社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

統計的音源予測に基づく電気式人工喉頭制御法のシミュレーションによる評価

田中　宏† 戸田　智基† グラム・ニュービッグ† サクリアニ・サクティ† 中村　哲†

† 奈良先端科学技術大学院大学情報科学研究科，
〒 630-0101，奈良県生駒市高山町 8916-5
E-mail: †{ko-t,tomoki,neubig,ssakti,s-nakamura}@is.naist.jp

あらまし　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 パターンの機械的な生成は極めて難しく，電気音声の自然性は著しく劣化する．この問題に対して，統計的音源予測に基づき，電気音声のスペクトル特徴量から通常音声の音源特徴量を予測し，ボコーダを用いて電気音声に付与することで強調処理を行う手法を提案している．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枠組みであるため，リアルタイム処理時には，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（電話など）では有効であるが，聞き手が話者に近く，両方の音声が相手に提示される場合（対面会話など）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法を提案する．本稿では，提案法を電気式人工喉頭に実装する前段階として，シミュレーション実験による評価を行う．実験結果から，提案法により，自然性が大幅に改善された電気音声の生成が可能となることを示す．
キーワード　電気式人工喉頭，電気音声，統計的音源予測，F_0 制御，シミュレーション評価

An Evaluation through Simulation for Direct F_0 Control of an Electrolarynx based on Statistical Excitation Feature Prediction

Kou TANAKA†, Tomoki TODA†, Graham NEUBIG†, Sakriani SAKTI†, and Satoshi NAKAMURA†

† Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama-cho, Ikoma-shi, 630-0101, Japan
E-mail: †{ko-t,tomoki,neubig,ssakti,s-nakamura}@is.naist.jp

Abstract　An electrolarynx is a device that artificially generates excitation sounds to enable laryngectomees to produce electrolaryngeal (EL) speech. Although proficient laryngectomees can produce quite intelligible EL speech, it sounds very unnatural due to the mechanical excitation produced by the device. To address this issue, we have proposed several EL speech enhancement methods using statistical excitation prediction, which was essential to significantly improve naturalness by predicting excitation parameters of normal speech. In these methods, the original EL speech is recorded with a microphone and the enhanced EL speech is presented from a loudspeaker in real time. This framework is effective for telecommunication but it is not suitable to face-to-face conversation because both the original EL speech and the enhanced EL speech are presented to listeners. In this paper, we propose direct F_0 control of the electrolarynx based on the statistical excitation prediction also effective for face-to-face conversation. A simulation experiment is conducted to evaluate the effectiveness of the proposed method. The experimental result shows that our proposed system enables laryngectomees to produce more natural EL speech.
Key words　electrolarynx, electrolaryngeal speech, statistical excitation prediction, Direct F_0 control, evaluation through simulation

1. まえがき

音声は，人々がお互いにコミュニケーションを取るうえで，基本的な手段の 1 つである．しかしながら，喉頭摘出者は音声を自然な形で発声することが難しい．通常の音声生成過程では，肺からの呼気により声帯を振動させることで音源信号を生成し，それを調音することで音声を生成する（図 1 左を参照）．一方で、喉頭摘出者は，多くの場合声帯を摘出するため，音源生成機能を失う．そのため，声帯振動を用いずに音源信号を生成する発声法が必要となり，深刻な発声障害を患う．

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．図 1 右に示す通り，外部から生成された音源信号が声道内に伝達し，調音されることで音声（電気音声）が生成される．電気式人工喉頭を用いた発声法は，1) 習得が容易である，2) 発声時に身体への負担が少ない，3) 他の代用発声法と比べ，比較的高い明瞭性を持つ音声を生成できる，といった利点がある．一方で，1) 電気式人工喉頭の音源信号自体が外部に漏れ出し，雑音として電気音声に混入するため，電気音声の品質が劣化する，2) 自然な音源信号を外部から機械的に生成するのは困難であり，電気音声の自然性は著しく低下する，といった欠点がある．特に，2 つ目の欠点は，電気音声が持つ最大の欠点であり，通常音声と大きく乖離させる主要因である．

