Vol.2010-SLP-82 No.2
2010/7/22

1 対多固有声変換に基づく無喉頭音声の音質及び話者性の改善

土井 啓成[†1] 中村 圭吾[†1] 戸田 智基[†1]
猿渡 洋[†1] 鹿野 清宏[†1]

喉頭摘出者は自身の声帯振動を用いた発声が不可能であるため，代用発声法で発声を行う．代用発声法により，喉頭摘出者は発声が可能になるが，生成される無喉頭音声は，健常者の通常音声と比較して，音質が低く，話者性も劣化してしまう．そのため本稿では，統計的手法による無喉頭音声の音質及び話者性の改善を試みる．無喉頭音声の音質改善には，統計的声質変換に基づく音質改善法 (AL-to-Spech) がこれまでに提案されている．本稿では，AL-to-Speech に対して，変換音声の声質を柔軟に制御することができる 1 対多固有声変換を導入することで，音質及び話者性の改善を行う．客観評価及び主観評価結果から，提案法が高い音質改善効果を持ちつつ，声質の制御が可能であることを示し，無喉頭音声の音声及び話者性の改善に有効であることを示す．

Improvement of sound quality and speaker individuality for alaryngeal speech based on one-to-many eigenvoice conversion

Hironori Doi,[†1] Keigo Nakamura,[†1] Tomoki Toda,[†1]
Hiroshi Saruwatari [†1] and Kiyohiro Shikano[†1]

This paper proposes the improvement method based on one-to-many eigenvoice conversion (EVC) for sound quality and speaker individuality of three types of alaryngeal speech: esophageal speech; electrolaryngeal speech; and body-conducted silent electrolaryngeal speech. Although alaryngeal speech allows laryngectomees to utter speech sounds, it suffers from lack of naturalness and speaker individuality. To improve the sound quality of alaryngeal speech, alaryngeal-speech-to-speech (AL-to-Speech) methods based on statistical voice conversion have been proposed. This paper further applies one-to-many EVC capable of flexibly adapting the conversion model to given target natural voices to the AL-to-Speech methods for recovering speaker individuality of alaryngeal speech. The experimental results of objective and subjective evaluations demonstrate that the proposed methods yield significant improvements of speech quality and make the converted voice quality similar to the given target voice quality.

1. はじめに

事故や喉頭癌等の病気で喉頭を摘出した喉頭摘出者（喉摘者）は，喉頭と共に声帯を失うため，自身の声帯振動を利用した発声が不可能になる．そのため，喉摘者は，声帯の代わりとなる器官や機器を用いて音源を生成する代用発声法により発声を行う．尚，代用発声法により生成される音声は，無喉頭音声と呼ばれる．

日本で広く使用されている代用発声法に食道発声法と電気発声法があり，それらで発声される無喉頭音声はそれぞれ，食道音声，電気音声と呼ばれる．また，近年，微弱音源と呼ばれるパワーの小さい電気式人工喉頭を用いて電気音声を生成し，それを NAM マイクロフォン[1] で収録する手法も提案されている[2]．尚，本稿では，微弱音源を用いた電気発声法で生成される音声を微弱電気音声と呼ぶこととする．

これらの無喉頭音声は，自然性や明瞭性，または利便性等において，それぞれ利点を有しているが，健常者の発声する通常音声と比較すると，その音質は総じて低く，話者に寄らず似たような声質を持つ．これらの問題は，喉摘者の社会復帰に対する大きな妨げとなっており，古くから様々な対策が講じられてきた．近年では，上記 3 種類の無喉頭音声に対して，統計的な音質改善法が提案され，その有効性が示されている．この手法は，統計的声質変換 (statistical voice conversion: VC)[3)–5)] に基づき，無喉頭音声を通常音声に変換することで音質改善を図るものであり，本稿では，VC に基づく Alryngeal speech-to-speech (AL-to-Speech)[2),6),7)] と呼ぶ．

