

スペクトル補正及び統計的音源生成に基づくハイブリッド電気音声強調

田中 宏[†] 戸田 智基[†] グラム・ニュービグ[†] 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, 生駒市

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）を生成できるという利点がある。一方で、自然な音源信号を機械的に生成するのは困難であり、特に発話内容に応じた自然な基本周波数パターンを生成するのは本質的に極めて困難な処理となる。結果として、電気音声の自然性は大きく劣化する。また、電気式人工喉頭から生成される音源信号自体が外部に漏れるため、雑音として電気音声に混入し、その品質を劣化させる。これらの問題に対処するため、本稿では、電気音声の聞き取りやすさを保持しながら自然性を大幅に改善する音声強調法として、従来の強調法である雑音抑圧に基づくスペクトル補正処理と統計的声質変換に基づく音源特徴量生成処理を組み合わせたハイブリッド法を提案する。提案法では、統計的手法によるスペクトルおよび有声無声情報への変換処理を回避することで、変換誤差により聞き取りやすさが劣化する事態を回避する。また、統計的手法により、通常音声の基本周波数パターンを予測することで、自然性を大幅に改善する。実験的評価結果から、本手法の有効性を示す。

キーワード 発声障害, 電気音声, スペクトル減算, 統計的声質変換, ハイブリッド法

A Hybrid Approach to Electrolaryngeal Speech Enhancement Based on Spectral Compensation and Statistical Source Excitation Generation

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 We present a hybrid approach to improving naturalness of electrolaryngeal (EL) speech while minimizing degradation in listenability. An electrolarynx is a device that artificially generates excitation sounds to enable laryngectomees to produce EL speech. Although proficient laryngectomees can produce quite intelligible EL speech, it sounds very unnatural due to the mechanical excitation produced by the device. Moreover, the excitation sounds produced by the device often leak outside, adding noise to EL speech. To address these issues, we propose a hybrid method using the noise reduction method for enhancing spectral parameters and voice conversion method for predicting excitation parameters. The experimental results demonstrate the proposed method yields significant improvements in naturalness compared with EL speech while keeping listenability high enough.

Key words speaking-aid, electrolaryngeal speech, spectral subtraction, voice conversion, hybrid approach

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。

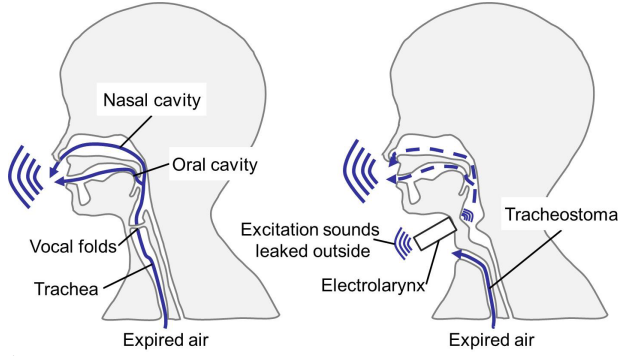


図 1 Speech production mechanisms of non-disabled people (left figure) and total laryngectomees (right figure).

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が生産される。本稿では、この代用発声法で生成される音声を電気音声と呼ぶ。電気式人工喉頭を用いた発声法は、1) 修得が容易である、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較し、比較的聞き取りやすい音声を生成できる、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 つ目の欠点に関しては、発話内容に沿った自然な基本周波数パターンを持つ音源信号を外部から機械的に生成する必要がある、本質的に極めて困難な処理となる。

これらの問題に対処するため、従来の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法 [1] や、統計的声質変換 [2] [3] に基づく通常音声への変換法 [4] [5] などが提案されている。前者は、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ことができるが、機械的な音源信号に起因する電気音声の音響的特徴は変化しないため、自然性改善効果は限定される。一方、後者は、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スペクトル特徴量および音源特徴量への変換を行う。フォルマント操作 [6] などの他の信号処理に基づく手法と比べて、複雑な変換処理が可能となり、より自然音声に近いスペクトル特徴量が得られる。また、発話内容に沿ったより自然な基本周波数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。その結果、電気音声の自然性は著しく向上する。一方で、少なからず変換誤差が生じるため、聞き取りやすさは若干劣化する傾向がある [5]。

