統計的手法に基づく電気音声変換における変換特徴量に関する調査*

☆ 田中 宏，戸田 智基，Graham Neubig，Sakriani Sakti，中村 哲（奈良先端大・情報）

1 はじめに

喉頭摘出者は，多くの場合声帯を摘出しており，音源生成機能の消失による深刻な発声障害を患う．喉頭摘出者の発声を手助けする有用な代替発声法の一つとして，電気式人工喉頭を用いた発声法が挙げられる．本発声法により生成される電気音声は，比較的明瞭性が高い一方で，その声質は電気的・機械的なものになり，自然性は大きく劣化する．この問題に対して，統計的手法に基づく電気音声変換が提案されている [1]．電気音声から通常音声への変換により，自然性は大幅に改善されるが，変換処理における歪みの影響により，明瞭性については若干劣化することが報告されている [2]．本稿では，統計的手法に基づく電気音声変換において明瞭性の劣化を引き起こす変換特徴量の使用を回避するために，雑音抑圧に基づく電気音声強調法 [3] とのハイブリッド変換法を提案し，その有効性を評価する．

2 電気音声の音響的特徴

電気式人工喉頭を用いた発声では，外部から音源が与えられる．自然な音源信号を機械的に生成するのは極めて困難であり，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となり，音源の周期性は一貫して強くなる．また，波形パワーは，無声子音等では大きく変化するが，有声音では変化が小さくなる．なお，発声区間のみ電気式人工喉頭のスイッチを ON にするため，パワー等により，有音区間と無音区間の区別をするのは比較的容易である．

喉頭摘出者の声帯以外の調音器官に関しては，正常に機能する場合が多い．その際，電気音声のスペクトル包絡に関しては，健常者の場合と同様，個々の音韻に応じて滑らかに変化する．ただし，電気式人工喉頭の音源信号自体が空気中に漏れ出すため，雑音として電気音声に混入する．結果，スペクトル包絡に関しても，健常者のものとは異なる特徴を持つ．

3 電気音声強調に関する従来法

3.1 雑音抑圧に基づく電気音声強調

電気音声に雑音として混入する電気式人工喉頭の音源信号を除去するために，適応フィルタによる雑音抑圧 [4] やスペクトル減算処理（spectral subtraction: SS）[5] による雑音抑圧 [3] を用いる手法が提案されている．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 SS に基づく手法について述べる．

単一チャネルの観測信号 o_t は以下で記述される．

$$o_t = s_t + d_t \tag{1}$$

ここで，s_t は音声信号，d_t は雑音信号を表す．短時間離散フーリエ変換による時間周波数領域においては，

$$O_{k,m} = S_{k,m} + D_{k,m} \tag{2}$$

と表現される．ここで，k は周波数を，m は時間フレームを表すインデックスである．SS 法では，雑音信号の定常性を仮定し，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を次式にて求める．

$$Y_{k,m} = (|O_{k,m}| - \beta E_m[|\hat{D}_{k,m}|]) \cdot e^{j\arg(O_{k,m})} \tag{3}$$

$$Y_{k,m} = \begin{cases} Y_{k,m} & (\text{if } |Y_{k,m}| > 0) \\ 0 & (\text{otherwise}) \end{cases} \tag{4}$$

ここで，$Y_{k,m}$ は音声強調された信号であり，β は処理強度を制御するパラメータ，$E_m[\cdot]$ は m に関する期待値演算である．式 (4) はフロアリングと呼ばれる補正処理である．通常，無音区間の雑音の振幅スペクトル $\hat{D}_{k,m}$ 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として利用する．一方で，電気音声強調においては，予め雑音信号を収録しておくことが可能である．本稿では，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．その際には，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，SS で使用する．

3.2 統計的手法に基づく電気音声変換

統計的手法に基づく電気音声変換 [1] では，電気音声のスペクトル特徴量（メルケプストラムセグメント）から，通常音声のスペクトル特徴量（メルケプストラム）および音源特徴量（対数 F_0 ／無声シンボルと非周期成分）への変換を行う．本枠組みは学習処理と変換処理で構成される．

学習処理では，電気音声と通常音声の同一発話データ（パラレルデータ）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 フレームから計算される入力特徴量を X_t とし，出力静的・動的特徴量を，$\boldsymbol{Y}_t = \left[\begin{array}{cc} \boldsymbol{y}_t^T, \triangle \boldsymbol{y}_t^T \end{array} \right]^T$ とする．パラレルデータに対して動的時間伸縮を行い，対応付けを行った結合ベクトル $\left[\begin{array}{cc} \boldsymbol{X}_t^T, \boldsymbol{Y}_t^T \end{array} \right]^T$ 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．

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

ここで，$\mathcal{N}\left(\cdot ; \boldsymbol{\mu}, \boldsymbol{\Sigma} \right)$ 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．混合数 M の GMM のモデルパラメータセット $\boldsymbol{\lambda}^{(X,Y)}$ は，各分布 m の混合重み $\boldsymbol{\alpha}_m$，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．電気音声の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，非周期成分との間において，計 3 つの GMM を学習する．

変換処理では，個々の GMM を用いて，最尤系列変換法 [6] により，電気音声の入力特徴量系列から自然音声の各出力特徴量系列へと変換する．得られた

---

*Investigation of converted acoustic features in statistical electrolaryngeal speech conversion, by TANAKA, Kou, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani, and NAKAMURA, Satoshi (Nara Institute of Science and Technology)

