ハイブリッド電気音声強調法における音源特徴量予測*

○田中宏, 戸田智基, Graham Neubig, Sakriani Sakti, 中村哲（奈良先端大）

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．本発声法により生成される音声である電気音声（ElectroLaryngeal speech: EL）は，明瞭性が比較的高いものの，自然性は著しく低い．この問題に対する代表的なEL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（Spectral Subtraction: SS）[1]と統計的手法に基づく声質変換（statistical Voice Conversion: VC）[2]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，明瞭性が劣化する．そこで，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，SSによる補正スペクトル特徴量とVCにより予測される音源特徴量を用いたハイブリッド方式[3]を提案し，その有効性を示した．

本稿では，ハイブリッド方式のさらなる改善を目指し，VCに基づく音源特徴量予測の精度向上に取り組む．連続F_0モデル[4]及びマイクロプロソディの除去処理[5]を導入し，さらに有声無声（Unvoiced/Voiced: U/V）情報の取り扱いについて検討する．

2　ハイブリッド電気音声強調法（SS+VC）

喉頭摘出者の調音器官は正常に機能する場合が多く，EL音声のスペクトル特徴量は，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する．一方で，EL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．特に，F_0パターンの差は大きく，EL音声の自然性を大きく劣化させる主要因といえる．そこで，ハイブリッド方式では，EL音声から得られるスペクトル特徴量を最大限に活用するSSと，通常音声の統計量を活用して自然音声に近い音源特徴量を予測するVCを組み合わせることで，強調処理を行う．

SSでは，外部に雑音として漏れ出す音源信号Lの定常性を仮定し，そ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，強調信号の振幅スペクトル$|\hat{S}_{(\omega,t)}|$を求める．

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

ここで，tは時間，ωは周波数，$\alpha(\alpha > 0)$は減算パラメータ，γは指数パラメータとする．

VCは学習処理と変換処理で構成される．学習処理では，EL音声と通常音声の同一発話データを用いて，変換モデルを学習する．時間フレームtにおいて，前後Cフレームから抽出されるEL音声のスペクトルセグメント特徴量を$\boldsymbol{X}_t$とし，通常音声の静的･動的音源特徴量を$\boldsymbol{Y}_t = [\boldsymbol{y}_t^{\top}, \Delta\boldsymbol{y}_t^{\top}]^{\top}$とする．学習データに対する動的時間伸縮（Dynamic Time Warping: DTW）により対応付けられた結合ベクトル$[\boldsymbol{X}_t^{\top}, \boldsymbol{Y}_t^{\top}]^{\top}$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[6]．

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

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$\boldsymbol{\Sigma}$を持つ正規分布を示す．また，$\boldsymbol{\lambda}$はモデルパラメータセットを示し，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．

変換処理では，最尤系列変換法[7]により，EL音声のスペクトルセグメント特徴量系列から通常音声の音源特徴量系列へと変換する．

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

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．

3　音源特徴量予測の改善

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，従来のハイブリッド方式[3]では，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[8]．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[4]．そこで，本稿では，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．

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．除去処理には，低域通過フィルタ（Low-Pass Filter: LPF）を用いる．

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

*Excitation feature prediction in a hybrid approach to electrolaryngeal speech enhancement. by TANAKA, Ko, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani and NAKAMURA, Satoshi (NAIST)

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．

4 実験的評価

4.1 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 [9] を用いる．GMM の混合数は 32（スペクトル変換用），32（F_0 推定用），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，以下に示す各システムによる音声について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
- 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 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．

4.2 実験結果

図 1 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 2 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．

図 3 に書き取り試験結果を示す．文献 [10] 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．なお，文献 [3] 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ことに注意する．

以上の結果から，ハイブリッド方式において，連続的な F_0 パターンを導入することで，F_0 予測精度を改善し，U/V 予測処理を回避することが可能となり，EL 音声の明瞭性を保持した音声強調処理を実現できることが分かる．

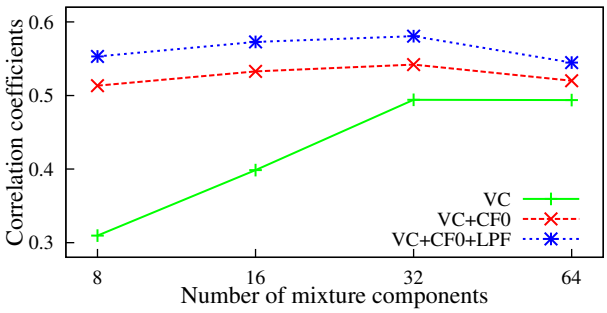

Fig. 1 各手法における F_0 推定精度

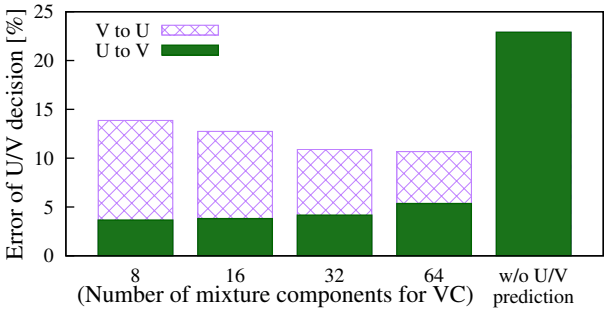

Fig. 2 U/V 予測処理の有無に対する U/V 予測誤差

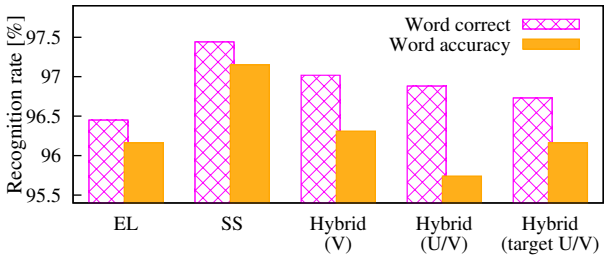

Fig. 3 書き取り試験結果

5 まとめ

ハイブリッド方式に基づく電気音声強調処理において，連続 F_0 モデルの導入，マイクロプロソディの除去，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，連続 F_0 モデルの有効性，マイクロプロソディの除去処理の有効性を示した．また，書き取り試験の結果から，U/V 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．

参考文献

[1] H. Liu *et al.*, *IEEE Trans. Biomedical Engineering*, 53(5), pp. 865–874, May 2006.
[2] K. Nakamura *et al.*, *SPECOM*, 54(1), pp. 134–146, Jan 2012.
[3] 田中宏 *et al.*, 信学技報, 113(76), SP2013-37, pp. 37–42, Jun. 2013.
[4] K. Yu *et al.*, *IEEE Trans. Audio, Speech, and Language*, 19(5), pp. 1071–1079, Jul 2011.
[5] A. Sakurai *et al.*, *ICSLP*, 2, pp. 817–820, Oct 1996.
[6] A. Kain *et al.*, *Proc. ICASSP*, pp. 285–288, May 1998.
[7] T. Toda *et al.*, *IEEE Trans. Audio, Speech, and Language*, 15(8), pp. 2222–2235, Nov 2007.
[8] T. Toda *et al.*, *IEEE Trans. Audio, Speech, and Language*, 20(9), pp. 2505–2517, Nov 2012.
[9] H. Kawahara *et al.*, *SPECOM*, 27(3-4), pp. 187–207, Apr 1999.
[10] H. Doi., *NAIST Doctoral Dissertation*, NAIST-IS-DD1061014, March 2013.