や構築に用いたソフトウェア，オンラインサービスの運用などついて述べる．

2.1 HMM 歌声合成システム

歌声合成の研究に関しては長い歴史があり，様々な方式が検討されてきた．最近では，Vocaloid[14] の技術を用いた歌声合成ソフトウェアが市販され，広く利用されるようになって

†1 名古屋工業大学
Nagoya Institute of Technology (NIT)
†2 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

きている．一般の人々の認知度も高くなってきており，自分の声や好きな歌い手の声で好きな曲を簡単に歌わせたいという要望が高まっている．このような歌声合成ソフトウェアにおいて自然に歌わせるために必要となる調整作業は，作品制作上の創造的な部分ではあるものの，一般ユーザーには敷居が高すぎるという問題があった．このため，自ら歌うことにより調整を行う VocaListener[15] 等の開発が行われた．このような流れの中で，我々は歌声合成システム Sinsy のオンラインデモを公開した．Sinsy では HMM を利用することにより，楽譜とそれに対応した歌声との関係をモデル化している．スペクトルや基本周波数，時間情報などを全て同時に自動学習する方式のため，声質や音量は元より，基本周波数の変化パターンによって表されるプレパレーション，オーバーシュートや，前ノリ，後ノリ等の歌唱スタイルに関しても自動的に学習される．

現在，一般に用いられている Vocaloid 等の歌声合成システムは波形接続型と呼ばれる方式を採用している．この方式は蓄積された歌声の波形データを接続することにより，任意の楽譜に対応する歌声を合成するものである．これに対し，HMM 歌声合成は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，声の特徴や歌唱表現を再現する歌声を合成することができ，調整作業が必要ない．
(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．
(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」[2],「声を混ぜる」[3] などの手法が開発されている．

図 1 に HMM 歌声合成システム[5] の概要を示す．学習部と合成部に分かれており，学習部では，歌声データからメルケプストラム係数[16] などのスペクトルパラメータと基本周波数パラメータが抽出され，コンテキスト依存[⋆1]HMM でモデル化される．また，状態継続長に関しても同時にモデル化される．なお，図では省略されているが，3.2 節で述べるビブラートに関する特徴もスペクトルパラメータや基本周波数パラメータと同様にモデル化される．合成部では，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・基本周波数パラメータがパラメータ生成アルゴリズム[17] によって生成され，MLSA フィルタ[18] により歌声が合成される．

⋆1 3.1 節を参照.

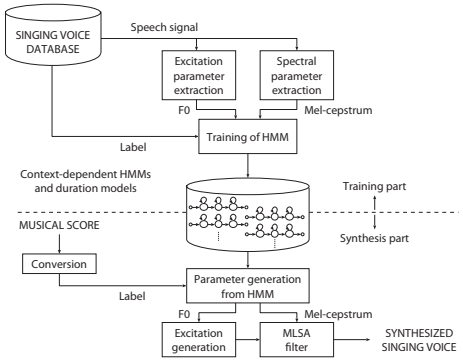


図 1 *HMM* 歌声合成システムの概要
Fig. 1 *The overview of the HMM-based singing voice synthesis system.*

2.2 ソフトウェア

Sinsy はオープンソースのソフトウェアを用いて構築されている．ここでは各モジュールに用いたソフトウェアについて述べる．

- Speech Signal Processing Toolkit (SPTK)
  SPTK-3.3[9] は UNIX 環境向けの信号処理ツールである．このソフトウェアは修正 BSD ライセンス[19] で配布されており，Sinsy の構築に関しては，メルケプストラム分析[16] などの SPTK のコマンドを利用した．
- STRAIGHT
  STRAIGHT V40[10), 20)] は音声の分析・合成のための信号処理ソフトウェアである．先進的な信号処理アルゴリズムが導入されており，Sinsy では音声スペクトルの抽出に STRAIGHT を利用している．なお，STRAIGHT は独自のライセンス形態で配布されている．
- HMM-based Speech Synthesis System (HTS)
  HTS-2.1.1[7), 21)] は HMM に基づく音声合成に関する研究・開発のためのプラットフォームを提供するオープンソースソフトウェアであり，Sinsy では歌声モデルの学習に用いている．本ソフトウェアは，ケンブリッジ大学で公開されている HMM ツールキット HTK[22] へのパッチの形で公開されており，国内外の様々な研究機関・企業で利用されている．なお，パッチは修正 BSD ライセンス[19] で配布されているが，一度パッチをあてると，ユーザーは HTK のライセンス[⋆2]に従わなければならないことに注意されたい．

