社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

統計的手法に基づくリアルタイム声質変換処理の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†

† 奈良先端科学技術大学院大学情報科学研究科，生駒市
†† フォスター電機株式会社
E-mail: †{takuto-m,tomoki,neubig,ssakti,s-nakamura}@is.naist.jp

あらまし　肉伝導音声を用いたサイレント音声コミュニケーションや，無喉頭音声による代用発声において，統計的手法に基づくリアルタイム声質変換処理を用いた音声強調技術が研究されている．これまでは，主にノートPCなどの十分な計算リソースが得られる環境下での動作が確認されているが，本技術を実環境で用いるには，より傾向性に優れたデバイスの使用が望まれる．本報告では，サイレント音声コミュニケーションのための非可聴つぶやき（Non-Audible Murmur: NAM）からささやき声への変換処理を，小型で低消費電力なデバイスであるDSP上へと実装する．DSP上でのリアルタイム変換処理を実現するために，変換精度を保ちつつ，演算量を削減する手法を提案する．実験的評価結果から，本手法の有効性を示す．
キーワード　統計的声質変換　非可聴つぶやき　リアルタイム処理　DSP　演算量削減

Implementation of real-time statistical voice conversion on a DSP

Takuto MORIGUCHI†, Tomoki TODA†, Motoaki SANO††, Hiroshi SANO††, Graham NEUBIG†,
Sakriani SAKTI†, and Satoshi NAKAMURA†

† Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma-shi, 630–0101, Japan
†† Foster Electric Company, Limited
E-mail: †{takuto-m,tomoki,neubig,ssakti,s-nakamura}@is.naist.jp

Abstract　Real-time statistical voice conversion is one of promising approach to developing technology for body–conducted speech in silent speech communication and alaryngeal speech produced by alternative speaking methods for laryngectomees. Although it has been successfully implemented on devices with sufficient computational resources such as laptop PCs, an implementation in enviromments with limited resources such as portable devices would greatly contribute its practical use. In this report, we implement real-time voice conversion from non-audible murmur into whisper on a DSP for silent speech interfaces. To achieve real-time processing, we propose some methods for reducing computational cost while keeping conversion accuracy high. We conduct an experimental evaluations, which shows effectiveness of the proposed methods.
Key words　statistical voice conversion, non-audible murmur, real-time processing, DSP, reduction of computational cost

1. まえがき

音声は基本的なコミュニケーション手段の1つであり，携帯電話などの普及により，その利便性はさらに改善され，いつでもどこでも使用することが容易となった．一方で，静かな環境下で音声を発声すると周囲の人に迷惑をかける，周囲に人がいる際には秘匿性の高い内容を発声するのに躊躇する，身体的な障害により発声機能が失われると音声を発声すること自体ができなくなる，といった音声コミュニケーションに内在する問題が浮き彫りとなっている．これらの解決を目指し，新たな収音デバイスや音声強調処理を用いた技術がいくつか研究されている [1] [2].

