統計的音源予測に基づく電気式人工喉頭制御法*

○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，習得が容易で，かつ，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，明瞭性を劣化させずに自然性を大幅に改善する方法として，雑音抑圧に基づくスペクトル補正 [1] と統計的声質変換 [2] に基づく統計的音源予測を用いたハイブリッドな電気音声強調法 [3] を提案している．この枠組みでは，発声された電気音声をマイクで収録し，強調音声をスピーカから出力する．そのため，発声された電気音声と強調音声が同時に外部に提示される．聞き手が話者から離れており，強調音声のみを提示できる状況（例えば電話など）では有効であるが，対面会話には不向きである．

本研究では，対面会話においても使用可能な電気音声強調法として，統計的音源予測を用いた電気式人工喉頭の直接制御を目指す．本稿では，その前段階としてシミュレーション実験を行い，実験結果から提案法の有効性を示す．

2 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音源特徴量の統計量に基づき、通常音声の音源特徴量を予測する。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・動的音源特徴量を抽出する．動的時間伸縮（Dynamic Time Warping; DTW）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（Global Variance; GV）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]．

3 電気式人工喉頭の F_0 パターン制御法

3.1 電気式人工喉頭の直接制御システム

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を提案する．提案法の処理過程を図 1 の左図に示す．なお，対面会話での使用を想定してリアルタイム統計的音源予測処理を用いる．

本システムは，1）電気式人工喉頭から生成される音源信号を調音器官により調音する過程と，2）発声された電気音声から F_0 をリアルタイムに予測する処理により構成される．後者では，リアルタイム予測処理において 50〜70 ms 程度の遅延が生じる [6]．従来の電気音声強調処理においては，全音声パラメータを同期させ，強調音声生成処理自体を遅延させることができるが，提案法においては，調音動作に対して F_0 パターンが遅延する．

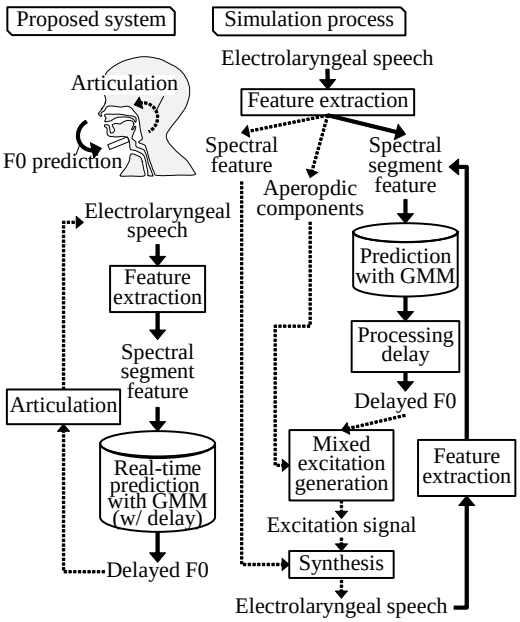

Fig. 1 電気式人工喉頭の直接制御システムとそのシミュレーションの流れ

提案法により生成される電気音声は，予測された F_0 の影響を受けたものとなり，それが次の時間における予測処理で用いられる．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，予測された 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．この問題に対して，STRAIGHT 分析 [7] および学習データ生成処理の導入を検討する．STRAIGHT 分析は，F_0 の影響を大幅に低減できる．また，分析時に予測 F_0 を直接用いることで，F_0 抽出処理を回避し，演算量を大幅に削減する．一方で，学習データ生成処理に関しては，従来通り FFT 分析を使用する．FFT 分析における F_0 の影響を考慮した GMM を構築するために，学習データとして用いる電気音声に対して，STRAIGHT 分析合成処理を施し，様々な高さの F_0 を持つ電気音声を合成して，それらを全て同時に学習データとして使用する．

3.2 シミュレーション

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．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．これらは，調音情報を有しており，生成される電気音声，および，電気式人工喉頭から外部に漏れ出す音源信号の両者の影響を受けたものとなる．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．

