一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2014-32(2014-05)

雑音抑圧及び統計的音源予測に基づくハイブリッドな電気音声強調法の評価

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，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に，電気式人工喉頭を用いた発声法があり，本発声法により生成される音声を電気音声と呼ぶ．熟練の喉頭摘出者は明瞭性の高い電気音声を生成できる．しかしながら，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，特に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ため，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，我々は，明瞭性の劣化を引き起こさずに自然性を大幅に改善する強調法として，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した．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の変換処理を回避することで，変換誤差による明瞭性劣化を防ぐ．また，統計的手法により，通常音声の基本周波数パターンへを生成することで，自然性を大幅に改善する．本報告では，提案法について述べるとともに，実験的評価結果からその有効性を示す．

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

An Evaluation of a Hybrid Approach to Electrolaryngeal Speech Enhancement Based on Noise Reduction and Statistical 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　This report presents our recent work on the development of a hybrid approach to improving naturalness of electrolaryngeal (EL) speech while minimizing degradation in intelligibility. 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 a noise reduction method for enhancing spectral parameters and a voice conversion method for predicting excitation parameters. The experimental results demonstrate that the proposed method yields significant improvements in naturalness compared with EL speech while keeping intelligibility 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，

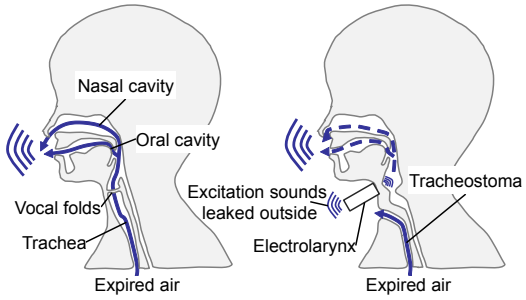

図 1　Speech production mechanisms of non-disabled people (left figure) and total laryngectomees (right figure).

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が生成される．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，比較的明瞭性の高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（Fundamental frequency; F_0）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り，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 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．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，明瞭性は若干劣化する傾向がある [5]．

本研究では，明瞭性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，変換誤差の影響を低減させるため，連続 F_0 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，電気音声の明瞭性を保ちながら，自然性を大幅に改善できることを示す．

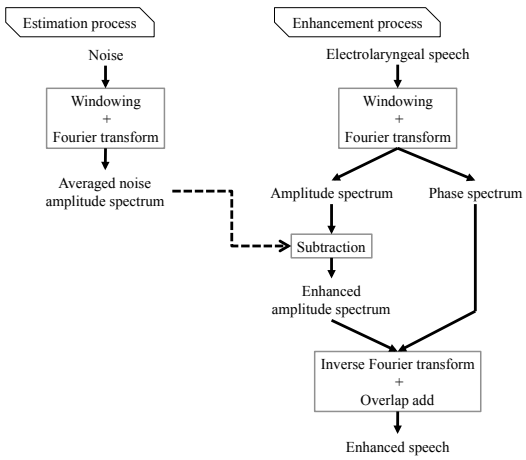

図 2　EL speech enhancement based on SS.

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．

$$Y(t) = S(t) + L(t) \tag{1}$$

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．

$$Y(\omega, t) = S(\omega, t) + L(\omega, t) \tag{2}$$

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．

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

ここで，$\alpha(\alpha > 1)$ は過減算パラメータ，$\beta(0 \leqq \beta \leqq 1)$ 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，無音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．一方で，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．

本報告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成した際に，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 $|L(\omega, t)|$ を求める．スペクトル減算処理で

は，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である F_0 と非周期成分（Aperiodic Components; AC）[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，F_0，AC 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（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．

$$\boldsymbol{X}_t = \boldsymbol{E}[\boldsymbol{x}_{t-C}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+C}^\top]^\top + \boldsymbol{f} \tag{4}$$

ここで，$\boldsymbol{x}$ 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．$\boldsymbol{E}$ 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，通常音声の個々の音響特徴量として，$\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top,\ \Delta\boldsymbol{y}_t^\top]^\top$ を使用する．ここで，動的特徴量 $\Delta\boldsymbol{y}$ は $\Delta\boldsymbol{y}_t = \boldsymbol{y}_t - \boldsymbol{y}_{t-1}$ 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 F_0 および帯域別平均 AC [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 $\boldsymbol{X}_t$ 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].

