隠れマルコフモデルに基づく音声合成システム

全 炳河　徳田 恵一
名古屋工業大学 情報工学専攻

1 概要

近年，隠れマルコフモデル (HMM) に基づく統計的パラメトリック音声合成方式が注目されている．本方式では，音声のスペクトル・励振源・継続長が統計モデルの一種である HMM により同時にモデル化される．音声合成時は，合成したい文章に対応する HMM からの出力確率が最大となるよう，継続長・スペクトル・励振源系列を再合成した後，音声合成フィルタを用いて波形を出力する．2002 年より我々は，HMM に基づく音声合成のための研究・開発ツール「HMM 音声合成システム (HTS)」を，オープンソースソフトウェアとして公開してきた．本稿では，その開発状況について述べる．

2 隠れマルコフモデルに基づく音声合成

テキスト音声合成は，音声認識の逆問題とみなせる．つまり，単語列 $\boldsymbol{w} = \{w_1, \ldots, w_N\}$ を入力として，音声波形列 $\boldsymbol{o} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$ を生成するような問題である．通常テキスト音声合成器は，言語解析部と波形生成部から構成される．言語解析部において入力単語列が読み・品詞・アクセント・句境界などが付与されたサブワード単位列に変換された後，波形生成部で音声波形が合成される．

現在の音声合成器のほとんどは，大量の音声データを用いて構築されている．このような手法は一般に，「コーパスベース音声合成」と呼ばれる．コーパスに基づく手法の導入により合成音の品質は，従来の規則ベースのものと比較して大きく向上した．

コーパスベース音声合成における最も一般的な手法は，波形接続音声合成である [1]．本手法ではまず，音声コーパスを半音素・音素・音節といったサブワード単位の音声素片に分割し，素片集合を構成する．合成時は，入力単語列に対するコストを最小にするよう，動的計画法を用いて素片集合から音声素片列を選択し，選ばれた素片列を連結・信号処理して出力する．本手法は音声波形をそのまま用いるので，その品質は非常に高いが，しばしば素片接続部で不連続が生じることがある．また，様々な話者性や発話スタイルを実現するには，大量の音声を収録する必要がある．

コーパスベース音声合成におけるもう一方の手法は，統計的パラメトリック音声合成である [2]．本手法では，入力単語列 $\boldsymbol{w}$ に対する確率が最大となる音声特徴量系列 $\boldsymbol{o} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$ を出力する．つまり，

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid \boldsymbol{w}). \quad (1)$$

である．$P(\boldsymbol{o} \mid \boldsymbol{w})$ としては様々な生成モデルが利用できるが．現在広く使われているものは，統計モデルの一種である隠れマルコフモデル (HMM) である．この場合を特に「HMM に基づく音声合成」と呼ぶ [3]．

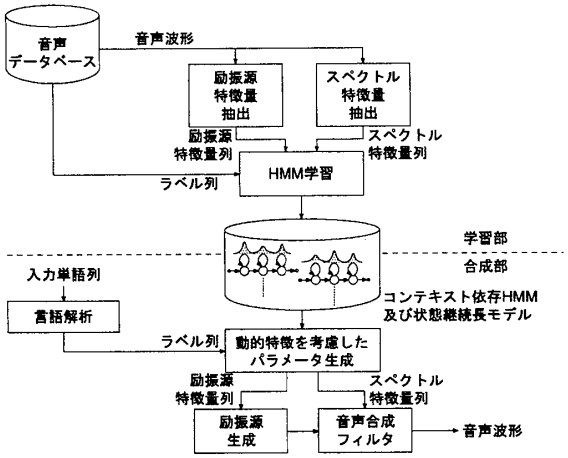

図 1 HMM に基づく音声合成の概要

図 1 に，HMM に基づく音声合成の概要を示す．学習部は，HMM を用いた音声認識で用いられるものとほぼ等価である．主な違いは，励振源特徴量を特徴ベクトルに含む・韻律情報もコンテキストに用いる，という点である．合成部では，まず入力単語列を言語解析してコンテキスト依存ラベル列に変換した後，ラベル列に従いサブワード単位 HMM を連結し，文章 HMM を構成する．次に，動的特徴を考慮した音声パラメータ生成アルゴリズム [4] を用いて，スペクトル及び励振源特徴量列を文章 HMM より生成する．最後に，音声合成フィルタにより音声波形を合成する．本手法の大きな利点として，HMM のパラメータを適切に変化させることで，その発話スタイルや話者性を柔軟に変化できることが挙げられる [5]．一方，その欠点としては，音質がボコーダ音となるという点があるが，近年開発が進み，その品質は単位接続型音声合成に近づきつつある．

