Vol.2010-SLP-82 No.1
2010/7/22

異なる収録機器に対応した肉伝導音声変換のための音響特性自動補正

出口大祐[†1] 土井啓成[†1] 戸田智基[†1]
猿渡洋[†1] 鹿野清宏[†1]

NAM マイクロフォンを用いて収録される肉伝導音声の品質を改善する技術として，統計的声質変換に基づく肉伝導音声変換法が提案されている．この音質改善法は，肉伝導音声と通常音声の同一発話により得た音響特徴量から変換モデルを構築することで，肉伝導音声から通常音声への変換を実現する．大幅な音質改善効果が得られるが，学習時と変換時において収録音声の音響特性は変化しないという仮定を必要とし，仮に音響特性が大きく異なると，大幅な音質劣化が生じる．実際の使用状況を想定すると，NAM マイクロフォンの圧着位置や異なる収録機器の使用など，収録条件を一定に保つのは困難である．これまでに，制約付き最尤線形回帰（CMLLR: Constrained Maximum Likelihood LinearRegression) に基づく教師なし音響補正法が提案され，NAM マイクロフォンの圧着位置に起因する音響特性変化に対して，その有効性が報告されている．本稿では，CMLLR 基づく音響特性補正法を，収録機器の違いによる音響特性変化に対する補正に応用し，その有効性を評価する．実験的評価により，提案法が収録機器の違い及び NAM マイクロフォンの圧着位置の変化による変換音声の音質低下を抑えることができることを示す．

Automatic Acoustic Compensation Method for Accepting Different Recording Devices in Body-Conducted Voice Conversion

DAISUKE DEGUCHI ,[†1] HIRONORI DOI ,[†1]
TOMOKI TODA,[†1] HIROSHI SARUWATARI[†1]
and KIYOHIRO SHIKANO[†1]

This paper presents an acoustic compensation method in body-conducted speech conversion for automatically compensating acoustic differences caused by changes of recording conditions. A statistical voice conversion technique has successfully been applied to an enhancement process of body-conducted speech recorded with Non-Audible Murmur (NAM) microphone. Speech waveforms are generated from acoustic parameters of normal speech estimated from those of body-conducted speech with a conversion model previously trained using stereo data of those two types of speech. This framework suffers from mismatched conditions between training and conversion processes. To alleviate this issue, an unsupervised acoustic compensation method based on constrained maximum likelihood linear regression (CMLLR) has been proposed and its effectiveness has been reported in the compensation of acoustic differences caused by the attachment location of NAM microphone. This paper further applies the CMLLR-based acoustic compensation method to the compensation of acoustic differences caused by different recording devices and evaluates its effectiveness. The experimental results demonstrate that the proposed method effectively reduces the quality degradation in the converted speech caused by the differences of recording devices as well as the attachment location of NAM microphone.

1. はじめに

近年，携帯電話などの普及により，様々な場所で音声を用いてコミュニケーションを行うことが容易となった．しかし，人ごみや工事現場などの騒音下では，雑音が混入し，相手に発声内容を伝えることが困難である．この問題に対し，体内を伝導する音声を用いた音声コミュニケーション法が提案されている[1),2)]．我々は，体内伝導音声の一つして，Non-Audible Murmur (NAM)[2)] マイクロフォンを用いて収録される肉伝導音声に着目し，研究を進めている．NAM マイクロフォンは，耳介後方に圧着することで，体表から皮下の筋肉振動を介して音声を収録することが可能なマイクロフォンである．NAM マイクロフォンを使用すれば，体表から直接肉伝導音声を収録できるため，騒音環境下においても，外部雑音の影響を大幅に低減することができる．一方で，肉伝導音声は，筋肉を介して収録されるため，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，空気伝導音声に比べ音質が劣化する．そのため，肉伝導音声のみを用いて，円滑なコミュニケーションを行うことは難しい．

NAM マイクロフォンで収録される肉伝導音声の音質を改善するために，統計的声質変換に基づく肉伝導音声変換[3)] が提案されている．統計的声質変換[4),5)] は，ある話者の任意の発話を別の話者のものに変換する手法であり，両話者の 50 文程度の同一発話を基に，事前

