ADMM を用いた NMF による雑音環境下での少量パラレルデータ声質変換*

☆李権俊, 相原龍, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

声質変換は, 入力した音声を音韻情報などは保ったまま, 話者性のような特定の情報のみを変換する技術であり, 話者変換や感情変換 [1, 2], 発話支援 [3], など様々なタスクへの応用が期待されている. これまで声質変換のための統計的手法が多く提案されているが, 中でも混合正規分布モデル (Gaussian Mixture Model: GMM) を用いた手法 [4] が広く用いられており, 多くの改良がされ続けている.

我々はこれまで, 従来の統計的手法とは異なる, スパース表現に基づく Exemplar-based な声質変換手法を提案してきた [5]. 近年スパース表現に基づく手法は信号処理の分野において注目されており, 音声信号処理の分野でも音声認識や音源分離, 雑音抑圧などにおいて, その有効性が報告されている [6, 7]. スパース表現の考え方においては, 与えられた信号は少量の学習サンプルや基底の線形結合で表現される. 音源分離に用いる場合, まず学習サンプルや基底を音源ごとにグループ (辞書) 化し, 混合音声をそれらのスパース表現にする. その後目的音声の辞書に対する重みベクトルのみを取り出して用いることで, 目的音声のみを分離する. Gemmeke ら [8] は雑音の重畳した音声を, クリーン音声辞書とノイズ辞書のスパース表現にし, クリーン音声辞書に対する重みを音声認識における Hidden Markov Model (HMM) の尤度として用いることで, 雑音にロバストな音声認識を行う手法を提案している.

我々の提案している声質変換手法では, スパースコーディングの代表的な手法として Non-negative Matrix Factorization (NMF) [9] を用いてきた. この手法では, 入力話者の音声辞書 (入力話者辞書) と出力話者の音声辞書 (出力話者辞書) からなる同一発話内容のパラレル辞書を構築する. 変換時には, 入力音声を NMF によって, 入力辞書に含まれる少量の基底からなるスパース表現にする. 得られた入力辞書の基底毎の重み係数 (アクティビティ) に基づいて, 入力話者辞書の基底を出力辞書内の基底と置き換え, 線形結合することで, 出力話者の音声へと変換する.

本研究では Alternating Direction Method of Multipliers (ADMM) を用いた Semi-NMF による声質変換手法を提案する. 我々が提案してきた従来の NMF による声質変換手法では, アクティビティを求める際の 計算コストが高く, 収束に時間がかかるという問題点があった. そこで本研究では辞書の次元数を削減するために特徴量としてメルケプストラムを用いる. メルケプストラムを用いるに当たって, NMF の非負制約を一部無くした Semi-NMF を用いる. さらに Semi-NMF を解く手法として従来の補助関数法ではなく, ADMM を用いる. ADMM は最適化手法の一種で, 複雑な最適化問題をいくつかの簡単な最適化問題に分割して解く手法である. この手法は近年様々な分野において注目されており, Dennis [10] らは ADMM を用いて NMF を解く手法を提案している. ADMM を用いることでアクティビティ推定における収束の早期化, よりスパースなアクティビティの推定が可能になる. 評価実験では, クリーンな音声及び雑音重畳音声を用いて提案手法の有効性を示す.

2 NMF を用いた声質変換

スパース表現の考え方において, 与えられた信号は少量の学習サンプルや基底の線形結合で表現される.

$$\mathbf{v}_l \approx \sum_{j=1}^{J} \mathbf{w}_j h_{j,l} = \mathbf{W}\mathbf{h}_l \tag{1}$$

$\mathbf{v}_l$ は観測信号の l 番目のフレームにおける D 次元の特徴量ベクトルを表す. $\mathbf{w}_j$ は j 番目の学習サンプル, あるいは基底を表し, $h_{j,l}$ はその結合重みを表す. 本手法では学習サンプルそのものを基底 $\mathbf{w}_j$ とする. 基底を並べた行列 $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_J]$ は “辞書” と呼び, 重みを並べたベクトル $\mathbf{h}_l = [h_{1,l} \dots h_{J,l}]^T$ は “アクティビティ” と呼ぶ. このアクティビティベクトル $\mathbf{h}_l$ がスパースであるとき, 観測信号は重みが非ゼロである少量の基底ベクトルのみで表現されることになる. フレーム毎の特徴量ベクトルを並べて表現すると式 (1) は二つの行列の内積で表される.