本研究では、聞き取りやすさの劣化を引き起こさずに自然性を大幅に改善す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処理と統計的声質変換に基づく音源特徴量生成処理を組み合わせたハイブリッド強調法を提案する。提案法では、変換誤差の影響を減らすために、統計的声質変換により変換する音響特徴量を限定する。具体的には、音源特徴量については統計的声質変換により推定し、スペクトル特徴量については雑音抑圧に基づく補正処理を用いる。さらに、変換誤差の影響を

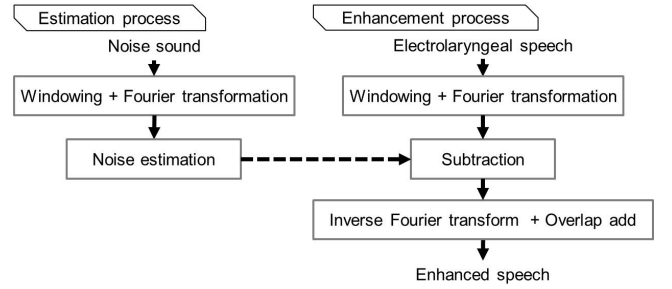


図 2 EL speech enhancement based on SS.

低減させるため、連続基本周波数パターン推定処理を導入する。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、電気音声の聞き取りやすさを保ちながら、自然性を大幅に改善できることを示す。

2. 電気音声に対する雑音抑圧に基づくスペクトル補正法

電気音声に対して、雑音として混入する空気中に漏れ出した電気式人工喉頭の音源信号自体を除去するために、適応フィルタを用いた雑音抑圧 [7] やスペクトル減算処理 (spectral subtraction: SS) [8] を用いた雑音抑圧などが提案されている。ここでは、非常に簡潔なアルゴリズムながらも、高い雑音抑圧性能を発揮するスペクトル減算処理を用いた雑音抑圧に基づくスペクトル補正法 [1] について述べる。

時間 t における観測信号 $Y(t)$ は以下で記述される。

$$Y(t) = S(t) + L(t) \quad (1)$$

ここで、 S は音声信号、 L は雑音信号を表す。短時間離散フーリエ変換により得られる時間 t かつ周波数 ω の時間周波数表現 $Y(\omega, t)$ は、以下で記述される。

$$Y(\omega, t) = S(\omega, t) + L(\omega, t) \quad (2)$$

一般化スペクトル減算法 [9] では、雑音信号 L の定常性を仮定し、推定された雑音の振幅スペクトルの期待値を観測信号の振幅スペクトルから減算することにより、雑音が抑圧された信号の振幅スペクトル $|\hat{S}(\omega, t)|$ を次式にて求める。

$$|\hat{S}(\omega, t)|^\gamma = \begin{cases} |Y(\omega, t)|^\gamma - \alpha |\hat{L}(\omega)|^\gamma & \left(\frac{|\hat{L}(\omega)|^\gamma}{|Y(\omega, t)|^\gamma} < \frac{1}{\alpha + \beta} \right) \\ \beta |\hat{L}(\omega)|^\gamma & (\text{otherwise}) \end{cases} \quad (3)$$

ここで、 $\alpha (\alpha > 1)$ は過減算パラメータ、 $\beta (0 \leq \beta \leq 1)$ はフロアリング強度の制御パラメータ、 γ は指数パラメータで $\hat{L}(\omega)$ は推定された雑音の振幅スペクトルの期待値とする。通常、非発話区間の雑音の振幅スペクトル $|L(\omega, t)|$ を時間平均したものを、全区間での雑音振幅スペクトルのプロトタイプ $\hat{L}(\omega)$ として利用する。

電気音声強調においては、雑音信号は電気式人工喉頭から生成される音源信号であるため、事前にその振幅特性を求めておくことが可能である。本稿における強調処理過程を図 2 に示す。電気式人工喉頭を通常通り喉元に押し当てて音源信号を生じた際に、口元のマイクで観測される雑音信号を事前に収録する。その際に、電気音声自体が収録されないように口は閉じ