VC に基づく AL-to-Speech は，学習部と変換部から成る．学習部では，無喉頭音声と通常音声の多数の同一内容発話から作成したパラレルデータセットを用いて，両音声の音響特徴量間の対応関係を学習する．この対応関係は，両音声特徴量の結合確率密度を表す混合正規分布 (Gaussian Mixture Model: GMM) でモデル化される．変換部では，学習された GMM を用いて，最尤基準により，無喉頭音声の音響特徴量から通常音声の音響特徴量へ，音韻情報を保ったまま変換する．変換された音響特徴量は，通常音声の統計量に基づき生成されるため，通常音声に近い音質を得る．VC に基づく AL-to-Speech において，喉頭摘出

†1 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of information Science, Nara Institute of Science and Technology

Vol.2010-SLP-82 No.2
2010/7/22

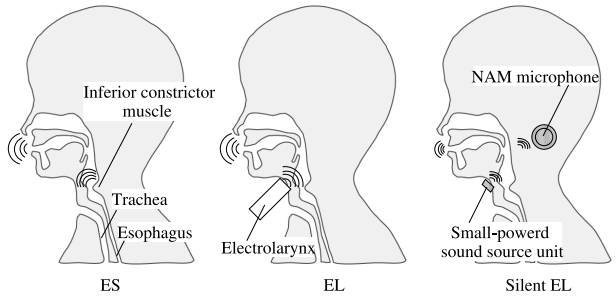

図 1 無喉頭音声 (食道音声:ES，電気音声:EL，微弱電気音声:silent EL) の発声過程
Fig. 1 Speaking methods for producing three types of alaryngeal speech (ES, EL, and silent EL).

以前の通常音声を再現するためには，その音声への変換モデルを学習する必要があるが，学習に十分な量の通常音声が保持されていることは稀であり，話者性の改善はしばしば困難である．

そこで本稿では，AL-to-Speech に対し，1 対多固有声変換 (Eigenvoice Conversion: EVC)[8] を導入することで，話者性の改善を行う．1 対多 EVC は，ある話者の声質を任意の話者の声質へと変換する手法であり，話者性を表すパラメータを手動操作，もしくは，少量の目標音声を用いて自動推定することにより，自由な声質制御を可能にする．したがって，喉頭摘出以前の通常音声を少量でも保持していれば，かつての声質を再現できる．また，喉頭摘出以前の通常音声が保持されていない場合でも，声質を手動制御することにより，話者独自の声質を獲得することが可能である．これまでに，食道音声に対しては，1 対多 EVC に基づく音質及び話者性の改善が提案されており，その有効性が示されている[9]．本稿では，食道音声に加えて，電気音声および微弱電気音声に対しても，1 対多 EVC に基づく音質及び話者性の改善を行う．無喉頭音声に対する 1 対多 EVC に基づく音質及び話者性の改善法を総じて 1 対多 EVC に基づく AL-to-Speech と呼ぶ．喉頭摘出以前の音声が利用可能な場合を想定し，提案法の有効性を主観的及び客観的に評価する．

2. 代用発声法

本節では，本稿で扱う 3 つの代用発声法について概説する．図 1 に各代用発声法を示す．

2.1 食 道 発 声

食道発声は，術後の残存器官を用いて音源を生成する手法の一つである．食道発声では，鼻腔あるいは口腔から頸部食道内に取り込んだ空気を逆流させて気流を生成し，残された下咽頭食道接合部の粘膜を振動させることにより音源を得る．その際，下咽頭収縮筋を緊張・緩和させることで，ある程度音源を制御することが可能である．そのため，食道発声は，音源の自然性及び調節性に優れ，発声と構音操作の協調を得やすく，代用発声法の中では，比較的高音質な音声を生成できる．しかしながら，音源の生成過程で，雑音の様に聞こえる独特の音が混入してしまうため，通常音声の音質には及ばない．食道発声は，外部機器を用いる必要がない反面，気流生成等が難しく，その習得には努力と時間を要する．

