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.

ハイブリッド式電気音声強調法における音源特徴量予測の評価

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 F_0 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．これらの問題に対して，我々は，スペクトル減算法に基づくスペクトル特徴量強調処理と，連続 F_0 モデルを考慮した統計的声質変換法による音源特徴量予測処理を併用した，ハイブリッド方式に基づく電気音声強調法を提案した．本報告では，ハイブリッド方式における音源特徴量予測処理に着目し，その精度を実験的に評価する．また，予測精度の改善を目指し，マイクロプロソディの除去処理及び有声無声予測の回避処理を導入し，その有効性を評価する．

キーワード　発声障害者補助，電気音声強調，ハイブリッド方式，統計的音源特徴量予測，有声無声情報

Evaluation of Excitation Feature Prediction in a Hybrid Approach to Electrolaryngeal Speech Enhancement

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 implement removing micro-prosody with low-pass filtering and avoiding Unvoiced/Voiced (U/V) prediction as part of a hybrid approach to improve statistical excitation prediction in the hybrid approach to electrolaryngeal (EL) speech enhancement. 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, in our previous work, we proposed a hybrid method using the noise reduction method for enhancing spectral parameters and voice conversion method for predicting excitation parameters. In this paper, we evaluate the effect of removing micro-prosody with low-pass filtering and avoiding U/V prediction in the hybrid enhancement process.

Key words　speaking-aid, electrolaryngeal speech, hybrid approach, statistical excitation prediction, unvoiced/voiced information

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 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，スペクトル減算法（Spectral Subtraction: SS）[1] [2] などを用いた雑音抑圧に基づくスペクトル補正法 [3] と，統計的声質変換（statistical Voice Conversion: VC）[4] [5] に基づく通常音声への変換法 [6] [7] がある．前者の手法は，明瞭性および自然性がわずかに向上するが，その改善効果は極めて限定的であり，特に自然性は依然として著しく低い．一方，後者の手法は，自然性を大幅に改善できるが，少なからず変換誤差が生じるため，明瞭性が劣化する [7].

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，スペクトル特徴量強調処理には SS に基づく電気音声強調処理を，音源特徴量予測処理には VC に基づく電気音声強調処理を用いたハイブリッドな電気音声強調法を新たに提案した [8]．さらに，不連続な F0 パターンのモデル化は比較的困難である [9] という問題点に対しては，主に統計的パラメトリック音声合成の分野においてその有効性が確認されている連続 F_0（Continuous F_0: CF0）モデル [10] を，ハイブリッドな電気音声強調法における F_0 パターン予測処理に導入した．これにより，予測精度は若干向上したものの，依然として十分に高い精度は得られていない．また，音源特徴量の一つである有声無声情報（Unvoiced/Voiced: U/V）の予測処理も統計的手法に基づいているため，少なからず発生する変換誤差による悪影響を受けるという問題がある．特に，有声フレームを無声フレームと推定する誤差は，明瞭性を劣化させる要因の一つと考えられる．

本研究では，ハイブリッドな電気音声強調処理において本質的に極めて重要な処理となる音源特徴量予測に着目し，予測精度を実験的に評価する．また，予測精度を改善するために，低域通過フィルター（Low-Pass Filter: LPF）を使用したマイクロプロソディの除去を導入する．さらに，VC に基づく U/V 予測処理の際に少なからず生じる変換誤差の影響を調査し，U/V 予測処理の取り扱いについて検討する．

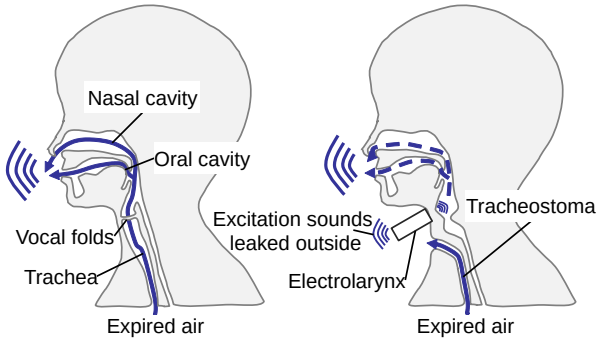

図 1　Speech production mechanisms of non-disabled people (left figure) and total laryngectomees (right figure).

