Vol.2009-SLP-77 No.18
2009/7/18

統計的声質変換を用いた食道発声音声の音質改善

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

喉頭摘出者が行う発声法の一つに，食道等を用いて音源を生成し，発声を行う食道発声法がある．食道発声法は国内で広く使用されているが，生成された音声は健常者の音声と比較して不自然である．本稿では，食道発声音声の音質改善を目指し，食道発声音声から健常者音声への統計的声質変換を用いた音質改善法 (ES-to-Speech) を提案する．健常者音声のスペクトル特徴量や F_0，非周期成分といった音源特徴量は，それぞれ食道発声音声のスペクトル特徴量から独立に推定する．変換音声の客観評価実験及び，主観評価実験結果から，ES-to-Speech は，食道発声と同等の明瞭性を保ったまま，自然性を大きく改善できることを示す．また，F_0 推定時に，入力特徴量として食道発声音声のスペクトルと F_0 の併用も試みることで，食道発声音声の F_0 情報を用いる効果を検証する．

Enhancement of Esophageal Speech Using Statistical Voice Conversion

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

This paper proposes a novel method of enhancing esophageal speech based on statistical voice conversion. Esophageal speech is one of the speaking methods for total laryngectomees to speak by generating sound excitations at their esophagus. Although esophageal speech is the major method in Japan, the generated voices sound unnatural. To improve naturalness of the esophageal speech, we propose a conversion method from esophageal speech to normal speech (ES-to-Speech) using a statistical voice conversion technique. Spectral features and excitation features, such as F_0 and aperiodic components, of the normal speech are independently estimated from the spectral features of the esophageal speech based on the maximum likelihood criterion. The effectiveness of ES-to-Speech is evaluated by conducting objective and subjective experiments to demonstrate that the proposed method yields significant improvements in naturalness of esophageal speech while keeping its intelligibility.

1. はじめに

音声は，人間の最も一般的なコミュニケーション手段の一つとして古くから用いられている．しかし，誰もがその恩恵を受けることができるわけではない．一般に発声障害者と呼ばれる者は音声の生成に何らかの障害を抱えており，音声コミュニケーションにおいて大きな困難を有する．発声障害者の中でも，事故や喉頭癌等によって喉頭を摘出された喉頭摘出者は，喉頭と共に声帯も失うため，自身の声帯振動による音源を用いた発声が不可能である．そのため，喉頭摘出者は健常者とは別の方法で音源を得る必要がある．

喉頭摘出者が行う代表的な発声法の一つに食道発声法がある．これは，胃に空気を飲み込み，吐き出す際に食道入り口付近の粘膜のひだを振動させることにより音源を生成し，口や舌の動きによって調音することで発声を行う手法である．食道発声法は音源を体内で生成するため，電気式人工喉頭のような他の手法とは異なり，生成される音声には肉声感があり，器具を使用せずに発声することが可能である．また，日本では，ボランティア団体による食道発声法の習得支援環境が充実されていることもあり，食道発声法は広く使用されている．一方で，食道発声音声には音声の生成過程で生じる特有の雑音が含まれ，さらに話者の意思通りに抑揚を表現することが困難な場合が多く，明瞭性や自然性が健常者音声と比較して劣ることが多い．これらの問題は話者の習熟度に依存するが，食道発声法の習得には困難を有し，また，多くの場合，上記の問題点に対して完全には解決に至らないというのが現状である．そのため，喉頭摘出者が食道発声音声で満足にスムーズな音声コミュニケーションを行うために，音質改善等の支援技術の構築が望まれている．

食道発声音声の音質改善法としては，櫛形フィルタ[1] や平滑化手法[2] 等を用いて，音響特徴量を修正することにより，食道発声音声特有の雑音の低減や抑揚の改善を行う手法が主流である．しかし，食道発声音声と健常者音声の間にある音響特徴量の違いは非常に複雑であり，また，そもそも食道発声音声の音響特徴量の抽出精度が低いことから，これらの手法で健常者音声と同等の音質を得ることは難しい．

