非負値行列因子分解に基づく声質変換のための Graph Embeddingを用いたパラレル辞書学習*

◎相原龍, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

我々はこれまで，従来一般的であった統計的手法による声質変換 [1] とは異なる，スパース表現に基づく非負値行列因子分解 (Non-negative Matrix Factorization : NMF) [2] を用いた Exemplar-based 声質変換手法を提案してきた [3]．声質変換とは，入力された音声に含まれる話者性・音韻性・感情性などといった多くの情報の中から，特定の情報を維持しつつ他の情報を変換する技術である．音韻情報を維持しつつ話者情報を変換する “話者変換” [4] を目的として混合正規分布モデル (Gaussian Mixture Model : GMM) を用いた手法を中心に広く研究されてきたが，NMF 声質変換は従来の声質変換のように統計的モデルを用いないため過学習がおこりにくいことに加え，高次元スペクトルを用いて変換するため，自然性の高い音声へと変換可能であると考えられる．さらに，NMF 声質変換は，NMF によるノイズ除去手法と組み合わせることでノイズロバスト性を有する．

NMF は，スパース行列分解手法のひとつであり入力信号 $\mathbf{V}$ は，辞書行列を $\mathbf{W}$，係数行列（アクティビティ）を $\mathbf{H}$ とすると，以下のような式で表される．

$$\mathbf{V} \approx \mathbf{WH}. \tag{1}$$

音声信号処理においても，NMF はシングルチャンネル音声分離 [5, 6]，歌声分離 [7] などに応用されてきた．

NMF は $\mathbf{W}$ と $\mathbf{H}$ を同時に推定する辞書推定による手法と，$\mathbf{W}$ を Exemplar で固定し $\mathbf{H}$ のみを推定する Exemplar-based の手法に分けることができる．辞書推定による手法は，コンパクトな辞書を推定することができるため計算コストを削減できるが，アクティビティのみならず辞書基底もスパースになる傾向があるため，音声のフォルマント構造が壊れてしまい高い精度が得られないという問題点があった．Exemplar-based 手法ではそのような現象を防ぐことができるが，辞書推定による手法と比較して，計算コストが高く，分解精度誤差も大きくなるという問題点がある．

NMF 声質変換はこれまで Exemplar-based によるものがほとんどであった．本稿では，NMF を声質変換における分解誤差削減による精度向上を目指し，識別的な Graph-embedded NMF (Discriminative Graph-embedded Non-negative Matrix Factorization: DGNMF) を提案し，この手法を用いたパラレル辞書学習を行う．類似した手法として Graph Regularized NMF (GRNMF) [8] が提案されているが，GRNMF は音素ラベルに基づく識別がないうえ，クラス間分散は考慮しておらず，真に識別的な学習が行われているとはいえない．DGNMF は音素ラベルを考慮し，クラス内クラス間分散制約に基づいて辞書を推定する識別的 NMF だといえる．

以下，第 2 章で先行研究について説明し，問題点を述べる．第 3 章で提案手法を説明する．第 4 章で評価実験を行い，第 5 章で本稿をまとめる．

2 NMF 声質変換

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．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．NMF のコスト関数を下に示す．

$$d_{KL}(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \quad s.t. \quad \mathbf{H}^s \geq 0 \tag{2}$$

式 (2) において，第 1 項は $\mathbf{V}^s$ と $\mathbf{W}^s\mathbf{H}^s$ の間の Kullback-Leibler（KL）ダイバージェンスであり，第 2 項はアクティビティをスパースにするための L1 ノルム正則化項である．λ はスパース重みを表す．辞書は固定し，アクティビティのみを推定する．コスト関数は，次のような更新式を用いて最小化される

$$\begin{aligned}\mathbf{H}^s \quad \leftarrow \quad & \mathbf{H}^s .* (\mathbf{W}^{s\mathsf{T}}(\mathbf{V}^s ./ (\mathbf{W}^s\mathbf{H}^s))) \\ & ./ (\mathbf{W}^{s\mathsf{T}}\mathbf{1}^{(I\times J)} + \lambda\mathbf{1}^{(K\times J)})\end{aligned} \tag{3}$$

