# リアルタイム肉伝導音声変換処理の DSP 上への実装 *

☆森口 拓人†，戸田 智基†，佐野 元明††，佐藤 宏††
グラム・ニュービッグ†，サクリアニ・サクティ†，中村 哲†
†奈良先端大・情報 ††フォスター電機

## 1 はじめに

非可聴つぶやき (noaudible murmur: NAM) マイクロフォン [1] を用いて収録される肉伝導音声を，統計的声質変換に基づき，より自然な音声へと変換する技術 [2] は秘匿性が高く，周囲に迷惑をかけないサイレント音声インターフェース [3] の 1 つとして有効である．同一話者による肉伝導音声と自然音声で構成される同一文対セット（パラレルデータ）を用いて，両音声の音響特徴量間の変換モデルを学習することにより，肉伝導音声から自然音声への変換を実現する．人対人のコミュニケーションへの応用を目指し，リアルタイム変換処理も提案されており，PC 上での動作が確認されている [4]．今後，本技術を実用化する上で，計算能力は限られるが，携行性に優れる DSP 等の小型デバイスへ実装することは有効であろう．

本稿では，肉伝導音声変換として，NAM からささやき声への変換を対象とし，DSP 上へのリアルタイム変換処理の実装に取り組む．演算量削減処理を導入することで，DSP 上でのリアルタイム動作が可能であることを示すとともに，得られる変換精度についても実験的に評価する．

## 2 リアルタイム肉伝導音声変換 [2]

本稿における肉伝導音声変換では，NAM のスペクトル特徴量からささやき声のスペクトル特徴量への変換を行う．

### 2.1 学習処理

時間フレーム $t$ における NAM のスペクトルセグメント特徴量を $\boldsymbol{X}_t$ とし，前後 $C$ フレームの情報を用いて，次式により抽出する．

$$\boldsymbol{X}_t = \boldsymbol{E}\left[\boldsymbol{x}_{t-C}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+C}^\top\right]^\top + \boldsymbol{f} \quad (1)$$

ここで $\boldsymbol{x}_t$ は時間フレーム $t$ におけるスペクトルパラメータ（メルケプストラム）を表し，精度は低いが高速に動作する分析手法により求める．また，$\boldsymbol{E}$ および $\boldsymbol{f}$ は各々変換行列およびバイアスベクトルを表し，学習データに対する主成分分析により求める．一方で，ささやき声のスペクトル特徴量として，$\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ を用いる．スペクトルパラメータ $\boldsymbol{y}_t$ の抽出には，演算量は多いが高精度な分析手法を用い，動的特徴量 $\Delta\boldsymbol{y}_t$ は $\Delta\boldsymbol{y}_t = \boldsymbol{y}_t - \boldsymbol{y}_{t-1}$ により計算する．

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 $\left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top$ を用いて，次式に示すとおり，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．

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t | \boldsymbol{\lambda}^{(X,Y)}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top; \boldsymbol{\mu}_m^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}\right) \quad (2)$$

ここで，$\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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 $\alpha_m$，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．$m$ 番目の分布において，平均ベクトル $\boldsymbol{\mu}_m^{(X,Y)}$ および共分散行列 $\boldsymbol{\Sigma}_m^{(X,Y)}$ は次式で表される．

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

ここで $\boldsymbol{\mu}_m^{(X)}$ および $\boldsymbol{\mu}_m^{(Y)}$ は NAM およびささやき声の平均ベクトルを表し，$\boldsymbol{\Sigma}_m^{(XX)}$，$\boldsymbol{\Sigma}_m^{(YY)}$，$\boldsymbol{\Sigma}_m^{(XY)}$ および $\boldsymbol{\Sigma}_m^{(YX)}$ は NAM およびささやき声の共分散行列，相互共分散行列を表す．

### 2.2 短遅延変換処理

時間フレーム 1 から $T$ までの NAM およびささやき声の音響特徴量系列をそれぞれ $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$，$\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$ とおく．このとき，変換後の静的特徴量系列 $\hat{\boldsymbol{y}} = \left[\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top\right]^\top$ は次式で計算される．

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} P\left(\boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right) \text{ subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \quad (4)$$

ここで，$\boldsymbol{W}$ は静的特徴量系列を静的・動的特徴量系列に写像する変換行列を表す [5]．短遅延変換処理では，各時間フレームにおいて準最適な分布を次式で決定する．

$$\hat{m}_t = \underset{m}{\operatorname{argmax}} P\left(m | \boldsymbol{X}_t, \boldsymbol{\lambda}^{(X,Y)}\right) \quad (5)$$