この問題に対処するため，電気式人工喉頭の生成する音源信号の基本周波数（fundamental frequency: F_0）を制御する方法として，1）呼気圧を用いて電気式人工喉頭を制御する方法 [1] や，2）スライダーボタンを用いて制御する方法 [2]，3）手の動きを用いて制御する方法 [3] などが提案されている．これらにより，いずれも自然性の改善した電気音声を生成可能となる．一方で，呼気圧や手などの動作から自然な F_0 パターンを生成するのは容易ではなく，また，発話内容に沿った自然な F_0 パターンを意識的に制御するのは極めて困難な処理となる．

我々は，特に意識せずとも，発話内容に沿った F_0 パターンを持つ音源信号を付与することができる電気音声強調法として，雑音抑圧に基づくスペクトル補正 [4] および統計的声質変換 [5] [6] に基づく統計的音源予測 [7] [8] を組み合わせたハイブリッド強調法 [9] を提案している．雑音抑圧処理により，電気音声に雑音として混入する音源信号の影響を緩和する．一方，統計的手法により，電気音声および通常音声の同一発話データから得られる統計量を用いて，電気音声のスペクトル特徴量から通常音声の音源特徴量への変換を行う．発話内容に沿ったより自然な F_0 パターンを予測することで，より自然な音源信号を生成することができる．得られた音源信号と雑音抑圧後のスペクトル包絡を用いてボコーダによる波形合成を行うことで，元の電気音声と比較し，明瞭性を保持しながら大幅に自然性が改善された強調音声の生成が可能となる．一方で，電気音声をマイク収録し，強調音声をスピーカーから出力するという枠組みであるため，リアルタイム動作時には，電気音声と強調音声の両方が同時に提示される．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（例えば電話や講演など）では有効であるが，聞き手が話者に近く，両方の音声が相手に提示される場合（例えば対面会話など）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自動制御法を提案する．提案法では，スピーカーを使用せず，電気音声をマイク収録し，統計的音源予測に基づいて F_0 の値を予測し，予測された F_0 値に応じて電気式人工喉頭から生成される音源を直接リアルタイム制御する．結果，喉頭摘出者は，発話内容に応じた F_0 パターンを持つ電気音声の発声が可能となる．本稿では，提案法を電気式人工喉頭に実装する前段階として，シミュレーション実験による評価を行う．客観評価実験および主観評価実験の結果から，提案法により生成される電気音声の自然性は大幅に改善されることを示す．

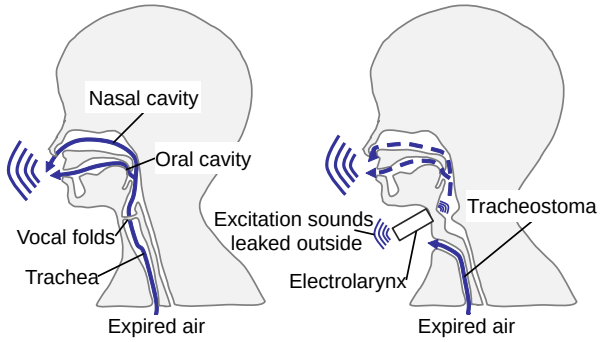

図 1　Speech production mechanisms of non-disabled people (left figure) and total laryngectomees (right figure).

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（図 2）．学習処理では，入力話者と目標話者の同一文発話対（パラレルデータ）を用いて，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対応関係をモデル化する．変換処理では，電気音声と通常音声の統計量に基づき，入力された電気音声のスペクトル特徴量に対して，対応する通常音声の音源特徴量を求める．

2.1 学習処理

時間フレーム t における電気音声のスペクトルセグメント特徴量（D_x 次元ベクトル）を $\boldsymbol{X}_t$ とし，前後 C フレームの情報を用いて，次式により抽出する．

$$\boldsymbol{X}_t = \boldsymbol{E}[\boldsymbol{x}_{t-C}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+C}^\top]^\top + \boldsymbol{f} \tag{1}$$

