2009－SLP－75 (2)
2009／2／6

# 音響特性補正の導入による肉伝導音声変換の収録環境適応

宮本 大輔 †， 中村 圭吾 †， 戸田 智基 †， 猿渡 洋 †， 鹿野 清宏 †

† 奈良先端科学技術大学院大学 情報科学研究科

肉伝導音声変換は Non-Audible Murmur (NAM) マイクロフォンで収録される肉伝導音声の音質向上に効果的である．この手法では，肉伝導音声から空気伝導音声へ変換するための確率モデルが事前に学習される．肉伝導音声の音響特性は，NAM マイクロフォンの圧着位置などの収録環境に敏感であるため，実際の使用においては学習時と変換時の音響特性の不一致により，しばしば変換音質が大きく劣化する．この問題に対して，我々は肉伝導音声変換のための Cepstrum Mean Subtraction (CMS) と Constrained Structural Maximum A Posteriori Linear Regression (CSMAPLR)，または Signal Bias Removal (SBR) と CSMAPLR の組み合わせに基づく教師無しの音響特性補正法を提案する．実験結果から，提案手法により音響特性の不一致に起因する変換音質の劣化が大幅に低減されることを示す．

# Adaptive Approach to Varying Recording Conditions in Body Transmitted Voice Conversion Based on Acoustic Compensation

Daisuke MIYAMOTO † Keigo NAKAMURA † Tomoki TODA † Hiroshi SARUWATARI †
Kiyohiro SHIKANO †

† Graduate school of Information Science, Nara Institute of Science and Technology

Body transmitted voice conversion is very effective for enhancing body transmitted speech recorded with Non-Audible Murmur (NAM) microphone. In this method, a probabilistic model to convert body transmitted speech into natural speech is trained previously. Because acoustic characteristics of body transmitted speech is sensitive to recording conditions such as a location of NAM microphone, significant degradation of the conversion performance is often caused in practical situations by acoustic mismatches between the training and the conversion processes. To alleviate this problem, we propose unsupervised acoustic compensation methods based on combination of Cepstrum Mean Subtraction (CMS) and Constrained Structural Maximum A Posteriori Linear Regression (CSMAPLR), or combination of Signal Bias Removal (SBR) and CSMAPLR for body transmitted voice conversion. Experimental results demonstrate that the proposed methods significantly reduce the quality degradation of the converted speech caused by the acoustic mismatches.

## 1 はじめに

近年，携帯電話の普及により様々な場所で音声コミュニケーションをとることが可能になった．しかし，人ごみや道路の側などの騒音環境下では相手に声を伝えることは困難である．また，図書館などの静穏環境下では，声を出すことで周囲に迷惑を掛けてしまうため小声で話さなければならず，同様に相手に声を伝えるのは困難である．このような状況下においても快適な音声コミュニケーションを実現するため，外部雑音に頑健であり，小さな声を収録可能な体内伝導音収録用マイクロフォンが開発されている [1, 2]．

我々は，体内伝導音収録用マイクロフォンの一つとして Non-Audible Murmur (NAM) マイクロフォン [2] に着目している．NAM マイクロフォンは耳介後方に圧着して，声道内の空気振動を皮下の筋肉組織を通じて収録する．このマイクロフォンの特徴は，通常の空気伝導マイクロフォンと比べてダイナミックレンジが大きく，通常の音量の音声から周囲に聞こえないほど小さな呟き声である NAM まで収録できることである．しかし，NAM マイクロフォンで収録される肉伝導音声は，口唇の放射特性の欠落や体内伝導によるローパス特性などの影響により，通常の空気伝導音声と比べて音質が大きく劣化する問題がある．

肉伝導音声の音質改善のために，統計的声質変換 [3] に基づく肉伝導音声変換が提案されている [4]．肉伝導音声変換では，同一内容を発声した空気伝導音声と肉伝導音声のパラレルデータを 50 文程度用いて，事前に変換モデルを学習しておく．この変換モデルを用いることで，肉伝導音声から空気伝導音声への変換が可能となる．学習時と変換時でデータの収録環境が同一であれば，肉伝導音声変換は極めて有効であり，音質が大きく改善される．しかし，肉伝導音声はマイクロフォンを体表に直接圧着して収録するため，NAM マイクロフォンの圧着位置などの収録環境が変化することで，収録される肉伝

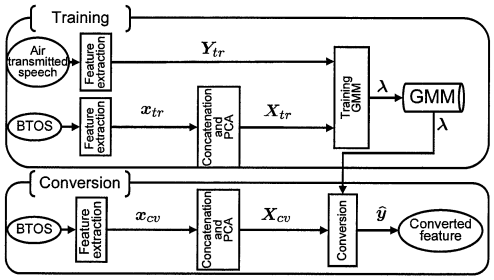


Fig. 1 Framework of body transmitted voice conversion.