2.　ハイブリッド電気音声強調法（SS+VC+CF0）

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，SS により得られるスペクトル特徴量を用いる．VC 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，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 F_0 および非周期成分に関しては，VC 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未だ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，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．強調処理の流れを図 2 に示す．

スペクトル特徴量強調処理においては，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，SS により除去する．時間 t における電気音声信号 $Y(t)$ は以下で記述される．

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

ここで，S は口から放射されたクリーンな音声信号，L は空気中に漏れ出した電気式人工喉頭の音源信号を表す．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．

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

強調処理部では，雑音信号 L の定常性を仮定し，推定された雑

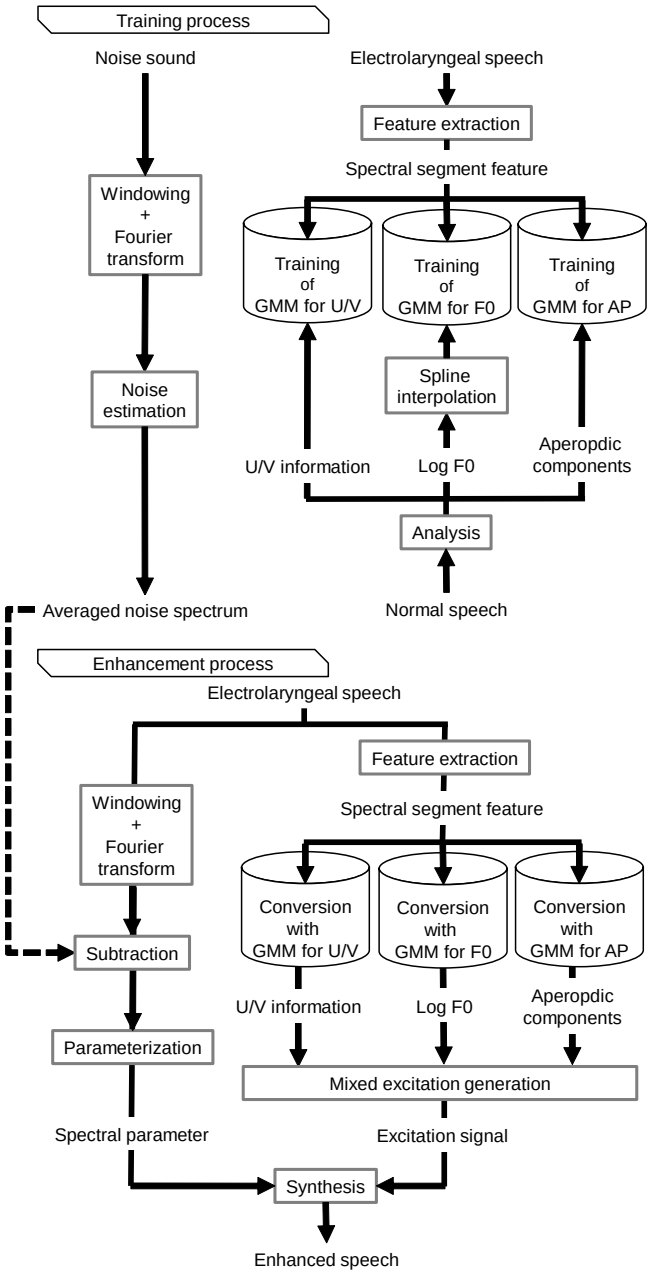

図 2　EL speech enhancement based on a hybrid approach.

音の振幅スペクトルの期待値 $|\hat{L}_{(\omega)}|$ を観測信号の振幅スペクトル $|Y_{(\omega,t)}|$ 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．

$$|\hat{S}_{(\omega,t)}|^{\gamma} = \begin{cases} |Y_{(\omega,t)}|^{\gamma} - \alpha|\hat{L}_{(\omega)}|^{\gamma} & (\frac{|\hat{L}_{(\omega)}|^{\gamma}}{|Y_{(\omega,t)}|^{\gamma}} < \frac{1}{\alpha}) \\ 0 & (\text{otherwise}) \end{cases} \tag{3}$$

ここで，α ($\alpha > 1$) は過減算パラメータ，γ は指数パラメータである．なお，$\hat{L}(\omega)$ は，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の，口元のマイクで観測される雑音信号の振幅スペクトル $|L(\omega,t)|$ 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．

音源特徴量予測処理では，CF0 を考慮した VC に基づいており，学習部と変換部からなる．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音源特徴量（U/V 情報，無声区間に対してスプライン補間処理を行うことで得られた連続的な F_0，および非周期成分 [11]）への変換モデルが，各々事前に学習される．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 $\boldsymbol{X}_t$ 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，$\boldsymbol{x}_t$ 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．$\boldsymbol{E}$ 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．音源特徴量として U/V 情報，対数 F_0，および帯域別平均非周期成分 [12] を用いる．ここで，U/V 情報に関しては，有音区間を対数 F_0，無音区間を 0 とした F_0 パターンを用いることで表現する．また，対数 F_0 に関しては，通常音声から抽出された F_0 パターンの無声区間に対してスプライン補間を行い，本来の F_0 パターンが有する不連続性を解消した連続的な F_0 パターンの値を使用する．

パラレルデータに対して動的時間伸縮（Dynamic Time Wraping: 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 [13].