⋆2 HTK のライセンスは商用利用を禁止している．

- CrestMuseXML Toolkit (CMX)
  歌詞付き楽譜を扱うために MusicXML-2.0[12] を採用し，その解析には CMX-0.50[11), 23)] を用いた．図 2 に MusicXML の例を示す．MusicXML では音符の音高や長さ，歌詞だけでなく，調や拍子，強弱記号等を記述することができる．
- HMM-based Speech Synthesis Engine (hts_engine API)
  HTK 及びそのパッチの形で与えられる HTS はそのライセンスの性格上，商用のアプリケーションに組み込むことは難しい．そこで Sinsy の波形生成部には，SourceForge のウェブサイトで公開されている HMM 音声合成エンジン hts_engine API-1.03[8] を用いた．HTK や HTS に依存しない形で開発されており，修正 BSD ライセンス[19] で配布されているため，商用のアプリケーションへ組み込む際にも問題が生じにくい．

2.3 オンラインサービス

頻繁なシステムアップデートが容易なことから，ウェブベースのユーザーインターフェースを採用した（図 3）．ユーザーは MusicXML で記述された歌詞付き楽譜をアップロードすることで任意の歌声を合成することができ，さらにウェブインターフェースから与えることのできるパラメータを変更することで声質やビブラートの強度を調整することができる．ただし，HMM 歌声合成の性質上，学習データに無い音高は合成できないため，音域を越えた MusicXML や，サーバーの負荷軽減のために 5 分以上の MusicXML を棄却している．

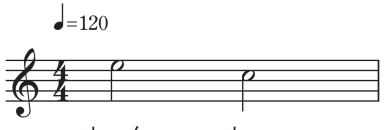
2010 年 1 月から 4 月までの Sinsy へのアクセスを計測したところ，そのページビューは合計一万件以上にのぼり，千種類以上の MusicXML ファイルがアップロードされていることから，歌声情報処理技術への期待の高さを垣間見ることができる．なお，アップロードされた MusicXML から正しく歌声が合成される割合は約 70%だった．残りの約 30%には，上記の制限による棄却以外にも，不正な MusicXML のアップロードにや，パーサーのバグに起因するものなどがあった．

3. 導 入 手 法

Sinsy では，HMM 音声合成システムをベースに歌声合成に特化した手法をいくつか導入している．本節ではそれらの手法について述べる．

3.1 コンテキストの定義

音声は連続的な観測系列であるため，同一音素でも文脈的な要因（コンテキスト）によって音響的な特徴量が変化することが知られている．従って，このようなコンテキストを考慮したモデル化により合成音声の品質が向上すると考えられる[24)]．HMM テキスト音声合成システム[1)] では，読み上げ音声に関わる音素や品詞，アクセントなどのコンテキストを定義している．しかし，歌声には音高やテンポ，調，拍子などのコンテキストが考えられるため，Sinsy では新しく歌声合成に特化したコンテキストを定義した[25)]．定義したコンテキストを




```
<?xml version="1.0" encoding="UTF-8">
 <!DOCTYPE score-partwise PUBLIC
 "-//Recordare//DTD MusicXML 2.0 Partwise//EN"
 "http://www.musicxml.org/dtds/partwise.dtd">
 ...
 <measure number="1">
  <attributes>
   <divisions>1</divisions>
   <key>
    <fifths>0</fifths>
    <fifths>major</fifths>
   </key>
   <time>
    <beats>4</beats>
    <beat-type>4</beat-type>
   </time>
 ...
  </attributes>
  <sound tempo="120"/>
  <note>
   <pitch>
    <step>E</step>
    <octave>5</octave>
   </pitch>
   <duration>2</duration>
   <type>half</type>
   <lyric>
    <text> しぃん </text>
   </lyric>
  </note>
  <note>
   <pitch>
    <step>C</step>
    <octave>5</octave>
   </pitch>
   <duration>2</duration>
   <type>half</type>
   <lyric>
    <text> しぃ</text>
   </lyric>
  </note>
 </measure>
</part>
</score-partwise>
```

