統計的音源予測に基づく電気式人工喉頭のリアルタイム制御*

☆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代用発声法の一つとして，電気式人工喉頭を用いた発声法がある．外部から機械的に生成される音源信号を用いて発声を行う方法であり，比較的聞き取りやすい音声（電気音声）を生成できる．一方で，発話内容に応じた自然な F_0 パターンを機械的に生成するのは極めて難しく，電気音声の自然性は著しく劣化する．この問題に対して，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 [3] を提案しており，シミュレーション実験により，自然性を大幅に改善できる可能性を示している．

本稿では，提案法の実装を行い，実機システム使用による性能を評価する．客観評価結果より，実機システムにおいても，シミュレーション実験結果と同等の音源予測精度が得られることを示す．

2 統計的音源予測を用いた電気式人工喉頭の音源制御

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連続 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リアルタイム予測処理の実現が可能となる [6]．また，変換処理遅延は，50〜70 ms 程度に抑えることができる．

2.2 電気式人工喉頭のリアルタイム音源制御

統計的音源予測により得られる F_0 パターンを用いて，電気式人工喉頭から生成される音源信号の F_0 を直接制御する [3]．本手法の処理過程を図 1 の左図に示す．

本システムを用いた発声は，1）喉頭摘出者が調音する過程と，2）発声された電気音声から F_0 値をリアルタイムに予測し電気式人工喉頭の音源信号を制御する処理により行われる．前者は，従来の電気式人工喉頭を用いた発声法における生成過程と同一である．一方，後者では，前者の生成過程で得られた電気音声からリアルタイム予測される F_0 値に応じて，電気式人工喉頭の音源信号の F_0 を制御する電圧を変化させる．結果，電気式人工喉頭からは発話内容に応じた F_0 パターンを持つ音源信号が生成され，喉頭摘出者はより自然な電気音声を発声することができる．喉頭摘出者による通常の発声動作に基づき F_0 パターンが予測される枠組みであるため，呼気や手の動作などに基づく意識的な F_0 制御 [7] [8] を必要とせず，従来の電気式人工喉頭と同様に使用することができる．また，通常音声の統計量を用いることで，より自然な F_0 パターンの予測が可能となる．一方で，調音動作と F_0 パターンの間には，リアルタイム予測処理に起因する遅延が必ず生じる．

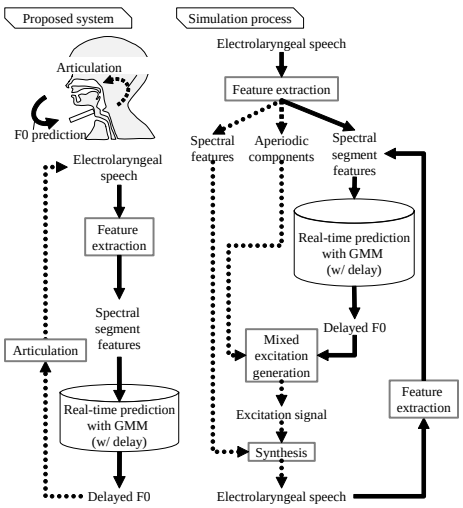

Fig. 1 電気式人工喉頭のリアルタイム制御システムとそのシミュレーションの流れ

我々は，これまでに，本システムのシミュレーション実験を行い，その性能を評価している．シミュレーション処理を図 1 の右図に示す．シミュレーション実験の結果から，本システムによりオフライン処理と同等の F_0 パターン予測精度が得られること（ただし，リアルタイム変換処理はオフライン変換処理と遅延付与処理により近似 [3]），調音動作に対する F_0 パターンの遅れは自然性に大きな影響を与えず，本システムにより電気音声の自然性を大幅に改善できることが示されている．

2.3 実機への実装

PC と D/A 変換器を用いて本システムの実装を行う．発声される電気音声をヘッドセットマイクで収録し，PC へと入力する．PC にて，入力される電気音声から連続 F_0 パターンを予測し，電気式人工喉頭へと入力する電圧（生成される音源信号の F_0 に対応）を決定する．そして，D/A 変換器を用いて，電圧を電気式人工喉頭へと入力する．実装に用いた機器等の情報を表 1 に示す．なお，D/A 変換器の仕様により，出力すべき電圧値の書き込みに約 50 ms 程度の時間を要する．書き込み開始時には，書き込む電圧値を確定しておく必要があるため，図 2 に示す処理により実装する．結果，F_0 パターンの遅延時間は，約 150 ms 程度となる．

3 実験的評価

3.1 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

*A Real-Time Control of Electrolarynx Based on Statistical Excitation Prediction. by TANAKA, Kou, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani and NAKAMURA, Satoshi (NAIST)