ここで，$\boldsymbol{x}_t$ は時間フレーム t におけるスペクトル特徴量を表す．$\boldsymbol{E}$ および $\boldsymbol{f}$ は各々変換行列およびバイアスベクトルを表し，学習データの全フレームにおけるスペクトル特徴量に対する主成分分析により求める．$\top$ は転置を表す．一方で，通常音声音源特徴量として，$\boldsymbol{Y}_t = [y_t,\ \Delta y_t]^\top$ を使用する．ここで，動的特徴量 Δy_t は $\Delta y_t = y_t - y_{t-1}$ により計算する．本研究では，スペクトル特徴量としてメルケプストラムを用い，音源特徴量として，無声音および非発話区間に対してスプライン補間を施したのち，低域通過フィルタを用いてマイクロプロソディを除去した，滑らかな連続対数 F_0 パターン [9] を用いる．

パラレルデータに対して，[10] に示す手順に従い動的時間伸縮（Dynamic time wrapping; DTW）を行い，入力特徴量 $\boldsymbol{X}_t$ と出力特徴量 $\boldsymbol{Y}_t$ の対応付けを行った結合ベクトル $[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$ を用いて，次式に示す通り，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 [11].

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda})$$

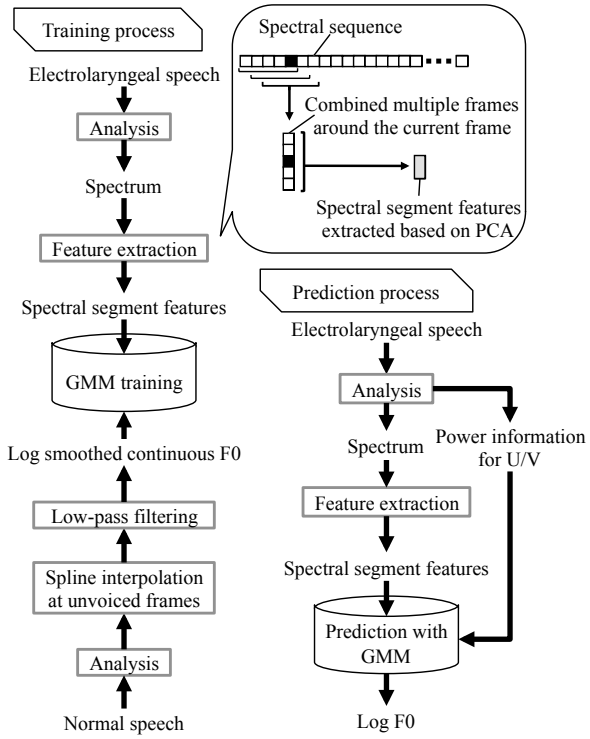

図 2　The training and prediction process.

$$= \sum_{m=1}^{M} \alpha_m \mathcal{N}\left([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top; \boldsymbol{\mu}_m^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}\right) \tag{2}$$

ここで，$\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は平均ベクトル $\boldsymbol{\mu}$，および共分散行列 $\boldsymbol{\Sigma}$ を持つ正規分布である．また，$\boldsymbol{\lambda}$ はモデルパラメータセットを表し，各分布 m の混合重み α_m，平均ベクトル $\boldsymbol{\mu}_m^{(X,Y)}$ および共分散行列 $\boldsymbol{\Sigma}_m^{(X,Y)}$ で構成される．ここで，m 番目の分布において，平均ベクトル $\boldsymbol{\mu}_m^{(X,Y)}$ および共分散行列 $\boldsymbol{\Sigma}_m^{(X,Y)}$ は次式で表される．

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \tag{3}$$

ここで，$\boldsymbol{\mu}_m^{(X)}$ および $\boldsymbol{\mu}_m^{(Y)}$ は入力特徴量および出力特徴量の平均ベクトルを表し，$\boldsymbol{\Sigma}_m^{(XX)}$ および $\boldsymbol{\Sigma}_m^{(YY)}$ は入力特徴量および出力特徴量の共分散行列，$\boldsymbol{\Sigma}_m^{(XY)}$ および $\boldsymbol{\Sigma}_m^{(YX)}$ は相互共分散行列を表す．

