Vol.2013-MUS-99 No.52
2013/5/12

Sinsy
～隠れマルコフモデルに基づく歌声合成の現状と今後～

大浦 圭一郎[1,a)]　南角 吉彦[1]　徳田 恵一[1]

概要：近年，音声合成関連の研究分野では，統計的パラメトリック音声合成と呼ばれる統計モデルに基づいた手法が広く研究されている．この中でも，統計モデルとして隠れマルコフモデル (Hidden Markov Model; HMM) を用いる HMM 音声合成方式は，理論的に整理されたアルゴリズムと利用しやすいソフトウェアツールが公開されており，広く普及してきている．従来の波形接続方式と比較すると HMM 音声合成方式は，発話の癖の再現や感情音声合成などの多様性，さらにそのフットプリントの小ささや言語依存性の低さなど，多くの優位性を持っている．一方，歌声合成関連の研究分野では従来の波形接続方式が広く用いられているものの，HMM 音声合成方式も徐々に使われてきている．このような流れの中，我々は Sinsy と名付けた HMM 歌声合成システムを構築し，そのオンラインデモを公開した．本稿では HMM 歌声合成方式を紹介し，現状の Sinsy のサービスや，今後の展望等を述べる．

1. はじめに

歌声の合成の研究に関しては長い歴史があり，これまで様々な方式が検討されてきた．最近では，VOCALOID [1] 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きている．また，一般の人々の認知度が高まるにつれて，より簡単に，好きな歌い手の声で，好きな曲を歌わせることのできる柔軟なシステムの需要が高まっている．実際，UTAU [2] 等の歌声合成のためのフリーソフトウェアにおいては，ユーザが作成した多くの歌手ライブラリが公開されている．また，従来の歌声合成システムにおいて，自然な歌声を実現するための調整作業は，作品制作の上で創造的な部分ではあるものの，一般ユーザには敷居が高すぎるとの声もあり，ユーザの歌唱データに基づいた自動調整手法 [3] 等が提案されてきた．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（図 1）．本稿では Sinsy が採用している歌声合成方式の概要とその特徴，現状のサービスや今後の展望について述べる．

2. HMM 歌声合成システム Sinsy

Sinsy は，音声合成方式の一つである隠れマルコフモデル (Hidden Markov Model; HMM) に基づく音声合成方式

図 1　*Sinsy オンラインデモ http://www.sinsy.jp/*

を歌声の合成に応用したものである [4]．HMM は，時系列を統計的にモデル化することのできる確率モデルの一種で，音声の特徴パラメータ系列のモデル化手法として音声認識等で広く利用されているモデルである．図 2 に HMM 歌声合成システム Sinsy の概要を示す．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の関係をモデル化する．スペクトル，基本周波数，時間情報を同時にモデル化する方式となっており，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，オーバーシュート等に関する特徴だけでなく，音符に対するタイミングも自動学習するため，前ノリ，後ノリ等の歌唱スタイルさえも自動学習によって再現することがで

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

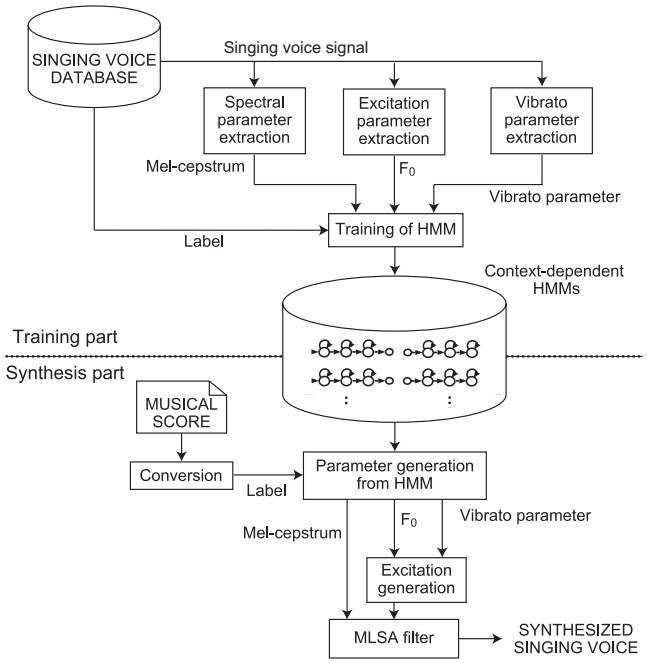

図 2 *HMM* 歌声合成システム *Sinsy* の概要

きる．

現在，一般に用いられている歌声合成システムの多くは，単位選択型あるいは素片連結型と呼ばれる音声合成方式と同じアイディアに基づいている．素片連結型の音声合成方式では，あらかじめ蓄積された音声波形データを素片に分割し，合成したいテキストに応じて素片を選択，連結することにより，音声を合成する．これに対し，HMM 歌声合成の元となった HMM 音声合成方式は，素片連結型には無い以下のような優れた特徴を持つ．

(1) 与えられた音声データに基づいてモデルを自動学習することにより，元話者の声の特徴を再現する合成音声を得ることができる．
(2) 比較的少ない量の学習データで高品質な合成音声を得ることができる．
(3) 学習用の音声データをランタイムのシステムに蓄積する必要がないため，軽量である．
(4) HMM のモデルパラメータを適切に変更することにより，様々な声の合成音声を得ることができる．

