HMM 音声合成における音声分析合成器 STRAIGHT と WORLD の比較*

○ 高道 慎之介, 戸田 智基 （奈良先端大）,
森勢 将雅 （山梨大）, 中村 哲 (奈良先端大)

1 はじめに

統計的パラメトリック音声合成の高品質化には，高品質な音声分析合成器の使用が不可欠であり，当該研究分野の昨今の発展は，特許ライセンスで提供されている分析合成器 STRAIGHT [1, 2] の貢献によるところが大きい．一方で近年，新たな高品質分析合成器 WORLD [3, 4, 5] が提案され，修正 BSD ライセンスにて公開された．本稿では，WORLD による統計的パラメトリック音声合成の技術的・産業的発展を見据え，合成法の一つである HMM 音声合成 [6] において STRAIGHT と WORLD の品質を比較する．実験的評価結果から，合成音声の知覚的音質は両分析合成器で同程度であることを示す．

2 HMM 音声合成における音声分析合成器 WORLD の利用

2.1 使用する特徴量

分析合成器 STRAIGHT と WORLD は，抽出された F_0 に基づいてスペクトル包絡・非周期性指標を抽出する点で共通するが，非周期性指標の特徴量の形式が異なる．STRAIGHT は全時間フレームにおいて，周波数ビン毎の非周期性指標を明示的に抽出する．HMM 音声合成では，抽出された非周期性指標を帯域毎に平均化 [7, 8] して次元圧縮したものを特徴量として用いる（Fig. 1(a)）．一方で WORLD は，有声フレームにおいてのみ，3 kHz 帯域毎の非周期性指標を明示的に推定し，周波数ビン毎の指標を決定する．この帯域毎の非周期性指標は，有声フレームにおいて帯域数[1]だけの次元数，無声フレームにおいて 0 次元の値を持つ特徴量とみなされる．故に，この特徴量を多空間確率密度関数 HMM（MSD-HMM） [9] でモデル化する (Fig. 1(b))．F_0 と非周期性指標の結合ベクトルを MSD-HMM でモデル化することも可能だが，本稿では，別ストリームでモデル化するものとする．

さらに本稿では，次の 2 点を検討する．

2.2 連続的な非周期性指標系列の使用

MSD-HMM を用いた F_0 系列のモデル化では，異なる尤度スケールを持つモデルで尤度を計算する問題があり，特に，有声と無声の切り替わるフレーム近傍において頑健な学習が行われない．これに対し，無声フレームにおける F_0 値を推定し，得られた連続的 F_0 系列をモデル化する手法が提案されている [10]．これにより得られるモデルは，頑健に学習されるのみならず，他の連続的系列（例えば，スペクトル系列）と同様の枠組みでの取り扱いが可能である [11]．

本稿では，連続的な非周期性指標系列を推定し，連続 HMM でモデル化する手法を検討する (Fig. 1(c))．無声フレームにおける非周期性指標の推定法として，WORLD で決定する F_0 候補を当該フレームの F_0 とみなして推定する方法が考えられるが，全フレームにおいて F_0 候補が存在する保証はない．そこで，得られた F_0 系列に対してスプライン法を用いて無声フレームの F_0 値を推定 [12] する．連続的な非周期性指標は，この連続 F_0 系列のもとで WORLD により推定される．得られた連続的な非周期指標の例を Fig. 2 に示す．有声フレームにおける指標値は通常の指標値と一致することに注意する．生成時には，生成された F_0 系列により有声／無声区間を決定し，有声区間における非周期性指標の値のみを使用する．

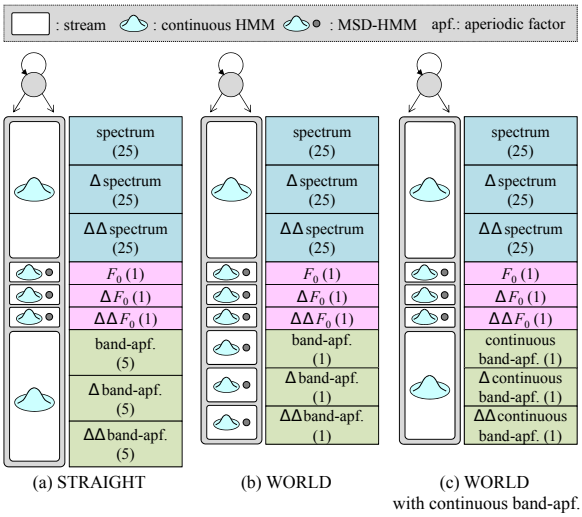

Fig. 1 STRAIGHT と WORLD を用いた HMM 音声合成の特徴量ベクトル（各特徴量における括弧内の数字は，本稿の実験条件における特徴量の次元数を表す.）