ここで，$.*$，$./$，$\mathbf{1}$ は要素積，要素商，全要素が 1 の配列を表す．

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．

$$\hat{\mathbf{V}}^t = \mathbf{W}^t\mathbf{H}^s \tag{4}$$

*Parallel Dictionary Learning for Voice Conversion Using Discriminative Graph-embedded Non-negative Matrix Factorization by Ryo AIHARA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

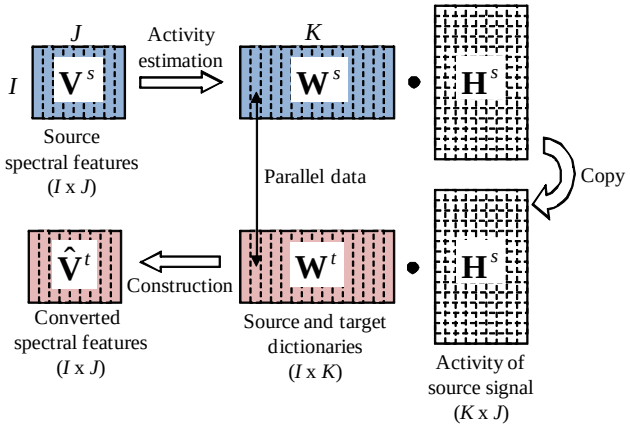

Fig. 1　Basic approach of NMF-based voice conversion

3　DGNMF を用いたパラレル辞書学習

3.1　Graph Embedding を考慮した NMF

学習データにおいて，クラス内分散グラフとクラス間分散グラフの隣接行列をそれぞれ次のように定義する．

$$\mathbf{A}^w_{ij} = \begin{cases} 1 & \left(\begin{array}{c} \mathbf{v}_i \in N_{k_w}(\mathbf{v}_i) \ or \ \mathbf{v}_j \in N_{k_w}(\mathbf{v}_j) \\ and \\ c_i = c_j \end{array} \right) \\ 0 & (otherwise) \end{cases} \quad (5)$$

$$\mathbf{A}^b_{ij} = \begin{cases} 1 & \left(\begin{array}{c} \mathbf{v}_i \in N_{k_b}(\mathbf{v}_i) \ or \ \mathbf{v}_j \in N_{k_b}(\mathbf{v}_j) \\ and \\ c_i \neq c_j \end{array} \right) \\ 0 & (otherwise) \end{cases} \quad (6)$$

ここで，$N_{k_w}(\mathbf{v}_i)$，$N_{k_b}(\mathbf{v}_i)$ はそれぞれ，クラス内分散グラフにおける学習データ $\mathbf{v}_i$ の k_w 近傍集合，クラス間分散グラフにおける $\mathbf{v}_i$ の k_b 近傍集合を表す．c_i と c_j はそれぞれ $\mathbf{v}_i$ の $\mathbf{v}_j$ 音素ラベルである．これらの識別行列を用いて，クラス内分散とクラス間分散のグラフラプラシアン行列を求める．

$$\mathbf{L}^w = \mathbf{D}^w - \mathbf{A}^w \quad (7)$$

$$\mathbf{L}^b = \mathbf{D}^b - \mathbf{A}^b \quad (8)$$

ここで，$\mathbf{D}^w$ と $\mathbf{D}^b$ はそれぞれ，対角成分に $\mathbf{A}^w$ と $\mathbf{A}^b$ の各行（あるいは列）の和をもつ重み行列である．

グラフラプラシアンを用いて，DGNMF のコスト関数を次のように定義する．

$$d_{KL}(\mathbf{V}, \mathbf{WH}) + \frac{\phi}{2}\mathbf{Tr}(\mathbf{H}^\mathsf{T}\mathbf{L}^w\mathbf{H}) - \frac{\psi}{2}\mathbf{Tr}(\mathbf{H}^\mathsf{T}\mathbf{L}^b\mathbf{H}) \quad s.t. \ \ \mathbf{W} \geq 0, \mathbf{H} \geq 0 \quad (9)$$