ておく．得られた雑音信号から振幅スペクトルの期待値を計算し，雑音振幅スペクトルのプロトタイプを求める．なお，過減算パラメータは $\alpha = 2$ とし，フロアリング強度の制御パラメータは $\beta = 0$ ，指数パラメータは $\gamma = 1$ とする．

3. 電気音声から通常音声への統計的手法に基づく声質変換法

統計的手法に基づく声質変換法 (statistical voice conversion: VC) に基づき，電気音声から通常音声へ変換することで，電気音声強調を行う [4]．本処理過程を図 3 に示す．学習部と変換部から構成されており，学習部では，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セグメント特徴量から通常音声のスペクトル特徴量および音源特徴量（基本周波数と非周期成分 [10]）への変換モデルが，各々事前に学習される．変換部では，得られた変換モデルを用いて，電気音声のスペクトルセグメント特徴量から，通常音声のスペクトル特徴量，基本周波数，非周期成分が推定される．推定された音響特徴量から，ボコーダにより強調音声信号を生成する．

3.1 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を $\mathbf{X}_t$ 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．

$$\mathbf{X}_t = \mathbf{E}[\mathbf{x}_{t-C}^\top, \dots, \mathbf{x}_t^\top, \dots, \mathbf{x}_{t+C}^\top]^\top + \mathbf{f} \quad (4)$$

ここで， $\mathbf{x}_t$ は時間フレーム t におけるスペクトル特徴量を表し，本研究ではメルケプストラムを用いる． $\mathbf{E}$ および $\mathbf{f}$ は各々変換行列およびバイアスペクトルを表し，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，通常音声の個々の音響特徴量として， $\mathbf{Y}_t = [\mathbf{y}_t^\top, \Delta \mathbf{y}_t^\top]^\top$ を使用する．ここで，動的特徴量 $\Delta \mathbf{y}$ は $\Delta \mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$ により計算する．スペクトル特徴量としてメルケプストラムを用い，音源特徴量として対数基本周波数および帯域別平均非周期成分 [11] を用いる．

パラレルデータに対して動的時間伸縮 (Dynamic Time Wrapping: DTW) を行い，入力特徴量 $\mathbf{X}_t$ と出力特徴量 $\mathbf{Y}_t$ の対応付けを行った結合ベクトル $[\mathbf{X}_t^\top, \mathbf{Y}_t^\top]^\top$ を用いて，次式に示すとおり，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 [12]．

$$P(\mathbf{X}_t, \mathbf{Y}_t | \boldsymbol{\lambda}) = \sum_{m=1}^M \alpha_m \mathcal{N}([\mathbf{X}_t^\top, \mathbf{Y}_t^\top]^\top; \boldsymbol{\mu}_m^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}) \quad (5)$$

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．

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \quad (6)$$

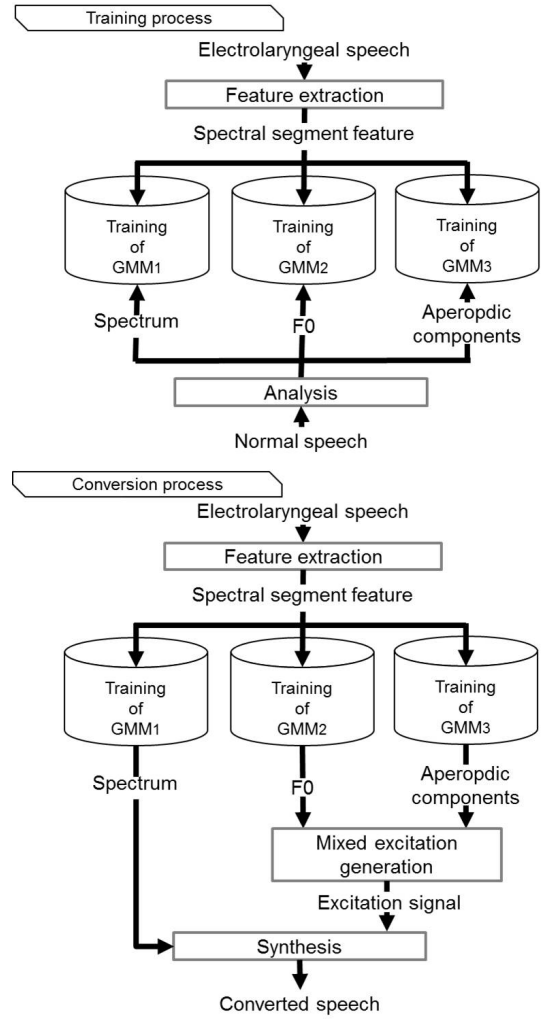


図 3 EL speech enhancement based on VC.

