統計的手法を用いた電気式人工喉頭制御における遅延時間と予測精度の調査 *

☆田中 宏，戸田 智基，ニュービッグ グラム，サクティ サクリアニ，中村 哲（奈良先端大）

1 はじめに

喉頭摘出者のための発声補助器具である電気式人工喉頭に対して，我々は，統計的音源予測 [1] [2] を用いた電気式人工喉頭の音源制御法を提案している．本手法は，生成される音声（電気音声）の音響特徴量に基づき，統計的手法を用いて実時間で音源の基本周波数（F_0）を予測・制御するものであり，電気音声の自然性を大幅に改善できる可能性を示している [3]．一方で，実時間 F_0 予測処理において予測精度を高めるためには，200 msec 程度の遅延時間を必要とするため，調音動作に対して F_0 パターンがその分遅延するという問題がある．

本稿では，予測精度を劣化させることなく遅延時間を短縮することを目指し，連続 F_0 パターンのセグメント化学習および遅延時間を考慮した F_0 パターン学習を検討する．客観評価実験の結果から，提案法により，予測精度を劣化させることなく遅延時間を短縮できることを示す．

2 統計的手法に基づく音源予測

2.1 統計的音源予測

統計的手法に基づき，電気音声のスペクトル特徴量から通常音声の F_0 パターン [2] を予測する．本手法は，学習処理と変換処理で構成される．

学習処理では，電気音声と通常音声の同一発話データを用いて，変換モデルを学習する．各時間フレームにおいて，前後数フレームから得られる電気音声のスペクトルセグメント特徴量と，通常音声の対数 F_0 に対する静的・動的特徴量を抽出する．スペクトル距離尺度に基づく動的時間伸縮によりこれらを対応付けた結合ベクトルを用いて，結合確率密度関数を混合正規分布モデル（Gaussian Mixture Model; GMM）でモデル化する [4]．

変換処理では，学習された GMM を用いて，系列内変動を考慮した最尤系列変換法 [5] により，電気音声のスペクトルセグメント特徴量系列から通常音声の対数 F_0 系列へと変換する．なお，最尤系列変換法に対し，準最適的な単一分布系列およびカルマンフィルタによる近似を導入することで，短遅延変換に基づく実時間予測処理の実現が可能となる [6]．

2.2 連続 F_0 パターン

音源予測における変換処理は，有声区間単位で動作するため，連続 F_0 パターン（Continuous F_0; CF_0）を用いることで，発話単位でのフレーム間相関を考慮することが可能となる．学習処理において，通常音声から分析された不連続な F_0 パターン（図 2 の (a) 参照）の無声区間に対して，スプライン補間を施し，低域通過フィルタを用いてマイクロプロソディを除去することで，滑らかな連続 F_0 パターン（図 2 の (b) 参照）を抽出し，GMM によりモデル化する．得られた GMM を用いて，最尤系列変換法により発話単位で連続 F_0 パターンを予測することで，予測精度の改善がもたらされる [2]．

一方で，実時間予測処理においては，カルマンフィルタによる近似の影響が大きくなる．その結果，不連続な F_0 パターンの予測では 50 msec 程度の遅延時間を要するのに対し，連続 F_0 パターンの予測では 200 msec 程度の遅延時間が必要となる [7]．予測精度と遅延時間の関係を図 1 に示す．連続 F_0 パターンによる予測精度を保ちつつ，遅延時間を減らす枠組みの構築が必要がある．