導音声の音響特性が大きく変化する．収録環境の変化に起因する音響特性の変化は空気伝導音声と比べて大きく，その結果，肉伝導音声変換による変換音質が大きく劣化してしまう．

この問題を解決するため，我々は肉伝導音声変換のための音響特性補正技術の研究に取り組んでいる [5]．これまでに，Cepstrum Mean Subtraction (CMS) [6]，Constrained Structural Maximum A Posteriori Linear Regression (CSMAPLR) [7] に基づく音響特性補正の導入を行い，その有効性を確認している [5]．CMS に基づく補正法では，1 発話ごとに入力データをバイアス補正することで変換音質の劣化を大きく抑制できる．また，CSMAPLR に基づく補正法では，適応データを用いて新しい収録環境に変換モデルを適応させることで，数文以上の適応発声で CMS を上回る性能が得られる．本稿では，1 文ごとのバイアス補正とモデル適応を組み合わせた手法として，CMS と CSMAPLR，Signal Bias Removal (SBR) [8] と CSMAPLR を組み合わせた手法をそれぞれ導入する．客観および主観評価実験結果から，各音響特性補正法の性能を示す．

## 2 肉伝導音声変換

本稿では通常発声を NAM マイクロフォンで収録した肉伝導通常音声 (Body Transmitted Ordinary Speech: BTOS) から，通常の空気伝導マイクロフォンで収録された通常音声への変換を扱う．Fig. 1 に肉伝導音声変換の枠組みを示す．

### 2.1 肉伝導通常音声 (BTOS)

空気伝導音声と比較すると，BTOS は特に高域の周波数成分が大きく減衰している．このため，BTOS は酷くこもった音声となる．さらに，無声摩擦音などの高域に大きなパワーをもつ音素は，音韻特徴が失われやすい．これらの要因から，BTOS は空気伝導音声に比べて音質が大きく劣化しており，明瞭性も低い．

### 2.2 音響特徴量

本稿で用いる肉伝導音声変換では，スペクトル特徴量としてメルケプストラムを使用する．肉伝導によって欠落した情報を補うため，入力特徴量としてスペクトルセグメントを用いる．時刻 $t$ におけるメルケプストラムベクトル $\boldsymbol{x}_t$ の前後 $n$ フレームを連結させたベクトル $\boldsymbol{c}_t = [\boldsymbol{x}_{t-n}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+n}^\top]^\top$ ($\top$ は転置) を作成し，主成分分析 (PCA) により次元圧縮することで時刻 $t$ におけるスペクトルセグメント $\boldsymbol{X}_t$ を得る．

$$\boldsymbol{X}_t = \boldsymbol{D}\boldsymbol{c}_t - \boldsymbol{d} \tag{1}$$

ここで $\boldsymbol{D}$ は PCA 変換行列，$\boldsymbol{d} = \boldsymbol{D}\bar{\boldsymbol{c}}$ であり，$\bar{\boldsymbol{c}}$ は学習データ全体における $\boldsymbol{c}_t$ の平均ベクトルを表す．また，時刻 $t$ における出力特徴量として，静的特徴量 $\boldsymbol{y}_t$ と動的特徴量 $\Delta\boldsymbol{y}_t$ を連結させた静的・動的特徴量 $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ を用いる．

### 2.3 最尤基準に基づく特徴量変換 [3]

学習処理では，入力と出力の時間方向のアライメントをとった特徴量の結合確率密度を以下のように Gaussian Mixture Model (GMM) でモデル化する．

$$P(\boldsymbol{Z}_t|\boldsymbol{\lambda}) = \sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(Z)}, \boldsymbol{\Sigma}_m^{(ZZ)}) \tag{2}$$

ここで $\boldsymbol{Z}_t$ は入力特徴量と出力特徴量の結合ベクトルであり $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$ で構成される．記号 $\mathcal{N}()$ は正規分布を表す．混合数 $M$ の GMM のモデルパラメータセット $\boldsymbol{\lambda}$ は各分布 $m$ の重み $w_m$，平均ベクトル $\boldsymbol{\mu}_m^{(Z)}$，共分散行列 $\boldsymbol{\Sigma}_m^{(ZZ)}$ で構成される．ここで $\boldsymbol{\mu}_m^{(Z)}$ と $\boldsymbol{\Sigma}_m^{(ZZ)}$ はそれぞれ以下のように構成される．

$$\boldsymbol{\mu}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix} \tag{3}$$

$$\boldsymbol{\Sigma}_m^{(ZZ)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \tag{4}$$

$\boldsymbol{\mu}_m^{(X)}$ と $\boldsymbol{\mu}_m^{(Y)}$ はそれぞれ分布 $m$ の入力特徴量の平均ベクトルと出力特徴量の平均ベクトルを示す．$\boldsymbol{\Sigma}_m^{(XX)}$ と $\boldsymbol{\Sigma}_m^{(YY)}$ はそれぞれ分布 $m$ の入力特徴量，出力特徴量の共分散行列であり，$\boldsymbol{\Sigma}_m^{(XY)}$，$\boldsymbol{\Sigma}_m^{(YX)}$ は相互共分散行列を示す．

特徴量変換を行う際は，入力および出力の特徴量系列ベクトルをそれぞれ $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$，$\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$ とし，以下の尤度関数の最大化に基づき特徴量変換を行う．

