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.

# リアルタイム音源予測に基づく電気式人工喉頭制御の実装

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**　喉頭全摘出者のための代用発声法の一つに，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，我々は，電気音声から $F_0$ パターンをリアルタイムに予測する統計的音源予測処理を用いた電気式人工喉頭の音源制御を提案しており，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，本提案法を実装した実機システムを構築し，その性能を評価する．また，リアルタイム音源予測精度改善のため，連続 $F_0$ パターンのセグメント化学習を検討する．実験結果より，1）実機システムにおいても，従来の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，2）シミュレーション実験は高い精度で実機システムの性能を模擬可能であること，3）リアルタイム音源予測において，連続 $F_0$ パターンのセグメント化学習を行うことで，変換処理遅延を抑えつつ予測精度を改善できることを示す．
**キーワード**　電気式人工喉頭，電気音声，$F_0$ 制御，リアルタイム統計的音源予測，遅延

# Implementation of Direct $F_0$ Control of an Electrolarynx based on Real-time Excitation 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 speaking aid device to artificially generate excitation sounds to help laryngectomees produce electrolaryngeal (EL) speech. Although EL speech is quite intelligible, its naturalness significantly suffers from unnatural $F_0$ pattern of the mechanical excitation sounds. To make it possible to produce more naturally sounding EL speech, we have proposed a method to automatically control the $F_0$ patterns of the excitation sounds generated from the electrolarynx based on the statistical $F_0$ prediction, which predicts the $F_0$ pattern pattern of natural speech from the produced EL speech in real-time. In our previous work, we have confirmed its effectiveness by an experimental evaluation through simulation. In this report, we develop a prototype system by implementing the proposed control method in an actual, physical electrolarynx and evaluate its performance. We also propose a training method using segmented continuous $F_0$ patterns to improve accuracy of real-time statistical $F_0$ prediction. The experimental results demonstrate that 1) the prototype system is capable of generating more naturally sounding EL speech, 2) its prediction accuracy is well simulated by our previously proposed simulation method, and 3) the use of segmented continuous $F_0$ patterns is effective for improving prediction accuracy while not increasing prediction processing delay.
**Key words**　electrolarynx, electrolaryngeal speech, $F_0$ control, real-time statistical $F_0$ prediction, processing delay

## 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，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．

この問題に対処するため，電気式人工喉頭が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，より自然性の高い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．そのため，自然性改善効果は限定される．

より自然な $F_0$ パターンを電気音声に付与するために，我々は，リアルタイム統計的音源予測 [4] [5] [6] を用いた電気式人工喉頭の音源制御法 [7] を提案している．図 1 に示すこの枠組みでは，話者による意図的な $F_0$ パターン操作を必要とせず，従来の発声行為のみを必要とし，生成された電気音声のみから $F_0$ パターンが予測される．統計的声質変換 [8] [9] の枠組みを応用することで，電気音声と通常音声の同一文発話対（パラレルデータ）から事前に得られる統計量に基づいて，発話内容に沿った $F_0$ パターンの予測が可能となる．これまでに，シミュレーション実験による性能評価の結果から，提案法により，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．

本稿では，本提案法を実装した実機システムを構築し，その性能を評価する．先に提案したシミュレーションによる評価法の精度を明らかにするとともに，リアルタイム音源予測精度のさらなる改善に向けて，連続 $F_0$ パターンのセグメント化学習も検討する．実験結果より，1）実機システムにおいても，従来の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，2）シミュレーション実験は高い精度で実機システムの性能を模擬可能であること，3）リアルタイム音源予測において，連続 $F_0$ パターンのセグメント化学習を行うことで，変換処理遅延を抑えつつ予測精度を改善できることを示す．

## 2. 統計的音源予測

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 $F_0$ パターン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 $F_0$ パターンを求める．

### 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，通常音声の $F_0$ 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 $\Delta y_t$ は $\Delta y_t = y_t - y_{t-1}$ により計算する．

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].