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \tag{2}$$

$$\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_L], \quad \mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_L]. \tag{3}$$

ここで L はフレーム数を表し, 本手法において, $\mathbf{W}$ は学習データで固定される.

本手法の概要を Fig. 1 に示す. $\mathbf{V}^s$ は入力話者スペクトル, $\mathbf{W}^s$ は入力話者辞書, $\mathbf{W}^t$ は出力話者辞書, $\hat{\mathbf{V}}^t$ は変換されたスペクトル, $\mathbf{H}^s$ は入力話者スペクトルから推定されるアクティビティを表す. D, J はそれぞれスペクトルの次元数, 辞書の基底数である.

*Voice Conversion using a Small Parallel Corpus based on NMF using ADMM in Noisy Environments, by Konjun I, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki (Kobe univ.)

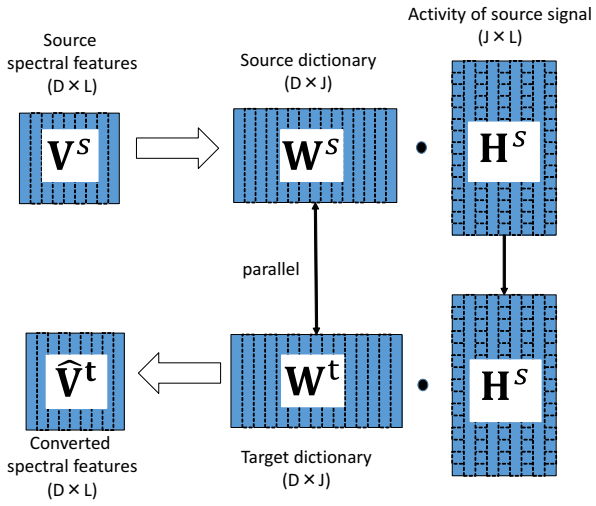

Fig. 1　Basic approach of exemplar-based voice conversion in a noisy environment

この手法では，パラレル辞書と呼ばれる入力話者辞書 $\mathbf{W}^s$ と出力話者辞書 $\mathbf{W}^t$ からなる辞書の対を用いる．この辞書の対は従来の声質変換法と同様，入力話者と出力話者による同一発話内容のパラレルデータに Dynamic Time Warping (DTW) を適用することでフレーム間の対応を取った後，入力話者と出力話者の学習サンプルをそれぞれ並べたものである．

入力スペクトル $\mathbf{V}^s$ は NMF によって $\mathbf{W}^s$ と $\mathbf{H}^s$ の積に分解される．本手法では，「パラレル辞書で推定したパラレルな発話のアクティビティは置き換え可能である」と仮定している．従って，変換スペクトル $\hat{\mathbf{V}}^t$ は，$\mathbf{W}^t$ と推定した $\mathbf{H}^s$ の積によって得られる．

3　ADMM を用いた Semi-NMF による最適化

本手法では メルクプストラムを特徴量として用い，アクティビティ行列の推定に Semi-NMF を用いる．Semi-NMF のアクティビティを求める最適化問題は，$\mathbf{V}$，$\mathbf{W}$，$\mathbf{H}$ を用いて以下のような式で表せる．

$$\begin{aligned} min \quad & d(\mathbf{V},\mathbf{WH}) + \lambda||\mathbf{H}||_1 \\ sub.to \quad & \mathbf{H} \geq 0 \end{aligned} \tag{4}$$

ここで，第 1 項は $\mathbf{V}$ と $\mathbf{WH}$ の間の Euclidean distance であり，第 2 項はアクティビティ行列をスパースにするための L1 ノルム制約項である．λ はスパース重みを表す．[10] にしたがって式 (4) に $\mathbf{H}_+ \geq 0$ を加えると式は以下のようになる．

$$\begin{aligned} min \quad & d(\mathbf{V},\mathbf{WH}) + \lambda||\mathbf{H}||_1 \\ sub.to \quad & \mathbf{H} = \mathbf{H}_+ \qquad \mathbf{H}_+ \geq 0 \end{aligned} \tag{5}$$