$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) \simeq P(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}) P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}) \tag{5}$$

ここで，$\boldsymbol{m} = [m_1, \cdots, m_T]^\top$ は分布系列を表す．まず，準最適な分布系列 $\hat{\boldsymbol{m}}$ を次式で決定する．

$$\hat{\boldsymbol{m}} = \arg\max_{\boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}) \tag{6}$$

この分布系列 $\hat{\boldsymbol{m}}$ を用いて変換特徴量の静的特徴量系列 $\hat{\boldsymbol{y}}$ を次式で得る．

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}) \tag{7}$$

このとき $\boldsymbol{Y} = \boldsymbol{E}\boldsymbol{y}$ であり，$\boldsymbol{E}$ は静的特徴量系列から静的・動的特徴量系列へと変換する行列を表す．さらに，系列内変動を考慮することで，大幅に変換音質が改善される [3]．

本変換法においては，確率密度 $P(\boldsymbol{Z}_t|\boldsymbol{\lambda})$ が学習時と変換時で変化しないという前提が必要となる．

## 3 音響特性補正法

NAM マイクロフォンは体表に直接圧着して使用するため，収録される肉伝導音声の音響特性は圧着位置や接触圧などの収録環境に敏感である．このような収録環境の変化に対応するために音響特性の補正を試みる．本稿ではより簡便で実用的な方法として，出力音声である空気伝導音声や言語情報を用いず，入力音声である肉伝導音声のみを用いた補正法を提案する．

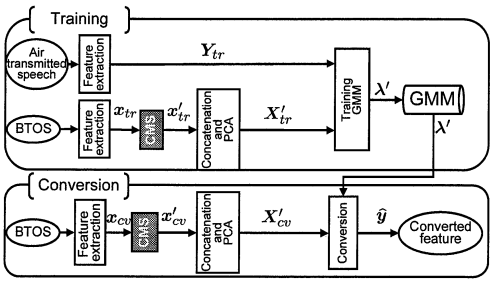


Fig. 2 Framework of body transmitted voice conversion with acoustic compensation based on CMS.

### 3.1 Cepstrum Mean Subtraction (CMS) に基づく補正法 [5]

CMS [6] は静的な乗法性歪みの影響を除去する手法である．学習時と変換時それぞれの入力特徴量に対してCMS を行い，収録環境の違いによって変化する静的な乗法性歪みの影響を除去し，変換音質の劣化を抑制する．

Fig. 2 に CMS に基づく音響特性補正を導入した肉伝導音声変換の枠組みを示す．Section 2.2 で示した入力メルケプストラムに対して CMS を導入すると，正規化された特徴量が次式で与えられる．

$$\boldsymbol{x}'_t = \boldsymbol{x}_t - \bar{\boldsymbol{x}} \tag{8}$$

$\bar{\boldsymbol{x}}$ は $\boldsymbol{x}_t$ の長時間平均であり，本稿では発話ごとに計算された平均ベクトルを用いる．この正規化されたメルケプストラムを用いてスペクトルセグメント $\boldsymbol{X}'_t$ を求める．

$$\boldsymbol{X}'_t = \boldsymbol{D}'\boldsymbol{c}'_t - \boldsymbol{d}' \tag{9}$$

このとき，$\boldsymbol{c}'_t = [\boldsymbol{x}'^{\top}_{t-n}, \cdots, \boldsymbol{x}'^{\top}_{t}, \cdots, \boldsymbol{x}'^{\top}_{t+n}]^{\top}$，$\boldsymbol{d}' = \boldsymbol{D}'\bar{\boldsymbol{c}}'$ であり，$\boldsymbol{D}'$ と $\bar{\boldsymbol{c}}'$ はそれぞれ PCA 変換行列と学習データ全体における $\boldsymbol{c}'_t$ の平均ベクトルを示す．

### 3.2 Constrained Structural Maximum A Posteriori Linear Regression (CSMAPLR) に基づく補正法 [5]

CSMAPLR [7] はモデル適応手法の一つであり，CMS と比べてより複雑な変換処理を行う．Fig. 3 に CSMAPLR に基づく音響特性補正を導入した肉伝導音声変換の枠組みを示す．

CSMAPLR は Constrained Maximum Likelihood Linear Regression (CMLLR) [9] と同様，入力特徴量に対してアフィン変換を行うことでモデルとデータの不一致を補正する．入力特徴量 $\boldsymbol{X}_t$ の変換特徴量は次式で与えられる．

$$\hat{\boldsymbol{X}}_t = \boldsymbol{A}\boldsymbol{X}_t + \boldsymbol{b} = \boldsymbol{W}\boldsymbol{\xi}_t \tag{10}$$