2.2 電 気 発 声

電気発声は，電気式人工喉頭と呼ばれる医療機器を用いて，音源を生成する手法である．電気式人工喉頭は，電気エネルギーにより，内部の振動子を振動させることで外部音源を発生させる．これを手で前頸部の皮膚に密着させ，振動を前頸部から軟部組織を介して咽頭に伝え，粘膜を振動させることで音源が生成される．電気発声は機器を使用するため，習得が容易であるが，反面，発声時には手がふさがるという欠点もある．また，振動の周波数が一定に固定されてしまうため，機械的で単調な音声に聞こえる．さらに，電気発声では，周りに聞こえるだけの音量を持つ電気音声を生成するため，十分な音量の音源を用いるが，その音源自体が周囲への雑音となってコミュニケーションの妨げとなることもある．

2.3 微弱音源を用いた電気発声

電気式人工喉頭で生成される音源そのものが雑音となる問題を緩和すべく，周囲に聴取されない程微弱な音源を用いた電気発声が提案されている．本手法では，微弱音源を用いて生成された音声を筋肉を介して NAM マイクロフォンで収録し，増幅した後に再生することで，周囲が聴取可能な音声を生成する．NAM マイクロフォンは，肉伝道音声の収録に特化したマイクロフォンであり，非可聴つぶやきなどの極めてパワーの小さな音声を収録することが可能であるため，筋肉を介した微弱電気音声の収録に適している．微弱音源を用いた発声法では，調音された微弱電気音声のみを周囲に伝えることが可能なため，音源自体の音を不快に感じることはない．しかしながら，筋肉を介して収録するため，口唇による放射特性の欠落や体内伝導による高域減衰特性などの影響により，音質は電気音声よりも劣化する．

3. 1 対多固有声変換（1 対多 EVC）

1 対多 EVC は，ある特定の話者の音声を任意の話者の音声へ変換する手法であり，両話者間の対応関係を固有声 GMM(Eigenvoice GMM: EV-GMM) でモデル化する．

3.1 Eigenvoice GMM: EV-GMM

時刻 t における元話者と任意の目標話者の静的特徴量をそれぞれ $\boldsymbol{x}_t = [x_t(1), \cdots,$

$x_t(D_x)]^\top$，$\boldsymbol{y}_t = [y_t(1), \cdots, y_t(D_y)]^\top$ とする．ここで，両特徴量はそれぞれ D_x, D_y 次元である．$\top$ は転置を表す．入力特徴量として，静的・動的特徴量や複数のフレームを結合した特徴量等の数フレームの時間情報を含む特徴量 $\boldsymbol{X}_t$ を用いる．出力特徴量として，目標話者の静的・動的特徴量 $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ を用いる．この時，時刻 t の入力特徴量 $\boldsymbol{X}_t$ と出力特徴量 $\boldsymbol{Y}_t$ の結合確率密度は，EV-GMM によりモデル化される．

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

$$\boldsymbol{\mu}_m^{(X,Y)}(\boldsymbol{w}) = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}(\boldsymbol{w}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{A}_m\boldsymbol{w} + \boldsymbol{b}_m \end{bmatrix} \tag{2}$$

$$\boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \tag{3}$$

ここで，$\boldsymbol{\lambda}^{(EV)}$ は規範 EV-GMM のパラメータであり，第 m 分布の分布重み α_m，入力平均ベクトル $\boldsymbol{\mu}_m^{(X)}$，共分散行列 $\boldsymbol{\Sigma}_m^{(X,Y)}$，$J$ 個の固有ベクトルから成る行列 $\boldsymbol{A}_m$，及びバイアスベクトル $\boldsymbol{b}_m$ で構成される．$\boldsymbol{w} = [w(1), \cdots, w(J)]^\top$ は，変換音声の声質を制御する話者依存重みパラメータであり，$\boldsymbol{w}$ を手動操作もしくは目標音声を用いて自動推定することで，EV-GMM の出力音声を任意の音声に適応させる．従って，EV-GMM は，話者依存の重みパラメータ及び，話者非依存な規範 EV-GMM パラメータから成る．