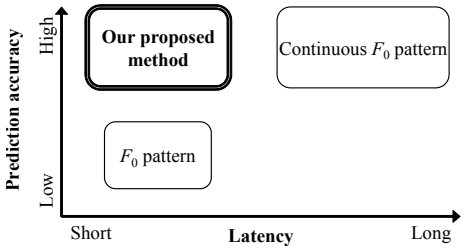

Fig. 1 予測精度と遅延時間の関係性．

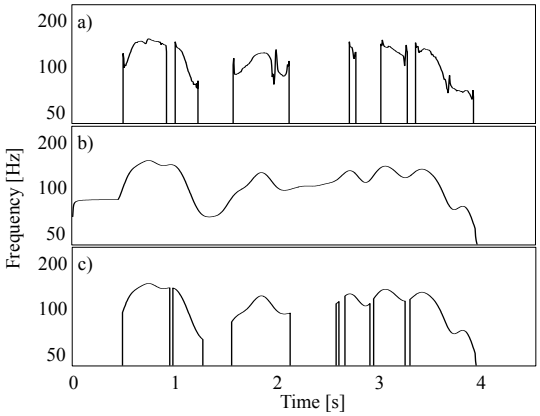

Fig. 2 a) 目標音声から抽出される F_0 パターン, b) 無声区間の補完およびマイクロプロソディの除去により得られる連続 F_0 パターン，および c) 波形パワーに基づくセグメント化連続 F_0 パターン．

3 遅延時間の低減

3.1 連続 F_0 パターンのセグメント化学習

遅延時間を低減するために，連続 F_0 パターンのセグメント化学習を提案する．発話単位の連続 F_0 パターンに対して，フレーム間相関を考慮するセグメントを決定し，個々のセグメントに対する連続 F_0 パターン（Segmented continuous F_0; Seg CF_0）を抽出する．本稿では，波形パワーに基づきセグメントを行う．波形パワーが閾値を下回るフレームについては無声と判定することで，セグメント化連続 F_0 パターン（図 2 の (c) 参照）を抽出する．得られたセグメント化連続 F_0 パターンを用いることで，GMM の学習を行う．変換時には，フレーム間相関を考慮する単位が各セグメントに限定されるため，カルマンフィルタによる近似の影響が抑えられる．

---

* An Investigation of Latency and Prediction Accuracy in Real-Time Control of Electrolarynx based on Statistical Excitation Prediction. by TANAKA, Kou, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani and NAKAMURA, Satoshi (NAIST)

3.2 遅延時間を考慮した F_0 パターン学習

実時間音源予測処理の際に，必要となる遅延時間の分だけ先読みした時間の F_0 パターンを予測することができれば，調音動作と同期した F_0 パターンを予測可能となる．そこで，学習処理において，必要となる遅延時間の分だけシフトさせた F_0 パターンを用いる．すなわち，実時間音源予測処理の遅延時間が 200 msec の際には，F_0 パターンを一律 -200 msec シフトさせることで，各時刻のスペクトルセグメント特徴量と 200 msec 先の F_0 特徴量を対応づけ，GMM の学習を行う．実時間変換処理においては，200 msec の遅延が発生するが，各時間フレームにおいて 200 msec 先の F_0 の予測が行われるため，調音動作との同期が保たれる．

4 実験的評価

4.1 実験条件

男性健常者 1 名による模擬電気音声を入力音声として使用し，女性健常者 1 名による通常音声を目標音声として使用する．学習データとして ATR 音素バランス文 A セットの 50 文中 40 文を用い，評価データとして残り 10 文を用いた 5 交差検定を行う．入力特徴量には，0〜24 次のメルケプストラム係数から得られるセグメント特徴量（前後 4 フレーム）を用いる．フレームシフト長は 5 ms とする．電気音声のスペクトル分析には，FFT 分析を用いる．電気式人工喉頭の F_0 は約 100 Hz である．連続 F_0 パターンを抽出する際に用いる低域通過フィルタのカットオフ周波数は 10 Hz とする．各種 F_0 パターンに対して，実時間音源予測処理の上限値として，発話単位の音源予測処理の予測精度を表 1 に示す．

4.2 連続 F_0 パターンのセグメント化学習の評価

目標音声の F_0 パターンと実時間音源予測により予測された各 F_0 パターンとの相関係数を図 3 に示す．なお，予測精度の収束する遅延時間を見るため，予測 F_0 パターンを遅延時間だけシフト補正した F_0 パターンとの間で相関係数を図る．連続 F_0 パターン予測において 200 msec 程度で，F_0 パターン予測において 50 msec 程度で予測精度が収束することがわかる．一方で，セグメント化連続 F_0 パターンでは，連続 F_0 パターンの予測精度を劣化させずに遅延時間を短縮可能であることがわかる．