†1 奈良先端科学技術大学院大学
Nara Institute of Science and Technology

© 2010 Information Processing Society of Japan

Vol.2010-SLP-82 No.1
2010/7/22

に変換モデルを学習し，学習した変換モデルを用いて，ある話者の任意の発話を別の話者に変換する．肉伝導音声変換では，同時に収録した肉伝導音声と空気伝導音声を用いて，肉伝導音声から空気伝導音声への変換モデルを学習する．学習したモデルを用いて，肉伝導音声を空気伝導音声に変換することで，音質の改善を図る．この時，変換音声は，空気伝導音声の統計量に基づき生成されるため，口唇の放射特性等の肉伝導音声では欠落する特徴が再現され，大きな音質改善が得られる．しかしながら，学習と変換で，使用した収録機器や圧着位置等が異なると，肉伝導音声の音響特性に変化が生じ，変換音声の音質が大きく劣化することがある．実際の利用を考えた場合，用途や状況に応じた機器の選択や，買い替え等による器機の変更，圧着位置の変化等は十分に起こりえる．そのため，収録環境の変化に対応した音響特性補正を行う必要がある．

圧着位置の変化による音質劣化に対しては，制約付き最尤線形回帰 （CMLLR: Constrained Maximum Likelihood Linear Regression) に基づく音響特性補正法[6] が提案されており，その有効性が示されている．この手法では，異なる圧着位置から収録された肉伝導音声のみを用いて，変換モデルの尤度最大化基準に基づき，学習時と変換時の音響特性の違いを補正する特徴量変換行列を推定する． そして，推定した変換行列を基に新たな肉伝導音声の音響特徴量を変換することで，変換モデルと音響特徴量との不一致を解消する．

本稿では，CMLLR に基づく音響特性補正法を応用し，NAM マイクロフォンの圧着位置の違いのみでなく，収録機器の違いに起因する音響特性の変化に対しても，補正を行う手法について検討する．異なる NAM マイクロフォンで収録された肉伝導音声を用いることで，収録条件の違いがもたらす音質劣化度合いを明らかにするとともに，本手法による補正効果を客観的及び主観的に評価する．

2. 肉伝導音声

NAM マイクロフォンは，非常に小さな音声から通常の音声まで，幅広い音声を収録することが可能である．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を取り扱う．

肉伝導音声は，空気伝導音声と異なるスペクトル構造を持ち，高帯域の成分は大幅に減衰する．そのため，肉伝導音声は，非常にこもった音声になる．音源特徴量の一つである非周期成分[7] においても，肉伝導音声と空気伝導音声では異なる特徴が見られる．一方で，基本周波数は空気伝導音声と同等のものが得られる．

本稿では，図 1 に示す三種類の異なる NAM マイクロフォンとアンプを用いて肉伝導音声を収録する．図 1 から，波形やスペクトルにおいて，使用したマイクロフォン毎に明らかな違いが生じており，収録機器の違いが音響特性に与える影響は大きいことが分かる．なお，これらのサンプルには，異なる圧着位置により生じる音響特性変化や収録日時の違いによる声質の変動なども含まれる．