3.2 EV-GMM の話者適応学習[10)]

本稿では，話者適応学習法 (Speaker Adaptive Training: SAT) により，EV-GMM の変換性能の改善を行う．SAT では，規範 EV-GMM 及び事前学習用話者の話者依存重みパラメータを同時に更新する．

$$\{\hat{\boldsymbol{\lambda}}^{(EV)}, \hat{\boldsymbol{\omega}}_{(1:S)}\} = \underset{\boldsymbol{\lambda}^{(EV)}, \boldsymbol{\omega}_{(1:S)}}{\operatorname{argmax}} \prod_{s=1}^{S}\prod_{t=1}^{T_s} P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{\omega}_s) \tag{4}$$

ここで，$\hat{\boldsymbol{\lambda}}^{(EV)}$ は更新された規範 EV-GMM パラメータセットであり，$\hat{\boldsymbol{\omega}}_{(1:S)} = \{\hat{\boldsymbol{\omega}}_1, \cdots, \hat{\boldsymbol{\omega}}_S\}$ は，S 人の事前学習用話者の更新された話者依存重みパラメータを表す．また，$\boldsymbol{Y}_t^{(s)}$ は，時刻 t における s 番目の事前学習用話者の出力特徴量ベクトルである．

3.3 EV-GMM の話者適応

話者依存重みパラメータ $\boldsymbol{w}$ は，目標とする話者の任意の発話を用いて，次式により自動的に推定することができる．

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \prod_{t=1}^{T} \int P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(tar)}|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) d\boldsymbol{X}_t \tag{5}$$

ここで，$\{\boldsymbol{Y}_1^{(tar)}, \cdots, \boldsymbol{Y}_T^{(tar)}\}$ は，与えられた目標話者の音響特徴量の時系列データである．

3.4 EV-GMM による変換

入力及び出力特徴量の時系列をそれぞれ $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$，$\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$ とする．変換された特徴量の時系列 $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ は，$\boldsymbol{X}$ が与えられた際の $\boldsymbol{Y}$ の条件付き確率密度の最大化によって求める．

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) \qquad \text{subject to}\, \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{6}$$

ここで，$\boldsymbol{W}$ は静的特徴量から静的・動的特徴量への変換行列である．変換音声の系列内変動 (Global Variance: GV)[5)] を考慮して変換を行うことで，変換音声の品質をさらに改善することが可能である．

4. 1 対多 EVC に基づく AL-to-Speech

無喉頭音声の音質及び話者性改善のため，本稿では，1 対多 EVC に基づく AL-to-Speech を提案する．提案法は，少量の喉頭摘出以前の通常音声を用いて，EV-GMM を適応させることで，喉摘者のかつての音声を再現する．また，喉頭摘出以前の音声が保持されていない場合でも，話者依存重みパラメータを手動で操作することで，話者のかつての音声の再現，もしくは，話者独自の音声の生成が期待できる．

無喉頭音声から抽出されるスペクトル，非周期成分[11)]，F_0 の 3 つの特徴量の内，有用な情報を含む特徴量は，スペクトルのみである．そのため，AL-to-Speech では，無喉頭音声のスペクトルから通常音声のスペクトル，非周期成分，F_0 をそれぞれ推定する．ただし，無喉頭音声のスペクトルは通常音声と比較すると情報量が乏しいため，より幅広い時間の情報を用いて音韻情報の補完を行うスペクトルセグメント特徴量を入力特徴量として用いる．スペクトルセグメント特徴量は次式により定義される．

$$\boldsymbol{X}'_t = \boldsymbol{C}(\boldsymbol{X}_t - \boldsymbol{d}) \tag{7}$$

ここで，$\boldsymbol{X}_t = [\boldsymbol{x}_{t-i}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+i}^\top]^\top$ は時刻 t 及び前後 i フレームにおけるスペクトル特徴量ベクトル $\boldsymbol{x}_t$ を結合したベクトルである．この結合ベクトルに対し，PCA による次元圧縮を行い，スペクトルセグメント特徴量を抽出する．$\boldsymbol{C}$ は PCA により求まった固有ベクトルで構成される変換行列，$\boldsymbol{d}$ は平均ベクトルである．これまでに，食道音声の変換に対して，スペクトルセグメント特徴量の有効性が示されている[7)]．