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

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，λ 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．

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

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 F_0，帯域別平均 AC の間において，計 3 つの GMM を学習する．

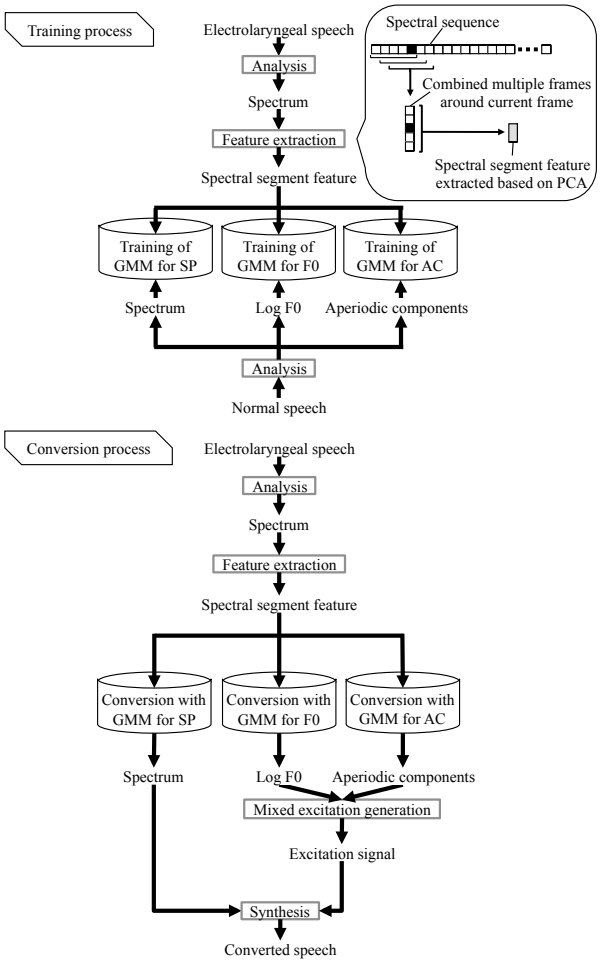

図 3　EL speech enhancement based on VC.

また，通常音声の音響特徴量（メルケプストラム）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 $\boldsymbol{v}^{(y)} = [v_1^{(y)}, \cdots, v_{D(y)}^{(y)}]^\top$ 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．

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

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 $\boldsymbol{y}_t$ の d 次元目の要素である．GV の確率密度関数 $P(\boldsymbol{v}^{(y)}|\boldsymbol{\lambda}^{(v)})$ 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．

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

3.2 処理過程：変換部

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 $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_t^\top, \cdots, \boldsymbol{X}_T^\top]^\top, \boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_t^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$ とおく．このとき，変換後の静的特徴量系列 $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ は次式で計算される．

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

$$\text{subject to}\, \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{9}$$

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 の尤度重みを表す．なお，通常音声の対数 F_0 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，体内伝導無声音声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 F_0 パターンおよび AC 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

4.1 従来の電気音声強調法の利点および欠点

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 F_0 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明瞭性も維持される．

統計的手法に基づく声質変換法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が生成される．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，強調音声の明瞭性は電気音声と比較して若干劣化する傾向がある．

4.2 ハイブリッド電気音声強調法

明瞭性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，比較的明瞭性の高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，F_0 パターンは発声区間でほぼ一定であり，AC は一貫して強いものとなる．そこで，音源特徴量である F_0 および AC 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が得られる．特に，F_0 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，元の電気音声が持つ人工的なものと比較すると，より自然音声に近いものが得られる．また，AC に関しても，音韻に応じてより自然に変化するものが得られる．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．