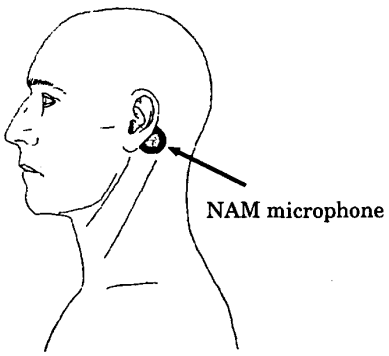

図 1　Setting position of NAM microphone

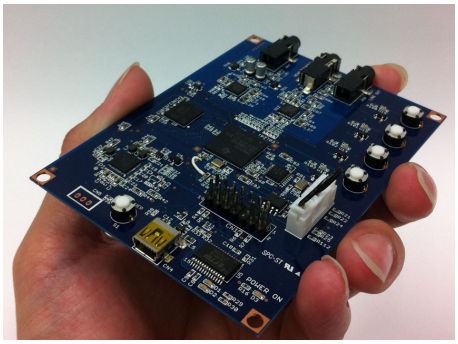

図 2　DSP used in this report

秘匿性が高く，周囲に迷惑をかけない発声を可能とするサイレント音声インタフェース [1] の一つとして，体内伝導音声の非可聴つぶやき（Non-Audible Murmur: NAM）が提案されている [3]．NAM は，「声帯振動ではなく気道の乱流雑音を音源とする無声呼気音が，発話器官の運動による音響的フィルタ特性変化により調音されて，人体頭部の主に軟部組織を伝導したもの」と定義される．気導音としては周囲の人が聞き取ることが困難なほど小さな無声音のつぶやきであるが，専用の収音機器である NAM マイクロフォンを図 1 に示すように耳介後方に圧着することで，体内を通じて体表から直接収録可能である．NAM マイクロフォンは，骨伝導マイクロフォンなどの他の体内伝導マイクロフォンと比較して，比較的高い品質の体内伝導音声を収録することができ，空気伝導マイクロフォンでは周囲の雑音に埋もれて収録が困難となる NAM のようなわずかな音声も収録可能といった特徴を持つ．一方で，NAM は体内を通じて収録されるため，口唇での放射特性の欠如や肉伝導によるローパス特性の影響により，空気伝導の音声と比較してその周波数特性は大きく異なる．その結果，自然性および明瞭性は大きく劣化する．

NAM の音質および明瞭性を改善する手法として，統計的声質変換に基づく NAM から自然な音声への変換法が研究されている [4]．自然な音声として通常音声への変換が究極的な目標ではあるものの，無声音声である NAM から基本周波数の推定を必要とする通常音声への変換処理は極めて困難であり，十分な変換精度が得られていない．そこで，基本周波数の推定を必要としない手法として，自然な無声音声であるささやき声への変換が提案されており，自然性および明瞭性を大幅に改善出来ることが確認されている．さらに，本技術を人対人のコミュニケーションへと応用するために，リアルタイム変換処理が提案されており，ノート PC など十分な計算リソースが得られる環境において，その動作が確認されている [5]．本技術をさらに発展し，歩行しながらの利用や長時間の利用など実環境での使用を可能とするためには，より傾向性に優れたデバイス上への実装が望まれる．

本報告では，携行性の高いデバイスとして，小型で低消費電力の DSP（図 2）に着目し，NAM からささやき声へのリアルタイム変換処理を実装する．DSP の演算能力は，ノート PC と比較すると限定される．そこで，変換精度を保ちつつ演算量を削減する手法を提案することで，DSP 上でのリアルタイム動作を実現する．実験的評価により，DSP による変換処理においても，従来のオフライン変換処理と同等の品質の変換音声が得られることを示す．

以下，2 節でリアルタイム肉伝導音声変換処理について述べ，3 節で演算量削減について述べる．4 節で実験的評価について述べ，最後に 5 節で本報告の結論と今後の課題について述べる．

2. NAM からささやき声へのリアルタイム変換処理

統計的手法に基づく声質変換に基づき，NAM からささやき声への変換 [4] を行う．予め収録された NAM とささやき声の同一文発話データ（パラレルデータ）を用いて，変換モデルを事前に学習することで，任意の発話に対する NAM からささやき声へのリアルタイム変換処理 [5] が可能となる．図 3 に，分析窓長を 25 ms，分析フレームシフト長を 5 ms とした際の NAM からささやき声へのリアルタイム変換処理を示す．各フレームにおいて，特徴量抽出処理，特徴量変換処理，波形合成処理が行われる．リアルタイム動作を実現するためには，これら一連の処理を，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．以下では，変換モデルの学習処理およびリアルタイム変換処理について述べる．

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（$D^{(X)}$ 次元ベクトル）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．

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

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を表し，本報告ではメルケプストラムを用いる．メルケプストラムの計算には，精度は低いが高速に動作する分析処理（FFT 分析，対数変換，一次の全域通過フィルタによる周波数軸変換）を用いる．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．$\top$ は転置を表す．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，高精度な分析手法として最尤推定に基づくメルケプストラム分析 [6] 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，入力特徴量 $\boldsymbol{X}_t$ と出力特徴量 $\boldsymbol{Y}_t$ の対応付けを行った結合ベクトル $\left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top$