ここで，$\mathbf{Tr}$，ϕ，ψ はそれぞれ行列のトレース，クラス内分散重み，クラス間分散重みを表す．式 (9) において，第 1 項は $\mathbf{V}$ と $\mathbf{WH}$ の間の KL ダイバージェンス，第 2 項はクラス内分散制約，第 3 項はクラス間分散制約を表す．

このコスト関数は次の更新式を用いて最小化できる．

$$\mathbf{W} \leftarrow \mathbf{W}. * ((\mathbf{V}./(\mathbf{WH}))\mathbf{H}^\mathsf{T})./(\mathbf{1}^{(J\times I)}\mathbf{H}^\mathsf{T}) \quad (10)$$

$$\mathbf{H} \leftarrow \frac{-\beta + \sqrt{\beta^2 + 4\alpha\gamma}}{2\alpha} \quad (11)$$

$$\alpha = ((\phi\mathbf{D}^w + \psi\mathbf{A}^b)\mathbf{H})./\mathbf{H} \quad (12)$$

$$\beta = \mathbf{W}^\mathsf{T}\mathbf{1}^{(I\times J)} - \mathbf{H}(\phi\mathbf{A}^w + \psi\mathbf{D}^b) \quad (13)$$

$$\gamma = \mathbf{H}. * (\mathbf{W}^\mathsf{T}(\mathbf{V}./(\mathbf{WH}))) \quad (14)$$

これらの更新式は GNMF と同様にして導出できる [8]．しかしながら，GNMF には音素ラベルに基づく識別ならびにクラス間分散は考慮しておらず，DGNMF と GNMF は異なるアプローチである．

3.2　パラレル辞書学習

パラレル制約付き DGNMF を用いて，コンパクトで識別的な辞書を推定する．Fig. 2 に，パラレル辞書学習の概要を示す．パラレル制約付き DGNMF の目的関数を次のように定義する．

$$\begin{aligned} & d_{KL}(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + d_{KL}(\mathbf{V}^t, \mathbf{W}^t\mathbf{H}^t) \\ & + \frac{\phi}{2}\mathbf{Tr}(\mathbf{H}^{s\mathsf{T}}\mathbf{L}^{sw}\mathbf{H}^s) + \frac{\phi}{2}\mathbf{Tr}(\mathbf{H}^{t\mathsf{T}}\mathbf{L}^{tw}\mathbf{H}^t) \\ & - \frac{\psi}{2}\mathbf{Tr}(\mathbf{H}^{s\mathsf{T}}\mathbf{L}^{sb}\mathbf{H}^s) - \frac{\psi}{2}\mathbf{Tr}(\mathbf{H}^{t\mathsf{T}}\mathbf{L}^{tb}\mathbf{H}^t) \\ & + \lambda||\mathbf{H}^s||_1 + \lambda||\mathbf{H}^t||_1 + \frac{\epsilon}{2}||\mathbf{H}^s - \mathbf{H}^t||_F^2 \\ & s.t. \ \ \mathbf{W}^s \geq 0, \mathbf{H}^s \geq 0, \mathbf{W}^t \geq 0, \mathbf{H}^t \geq 0 \end{aligned} \quad (15)$$

ここで，$\mathbf{V}^s$，$\mathbf{V}^t$，$\mathbf{W}^s$，$\mathbf{W}^t$，$\mathbf{H}^s$，$\mathbf{H}^t$ は入力話者学習データ，出力話者学習データ，入力話者辞書行列，出力話者辞書行列，入力アクティビティ，出力アクティビティを表す．さらに $\mathbf{L}^{sw}$，$\mathbf{L}^{sb}$ は入力話者学習データそれぞれクラス内分散，クラス間分散のグラフラプラシアンを表し，$\mathbf{L}^{tw}$，$\mathbf{L}^{tb}$ は出力話者学習データのグラフラプラシアンを表す．入力話者学習データ，出力話者学習データはそれぞれ DTW で対応を取り，同一フレーム数となったものを用いる．ϵ，λ はそれぞれパラレル制約，スパース制約の重みである．式 (15) において，第 1 項から第 6 項は式 (9) を拡張したものである．第 7 項と第 8 項は $\mathbf{H}^s$ と $\mathbf{H}^t$ におけるスパース制約，最終項は $\mathbf{H}^s$ と $\mathbf{H}^t$ の間のパラレル制約を表す．