2.3 統計的パラメトリック音声合成方式における高音質化手法の効果

HMM 音声合成の通常の学習・生成の枠組みは，低音質の音声を生成する傾向にある．この問題に対し，様々な高品質化手法が提案されているが，これらの音質改善効果は，音声分析合成器の品質に依存する．そこで本稿では，高音質化手法として，HMM 音声合成システム（HTS）[13] ver. 2.3 beta の Speaker dependent training demo に搭載されている，変調スペクトルに基づくポストフィルタ [14, 15] [2]の効果を検証する．

3 実験的評価

3.1 実験条件

学習データは男女各 2 名（RMS, SLT, BDL, CLB）による CMU ARCTIC speech database [17] a セット 593 文とし，各話者毎に 5 状態 left-to-right 型の隠れ

*STRAIGHT vs. WORLD, Comparison of Speech Anaylysis-Synthesis Systems in HMM-Based Speech Synthesis. By TAKAMICHI, Shinnosuke, TODA, Tomoki (NAIST), MORISE, Masanori (Univ. of Yamanashi), NAKAMURA, Satoshi (NAIST)

[1]帯域数は音声のサンプリング周波数に依存し，16 kHz サンプリングの場合は，1 帯域となる．

[2]F_0 に対するフィルタ処理は搭載されてないが，ローパスフィルタによるマイクロプロソディ [16] 除去及び連続 F_0 の推定により，実装可能である．

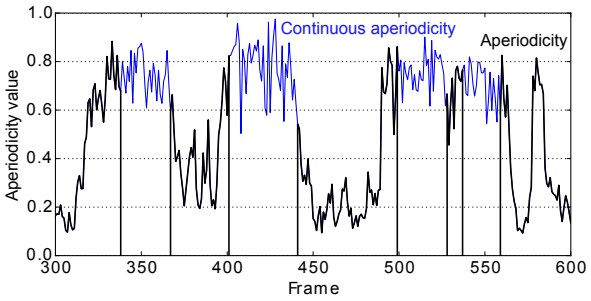

Fig. 2　WORLD で推定した連続的な非周期性指標系列の例

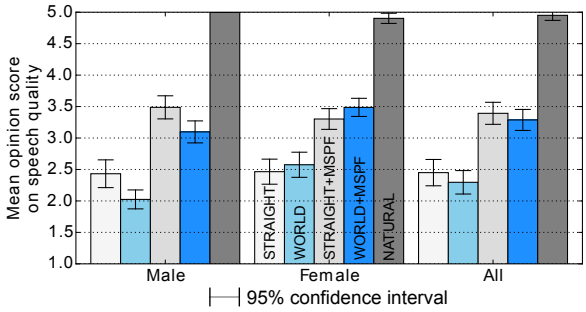

Fig. 3　音声の音質に関する主観評価結果

セミマルコフモデル（Hidden Semi-Markov Model：HSMM）[18] の学習を行う．評価データは同 b セット 100 文を使用する．学習データのサンプリング周波数は 16 kHz，フレームシフトは 5 ms とする．スペクトル特徴量は，各分析合成器で抽出されたスペクトル包絡を周波数伸縮して計算した，0 次から 24 次のメルケプストラム係数とする．音源特徴量は，対数 F_0，帯域毎の非周期性指標を使用する．STRAIGHT と WORLD の帯域非周期性指標の次元数はそれぞれ，5 及び 1 である．STRAIGHT を使用する場合の波形合成には，MLSA フィルタ [19] を使用する．スペクトル特徴量，対数 F_0，帯域毎の非周期成分のストリーム重みはそれぞれ，1.0，1.0，及び 0.0 とする．コンテキストクラスタリングには，MDL 基準による決定木クラスタリング [20] を用いる．ただし，MDL factor は，全ての特徴量に対して通常の設定（1.0）とする．音声パラメータの生成には，最尤パラメータ生成法 [21] を用いる．変調スペクトルを計算する際のフレーム長，シフト長，フーリエ変換長はそれぞれ，25，12，及び 64 フレームとする．変調スペクトルに基づくポストフィルタは，スペクトル及び F_0 にのみ適用する．ただし，フィルタ強度係数は 1.0 とする．