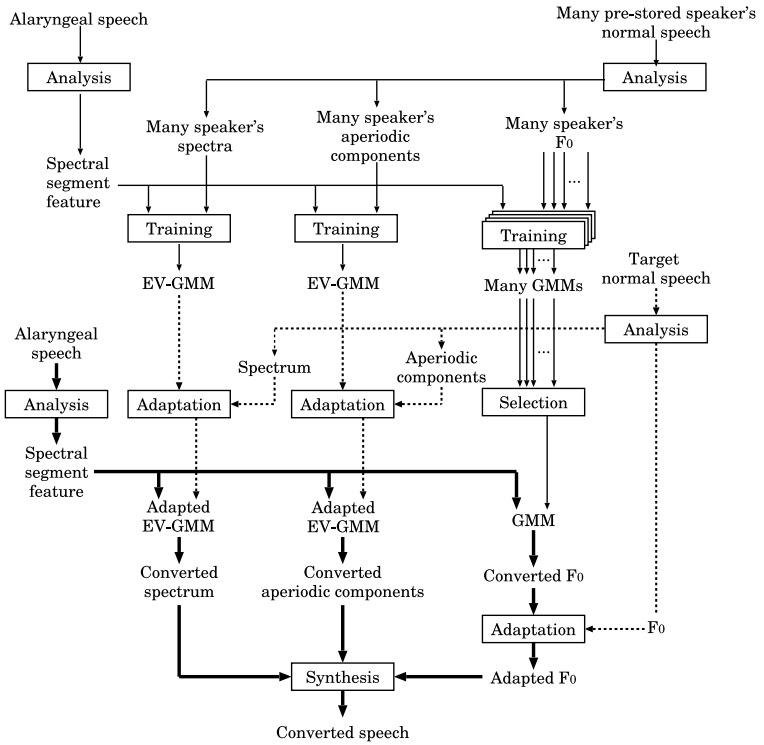

図 2 1 対多 EVC に基づく AL-to-Speech の処理過程．細実線は学習処理を，点線は適応処理を，太実線は変換処理をそれぞれ表す．

Fig. 2 Process of AL-to-Speech based on one-to-many EVC. fine line, dashed line and bold line represents training, adaptation and conversion process, respectively.

本稿で提案する 1 対多 EVC に基づく AL-to-Speech は，学習部，適応部，及び変換部からなる．図 2 に提案法の処理を示す．学習部では，スペクトル推定 EV-GMM，非周期成分推定 EV-GMM，及び F_0 推定 GMM をそれぞれ独立に学習する．スペクトル推定 EV-GMM と非周期成分推定 EV-GMM は，1 人の喉摘者の無喉頭音声から抽出されたスペクトルセグメント特徴量と，複数の事前学習用話者の通常音声から抽出したスペクトルもしくは非周期成分から成るパラレルデータセットを用いて，それぞれ独立に学習する．また，無喉頭音声のスペクトルセグメント特徴量から各事前学習用話者の通常音声の F_0 への変換 GMM を個別に学習し，得られた多数の話者依存 F_0 推定 GMM の中から，最も自然な変換 F_0 が得られる GMM を選択し，提案法における F_0 推定モデルとして用いる．適応部では，目標とする通常音声のスペクトルと非周期成分から，それぞれ独立に推定した話者依存重みパラメータを用いて，スペクトル推定 EV-GMM 及び非周期成分 EV-GMM の適応を行う．変換部では，適応 EV-GMM を用いて，スペクトルと非周期成分をそれぞれ独立に推定する．また，F_0 変換では，GMM を用いて推定した対数 F_0 が，目標とする通常音声の対数 F_0 の平均 μ_x と分散 σ_x を持つように，次式により適応する．