図 2 *MusicXML* の例
Fig. 2 *An example of MusicXML.*

以下に示す．

- 音素
  - 当該音素から前後 2 個までの音素
- モーラ

Vol.2010-MUS-86 No.1
2010/7/28

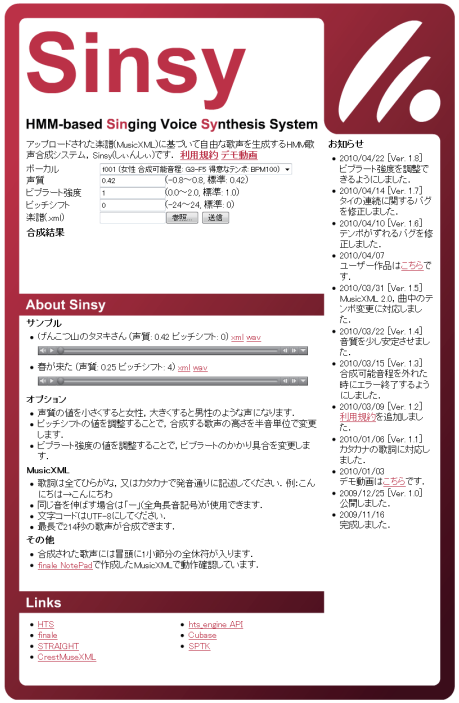


図 3 *HMM* 歌声合成システム「*Sinsy*」
Fig. 3 *HMM-based Speech Synthesis System — Sinsy.*

  - 先行／当該／後続モーラ内の音素数
  - 先行／当該／後続モーラの音符内位置
- 音符
  - 先行／当該／後続音符の音高，調，拍子，テンポ，音符長，小節内位置，フレーズ内位置
  - 当該音符と先行／後続音符間のタイ，スラー
  - 当該音符の強弱記号
  - 当該音符からアクセント，スタッカートまでの長さ
  - 当該音符のクレッシェンド，デクレッシェンド内位置
  - 先行音符／後続音符に対する当該音符の音程
- フレーズ

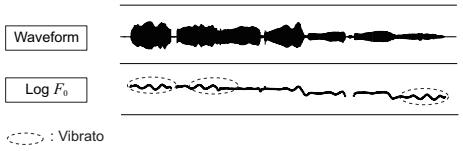


図 4 対数基本周波数系列の中のビブラート区間
Fig. 4 *An example of vibrato parts in a log F_0 sequence.*

  - 先行／当該／後続フレーズ内の音素数，モーラ数
- 曲
  - 曲内のモーラ数，音符数，フレーズ数

これらのコンテキストは歌詞付き楽譜から自動で決定される．階層毎に整理して対称性を考慮することにより，歌声に必要なコンテキストを網羅することができたと考えられる．

3.2 ビブラートモデル

ビブラート[26), 27)] は音高や音量等を周期的に揺らす歌唱表現である．その対数基本周波数の軌跡の例を図 4 に示す．ビブラートのかかるタイミングやその変動は歌い手によって異なるため，楽譜に含まれていないものの，ビブラートはモデル化すべき重要な歌唱表現の一つであると考えられる．しかし，従来の HMM 歌声合成システムでは，ビブラートのような音符内で発生する微細な変動はモデルの学習過程で平滑化されてしまい，ビブラートを含む歌声を精度よく再現することはできなかった．そこで，Sinsy ではビブラートを歌声データから自動的に学習するためにビブラートモデル[28)] を導入した．