予備実験において，WORLD による非周期性指標（Fig. 1(b)）と連続的な非周期性指標（Fig. 1(c)）を用いた手法間で，知覚的な音質の差は認められなかった．そのため，Fig. 1(c) を以降の評価に使用するものとする．音質評価は，各音声分析合成器を用いた手法（“STRAIGHT”，“WORLD”）の合成音声，更にポストフィルタを適用した手法（“*+MSPF”）の合成音声，及び自然音声（“NATURAL”）の計 5 種を用いた 5 段階 MOS 評価であり，8 名の評価者に対して，各話者・各評価データをランダムに提示し，評価させた．

3.2　主観評価結果

音質に関する主観評価結果を Fig. 3 に示す．全話者（All）の結果を見ると，STRAIGHT と WORLD で同程度の音質が得られていることから，HMM 音声合成における WORLD の使用の有効性が確認できる．一方で性別毎の結果を見ると，女性話者に対して WORLD，男性話者に対して STRAIGHT の音質がより高いことが分かる．

4　まとめ

本稿では，音声分析器 STRAIGHT と WORLD の品質を HMM 音声合成を用いて評価した．その結果，各分析合成器による合成音声は同程度の音質であることが確認されたため，修正 BSD ライセンスで公開されている WORLD による，統計的パラメトリック音声合成の技術的・産業的発展が期待される．今後は，性別依存性の調査，より頑健な合成へ向けた処理を検討する．

謝辞　本研究の一部は，JSPS 科研費 26・10354 および 26280060 の助成を受け実施したものである．

参考文献

[1] H. Kawahara *et al.*, *Speech Commun.*, Vol. 27, No. 3–4, pp. 187–207, 1999.
[2] H. Kawahara *et al.*, *MAVEBA 2001*, pp. 1–6, 2001.
[3] M. Morise, *Speech Commun.*, Vol. 67, pp. 1–7, 2015.
[4] 森勢 他, 信学論, Vol. J94-D, No. 7, pp. 1079–1087, 2011.
[5] 森勢 他, 信学技報, SP2015-3, pp. 217–222, 2015.
[6] K. Tokuda *et al.*, *Proc. of the IEEE*, Vol. 101, No. 5, pp. 1234–1252, 2013.
[7] Y. Ohtani *et al.*, *Proc. INTERSPEECH*, pp. 2266–2269, 2006.
[8] T. Yoshimura *et al.*, *Proc. Eurospeech*, pp. 2259–2262, 2001.
[9] K. Tokuda *et al.*, *IEICE Trans. Inf. & Syst.*, Vol. E85-D, No. 3, pp. 455–464, 2002.
[10] K. Yu *et al.*, *IEEE Trans. Audio, Speech, and Language*, Vol. 19, No. 5, pp. 1071–1079, 2011.
[11] 高道 他, 信学技報, SP2014-140, pp. 31–36, 2015.
[12] K. Tanaka *et al.*, *Proc. INTERSPEECH*, pp. 3067–3071, 2013.
[13] HMM-based Speech Synthesis System (HTS) `http://hts.sp.nitech.ac.jp/`
[14] S. Takamichi *et al.*, *Proc. ICASSP*, pp. 290–294, 2014.
[15] S. Takamichi *et al.*, *Proc. GlobalSIP*, pp. 710–714, 2014.
[16] P. Taylor, “Text-To-Speech Synthesis”, Cambridge Univ. Press, 2009.
[17] J. Kominek *et al.*, *Tech. Rep. CMU-LTI-03-177*, 2003.
[18] H. Zen *et al.*, *IEICE Trans., Inf. and Syst.*, E90-D, Vol. 90, No. 5, pp. 825–834, 2007.
[19] S. Imai *et al.*, *Electronics and Communications in Japan (Part I: Communications)*, Vol. 66, No. 2, pp. 10–18, 1983.
[20] K. Shinoda *et al.*, *J. Acoust. Soc. Jpn. (E)*, Vol. 21, No. 2, pp. 79–86, 2000.
[21] K. Tokuda *et al.*, *Proc. ICASSP*, pp. 1315–1318, 2000.