生成過程モデルによる基本周波数パターンの階層表現とHMM音声合成のマルチストリーム学習*

☆島田智大, 百武恭汰, 橋本浩弥, 齋藤大輔, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

近年、テキストを音声に変換するテキスト音声合成技術は、様々な分野で活用されている。音声合成技術の手法の一つである HMM (Hidden Markov Model) 音声合成 [1] は特徴量を統計モデルとして扱うため、録音された音声を波形の断片として連結して合成する波形接続型の音声合成に対して少ない学習データでより柔軟な音声合成が実現できることが期待される。その一方で、HMM 音声合成は特徴量をフレーム単位で扱っているため、アクセント、イントネーションなどの、より長い時間単位にわたって現れる韻律的特徴を的確に表現しにくく、音質の劣化につながる要因となっていた。韻律は話者の意図や感情を伝達するのに大きな役割を持っており、それを正しく表現することは自然な音声を合成する上では不可欠である [2]。このため、HMM 音声合成においてこの韻律的特徴を明示的にモデル化することは重要である。HMM 音声合成におけるこの問題を解決する方法として基本周波数パターン生成過程モデル (以下 F_0 モデル) を利用したアプローチが検討されている [3]。このモデルは人間の生理的・物理的特性に基づいて少数のパラメータで音声の基本周波数パターン (以下 F_0 パターン) を表すことが可能になっている。また、言語情報を利用することにより、高精度にモデルパラメータを自動抽出する手法が提案されており、これを利用して音声データから抽出された F_0 パターンではなく F_0 モデルによって修正した F_0 パターンを使うことにより合成音声の品質が改善されることが実験的に確認されている [4]。また、音声から抽出された対数 F_0 パターンと F_0 モデルによる対数 F_0 パターンとの差である F_0 残差は F_0 モデルでは表現されないが、これを考慮した HMM 音声合成の検討もなされている [5]。

F_0 パターンは文節の係り受け、アクセント型、音素など様々な要因と関係するが、F_0 モデルにより、フレーズ成分、アクセント成分、F_0 残差に分離することにより、個々の要因とのより明確な対応が得られやすくなると考えられる。本稿では、従来の F_0 モデルを用いた HMM 音声合成では区別されていなかったこれらの 3 成分を、別々の特徴量ストリームとして取り扱うマルチストリーム学習を行うことで、HMM 音声合成の性能向上を図り、合成音声の聴取実験等により確認する。

2 提案手法

音声を分析して F_0 パターンを抽出し、得られた F_0 パターンから橋本らの手法により F_0 モデルのフレーズ指令とアクセント指令の自動推定を行う [4]。推定されたパラメータを元に F_0 モデルの F_0 パターンを算出し、F_0 残差を求める。そして、F_0 モデルのアクセント成分・フレーズ成分と F_0 残差を別々のストリームで HMM の学習を行い、得られたモデルにより合成文のパラメータ生成を行う。生成されたフレーズ成分、アクセント成分、F_0 残差を加算して得られる対数 F_0 パターンと HMM により生成されたスペクトル包絡特徴量、非周期性指標を用いて音声波形を生成する。

3 実験

3.1 実験条件

対数 F_0 パターンをフレーズ成分、アクセント成分、F_0 残差に分け、マルチストリーム学習・生成する HMM 音声合成を提案手法 (F_0-multi)、分けずに学習・生成する手法を従来手法 (HTS) とし、合成音声の聴取実験による音声品質の比較 (主観評価) と、F_0-RMSE による F_0 パターン比較 (客観評価) とを行った。音声データには、ATR 日本語データベースの音素バランス全 503 文の読み上げデータから話者 MHT を選んで使用した [6]。全 503 文のうちサブセット A から I までの 450 文で HMM を学習し、サブセット J の 53 文を合成した。音声の分析は STRAIGHT を用いて、F_0 パターン、スペクトル包絡特徴量、非周期性指標を抽出した [7]。分析条件は, フレーム周期 5 [ms] である。HMM に用いた特徴量は 0 から 39 次元までのメルケプストラムと 5 帯域の平均非周期性指標、F_0 モデルのアクセント成分、フレーズ成分、F_0 残差、およびそれらの Δ、Δ^2 を含めた 144 次元のベクトルとした。

メルケプストラムと平均非周期指標は、スペクトル

*Hierarchical expression of speech F_0 contours by the F_0 model and its use for multi-stream HMM-based speech synthesis by SHIMADA Tomoharu, HYAKUTAKE Kyota, HASHIMOTO Hiroya, SAITO Daisuke, MINEMATSU Nobuaki, and HIROSE Keikichi (The University of Tokyo)