また，目標特徴量である通常音声の音源特徴量（F_0 パターン）に対しては，系列内変動（Global variance; GV）[6] の確率密度関数も学習する．ここで，GV $v^{(y)}$ は，通常音声の静的特徴量系列に対して，発話ごとに次式で計算される．

$$v^{(y)} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \frac{1}{T}\sum_{\gamma=1}^{T} y_\gamma)^2 \tag{4}$$

ここで，y_t はフレーム t の通常音声の静的特徴量である．GV の確率密度関数 $P(v^{(y)}|\boldsymbol{\lambda}^{(v)})$ は，平均 $\mu^{(v)}$ および分散 $s^{(vv)}$ の正規分布を用いて，以下のようにモデル化する．

$$P(v^{(y)}|\boldsymbol{\lambda}^{(v)}) = \mathcal{N}\left(v^{(y)}; \mu^{(v)}, \sigma^{(vv)}\right) \tag{5}$$

2.2　変 換 処 理

変換部では，学習された GMM を用いて，最尤系列変換法 [6] により，電気音声の入力特徴量系列から通常音声の出力特徴量系列へと変換する．時間フレーム 1 から T までの電気音声および通常音声の特徴量系列を $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_t^\top, \cdots, \boldsymbol{X}_T^\top]^\top$, $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_t^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$ とおく．このとき，変換後の静的特徴量系列 $\hat{\boldsymbol{y}} = [\hat{y}_1, \cdots, \hat{y}_t, \cdots, \hat{y}_T]^\top$ は次式で計算される．

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) P(v^{(y)}|\boldsymbol{\lambda}^{(v)})^\omega$$
$$\text{subject to}\, \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{6}$$

ここで，$\boldsymbol{W}$ は静的特徴量系列 $\boldsymbol{y}$ を静的・動的特徴量系列 $\boldsymbol{Y}$ に写像する変換行列変換行列を表す．また，ω は GV の尤度重みを表し，本稿では $1/2T$ とする．なお，電気音声強調においては，発話区間は全て有声とする F_0 パターンの有効性が確認されている [9]．本稿では，発話区間／非発話区間は入力される電気音声波形のパワー情報から決定する．

電気音声強調では，最尤系列変換法に近似を導入したリアルタイム変換処理 [12] を用いる．まず，準最適な分布系列 $\hat{\boldsymbol{m}} = [\hat{m}_1, \cdots, \hat{m}_t, \cdots, \hat{m}_T]^\top$ を各時間フレームにおいて独立に決定する．

$$\hat{m}_t = \underset{m}{\operatorname{argmax}}\, P(m|\boldsymbol{X}_t, \boldsymbol{\lambda})$$
$$= \underset{m}{\operatorname{argmax}}\, \mathcal{N}(\boldsymbol{X}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_{\boldsymbol{m}}^{(\boldsymbol{XY})}) \tag{7}$$

その後，式（6）の $P(Y|X, \lambda)$ の最大化処理に対し，カルマンフィルタによる近似を導入することで，各フレームにおいて L フレーム前の変換静的特徴量を決定する短遅延変換処理 [13] を実現する．また，式（6）の $P(v^{(y)}|\lambda^{(v)})$ の最大化処理に対しては，GV に基づくポストフィルタ処理 [12] により近似する．これらの近似処理により，50〜70 msec 程度の遅延時間でリアルタイム変換処理が可能となる．

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音源信号の 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 3 の左図に示す．

本システムは，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構成される．前者は従来の電気音声発声法における生成過程と違いはない．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，従来の F_0 制御法 [1] [2] [3] とは異なり，意識的な F_0 制御を必要とせず，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．

提案システムで生じ得る問題点として，1）調音動作に対する F_0 パターンの遅延の影響と，2）F_0 変動を伴う電気音声から抽出されるスペクトル特徴量の影響による F_0 予測精度の劣化が挙げられる．前者の問題点は，リアルタイム変換処理において過去の F_0 値が予測されるため，必ず現在の調音動作に対