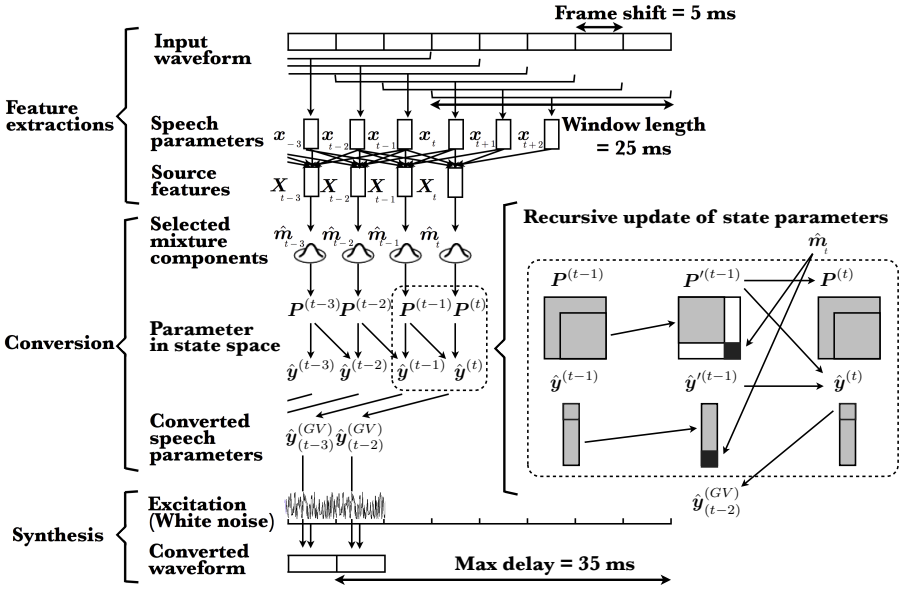

図 3　Frame-by-frame processing in real-time voice conversion ($C = 2$, $L = 2$)

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．

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

ここで，$\mathcal{N}\big(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}\big)$ 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 α_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．

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

$$\boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \quad (4)$$

ここで $\boldsymbol{\mu}_m^{(X)}$ および $\boldsymbol{\mu}_m^{(Y)}$ は入力特徴量および出力特徴量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 $\boldsymbol{\Sigma}_m^{(YX)}$ は入力特徴量および出力特徴量の共分散行列，相互共分散行列を表す．

2.2　リアルタイム変換処理

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．

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P\Big(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\Big) \text{ subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \quad (5)$$

ここで，$\boldsymbol{W}$ は静的特徴量系列 $\boldsymbol{y}$ を静的・動的特徴量系列 $\boldsymbol{Y}$ に写像する変換行列を表す [6]．リアルタイム変換処理では，短遅延変換法 [7] を導入することで，式 (5) を近似的に解く．まず，各時間フレームにおいて準最適な分布系列 $\hat{\boldsymbol{m}} = \{\hat{m}_1, \cdots, \hat{m}_T\}$ を次式で決定する．

$$\begin{aligned} \hat{m}_t &= \underset{m}{\operatorname{argmax}}\, P\Big(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(X,Y)}\Big) \\ &= \underset{m}{\operatorname{argmax}}\, \mathcal{N}\Big(\boldsymbol{X}_t; m, \boldsymbol{\lambda}^{(X,Y)}\Big) \end{aligned} \quad (6)$$

そして，式 (5) 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とカルマンフィルタによる近似を導入することで，L フレーム前における変換静的特徴量 $\hat{\boldsymbol{y}}_{t-L}$（本報告では $L = 3$ 程度）を決定する [7]．$\boldsymbol{X}_t$ が与えられた際の $\boldsymbol{Y}_t$ に対する条件付き確率密度関数と各フレームの m 番目の混合要素は正規分布によりモデル化され，その平均ベクトルと共分散行列はそれぞれ次式で与えられる．

$$\boldsymbol{\mu}_{m,t}^{(Y|X)} = \boldsymbol{\mu}_m^{(Y)} + \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)^{-1}} \Big(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}\Big) \quad (7)$$

$$\boldsymbol{\Sigma}_m^{(Y|X)} = \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)^{-1}} \boldsymbol{\Sigma}_m^{(XY)} \quad (8)$$