現在の実装では，簡単のため，ビブラート現象の中で音高の周期的な揺らぎのみを扱い，音量等については考慮しないこととしている．時刻 $t = 0, 1, \ldots, T+1$ がビブラート区間 $i = 0, 1, \ldots, M-1$ に含まれるとき，ビブラート $v(\cdot)$ は次式でモデル化できる．

$$v\left(m_a(t), m_f(t), i\right) = m_a(t) \sin\left(2\pi m_f(t) f_s \left(t - t_0^{(i)}\right)\right), \tag{1}$$

ただし，f_s はフレーム周期，$t_0^{(i)}$ はビブラート区間 i の開始時間，$m_a(t)$ と $m_f(t)$ はそれぞれ時刻 t におけるビブラートの振幅と周波数である．対数基本周波数系列を用いたビブラートの推定[27)] を図 5 に示す．ここで，$c = \log 2/1200$ であり，これは cent 単位から対数基本周波数単位へのスケーリングのための定数である．なお，観測ベクトル $\boldsymbol{o}_t$ が，スペクトルパラメータ $\boldsymbol{o}_t^{(spec)}$，基本周波数パラメータ $\boldsymbol{o}_t^{(F_0)}$，ビブラートパラメータ $\boldsymbol{o}_t^{(vib)}$ で構成されるとき，s 番目の状態の状態出力確率 $b_s(\boldsymbol{o}_t)$ は次の式で与えられることになる．

$$b_s(\boldsymbol{o}_t) = p_s^{\gamma_{spec}}\left(\boldsymbol{o}_t^{(spec)}\right) \cdot p_s^{\gamma_{F_0}}\left(\boldsymbol{o}_t^{(F_0)}\right) \cdot p_s^{\gamma_{vib}}\left(\boldsymbol{o}_t^{(vib)}\right) \tag{2}$$

ただし，$\gamma_{spec}, \gamma_{F_0}, \gamma_{vib}$ は 3 種類の観測データの出力確率それぞれに対する重みである．

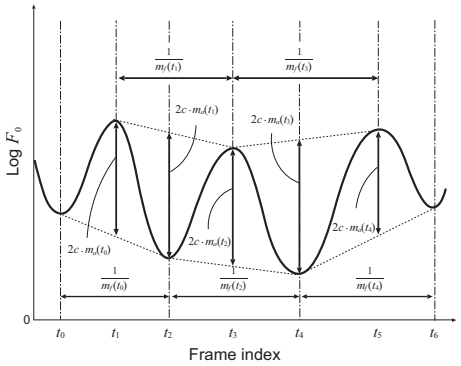


図 5 ビブラートの推定
Fig. 5 *The vibrato parameter analysis.*

3.3 音高シフトによる疑似学習データ

3.1 節で述べたコンテキスト依存モデルは，コンテキストの組みわせにより膨大な数になり，それらすべてが学習データに現れることは期待できない．この問題は，コンテキストクラスタリングの手法[24] により，学習データに現われないコンテキストに対応するモデルを生成することで解決することができる．ところが，音高に関しては，異なる音高に対応するモデルを代替モデルとして用いることはできないという問題がある．なぜなら，代替モデルを用いた場合には，楽譜と異なる音高を生成することになってしまうためである．このため音高に関しては，合成時に必要となる音高全てのそれぞれに関してモデル学習に十分な数だけ現われる必要がある．そこで Sinsy では，音高シフトによる疑似学習データを追加する手法[30] を導入している．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，音高をシフトした疑似学習データは対数基本周波数パラメータを上下に半音ずつシフトするだけで簡単に用意でき，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 6 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カバーできていることが確認できる．ただし，スペクトル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した．

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[31] に基づくコンテキストクラスタリングを用いて決定木を構築すると，約 3 倍の大きさの決定木が構築されてしまう．Sinsy では音高シフト前の学習データ量に適したパラメータ数に抑えるため，MDL 基準に適切な重みを設定している．HMM 歌声合成において，決定木に基づくコンテキストクラスタリングでは，分割前後の記述長を計算し，