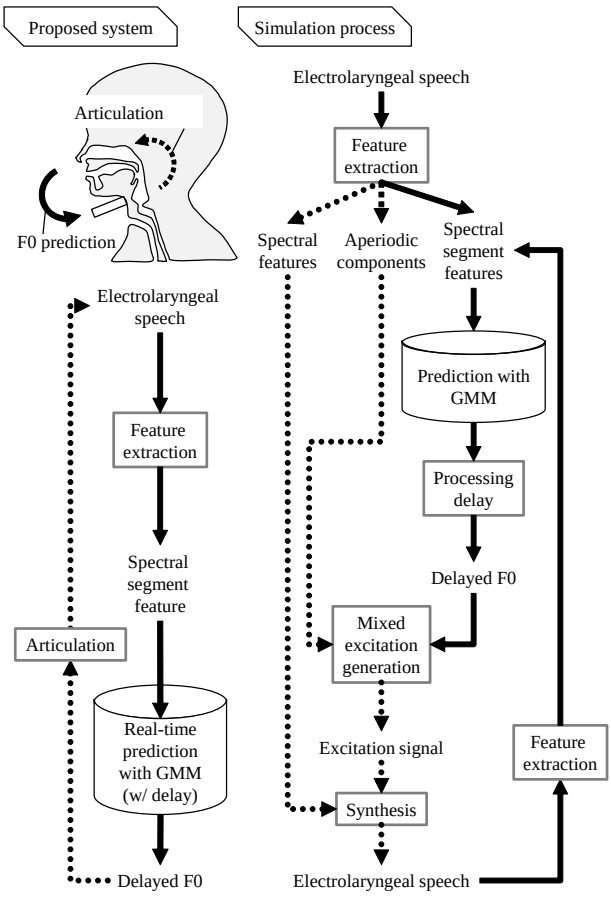

図 3　The proposed system and its simulation implementation.

して遅延が生じる．従来のハイブリッド電気音声強調法のように，ボコーダによる音声合成処理を用いて強調音声を生成する際には，音源信号にあわせてスペクトル特徴量も遅延させることで同期をとることが可能であるが，提案システムでは喉頭摘出者の調音により強調音声が生成されるため，同期をとることは不可能となる．本稿では，この遅延が強調音声に与える影響に関して，主観評価実験により調査する．

一方で，後者の問題点は，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ために生じる．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 分析が用いられるため，F_0 に応じた調波成分の影響を受けやすい．その結果，学習データとの不一致が生じ，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 [14] 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 適応型の分析処理を行うことで，スペクトル特徴量に対して音源信号の周期成分が与える影響を大幅に低減できる．通常は，F_0 抽出処理が必要となるが，提案システムでは予測 F_0 を直接用いることでそれを回避し，演算量を大幅に削減する．

一方，学習データ生成処理では，従来通り FFT 分析を使用するが，FFT 分析における F_0 の影響を考慮した GMM を構築する．まず，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する．その後，それらを全て同時に学習データとして使用することで，様々な F_0 の影響を受けたスペクトル特徴量に対応した GMM を学習する．

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 3 の右図に示す．

事前段階として，提案システムにおける調音動作に相当するスペクトル特徴量を STRAIGHT 分析を用いて電気音声から抽出する．また，提案システムにおける電気式人工喉頭の生成する音源信号を仮想的に生成するために，非周期成分 [15] に関しても事前に抽出しておく．これらは，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．これらの抽出パラメータと所望の F_0 パターンを用いてボコーダによる波形合成を行うことで，その F_0 パターンを用いて電気式人工喉頭の音源を制御した際に得られる電気音声を仮想的に生成する．

提案システムで得られる電気音声を模擬するために，以下の処理を行う．まず，1）電気音声からスペクトル特徴量を分析し，スペクトルセグメント特徴量を抽出したのち，オフライン統計的音源予測に基づいて F_0 パターンを予測する．2）予測された F_0 パターンは，調音動作に同期しているため，リアルタイム予測処理による遅延時間を考慮し，F_0 を遅延させる．3）得られた遅延 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 [16] により，音源信号を生成する．4）生成された音源信号に対して，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．5）前の時間の予測 F_0 が次の時間の予測 F_0 に与える影響を考慮するため，生成された電気音声を新たな入力とし，F_0 予測結果が安定するまで 2〜6 の処理を反復的に繰り返し，提案システムにより生成可能な電気音声を仮想的に生成する．