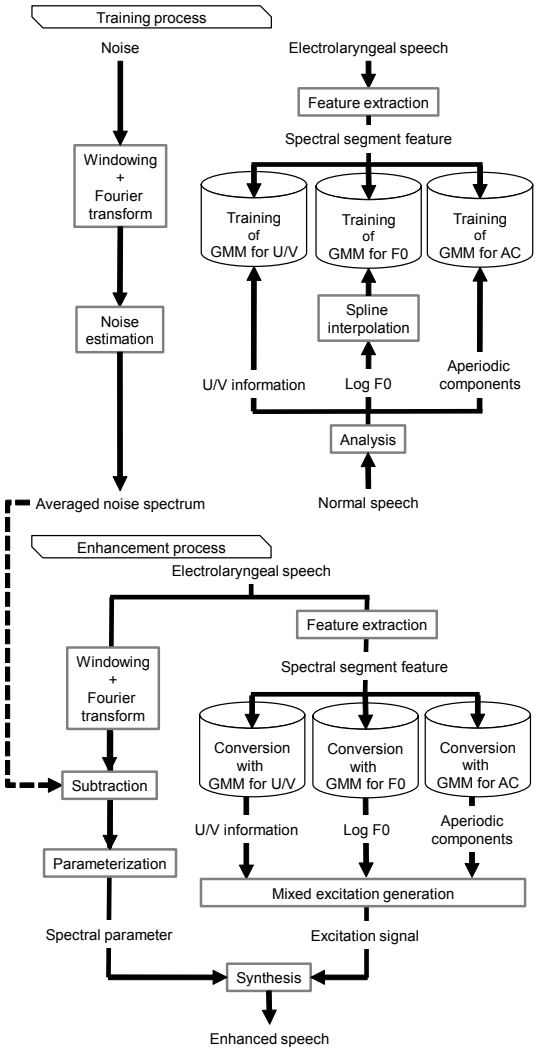

図 4　EL speech enhancement based on a hybrid approach.

4.3 音源特徴量予測の改善

4.3.1 連続 F_0 モデルの導入（CF0）

無声区間では F_0 が観測できないため，F_0 パターンは不連続なものとなる．例えば，無声区間における F_0 の値として，有声区間で観測される値とは明らかに異なる値（例えば 0 など）を用いる [13]．このような不連続な F_0 パターンをモデル化するのは容易ではなく，複雑なモデルが必要となる．

これに対して，主に統計的パラメトリック音声合成の分野において，無声区間においても連続的な F_0 パターンが観測できるものとしてモデル化を行う連続 F_0（Continuous F_0; CF0）モデルが提案されており，その有効性が報告されている [14]．そこで，本報告では，F_0 パターン予測に連続 F_0 モデルを導入する．無声区間に対してスプライン補間処理を行うことで，連続的な F_0 パターンを生成した後に，GMM によるモデル化を行う．なお，U/V 情報に関しては，F_0 パターンとは別の GMM によりモデル化する．

4.3.2 マイクロプロソディの除去（LPF）

通常音声から抽出される F_0 パターン上では，マイクロプロソディと呼ばれる急峻な変化がしばしば観測される．一方で，ハイブリッド方式において，マイクロプロソディを精度良く予測するのは容易ではなく，より複雑なモデルが必要となる．そこで，現状のモデル（GMM）で上手くモデル化できないマイ

クロプロソディに関しては，ノイズとみなし，モデル学習の前段で除去する．これにより，F_0 パターンの概形をより精度良くモデル化するように GMM が学習されると期待される．なお，除去処理には，低域通過フィルタ（Low-pass filter; LPF）を用いる．

4.3.3 U/V 予測の回避

自然な F_0 パターンを生成するためには，U/V 情報を予測し付与する必要がある．しかしながら，ハイブリッド方式における U/V 予測処理は本質的に困難な処理であり，少なからず推定誤差が生じる．この推定誤差は，強調音声の品質劣化を引き起こす要因となり得る．特に，有声音を無声音とする予測誤差（V to U）が強調音声の品質に与える影響は大きい．

EL 強調処理において，強調前の EL 音声は，音源信号が生成されていない無音区間を除き，全て有声音である．そのため，無声区間を持たない連続 F_0 パターンを用いたとしても，強調前と比べて，悪影響は生じない．逆に，V to U の予測誤差による品質劣化を回避できるという利点がある．そこで，U/V 予測を行わず，連続 F_0 パターンを用いて強調音声を生成する．なお，無音区間に関しては，EL 音声の波形パワーを用いて自動的に検出し，無声フレームとして合成する．

5. 実験的評価

5.1 実験条件

喉頭摘出者 1 名の EL 音声と，健常者 1 名の通常音声を用いる．学習データとして ATR 音素バランス文セット中の 50 文中 40 文を用い，評価データとして残りの 10 文を用い，交差検定を行う．サンプリング周波数は 16 kHz，分析フレーム長は 25 ms，分析フレームシフトは 5 ms とする．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 EL 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 [15] を用いる．GMM の混合数は 32（スペクトル変換用），16（非周期成分推定用）とする．LPF のカットオフ周波数は 10 Hz とする．

客観評価実験では，学習データにおける F_0 パターンが F_0 推定精度に与える影響を調査する．その際に，F_0 推定用 GMM の混合数を 8，16，32，64 と変化させる．主観評価実験では，以下に示すシステムの音声について，自然性に関する 5 段階オピニオン評定とプリファレンステスト，及び明瞭性に関する書き取り試験を行う．