共分散行列 $\boldsymbol{\Sigma}_m^{(Y|X)}$ の対角要素のみを用いるため，$\hat{\boldsymbol{y}}_{t-L}$ は次元毎に独立に求める．$(L+1) \times (L+1)$ 次元の状態共分散行列 $\boldsymbol{P}_d^{(0)}$ と $(L+1)$ 次元の状態ベクトル $\hat{\boldsymbol{y}}_d^{(0)}$ をそれぞれ零行列と零ベクトルに初期化する．そして，それらを各時間フレームごとに次式で再帰的に更新を行う．

$$\boldsymbol{P'}_d^{(t-1)} = \boldsymbol{J}_L \boldsymbol{P}_d^{(t-1)} \boldsymbol{J}_L^\top + \mathrm{diag}\Big[\boldsymbol{0}_{1\times L}, \Sigma_{m_t,d}^{(y|X)}\Big] \quad (9)$$

$$\hat{\boldsymbol{y}}_d^{\prime(t-1)} = \boldsymbol{J}_L \hat{\boldsymbol{y}}_d^{(t-1)} + \Big[\boldsymbol{0}_{1\times L}, \mu_{m_t,t,d}^{(y|X)}\Big]^\top \quad (10)$$

$$\boldsymbol{P}_d^{(t)} = \Big(\boldsymbol{I} - \boldsymbol{k}_d^{(t)} \boldsymbol{w}_L\Big) \boldsymbol{P'}_d^{(t-1)} \quad (11)$$

$$\hat{\boldsymbol{y}}_d^{(t)} = \hat{\boldsymbol{y}}_d^{\prime(t-1)} + \boldsymbol{k}_d^{(t)} \Big(\mu_{m_t,t,d}^{(\Delta y|X)} - \boldsymbol{w}_L \hat{\boldsymbol{y}}_d^{\prime(t-1)}\Big) \quad (12)$$

ここで $(L+1)$ 次元のベクトル $\boldsymbol{k}_d^{(t)}$ はカルマンゲインを表し，次式で計算される．

$$\boldsymbol{k}_d^{(t)} = \boldsymbol{P}_d^{(t-1)}\boldsymbol{w}_t^\top\left(\Sigma_{m_t,d}^{(\Delta y|X)} + \boldsymbol{w}_L\boldsymbol{P}_d^{(t-1)}\boldsymbol{w}_L^\top\right)^{-1} \tag{13}$$

そして $(L+1)$ 次元の行ベクトル $\boldsymbol{w}_L$ および $(L+1)\times(L+1)$ 次元の行列の $\boldsymbol{J}_L$ はそれぞれ次式で与えられる．

$$\boldsymbol{w}_L = \Big[\boldsymbol{0}_{1\times(L-1)}, -1, 1\Big], \boldsymbol{J}_L = \begin{bmatrix} 0 & \boldsymbol{I}_{L\times L} \\ 0 & \boldsymbol{0}_{1\times L} \end{bmatrix} \tag{14}$$

平均ベクトル $\boldsymbol{\mu}_{m,t}^{(Y|X)}$ における d 次元目の静的特徴量に対する要素 $\mu_{m,t,d}^{(y|X)}$ と，共分散行列 $\boldsymbol{\Sigma}_m^{(Y|X)}$ における d 次元目の静的特徴量に対する対角要素 $\Sigma_{m,d}^{(y|X)}$ が，式 (9)，(10) に示す状態共分散行列と状態ベクトルを予測するのに用いられる．また，それらの動的特徴量に対する要素 $\mu_{m,t,d}^{(\Delta y|X)}$ と $\Sigma_{m,d}^{(\Delta y|X)}$ は，式 (13) に示すカルマンゲインの最適化と式 (11)，(12) に示す状態共分散行列と状態ベクトルの更新を行うために用いられる．更新された状態ベクトル $\hat{\boldsymbol{y}}_d^{(t)}$ の一次元目の要素が，フレーム $t-L$ における変換静的特徴量の d 次元目の要素 $\hat{y}_{t-L,d}$ として出力される．結果，フレーム遅延量は，式 (1) における先読みフレーム数 C と合わせて，$L+C$ となる．

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) [8] を考慮したポストフィルタ処理 [5] を導入する．GV ベクトル $\boldsymbol{v}^{(y)} = \left[v_1^{(\boldsymbol{y})}, \ldots, v_{D(y)}^{(\boldsymbol{y})}\right]^\top$ は，ささやき声の静的特徴量系列に対して，各発話ごとに次式で計算される．

$$v_d^{(\boldsymbol{y})} = \frac{1}{T}\sum_{t=1}^{T}\left(y_{t,d} - \frac{1}{T}\sum_{\tau=1}^{T} y_{\tau,d}\right) \tag{15}$$