4.3 遅延時間を考慮した F_0 パターン学習の評価

目標音声の F_0 パターンと実時間音源予測により得られる F_0 パターンおよび遅延時間を考慮した学習を利用した際に実時間予測される F_0 パターンとの相関係数を図 4 に示す．F_0 パターンとしては，連続 F_0 パターンを用いる．また，相関係数を計算する際には，遅延時間の補正は行わない．遅延時間を考慮せずに学習を行った際には，遅延時間が大きくなるほど入力音声との時間同期がずれるため，予測精度が劣化することが分かる．一方で，遅延時間を考慮した学習を利用することで，時間同期のずれを補正できることが分かる．250 msec 程度の遅延時間を想定すること

Table 1 発話単位における音源予測精度

	F_0	CF_0	Seg CF_0
The number of mixture components	32	16	16
F_0 correlation coefficients	0.40	0.48	0.46

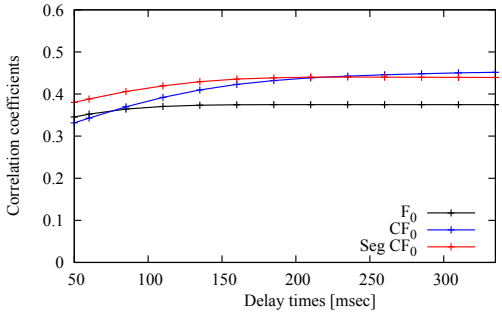

Fig. 3 目標音声の F_0 と実時間音源予測における予測 F_0 との相関係数（評価時に遅延時間補正あり）．

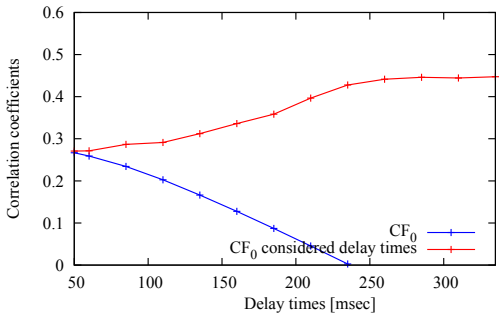

Fig. 4 目標音声の F_0 と実時間予測された F_0 との相関係数による従来の学習処理と遅延時間を考慮した学習処理の比較．

で，発話単位の音源予測処理と同程度の予測精度が達成できる．

5 おわりに

本稿では，実時間音源予測処理における予測精度を劣化させることなく遅延時間を短縮することを目指し，連続 F_0 パターンのセグメント化学習および遅延時間を考慮した F_0 パターン学習を提案した．客観評価実験の結果から，提案法により，予測精度を劣化させることなく遅延時間を低減できることを示した．今後は，実時間変換処理におけるカルマンフィルタを用いた近似の影響を考慮した予測モデルの最適化に取り組む．

謝辞 本研究の一部は，JSPS 科研費 15J10727 および 26280060 の助成を受け実施したものである．

参考文献

[1] K. Nakamura *et al.*, *SPECOM,* 54(1), pp. 134–146, Jun. 2012.
[2] K. Tanaka *et al.*, *IEICE Transactions on Information and Systems,* Vol. E97-D, No. 6, pp. 1429–1437, Jan. 2014.
[3] K. Tanaka *et al.*, *Proc. INTERSPEECH,* pp. 31–35, Sep. 2014.
[4] A. Kain *et al.*, *Proc. ICASSP,* pp. 285–288, May 1998.
[5] T. Toda *et al.*, *IEEE Trans. Audio, Speech, and Language,* 15(8), pp. 2222–2235, Nov. 2007.
[6] T. Toda *et al.*, *Proc. INTERSPEECH,* Sep. 2012.
[7] 田中 宏 他, **信学技報**, Vol. 115, No. 99, SP2015-9, pp. 4752, Jun. 2015.