4.　実験的評価

4.1　実 験 条 件

入力音声として喉頭摘出者男性 2 名の電気音声と健常者男性 1 名が模擬的に発声した電気音声を用いる．目標音声には，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 文を学習データとして用いる．なお，その際の 100 Hz 電気音声は収録されたものを，その他の 150，200，250 Hz 電気音声は分析再合成後の電気音声とする．学習データ中の F_0 パターンにおけるマイクロプロソディ除去処理には，カットオフ周波数は 10 Hz の低域通過フィルタを用いる．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．GMM 学習における入力音声と目標音声のアライメント情報は，スペクトル特徴量に基づいて行

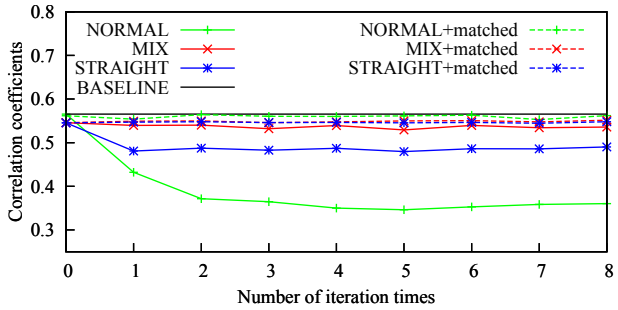

図 4　Prediction accuracy for F_0 correlation coefficient of laryngectomee A.

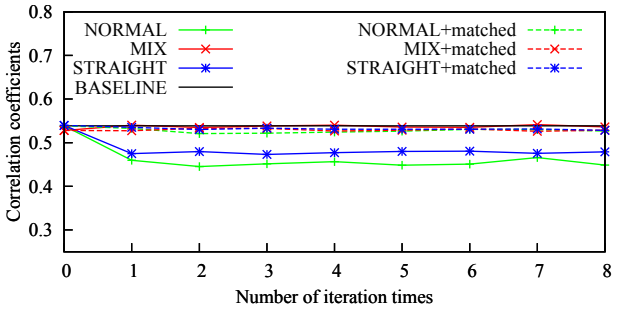

図 5　Prediction accuracy for F_0 correlation coefficient of laryngectomee B.

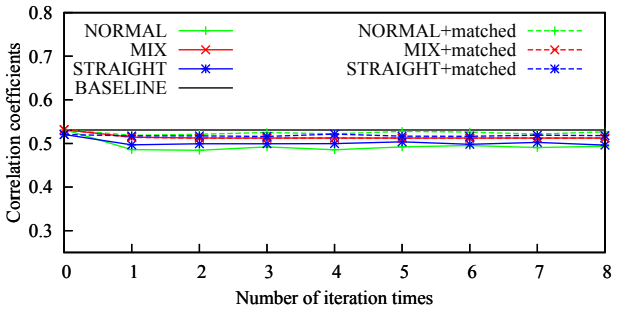

図 6　Prediction accuracy for F_0 correlation coefficient of non-disabled person.

うため [10]，システムに関係なく，話者対ごとに同じものを用いる．

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 予測精度を評価する．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により評価する．本研究における聞き取りやすさとは発話内容が理解できるかどうかという指標を表し，書き取り評価による明瞭性とは異なる．評価する音声は以下の 4 つである．

- EL: 元の電気音声
- BASELINE: オフライン統計的音源予測に基づく予測 F_0 を用いた音源信号に電気音声のスペクトル特徴量を畳み込んだ遅延なしの強調音声（[9] 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際に，反復シミュレーションにより仮想的に生成された提案法による強調音声
- STRAIGHT: STRAIGHT 分析を用いた際に，反復シミュレーションにより仮想的に生成された提案法による強調音声