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

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 $\alpha_m$，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．

### 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 [9] により，電気音声のスペクトル特徴量系列から通常音声の $F_0$ パターンへと変換する．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．

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) \quad \text{subject to}\, \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{3}$$

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を表す．リアルタイム音源予測処理 [12] では，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．

$$\hat{m}_t = \underset{m}{\operatorname{argmax}}\, P(m|\boldsymbol{X}_t, \boldsymbol{\lambda})$$

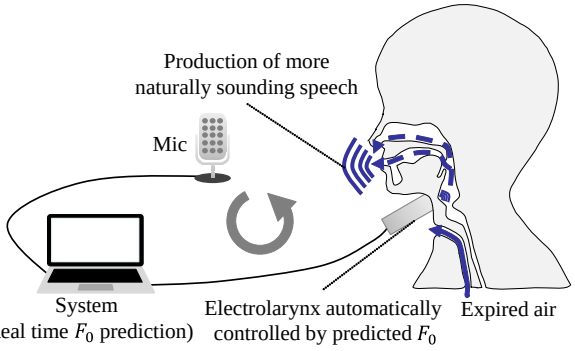


図 1 Proposed system to directly control electrolarynx using real time statistical $F_0$ prediction for laryngectomees.

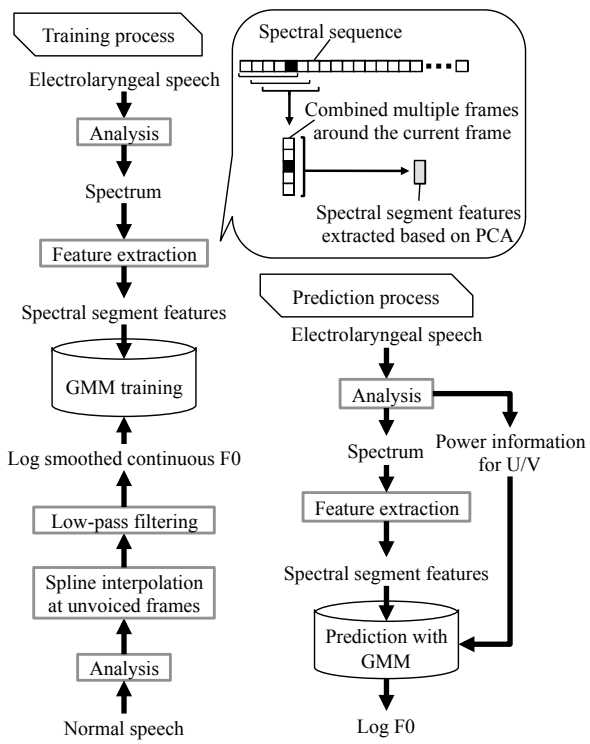


図 2　The training and prediction process.

$$= \underset{m}{\operatorname{argmax}} \mathcal{N}(\boldsymbol{X}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_{\boldsymbol{m}}^{(\boldsymbol{XY})}) \tag{4}$$

その後，式（3）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．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 [7]．本手法の処理過程を図 3 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 [1] [2] [3] 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．