そして，式 (4) の最大化処理に対して，現在の時間フレーム $t$ までの準最適な分布系列 $[\hat{m}_1, \cdots, \hat{m}_t]$ とカルマンフィルタによる近似を導入することで，数フレーム前における変換静的特徴量 $\hat{\boldsymbol{y}}_{t-D}$（本稿では $D = 3$ 程度）を決定する [6]．

また，変換音声の品質を向上させるために，系列内変動（Global Variance: GV）をポストフィルタリング処理として考慮したリアルタイム変換処理を用いる [4]．ささやき声の音響特徴量系列に対する GV の平均ベクトル $\boldsymbol{\mu}^{(v)} = \left[\mu_1^{(v)}, \cdots, \mu_D^{(v)}\right]^\top$ と，学習データ中の NAM を GMM により変換して得られる音響特徴量系列に対する GV の平均ベクトル $\hat{\boldsymbol{\mu}}^{(v)} = \left[\hat{\mu}_1^{(v)}, \cdots, \hat{\mu}_D^{(v)}\right]^\top$ およびバイアスベクトル $\langle\hat{\boldsymbol{y}}\rangle = [\langle\hat{y}_1\rangle, \cdots, \langle\hat{y}_D\rangle]^\top$ を予め計算しておく．リアルタイム変換処理では，$d$ 次元目の変換静的特徴量 $\hat{y}_{t,d}$ を次式にて強調する．

$$\hat{y}_{t,d} = \mu_d^{(v)^{\frac{1}{2}}} \hat{\mu}_d^{(v)^{-\frac{1}{2}}} (\hat{y}_{t,d} - \langle\hat{y}_d\rangle) + \langle\hat{y}_d\rangle \quad (6)$$

### 2.3 共分散行列の対角化による演算量削減

肉伝導音声変換では全共分散行列が使用されるため，式 (5) に示す分布選択処理における演算量は多い．変換精度の劣化を抑えつつ演算量を削減する手法として，最尤基準に基づく共分散行列の対角化が有効である [4]．$m$ 番目の分布における共分散行列 $\boldsymbol{\Sigma}_m^{(XX)}$ は，全ての混合分布に共通の変換行列 $\boldsymbol{A}$ と各分布に依存する対角行列 $\boldsymbol{\Sigma}_{m,diag}^{(XX)}$ を用いて以下の式の様にモデル化する．

$$\boldsymbol{\Sigma}_m^{(XX)} \simeq \boldsymbol{A}^{-1} \boldsymbol{\Sigma}_{m,diag}^{(XX)} \boldsymbol{A}^{-\top} \quad (7)$$

---

* “Implementation of real-time body-conducted voice conversion on DSP”
by T. Moriguchi†, T. Toda†, M. Sano††, H. Sato††, G. Neubig†, S. Sakti†, S. Nakamura†
†Nara Institute of Science and Technology ††Foster Electronics

このモデル構造をとることにより，対角共分散行列を用いた際と同等の演算量を達成できる．

## 3 DSP 上への実装

リアルタイム変換処理では，特徴量抽出処理，変換処理，波形合成処理を分析フレームシフト長の時間内で終わらせる必要がある．2 節で述べた肉伝導音声変換は PC などの十分な計算リソースのある環境では，2．3 節で述べた共分散行列の対角化を用いずとも，十分にリアルタイムで動作する．一方，DSP のように計算リソースが限られる環境においてリアルタイムで動作させるためには，共分散行列の対角化に加え，さらなる演算量削減が必要となる．

DSP 上への実装を行うために，まずプログラムの高速化（除算から乗算への置き換え，ビットシフト演算処理への置き換え，FFT における回転因子のテーブル化など）を行う．演算量の多い対数計算や指数計算に対しては，区分線形関数による近似計算を導入し，ケプストラムからメルケプストラムへ変換を行うオールパスフィルタ演算においては，高次のケプストラム係数を 0 で近似する．また，DSP 用コンパイラの組み込み関数の使用による高速化も行う．

上記の処理に加え，さらに演算量を削減するためには，アルゴリズムの変更が必要となる．具体的には，分析フレームシフトを長くすることで，特徴量抽出処理および変換処理の実行回数を減らす．

## 4 実験的評価

### 4.1 実験条件

同一話者に対して，NAM マイクロフォンによる NAM 収録と，空気伝導マイクによるささやき声収録を行う．話者は男性 2 名，女性 1 名であり，各話者において，学習データとして ATR 音素バランス文セット中の約 50 文，評価データとして新聞記事約 150 文を用いる．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．スペクトル分析は NAM に対しては FFT 分析を用い，ささやき声に対してはメルケプストラム分析 [7] を用いる．分析フレームシフトは 5 ms および 10 ms とする．5 ms シフトの際には，スペクトルセグメント特徴量抽出には前後 4 フレーム（C=4）を使用し，短遅延変換処理における遅延フレーム数は 3（D=3）とする．一方で，10 ms シフトの際には，C=2，D=2 とする．浮動小数点版の DSP として，TI 社の TMS320C6748（375 MHz）を用いる．