このコスト関数は次の更新式を用いて最小化できる．

$$\mathbf{W}^s \leftarrow \mathbf{W}^s. * ((\mathbf{V}^s./(\mathbf{W}^s\mathbf{H}^s))\mathbf{H}^{s\mathsf{T}})./(\mathbf{1}^{(J\times I)}\mathbf{H}^{s\mathsf{T}}) \quad (16)$$

$$\mathbf{W}^t \leftarrow \mathbf{W}^t. * ((\mathbf{V}^t./(\mathbf{W}^t\mathbf{H}^t))\mathbf{H}^{t\mathsf{T}})./(\mathbf{1}^{(J\times I)}\mathbf{H}^{t\mathsf{T}}) \quad (17)$$

$$\mathbf{H}^s \leftarrow \frac{-\beta^s + \sqrt{\beta^{s2} + 4\alpha^s\gamma^s}}{2\alpha^s} \quad (18)$$

$$\alpha^s = ((\phi \mathbf{D}^{sw} + \psi \mathbf{A}^{sb})\mathbf{H}^s)./\mathbf{H}^s + \epsilon \quad (19)$$

$$\beta^s = \mathbf{W}^{s\mathsf{T}}\mathbf{1}^{(I\times J)} - \mathbf{H}^s(\phi \mathbf{A}^{sw} + \psi \mathbf{D}^{sb}) - \epsilon \mathbf{H}^t + \lambda \quad (20)$$

$$\gamma^s = \mathbf{H}^s .* (\mathbf{W}^{s\mathsf{T}}(\mathbf{V}^s./(\mathbf{W}^s\mathbf{H}^s))) \quad (21)$$

$$\mathbf{H}^t \leftarrow \frac{-\beta^t + \sqrt{\beta^{t2} + 4\alpha^t\gamma^t}}{2\alpha^t} \quad (22)$$

$$\alpha^t = ((\phi \mathbf{D}^{tw} + \psi \mathbf{A}^{tb})\mathbf{H}^t)./\mathbf{H}^t + \epsilon \quad (23)$$

$$\beta^t = \mathbf{W}^{t\mathsf{T}}\mathbf{1}^{(I\times J)} - \mathbf{H}^t(\phi \mathbf{A}^{tw} + \psi \mathbf{D}^{tb}) - \epsilon \mathbf{H}^s + \lambda \quad (24)$$

$$\gamma^t = \mathbf{H}^t .* (\mathbf{W}^{t\mathsf{T}}(\mathbf{V}^t./(\mathbf{W}^t\mathbf{H}^t))) \quad (25)$$

$\mathbf{A}^{sw}$ と $\mathbf{A}^{sb}$ は入力話者学習データにおけるクラス内分散グラフとクラス間分散グラフの隣接行列，$\mathbf{D}^{sw}$ と $\mathbf{D}^{sb}$ はそれらの重み行列である．$\mathbf{A}^{tw}$，$\mathbf{A}^{tb}$，$\mathbf{D}^{tw}$，$\mathbf{D}^{tb}$ はそれらのそれぞれ出力話者学習データである．