そこで，本稿では統計的声質変換[3),4)] を用いて食道発声音声を健常者音声へ変換（Esophageal-Speech-to-Speech: ES-to-Speech）することで，食道発声音声の音質改善を

†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.2009-SLP-77 No.18
2009/7/18

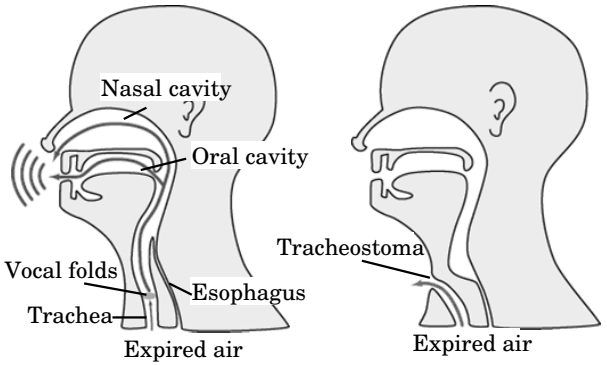

図 1 健常者 (左図) と喉頭摘出者 (右図) の呼気の流れ
Fig. 1 Air flows from lungs in non-laryngectomees (left) and total laryngectomees (right).

目指す．統計的声質変換は，ある話者 (元話者) の音声を別の話者 (目標話者) が発声しているかのように変換する技術であり，言語情報を必要とせずに，発話内容を保ったまま声質のみを変換することが可能である．統計的声質変換は，学習部と変換部から構成される．学習部では，元話者と目標話者の同一内容発話に対して時間方向にアライメントをとったパラレルデータセットを用いて，話者間の音響特徴量の対応関係をモデル化する．変換部では，作成されたモデルを基に変換を行う．本稿で用いる統計的声質変換では，混合正規分布「Gaussian mixture model: GMM」 を用いて，元話者と目標話者の音響特徴量の結合確率密度をモデル化し，そのモデルを用いて最尤基準で元話者の音声特徴量から目標話者の音響特徴量を推定する．本稿では，元話者の音声を食道発声音声，目標話者の音声を健常者音声として，統計的声質変換を適用し，特徴量の修正だけでは到達が困難であった，より健常者に近い音質への変換を目指す．提案手法を客観的及び主観的に評価することにより，提案手法の有効性を示す．

以下，**2**．で食道発声法を，**3**．で統計的声質変換手法を説明し，**4**．で食道発声音声への統計的声質変換の適用 (ES-to-Speech) について述べる．**5**．では，ES-to-Speech の有効性を実験的に評価する．最後に **6**．で本稿をまとめる．

2. 食道発声法

図 1 に健常者と喉頭摘出者の呼気の流れをそれぞれ示す．喉頭には，気管への飲食物の流入を防ぐ働きがあるため，喉頭を摘出した場合，図 1 右図のように気管と食道を完全に分離

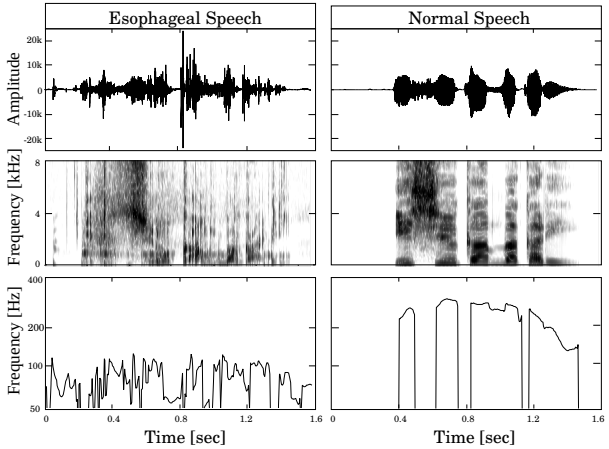

図 2 同一内容の文を食道発声話者と健常者がそれぞれ発声した際の音声波形，スペクトログラム，及び基本周波数 (F_0)
Fig. 2 An example of waveforms, spectrograms and F_0 counters of both esophageal and normal speech.