EL: 電気音声

SS: 雑音抑圧に基づくスペクトル補正処理音声

VC: 統計的手法に基づく強調音声

SS+VC: 音源特徴量予測に対する改善処理を施していないハイブリッド法に基づく強調音声

SS+VC+CF0: 連続 F_0 パターンを考慮したハイブリッド法に基づく強調音声

Hybrid (V): 発話区間を全て有声音としたハイブリッド法に基づく強調音声

Hybrid (U/V): VC に基づく推定 U/V 情報を用いたハイブリッド法に基づく強調音声

Hybrid (target U/V): 理想的な U/V 情報を用いたハイブリッド法に基づく強調音声

ここで，Hybrid (V, U/V, target U/V) におけるハイブリッド方式においては，SS+VC に対して CF0 および LPF を導入したものを用いる．また，理想的な U/V 情報は，VC に基づく EL 強調音声と通常音声との間で DTW を行うことで得る．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．

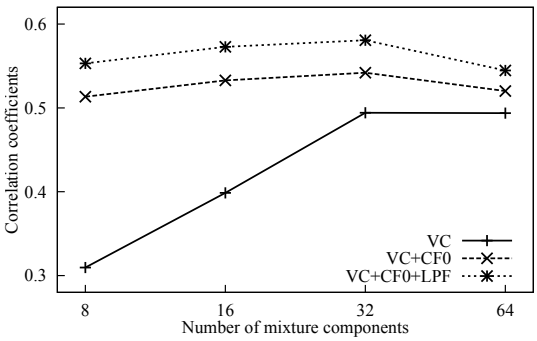

図 5 各手法における F_0 推定精度

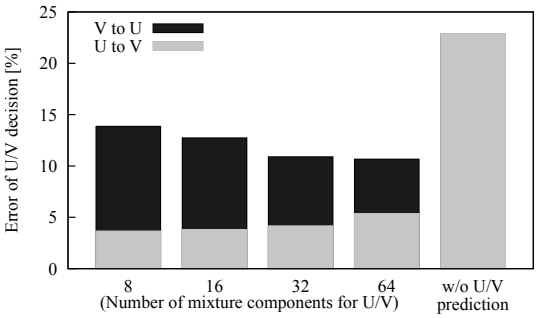

図 6 U/V 予測処理の有無に対する U/V 予測誤差

5.2 客観評価実験結果

図 5 に音源特徴量予測時における各手法における F_0 推定精度を示す．CF0 及び LPF の導入により相関係数が改善する．これより，学習データ中の F_0 パターンに対して，無声区間を補間し，マイクロプロソディを除去することは有効であると言える．また，最適な混合数は 32 である．

図 6 に音源特徴量予測時における U/V 予測処理の有無に対する U/V 予測誤差を示す．U/V 予測処理の回避により，V to U の予測誤差は 0 となるが，U to V の予測誤差は増大する．なお，EL 音声も同様の予測誤差を持つと考えられる．

5.3 主観評価実験結果

図 7 に自然性に関する MOS 評価結果を示す．SS による自然性改善効果は極めて小さい．一方で，VC は，著しい自然性改善をもたらす．提案法は，VC 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，CF0 の導入により，自然性が若干改善する傾向が見られる．

図 8 に書き取り試験結果を示す．文献 [5] において，VC に基づく EL 音声強調は明瞭性を劣化させること（単語正解率で約 3%劣化）が報告されているが，ハイブリッド方式は明瞭性劣化をもたらさないことが分かる．また，ハイブリッド方式において，U/V 予測を回避した際においても，理想的な U/V 情報を用いた場合と同等の明瞭性が得られていることから，必ずしも U/V 予測が必要ではないことが分かる．一方で，SS と比較すると，明瞭性が若干低下する傾向が見られる．この原因として，ボコーダによる波形合成の影響が考えられる．

図 9 にプリファレンステストの結果を示す．理想的な U/V 情報を用いた場合と U/V 予測処理を回避した場合とで有意差が見られないことから，U/V 予測処理の回避が，ハイブリッドな手法に基づく強調音声の自然性に与える影響はないことがわかる．なお，図 8 同様に，理想的な U/V 情報を用いた場合とその他の場合と比較した際に有意差が見られない原因としては，ボコーダによる波形合成の際のスペクトル特徴量と U/V 情報の不一致が考えられる．電気音声の発話中は必ず有声音となるため，SS を用いて電気音声から抽出されるスペクトル特徴量は全て有声音のスペクトル特徴量となる．一方で，通常音声の