このとき，$\boldsymbol{W} = [\boldsymbol{b}, \boldsymbol{A}]$ は変換行列であり，$\boldsymbol{\xi}_t$ は特徴量の拡張ベクトル $\boldsymbol{\xi}_t = [1, \boldsymbol{X}_t^{\top}]^{\top}$ である．ここで，Fig. 4 に示すような回帰木を用いて，適応データ量に従って複数の変換行列を推定し，変換を行うことで複雑な変換を実現する．各分布に使用する変換行列は，閾値以上のデータ量を持つ上位ノード (クラス)$r$ で推定される変換行列 $\boldsymbol{W}_r = [\boldsymbol{b}_r, \boldsymbol{A}_r]$ を使用する．CSMAPLR では，変換行列を推定する際に木構造を利用した MAP 推定を行う．Fig. 4 に示すように，ハイパーパラメータとして親ノードの変換行列の推定値を使用することで，過学習を抑制する．

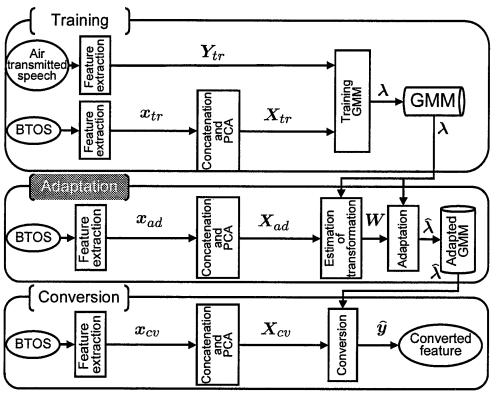


Fig. 3 Framework of body transmitted voice conversion with acoustic compensation based on CSMAPLR.

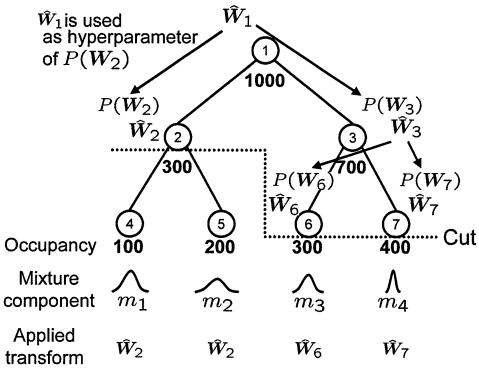


Fig. 4 An example of usage of regression tree with CSMAPLR when setting occupancy threshold to 300.

出力特徴量を用いない教師無し適応を行うため，適応データ特徴量系列 $\boldsymbol{X}$ の周辺分布の尤度を最大化する変換行列 $\hat{\boldsymbol{W}}_r$ を推定する．

$$\hat{\boldsymbol{W}}_r = \arg\max_{\boldsymbol{W}_r} \int P(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{W}_r, \boldsymbol{\lambda}) P(\boldsymbol{W}_r) d\boldsymbol{Y} \tag{11}$$

ここで事前分布 $P(\boldsymbol{W}_r)$ は行列変量正規分布で表され，次式で定義される．

$$P(\boldsymbol{W}_r) \propto |\boldsymbol{\Psi}|^{-(d+1)/2} |\boldsymbol{\Phi}|^{-d/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\{(\boldsymbol{W}_r - \boldsymbol{H})^{\top}\boldsymbol{\Psi}^{-1}(\boldsymbol{W}_r - \boldsymbol{H})\boldsymbol{\Phi}^{-1}\}\right] \tag{12}$$

ここで，$d$ は入力特徴量の次元数であり，$\boldsymbol{\Psi}$，$\boldsymbol{\Phi}$，$\boldsymbol{H}$ はそれぞれハイパーパラメータである．$\boldsymbol{H}$ は回帰木の親ノードで推定された変換行列を用い，$\boldsymbol{\Psi}$，$\boldsymbol{\Phi}$ はそれぞれ $\boldsymbol{\Psi}^{-1} = C \cdot \boldsymbol{I}$，$\boldsymbol{\Phi}^{-1} = \boldsymbol{I}$ とする．$\boldsymbol{I}$ は単位行列を示

し，$C$ は $P(\boldsymbol{W}_r)$ の大きさを決定する係数である．$\hat{\boldsymbol{W}}_r$ は EM アルゴリズムで求められ

$$\hat{\boldsymbol{w}}_{ri} = (\alpha \boldsymbol{c}_i + \boldsymbol{n}^{(ri)}) \boldsymbol{V}^{(rii)^{-1}} \tag{13}$$

で与えられる．ここで，$\hat{\boldsymbol{w}}_{ri}$ は $\hat{\boldsymbol{W}}_r$ の $i$ 行成分，$\boldsymbol{c}_i$ は $\boldsymbol{A}_r$ の $i$ 行目の余因子ベクトルの拡張ベクトルである．$\alpha$ は二次方程式を解くことで得られる値である [9]．$\boldsymbol{n}^{(ri)}$ と $\boldsymbol{V}^{(rii)}$ はそれぞれ以下の式で与えられる．

$$\boldsymbol{V}^{(rii)} = \boldsymbol{G}^{(rii)} + C \cdot \boldsymbol{I} \tag{14}$$

$$\boldsymbol{n}^{(ri)} = \boldsymbol{k}^{(ri)} + C \cdot \boldsymbol{h}(i) \tag{15}$$

$$\boldsymbol{G}^{(rij)} = \sum_{m=1}^{M_r} p_m(i,j) \sum_{t=1}^{T} \gamma_m(t) \boldsymbol{\xi}_t \boldsymbol{\xi}_t^\top \tag{16}$$

$$\boldsymbol{k}^{(ri)} = \sum_{m=1}^{M_r} \boldsymbol{p}_m(i) \boldsymbol{\mu}_m \sum_{t=1}^{T} \gamma_m(t) \boldsymbol{\xi}_t^\top - \sum_{j=1, j\neq i}^{d} \boldsymbol{w}_{rj} \boldsymbol{G}^{(rij)} \tag{17}$$