するのが一般的である．その結果，喉頭摘出者の多くは，声帯振動を失うだけでなく呼気を用いた乱流雑音を得ることすら不可能になり，音源を失くしてしまう．従って，喉頭摘出者が発声するためには，食道発声法等によって，代わりとなる音源を得る必要がある．

食道発声法とは，いわゆるゲップを利用した発声法であり，その発声は次の手順で行われる．まず，呼吸器官のかわりに食道や胃に空気を取り込む．次に，その空気を吐き出す際に，発声器官（喉頭）のかわりに咽頭または食道粘膜を震わせて音源を得る．最後に，健常者と同様に構音器官の動きによって音源を調音し，放射することで発声を行う．

図 2 に，ある同一内容の発話を食道発声話者と健常者がそれぞれ発声した際の音声波形，スペクトログラム，及び基本周波数 (F_0) を示す．ただし，食道発声音声の F_0 の値は **4**．に示す手法で抽出されたものである．図 2 から，食道発声音声は健常者音声と比較して，いくつかの点において大きく異なっていることが分かる．食道発声音声では，本来無音であるはずの箇所においても音が存在し，有音部においてもパワーが不安定である．このことが，食道発声特有の雑音の原因だと考えられる．また，食道発声音声を分析合成した際，音質が著しく低下する．これは，聴感上のピッチに対応した F_0 を抽出することが困難であり，また，有声無声判定の判定精度が低いためである．その他の食道発声音声の特徴としては，無

声摩擦音/h/等のいくつかの音素を発声することが困難であることや F_0 が男女共に低くなるということがあり，これらも健常者音声との大きな違いの一つである．

食道発声音声の音質が健常者音声と比較してどの程度劣化するかは，話者の熟練度に依存する．しかし，食道発声音声特有の雑音をはじめとするいくつかの特徴は食道発声法の原理に起因する問題であり，話者の食道発声技術が上達したからといって必ずしも解決するわけではない．

3. 最尤基準に基づく統計的声質変換

本稿では，ES-to-Speech を実現するために，動的特徴量と発話内変動 (Global Variance: GV)[4] を考慮した GMM に基づく最尤変換を用いる．本手法は GMM の学習部と，それを用いた変換部から構成される．

3.1 GMM の学習手順

まず，元話者と目標話者から同一内容の発話データを収録し，各話者の音響特徴量を時間的に対応づける．時間フレーム t における元話者及び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．さらに，両話者の静的特徴量と動的特徴量を結合した静的・動的特徴量 $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$，$\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ をそれぞれ入力特徴量，出力特徴量として用いる．ここで，$\top$ は転置である．

学習データとしてフレーム毎に対応付けられた両特徴量を用いて，結合確率密度分布 $P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda})$ を GMM でモデル化する[6]．モデル化された結合確率密度分布 $P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda})$ は次式で表される．

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

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \qquad \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \tag{2}$$

ここで，$\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は平均ベクトル $\boldsymbol{\mu}$ と共分散行列 $\boldsymbol{\Sigma}$ から成る正規分布を表し，混合数は M である．また，$\boldsymbol{\lambda}$ は GMM のモデルパラメータを表し，各分布 m に対する重み w_m，平均ベクトル $\boldsymbol{\mu}_m^{(X,Y)}$，全共分散行列 $\boldsymbol{\Sigma}_m^{(X,Y)}$ から成る．

目標話者の静的特徴量系列全体 $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top$ の分散を GV $\boldsymbol{v}(\boldsymbol{y}) = [v(1), \cdots, v(D_y)]^\top$ とし，その確率密度を次式にて表す．

$$P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}) = \mathcal{N}(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(v)}) \tag{3}$$

ここでモデルパラメータ $\boldsymbol{\lambda}^{(v)}$ は，平均ベクトル $\boldsymbol{\mu}^{(v)}$ と対角共分散行列 $\boldsymbol{\Sigma}^{(v)}$ から成る．また，GV の d 次元目の値は次式で算出される．

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t(d) - \frac{1}{T}\sum_{\tau=1}^{T} y_\tau(d)\right)^2 \tag{4}$$

