統計的音源予測を用いた電気式人工喉頭制御法の評価*

◎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，自然性が大幅に改善されることを示している．しかしながら，これは喉頭摘出者 1 名に対してのみの評価結果であり，提案法の話者依存性については未だ不明である．電気音声の品質自体が発声法の習熟度に大きく依存するため，複数話者による評価は重要である．また，明瞭性を劣化させずに自然性を改善させる技術が望まれているため，明瞭性に関する評価は必要不可欠である．

本稿では，複数の話者に対するシミュレーションを用いて，提案法の話者依存性を調査するとともに，明瞭性の詳細な評価を行う．評価実験より，提案法は話者に依存せず，電気音声の明瞭性を維持したまま，自然性を大幅に改善できることを示す．

2 統計的音源予測を用いた制御法

2.1 統計的音源予測

電気音声のスペクトル特徴量と通常音声の連続 F_0 パターン [2] の統計量に基づき、通常音声の連続 F_0 パターンを予測する．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，通常音声の静的·動的特徴量を抽出する．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，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音源特徴量系列へと変換する．なお，最尤系列変換法に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変換処理の実現が可能となる [6]．

2.2 電気式人工喉頭の直接制御法

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 F_0 を直接制御する手法 [3] の処理過程を図 1 の左図に示す．

本システムでは，発声された電気音声から F_0 をリアルタイム予測する処理において，50〜70 ms 程度の遅延が生じるため [6]，調音動作に対して F_0 パターンが遅延した音声が生成される．また，予測された F_0 に基づいて生成された電気音声が，次の時刻における入力として用いられる．通常，入力特徴量である電気音声のスペクトルセグメント特徴量を抽出する際に，演算量の少ない FFT 分析が用いられるが，F_0 の影響を受けやすいため，予測精度が劣化する可能性がある．この影響を緩和するため，STRAIGHT 分析 [7] や学習データ生成処理 [3] を導入する．

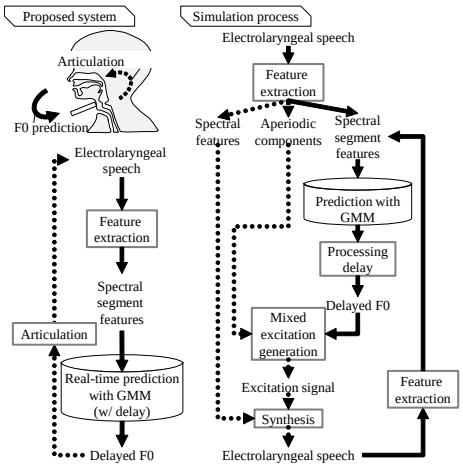

Fig. 1 電気式人工喉頭の直接制御システムとそのシミュレーションの流れ

3 シミュレーション実験

提案処理を電気式人工喉頭に組み込む前段階として，シミュレーションを行う [3]．本シミュレーション処理過程を図 1 の右図に示す．まず事前段階として，1）電気音声に対して STRAIGHT 分析を行い，スペクトル特徴量および非周期成分を抽出しておく．そして，2）電気音声からスペクトルセグメント特徴量を抽出し，F_0 予測を行う．3）リアルタイム予測処理による遅延時間を考慮するため，F_0 を遅延させる．4）得られた F_0 と事前に抽出しておいた非周期成分を用いて，混合励振源モデル [8] により音源信号を生成する．5）音源信号に対して，事前に抽出しておいたスペクトル特徴量を畳み込むことで，予測 F_0 による電気式人工喉頭制御を行った際の電気音声を仮想的に生成する．6）生成された電気音声を新たな入力とし，予測結果が安定するまで 2〜6 の処理を反復的に繰り返す．

4 実験的評価

4.1 実験条件

喉頭摘出者 2 名の電気音声と健常者 1 名による模擬電気音声を入力音声に，健常者 1 名の通常音声を目標音声に用いる．学習データに ATR 音素バランス文 A セットの 50 文中 40 文を用い，評価データに残り 10 文を用い，5 交差検定を行う．入力特徴量として，0〜24 次のメルケプストラムセグメント特徴量（前後 4 フレーム）を用いる．スペクトル分析は，電気音声に対しては FFT 分析及び STRAIGHT 分析を，通常音声に対しては STRAIGHT 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，目標とする健常者の F_0 平均は約 220 Hz である．学習データ生成処理では，電気音声の F_0 を 150，200，250 Hz とシフトさせた分析再合成音声を作成し，元の 100 Hz のものとあわせて計 160 文を用いる．リアルタイム予測処理に起因する遅延時間は 70 ms とする．

シミュレーションにより得られる強調音声を客観評価実験および主観評価実験により評価する．客観評価実験では，目標音声の F_0 と予測 F_0 間の相関係数により，F_0 推定精度を評価する．主観評価実験では，強調音声について，明瞭性に関する書き取り試験および自然性に関する 5 段階オピニオン評定により評価する．評価する音声は以下の 4 つである．

*Evaluation of F_0 control of electrolarynx based on statistical excitation feature prediction. by TANAKA, Kou, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani and NAKAMURA, Satoshi (NAIST)