特に (4) は HMM 音声合成方式の重要な優位点の一つであり，実際に「声を真似る」話者適応手法，「声を混ぜる」話者補間手法，「声を作る」固有声手法などが提案されている．HMM 音声合成方式は理論的に整理されたアルゴリズムと利用しやすいソフトウェアツールが広く普及しており，Sinsy は HTS [5] や STRAIGHT [6] を軸としたオープンソースソフトウェアを利用して構築されている．

Sinsy オンラインデモ（図 1）では，MusicXML 形式で記述された歌詞付きの楽譜をアップロードすることにより，合成された歌声波形を得ることができる．楽譜の表現に MusciXML を採用した理由は，オープンなフォーマットであること，楽譜を正確に表現可能であること等による．Finale NotePad や MuseScore など，MusicXML ファイルのエクスポート機能を持った楽譜編集ツールを用いることにより，ユーザは容易に MusicXML ファイルを作成することができる．

3. Sinsy の現状と今後

HMM 音声合成方式を歌声に応用するにあたり，歌声特有の現象を高精度にモデル化するため，現状の Sinsy では以下の手法を導入している．

- 音符と発声の開始時間及び終了時間の差分のモデル化 [7].
- 楽譜の時間情報を用いた学習アルゴリズムの安定化及び高速化手法 [8].
- ビブラートの振幅，周期のモデル化 [9].
- 学習データの音高出現頻度の偏りに頑健な音高正規化学習 [10].
- 英語歌唱のモデル化.

また，さらなる高精度化に向けて，以下の手法の導入を検討している．

- 一音素あたりの表現能力が入力楽譜に基づいて適切に変動するモデル構造 [11].
- 複数の歌唱者のデータを同時に用いることによって頑健なモデル化が可能な歌唱者正規化学習 [12].
- 調子外れの補正手法 [13].

Sinsy の公開をきっかけに UTAU, VOCALOID, Cadencii 等の楽譜編集機能をもった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有志によって複数制作され，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．Sinsy の公開が，このような「CGM 的」あるいは「オープンソース的」な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．今後，Sinsy オンラインデモを利用した楽曲が動画サイト等に多く投稿され，二次創作の輪が更に広がることを期待している．MusicXML で記述された楽譜とそれに対応した歌声データのセットが，数十分程あれば，Sinsy の歌声モデルの学習は，ほぼ自動で行うことができる．将来，ユーザが独自の歌声データをアップロードすると，それに基づいて自動学習した歌声モデルが，Sinsy オンラインデモに追加されるという機能を実現したいと考えている．

謝辞 本研究の一部は，公益財団法人堀科学芸術振興財団，及び JST CREST uDIalogue プロジェクトの助成による．

参考文献

[1] H. Kenmochi and H. Ohshita, “VOCALOID — Commercial Singing Synthesizer based on Sample Concatenation,” Proc. of Interspeech, Special session, 2007.
[2] “歌声合成ツール UTAU,” http://utau2008.web.fc2.com/.
[3] Tomoyasu Nakano and Masataka Goto, “VocaListener: A Singing-to-Singing Synthesis System Based on Iterative Parameter Estimation,” Proc. of Sound and Music Computing Conference, pp. 343–348, 2009.
[4] Keiichiro Oura, Ayami Mase, Tomohiko Yamada, Satoru Muto, Yoshihiko Nankaku, and Keiichi Tokuda, “Recent Development of the HMM-based Singing Voice Synthesis System — Sinsy,” Proc. of Speech Synthesis Workshop, pp. 211–216, 2010.
[5] 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 Asia Pacific Signal and Information Processing Association, pp. 121–130, 2009.
[6] 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.
[7] Keijiro Saino, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, and Keiichi Tokuda, “An HMM-based Singing Voice Synthesis System,” Proc. of ICSLP, pp. 1141–1144, 2006.
[8] 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,” 音響学会講論集, pp. 347–348, 2010.
[9] 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,” 情報処理学会研究会, vol. 2009-MUS-80, no. 5, pp. 1–6, 2009.
[10] Keiichiro Oura, Ayami Mase, Yoshihiko Nankaku, and Keiichi Tokuda, “Pitch adaptive training for HMM-based singing voice synthesis,” Proc. of ICASSP 2012, vol. 1, pp. 5377–5380, 2012.
[11] 大浦圭一郎, 南角吉彦, 徳田恵一, “HMM 歌声合成における状態数可変のモデル構造の検討,” 日本音響学会秋季講論集, vol. I, 2-2-2, pp. 275–276, 2012.
[12] 城田佳菜子, 中村和寛, 大浦圭一郎, 南角吉彦, 徳田恵一, “HMM 歌声合成における歌唱者適応学習の検討,” 日本音響学会春期講論集, vol. I, 2-7-11, pp. 339–340, 2013.
[13] 喜多村翔斗, 中村和寛, 橋本佳, 大浦圭一郎, 南角吉彦, 徳田恵一, “音高正規化学習を用いた HMM 歌声合成における調子外れの補正,” 日本音響学会春期講論集, vol. I, 2-7-10, pp. 337–338, 2013.