ここで $y_{t,d}$ はフレーム t のささやき声の静的特徴量ベクトル $\boldsymbol{y}_t$ の d 次元目の要素である．事前に学習データから，ささやき声の特徴量系列に対する GV の平均ベクトル $\boldsymbol{\mu}^{(v)} = \left[\mu_1^{(v)}, \cdots, \mu_D^{(v)}\right]^\top$ を求めておく．また，学習データ中の NAM に対して，短遅延変換処理を行うことで変換特徴量系列を求め，それに対するバイアスの平均ベクトル $\langle\hat{\boldsymbol{y}}\rangle = [\langle\hat{y}_1\rangle, \cdots, \langle\hat{y}_D\rangle]^\top$ および GV の平均ベクトル $\hat{\boldsymbol{\mu}}^{(v)} = \left[\hat{\mu}_1^{(v)}, \cdots, \hat{\mu}_D^{(v)}\right]^\top$ を予め計算しておく．リアルタイム変換処理では，これらの平均ベクトルの値を用いて，d 次元目の変換静的特徴量 $\hat{y}_{t,d}$ をフレームごとに次式にて強調する．

$$\hat{y}_{t,d}^{(GV)} = \mu_d^{(v)^{\frac{1}{2}}}\hat{\mu}_d^{(v)^{-\frac{1}{2}}}\left(\hat{y}_{t,d} - \langle\hat{y}_d\rangle\right) + \langle\hat{y}_d\rangle \tag{16}$$

3. 演算量削減

NAM からささやき声へのリアルタイム変換処理を DSP 上へと実装する．リアルタイム動作を実現するために，変換音声の品質劣化を最小限に抑えつつ，演算量の削減を行う．

3.1 共分散行列の対角化による演算量削減

NAM からささやき声への変換処理では，全共分散行列が使用されるため，式 (6) 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，肉伝導有声音声に対する変換処理でその有効性が確認されている最尤基準に基づく共分散行列の対角化 [5] を導入する

本報告では，制約付き最尤線形回帰（Constrained Maximum Likelihood Linear Regression: CMLLR）に基づく話者適応学習 [7] の枠組みを応用して，共分散行列の対角化を行う．対角化が行われた際の結合確率密度関数は次式で与えられる．

$$\begin{aligned} &P\Big(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}^{(X,Y)}, \boldsymbol{A}, \boldsymbol{b}\Big) \\ &= \sum_{m=1}^{M}\alpha_m\mathcal{N}\Big(\boldsymbol{X}_t; \hat{\boldsymbol{\mu}}_m^{(X)}, \hat{\boldsymbol{\Sigma}}_m^{(XX)}\Big)\mathcal{N}\Big(\boldsymbol{Y}_t; \boldsymbol{\mu}_{m,t}^{(Y|X)}, \boldsymbol{\Sigma}_m^{(Y|X)}\Big) \end{aligned} \tag{17}$$

ここで，$\boldsymbol{A}$ および $\boldsymbol{b}$ は CMLLR の変換パラメータである．入力特徴量に対する確率密度関数に対して共分散行列の対角化が行われており，その平均ベクトル $\hat{\boldsymbol{\mu}}_m^{(X)}$ および共分散行列 $\hat{\boldsymbol{\Sigma}}_m^{(XX)}$ は次式にて表される

$$\hat{\boldsymbol{\mu}}_m^{(X)} = \boldsymbol{A}^{-1}\boldsymbol{\mu}_m^{(X')} - \boldsymbol{A}^{-1}\boldsymbol{b} \tag{18}$$

$$\hat{\boldsymbol{\Sigma}}_m^{(XX)} = \boldsymbol{A}^{-1}\boldsymbol{\Lambda}_m^{(X'X')}\boldsymbol{A}^{-\top} \tag{19}$$

各混合要素に依存する全共分散行列 $\hat{\boldsymbol{\Sigma}}_m^{(XX)}$ は，各混合要素に依存する対角共分散行列 $\boldsymbol{\Lambda}_m^{(X'X')}$ と全混合要素で共通の変換行列 $\boldsymbol{A}$ で表される．CMLLR の変換パラメータ $\{\boldsymbol{A}, \boldsymbol{b}\}$ と各混合要素に依存するパラメータ $\{\boldsymbol{\Lambda}_m^{(X'X')}, \boldsymbol{\mu}_m^{(X')}\}$ は，学習データを用いて最尤推定により最適化される．なお，その際に用いる十分統計量は，元の全共分散行列を持つ結合 GMM と学習データ中の結合ベクトルを用いて計算する．