ここで，$\boldsymbol{p}_m(i)$ と $p_m(i,j)$ はそれぞれ共分散行列 $\boldsymbol{\Sigma}_m^{(XX)}$ の $i$ 行成分と $(i,j)$ 要素を示す．また，$M_r$ と $\gamma_m(t)$ はそれぞれクラス $r$ に属する分布数と分布 $m$ の時間 $t$ における生起確率 $\gamma_m(t) = P(m|\boldsymbol{X}_t, \boldsymbol{W}_r, \boldsymbol{\lambda})$ を示す．

本稿では CSMAPLR で推定された変換行列をモデル空間に適用し，適応された変換モデルを用いて変換処理を行う．適応されたモデルパラメータは以下の形で与えられる．

$$\hat{\boldsymbol{\mu}}_m^{(Z)} = \begin{bmatrix} \boldsymbol{A}_r' \boldsymbol{\mu}_m^{(X)} - \boldsymbol{b}_r' \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix} \tag{18}$$

$$\hat{\boldsymbol{\Sigma}}_m^{(Z)} = \begin{bmatrix} \boldsymbol{A}_r' \boldsymbol{\Sigma}_m^{(XX)} \boldsymbol{A}_r'^\top & \boldsymbol{A}_r' \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{A}_r'^\top & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \tag{19}$$

このとき，$\hat{\boldsymbol{\mu}}_m^{(Z)}$ と $\hat{\boldsymbol{\Sigma}}_m^{(ZZ)}$ はそれぞれ適応された分布 $m$ の平均ベクトルと共分散行列を表す．また，$\boldsymbol{A}_r' = \boldsymbol{A}_r^{-1}$，$\boldsymbol{b}_r' = \boldsymbol{A}_r' \boldsymbol{b}_r$ である．

### 3.3 バイアス補正と CSMAPLR を組み合わせた補正法

1 文ごとにバイアス補正を行う CMS は，適応データ数文で変換を推定する CSMAPLR と比べると，発話ごとの変動に対して頑健であると考えられる．そこで，Fig. 5 に示すように CMS と CSMAPLR を組み合わせることで，更なる性能改善を試みる．学習時は，CMS のみを用いた補正と同様に変換モデルを学習する．適応時は，適応データから得られる CMS によって補正された特徴量を用いて，CSMAPLR によって回帰行列を推定し，変換モデルを適応する．変換時は，適応された変換モデルを用い，CMS のみを用いた補正と同様に変換処理を行う．

CMS によって得られるバイアスは 1 文の平均として計算される．この場合，後段の CSMAPLR で用いられる評価基準を一切考慮しておらず，CSMAPLR における回帰行列の推定に悪影響を及ぼす可能性がある．そこで，バイアスを尤度基準で推定する Signal Bias Removal (SBR) [8] を導入する．また，本稿では SBR によってバイアスを推定する際に CMS の結果を初期値として与える．Fig. 6 に SBR と CSMAPLR を組み合わせた補正を導入した肉伝導音声変換の枠組みを示す．

学習時は，まず CMS のみを用いて補正を行い GMM を学習する．得られた GMM を用いて，1 文ごとに SBR

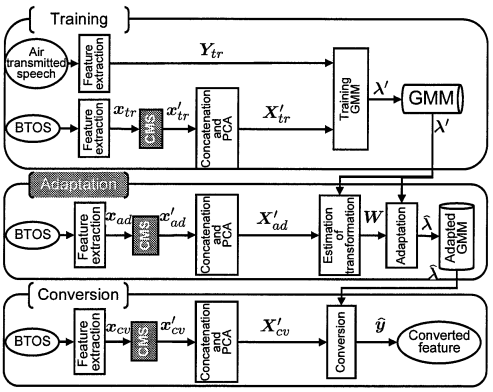


Fig. 5 Framework of body transmitted voice conversion with acoustic compensation based on the method combining CMS and CSMAPLR.