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

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，$\boldsymbol{\lambda}$ 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．

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

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 U/V 情報，対数 F_0，および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．

変換部では，得られた変換モデルを用いて，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から，通常音声の U/V 情報，F_0，及び非周期成分が推定される．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．

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) \text{ subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \qquad (7)$$

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．なお，通常音声の対数 F_0 への変換処理においては，U/V 情報推定用の GMM により推定された U/V 情報を使用する．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 [9] で行われるものと同一である．

変換後の F_0 パターンおよび非周期成分を用いて混合励振源モデル [12] により音源信号を生成する．一方，スペクトル特徴量に関しては，SS により雑音の影響を抑圧されたスペクトルから包絡成分を抽出する．VC に基づいて生成された音源信号と SS に基づいて雑音抑圧されたスペクトル包絡を畳み込むことで，最終的に強調音声を生成する．

3. 音源特徴量予測の改善

ハイブリッドな電気音声強調法において，音源特徴量予測処理は強調音声の品質に大きな影響を与える．本報告では，より精度の高い予測処理の実現及び明瞭性の劣化を極力防ぐ処理の実現を目指し，学習に用いる F_0 パターンの改善及び U/V 予測処理に関する検討を行う．

3.1 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，GMM に基づく統計的予測処理において，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，マイクロプロソディに関しては，ノイズとみなし，GMM 学習の前段で除去する．除去処理には，低域通過フィルタ [14] を用い，連続的な F_0 パターンを平滑化する．処理後は，通常の F_0 パターンのモデル化同様，平滑化された連続的な F_0 パターンを GMM でモデル化する．実際の F_0 パターンに対して，この処理を行ったときの一例を図 3 に示す．

3.2 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，電気音声強調における 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と予想される．

電気音声強調処理において，強調前の電気音声は，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，電気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．なお，文献 [8] と本研究において違う点は，U/V 情報が強調音声に与える影響をより詳細に評価した点にある．

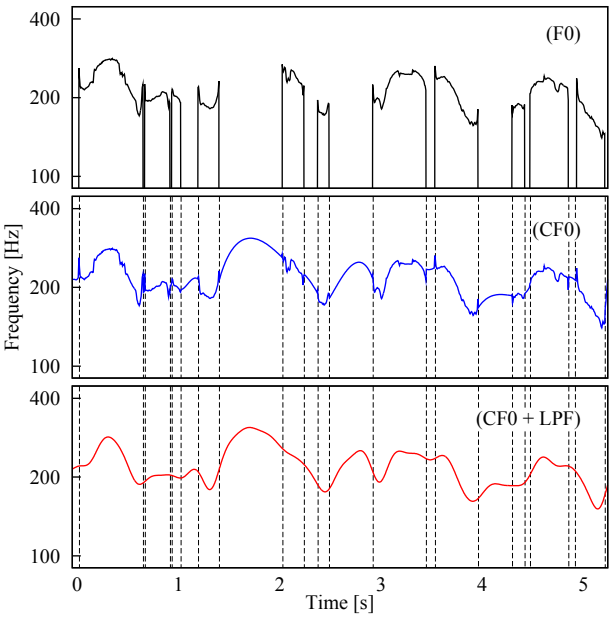

図 3 Each type of F_0 pattern. Top figure is a target F_0 pattern, the middle is a continuous F_0 pattern using spline interpolation, and the bottom is a continuous F_0 pattern smoothed using the low-pass filter (cutoff frequency = 10 Hz).