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．電気音声のスペクトルセグメント特徴量（メルケプストラムセグメント）と通常音声のメルケプストラム，対数基本周波数，帯域別平均非周期成分の間において，計 3 つの GMM を学習する．

また，通常音声の音響特徴量に対しては，系列内変動 (Global Variance: GV) [3] の確率密度関数も学習する．ここで，GV ベクトル $\mathbf{v}^{(y)} = [v_1^{(y)}, \dots, v_{D(y)}^{(y)}]^\top$ は，通常音声の静的特徴量系列に対して，各発話ごとに次式で計算される．

$$v_d^{(y)} = \frac{1}{T} \sum_{t=1}^T (y_{t,d} - \frac{1}{T} \sum_{\gamma=1}^T y_{\gamma,d}) \quad (7)$$

ここで， $y_{t,d}$ はフレーム t の通常音声の静的特徴量ベクトル $\mathbf{y}_t$ の d 次元目の要素である．GV の確率密度関数 $P(\mathbf{v}^{(y)} | \boldsymbol{\lambda}^{(v)})$ は，平均ベクトル $\boldsymbol{\mu}^{(v)}$ および共分散行列 $\boldsymbol{\Sigma}^{(vv)}$ の正規分布を用いて，以下のようにモデル化する．

$$P(\mathbf{v}^{(y)} | \boldsymbol{\lambda}^{(v)}) = \mathcal{N}(\mathbf{v}^{(y)}; \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)}) \quad (8)$$

3.2 変換処理

変換部では，個々の GMM を用いて，最尤系列変換

法[3]により、電気音声の入力特徴量系列から通常音声の各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個々の特徴量系列をそれぞれ $\mathbf{X} = [\mathbf{X}_1^\top, \dots, \mathbf{X}_t^\top, \dots, \mathbf{X}_T^\top]^\top$, $\mathbf{Y} = [\mathbf{Y}_1^\top, \dots, \mathbf{Y}_t^\top, \dots, \mathbf{Y}_T^\top]^\top$ とおく．このとき、変換後の静的特徴量系列 $\hat{\mathbf{y}} = [\hat{\mathbf{y}}_1^\top, \dots, \hat{\mathbf{y}}_t^\top, \dots, \hat{\mathbf{y}}_T^\top]^\top$ は次式で計算される．

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} P(\mathbf{Y}|\mathbf{X}, \lambda) P(\mathbf{v}^{(y)}|\lambda^{(v)})^\omega$$

subject to $\mathbf{Y} = \mathbf{W}\mathbf{y}$ (9)

ここで、 $\mathbf{W}$ は静的特徴量系列 $\mathbf{y}$ を静的・動的特徴量系列 $\mathbf{Y}$ に写像する変換行列を表す．また、 ω は GV の尤度重みを表す．なお、通常音声の対数基本周波数への変換処理の際には、有声／無声情報も同時に推定する．この処理は、体内伝導無声音から通常音声への統計的変換法[13]で行われるものと同一である．

変換後の基本周波数パターンおよび非周期成分を用いて混合励振源モデル[11]により音源信号を生成する．変換後のスペクトル特徴量を音源信号に畳み込むことで、強調音声を生成する．

4. スペクトル補正及び統計的音源生成に基づくハイブリッド電気音声強調法

雑音抑圧に基づくスペクトル補正法では、外部に漏れ出す音源信号の影響を低減できるが、仮に完全な雑音除去処理が行えたとしても、強調された電気音声は依然として機械的に生成された音源信号の影響を強く受ける．特に、人工的な基本周波数パターンの影響は大きく、自然性の著しい劣化は本質的に避けられない．結果として、補正処理による品質改善効果は限られる．一方で、本処理で生じる悪影響は極めて小さく、電気音声の比較的高い聞き取りやすさも維持される．