$$\log y_t = \frac{\sigma_y}{\sigma_x}(\log x_t - \mu_x) + \mu_y \tag{8}$$

ここで，x_t と y_t はそれぞれ，時刻 t における，GMM で推定された F_0 及び，目標音声に適応した F_0 である．

5. 実験による評価

1 対多 EVC に基づく AL-to-Speech の有効性を客観的及び主観的に評価する．

5.1 実 験 条 件

元音声として，男性喉摘者 1 名の食道音声，及び，別の男性喉摘者 1 名の電気音声と微弱電気音声をそれぞれ収録する．また，健常者男性 27 名，女性 13 名の計 40 名の通常音声をそれぞれ収録する．発話内容は全話者同一の音素バランス文 50 文である．収録した 40 名分の通常音声の内，30 名分 (男性 22 名，女性 8 名) を EV-GMM の学習に，残り 10 名分 (男性 5 名，女性 5 名) を評価用の目標音声に使用する．また，話者毎に収録された 50 文の発話の内，40 文を学習または適応に，残り 10 文を評価に用いる．この時，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．無喉頭音声に対しては，メルケプストラム分析[12] を，通常音声に対しては，STRAIGHT 分析[13] をそれぞれ用いる．この時，シフト長は 5 ms とする．食道音声のスペクトルセグメント特徴量は，スペクトル推定及び非周期成分推定時においては当該フレーム及び ±8 フレームを用いて生成し，F_0 推定においては当該フレーム及び ±16 フレームを用いて生成する．電気音声及び微弱電気音声のスペクトルセグメント特徴量は，どの特徴量推定においても，当該フレーム及び ±8 フレームを用いて生成する．スペクトルセグメント特徴量の次元数は 50 とする．また，音源特徴量として，STRAIGHT[14] によって抽出された対数 F_0 と 5 帯域 (0-1, 1-2, 2-4, 4-6, 6-8 kHz) の非周期成分を用いる．

食道音声，電気音声及び微弱電気音声のスペクトル変換と非周期成分変換のために，喉摘者 1 名の無喉頭音声と健常者 30 名の通常音声を用いて EV-GMM をそれぞれ学習する．こ

の時，各 EV-GMM の混合数は 64，固有ベクトル数は 29 とする．目標話者 10 名の通常音声 1，2，4，8，16，32 文を用いて，目標話者毎に各 EV-GMM をそれぞれ適応する．さらに，従来法として VC に基づく AL-to-Speech を行う．VC に基づく AL-to-Speech では，喉摘者 1 名の無喉頭音声と健常者 1 名の通常音声を用いて，スペクトル推定用及び非周期成分推定用 GMM をそれぞれ学習する．この時，1，2，4，8，16，32 文の発話対を用いて GMM をそれぞれ学習する．その際に，GMM の混合数は，各学習データ量に応じて事後的に最適化する．

5.2 客観評価実験

客観評価では，従来法と提案法によって生成された変換音声と目標通常音声のメルケプストラムひずみ及び非周期成分ひずみを計測し，各 AL-to-Speech の性能を評価する．この時，メルケプストラムひずみは，パワーを除いた 1 次から 24 次のメルケプストラム係数から計算する．

図 3 に，提案法と従来法のメルケプストラムひずみを，図 4 に非周期成分ひずみを示す．これらの図において，ほとんどの無喉頭音声に対し，提案法は，1 文適応で，16 文で学習した従来法と同等以上の変換性能を示している．また，従来法では変換モデルの学習にパラレルデータが必要であるのに対し，提案法では任意の目標音声サンプルを用意するだけでよいことを考慮すれば，提案法は従来法よりも高い有用性を持つと言える．これらのことから，提案法は高い変換精度を保ちつつ，容易に声質制御が行え，喉頭摘出以前の通常音声の再現に対し，非常に有効であることが分かる．