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．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，残りの 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．EL 音声に対するスペクトル分析は FFT 分析を用い，通常音声に対する非周期成分分析は，STRAIGHT 分析 [15] 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用いる [12]．分析フレーム長は 25 ms とし，分析フレームシフト長は 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．上記二話者の音声データを用いて，話者対依存 GMM を学習する．非周期成分推定用 GMM の混合数は 16 とする．なお，その際の非周期成分変換精度は 3.19 dB である．

客観評価実験では，統計的音源生成に基づく音源特徴量予測において，F_0 パターンに対するマイクロプロソディの除去が F_0 推定精度に与える影響を調査する．評価尺度として，強調音声と通常音声間の有声／無声推定誤り，基本周波数パターン間の相関係数を用いる．その際の LPF のカットオフ周波数は 5 Hz，10 Hz，20 Hz，25 Hz と変化させ， F_0 予測用および U/V 予測用 GMM の混合数は 8，16，32，64 と変化させる．

主観評価実験では，強調音声の明瞭性についての書き取り試験，及び，自然性におけるプレファレンステストを行う．明瞭性についての書き取り試験では，ハイブリッドな電気音声強調

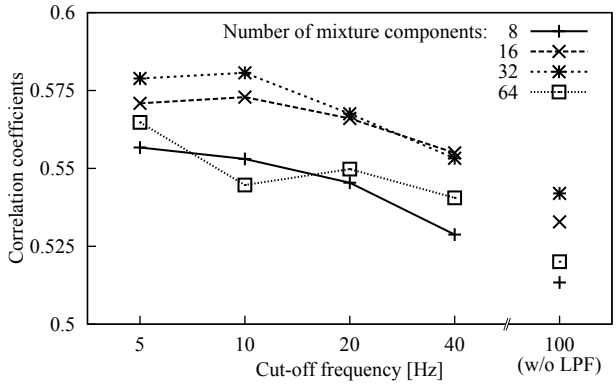

図 4　Relationship between cut-off frequency of LPF and F_0 correlation coefficients.

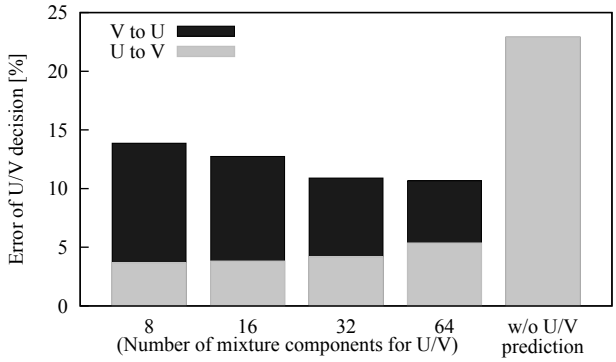

図 5　U/V error rate of each system.

処理及び U/V 情報が明瞭性に与える影響を調査するために，以下に示す各システムを評価する．

- 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
- Hybrid (V): 発話区間が全て有声音
- Hybrid (U/V): VC に基づく推定 U/V 情報
- Hybrid (target U/V): 理想的な U/V 情報

ここで，ハイブリッド方式においては，SS+VC+CF0 に対して 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． 自然性におけるプリファレンステストでは，U/V 情報が自然性に与える影響を調査するために，以下に示す各システムを総当たりで対比較する．

- Hybrid (V)
- Hybrid (U/V)
- Hybrid (target U/V)

各主観評価実験の被験者は男性 5 名であり，1 人あたり各システムにつき 10 サンプルを受聴する．また，この際に用いた LPF のカットオフ周波数は 10 Hz で，F_0 予測用および U/V 予測用 GMM の混合数は 32 である．

4.2　実 験 結 果

図 4 に，F_0 モデル化混合数ごとに LPF のカットオフ周波数を変化させたときの音源特徴量予測精度を示す．混合数が 32 でかつカットオフ周波数が 10 Hz のときが最適であることがわかる．また，このことから，GMM を用いた推定精度を劣化させる要因となるマイクロプロソディなどの急須な変化は，10 Hz 以上に存在することがわかる．

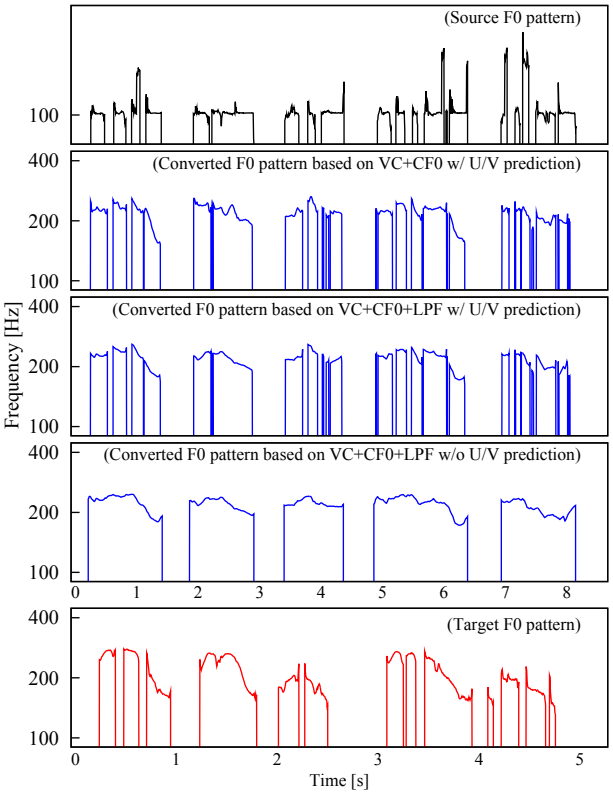

図 6　Each type of F_0 pattern.