統計的声質変換に基づく電気音声強調処理では、通常音声の統計量に基づいて強調音声生成される．スペクトル特徴量のみでなく音源特徴量に対しても高度で複雑な変換処理が可能となるため、大幅な自然性改善効果が得られる．一方で、完全に自然な通常音声の音響特徴量へと変換するのは困難であり、少なからず変換誤差が生じる．結果として、強調音声の聞き取りやすさは電気音声と比較して若干劣化する傾向がある．

4.1 ハイブリッド電気音声強調法

聞き取りやすさの劣化を引き起こさずに自然性を大幅に改善す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処理と統計的声質変換に基づく音源特徴量生成処理を組み合わせたハイブリッド法を提案する．提案法の処理過程を図4に示す．

喉頭摘出者の調音器官は正常に機能する場合が多く、それ故に、比較的聞き取りやすい電気音声の生成が可能となる．すなわち、電気音声のスペクトル特徴量に関しては、生成過程の相違や音源信号の外部漏れの影響はあるものの、通常音声のスペクトル特徴量に比較的類似したものとなる．そこで、電気音声のスペクトル特徴量に関しては最大限に活用することを考え、雑音抑圧に基づくスペクトル補正処理により得られるスペクトル

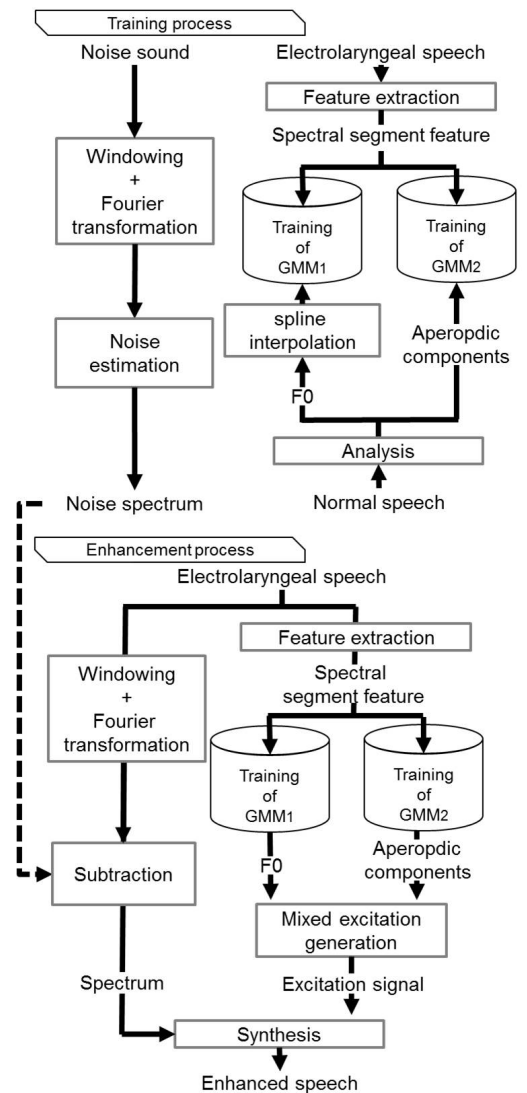


図4 EL speech enhancement based on a hybrid approach.