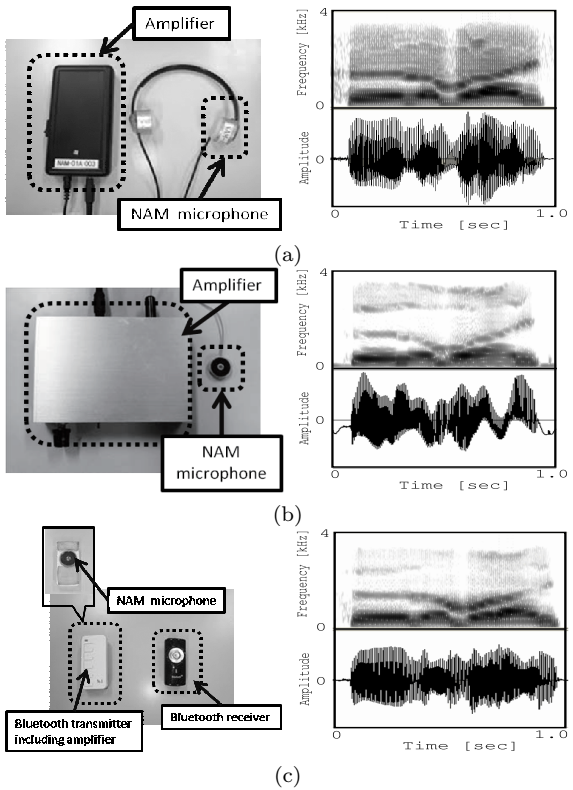


図 1 3 種類の NAM マイクロフォンとそれぞれの NAM マイクロフォンを用いて収録された肉伝導音声のスペクトログラムの例．(a) ネックバンド持つ有線タイプ，(b) ネックバンドを持たない有線タイプ，(c) 無線タイプ

Fig. 1 Three types of NAM microphone, (a) wired-type with neckband, (b) wired-type without neckband, and (c) wireless-type, and an example of spectrogram of body-conducted speech recorded with each NAM microphone.

© 2010 Information Processing Society of Japan

Vol.2010-SLP-82 No.1
2010/7/22

3. 肉伝導音声変換

本稿における肉伝導音声変換では，肉伝導音声のスペクトルから空気伝導音声のスペクトルへの変換，肉伝導音声の非周期成分から空気伝導音声の非周期成分への変換を行う．変換モデルは，それぞれ独立に学習する．また，基本周波数 F_0 は，肉伝導音声から得られたものをそのまま用いる．肉伝導音声変換では，変換されたスペクトル及び非周期成分，肉伝導音声から抽出された基本周波数 F_0 を用いて，変換音声を合成する．

3.1 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 $\boldsymbol{c}_t = [\boldsymbol{x}_{t-n}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+n}^\top]^\top$ を作成し，主成分分析 （Principal Component Analysis : 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．ここで，$\top$ は転置を表す．

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

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．

3.2 最尤基準に基づく特徴量変換[5)]

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混合正規分布モデル（GMM : Gaussian Mixture Model）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，$\boldsymbol{m} = \{m_1, \cdots, m_T\}$ 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．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り，変化した場合，変換した音声の劣化が起こる．

4. 制約付き最尤線形回帰 (CMLLR: Constrained Maximum Likelihood Linear Regression) に基づく音響特性補正[6)]

実環境下における NAM マイクロフォンの使用を考えると，マイク特性の違いやマイク圧着位置の違いにより生じる音響特性の変化を補正する必要がある．本稿では，CMLLR 適応処理を用いて，それらの補正を行う．なお，より利便性に優れた補正処理を実現するために，新たな収録条件下において，肉伝導音声と空気伝導音声の同時収録を必要とせず，任意の発声の肉伝導音声のみを用いた教師なし CMLLR 適応を行う．これにより，収録条件が異なっても，NAM マイクロフォンを使用するだけで，自動的に補正処理を行うことが可能となる．

CMLLR は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 $\boldsymbol{X}_t$ の補正特徴量は

$$\hat{\boldsymbol{X}}_t = \boldsymbol{A}\boldsymbol{X}_t + \boldsymbol{b} = \boldsymbol{W}\boldsymbol{\xi}_t \tag{8}$$

で与えられる．このとき，$\boldsymbol{W}$ は変換行列 $\boldsymbol{W} = [\boldsymbol{b}, \boldsymbol{A}]$ であり，$\boldsymbol{\xi}_t$ は拡張ベクトル

© 2010 Information Processing Society of Japan

Vol.2010-SLP-82 No.1
2010/7/22

$\boldsymbol{\xi}_t = [1,\ \boldsymbol{X}_t^\top]^\top$ である．本稿では，1 クラス CMLLR を行う．

変換行列は，変換されたデータに対するモデルの尤度が最大となるように推定される．出力特徴量を用いない教師無し適応を行うため，適応データの特徴量系列 $\boldsymbol{X}$ の周辺分布尤度を最大化する変換行列 $\hat{\boldsymbol{W}}$ を推定する．

$$\hat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W}} \prod_{t=1}^{T} \int P(\boldsymbol{X}_t, \boldsymbol{Y}_t | \boldsymbol{W}, \boldsymbol{\lambda}) d\boldsymbol{Y}_t \tag{9}$$