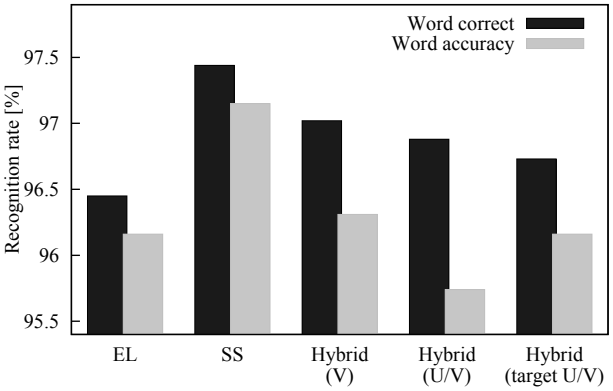

図 7　Result of dictation test on intelligibility.

図 5 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．また，EL 音声も同様の予測誤差を持つと考えられる．なお，実際に推定された F_0 パターンの一例を図 6 に示す．

図 7 に書き取り試験結果を示す．文献 [16] において，VC に基づく EL 音声強調は電気音声と比較して，約 3% の明瞭性を劣化させること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

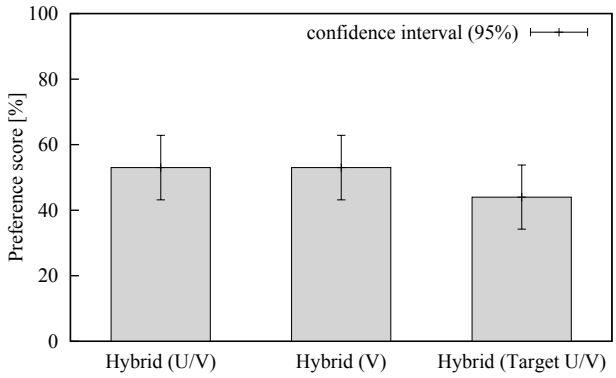

図 8　Result of preference test on naturalness.

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，スペクトル特徴量及び音源特徴量をボコーダにより波形合成した影響が考えられる．なお，文献 [8] で報告されている通り，SS のみの自然性はハイブリッド方式と比べて著しく低い．

図 8 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．一方で，理想的な U/V 情報を用いた場合と U/V 情報を VC に基づいて推定した場合とで有意差が見られないことから，VC に基づく U/V 情報推定処理は，自然性に関して，十分な精度を保っていることがわかる．なお，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音源特徴量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，理想的な U/V 情報を用いた音源特徴量に関しては，通常音声同様に有声音及び無声音の音源特徴量となる．それ故に，有声音のスペクトル特徴量と無声音の音源特徴量が畳み込まれる場合があり，現状の電気音声強調に対して理想的な U/V 情報を用いても音質改善につながらないと考えられる．

5. ま と め

従来の連続 F_0 モデルを考慮したハイブリッド方式に基づく電気音声強調法において，F_0 推定精度に不十分さがあることから，統計的音源生成の学習処理に用いる連続的な F_0 パターンに対するマイクロプロソディの除去を提案した．また，その際の U/V 情報の取り扱いについて検討した．客観評価実験の結果から，マイクロプロソディの除去処理の有効性を示した．また，主観評価実験の結果から，U/V 予測処理を行う必要はないといえる．

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] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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，“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,” **信学技報**, 113(76), SP2013-37, pp. 37–42, Jun. 2013.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] A. Sakurai and K. Hirose, “Detection of phrase boundaries in Japanese by low-pass filtering of fundamental frequency contours,” *Proc. ICSLP,* Vol. 2, pp. 817–820, Oct 1996.

[15] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, “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,” *SPECOM,* Vol. 27, No. 3-4, pp. 187–207, Apr 1999.

[16] H. Doi., “Augmented speech production beyond physical constraints using statistical voice conversion -Alaryngeal speech enhancement and singing voice quality control-,” *NAIST Doctoral Dissertation,* NAIST-IS-DD1061014, Mar 2013.