3.2 最 尤 変 換

入力特徴量及び出力特徴量の時系列をそれぞれ $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_t^\top, \cdots, \boldsymbol{X}_T^\top]^\top$，$\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_t^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$ とする．変換された特徴量の時系列 $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ は，$\boldsymbol{X}$ が与えられた際の $\boldsymbol{Y}$ の条件付確率密度と GV の確率密度の積で表される尤度関数の最大化により求められ，次式で表される．

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})^\omega P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}) \tag{5}$$

ここで，$\boldsymbol{Y}$ は，静的特徴量から静的・動的特徴量への変換行列 $\boldsymbol{W}$[7] を用いて，$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$ で表される．また，$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$ と $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$ の比率は重み ω により制御される．本稿では $\omega = \frac{1}{2T}$ とする．

4. 食道発声音声から健常者音声への変換: ES-to-Speech

本稿で提案する ES-to-Speech では，健常者音声の音響特徴量を推定するために，入力として用いる食道発声音声の音響特徴量の抽出は避けられない．しかしながら **2**. で述べた通り，食道発声音声の音響特徴量は不安定であり，通常の特徴量抽出処理では高い推定精度が得られない可能性がある．そこで本稿では，食道発声音声に特化したスペクトル及び F_0 抽出法をを検討する．

4.1 スペクトル抽出

食道発声音声のスペクトルは不安定であり，さらに，原理的に発声できない音素が存在するため，各フレームで得られるスペクトルだけでは，情報量が不足する危険性がある．そこで，本稿では食道発声音声のスペクトルの不安定さを軽減し特徴量の欠落を補うため，スペクトルのセグメント化[5] を行い，より広範囲にわたるスペクトル情報を抽出する．セグメント特徴量は，当該フレームとその前後数フレームを結合し，主成分分析 (Principal component analysis: PCA) により次元圧縮を行うことで抽出する．

4.2 F_0 抽 出

食道発声音声では F_0 の抽出も困難であり，通常の F_0 抽出法では，非常に広範囲にわたって無声区間だと判断され，F_0 が連続的に抽出されない．そこで，本稿では各フレームごとにいくつかの F_0 候補を抽出し，当該フレームの前 5 フレームの平均に最も近い F_0 候補を当該フレームの F_0 とする．

4.3 変換モデル及び入力特徴量

本稿では，ES-to-Speech を実現するために，スペクトル，F_0，非周期成分[8] を推定するための GMM をそれぞれ用いる．ここで，非周期成分とは，音源の各周波数帯域における雑音成分の強さを表す．スペクトルの推定には，食道発声音声のスペクトルセグメント特徴量から健常者音声のスペクトルへと変換する GMM を用いる．F_0 の推定には，食道発声音声のスペクトルセグメント特徴量から健常者音声の F_0 へと変換する GMM を用いる．また，食道発声音声から抽出された F_0 情報の有効性を調査するため，食道発声音声のスペクトルと F_0 を結合し，セグメント化した特徴量を入力とする GMM の使用も試みる．非周期成分の推定には，食道発声音声のスペクトルセグメント特徴量から健常者音声の非周期成分へと変換する GMM を用いる．

スペクトル，F_0，非周期成分が各 GMM から推定された後，これらを用いて音声の合成を行う．まず，推定された F_0 と非周期成分を元に混合励振源[8] を生成する．得られた混合励振源に対して，推定されたスペクトルでフィルタリングを行うことで音声の生成を行う．

5. 実験による評価

ES-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

元話者及び目標話者の音声として，男性喉頭摘出者 1 名の食道発声音声，及び男性健常者 1 名の音声をそれぞれ録音する．発話内容はいずれも同一の ATR 音素バランス文 50 文である．分析によって失われる食道発声音声のピッチ情報を声質変換により再現するために，目標話者は音声収録の際に，食道発声音声を注意深く聞きながら，可能な限り食道発声音声のピッチを模擬して発声する．サンプリング周波数は 16 kHz とする．録音した音声データのうち，40 文を学習用データとし，それ以外の 10 文をテスト用データとして，5 パターンのクロスバリデーションを行う．