3 HMM 音声合成システム

HMM 音声合成システム (HTS) [6] 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る．本ソフトウェアは，ケンブリッジ大学により開発・公開されている HMM ツールキット (HTK) [7] へのパッチの形で BSD ライセンスの下公開されており，国内外の様々な研究機関・ベンチャー企業で利用されている．HTK からの変更点の概略を以下に示す．

- バージョン 1.0 (2002 年 12 月)
  - MDL 基準に基づく決定木によるクラスタリング [3]
  - ストリーム依存のクラスタリング
  - 多空間確率分布にる F_0 系列のモデル化 [8]

An HMM-based speech synthesis system
Heiga ZEN, Keiichi TOKUDA
Department of Computer Science, Nagoya Institute of Technology
466-8555, Nagoya, Japan
zen@sp.nitech.ac.jp, tokuda@nitech.ac.jp

  - 状態継続長分布のモデル化 [3]
  - 動的特徴を考慮したパラメータ生成アルゴリズム [4]
- バージョン **1.1 (2003 年 5 月)**
  - ランタイム音声合成エンジン
  - Festival 音声合成システム用 Voice の提供
- バージョン **1.1.1 (2003 年 12 月)**
  - 多空間確率分布の分散のフロアリング
  - ポストフィルタリング
  - 名工大日本語音声データベースを用いたデモ
- バージョン **2.0 (2006 年 12 月)**
  - EM 型パラメータ生成アルゴリズム [4]
  - 話者適応・適応学習 [9]
- バージョン **2.0.1 (2007 年 10 月)**
  - 話者補間
  - LSP 型スペクトル特徴量のサポート
  - API 版ランタイム合成エンジン
- バージョン **2.1 (2008 年 3 月)**
  - 隠れセミマルコフモデル [10]
  - 発話内変動を考慮したパラメータ生成 [11]
  - 高度な適応アルゴリズム [12]

HTS バージョン 2.1 と STRAIGHT 分析合成系 [13] を組み合わせることで，共通の音声データベースを用いた音声合成器の国際的な評価会 Blizzard Challenge [14] 向けに我々のグループが準備した，最新の HMM 音声合成器を構築することができる [15–17].

4 音声合成以外への応用

HTS は HMM に基づく音声合成のためのプラットフォームを提供することを主な目的として開発されているが，その他様々な形で利用されている．以下に例を示す．

- 動作生成 [18–20]
- 顔画像生成 [21]
- 音声・動画同時生成 [22, 23]
- 調音運動から音声への変換 [24]
- 韻律認識 [25, 26]
- 語学学習システムにおける発音誤りの検出 [27]
- 極低ビットレート音声符号化 [28]
- 音声認識システムの自動評価 [29].
- オンライン文字認識 [30]

5 まとめ

本稿では，HMM に基づく音声合成の概要と，その研究プラットフォームである HMM 音声合成システム (HTS) について述べた．本ソフトウェアは，音声合成及び他分野で利用が広がっており，今後更なる発展が期待される．

