無喉頭音声から通常音声への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†
†奈良先端大・情報 ††フォスター電機

1 はじめに

喉頭癌等により喉頭を手術で摘出した喉頭摘出者は，多くの場合声帯も除去するため，音源生成機能を失う．代用発声法により，無喉頭音声の発声は可能となるが，その自然性は通常音声と比較して大きく劣化する．この問題に対して，統計的手法を用いて，無喉頭音声から通常音声へと変換する手法 [1] が提案されている．本変換処理は，ラップトップ PC など十分な計算リソースが利用可能な条件下では，リアルタイム処理 [2] が可能である．本技術の実用化を進める上で，計算リソースは限られるが携行性の高い DSP 等の小型デバイス上で実装することは，有効である．

本稿では，電気式人工喉頭を用いて発声される電気音声から通常音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．実験的評価結果から，演算量削減により，変換処理の効果を保ちつつ，DSP 上でのリアルタイム動作が可能であることを示す．

2 リアルタイム無喉頭音声変換

本稿における無喉頭音声変換では，無喉頭音声のスペクトル特徴量（メルケプストラムセグメント）から通常音声のスペクトル特徴量（メルケプストラム）および音源特徴量（対数 F_0 ／無声シンボルおよび非周期成分）の推定を行う．

2.1 学習処理

時間フレーム t において，前後 C フレームから計算される入力特徴量を $\boldsymbol{X}_t$ とし，出力静的・動的特徴量を $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ とする．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．

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

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．無喉頭音声のメルケプストラムセグメントと，通常音声のメルケプストラム，対数 F_0，非周期成分との間において，計 3 つの GMM を学習する．

2.2 リアルタイム変換処理

時間フレーム 1 から T までの電気音声および通常音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．

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P\Big(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\Big) \text{ subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \quad (2)$$

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 [3]．まず，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．

$$\hat{m}_t = \underset{m}{\operatorname{argmax}}\, P\Big(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(X,Y)}\Big) \quad (3)$$

そして，式 (2) 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 [4]．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）を考慮したポストフィルタ処理を用いる [2]．

変換後の対数 F_0 に基づき有声音源（パルス列）を生成した後に，変換後の非周期成分に基づき，各周波数帯域において，有声音源と無声音源（白色雑音）を混合する重みを決定することで，励振源波形を生成する [5]．変換後のメルケプストラムに基づき励振源をフィルタリングすることで，強調音声を得る．

3 演算量削減

電気音声から通常音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削減を行う．

3.1 従来の演算量削減法 [6] の導入

非可聴つぶやきからささやき声へのリアルタイム変換処理を DSP 上に実装した際に用いた演算量削減法 [6] を，電気音声から通常音声へのリアルタイム変換処理に導入する．

共分散行列の対角化：式 (3) に示す分布選択処理において，変換精度の劣化を抑えつつ演算量を削減するために，最尤基準に基づく共分散行列の対角化を適用する [2]．m 番目の分布における入力特徴量に対する全共分散行列を，全ての混合分布に共通の変換行列と各分布に依存する対角行列を用いてモデル化する．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．

プログラムの高速化：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する．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．

分析フレームシフト長の変更：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．

3.2 混合励振源モデルの単純化

2.2 節で述べた混合励振源モデルは，高い品質の音声を合成可能であるものの，演算量は多い．そこで，Harmonic plus Noise Model (HNM)[7] で用いられる単純な混合励振源モデルを適用する．このモデルでは，最大有声周波数を境に，低域はパルス列，高域は白色雑音を用いて，励振源信号を生成する．本稿では，最大有声周波数を固定することで，さらに演算量を削減する．

---

* "Implementation of real-time alaryngeal-speech-to-speech 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

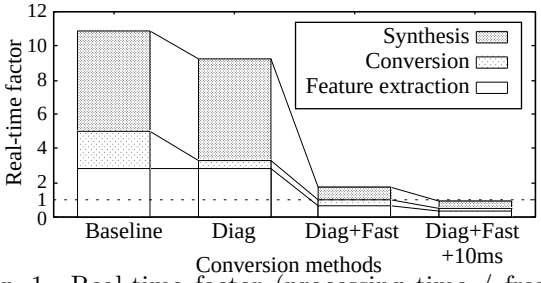

Fig. 1 Real-time factor (processing time / frame shift) of DSP.

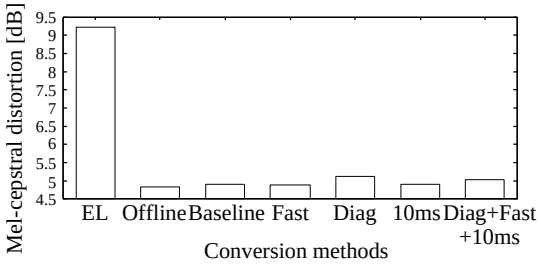

Fig. 2 Mel-cepstral distortion in each system.