確率密度は GMM でモデル化されているため，$\hat{\boldsymbol{W}}$ は EM アルゴリズムで求められる．更新された変換行列は

$$\hat{\boldsymbol{w}}_i = (\alpha \boldsymbol{c}_i + \boldsymbol{k}^{(i)}) \boldsymbol{G}^{(ii)-1} \tag{10}$$

で与えられる．ここで，$\hat{\boldsymbol{w}}_i$ は $\hat{\boldsymbol{W}}$ の i 行成分であり，$\boldsymbol{c}_i$ は $\boldsymbol{A}$ の i 行目の余因子ベクトルの拡張ベクトルである．α は

$$\alpha^2 \boldsymbol{c}_i \boldsymbol{G}^{(ii)-1} \boldsymbol{c}_i^\top + \alpha \boldsymbol{c}_i \boldsymbol{G}^{(ii)-1} \boldsymbol{k}^{(i)\top} - \beta = 0 \tag{11}$$

の二次方程式を解くことで得られる値である[8]．このとき，β, $\boldsymbol{k}^{(i)}$, $\boldsymbol{G}^{(ii)}$ はそれぞれ次式で与えられる．

$$\beta = \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_m(t) \tag{12}$$

$$\boldsymbol{G}^{(ii)} = \sum_{m=1}^{M} p_m(i,j) \sum_{t=1}^{T} \gamma_m(t) \boldsymbol{\xi}_m(t) \boldsymbol{\xi}_m(t)^\top \tag{13}$$

$$\boldsymbol{k}^{(i)} = \sum_{m=1}^{M} \boldsymbol{p}_m(i) \boldsymbol{\mu}_m^{(X)} \sum_{t=1}^{T} \gamma_m(t) \boldsymbol{\xi}_m(t)^\top - \sum_{j=1, j\neq i}^{d} \boldsymbol{w}_j \boldsymbol{G}^{(ij)} \tag{14}$$

ここで $\boldsymbol{p}_m(i)$ と $p_m(i,j)$ はそれぞれ共分散行列の逆行列 $\boldsymbol{\Sigma}_m^{(XX)-1}$ の i 行成分と (i,j) 要素を示す．また，$\gamma_m(t)$ は時間 t における生起確率

$$\gamma_m(t) = P(m | \boldsymbol{X}_t, \boldsymbol{\lambda}) \tag{15}$$

を示す．

本稿では CMLLR で推定された変換行列をモデル空間に適用する．適応されたモデルパラメータは以下の形で与えられる．

$$\hat{\boldsymbol{\mu}}_m^{(Z)} = \begin{bmatrix} \boldsymbol{A}' \boldsymbol{\mu}_m^{(X)} - \boldsymbol{b}' \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix} \tag{16}$$

$$\hat{\boldsymbol{\Sigma}}_m^{(Z)} = \begin{bmatrix} \boldsymbol{A}' \boldsymbol{\Sigma}_m^{(XX)} \boldsymbol{A}'^\top & \boldsymbol{A}' \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{A}'^\top & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \tag{17}$$

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，$\boldsymbol{A}' = \boldsymbol{A}^{-1}$，$\boldsymbol{b}' = \boldsymbol{A}' \boldsymbol{b}$ である．

表 1 実験的評価に用いる学習時と変換時の音声データ．図 1 に示す異なる収録機器（(a), (b), (c)）を用いて同時収録された肉伝導音声および空気伝導音声を，それぞれ肉伝導音声 a，b，c，空気伝導音声 A，B，C とする．

Table 1 Several conditions evaluated in experiments. Body-conducted speech (a, b, and c) and air-conducted speech (A, B, and C) are simultaneously recorded in three different conditions using different recording devices ((a), (b), and (c) in Figure 1)