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．抽出法には，食道発声音声ではメルケプストラム分析[9] を，健常者音声では STRAIGHT 分析[10] をそれぞれ用いる．また，音源特徴量として，STRAIGHT[11] 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[8] を用いる．フレームシフト長は 5 ms であり，メルケプストラム分析のフレーム長は 25 ms，セグメント特徴量の次元数は 50 である．

客観評価では，F_0 推定に用いる入力特徴量として，食道発声音声のスペクトル，及びスペクトルと F_0 の併用を試み，食道発声音声の F_0 情報を用いる効果を検証する．また，セグメント化がスペクトル変換精度に与える影響及び有効性を見るため，セグメント長を $\pm 2, \pm 4, \pm 8, \pm 16$ フレームとした場合及び，± 1 フレームから計算した静的・動的特徴量を用いた場合の各音素カテゴリー毎のメルケプストラム歪を計測する．メルケプストラム歪は，パワーを除いた 1 次から 24 次のメルケプストラムから計算する．この時，各 GMM の混合数を 32 とする．

主観評価では，ES-to-Speech の結果得られた変換音声の有効性を，明瞭性，自然性の観点からそれぞれ独立に評価する．以下の 5 つの音声に対し，平均オピニオン評定にて評価する．

$-ES$: 食道発声音声 (分析合成されていない収録データ)
$-NS$: 健常者音声 (分析合成されていない収録データ)
$-EstSpq$: 推定されたメルケプストラムと食道発声音声から抽出された F_0 を用いて合成された音声
$-EstF_0$: 食道発声音声から抽出されたメルケプストラムと推定された F_0 を用いて合成された音声
$-EstSpq - EstF_0$: 推定されたメルケプストラムと F_0 を用いて合成された音声

尚，上記 3 つの合成音声に関しては，音源として，推定された非周期成分を用いた混合励振源を使用する．各音声を評価する際には，任意の回数の聞き直しを認める．被験者は日本人成人男女 7 名である．防音室内にてヘッドフォン両耳受聴により実験を行う．この時，各 GMM の混合数を 32，スペクトルと非周期成分の推定に用いるセグメント長を当該 ± 8 フレーム，F_0 の推定に用いるセグメント長を当該 ± 16 フレームとする．

5.2 客観評価結果

表 1 に，変換後の音声と健常者音声の F_0 の相関係数，有声/無声判定エラーを示す．尚，変換前の F_0 と健常者音声の F_0 の相関係数は 0.12 である．F_0 の推定にスペクトルと F_0 の両方を用いても，相関係数ならびに有声/無声判定エラーの値は，スペクトルのみを用いた場合とほとんど変わらない．そのため，以降の実験評価においては，F_0 の推定にはスペクトルのみを使用する．

Vol.2009-SLP-77 No.18
2009/7/18

表 1 F_0 推定精度
Table 1 Estimation accuracy of F_0

Input Feature	Correlation	Voiced/Unvoiced Error [%]
Spectrum	0.68	8.36 ($V \rightarrow U : 4.30, U \rightarrow V : 4.05$)
Spectrum & F_0	0.68	8.39 ($V \rightarrow U : 4.35, U \rightarrow V : 4.04$)