謝辞 本研究の一部は，文部科学省リーディングプロジェクト e-Sociey によった．HTS の開発に携わった全ての方々に感謝する．

参考文献

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," Proc. ICASSP, pp.373–376, 1996.
[2] A. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," Proc. ICASSP, pp.1229–1232, 2007.
[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J83-D-II, no.11, pp.2099–2107, Nov. 2000.
[4] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc. ICASSP, pp.1315–1318, 2000.
[5] J. Yamagishi, Average-Voice-Based Speech Synthesis, Ph.D. thesis, Tokyo Institute of Technology, 2006.
[6] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. Black, and T. Nose, "The HMM-based speech synthesis system (HTS)." `http://hts.sp.nitech.ac.jp/`.
[7] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X.Y. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The hidden Markov model toolkit (HTK) version 3.4," 2006. `http://htk.eng.cam.ac.uk/`.
[8] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," IEICE Trans. Inf. & Syst., vol.E85-D, no.3, pp.455–464, Mar. 2002.
[9] J. Yamagishi, K. Ogata, Y. Nakano, J. Isogai, and T. Kobayashi, "HSMM-based model adaptation algorithms for average-voice-based speech synthesis," Proc. ICASSP, pp.77–80, 2006.
[10] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," IEICE Trans. Inf. & Syst., vol.E90-D, no.5, pp.825–834, 2007.
[11] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," IEICE Trans. Inf. & Syst., vol.E90-D, no.5, pp.816–824, 2007.
[12] Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis," Proc. Interspeech, pp.2286–2289, 2006.
[13] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "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.
[14] K. Tokuda and A. Black, "Speech synthesis research in a new age of cooperation and competition – The Blizzard Challenge," Journal of ASJ, vol.62, no.6, pp.466–470, 2006. (in Japanese).
[15] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," IEICE Trans. Inf. & Syst., vol.E90-D, no.1, pp.325–333, Jan. 2007.
[16] H. Zen, T. Toda, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," Blizzard Challenge Workshop, 2006.
[17] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system – HTS-2007 system for the Blizzard Challenge 2007," Proc. Blizzard Challenge 2007, 2007.
[18] K. Mori, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura, "Motion generation for Japanese finger language based on hidden Markov models," Proc. FIT, pp.569–570, 2005. (in Japanese).
[19] N. Niwase, J. Yamagishi, and T. Kobayashi, "Human walking motion synthesis with desired pace and stride length based on HSMM," IEICE Trans. Inf. & Syst., vol.E88-D, no.11, pp.2492–2499, 2005.
[20] G. Hofer, H. Shimodaira, and J. Yamagishi, "Speech driven head motion synthesis based on a trajectory model," Proc. SIGGRAPH, 2007.
[21] O. Govokhina, G. Bailly, G. Breton, and P. Bagshaw, "TDA: a new trainable trajectory formation system for facial animation," Proc. Interspeech, pp.1274–1247, 2006.
[22] M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi, "Text-to-audio-visual speech synthesis based on parameter generation from HMM," Proc. Eurospeech, pp.959–962, 1999.
[23] S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "HMM-based text-to-audio-visual speech synthesis," Proc. ICSLP, pp.25–28, 2000.
[24] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," Proc. of Interspeech, pp.577–580, 2006.
[25] K. Emoto, H. Zen, K. Tokuda, and T. Kitamura, "Accent type recognition for automatic prosodic labeling," Proc. Autumn Meeting of ASJ, pp.225–226, 2003. (in Japanese).
[26] H.L. Wang, Y. Qian, F. Soong, J.L. Zhou, and J.Q. Han, "A multi-space distribution (MSD) approach to speech recognition of tonal languages," Proc. of Interspeech, pp.125–128, 2006.
[27] L. Zhang, C. Huang, M. Chu, F. Soong, X. Zhang, and Y. Chen, "Automatic detection of tone mispronunciation in Mandarin," Proc. ISCSLP, pp.590–601, 2006.
[28] T. Hoshiya, S. Sako, H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Improving the performance of HMM-based very low bitrate speech coding," Proc. ICASSP, pp.800–803, 2003.
[29] R. Terashima, T. Yoshimura, T. Wakita, K. Tokuda, and T. Kitamura, "An evaluation method of ASR performance by HMM-based speech synthesis," Proc. Spring Meeting of ASJ, pp.159–160, 2003. (in Japanese).
[30] L. Ma, Y.J. Wu, P. Liu, and F. Soong, "A MSD-HMM approach to pen trajectory modeling for online handwriting recognition," Proc. ICDAR, pp.128–132, 2007.