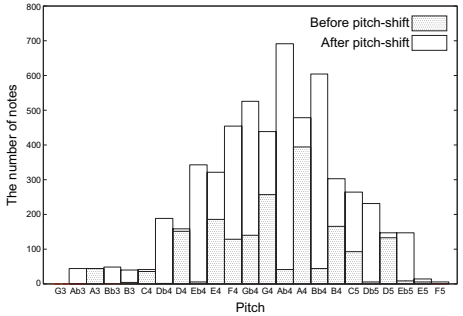


図 6 学習データ *10* 曲に含まれる音高の音符数の分布
Fig. 6 *The distribution of pitch in training data (10 songs).*

その差が閾値を越える場合に分割が行われ，閾値を越える分割が無くなるまで分割が繰り返される．従って，小さい閾値を用いた場合には大きな決定木になり，逆に大きな閾値を用いた場合には小さな決定木になる．ノード S が質問 q でノード S_{q+} と S_{q-} に分割されるとき，記述長の差 Δ_q は次のように計算される．

$$\Delta_q = \mathcal{L}(S) - \left\{\mathcal{L}\left(S_{q+}\right) + \mathcal{L}\left(S_{q-}\right)\right\} + \alpha\frac{N}{2}\log\Gamma(S_0) \tag{3}$$

ここで，$\mathcal{L}(\cdot)$ は対数尤度，S_0 はルートノード，α は重み，N は分割によって増えるパラメータ数，$\Gamma(\cdot)$ は事後確率である．

3.4 楽譜情報を用いたモデル学習の高速化

HMM 歌声合成システムでは，HMM テキスト音声合成で用いる読み上げ音声に比べて一つのデータの長さが長いため，モデルの学習に，より大きな計算量が必要になる．HMM の学習では EM アルゴリズムにより最尤パラメータを推定するが，フォワードバックワードアルゴリズムを用いた尤度計算時に，全ての時刻 $t = 0, \ldots, T+1$ と状態 $i = 1, \ldots, M$ に対して前向き確率 $\alpha_i(t)$ と後ろ向き確率 $\beta_i(t)$ を計算する必要があり，長い観測系列を扱う際にはその組み合わせ $M(T+1)$ が膨大となる．これまで，計算コストの削減のために尤度差による枝刈りが広く用いられてきたが，大幅な枝刈りは最尤パスの推定誤りを引き起こす問題があった．そこで Sinsy では，楽譜情報を用いた枝刈り手法[32] を導入した（図 7）．歌声はその性質上，音符の位置と実際の発声が大きくずれることがないため，時刻 t の観測が属する状態 i は音符の位置により限定される．楽譜上の k 番目の音符の開始時間を $Note_{start}(k)$，終了時間を $Note_{end}(k)$ とし，楽譜と実際の歌声とのずれを吸収するためのマージンをそれぞれ $Margin_{start}(k)$，$Margin_{end}(k)$ とすると，時刻 t が

$$t = Note_{start}(k) - Margin_{start}(k) \tag{4}$$

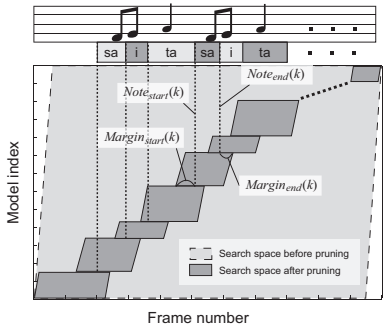


図 7 楽譜情報を用いた枝刈り
Fig. 7 *The pruning approach using note boundaries.*

から

$$t = Note_{end}(k) + Margin_{end}(k) \tag{5}$$

の区間における当該音符の歌詞に対応した状態のみ探索すればよく，効率的な枝刈りが期待できる．この手法により，状態系列の推定誤りを抑制し，さらに計算量を削減することが可能となる．