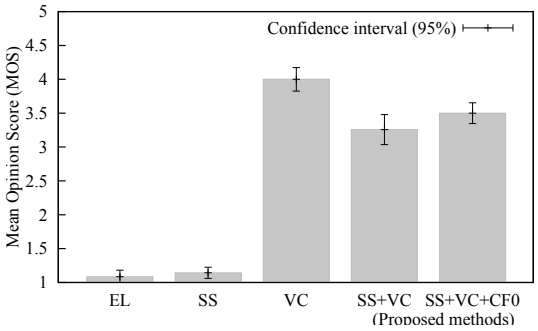

図 7　自然性に関する MOS 評価結果

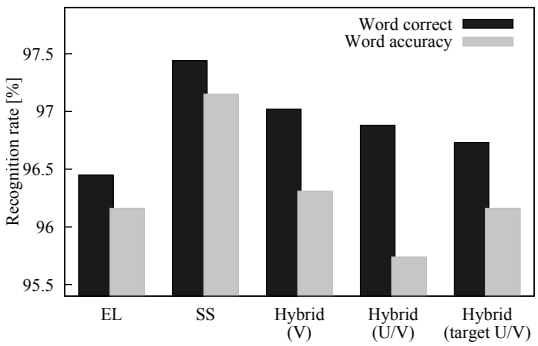

図 8　明瞭性に関する書き取り試験結果

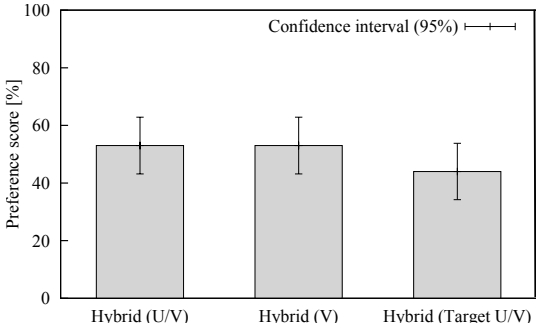

図 9　自然性に関するプリファレンステスト結果

場合，有声音と無声音で，スペクトル特徴量（例えばスペクトル傾斜）は大きく変化する．現状のハイブリッド法では，U/V 情報に基づくスペクトル特徴量の補正処理を行っていないので，有声音のスペクトル特徴量と無声音の音源信号が畳み込まれる場合がある．そのため，理想的な U/V 情報を音源特徴量に反映させるだけでは，音質改善につながらないと考えられる．

6. 終 わ り に

本報告では，明瞭性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，我々が提案した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について述べた．客観評価実験の結果から，連続 F_0 モデルの導入，マイクロプロソディの除去の有効性を示した．また，主観評価実験の結果から，提案法の有効性を示し，電気音声強調において有声無声情報の予測処理を回避できることを示した．

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. 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.

[2] 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.

[3] 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.

[4] K. Nakamura, T. Toda, H. Saruwatari, and 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, 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] H.R. Sharifzadeh, I.V. McLoughlin, and F. Ahmadi, “Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec,” Biomedical Engineering, IEEE Transactions on, vol.57, no.10, pp.2448–2458, Oct. 2010.

[7] C.Y. Espy-Wilson, V.R. Chari, J.M. MacAuslan, C.B. Huang, and M.J. Walsh, “Enhancement of electrolaryngeal speech by adaptive filtering,” Journal of Speech, Language, and Hearing Research, vol.41, no.6, p.1253, Dec. 1998.

[8] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” Acoustics, Speech and Signal Processing, IEEE Transactions on, vol.27, no.2, pp.113–120, April 1979.

[9] B.L. Sim, Y.C. Tong, J.S. Chang, and C.T. Tan, “A parametric formulation of the generalized spectral subtraction method,” Speech and Audio Processing, IEEE Transactions on, vol.6, no.4, pp.328–337, July 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 synthesis system straight,” Proc. MAVEBA, pp.13–15, Sept. 2001.

[11] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation,” Proc. INTERSPEECH, pp.2266–2269, Sept. 2006.

[12] A. Kain and M.W. Macon, “Spectral voice conversion for text-to-speech synthesis,” Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, vol.1, pp.285–288, May 1998.

[13] 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.

[14] K. Yu and S. Young, “Continuous F0 modeling for HMM based statistical parametric speech synthesis,” Audio, Speech, and Language Processing, IEEE Transactions on, vol.19, no.5, pp.1071–1079, July 2011.

[15] 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,” SPECOM, vol.27, no.3, pp.187–207, April 1999.