Condition	Parallel data set in training	Body-conducted speech in conversion
Fully-matched	a-A	a
	b-B	b
	c-C	c
Mismatched	a-A	b or c
	b-B	a or c
	c-C	a or b
Adapted	a-A	b or c (adapted to a)
	b-B	a or c (adapted to b)
	c-C	a or b (adapted to c)
Matched	a-B	a
	a-C	a
	b-A	b
	b-C	b
	c-A	c
	c-B	c

5. 評価実験

5.1 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用いて，肉伝導音声と空気伝導音声を同時に収録する．発話内容は，音素バランス文 100 文である．話者は日本人男性 1 名である．図 1 に示した 3 種類の NAM マイクロフォンを用いて，別々に収録を行う．収録した音声は以下の通りである．

- マイク (a) を用いて収録された肉伝導音声 a と，同時収録された空気伝導音声 A
- マイク (b) を用いて収録された肉伝導音声 b と，同時収録された空気伝導音声 B
- マイク (c) を用いて収録された肉伝導音声 c と，同時収録された空気伝導音声 C

この時，肉伝導音声と空気伝導音声共に，サンプリング周波数は 8 kHz とする．各マイクロフォンで収録した音声 100 文の内，50 文を学習及び適応に，残り 50 文を評価に用いる．

本稿では，スペクトル特徴量として，STRAIGHT[9] で抽出された 0 次から 16 次までの

© 2010 Information Processing Society of Japan

Vol.2010-SLP-82 No.1
2010/7/22

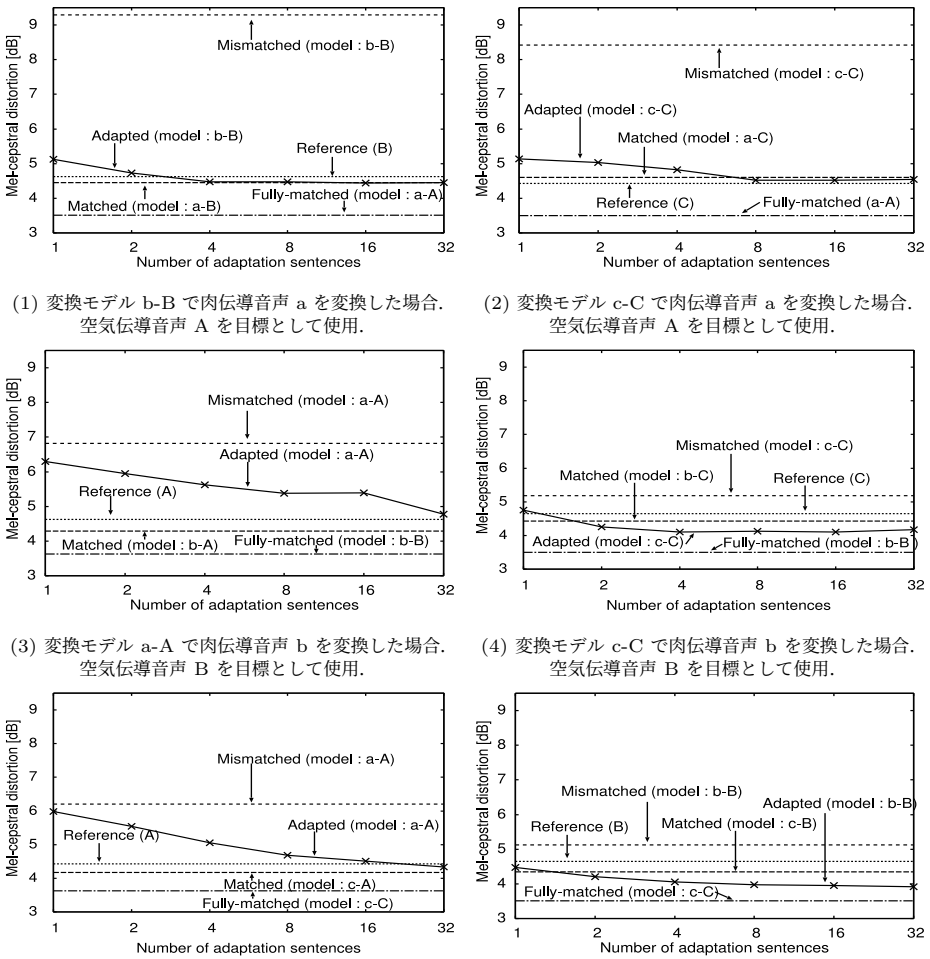
メルケプストラムを用いる．スペクトル推定用 GMM の入力特徴量として，スペクトル特徴量を当該フレーム及びその前後 4 フレームを結合し，PCA により 34 次元まで圧縮したスペクトルセグメント特徴量を用いる．フレームシフトは，5 ms とする．GMM の混合数は，32 とする．また，非周期成分及び F_0 についても STRAIGHT[10] により抽出する．