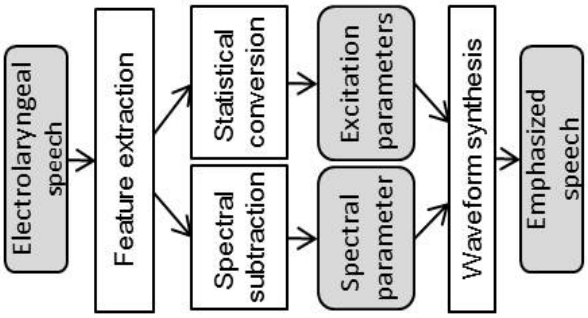

Fig. 1　提案法による電気音声強調処理の流れ

出力特徴量系列から音声波形を合成することで，強調音声を得る．

4　雑音抑圧及び統計的音源生成に基づくハイブリッド電気音声変換

統計的手法に基づく電気音声変換では，通常音声の統計量に基づいて各音声パラメータに対する変換特徴量を生成することができるため，得られる変換音声の特徴は通常音声のものに類似する．一方で，変換処理で生じるひずみの影響を完全に回避することは困難であり，少なからずひずみを含む変換特徴量を使用することになる．その結果，現状の変換精度では，変換前の電気音声と比較し，明瞭性が劣化する．

この問題に対して，明瞭性の劣化は主にスペクトル特徴量のひずみにより生じると予想し，統計的電気音声変換処理において，雑音抑圧に基づく電気音声強調で得られるスペクトル特徴量を使用するハイブリッド方式を提案する．提案法の処理を Fig. 1 に示す．提案法では，電気音声の自然性を大きく低下させる主要因である音源信号に関しては，統計的電気音声変換で得られる音源特徴量を用いることで，自然性を改善する．一方で，スペクトル特徴量に関しては，喉頭摘出者も調音器官は正常に機能するという点や，比較的高い明瞭性が電気音声で得られるという点に着目し，電気音声のものを最大限に活用する．

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，評価データとして他の 10 文を用いる．サンプリング周波数は 16 kHz とする．メルケプストラムセグメント特徴量として 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 [7] を用いる．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（スペクトル変換用），16（F_0 変換用），16（非周期成分変換用）とし，特定話者モデルを用いる．

客観評価実験により，統計的手法に基づく電気音声変換における変換精度を明らかにする．また，主観評価実験により，以下の 4 種類の音声に対して，明瞭性と自然性を 5 段階オピニオン評定により評価する．

- 電気音声（EL）
- 雑音抑圧に基づく強調音声（SS）
- 統計的手法に基づく強調音声（VC）
- 雑音抑圧と統計的音源生成のハイブリッド法に基づく強調音声（SS+VC）

Table 1　統計的手法に基づく電気音声変換の変換精度

Mel-cepstral distortion	4.3 dB
U/V error rate	12.6 %
F_0 correlation coefficient	0.33
Aperiodic distortion	3.2 dB

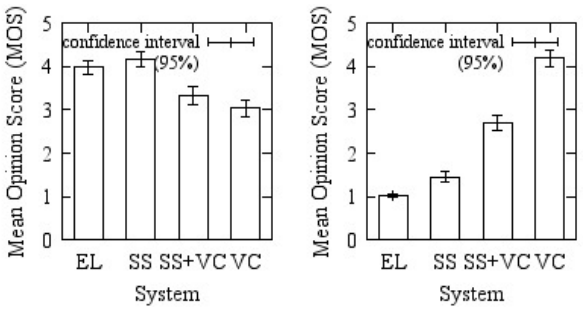

Fig. 2　主観評価実験結果（左図が明瞭性，右図が自然性に関する評価結果）．

被験者は男性 6 名であり，1 人あたり各システムにつき 20 サンプルの計 80 サンプルを受聴する．

5.2　実験結果

表 1 に統計的手法に基づく電気音声変換の変換精度を示す．また，図 2 に主観評価実験結果を示す．明瞭性に関しては，EL と比較して，SS では改善効果が得られるが，VC では逆に劣化する．提案法である SS+VC では，SS によるスペクトル特徴量を用いることで，その劣化を低減させることができる．ただし，EL と比べると依然として劣化している．この原因として，10 %以上存在する有声無声誤りの影響が考えられる．一方で，自然性に関しては，VC が最も高くなる．SS+VC では，音源特徴量変換の効果により，SS 単体よりも大きな改善効果が得られる．

6　終わりに

本稿では，統計的手法に基づく電気音声変換における明瞭性の劣化を低減させるために，雑音抑圧処理により得られるスペクトル特徴量と，統計的手法により得られる音源特徴量を併用したハイブリッド方式を提案した．実験的評価の結果，提案法は電気音声の自然性を大幅に改善しつつ，明瞭性の劣化を低減できることが分かった．一方で，未だ電気音声と比較すると明瞭性の劣化が生じているため，今後さらなる検討が必要である．

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．

参考文献

[1] K. Nakamura *et al.*, *Speech Communication*, 54(1), pp. 134–146, Jan. 2012.
[2] H. Doi *et al.*, *Proc. ICASSP*, pp. 5136–5139, Prague, Czech Republic, May. 2011.
[3] H. Liu *et al.*, *J. Acoust. Soc. Am.*, 120(1), pp. 398–406, 2006.
[4] C.Y. Espy-Wilson *et al.*, *J. Speech, Lang., Hear. Res.*, 41(6), pp. 1253–1264, 1998.
[5] S. F. Boll *et al.*, *IEEE Trans. Acoustic, Speech, Signal Proc.*, No.2, pp. 113–120, 1979.
[6] T. Toda *et al.*, *IEEE Trans. Audio, Speech, Lang Process Proc.*, 15(8), pp. 2222–2235, 2007.
[7] H. Kawahara *et al.*, *Speech Communication*, 27(3-4), pp. 187–207, 1999.