により補正を行う．SBR によって得られる特徴量は，バイアス $\boldsymbol{g}$ を用いて

$$\tilde{\boldsymbol{X}}_t = \boldsymbol{X}_t' - \boldsymbol{g} \tag{20}$$

で与えられる．学習時は出力特徴量の情報が得られるため，尤度を最大化する入力特徴量のバイアス $\hat{\boldsymbol{g}}$ は

$$\hat{\boldsymbol{g}} = \arg\max_{\boldsymbol{g}} P(\boldsymbol{X}', \boldsymbol{Y} | \boldsymbol{g}, \boldsymbol{\lambda}) \tag{21}$$

として，各発話に対して推定される．$\hat{\boldsymbol{g}}$ は EM アルゴリズムで求められ，次式で与えられる．

$$\hat{\boldsymbol{g}} = \boldsymbol{B}^{(Z)^{-1}} (\boldsymbol{l}^{(Z)} - \boldsymbol{q}^{(Z)}) \tag{22}$$

このとき $\boldsymbol{q}^{(Z)}$，$\boldsymbol{l}^{(Z)}$，$\boldsymbol{B}^{(Z)}$ は以下で与えられる．

$$\boldsymbol{q}^{(Z)} = \sum_{m}^{M} \left[ \boldsymbol{P}_m^{(XX)} \ \boldsymbol{P}_m^{(XY)} \right] \boldsymbol{\mu}_m^{(Z)} \sum_{t}^{T} \tilde{\gamma}_m^{(Z)}(t) \tag{23}$$

$$\boldsymbol{l}^{(Z)} = \sum_{m}^{M} \left[ \boldsymbol{P}_m^{(XX)} \ \boldsymbol{P}_m^{(XY)} \right] \sum_{t}^{T} \tilde{\gamma}_m^{(Z)}(t) \boldsymbol{Z}_t' \tag{24}$$

$$\boldsymbol{B}^{(Z)} = \sum_{m}^{M} \boldsymbol{P}_m^{(XX)} \sum_{t}^{T} \tilde{\gamma}_m^{(Z)}(t) \tag{25}$$

ここで，$\tilde{\gamma}_m^{(Z)}(t) = P(m|\boldsymbol{Z}_t', \boldsymbol{g}, \boldsymbol{\lambda})$ であり，$\boldsymbol{Z}_t' = [\boldsymbol{X}_t'^\top, \boldsymbol{Y}_t^\top]^\top$ である．また，$\boldsymbol{P}_m^{(XX)}$，$\boldsymbol{P}_m^{(XY)}$，$\boldsymbol{P}_m^{(YX)}$，$\boldsymbol{P}_m^{(YY)}$ はそれぞれ $\boldsymbol{\Sigma}_m^{(XX)}$，$\boldsymbol{\Sigma}_m^{(XY)}$，$\boldsymbol{\Sigma}_m^{(YX)}$，$\boldsymbol{\Sigma}_m^{(YY)}$ と同じサイズであり，$\boldsymbol{\Sigma}_m^{(ZZ)^{-1}}$ を構成する行列である．

$$\boldsymbol{\Sigma}_m^{(ZZ)^{-1}} = \begin{bmatrix} \boldsymbol{P}_m^{(XX)} & \boldsymbol{P}_m^{(XY)} \\ \boldsymbol{P}_m^{(YX)} & \boldsymbol{P}_m^{(YY)} \end{bmatrix} \tag{26}$$

得られたバイアスの最尤推定値を用いて特徴量の補正を行い，GMM のパラメータを再学習する．さらに，得られた GMM を用いて上記処理を繰り返し行う．本学習処理は，発話毎の適応パラメータと GMM パラメータを同一の尤度関数に基づいて更新するものであり，発話適応学習とみなせる．

適応時は学習時に得られた GMM を用いて SBR によるバイアス推定を行う．また，変換時は CSMAPLR で

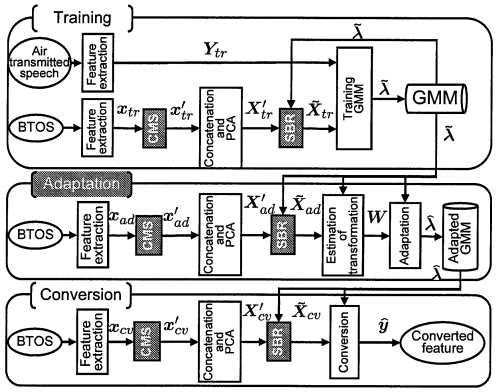


Fig. 6 Framework of body transmitted voice conversion with acoustic compensation based on the method combining SBR and CSMAPLR.

適応された GMM を用いて SBR によるバイアス推定を行う．適応および変換時は，入力特徴量のみを用いてバイアスを推定するため，$\boldsymbol{X}'$ の周辺分布を最大化する $\hat{\boldsymbol{g}}$ を各文ごとに推定する．

$$\hat{\boldsymbol{g}} = \arg\max_{\boldsymbol{g}} \int P(\boldsymbol{X}', \boldsymbol{Y}|\boldsymbol{g}, \boldsymbol{\lambda}) d\boldsymbol{Y} \tag{27}$$

$\hat{\boldsymbol{g}}$ は EM アルゴリズムで求められ次式で与えられる．

$$\hat{\boldsymbol{g}} = \boldsymbol{B}^{(X)^{-1}}(\boldsymbol{l}^{(X)} - \boldsymbol{q}^{(X)}) \tag{28}$$