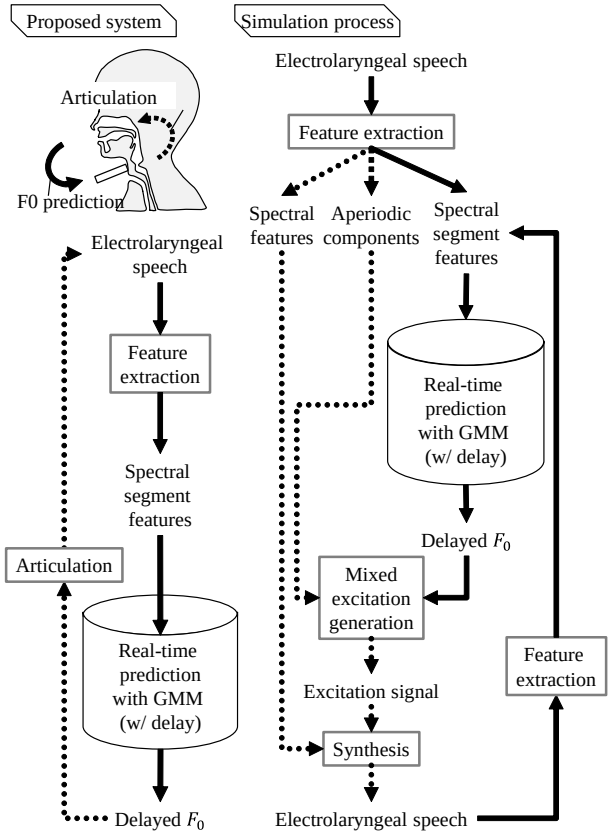


図 3　The proposed system and its simulation implementation.

### 3.2　シミュレーション

本システムのシミュレーション処理を図 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 [14] を用いて電気音声から抽出する．また，電気式人工喉頭が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，リアルタイム統計的音源予測に基づいて $F_0$ パターン（遅延あり）を予測する．2) 得られた予測 $F_0$ と事前に抽出しておいた非周期成分を用いて，混合励振源モデルにより，音源信号を生成する．3）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．4）予測 $F_0$ が入力特徴量さらには変換精度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 2〜4 の処理を反復的に繰り返し，提案システムにより生成される電気音声を仮想的に生成する．

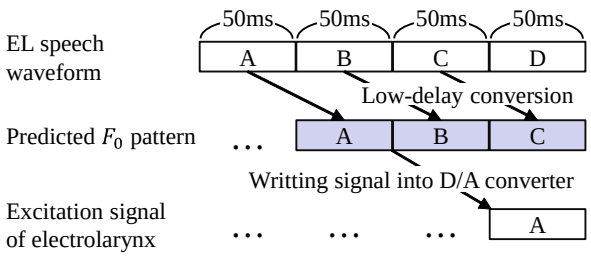


図 4　Latency caused by each process

## 4. 実機システムの構築

### 4.1 実機への実装

表 1 に示される PC と D/A 変換器を用いて，提案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，リアルタイム音源予測を用いて，入力される電気音声から連続 $F_0$ パターンを予測する．その後，電気式人工喉頭の生成する音源信号の $F_0$ パターンを制御するため，予測 $F_0$ 値に線形変換を施すことで，電気式人工喉頭へと入力する電圧を決定する．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．最終的に，予測 $F_0$ パターンに基づき電気式人工喉頭が生成する音源信号を，喉頭摘出者自身が調音することで，$F_0$ パターンの付与された電気音声が発声される．

前章で述べた通り，リアルタイム音源予測処理において生じる遅延の影響を受けて，$F_0$ パターンは調音動作に対して遅延する．さらに，本稿における実機システムでは，D/A 変換器において生じる遅延の影響も受ける．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，D/A 変換処理部では約 100 ms の遅延が生じる．結果，図 4 に示す通り，本実装では，リアルタイム音源予測処理において生じる遅延時間 50 ms と D/A 変換器で生じる遅延時間 100 ms の合計 150 ms 程度の遅延が生じる．なお，リアルタイム音源予測処理を実装した DSP [16] を用いて全処理系統を電気式人工喉頭へ組み込む際には，本実装で用いた D/A 変換器を使用する必要はなく，総遅延時間を 50 ms 程度にまで減らせると予想される．

本提案法の実機システムでは，予測される $F_0$ パターンは，予測された $F_0$ パターン自身を使用した電気音声により予測される．そのため，遅延時間の変更や GMM の混合数変更といったシステムの条件設定を変更し，$F_0$ 推定精度を比較評価するためには，個々の条件設定の下で実際に実機システムを使用し，電気音声を別途収録する必要がある．一方で，シミュレーションシステムにおいては，収録済みの電気音声を用いて，様々な条件設定下で実機システムを使用した際の電気音声を仮想的に生成することができる．実機システムの性能を容易に評価するために，高精度なシミュレーションシステムの実現が重要となる．