5.3 主観評価実験

主観評価では，無喉頭音声，従来法で生成した変換音声, 提案法で生成した変換音声の音質を，5 段階の平均オピニオン評定で評価する．この時，従来法では 32 文の発話対で学習した GMM を用い，提案法では，目標通常音声 1 文で適応された EV-GMM を用いて変換を行う．被験者は，防音室内にてヘッドフォン両耳受聴により，各代用発声法における上記 3 種類の音声（計 9 種類）を評価する．被験者は，男性 8 名，女性 2 名の計 10 名である．各被験者は，それぞれ 135 サンプルの音声を評価する．

図 5 に主観評価結果を示す．全ての無喉頭音声において，従来法及び提案法はその変換元の無喉頭音声よりも高い音質を示している．特に，食道音声と電気音声では，提案法による変換音声は，非常に高い音質を示しており，その有効性が窺える．微弱電気音声では，提案法で生成された変換音声の声質は，他の変換音声と比較し低い．それでも，どの無喉頭音声よりも高い音質を示しており，微弱電気音声が，変換音声のみを周囲に聴かせることが

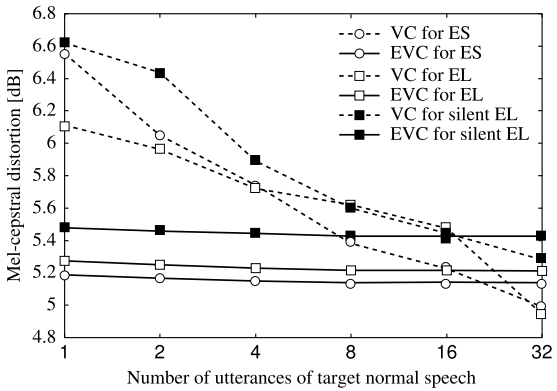

図 3 メルケプストラムひずみ．横軸は VC における学習分数及び EVC における適応分数を表す
Fig. 3 Mel-cepstral distortion as a function of the number of utterances of target normal speech (i.e., utterance-pairs in VC or adaptation utterances in EVC).

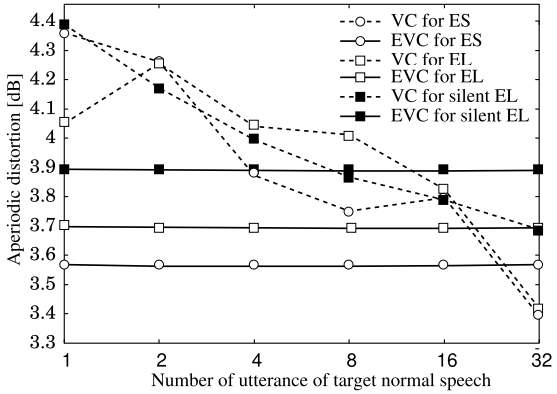

図 4 非周期成分ひずみ．横軸は VC における学習分数及び EVC における適応分数を表す
Fig. 4 Aperiodic distortion as a function of the number of utterances of target normal speech (i.e., utterance-pairs in VC or adaptation utterances in EVC).

可能な音声であることを鑑みれば，対面コミュニケーションにおける有効性は非常に高い．また，図では従来法による変換音声も高い音質を示しているが，モデル学習には 32 文の発話対が必要であり，1 文の目標通常音声で適応可能な提案法の方が，より利便性が高いと言える．これらのことから，提案法は従来法と同様に高い音質改善効果を持ち，その利便性は従来法よりも高いと言える．

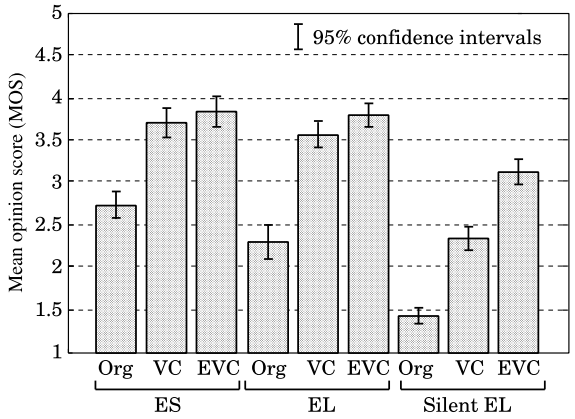

図 5 音質に関する主観評価結果．“Org” は無喉頭音声，“VC” は VC に基づく AL-to-Speech による変換音声，“EVC” は 1 対多 EVC に基づく AL-to-Speech による変換音声を表す

Fig. 5 Result of opinion test of speech quality. “Org”, “VC”, and “EVC” show original alaryngeal speech, converted speech by the conventional method, and converted speech by the proposed method, respectively.