この条件式の拡張ラグランジュ関数は以下のように定義できる．

$$\begin{aligned} d(\mathbf{V},\mathbf{WH}) + \lambda||\mathbf{H}||_1 + \langle \alpha_H, \mathbf{H} - \mathbf{H}_+ \rangle \\ + \frac{\rho}{2}||\mathbf{H} - \mathbf{H}_+||_2^F + \lambda||\mathbf{H}||_1 \end{aligned} \tag{6}$$

コスト関数を最小化するアクティビティは, (6) 式をそれぞれの変数で最適化する事により以下の更新式で求められる．

$$\begin{aligned} \mathbf{H} \leftarrow & (2\mathbf{W}^T + \rho)^{-1} \\ & (2\mathbf{W}^T\mathbf{V} - \alpha_H + \rho\mathbf{H}_+ - \lambda) \end{aligned} \tag{7}$$

$$\mathbf{H}_+ \leftarrow max(\mathbf{H} + \alpha_{\mathbf{H}}/\rho, 0) \tag{8}$$

$$\alpha_{\mathbf{H}} \leftarrow \alpha_{\mathbf{H}} + \rho(\mathbf{H} - \mathbf{H}_+) \tag{9}$$

$$\mathbf{W} = \mathbf{VW}(\mathbf{WW}^T)^{-1} \tag{10}$$

4　話者適応を用いた声質変換

本稿では少量のパラレルデータを用いて声質変換を行うために, 出力話者の辞書を入力話者の辞書から作成する話者適応を行う．Fig. 2 に話者辞書の適応を用いたパラレル辞書作成の概要を示す．適応データとして, 入力話者と出力話者の同一内容の発話 $\mathbf{V}^s$, $\mathbf{V}^t$ を用意する．入力話者の適応データのメルケプストラム $\mathbf{V}^s$ と入力話者辞書 $\mathbf{W}^s$ を用いて以下の式を最小化するアクティビティ $\mathbf{H}^s$ を推定する．

$$d(\mathbf{V},\mathbf{WH}) + \lambda||\mathbf{H}^s||_1 \tag{11}$$

ここで, 入力話者辞書は入力話者の音声から抽出したメルケプストラムを並べたものである．適応データである出力話者の メルケプストラム $\mathbf{V}^t$ は, $\mathbf{V}^s$ のパラレルデータなので , $\mathbf{V}^t$ は 出力話者辞書 $\mathbf{W}^t$ と 推定された $\mathbf{H}^s$ を用いて以下のように表せる．

$$\mathbf{V}^t = \mathbf{W}^t\mathbf{H}^s \tag{12}$$

ここで $\mathbf{W}^t$ が適応行列 $\mathbf{A}$ と $\mathbf{W}^s$ の積によって表現できると考えると

$$\mathbf{V}^t \simeq \mathbf{AW}^s\mathbf{H}^s \tag{13}$$

となる．適応行列 $\mathbf{A}$ は 式 (10) の $\mathbf{V}$ に $\mathbf{V}^t$ を, $\mathbf{W}$ に $\mathbf{A}$ を, $\mathbf{H}$ に $\mathbf{W}^s\mathbf{H}^s$ を代入することで得られる．ここで得られた $\mathbf{A}$ と $\mathbf{W}^s$ の積でパラレルな出力話者辞書 $\hat{\mathbf{W}}^t$ が得られる．