3.5 タイミングモデル

HMM テキスト音声合成には無い，HMM 歌声合成独自の要素の一つがタイミングモデル[5] である．読み上げ文章の合成と異なり，歌声は曲のテンポやリズムが無視されてはならない．そのため，歌声が開始されるタイミングやその中の音素の継続長は楽譜情報に基づいて決定される．しかし，図 8 のように歌声は音符に記述されている開始時間より少し早く発声されることが多く，歌声の品質の劣化につながっていた．この問題を解決するため，歌声と楽譜のタイミングのずれを 1 次元のガウス分布でモデル化したタイミングモデルが提案された．Sinsy では，WFST[33] を用いた HSMM[34] に基づく音素アライメント情報を元に，歌声の各音素の開始タイミングと楽譜上の開始タイミングのずれを抽出し，システム内の他のモデルと同様にコンテキスト依存モデルとしてモデル化している．合成時には，タイミングモデルにより k 番目の音符の発声の長さ T'_k は次のように決定される．

$$T'_k = T_k - g_{k-1} + g_k, \tag{6}$$

ここで T_k は楽譜上の音符長，g_k はモデル化した音符と発声の開始タイミングのずれである．

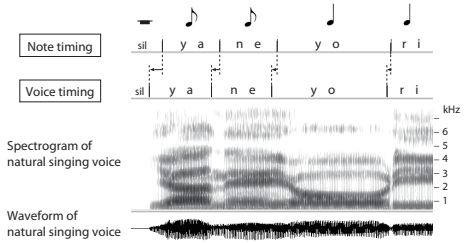


図 8 タイミングの例
Fig. 8 *An example of timings.*

4. 話者適応

Sinsy の歌声モデルには女性 1 名の歌い手による歌声 70 曲で学習した特定話者モデルを用いており，一般ユーザーが同様の歌声モデルを追加するための歌声収録は容易ではなく，十分な量の歌声データを用意することは難しい場合がある．そこで本稿では，CMLLR[13] に基づいた話者適応を歌声合成システムに導入する．CMLLR では，HMM の状態 i における平均ベクトルと対角共分散行列がそれぞれ $\boldsymbol{\mu}_i, \mathbf{U}_i$ のとき，適応後の平均ベクトル $\hat{\boldsymbol{\mu}}_i$ と対角共分散行列 $\hat{\mathbf{U}}_i$ は変換行列 $\mathbf{H}_i$ とバイアスベクトル $\mathbf{b}_i$ を用いてそれぞれ次のように求められる．

$$\hat{\boldsymbol{\mu}}_i = \mathbf{H}_i \boldsymbol{\mu}_i + \mathbf{b}_i \tag{7}$$

$$\hat{\mathbf{U}}_i = \mathbf{H}_i \mathbf{U}_i \mathbf{H}_i^{\top} \tag{8}$$

適応データに対する尤度を最大にする $\mathbf{H}_i$ 及び $\mathbf{b}_i$ は EM アルゴリズムを用いて推定する．また，CMLLR では適応データ量に応じて $\mathbf{H}_i, \mathbf{b}_i$ を複数の状態間で共有する必要があるため，本稿では回帰木と呼ばれる二分木を作成し，リーフノードの分布集合において $\mathbf{H}_i, \mathbf{b}_i$ を共有した．回帰木は学習時の決定木を用いて構築し[35]，回帰木のサイズはデータ量に閾値を設けることにより決定した．

5. 主観評価実験

5.1 実験条件

学習には女性 f001 による童謡など 70 曲，間奏を含めて合計約 70 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 のモノラル音声である．スペクトルパラメータとしては，STRAIGHT[20] 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[16] を適用することにより得られた合計 50 次元のメルケプストラム係数とその 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，ビ

ブラートパラメータにはビブラートの振幅 (cent) と周波数 (Hz) とそれらの Δ, Δ^2 を用いた．ビブラートの振幅と周波数はそれぞれ 30-150cent，5-8Hz に制限している[36), 37)]．HMM は 5 状態の left-to-right 型 HSMM[34)] を用いた．フレーム周期は 5ms とし，音高シフトによる疑似学習データの追加は上下に半音ずつとした．コンテキストクラスタリングの停止基準には MDL 基準[31)] を用い，その重み α（式 (3)）は 5.0 とした．歌声モデルを含む合成部のファイルサイズを表 1 に示す．HMM 歌声合成は比較的小さなフットプリントでも動作するため，合計ファイルサイズは 2.6MBytes 程度に収まっていることが確認できる．