リアルタイム変換処理では，全混合要素に対して共通の CMLLR 変換パラメータを，モデルパラメータに対してではなく，特徴量ベクトルに対して適用する．

$$\boldsymbol{X}_t' = \boldsymbol{A}\boldsymbol{X}_t + \boldsymbol{b} \tag{20}$$

この計算を行うために，CMLLR の変換パラメータを，式 (1) で使用するパラメータ $\boldsymbol{E}$ と $\boldsymbol{f}$ に対して，次式のように事前に適用しておく

$$\boldsymbol{E}' = \boldsymbol{A}\boldsymbol{E} \tag{21}$$

$$\boldsymbol{f}' = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{b} \tag{22}$$

また，式 (7) で表される条件付き確率密度関数の平均ベクトルの計算においても，変換特徴量ベクトル $\boldsymbol{X}_t'$ に対応させるため，次式のようにパラメータを事前に変形しておく．

$$\boldsymbol{\Sigma}_m^{(YX')} = \boldsymbol{\Sigma}_m^{(YX)}\boldsymbol{A}^\top \tag{23}$$

$$\boldsymbol{\Sigma}_m^{(X'X')} = \boldsymbol{A}\boldsymbol{\Sigma}_m^{(XX)}\boldsymbol{A}^\top \tag{24}$$

$$\boldsymbol{\mu}_m^{(X')} = \boldsymbol{A}\boldsymbol{\mu}_m^{(X)} + \boldsymbol{b} \tag{25}$$

これにより，変換時における式 (1) と式 (7) に関する演算量は一切増加しない．一方で，変換特徴量ベクトル $\boldsymbol{X}'$ および対角化処理において最適化されたモデルパラメータ $\hat{\boldsymbol{\Lambda}}^{(X'X')}$ を用いることで，式 (6) は次式にて表せる．

$$\hat{m}_t = \underset{m}{\operatorname{argmax}}\,\alpha_m\sqrt{|\boldsymbol{A}|^2}\mathcal{N}\Big(\boldsymbol{X}_t'; \hat{\boldsymbol{\mu}}_m^{(X')}, \hat{\boldsymbol{\Lambda}}_m^{(X'X')}\Big) \tag{26}$$

計算オーダーは，全共分散行列 $\mathbf{\Sigma}_m^{(XX)}$ を用いた場合の $O(D^{(X)^2})$ と比較し，対角共分散行列 $\hat{\mathbf{\Lambda}}_m^{(X'X')}$ の使用により $O(D^{(X)})$ となる．

3.2 プログラムの高速化

DSP 上でのリアルタイム動作を実現するために，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）を行う．また，DSP 用コンパイラの組み込み関数を使用することで高速化を行う．これらの処理は，基本的に変換精度に一切影響を与えない．

さらなる高速化を行うために，演算量の多い指数計算や対数計算に対しては，区分線形関数による近似計算を導入する．また，ケプストラムからメルケプストラムへの変換を行う一次の全域通過フィルタ演算においては，高次のケプストラム係数を 0 で近似することで，演算量を削減する．これらの処理は，変換精度に影響を与える可能性がある．これについては，次節の実験的評価において調査する．

3.3 分析フレームシフト長の変更

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ため，分析フレームシフト長を長くすることで，必要な演算量を大幅に削減することができる．例として，分析フレームシフト長を 5 ms から 10 ms に変更した際には，特徴量抽出部と変換部については，処理回数は変わらないため，リアルタイムファクタ（処理時間／分析フレームシフト長の時間）は半減する．一方で，合成部に関しては，分析フレームシフト長に相当する変換音声波形をフィルタリング処理で生成するため，本処理のリアルタイムファクタは変化しない．なお，分析フレームシフト長の変更が変換精度に与える影響については，次節の実験的評価において調査する．

4. 実験的評価

DSP 上に実装した NAM からささやき声へのリアルタイム変換処理の有効性を検証するために，客観評価実験及び主観評価実験を行う．

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 [9] 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．GMM の混合数は 32 とし，特定話者モデルを用いる．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．