Table 1　評価実験の結果

手法	聴取実験 (95%信頼区間)	F_0-RMSE
HTS	3.209±0.206	0.2102
F_0-multi	3.255±0.226	0.1590

包絡特徴量と非周期性指標からそれぞれ SPTK-3.6[1] を用いて求めた。F_0 モデルにおける基底周波数は 60 [Hz] とした。HMM は HTS-2.2[2] を用いて構築した。状態継続長を明示的に含んだ 5 状態 left-to-right HSMM を用い、各状態の出力は単一の対角共分散ガウス分布とし、決定木によるコンテキストクラスタリングを行い、木の停止基準には MDL 基準を用いた。

聴取実験ではサブセット J の 53 文からランダムに 10 文を選び、HTS による音声と F_0-multi による音声の 2 種類、計 20 個の文音声についてどの音声がどの手法によるものかを伝えず 11 名の日本語母国語話者に聴取させ、自然性を評価させた。評価は 5 段階とし、自然音声に近い品質である場合を 5、音声として十分に許容できる品質である場合を 4、音声として許容できる場合を 3、音声として何とか許容できる場合を 2、音声として許容できない場合を 1 としてスコアをつけた。

また、サブセット J の 53 文について各手法によって生成された F_0 パターンの対数値を F_0-RMSE によって評価した。F_0-RSME は次式で表される。

$$F_0\text{-RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\ln f_n - \ln f'_n)^2} \qquad (1)$$

ここで f_n、f'_n はそれぞれ、生成された対数 F_0 のフレームごとの値、参照音声の F_0 のフレームごとの値、N は合成文 53 文の合計のフレーム数である。提案手法による音声と参照音声による音声がどちらも有声であるフレームのみを考慮した。

3.2　結果

各評価による結果は Table 1 のようになった。聴取実験では提案法のほうが高い評価が出る傾向にあったものの、有意な差は見られなかった (t 検定による t 値が 0.745)。F_0-RMSE では提案法の方が良い結果となった。

コンテキストクラスタリングにおいて、フレーズ成分、アクセント成分、F_0 残差、それぞれの決定木の質問の選ばれ方について異なる傾向が見られた。フレーズ成分は呼気段落長に関する質問、アクセント成分はアクセント型に関する質問、F_0 残差は音素に関する質問が多く選ばれた。フレーズ成分は句頭から句末に向けての緩やかな下降に対応する成分、アクセント成分は日本語のアクセントに対応した成分、F_0 残差はフレーズ成分やアクセント成分と比べて短時間の情報に対応した成分であると考えられるため、コンテキストクラスタリングの結果はこうした各成分の特徴に対応した結果となっており、当初想定した傾向になることが確認できた。

4　おわりに

本稿では F_0 モデルのパラメータであるフレーズ成分、アクセント成分、F_0 残差を別々の特徴量ストリームとして学習・合成を行う HMM 音声合成の手法を提案し、比較評価によりその効果を確認した。主観評価では有意な差が得られなかったものの、コンテキストクラスタリングの結果を踏まえると、フレーズ成分、アクセント成分、F_0 残差に分けることで、合成音声の品質を下げることなくそれらを別々に取り扱うことができるため焦点付与などに応用できることが期待される。

フレーズ成分の先頭に位置するアクセント成分は大きい傾向にあるなど、両者には関係があるが、マルチストリーム学習では明示的に表現することが困難である。この対抗策としてマルチストリームによる学習を行うのではななくフレーズ成分を別枠で生成しその情報をアクセント成分を生成する際の HMM のコンテキストクラスタリングに含める手法などが考えられ、今後の課題として取り組んでいきたい。

参考文献

[1] K. Tokuda *et al.*, ICASSP, pp. 229–232, 1999.

[2] 広瀬 啓吉 “韻律と音声言語情報処理 アクセント・イントネーション・リズムの科学”, 丸善株式会社, 2006.

[3] H. Fujisaki *et al.*, J. Acoust. Soc. Japan (E), vol. 5, no. 4, pp. 233–242, 1984.

[4] H. Hashimoto *et al.*, INTERSPEECH, 2012.

[5] 百武恭汰他 “生成過程モデルにおける F0 パターン差分を考慮した HMM 音声合成の実験的検討”, 日本音響学会春季講演論文集, pp. 407-408, 2014.

[6] A. Kurematsu *et al.*, Speech Communication, vol. 9, pp. 357–363, 1990.

[7] H. Kawahara *et al.*, Speech Communication, vol. 27, no. 3–4, pp. 187–207, 1999.

[1] SPTK, http://sp-tk.sourceforge.net

[2] HTS, http://hts.sp.nitech.ac.jp