このとき $\boldsymbol{q}^{(X)}$，$\boldsymbol{l}^{(X)}$，$\boldsymbol{B}^{(X)}$ は，生起確率 $\tilde{\gamma}_m^{(X)}(t) = P(m|\boldsymbol{X}'_t, \boldsymbol{g}, \boldsymbol{\lambda})$ を用いて以下で与えられる．

$$\boldsymbol{q}^{(X)} = \sum_m^M \boldsymbol{\Sigma}_m^{(XX)^{-1}} \boldsymbol{\mu}_m^{(X)} \sum_t^T \tilde{\gamma}_m^{(X)}(t) \tag{29}$$

$$\boldsymbol{l}^{(X)} = \sum_m^M \boldsymbol{\Sigma}_m^{(XX)^{-1}} \sum_t^T \tilde{\gamma}_m^{(X)}(t) \boldsymbol{X}'_t \tag{30}$$

$$\boldsymbol{B}^{(X)} = \sum_m^M \boldsymbol{\Sigma}_m^{(XX)^{-1}} \sum_t^T \tilde{\gamma}_m^{(X)}(t) \tag{31}$$

## 4 評価実験

### 4.1 実験条件

実験用音声として NAM マイクロフォンとヘッドセットマイクロフォンを同時に装着し，BTOS と空気伝導音声を同時に収録した．2 名の男性話者から，NAM マイクロフォンの圧着位置のみ変化させ（首の左右を入れ替え），他の収録環境は可能な限り保持した 2 つの環境で同一内容の発話を収録した．まず，NAM マイクロフォンの圧着位置を固定したまま音素バランス文を 100 文収録した．次に NAM マイクロフォンの圧着位置を変更し，再度同一内容の音素バランス文を 100 文収録した．1 話者は 1 組 (100 × 2 文)，もう 1 話者は 2 組の音声を収録し，合計 3 組の音声セットを実験に用いた．音声のサンプリング周波数は 8 kHz とした．

収録環境が異なる条件での変換音質の評価を行うために，Fig. 7 に示すようなデータセットを用いた．一方の

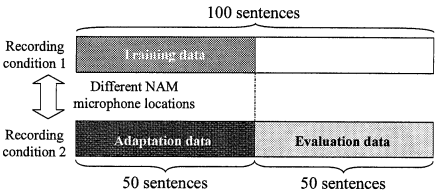


Fig. 7 Data sets for experiments.

環境の 50 文を変換モデル学習に用い，もう一方の環境の，モデル学習に使用しなかった内容の 50 文を用いて評価を行った．CSMAPLR の変換行列推定に用いる適応文は評価文と同一環境の，評価に用いない 50 文から必要量を選択し使用した．この適応データ 50 文は，同一環境の変換モデル学習時にも使用した．CMS 及び SBR による補正では，各文ごとにバイアスを算出した．交差検定として，学習環境と評価環境を入れ替えて同一の評価を行った．

音響特性補正法として，収録環境が変化した場合の，1) 補正無し，2) CMS，3) CSMAPLR，4) CMS と CSMAPLR の組み合わせ，5) SBR と CSMAPLR の組み合わせを導入した肉伝導音声変換をそれぞれ評価した．また，6) 同一の収録環境で変換を行った場合の評価も行った．これらの変換手法に対して客観的・主観的に評価を行った．客観評価実験の評価指標は目標音声と変換音声のパワー項を含むメルケプストラム歪みを用いた．主観評価実験は，一対比較法で音質を評価した．刺激音声として，上述の 1)，2)，3)，5)，6) を行った場合の変換音声を提示した．CSMAPLR に使用する適応文は 1 文あるいは 10 文を用いた場合の変換音声をそれぞれ評価した．刺激音声は，異なる補正法を用いて変換された 2 種類の音声をランダムに提示した．被験者は補正法のすべての組み合わせを評価し，1 文あたり 20 通りの組み合わせで補正法を評価した．被験者は 9 名とした．

スペクトル特徴量として，0 から 16 次までのメルケプストラム係数を用いた．また，入力特徴量として前後 4 フレーム連結したベクトルを 34 次元に圧縮したスペクトルセグメントを用いた．GMM の混合数は 64 とした．CSMAPLR において，クラス決定のためのデータ量の閾値を 1000 とし，ハイパーパラメータ $C$ の値は $10^5$ として実験を行った．

### 4.2 実験結果

Fig. 8 に客観評価実験結果を示す．この結果より，適応文数が少ない場合 CMS の方が性能が高く，適応文数が多くなると CSMAPLR は CMS を上回る性能が得られることが分かる．CMS と CSMAPLR を組み合わせた補正では，CSMAPLR のみの補正と比べると，適応文数が極端に少ない場合は性能が高いが，適応文数が多くなると性能が悪化することが分かる．これは，CMS によるバイアス推定が不安定であるため，CSMAPLR の回帰行列推定精度が低下していることが原因であると考えられる．一方で，SBR と CSMAPLR を組み合わせることで，適応文数が大きくなっても CSMAPLR のみの補正と同程度の性能が得られる．適応文が十分に大きくなると CSMAPLR のみの補正を導入した枠組みで変換した結果に漸近することから，発話ごとの変動は GMM