- Offline：オフライン変換処理（GV を考慮した系列単位のバッチ変換処理）を用いるシステム [4]
- Baseline：2.2 節で述べた従来のリアルタイム変換システム（分析フレームシフトは 5 ms）[5]
- Diag：Baseline に対して 3.1 節で述べた共分散行列の対角化を導入したシステム
- Fast：Baseline に対して，3.2 節で述べたプログラムの高速化を導入したシステム
- 10ms：Baseline において，分析フレームシフト長を 10 ms としたシステム
- Diag+Fast：Diag に対して 3.2 節で述べたプログラムの高速化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長を 10 ms としたシステム
- DSP：Diag+Fast+10ms を DSP 上で動作させたシステム

なお，DSP 上で変換処理を動作させる際には，NAM データのライン入力を行う．本稿で用いる DSP の音声入力システムの特性として，高域が減衰する傾向がある．そのため，事前に音声入力システムのインパルス応答を測定しておき，それを学習データの NAM に畳み込むことで，入力特性を加味した NAM データを作成する．得られた NAM データを用いて GMM を学習することで，入力特性による変換音声の品質劣化を解消する．

客観評価実験として，個々のシステムにおいて，各フレームにおける処理時間と分析フレームシフト長の時間比から計算されるリアルタイムファクタを計算する．また，スペクトル変換精度を評価するために，変換特徴量と目標特徴量間のメルケプストラムひずみを計算する．主観評価実験では，各システム（Diag+Fast は除く）による変換音声の聞き取りやすさに関して，オピニオン評価を行う．オピニオンスコアは 5 段階（1：非常に悪い〜5：非常に良い）に設定する．被験者は男性 10 名であり，1 人あたり各システムにつき 15 サンプル，計 105 サンプルを受聴する．提示するサンプルについては，評価データの中から被験者毎にランダムに選択する．

4.2 リアルタイム性能に関する客観評価結果

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，図 4 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，さらにプログラムの高速化（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．

4.3 変換精度に関する客観評価結果

オフライン変換システム（Offline）と比較して，カルマン

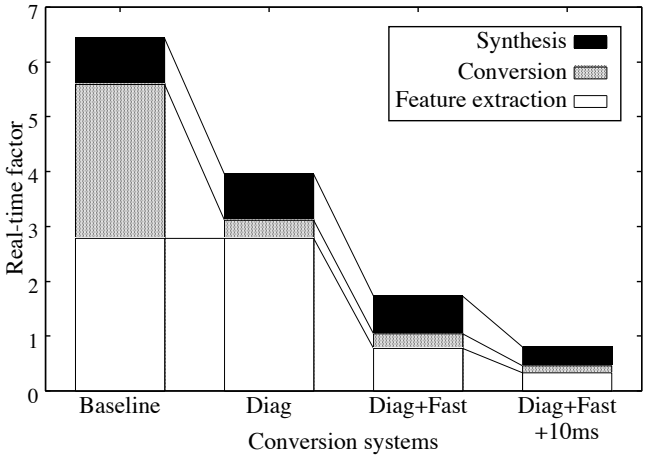

図 4 Real-time factor calculated as (processing time)/(shift length) in each system

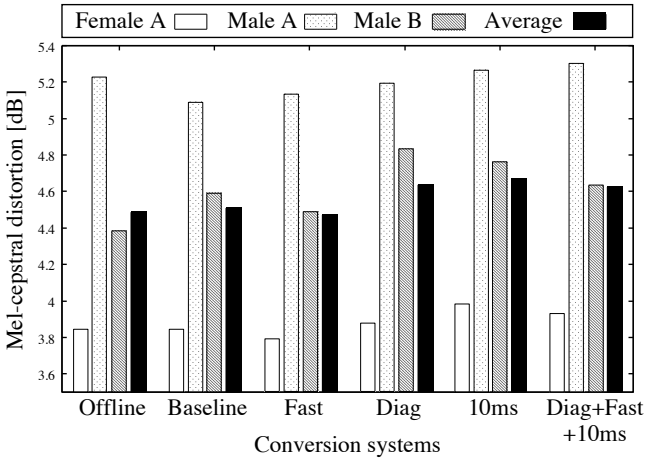

図 5 Mel-cepstral distortion in each system

フィルタや GV のポストフィルタリング処理を導入したオンライン変換システム（Baseline）は，平均的にほぼ同等の変換精度が得られることが分かる．オンライン変換システムにおいて，プログラムの高速化（Fast）が変換精度に与える影響は極めて小さい．一方で，共分散行列の対角化（Diag）と分析フレームシフト長の拡大（10ms）に関しては，若干変換精度を劣化させる傾向がある．ただし，DSP 上でリアルタイム動作するシステム（Diag+Fast+10ms）と比較すると，これら二つの要因は加算的に変換精度劣化をもたらすものではなく，個々の要因のみを導入した際とほぼ同等の変換精度が得られることが分かる．