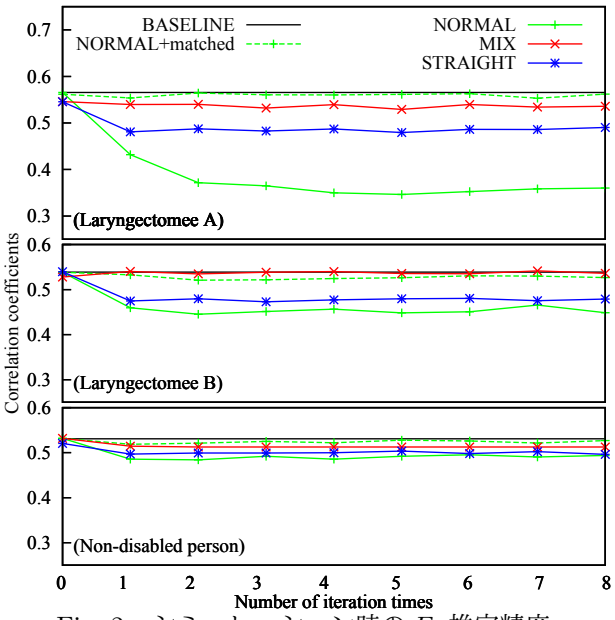

Fig. 2　シミュレーション時の F_0 推定精度

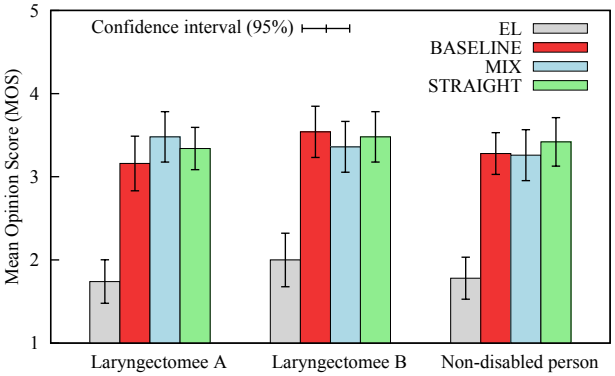

Fig. 3　自然性の主観評価結果

- EL: 元の電気音声
- BASELINE: 統計的音源予測に基づく予測 F_0 に電気音声のスペクトル特徴量を畳み込んだ遅延なし強調音声（ [2] のハイブリッドシステムにおける雑音抑圧処理なしに相当）
- MIX: 学習データ生成処理を用いた提案法によるシミュレーション強調音声
- STRAIGHT: STRAIGHT 分析を用いた提案法によるシミュレーション強調音声

なお，客観評価においては，予測 F_0 が予測精度に与える影響を明らかにするため，学習データ生成処理および STRAIGHT 分析を用いずに，単なる FFT 分析を行う際の提案法によるシミュレーション強調音声（NORMAL）も評価する．また，予測 F_0 に対して，平均が 100 Hz になるようにシフト処理を施した際も併せて評価する（+matched）．

4.2　実験結果

図 2 に F_0 推定精度を示す．提案法において，予測 F_0 をシフトした際（+matched）は，従来法（BASELINE）とほぼ同等の推定精度が得られる．これは，生成される電気音声と学習時に用いる電気音声との間に，大きな F_0 の差が生じないためである．一方で，FFT 分析使用時に予測 F_0 をシフトしない場合（NORMAL），推定精度が大きく劣化する．なお，その際の精度の劣化程度は話者依存であることがわかる．このことから，FFT 分析では，入力特徴量抽出時に F_0 の影響を強く受けることが分かる．これに対して，STRAIGHT 分析（STRAIGHT）を導入することで，推定精度の劣化を抑えることが可能であり，

Table 1　明瞭性に関する書き取り試験結果

	Word correct rate [%]	Word accuracy [%]	Number of replays
Laryngectomee A			
EL	97.88	98.15	1.39
BASELINE	96.72	96.45	1.79
MIX	98.43	97.73	1.21
STRAIGHT	98.29	98.15	1.40
Laryngectomee B			
EL	97.03	92.76	1.58
BASELINE	94.85	88.92	1.67
MIX	95.94	93.89	1.46
STRAIGHT	96.80	94.46	1.67
Non-disabled person			
EL	97.09	94.74	1.87
BASELINE	96.76	93.18	2.27
MIX	97.11	95.60	1.58
STRAIGHT	97.21	94.03	1.87

学習データ生成処理 (MIX) においては BASELINE の制度を維持している．

図 3 に自然性に関する主観評価結果を示す．話者に限らず電気音声（EL）の自然性は著しく低いのに比べて，他の手法では大幅な改善がみられる．また，BASELINE，MIX，及び STRAIGHT の間で有意差がないことから，各手法で予測された F_0 に相違はあるものの，自然性における違いは知覚されず，また，F_0 パターンの遅延は影響を及ぼさないことがわかる．

表 1 に明瞭性に関する書き取り試験結果を示す．話者に依存せず，EL は高い明瞭性を持つのに対して，BASELINE では僅かに劣化する傾向がある．一方で，単語正解精度において，MIX は話者に限らず電気音声の高い明瞭性を保持している．また，聞きなおし回数に関しても MIX が一番少ないことより，学習データ生成処理の導入は有効であるといえる．

5　おわりに

本稿では，統計的音源予測を用いた電気式人工喉頭制御法を複数の話者に適用し，シミュレーション実験を通して明瞭性および自然性に関して主観評価を行った．客観評価実験結果から提案システムは話者に限らず頑健に動作可能であることを示した．また，主観評価実験結果より，本システムは電気音声の高い明瞭性を劣化させずに，自然性の大幅な改善が可能であることを示した．今後は，統計的音源予測における予測精度の改善，及び提案システムの構築を行う．

謝辞　本研究の一部は，JSPS 科研費 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, Jan 2012.
[2] K. Tanaka *et al.*, *IEICE Transactions on Information and Systems,* Vol. E97-D, No. 6, pp. 1429–1437, Jun. 2014.
[3] 田中 宏 他, **信学技報**, Vol. 114, No. 91, SP2014-52, pp. 33–38, Jun. 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] H. Kawahara *et al.*, *SPECOM,* 27(3-4), pp. 187–207, Apr 1999.
[8] 大谷 大和 他, **信学論**, Vol. J91-D, No. 4, pp. 1082–1091, Apr. 2008.