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

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 1 名の電気音声と，健常者 1 名の通常音声を用いる．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．サンプリング周波数は 16 kHz，分析フレームシフト長は 5 ms，FFT 分析におけるフレーム長は 25 ms とする．入力特徴量に，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．収録に用いた電気式人工喉頭の F_0 はほぼ一定であり，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，元の 100 Hz のものとあわせて計 160 文を用いる．F_0 推定用 GMM の混合数は 32 とする．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，聞き取りやすさ及び自然性に関する 5 段階オピニオン評定より評価する．評価する音声は以下の 4 つである．

- 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（ [3] 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，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（F0fix）．

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（F0fix）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．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）や学習データ生成処理 (MIX) 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る．

図 3 上段に聞き取りやすさに関する主観評価結果を示す．BASELINE は電気音声（EL）と同程度の聞き取りやすさを持つ．また，STRAIGHT も同様に高い聞き取りやすさを持つことが分かる．一方で，MIX では若干劣化する傾向が見られる．MIX では，不安定な F_0 パターンが生成されることがあり，それが聞き取りやすさを劣化させる要因になったと考えられる．この原因については，今後さらに検討が必要である．

図 3 下段に自然性に関する主観評価結果を示す．電気音声の自然性は著しく低いのに比べて，他の手法で

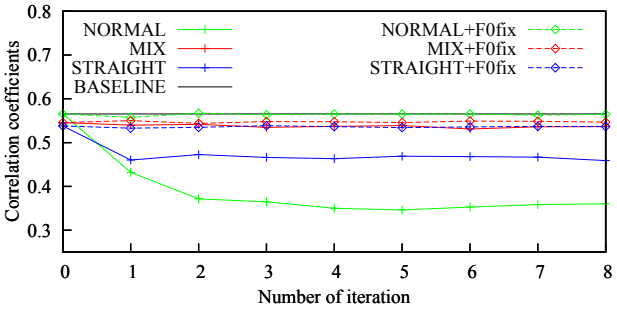

Fig. 2　シミュレーション時の F_0 推定精度

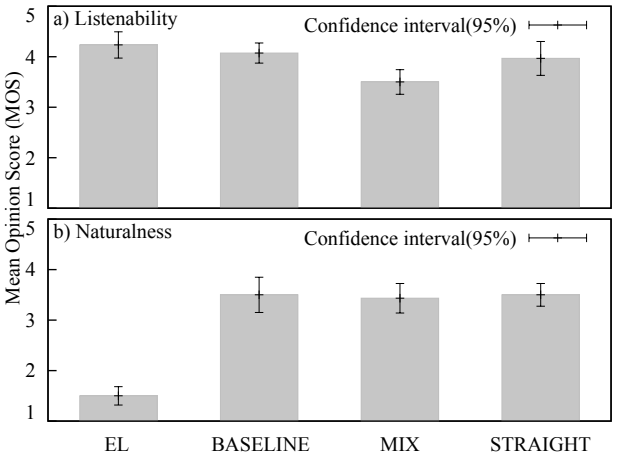

Fig. 3　聞き取りやすさ及び自然性の主観評価結果

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ないことが分かる．

5　おわりに

本稿では，対面会話において使用可能な電気音声強調法として，統計的音源予測を用いた電気式人工喉頭の直接制御を目指し，その前段階として，その際に生じる遅延時間及び予測 F_0 が与える影響をシミュレーション実験により調査した．客観評価実験結果から提案システムは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，提案システムは従来のハイブリッドな電気音声強調法 [3] と同等の明瞭性及び自然性を有することが分かった．今後は，書き取り試験による明瞭性の評価，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 22680016 の助成を受け実施したものである．

参考文献

[1] H. Liu *et al.*, *IEEE Trans. Biomedical Engineering,* 53(5), pp. 865–874, May 2006.
[2] K. Nakamura *et al.*, *SPECOM,* 54(1), pp. 134–146, Jan 2012.
[3] K. Tanaka *et al.*, *Proc. INTERSPEECH,* pp.3067–3071, Aug. 2013.
[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.