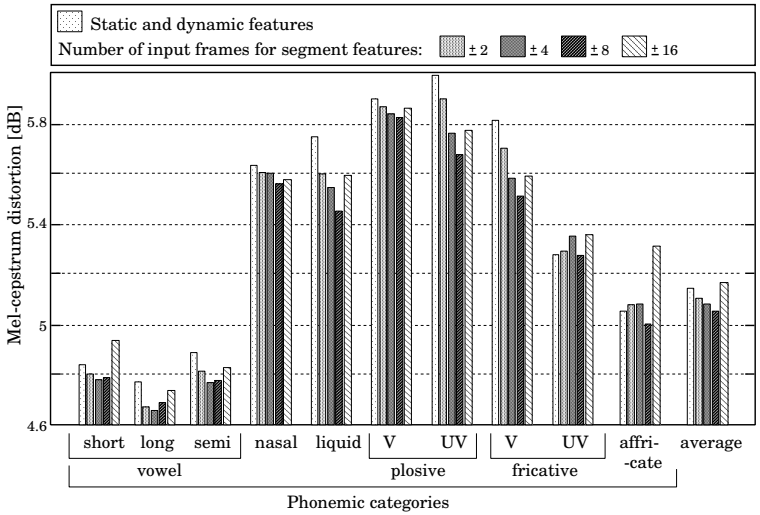

図 3 音素カテゴリー毎のパワー無しメルケプストラム推定精度．‘V” は有声を，“UV” は無声を表す．
Fig. 3 Estimation accuracy of mel-cepstrum without power information in each phoneme category. The notation “V” denotes voiced phonemes, and “UV” denotes unvoiced phonemes.

図 3 に，各音素カテゴリー毎のメルケプストラム歪を示す．スペクトルの推定精度改善において，入力スペクトル特徴量のセグメント化が有効であることが分かる．特に，流音や無声破裂音，有声摩擦音において有効性が顕著に現れている．また，セグメント化に使用するフレーム数が増えるほど，メルケプストラム歪が小さくなる傾向にある．しかしながら，フレーム数を増やしすぎると，PCA で次元圧縮を行った際に欠落する情報量が増加してしまうため，メルケプストラム歪が増加する．

5.3 主観評価結果

5.3.1 明瞭性に関する主観評価

図 4 に，明瞭性の主観評価結果を示す．図 4 より，提案法 (EstSpq-EstF_0) による変換音

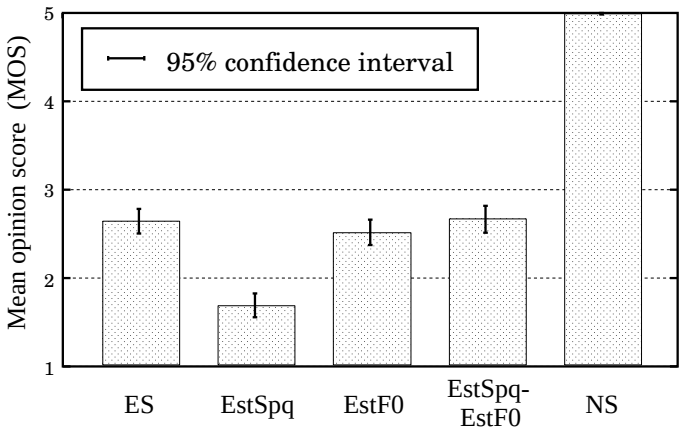

図 4 明瞭性に関する主観評価結果
Fig. 4 Mean opinion score on intelligibility.

声は，食道発声音声 (ES) と同等の明瞭性を保っていることが分かる．また，EstSpq-EstF_0 と EstF_0 を比較した時，スペクトル推定によって，聴感上では食道発声音声特有の雑音が上手く除去されているが，そのことが明瞭性の改善にはほとんど効果をもたらしていないことが分かる．一方，EstSpq の明瞭性が ES と比較して，大きく減少していることが分かる．これは，食道発声音声から F_0 を上手く抽出できないためだと考えられる．以上のことから，食道発声音声の F_0 抽出は困難なものの，F_0 推定を行うことで，元の音声と比較しても遜色のないピッチを生成可能であること，また，声質を変換しても明瞭性は劣化しないことが分かる．

5.3.2 自然性に関する主観評価結果

図 5 に自然性の主観評価結果を示す．ES と EstF_0 を比較すると，F_0 推定することにより，自然性が改善していることが分かる．一方で，EstSpq に注目すると，ES と比べて大きく自然性が減少している．これも明瞭性同様，食道発声音声から F_0 を上手く抽出できないことが原因であると考えられる．また，EstSpq-EstF_0 において，自然性が大きく改善されており，これは，スペクトルの推定により食道発声音声特有の雑音が上手く除去されたためであると考えられる．以上のことから，自然性を改善するためには，F_0 とスペクトル両方の推定が重要であると言える．