表 1　Electronic devices on the prototype system

| | |
|---|---|
| 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 |

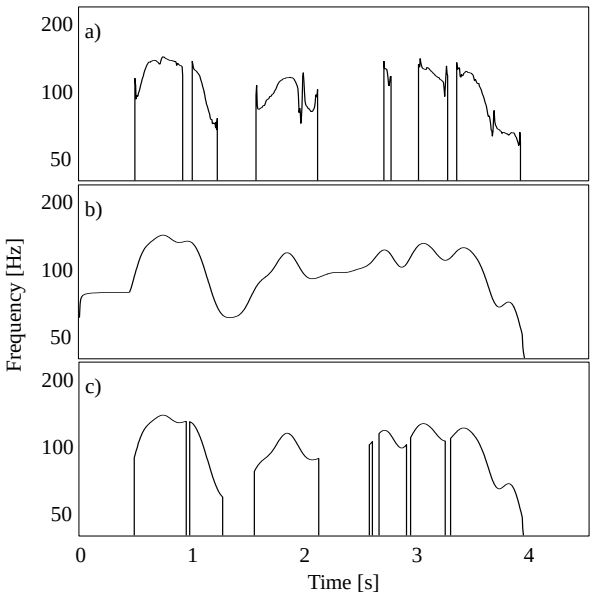


図 5　Example of $F_0$ patterns a) of target natural speech, b) interpolated at unvoiced frame and removed micro prosody from (a), and c) extracted at speech section from (b).

### 4.2 連続 $F_0$ パターンのセグメント化学習

統計的音源予測では，連続 $F_0$ パターンを用いた学習が行われる [17]．図 5 に，$F_0$ パターンの例を示す．図 5(b) に示される連続 $F_0$ パターンは，図 5(a) に示される自然音声から抽出された $F_0$ パターンに対して，無声音および非発話区間における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することで得られる．式 (3) で示される音源予測処理は，有声区間単位で動作するため，連続 $F_0$ パターンを用いることで，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り，予測精度の改善がもたらされる．一方で，リアルタイム音源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ため，高い変換精度を得るためにはより長い遅延フレーム数を要する傾向がある．

本稿では，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．電気音声の波形パワーを用いて発話区間を自動抽出し，無音区間と判定されたフレームについては無声とすることで，連続 $F_0$ パターンをセグメント化する．セグメント化連続 $F_0$ パターン（Segmented C$F_0$）を図 5(c) に示す．学習時にセグメント化連続 $F_0$ パターンを用いることで，フレーム間相関を考慮する単位を限定し，カルマンフィルタによる近似の影響を抑えることで，必要となる遅延フレーム数の低減を図る．

## 5. 実験的評価

### 5.1 実 験 条 件

男性健常者 1 名による従来の電気式人工喉頭使用時の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

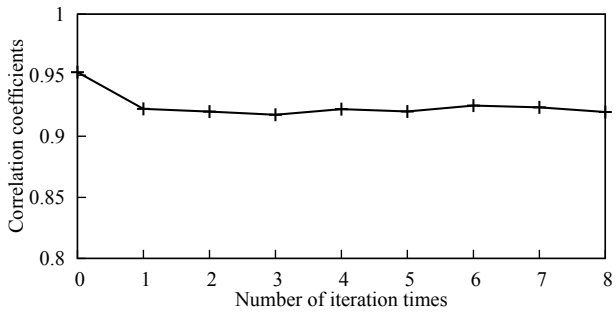


図 6　$F_0$ correlation coefficients between $F_0$ patterns predicted by the prototype system and those predicted in each iteration step by the simulated system.