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 分析を行った際の反復シミュレーションにより仮想的に生成された強調音声（NORMAL）も評価する．また，統計的音源予測処理において，予測 F_0 がスペクトル特徴量に与える影響を調査するため，平均値が学習データのもの（100 Hz）と一致するようにシフト処理を施した際も併せて評価する（matched）．一方で，主観評価においては，予測された F_0 パターンを，平均値が男性の平均的な F_0 値（125 Hz）に合うようにシフトさせた後に，強調音声を合成する．GMM の混合数は，32（BASELINE，NORMAL，MIX）および 16（STRAIGHT）とする．

4.2　実 験 結 果

図 4 に喉頭摘出者 A に対する 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，どのシステムにおいても，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，予測 F_0 をシフトしない際は，学習データとの不一致が大きくなり，FFT 分析使用時（NORMAL）の推定精度は大きく劣化する．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．なお，STRAIGHT 分析よりも学習データ生成処理の方が精度が高い傾向が見られる．特に，反復 1 回目において，STRAIGHT 分析では精度が劣化するのに対し，学習データ生成処理ではその劣化が見られない．反復 0 回目では実際に収録された電気音声が分析対象となるが，反復 1 回目からは分析合成音声が分析対象となることや，学習データ処理では分析合成音声が学習データに含まれていることを考えると，この劣化は分析合成処理に起因する可能性がある．

図 5 に喉頭摘出者 B に対する結果を，図 6 に健常者に対する結果を示す．全体的な傾向は，喉頭摘出者 A と類似しているが，そもそも FFT 分析（NORMAL）による劣化が小さいことから，予測 F_0 が予測精度に与える影響には個人差があることが分かる．

図 7 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，他の手法では大幅な改善がみられることから，文献 [9] で報告されている通り，統計的音源予測に基づく発話内容に応じた F_0 パターン予測は自然性の改善に非常に有効であることがわかる．また，話者ごとに同様の結果が得られており，かつ，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．また，BASELINE と比較し，提案法（MIX と STRAIGHT）では，調音動作に対して F_0 パターンが遅延するが，70 ms 程度の遅延であれば自然性に大きな影響を与えないことが分かる．

図 8 に聞き取りやすさに関する主観評価結果を示す．自然性に関する結果と異なり，個人差が大きいことが分かる．喉頭摘出者 A は，電気式人工喉頭を用いた発声に熟練しており，電気音声（EL）の聞き取りやすさが高い．この場合，全システムの間で有意差は見られず，MIX および STRAIGHT は電気音

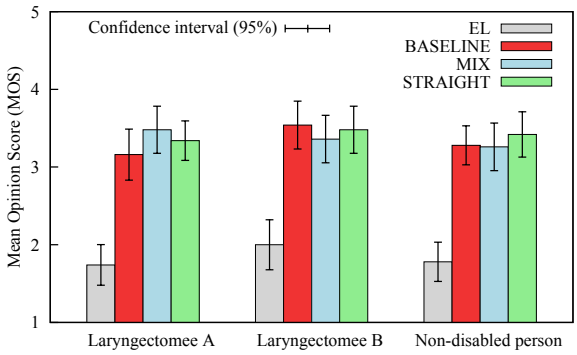

図 7　Result of opinion test on naturalness.

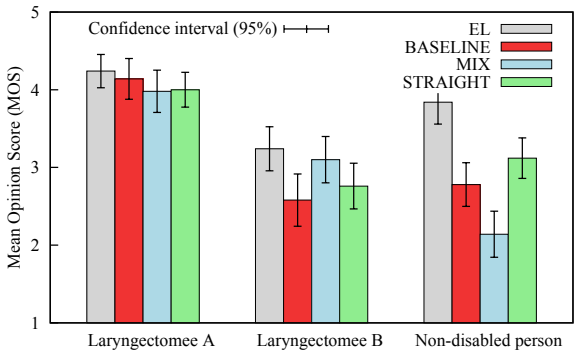

図 8　Result of opinion test on listenability.