6. ま と め

本稿では，食道音声，電気音声，微弱電気音声の 3 種の無喉頭音声の音質及び話者性の改善のために，1 対多 EVC に基づく AL-to-Speech を提案した．提案法は，特定の喉摘者の無喉頭音声から，任意の健常者の通常音声への変換を可能にする手法であり，手動もしくは少量の目標音声を用いて，変換モデルを適応することにより，声質の制御を行う．そのため，喉頭摘出以前の通常音声が少量でも保持されていれば，喉摘者自身のかつての音声を再現することが期待できる．客観評価及び主観評価結果から，提案法は，音質改善を行いつつ，容易に声質制御が行えることが分かった．今後，明瞭性や声質再現精度などに関して，さらなる評価実験を行う必要がある．

謝辞 本研究の一部は，総務省 SCOPE により実施したものである．STRAIGHT の使用を許可して頂いた和歌山大学河原英紀教授に感謝いたします．

参 考 文 献

1) Y.Nakajima, H.Kashioka, K.Shikano, N.Campbell, ”Remodeling of the sensor for Non-Audible Murmur (NAM),” *INTERSPEECH*, pp.389–392, September 2005.
2) K. Nakamura, T. Toda, Y. Nakajima, H. Saruwatari and K. Shikano, “Evaluation of speaking-aid system with voice conversion for laryngectomees toward its use in practical environments,” *INTERSPEECH*, pp.2209–2212, Sep, 2008.
3) 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, 1998.
4) A.Kain and M.W.Macon, “Spectral voice conversion for text-to-speech synthesis,” *Proc. ICASSP*, pp. 285–288, Seattle, USA, May 1998.
5) T. Toda, A.W. Black, and K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” *IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2235, Nov. 2007.
6) K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, “Electrolaryngeal speech enhancement based on statistical voice conversion,” *INTERSPEECH*, pp. 1431–1434, Brighton, UK, Sep. 2009.
7) H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Enhancement of esophageal speech using statistical voice conversion,” *APSIPA*, pp. 805–808, Sapporo, Japan, Oct. 2009.
8) T. Toda, Y. Ohtani, and K. Shikano, “One-to-many and many-to-one voice conversion based on eigenvoices,” *Proc. ICASSP*, pp. 1249–1252, Hawaii, USA, Apr. 2007.
9) H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, “STATISTICAL APPROACH TO ENHANCING ESOPHAGEAL SPEECH BASED ON GAUSSIAN MIXTURE MODELS,” *Proc. ICASSP*, pp. 4250–4253, Dallas, U.S.A., March 2010.
10) Y.Ohtani, T.Toda, H.Saruwatari, K.Shikano, “Adaptive training for voice conversion based on eigenvoices,” *IEICE Trans. Information and Systems*, vol. E93-D, no. 6, pp.1589–1598, June 2010.
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,” *MAVEBA*, Florence, Italy, Sept. 2001.
12) K.Tokuda, T.Kobayashi, T.Masuko, and S.Imai. “Mel-generalized cepstral analysis – a unified approach to speech spectral estimation,” *Proc. ICSLP*, pp. 1043–1045, Yokohama, Japan, Sep. 1994.
13) 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,” *Speech Communication*, Vol. 27, No. 3-4, pp. 187–207, 1999.
14) H.Kawahara, H.Katayose, A.Cheveigne, and R.D.Patterson, “Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F_0 and periodicity,” *Proc. EUROSPEECH*, pp. 2781–2784, Budapest, Hungary, Sept. 1999.