これら二つの主観評価の結果から，提案法 (EstSpq-EstF_0) は食道発声音声と同等の明瞭

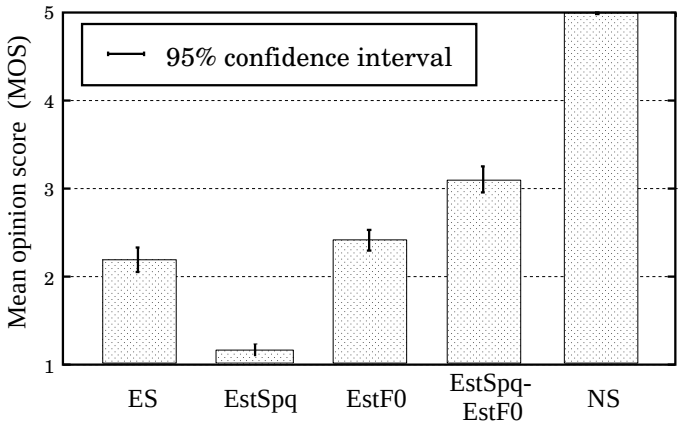

図 5 自然性に関する主観評価結果
Fig. 5 Mean opinion score on naturalness.

性を保ったまま自然性を改善することが可能であるため，食道発声音声の音質改善において非常に有効であると言える．

6. ま と め

本稿では，食道発声音声の新たな音質改善法として，統計的声質変換を用いて食道発声音声を健常者音声に変換する ES-to-Speech を提案した．健常者音声のスペクトル，F_0，非周期成分を，それぞれ別々の GMM を用いて食道発声音声のスペクトルセグメント特徴量から推定し，それらを用いて変換音声を生成した．変換音声の明瞭性と自然性に関する実験結果から，変換音声は食道発声と同等の明瞭性を保ちながら，自然性を大きく改善できることが分かった．

ES-to-Speech の実用化に向けた課題は山積みである．本稿では F_0 推定のための入力特徴量として，スペクトルと F_0 の併用を試みたが，推定精度はスペクトルのみを用いた場合と比べて，大きな違いは見られなかった．今後，入力特徴量として有効な F_0 情報を抽出する方法や，新たな特徴量の併用などを検討していく必要がある．また，本稿では特定の話者への変換を試みたが，今後，喉頭摘出者が好みの声質を得られるような ES-to-Speech の確立が必要である．

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) A. Hisada, H. Sawada. Real-time clarification of esophageal speech using a comb filter. International Conference on Disability, Virtual Reality and Associated Technologies, pp.39–46, 2002.
2) K. Matui, N. Hara, N. Kobayashi, H. Hirose. Enhancement of esophageal speech using formant synthesis. *Proc. ICASSP*, pp. 1831–1834, Phoenix, Arizona, May. 1999
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) 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.
5) T. Toda, K. Shikano, NAM-to-Speech Conversion with Gaussian Mixture Models. *Proc. INTERSPEECH*, pp. 1957–1960, Lisbon, Portugal, Sep. 2005.
6) A. Kain and M.W. Macon. Spectral voice conversion for text-to-speech synthesis. *Proc. ICASSP*, pp. 285–288, Seattle, USA, May 1998.
7) K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. *Proc. ICASSP*, pp. 1315–1318, Istanbul, Turkey, June 2000.
8) 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 2001*, Firentze, Italy, Sep. 2001.
9) 徳田恵一, 小林隆夫, 深田俊明, 今井 聖. 音声の適応メルケ プストラム分析, 電子情報通信学会論文誌 (A), vol.J74-A, no.8, pp.1249-1256, Aug. 1991
10) H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne. 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 *Speech Communication*, Vol. 27, No. 3-4, pp. 187–207, 1999.
11) H. Kawahara, H Katayose, A. Cheveigne and R. D. Patterson. FIEXED POINT ANALSYS OF FREQUENCY TO INSTANTANEOUS FREQUENCY MAPPING FOR ACCURATE ESTIMATION OF F0 AND PERIODICITY. *Proc. EUROSPEECH*, pp. 2781–2784, Budapest, Hungary, Sep. 1999.