4.4 主観評価結果

図 6 に主観評価実験の結果を示す．個々のシステムにおいて，変換音声の聞き取りやすさには大きな差がないことが分かる．結果，オフラインシステムと同等の品質を保ったまま，DSP 上でリアルタイム変換が可能であることが分かる．

5. 結　　び

本報告では，非可聴つぶやき（Non-audible murmur: NAM）からささやき声へのリアルタイム変換処理に対して，演算量削減処理を導入し，DSP 上への実装を行った．客観評価実験と主観評価実験の結果，変換精度劣化を最低限に抑えつつ，DSP 上でリアルタイム動作する変換処理を実現できることが分かった．

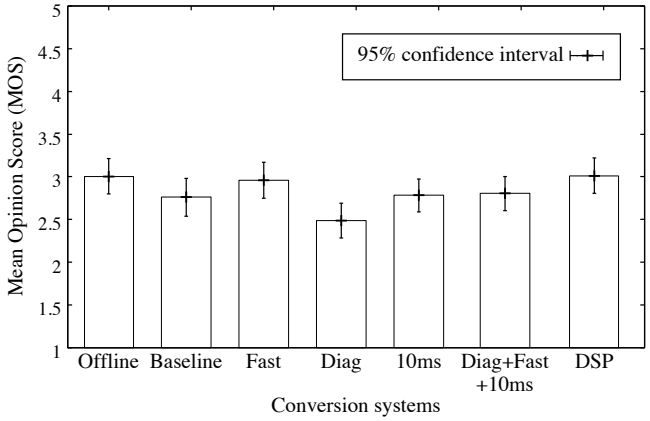

図 6 Result of opinion test in each system

リアルタイム声質変換の他の応用例として，無喉頭音声から通常音声への変換に基づく喉頭摘出者のための発声補助 [10] がある．NAM からささやき声への変換処理と比較し，通常音声への変換処理では，より多くの演算量を必要とする．今後，通常音声へのリアルタイム変換処理を DSP 上へ実装する予定である．さらに，これらのリアルタイム変換システムの実環境下への適用を目指す．

謝辞　本研究の一部は，科研費補助金若手研究（A）により実施したものである．

文　　献

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg. “Silent speech interfaces.” Speech Commun. 52, pp. 270–287, Apr. 2010.

[2] A. Subramanya, Z. Zhang, Z. Liu, and A. Acero. “Multisensory processing for speech enhancement and magnitude-normalized spectra for speech modeling.” Speech Communication, Vol. 50, No. 3, pp. 228–243, 2008.

[3] 中島淑貴, 柏岡秀紀, ニックキャンベル, 鹿野清宏, “非可聴つぶやき認識”, 信学論, Vol. J87-D-II, No.9, pp. 1757–1764, 2004.

[4] T. Toda, M. Nakagiri, K. Shikano. “Statistical voice conversion techniques for body-conducted unvoiced speech enhancement.” IEEE Transactions on Audio, Speech and Language Processing, Vol. 20, No. 9, pp. 2505–2517, Sep. 2012.

[5] T. Toda, T. Muramatsu, H. Banno. “Implementation of computationally efficient real-time voice conversion.”, Proc. INTERSPEECH, Portland, USA, Sep. 2012.

[6] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis”, Proc. ICASSP, pp.1315–1318, Jun. 2000.

[7] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, K. Shikano. “Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory.”, Proc. INTERSPEECH, pp. 1076–1079, Brisbane, Australia, Sep. 2008.

[8] T. Toda, K. Tokuda. “A Speech parameter generation algorithm considering global variance for HMM-based speech synthesis.” IEICE Transactions on Information and Systems, Vol. E90-D, No. 5, pp. 816–824, May. 2007.

[9] 徳田恵一, 小林隆夫, 深田俊明, 斎藤博徳, 今井 聖, “メルケプストラムをパラメータとする音声のスペクトル推定,”信学論, Vol. J74-A, No. 8, pp. 1240–1248, 1991.

[10] K. Nakamura, T. Toda, H. Saruwatari, K. Shikano. “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech.” Speech Communication, Vol. 54, No. 1, pp. 134–146, Jan. 2012.