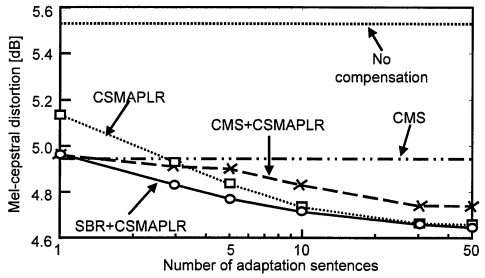


Fig. 8 Mel-cepstral distortion as a function of the number of adaptation sentences. The mel-capstral distortion in the matched condition is 3.93 dB.

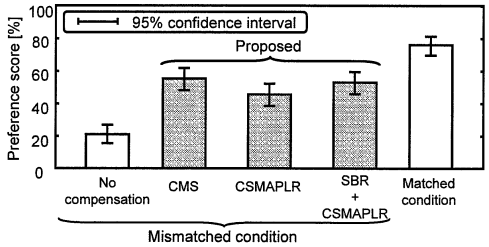


Fig. 9 Results of preference test on speech quality. One sentence was used for adaptation data in the CSMAPLR compensation.

で表現できる程度であり，発話適応処理を行う必要が無いことが分かる．また，適応文数が極端に少ない場合にはCSMAPLRのみの補正と比べて大きく性能が向上していることが分かる．このことから，適応分数が少ない場合にCSMAPLRで生じる過学習の影響は，SBRを用いてCSMAPLRの推定精度を改善させることにより，効果的に緩和できることが分かる．

Fig. 9に適応文を1文用いた場合，Fig. 10に適応文を10文用いた場合の主観評価実験結果を示す．適応文が1文の場合，CMSに基づく補正とSBRとCSMAPLRの組み合わせに基づく補正が同程度の性能を示しており，CSMAPLRのみの補正を用いる場合よりも音質が改善される傾向がある．また，適応文が10文程度得られた場合，CSMAPLR単体で用いる場合とSBRと組み合わせて用いた場合で同程度の音質であり，これらはCMSのみを用いた場合よりも音質が良いことが分かる．これらの結果は客観評価実験結果と同様の傾向を示しており，妥当であると考えられる．

## 5 おわりに

学習時と変換時で音響特性が異なる場合における肉伝導音声変換の音質改善のため，CMSとCSMAPLRの組み合わせ，およびSBRとCSMAPLRの組み合わせに基づく音響特性補正の提案を行った．客観および主観評価実験結果から，バイアス補正による発話ごとの変動を補正する必要性は低いが，SBRがCSMAPLRの推定精度を向上させ，極端に適応文が少ない場合において

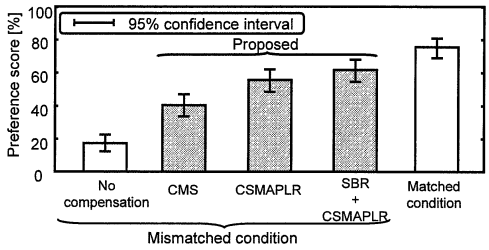


Fig. 10 Results of preference test on speech quality. Ten sentences were used for adaptation data in the CSMAPLR compensation.

特にその効果が顕著に見られることが分かった．

## 謝辞

本研究の一部は，総務省のSCOPEおよび文部科学省の科研費若手(A)により実施したものである．

## 参考文献

1) Y. Zheng, Z. Liu, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X. Huang, "Air- and Bone-Conductive Integrated Microphones for Robust Speech Detection and Enhancement", *Proc. ASRU*, pp. 249–254, 2003

2) Y. Nakajima, H. Kashioka, N. Campbell, and K. Shikano, "Non-Audible Murmur Recognition", *IEICE Trans. Information and Systems*, Vol. E89–D, No. 1, pp. 1–8, 2006

3) T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory", *Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2235, 2007

4) T. Toda, and K. Shikano, "NAM-to-Speech Conversion with Gaussian Mixture Models", *Proc. INTERSPEECH*, pp. 1957–1960, 2005

5) 宮本 大輔，中村 圭吾，戸田 智基，猿渡 洋，鹿野 清宏，"肉伝導音声変換のための音響特性補正法"，信学技報，Jan. 2009

6) B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification", J. Acoust. Soc. America, Vol. 55, No. 6, pp. 1304–1312, 1974

7) Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Constrained Structural Maximum A Posteriori Linear Regression for Average-Voice-Based Speech Synthesis", *Proc. INTERSPEECH*, pp. 2286–2289, 2006

8) M. Rahim, and B.H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition", *IEEE Trans. Speech Audio Processing* , Vol. 4, pp. 19–30, 1996

9) M. J. F. Gales, "Maximum Likelihood Linear Transformation for HMM-based Speech recognition", Computer Speech and Language, Vol. 12, No. 2, pp. 75–98, 1998