$$\hat{\mathbf{W}}^t = \mathbf{AW}^s \tag{14}$$

変換する音声のアクティビティを推定し, $\hat{\mathbf{W}}^t$ との積をとることによって変換後の音声が得られる．

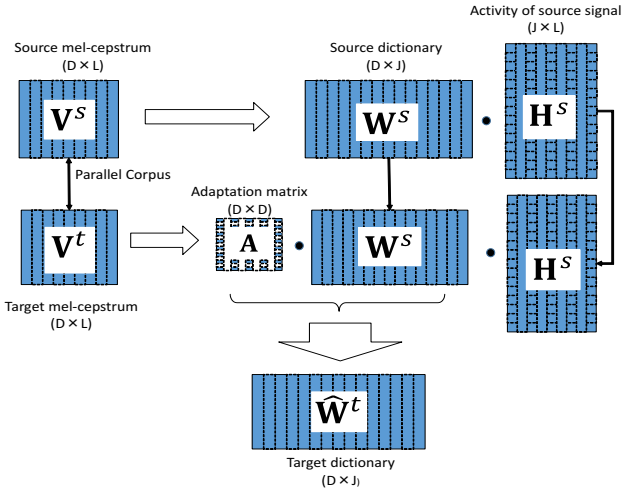

Fig. 2 Estimation of parallel dictionary using a speaker transformation matrix

5 雑音重畳音声の声質変換

雑音重畳音声の声質変換では入力話者辞書を雑音に適応させるため, 少量の入力話者の雑音重畳音声を用いて, 辞書適応を行う. 適応データとして 4 節の $\mathbf{V}^s$ に雑音を重畳した音声のメルケプストラム $\mathbf{V}^{sn}$ を用意し, 話者適応と同様に $\mathbf{V}^{sn}$, 入力話者辞書 $\mathbf{W}^s$ と, 4 節で求めたアクティビティ $\mathbf{H}^s$ を用いて以下の式を最小化する $\mathbf{A}^n$ を推定する.

$$d(\mathbf{V}^{sn}, \mathbf{A}^n\mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \tag{15}$$

ここで得られた適応行列 $\mathbf{A}^n$ と入力話者辞書 $\mathbf{W}^s$ の積によって雑音環境下に適応させた入力話者辞書 $\hat{\mathbf{W}}^s$ が生成される.

$$\hat{\mathbf{W}}^s = \mathbf{A}^n\mathbf{W}^s \tag{16}$$

この $\hat{\mathbf{W}}^s$ を用いて入力話者の雑音重畳音声 $\mathbf{V}^s$ のアクティビティを求める.

$$\mathbf{V}^s = \hat{\mathbf{W}}^s\mathbf{H}^s \tag{17}$$

6 評価実験

6.1 実験条件

ATR 研究用日本語音声データベースセット [11] を用いて話者変換を行い，提案手法を従来の NMF 声質変換, 辞書適応と雑音辞書の構築を行った NMF 声質変換 (Ada-NMF) [12], 及び GMM 声質変換と比較を行った. テストデータには雑音のないクリーンな音声と雑音重畳音声を用いた. 入力話者は男性, 出力話者は女性, サンプリング周波数は 8kHz とした. 従来の NMF 変換で用いるパラレル辞書の構築に 50 単語, 提案手法及び Ada-NMF の入力話者辞書の構築に 50 単語, 適応データとしてパラレルな 10 単語を使用した.

比較手法である GMM に基づく声質変換のための学習サンプルには, 辞書を構築したのと同様音声のケプストラムをフレーム間同期を取る事で作った 50 単語のパラレルデータを用いた. ケプストラムは STRAIGHT スペクトルから計算される線形ケプストトラムで, 次元数は 40 である. GMM の混合数は 30 とした.

テストデータには比較・提案手法ともにパラレル辞書内に含まれない 50 単語, 及びそれに雑音信号を重畳したものを用いた. 雑音信号は CENSREC-1-C データベースにて駅で収録された音声の無音声部分の雑音を用いた. 雑音信号の平均 SNR は 10dB とした. 提案手法のテスト時の入力音声及び入力話者辞書の構築には 256 次元の振幅スペクトルの メルケプストラム 24 次元を, 出力音声の生成及び出力話者辞書の構築には 513 次元の STRAIGHT スペクトルのメルケプストラム 24 次元を用いた. これは, 音声信号の分析合成ツールである STRAIGHT [13] では雑音重畳音声の雑音を上手く表現できないという問題があるためである.

提案手法の有効性を確かめるため，客観評価実験を行った. 客観評価はメルケプストラム 24 次元を用いて, 式 (18) で表されるメルケプストラム歪 (Melcepstrum distortion : MCD) [dB] によって各手法を比較した.

$$MCD = (10/\log 10)\sqrt{2\sum_{d}^{24}(mc_d^{conv} - mc_d^{tar})^2} \tag{18}$$

ここで，mc_d^{conv}，mc_d^{tar} は d 次元目の変換後のケプストラム，目標音声のケプストラムを表す.

6.2 実験結果・考察

Fig. 3 にクリーンな音声の変換音声の MCD による比較, Fig. 4 に雑音重畳音声の変換音声の MCD による比較を, またその変換にかかった 1 単語あたりの平均計算時間を Table 1 に示す.