ル特徴量を用いる．統計的声質変換に基づく手法と比較すると、電気音声の持つ独特の機械的な声色は残るものの、変換誤差の影響を回避することができる．また、喉頭摘出者本人のスペクトル特徴量を使用するという利点もある．

一方で、電気音声の音源特徴量に関しては、完全に機械的に生成されたものであり、通常音声の音源特徴量とは大きく異なる．本研究で用いる電気式人工喉頭においても、基本周波数パターンは発声区間でほぼ一定であり、非周期成分は自然音声のものと大きく異なる特徴を持つ．そこで、音源特徴量である基本周波数および非周期成分に関しては、電気音声から通常音声への統計的声質変換法により推定されたものを用いる．通常音声から得られる統計量の使用により、より自然な音源特徴量を持つ強調音声を得られる．特に、基本周波数パターンに関しては、現状の統計的変換技術では十分に高い推定精度は得られないものの、元の電気音声の持つ人工的なものと比較すると、より自然な音に近いものが得られる．また、非周期成分に関しても、より自然なものが得られる．これにより、電気音声の自然性を大幅に改善することができる．

4.2 連続基本周波数パターンの導入

電気式人工喉頭を用いた代用発声では、発声中は電気式人工

喉頭のスイッチを ON にするため、生成される音韻は全て有声音となる。これに対して、統計的声質変換による音源特徴量推定では、有聲無聲情報を推定することで、無聲音の発声も可能とする。しかしながら、本推定処理は本質的に困難なものであるため、推定誤差が少なからず生じる。有聲無聲情報は音韻の識別に直接影響を与えるため、推定精度によっては逆に聞き取りやすさを下げる要因となり得る。

有聲無聲情報の推定誤差による聞き取りやすさの劣化を回避するために、無声区間を持たない連続的な基本周波数パターンの推定処理を導入する。元の電気音声と同様に、強調音声の音韻も全て有声音となるため、有聲音を無聲音とする推定誤差は生じない。学習処理では、通常音声の基本周波数パターンに対して、スプライン補間処理を施すことにより得られる連続的な基本周波数パターンを用いて、変換用 GMM を学習する。変換処理では、全フレームにわたって基本周波数パターンを推定する。ただし、電気式人工喉頭のスイッチの ON・OFF に関しては、電気音声の波形パワー等を用いることで容易に検出できるため、スイッチが OFF の区間（無音区間）に関しては無声フレームに置き換える。なお、非周期成分に関しては、統計的声質変換による推定処理を用いる。

5. 実験的評価

5.1 実験条件

喉頭摘出者 1 名の電気音声と、健康者 1 名の通常音声を用いる。ATR 音素バランス文セット中の 50 文を用い、40 文を学習データ、他の 10 文を評価データとする 5 セット交差検定法により、評価を行う。サンプリング周波数は 16 kHz とする。スペクトル特徴量として、0 次から 24 次のメルケプストラム係数を用いる。スペクトル分析は、電気音声に対しては FFT 分析を用い、通常音声に対しては STRAIGHT 分析 [14] を用いる。また、非周期成分は、STRAIGHT 分析により抽出されたものを、5 周波数帯域で平均したものをを用いる [11]。分析フレームシフトは 5 ms とする。スペクトルセグメント特徴量抽出には前後 4 フレームを使用する。GMM の混合数は 32（スペクトル推定用）、64（基本周波数推定用）、32（非周期成分推定用）とし、特定話者モデルを用いる。また、連続基本周波数パターンを導入した際の基本周波数推定用 GMM の混合数は 32 とする。

客観評価実験および主観評価実験を行う。客観評価実験では、統計的声質変換に基づく強調法における変換精度を調査する。評価尺度として、強調音声と通常音声間のメルケプストラム歪み、有聲／無聲推定誤り、基本周波数パターン間の相関係数、および、非周期成分歪みを用いる。主観評価実験では、以下に示す各システムによる強調音声の自然性および聞き取りやすさについて、5 段階オピニオン評価（1：非常に悪い、2：悪い、3：普通、4：良い、5：非常に良い）を行う。

- 電気音声 (EL)
- 雑音抑圧に基づくスペクトル補正法による強調音声 (SS)
- 統計的手法に基づく声質変換法による強調音声 (VC)
- スペクトル補正及び統計的音源生成に基づくハイブリッド電気音声強調法による強調音声 (SS+VC)
- スペクトル補正及び統計的音源生成に基づくハイブリッド電気音声強調法による強調音声＋連続基本周波数パタ

表 1 Conversion accuracy in enhancement methods with VC.