声（EL）と同程度の聞き取りやすさを保持できることが分かる．一方で，喉頭摘出者 B や健常者に関してはシステム間で差が生じており，BASELINE が電気音声（EL）よりも劣化していることがわかる．なお，喉頭摘出者 B の学習データ生成処理（MIX）では，喉頭摘出者 A と同様に電気音声と有意差が生じていない．また，健常者に関しては，STRAIGHT 分析（STRAIGHT）により，BASELINE と同等以上の精度が得られている．一貫した結果が見られない原因として，電気式人工喉頭を用いた発声の熟練度の影響や，他の要因が結果に影響を与えている可能性が考えられる．そのため，今後，個々の評価音声を詳細に調べ，劣化を引き起こす要因を分析する必要がある．また，書き取り評価による明瞭度調査も必要である．

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法を提案した．また，電気式人工喉頭への実装を行う前段階として，提案法において生じる調音動作に対する F_0 パターンの遅延の影響と，F_0 変動を伴う電気音声が F_0 予測精度に与える影響を，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 [9] と同等の自然性を有することが分かった．一方で，聞き取りやすさに関しては個人差が見られた．今後は，書き取りやすさの主観評価結果に対する詳細な分析，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，提案法の電気式人工喉頭への実装を行う．

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．

文　献

[1] N. Uemi, T. Ifukube, M. Takahashi, and J. Matsushima, “Design of a new electrolarynx having a pitch control function,” Proc. 3rd IEEE International Workshop of Robot and Human Communication, pp.198–203, Jul. 1994.

[2] Y. Kikuchi and H. Kasuya, “Development and evaluation of pitch adjustable electrolarynx,” Proc. Speech Prosody 2004, International Conference., pp.761–764, Mar. 2004.

[3] K. Matsui, K. Kimura, Y. Nakatoh, and Y.O. Kato, “Development of electrolarynx with hands-free prosody control,” Proc. SSW8, pp.273–277, Aug. 2013.

[4] H. Liu, Q. Zhao, M. Wan, and S. Wang, “Enhancement of electrolarynx speech based on auditory masking,” Biomedical Engineering, IEEE Transactions on, vol.53, no.5, pp.865–874, May. 2006.

[5] Y. Stylianou, O. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” Speech and Audio Processing, IEEE Transactions on, vol.6, no.2, pp.131–142, Mar. 1998.

[6] T. Toda, A.W. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” Audio, Speech, and Language Processing, IEEE Transactions on, vol.15, no.8, pp.2222–2235, Nov. 2007.

[7] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech,” Speech Communication, vol.54, pp.134–146, Jan. 2012.

[8] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, “Alaryngeal speech enhancement based on one-to-many eigenvoice conversion,” Audio, Speech, and Language Processing, IEEE/ACM Transactions on, vol.22, no.1, pp.172–183, Jan. 2014.

[9] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, “An evaluation of excitation feature prediction in a hybrid approach to electrolaryngeal speech enhancement,” Proc. ICASSP, pp.4521–4525, May. 2014.

[10] T. Toda, M. Nakagiri, and K. Shikano, “Statistical voice conversion techniques for body-conducted unvoiced speech enhancement,” Audio, Speech, and Language Processing, IEEE Transactions on, vol.20, no.9, pp.2505–2517, Nov. 2012.

[11] A. Kain and M.W. Macon, “Spectral voice conversion for text-to-speech synthesis,” Proc. ICASSP, vol.1, pp.285–288, May. 1998.

[12] T. Toda, T. Muramatsu, and H. Banno, “Implementation of computationally efficient real-time voice conversion.,” Proc. INTERSPEECH, Sep. 2012.

[13] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” Proc. INTERSPEECH, pp.1076–1079, Sep. 2008.

[14] H. Kawahara, I. Masuda-Katsuse, and A. deCheveigné, “Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol.27, pp.187–207, Elsevier, Apr. 1999.

[15] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT,” Proc. MAVEBA, pp.13–15, Sep. 2001.

[16] 大谷大和，戸田智基，猿渡洋，鹿野清宏，“STRAIGHT 混合励振源を用いた混合正規分布モデルに基づく最ゆう声質変換法，” 電子情報通信学会論文誌，vol.J91-D, no.4, pp.1082–1091, Jan. 2008.