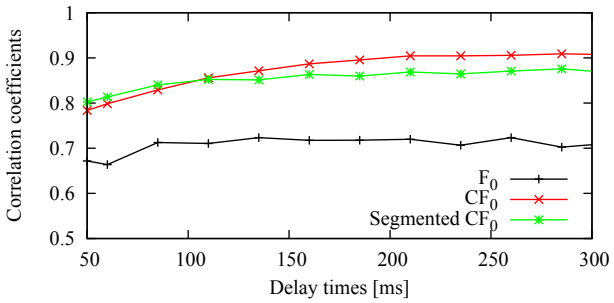


図 7　Effects of delay times on real-time $F_0$ prediction accuracy in the simulated system. Correlation coefficients are calculated between $F_0$ patterns predicted by real-time processing and those predicted by batch-type processing.

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．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，カットオフ周波数が 10 Hz の低域通過フィルタを用いる．実機システムにおける総合遅延時間は 150 ms である．

### 5.2　実装の妥当性評価

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 3 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．なお，各 $F_0$ パターンにおけるリアルタイム音源予測の上限値となるオフライン音源予測精度を表 2 に示す．

実機システム使用時の予測 $F_0$ パターンと，シミュレーションシステム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 6 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，提案法は実機上で動作可能であること，また，シミュレーション精度は十分に高いことが分かる．

### 5.3　リアルタイム音源予測における遅延時間と予測精度

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 [13]，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，シミュレーション実験により、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．

表 2　$F_0$ correlation coefficients between each $F_0$ patterns extracted from target speech and those predicted by the batch-type processing.

| | $F_0$ | C$F_0$ | Segmented C$F_0$ |
|---|---|---|---|
| The number of mixture components | 32 | 16 | 16 |
| $F_0$ correlation coefficients | 0.40 | 0.48 | 0.46 |

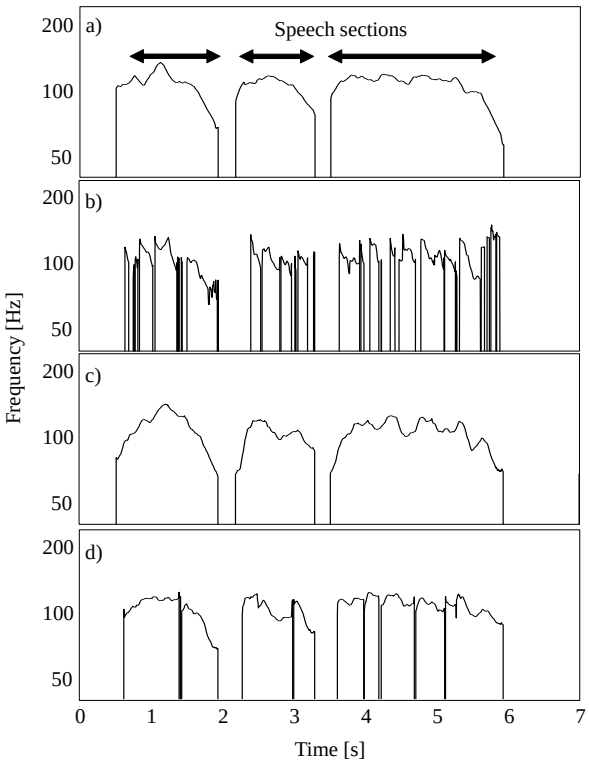


図 8　Example of $F_0$ patterns a) predicted by batch-type processing, b) predicted by $F_0$ model, c) predicted by C$F_0$ model, and d) predicted by Segment C$F_0$ model. The processing delay is set to 100 ms in real-time processing.