	VC or SS+VC	SS+VC+CF0
Mel-cepstral distortion without power infomation	5.09 dB	
U to V error rate	4.04 %	17.9 %
V to U error rate	4.42 %	0.09 %
F_0 correlation coefficient	0.53	0.55
Aperiodic distortion	3.19 dB	

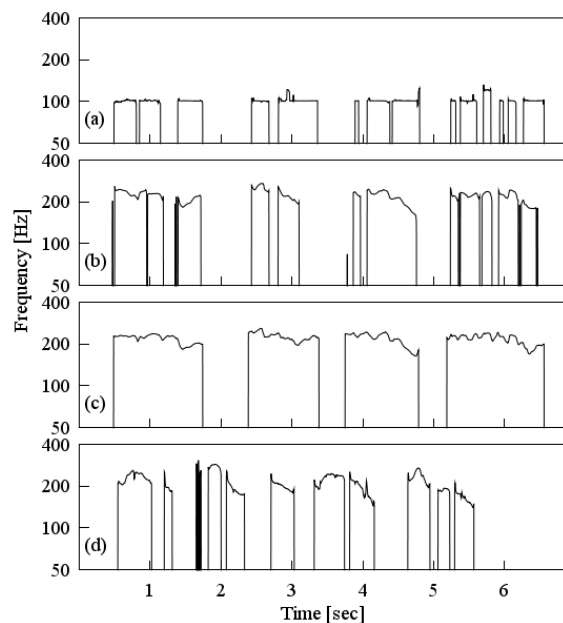


図 5 Example of F_0 patterns of a) original electrolaryngeal speech, b) enhanced speech with VC or SS+VC, c) enhanced speech with SS+VC+CF0, and d) target normal speech.

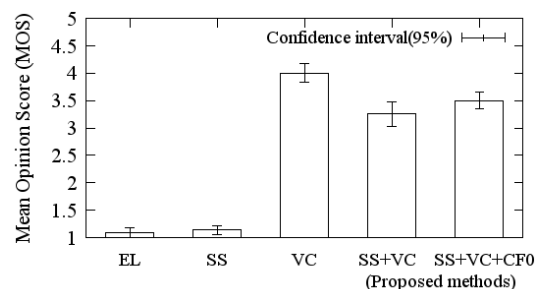


図 6 Result of opinion test on naturalness.

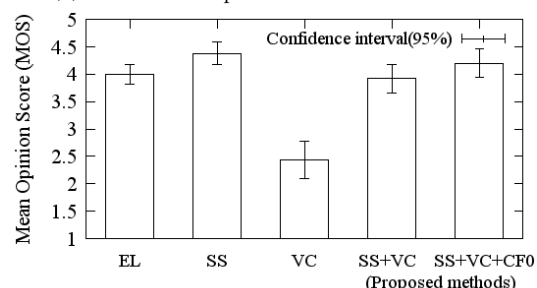


図 7 Result of opinion test on listenability.

ーン (SS+VC+CF0)

被験者は男性 5 名であり、1 人あたり各システムにつき 10 サンプルの計 50 サンプルを受聴する。

5.2 実験結果

客観評価実験結果を表 1 に示す。統計的声質変換に基づく有聲無聲情報推定処理では、有聲フレームを無聲フレームと

誤る確率が4.42%, 無声フレームを有声フレームと誤る確率が4.04%程度の推定精度が得られる。連続基本周波数パターンを導入すると、有声フレームを無声フレームと誤る確率はほぼ0となるが、元の電気音声と同様、発話区間中は全て有声音となみなすため、17.9%の無声フレームが有声フレームとして取り扱われる。また、連続基本周波数パターンの導入により、相関係数は若干改善する傾向が見られる。図5に基本周波数パターンの一例を示す。ここで、電気音声において発話中に短い無声区間が観測されるが、電気式人工喉頭をON・OFFしているわけではなく、基本周波数抽出誤差の影響である。従来の統計的手法に基づく声質変換により推定される基本周波数パターンでは、極めて短い有声区間の存在が見受けられる。また、その区間において推定される基本周波数は、値が大きく外れる傾向がしばしば見られる。一方で、連続基本周波数パターンを導入した際には、そのような問題は見受けられない。なお、連続基本周波数パターンのモデル化の効果は、隠れマルコフモデルに基づく音声合成においても報告されている[15]。