図より, 提案手法はクリーンな音声及び, 雑音重畳音声の変換において, 比較手法と比べて同程度の精度で変換ができていることが確認できる. さらに表から比較手法と比べて計算時間が大きく削減されていることが確認できる.

Table 1 Computation time [s]

SemiNMF	NMF	Ada-NMF
2.26	35.124	36.90

7 おわりに

本稿では 雑音重畳音声に対して ADMM を用いた Semi-NMF による少量パラレルデータのみを用いた

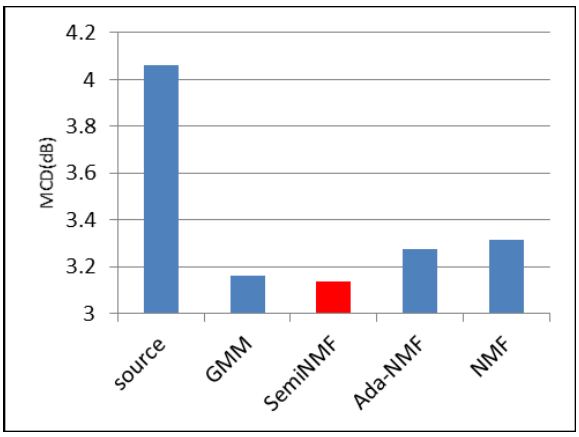

Fig. 3　MCD for converted voice in a clean environment

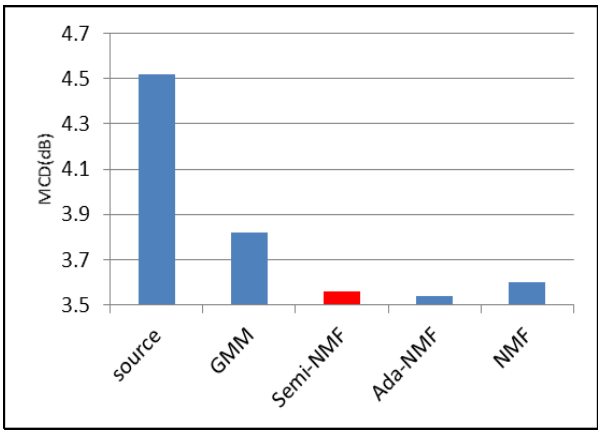

Fig. 4　MCD for converted voice in a noisy environment

声質変換を行う手法を提案した. 実験結果より, 提案手法は従来の NMF 声質変換 と比べて同程度の変換精度を持つとともに, 計算時間においては大きく削減できていることを示した.

今後は, Semi-NMF を用いた変換手法におけるアクティビティの推定において, 帯域ごとに分割して推定する手法の検討を進めていく.

参考文献

[1] Y. Iwami *et al.*, "GMM-based voice conversion applied to emotional speech synthesis," IEEE Trans. Speech and Audio Proc., vol. 7, pp. 2401–2404, 1999.

[2] C. Veaux and X. Robet, "Intonation conversion from neutral to expressive speech," in Proc. Interspeech, pp. 2765–2768, 2011.

[3] K. Nakamura *et al.*, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," Speech Communication, vol. 54, no. 1, pp. 134–146, 2012.

[4] Y. Stylianou *et al.*, "Continuous probabilistic transform for voice conversion," IEEE Trans. Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.

[5] R. Takashima *et al.*, "Exemplar-based voice conversion in noisy environment," in Proc. SLT, pp. 313–317, 2012.

[6] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 3, pp. 1066–1074, 2007.

[7] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in Proc. Interspeech, 2006.

[8] J. Gemmeke *et al.*, "Exemplar-based sparse representations for noise robust automatic speech recognition," IEEE Trans. Audio, Speech, Lang. Process., pp. 2067–2080, 2011.

[9] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," Neural Information Processing System, pp. 556–562, 2001.

[10] D. L.Sun and C. Fevotte, "Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence," in Proc. ICASSP, pp. 6201–6205, 2014.

[11] A. Kurematsu *et al.*, "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, vol. 9, pp. 357–363, 1990.

[12] R. Aihara *et al.*, "Noise-robust voice conversion using a small parallel data based on non-negative matrix factorization," Eurasip journal., 2015.

[13] H. Kawahara *et al.*, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequencybased F0 extraction: possible role of a repetitive structure in sounds," Speech Communication, vol. 27, no. 3-4, pp. 187–207, 1999.