以下のシステムに対して，DSP 上での処理時間および変換精度を評価する．

- Baseline：2.2 節で述べた従来の変換システム（分析フレームシフトは 5 ms）
- Diag：Baseline に対して 2.3 節で述べた共分散行列の対角化を導入したシステム
- Diag ＋ Fast：Diag に対して，3 節で述べた演算量削減処理を導入したシステム
- Diag+Fast+10ms：Diag+Fast において，分析フレームシフトを 10 ms としたシステム

GMM の混合数は 32 とし，特定話者モデルを用いる．

### 4.2 実験結果

各システムにおいて，特徴量抽出処理，変換処理，波形合成処理に要する時間（リアルタイムファクタ：処理時間／分析フレームシフト長）を，Fig. 1 に示す．共分散行列の対角化（Diag）を用いることで変換処理時間が大幅に減少し，演算量削減処理（Diag ＋ Fast）を導入することで特徴量抽出処理時間が大幅に減少する．しかしながら，リアルタイムで動作するまでには至らない．さらに，分析フレームシフト長を 10 ms にすることで，リアルタイム動作を実現出来ることが分かる．

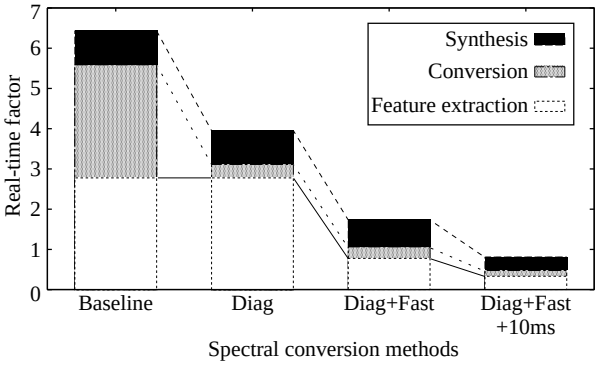


Fig. 1 Real-time factor calculated as (processing time)/(shift length) in each system

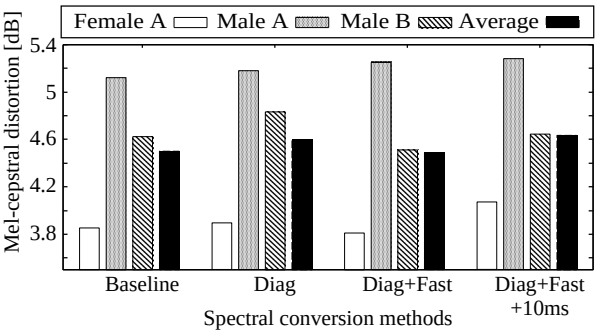


Fig. 2 Mel-cepstral distortion in each system

各システムの変換精度を評価するために，個々の話者に対するメルケプストラム歪みとその平均値を Fig. 2 に示す．なお，変換前の歪みは平均値で 9.28 dB である．リアルタイム動作を実現するための高速化により，メルケプストラム歪みが若干大きくなる傾向がみられるが，依然高い変換精度が保たれていることが分かる．

## 5 終わりに

本稿では，リアルタイム肉伝導音声変換に対して演算量削減処理を導入し，DSP 上への実装を行った．実験評価の結果，大幅な変換精度劣化を生じさせることなく，DSP 上でリアルタイム動作する肉伝導音声変換処理を実現できることが分かった．

**謝辞** 本研究の一部は，科研費補助金若手研究（A）および JST A-STEP 探索タイプにより実施したものである．

## 参考文献

[1] Y. Nakajima *et al.*, *IEICE Trans. Information and Systems*, Vol. E89-D, No. 1, pp. 1–8, 2006.

[2] T. Toda *et al.*, *IEEE*, Vol. 20, No. 9, 2012.

[3] B. Denby *et al.*, *Speech Communication*, Vol. 52, No. 4, pp. 270–287, 2010.

[4] T. Toda *et al.*, *Proc. INTERSPEECH*, Portland, USA , Sep. 2012.

[5] K. Tokuda *et al.*, *Proc. ICASSP*, pp. 1315–1318, Istanbul, Turkey, June 2000.

[6] T. Muramatsu *et al.*, *Proc. INTERSPEECH*, pp. 1076–1079, Brisbane, Australia, Sep. 2008.

[7] 徳田 他, 信学論 (A), Vol. J74-A, No. 8, pp. 1240–1248, 1991.