5.2 実験結果

「f001」による特定話者モデルを初期モデルとし，Vocaloid2[14)] を用いて合成した女性「初音ミク」の無伴奏の歌声（自然性がやや増すように人手でパラメータを調整済）と，RWC 研究用音楽データベースのポピュラー音楽データベース (RWC-MDB-P-2001)[38)] の女性「凛 (Rin)」の無伴奏の歌声を適応データとして，話者適応を行った．ただし，楽曲や曲数はそれぞれの歌い手で異なる．適応データに用いた歌い手と合計曲長（無音部分を含む）を表 2 に示す．なお，適応時のクラス数は予備実験で求めた閾値により決定した．評価のために，まず上記のポピュラー音楽データベースの「凛」以外の歌い手による 17 の楽曲から 18 フレーズをランダムに選択した．選択した 18 フレーズをランダムに 3 グループに分け，表 2 の 3 つのモデルを用いて 6 フレーズずつ合成した合計 18 フレーズの歌声を評価データとした．被験者 10 名に，提示した各 18 フレーズの合成歌声がどの歌い手に似ているかを強制的に選択させた結果を図 9 に示す．実験結果より，特定話者モデル (f001) には及ばないものの，話者適応手法によってターゲットの歌い手の特徴を持った歌声が得られることが確認できた．

表 1 ファイルサイズ *(KBytes)*
Table 1 *The total file sizes of Sinsy (KBytes).*

Front-end program (CMX)	456
Phoneme table	3
Back-end program (hts_engine API)	677
Acoustic model	1652
Total file size of Sinsy	2588

表 2 適応データ
Table 2 *Adaptation data.*

Singer	Model	Total song length for adaptation
f001 (Sinsy)	Speaker-dependent model	-
Miku Hatsune (CV01)	Speaker adapted model	9m28s (5 songs)
Rin (RWC-MDB-P-2001)	Speaker adapted model	17m59s (7 songs)

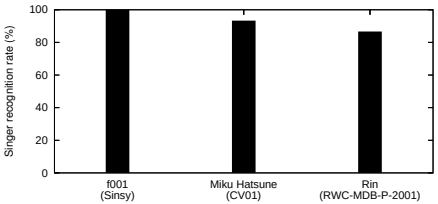


図 9 主観評価実験結果
Fig. 9 *Subjective evaluation results.*

6. むすび

本稿では，歌声データと楽譜から自動的に歌い手の特徴を学習できる HMM 歌声合成システム「Sinsy」を紹介し，構築に用いたソフトウェアやオンラインデモページの運用状況，HMM 歌声合成に特化した手法，歌声モデルの学習条件などをまとめた．HMM 歌声合成方式を採用することにより，調整作業の必要がないため，ユーザーは Sinsy のオンラインデモページに楽譜をアップロードすることで，楽譜に対応した任意の歌声を合成することができる．さらに，話者適応手法によるモデル化によって，より少量のデータから所望の人物の特徴を再現した歌声を合成可能であることを示した．Sinsy の公開をきっかけに他の歌声合成ツールのファイル形式を MusicXML 形式に変換するためのツールが複数制作されたり，ツール自体から MusicXML 形式のファイルを出力する機能が追加されたりしたようである．Sinsy の試みが，歌声情報処理技術の普及と関連コミュニティの「CGM 的」あるいは「オープンソース的」アクティビティ活性化の一助となっているとすれば大変喜ばしいことである．今後は，平均声モデルを用いた話者適応学習による品質の向上，歌唱スタイルの適応や，多言語化などを検討していきたい．

謝辞 本稿で述べた研究開発の一部は SCOPE による．5.1 節の学習で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．