自然性に関する主観評価実験の結果を図6に示す。雑音抑圧に基づくスペクトル補正法による自然性改善効果は極めて小さい。一方で、統計的手法に基づく声質変換による強調法は、著しい自然性改善をもたらす。提案法は、統計的手法に基づく声質変換による強調法には及ばないまでも、電気音声の自然性を大幅に改善できることが分かる。また、連続基本周波数パターンの導入により、自然性が若干改善する傾向が見られる。

聞き取りやすさに関する主観評価実験の結果を図7に示す。雑音抑圧に基づくスペクトル補正法により、聞き取りやすさが若干改善する傾向が見られる。また、[5]の報告と同様、統計的手法に基づく声質変換による強調法は、聞き取りやすさを劣化させる傾向がある。これに対して、提案法では、聞き取りやすさの劣化がほぼ生じないことが分かる。このことから、スペクトル特徴量の変換誤差は聞き取りやすさに大きな影響を与えることが分かる。また、連続基本周波数パターンを導入することで、若干の改善が得られることから、有声フレームを無声フレームと誤って推定すると、聞き取りやすさに悪影響を与えることが分かる。

以上の結果から、提案法である連続基本周波数パターンを導入したハイブリッドな電気音声強調法は、電気音声の聞き取りやすさを保持し、自然性を大幅に改善できることが分かる。

6. 終わりに

本稿では、聞き取りやすさの劣化を引き起こさずに自然性を大幅に改善す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処理と統計的声質変換に基づく音源特徴量生成処理を組み合わせたハイブリッド法を提案した。また、連続基本周波数パターン推定法も導入することで、さらなる性能改善を達成した。客観評価実験及び主観評価実験の結果から、提案法の有効性を示した。今後は、書き取り試験により明瞭性に関する評価を行う予定である。また、提案法の実環境下への適用に向けて、耐雑音性の向上などに取り組む予定である。

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。

文 献

- [1] H. Liu, Q. Zhao, M.X. Wan, and S.P. Wang, "Enhancement of electrolarynx speech based on auditory masking," *IEEE Trans. Biomedical Engineering*, vol. 53, no. 5, pp. 865–874, May 2006.
- [2] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar 1998.
- [3] T. Toda, A.W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, and Language*, Vol. 15, No. 8, pp. 2222–2235, Nov 2007.
- [4] K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *SPECOM*, vol. 54, no. 1, pp. 134–146, Jan 2012.
- [5] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "An evaluation of alaryngeal speech enhancement methods based on voice conversion techniques," *Proc. ICASSP*, pp. 5136–5139, May 2011.
- [6] H.R. Sharifzadeh, I.V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *IEEE Trans. Biomedical Engineering*, vol. 57, no. 10, pp. 2448–2458, Oct 2010.
- [7] C.Y. Espy-Wilson, V.R. Chari, J.M. MacAuslan JM, C.B. Huang CB and M.J. Walsh, "Enhancement of electrolaryngeal speech by adaptive filtering," *Journal of Speech, Language, and Hearing Research* Vol. 41, No. 6, pp. 1253–1264, Dec 1998.
- [8] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr 1979.
- [9] B.L. Sim, Y.C. Tong, J.S. Chang, and C.T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 4, pp. 328–337, Jul 1998.
- [10] 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 system STRAIGHT," *Proc. 2nd MAVEBA*, Sep 2001.
- [11] 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," *信学論*, Vol. J91-D, No. 4, pp. 1082–1091, Apr. 2008.
- [12] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. ICASSP*, pp. 285–288, May 1998.
- [13] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. Audio, Speech, and Language*, Vol. 20, No. 9, pp. 2505–2517, Nov 2012.
- [14] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Re-structuring 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," *SPECOM*, Vol. 27, No. 3-4, pp. 187–207, Apr 1999.
- [15] K. Yu and S. Young, "Continuous F_0 modelling for HMM based statistical parametric speech synthesis," *IEEE Trans. Audio, Speech, and Language*, Vol. 19, No. 5, pp. 1071–1079, Jul 2011.