4 実験的評価

4.1 実験条件

ATR 音素バランス文セット中から学習データとして 40 文，評価データとして別の 10 文の電気音声と通常音声を用いる．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．スペクトル分析は電気音声に対しては FFT 分析を用い，通常音声に対しては STRAIGHT 分析 [8] 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，メルケプストラム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）とし，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．GMM の混合数は 32 とし，特定話者モデルを用いる．

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．

- EL：変換前の電気音声
- Offline：オフライン変換処理を用いるシステム（GV を考慮した系列単位のバッチ変換処理）
- 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，共分散行列の対角化を導入したシステム
- Fast：Baseline に対して，プログラムの高速化・混合励振源の単純化を行ったシステム
- 10ms：Baseline に対して，分析フレームシフトを 10 ms としたシステム
- Diag+Fast：Diag に対して，プログラムの高速化・混合励振源の単純化を行ったシステム
- Diag+Fast+10ms：Diag+Fast に対して，分析フレームシフトを 10 ms としたシステム

主観評価実験では受聴音声の自然性に対して 5 段階（1：非常に悪い〜5：非常に良い）のオピニオンテストを行う．被験者は男性 12 名で 1 人あたり各システムにつき 10 サンプルの計 40 サンプルを受聴する．

4.2 実験結果

Fig. 1 にリアルタイムファクタ（実行時間 / 分析フレームシフト）を示す．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，プログラムの高速化・混合励振源の単純化を行うことにより特徴量抽出・合成時間が大幅に減少する．分析フレームシフトを 10 ms にすることでリアルタイムファクタが 0.93 となり，リアルタイム動作を実現できる．

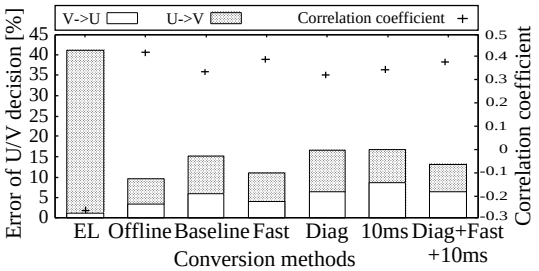

Fig. 3 Error of U/V decision and correlation coefficient in each system.

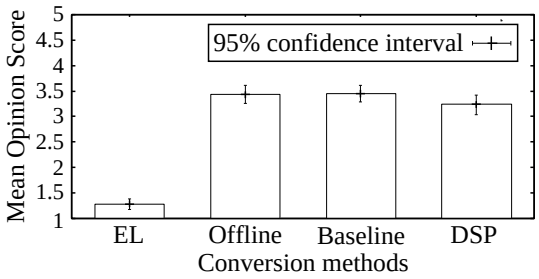

Fig. 4 Result of opinion test on naturalness.

スペクトル変換精度の評価として，Fig. 2 にメルケプストラム歪みを示す．オンライン処理（Baseline）を行うことで若干変換精度が落ち，共分散行列の対角化によりさらに劣化する．結果，リアルタイム動作可能なシステム（Diag+Fast+10ms）の変換精度は，従来のオフライン変換処理システムと比べると若干劣るものの，依然として高い変換精度が保たれていることが分かる

F_0 変換精度の評価として，Fig. 3 に相関係数と有声／無声判定エラー率を示す．オンライン処理（Baseline）を行うことで若干変換精度の劣化が見られるが，リアルタイム動作可能なシステム（Diag+Fast+10ms）においてもその劣化は最小限に抑えられていることが分かる．

Fig. 4 に主観評価実験の結果を示す．電気音声を変換することにより自然性が向上する．また，DSP 上の変換も演算量削減前とほぼ同等の自然性が得られることが分かる．

5 おわりに

本稿では，リアルタイム無喉頭音声変換に対して，演算量削減処理を導入し，DSP 上への実装を行った．実験的評価の結果，変換精度劣化を最小限に抑えつつ，DSP 上でリアルタイム動作する処理を実現できることを示した．

謝辞 本研究の一部は，JSPS 科研費 22680016 の助成を受け実施したものである．

参考文献

[1] H. Doi *et al.*, *Proc. ICASSP*, pp. 5136–5139, Prague, Czech Republic, May 2011.
[2] T. Toda *et al.*, *Proc. INTERSPEECH*, Portland, USA , Sep. 2012.
[3] K. Tokuda *et al.*, *Proc. ICASSP*, pp. 1315–1318, Istanbul, Turkey, June 2000.
[4] T. Muramatsu *et al.*, *Proc. INTERSPEECH*, pp. 1076–1079, Brisbane, Australia, Sep. 2008.
[5] Y. Ohtani *et al.*, *IEICEJ*, Vol. 91, No. 4, pp. 1082–1091, Apr. 2008.
[6] 森口 他，信学技報，SP2012-73, pp. 7–12, Nov. 2012.
[7] Y. Stylianou, *et al.*, *Proc. IEEE Trans.*, Vol 9, No. 1, pp 21–29, Jan. 2001.
[8] H. Kawahara *et al.*, *Speech Commun.*, Vol. 27, No 3–4, pp. 187–207, 1999.