識別的なパラレル辞書が推定された後，入力スペクトルは第 2 章で示した手法と同様にして変換される．

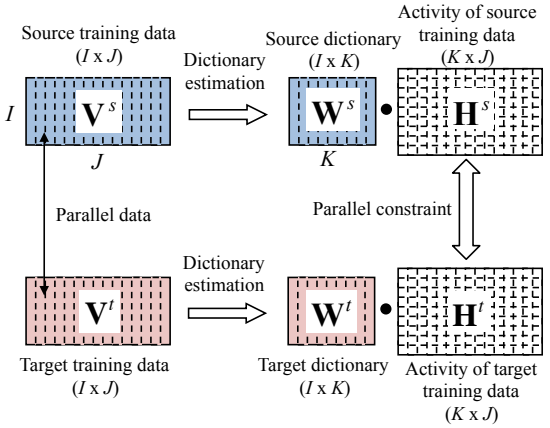

Fig. 2　Parallel dictionary learning

4　評価実験

4.1　実験条件

提案手法は，クリーン環境下での話者変換をタスクとし，従来の NMF 声質変換 [3]，GMM 学習声質変換と比較した．

ATR 研究用日本語音声データベースに含まれる男性 1 名を入力話者，女性 1 名を出力話者とした．サンプリング周波数は 12kHz である．音素バランス文 50 文を学習データとし，学習データに含まない音素バランス文 50 文をテストデータとして用いた．$\epsilon, \phi, \psi, k_w, k_b, \lambda$ はそれぞれ 40, 1, 10, 1024, 8192, 0.1 とした．NMF の更新回数は辞書学習時には 10，変換時には 300 とした．これらのパラメータは実験的に求められたものである．

提案手法と NMF 声質変換では，音声分析合成手法 STRAIGHT [**?**] によって推定されたスペクトルと前後 1 フレームを含むセグメント特徴量を用いた．この次元数は 1,539 である．GMM 声質変換において，STRAIGHT を用いて推定されたスペクトルから計算した Mel-cepstrum と前後 1 フレームを考慮した Δ パラメータを特徴量として用いた．特徴量の次元数は 48 である．GMM の混合数は 64 とした．

本稿では，F0 には平均，分散を考慮した線形変換を適用し [1]，非周期成分は入力発話のものを用いた．

4.2　実験結果

客観評価指標として，メルケプストラム歪（Mel-cepstral distortion: MelCD）[dB] を用いた．

$$MelCD = (10/\log 10)\sqrt{2\sum_{d}^{24}(mc_d^{conv} - mc_d^{tar})^2} \quad (26)$$

ここで，mc_d^{conv}，mc_d^{tar} は変換メルケプストラム，目標メルケプストラムにおける d 次元目の特徴量を表す．

Table 1 にそれぞれの手法による MelCD と計算コストを表す．Conv. は目標メルケプストラムと変換メルケプストラムの間の歪，Recon. は入力メルケプストラムと分解再合成時のメルケプストラム（式 (2) における $\mathbf{W}^s\mathbf{H}^s$）を表し，これは行列分解における分解誤差にあたる．括弧内の数字はパラレル辞書の基底数である．PDL は，Graph Embedding を考慮していないパラレル辞書学習（DGNMF における $\phi = 0, \psi = 0$）を表す．

まず，従来の NMF 声質変換 (Exemplar-based) において，基底数を削減していった場合を見ると，5,000 基底の場合と全基底を用いた場合の歪がほぼ一致するという結果になった．分解誤差は全基底を用いたときが最も小さく，次いで 5,000 基底を用いたときが小さい．

PDL でパラレル辞書を学習した場合は，NMF でランダムに基底を削減した場合と比較して歪が若干大きくなった．分解誤差も大きくなっていることがわかる．DGNMF でパラレル辞書を学習した場合，変換歪は従来の NMF とほぼ同等であり，分解誤差はこれらの手法のなかで最も小さくなった．この結果は，Graph Embedding を考慮することの効果を示している．

また，計算コストは辞書の基底数を削減するに従って小さくなっている．

主観評価実験は，10 人の日本語話者が 25 文のテストデータについてそれぞれの手法で変換した音声を評価した．本論文では，音質と話者性の 2 つの観点において評価実験を行った．音質に関しては 5 段階評価 (5: excellent, 4: good, 3: fair, 2: poor, 1: bad) で，話者性については，X を目標話者とし 2 つの変換手法のうちよい方を選ぶ XAB 法で評価した．従来の NMF 声質変換について基底数は 5,000 とした．