表 1 に，変換モデル学習時に用いる音声データと変換時に用いる音声データの組み合わせを示す．“Fully-matched” では，同時収録した肉伝導音声と空気伝導音声で学習した変換モデル（例えば，a-A）を用いて，学習時と同じ NAM マイクロフォンで収録した肉伝導音声（例えば，a）を変換する．“Mismatched” では，学習時（例えば，b-B）と異なる NAM マイクロフォンで収録した肉伝導音声（例えば a）を変換し，“Adapted” では，その条件下で提案法による補正処理を行う．補正の際には，適応文として，1, 2, 4, 8, 16, 32 文の肉伝導音声を用いる．なお，“Fully-matched” では，学習用音声と評価用音声は同一セッション内で収録された音声であるが，“Mismatced” 及び “Adapted” では，学習用音声と評価用音声は別々の日時に収録した音声となる．同じ話者の空気伝導音声ではあるが，収録日時の違いによる声質の変動が存在する．そのため，提案法による補正処理の上限をより明確にするために，“Matched” では，別々に収録された肉伝導音声と空気伝導音声で学習した変換モデル（例えば，a-B）を用いて，学習時と同じマイクロフォンで収録した肉伝導音声 (a) を変換する．

客観評価実験では，変換時の肉伝導音声と同時収録した空気伝導音声を評価用目標音声とし，変換音声と評価用目標音声のメルケプストラムひずみを求め，スペクトル変換精度を評価する．主観評価実験では，6 種類の変換音声，すなわち，“Fully-matched” における変換音声，“Mismatched” における変換音声，“Adapted” における 2, 8, 16 文で補正を行った際の変換音声，“Matched” における変換音声に対し，5 段階平均オピニオン評定により音質の評価を行った．オピニオンスコアは，(1：非常に悪い ～ 5：非常に良い) とした．被験者は，男性 5 名，女性 2 名の計 7 名であり，一人あたり各種 18 サンプル，計 108 サンプル受聴した．

5.2 実験結果

図 2 に客観評価実験の結果を示す．図 2 から，すべての “Mismatched” において，メルケプストラムひずみが “Fully-matched” より大きくなっていることが確認できる．この結果により，肉伝導音声変換の性能が収録条件に大きな影響を受けていることを確認できる．この劣化は，提案法（“Adapted”）により大幅に改善することができ，1 文のみを適応に用いた場合でもメルケプストラムひずみは，“Mismatched” と比較して大きく低下する．



(1) 変換モデル b-B で肉伝導音声 a を変換した場合．空気伝導音声 A を目標として使用．

(2) 変換モデル c-C で肉伝導音声 a を変換した場合．空気伝導音声 A を目標として使用．

(3) 変換モデル a-A で肉伝導音声 b を変換した場合．空気伝導音声 B を目標として使用．

(4) 変換モデル c-C で肉伝導音声 b を変換した場合．空気伝導音声 B を目標として使用．

(5) 変換モデル a-A で肉伝導音声 c を変換した場合．空気伝導音声 C を目標として使用．

(6) 変換モデル b-B で肉伝導音声 c を変換した場合．空気伝導音声 C を目標として使用．

図 2 個々の条件における適応データ数とメルケプストラムひずみの関係
Fig. 2 Mel-cepstral distortion as a function of the number of adaptation sentences in each condition.