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.Saino, H.Zen, Y.Nankaku, A.Lee, and K.Tokuda, “An HMM-based Singing Voice Synthesis System,” Proc.of ICSLP, pp.1141–1144, 2006.
6) HMM 歌声合成システム: Sinsy, http://www.sinsy.jp/.
7) HMM-based Speech Synthesis System (HTS), http://hts.sp.nitech.ac.jp/.
8) HMM-based Speech Synthesis Engine (hts_engine API), http://hts-engine.sourceforge.net/.
9) Speech Signal Processing Toolkit (SPTK), http://sp-tk.sourceforge.net/.
10) A Speech Analysis, Modification and Synthesis System (STRAIGHT), http://www.wakayama-u.ac.jp/kawahara/STRAIGHTadv/index_e.html.
11) CrestMuseXML Toolkit (CMX), http://cmx.sourceforge.jp/.
12) MusicXML Definition, http://musicxml.org/.
13) M.J.F. ales, “Maximum Likelihood Linear Transformations for HMM-based Speech Recognition,” Proc.of Computer Speech & Language, vol.12, no.2, pp.75–98, 1998.
14) H.Kenmochi and H.Ohshita, “VOCALOID — Commercial Singing Synthesizer based on Sample Concatenation,” Proc.of Interspeech, 2007.
15) 中野倫靖, 後藤 真孝, “VocaListener: ユーザ歌唱を真似る歌声合成パラメータを自動推定するシステムの提案,” 情報処理学会研究報告, vol.2008-MUS-75, no.50, pp.49–56, 2008.
16) 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.
17) 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.
18) S.Imai, “Cepstral Analysis Synthesis on the Mel Frequency Scale,” Proc.of ICASSP, pp.93–96, 1983.
19) A New and Simplified BSD License, http://www.opensource.org/licenses/bsd-license.php.
20) 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.
21) H. Zen, K. Oura, T. Nose, J. Yamagishi, S. Sako, T. Toda, T. Masuko, A. W. Black, and K. Tokuda, “Recent Development of the HMM-based Speech Synthesis System (HTS),” Proc.of APSIPA, pp.121-130, 2009.
22) The Hidden Markov Model Toolkit (HTK), http://htk.eng.cam.ac.uk/.
23) 北原鉄朗, 片寄晴弘, “CrestMuseXML (CMX) Toolkit ver.0.40 について,” 情報処理学会研究報告, vol.2008-MUS-75, no.17, pp.95–100, 2009.
24) J.Odell, “The Use of Context in Large Vocabulary Speech Recognition,” Ph.D.thesis, Cambridge University, 1995.
25) 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, 2010 (to be published).
26) 齋藤毅, 辻直也, 鵜木祐史, 赤木正人, “歌声らしさの知覚モデルに基づいた歌声特有の音響特徴量の分析,” 日本音響学会誌, vol.64, no.5, pp.267–277, 2008.
27) T.Nakano, M.Goto, and Y.Hiraga, “An Automatic Singing Skill Evaluation Method for Unknown Melodies Using Pitch Interval Accuracy and Vibrato Features”, Proc.of Interspeech, pp.1706–1709, 2006.
28) 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.309–312, 2009.
29) 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.
30) A.Mase, K.Oura, Y.Nankaku, and K.Tokuda, “HMM-based Singing Voice Synthesis System using Pitch-Shifted Pseudo Training Data,” Proc.of Interspeech, 2010 (to be published).
31) 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.
32) 武藤聡, 大浦圭一郎, 南角吉彦, 徳田恵一, “HMM 歌声合成における話者適応および楽譜情報を用いたモデル学習高速化,” 日本音響学会春季講論集, vol.I, 2-7-8, pp.347–348, 2010.
33) K. Oura, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, “A Fully Consistent Hidden Semi-Markov Model-Based Speech Recognition System,” Proc.of IEICE Trans., vol.E91-D, no.11, pp.2693–2700, 208.
34) 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.
35) J.Yamagishi, M.Tachibana, T.Masuko, and T.Kobayashi, “Speaking Style Adaptation Using Context Clustering Decision Tree for HMM-based Speech Synthesis,” Proc.of ICASSP 2004, pp.5–8, 2004.
36) J.Sundberg, “The Science of the Singing Voice,” Northern Illinois University Press, 1987.
37) C.E.Seashore, “A Musical Ornament, the Vibrato,” Proc.of Psychology of Music, McGrew-Hill Book Company, pp.33–52, 1938.
38) 後藤真孝, 橋口博樹, 西村拓一, 岡隆一, “RWC 研究用音楽データベース:研究目的で利用可能な著作権処理済み楽曲・楽器音データベース,” 情報処理学会論文誌, vol.45, no.3, pp.728–738, 2004.