オフライン予測処理で予測される $F_0$ パターンと，シミュレーションシステムにおけるオンライン予測処理の遅延フレーム数（フレームシフト長は 5 ms）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 7 に示す．また，各種 $F_0$ パターンを用いて学習された変換モデルにより予測された $F_0$ パターンを図 8 に示す．連続 $F_0$ パターン予測において，短遅延変換処理における近似の影響を低減するためには，200 ms 程度（40 フレーム相当）遅延させる必要があることが

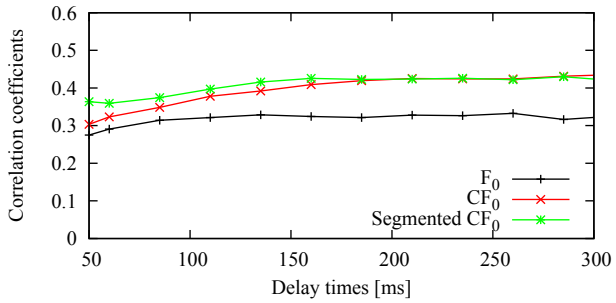


図 9　$F_0$ correlation coefficient between real-time prediction and target as a function of the delay times.

分かる．スペクトル予測処理では 15〜25 ms 遅延（3〜5 フレーム相当）で十分な結果が得られること [13] 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．一方で，セグメント化連続 $F_0$ パターンでは，連続 $F_0$ パターンよりも，必要となる遅延フレーム数が減少し，100 ms 程度（20 フレーム相当）の遅延で相関係数が概ね収束する．通常の $F_0$ パターンを用いた際には，必要な遅延フレーム数がさらに減少するが，相関係数も低下する．

各手法のリアルタイム予測精度を評価するために，目標話者の $F_0$ パターンと，シミュレーションシステムにおけるオンライン予測処理の遅延時間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 9 に示す．通常の $F_0$ パターンを学習データとして用いた場合は，相関が低いことが分かる．一方で，連続 $F_0$ にすることで，予測精度は改善するが，遅延フレーム数が少ない場合には改善効果が低下する．これに対し，セグメント化連続 $F_0$ を用いることで，遅延フレーム数の減少に伴う改善効果の低下を抑えることができる．

## 6. おわりに

本稿では，リアルタイム音源予測に基づく電気式人工喉頭制御法を実装した実機システムを構築し，その性能を評価した．また，リアルタイム音源予測精度改善のため，連続 $F_0$ パターンのセグメント化学習を検討した．実験結果より，1）実機システムにおいても，従来の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，2）シミュレーション実験は高い精度で実機システムの性能を模擬可能であること，3）リアルタイム音源予測において，連続 $F_0$ パターンのセグメント化学習を行うことで，変換処理遅延を抑えつつ予測精度を改善できることを示した．

**謝辞**：本研究の一部は，JSPS 科研費 26280060 および 15J10727 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．

### 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, July 1994.

[2] Y. Kikuchi and H. Kasuya, “Development and evaluation of pitch adjustable electrolarynx,” Proc. Speech Prosody 2004, International Conference., pp.761–764, March 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] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech,” Proc. Speech Communication, vol.54, pp.134–146, Jan. 2012.

[5] 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.

[6] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, “A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation,” IEICE Transactions on Information and Systems, vol.E97-D, no.6, pp.1429–1437, June 2014.

[7] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, “Direct f0 control of an electrolarynx based on statistical excitation feature prediction and its evaluation through simulation,” Proc. INTERSPEECH, pp.31–35, Sept. 2014.

[8] 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, March 1998.

[9] 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.

[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, Sept. 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, Sept. 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 F0 extraction: Possible role of a repetitive structure in sounds,” Proc. Speech Communication, vol.27, pp.187–207, April 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, Sept. 2001.

[16] T. Moriguchi, T. Toda, M. Sano, H. Sato, G. Neubig, S. Sakti, and S. Nakamura, “A digital signal processor implementation of silent/electrolaryngeal speech enhancement based on real-time statistical voice conversion,” Proc. INTERSPEECH, pp.3072–3076, Aug. 2013.

[17] 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.