© 2010 Information Processing Society of Japan

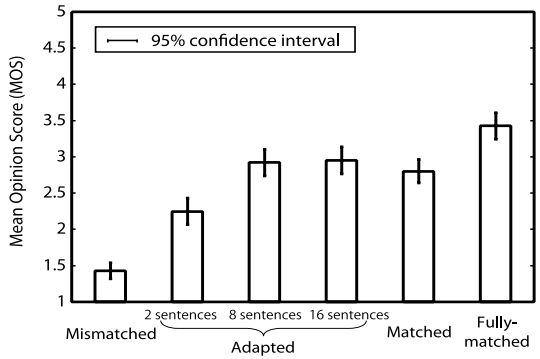


図 3 音質に関する主観評価実験結果
Fig. 3 Result of opinion test on speech quality.

また，適応文数の増加に伴い，ひずみが低減していき，“Matched” 条件で示される理想的な結果と同等の結果が得られることが分かる．このことから，提案法は，収録条件の違いにより生じる肉伝導音声の音響特性変化を効果的に補正することができ，また，1 クラスの CMLLR でも十分に高い性能が得られることが確認できる．なお，“Fully-matched” と “Matched” には明らかな差が見られる．参考までに，異なる日時に収録された空気伝導音声間におけるメルケプストラムひずみ（Reference）についても，図 2 に示す．今回収録した話者はプロナレータではなく，収録日時の違いによる声質変動が大きいことが分かる．また，“Adapted” と “Matched” は共に，“Reference” と同等のメルケプストラムひずみを示していることから，ひずみという観点から見れば，十分な変換精度が得られていることが分かる．

図 3 は，主観評価実験の結果を示す．主観評価においても，図 2 における “Mismatched” と “Fully-matched” 差が観測でき，収録機器の違いや NAM マイクロフォンの圧着位置の変化は，重大な音質の低下を起こすことが確認できる．また，“Mismatched” と “Adapted” を比較することで，提案する音響特性補正法が，変換音声の音質を効果的に変換音声の音質を効果的に改善できることが分かる．適応文数が 2 文から 8 文に増加することで，さらなる音質改善がなされており，8 文の適応データを用いた際には，“Matched” における変換音声と同等のものが得られることが分かる．これは，客観評価結果と同様の結果である．なお，“Fully-matched” と “Matched” の間には，音質のおいても有意な差がある．この原因として，声質の異なる音声間において，スペクトルと非周期成分以外の特徴量を変換していないことや，学習時においてフレームアライメント処理が必要になることなどが考えられる．

6. ま と め

本稿では，肉伝導音声変換おいて，異なる収録条件によって生じる音響特性の違いを補正するために，CMLLR に基づく教師なし音響特性補正法を適用した．異なる肉伝導音声収録機器を用いて音声収録を行うことで，学習時と変換時における収録環境の違いが肉伝導音声変換の品質に与える影響について調査し，提案法による補正効果を評価した．客観評価と主観評価の実験結果により，1) 学習時と変換時で異なる収録機器を用いると大幅に変換性能が低下すること，2) 提案法により変換性能の低下を効果的に軽減できること，を示した．

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．

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, Virgin Islands, U.S.A , 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, K. Nakamura, H. Sekimoto and K. Shikano, “Voice Conversion for Various Types of Body Transmitted Speec”, *Proc. ICASSP*, pp. 3061–3064, Taipei, Taiwan, 2009.
4) Y. Stylianou, O. Capp ´e and E. Moulines, “Continuous Probabilistic Transform for Voice Conversion”,*IEEE Trans. Speech and Audio Processing* SAP, Vol. 6, No. 2, pp. 131–142, 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, 2007
6) D. Miyamoto, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Acoustic Compensation Mmethods for Body Transmitted Speech Conversion”, *Proc. ICASSP*, pp. 3901–3904, Taipei, Taiwan, 2009.
7) 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”, *Proc. MAVEBA*, Florence, Italy, Sept. 2001.
8) 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.
9) H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, “Restructuring Speech Represantations Using A Pitch-adaptive Time-frequency Smoothing Repetitive structure in Sounds”, *Speech Communication*, Vol. 27, No.3-4, pp. 187–207,1999.
10) 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.

© 2010 Information Processing Society of Japan