Fig. 3 に音質の評価結果を示す．エラーバーは，95% 信頼区間を示す．提案手法は，従来の 2 つの手法と比較して音質が向上していることがわかる．この結果は t 検定により有意であることが示されている．これは，提案手法が NMF の声質変換の分解誤差を削減しつつパラレル辞書のマッピング精度を向上させることができたことによるものだと考えられる．

Fig. 4 に話者性による主観評価実験の結果を示す．提案手法と GMM 声質変換の間には有意差が見られないが，提案手法は NMF 声質変換と比較して有意に話者性を向上させていることがわかる．この結果は t 検定により有意であることがが示されている．

Table 1　MelCD [dB] and computational times [s] of each method

	Conv.	Recon.	times
GMM	2.96	-	2
NMF (all)	3.05	1.37	890
NMF (10,000)	3.11	1.59	680
NMF (5,000)	3.05	1.56	310
NMF (1,000)	3.11	1.59	71
PDL (1,000)	3.11	1.85	71
DGNMF (1,000)	3.08	1.09	71

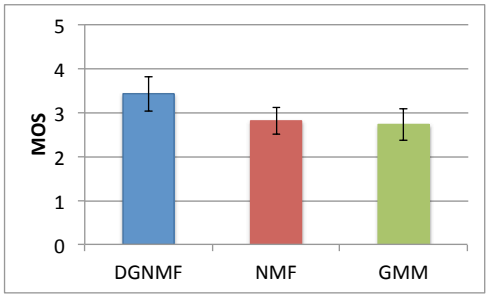

Fig. 3　MOS test on speech quality

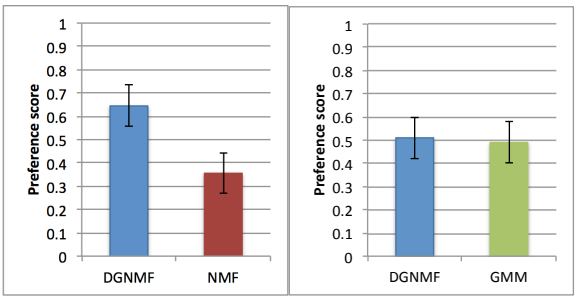

Fig. 4　Preference score on individuality

5　Conclusions

NMF 声質変換の変換精度を向上させるため，DGNMF を用いたパラレル辞書学習を提案した．提案手法である DGNMF はクラス間識別を導入した GNMF [8] に対して，音素ラベルによる識別とクラス内分散を導入したもので，NMF ベースの辞書学習における，辞書の過学習を防ぐことができる．実験結果により，DGNMF によるパラレル辞書学習は NMF を声質変換の精度を向上させたのみならず，基底数削減により計算コストの削減も可能にした．今後は，この手法を多対多声質変換 [10] や構音障害者のための声質変換 [11] に応用させていく予定である．

参考文献

[1] T. Toda *et al.*, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 8, pp. 2222–2235, 2007.

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

[3] R. Takashima *et al.*, "Exemplar-based voice conversion in noisy environment," in Proc. SLT, pp. 313–317, 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] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in Proc. Interspeech, 2006.

[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] H. Sawada *et al.*, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization," in Proc. ICASSP, pp. 261–264, 2012.

[8] D. Cai *et al.*, "Graph regularized nonnegative matrix factorization for data representation," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 33, no. 8, 2010.

[9] H. Kawahara, "STRAIGHT, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," Acoustical Science and Technology, pp. 349–353, 2006.

[10] R. Aihara *et al.*, "Multiple non-negative matrix factorization for many-to-many voice conversion," IEEE/ACM Trans. on Audio, Speech, and Language Processing, vol. 24, no. 7, pp. 1175–1184, 2016.

[11] R. Aihara *et al.*, "A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2014:5, doi:10.1186/1687-4722-2014-5, 2014.