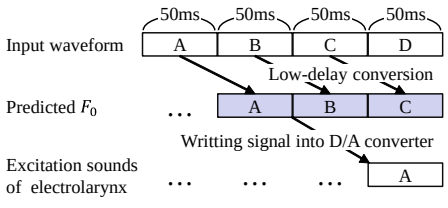

Fig. 2　実機実装に関する処理の流れとその遅延時間

Table 1　実装に用いた機器

電気式人工喉頭	ユアトーン 2 ・ゆらぎ
PC の CPU	Intel(R) Core(TM) i5-4200U
D/A 変換器	AIO-160802AY-USB

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．一方で，目標とする女性健常者の F_0 平均は約 220 Hz であるため，予測 F_0 に対して平均シフトを施し，約 100 Hz の平均を持つ F_0 パターンを生成する．

実機実装したシステムを用いて予測された F_0 パターンと，図 1 の右図に示すシミュレーションにより予測された F_0 パターンの比較を行う．シミュレーションにおいても，実機システムと同様に，リアルタイム予測処理を行う．リアルタイム予測処理における遅延フレーム数は 3 とする．

3.2　実装の妥当性評価

実機システムにより予測される F_0 パターンと，シミュレーションにおける各反復時に予測される F_0 パターン間の相関係数を，図 3 に示す．本実機システムにより，シミュレーション結果と十分に高い相関を持つ F_0 パターンが得られる．このことから，提案システムは実機上で動作可能であること，また，シミュレーション処理の精度は十分に高いことが分かる．

3.3　リアルタイム音源予測における遅延フレーム数の調査

実機システムによる発話を行ったところ，電気音声の高い明瞭性を損なうことなく，自然性が大きく改善されることを確認した．一方で，オフライン予測処理で予測される F_0 パターンと比較すると，実機システムで予測される F_0 パターンは，若干の自然性低下が感じられた．リアルタイム予測処理における遅延フレーム数が予測精度に与える影響については，スペクトル予測では検討がなされているが [9]，連続 F_0 パターン予測ではこれまでに詳細な検討がなされていない．そこで，本稿では，遅延フレーム数が予測 F_0 パターンに与える影響を明らかにする．

オフライン予測処理で得られる予測 F_0 パターンと，オンライン予測処理の遅延フレーム数を変化させて得られる予測 F_0 パターンの間の相関係数を調査した結果を，図 4 に示す．連続 F_0 パターン予測において，オフライン予測処理と同様の結果を得るためには，数十フレーム遅延させる必要があることが分かる．スペクトル予測処理では 3〜5 フレーム遅延で十分な結果が得られること [9] と比較すると，明らかにより多くの遅延フレーム数を必要とすることが分かる．本システムにおいて，遅延フレーム数の増加は，調音動作に対する F_0 パターンの遅延を直接増加させる要因となるため，聴感上許容可能な最大遅延フレーム数の調査や必要となる遅延フレーム数の減少に取り組む必要がある．

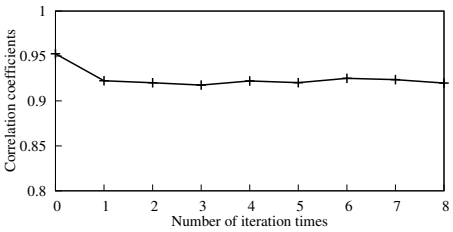

Fig. 3　シミュレーションにおける予測 F_0 と実機システムにおける予測 F_0 との相関係数

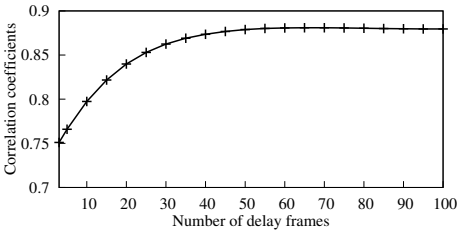

Fig. 4　遅延フレーム数が F_0 推定精度に与える影響

4　おわりに

本稿では，統計的音源予測に基づく電気式人工喉頭のリアルタイム制御の実装を行い，実機システム使用時の性能を評価した．客観評価実験結果より，実機システムはシミュレーションと同等の性能を有することを示し，実装の妥当性を示した．今後は，強調音声の主観評価実験およびリアルタイム音源予測処理の改善を行う．

謝辞　本研究の一部は，JSPS 科研費 26280060 および 23240023 の助成を受け実施したものである．ユアトーン 2・ゆらぎへの電圧入力に関する助言を頂いた株式会社電制の須貝保徳氏に感謝する．

参考文献

[1] K. Nakamura *et al.*, *SPECOM,* 54(1), pp. 134–146, Jan 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] N. Uemi *et al.*, *Proc. 3rd IEEE International Workshop of Robot and Human Communication,* pp.198–203, Jul. 1994.
[8] K. Matsui *et al.*, *Proc. SSW8,* pp.273–277, Aug. 2013.
[9] T. Muramatsu *et al.*, *Proc. INTERSPEECH,* pp. 1076–1079, Sep. 2008.