Alternating Direction Method of Multipliers による NMF 声質変換のためのパラレル辞書学習 *

◎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]．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．

しかしながら，NMF 声質変換には，計算コストが高いという問題があった．その理由には，大きく分けて以下の３つが存在する．

1. 高次元スペクトルの利用：非負制約のため，NMF 声質変換では使用できる特徴量が限定され，Mel-cepstrum といった負値を含む低次元特徴量を使用できず，計算コストが大きくなる．
2. 膨大な基底数：基本的に，NMF 声質変換では学習データの全てのスペクトルを基底として用いるため辞書の基底数が大きくなる．
3. 最適化手法： 従来の NMF 声質変換では，補助関数法を用いた最適化を行うため，コスト関数の収束に時間が掛かる．

本論文では，上記の問題点を解決するため，以下の３つの手法を採用する．

1. **Semi-Non-negative Matrix Factorization (Semi-NMF)**：入力特徴量，辞書行列の非負制約を取り除き，負値を含むコンパクトな特徴量を利用する．
2. 辞書学習： Semi-NMF を用いた辞書学習を提案し，基底数の削減を行う．
3. **Alternating Direction Method of Multipliers (ADMM)**：補助関数法ではなく，制約付き最適化手法のひとつである ADMM を用いることで，収束速度を向上させる．

補助関数法を用いた Semi-NMF は文献 [4] で提案されているが，これまで声質変換に Semi-NMF が用いられた例はない．ADMM を用いた NMF は文献 [5] において提案されており，本研究ではこの手法を Semi-NMF へ拡張する．

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 先行研究

2.1 NMF 声質変換

2.1.1 概要

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．NMF 声質変換では，基底は学習データのスペクトルであり，基底の集合 $\mathbf{W}$ を “辞書”，基底の線形結合重みの集合 $\mathbf{H}$ を “アクティビティ” と呼ぶ．このアクティビティがスパースであるとき，観測信号 $\mathbf{V}$ 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．

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

$$\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_L], \quad \mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_L]. \tag{2}$$

ここで L はフレーム数を表す．本手法において，$\mathbf{W}$ は学習データで固定され，NMF [2] のアルゴリズムを用いて入力スペクトルから $\mathbf{H}$ を推定する．

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．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．

2.1.2 問題点

NMF における非負制約のため，NMF 声質変換において用いられる特徴量は負値を含まない線形スペクトルなどに限定される．したがって，Δ や $\Delta\Delta$ 特徴量といった動的特徴量も使用できない．文献 [6] においては，513 次元の線形スペクトルとそのセグメント

* Parallel Dictionary Learning for NMF-based Voice Conversion Using Alternating Direction Method of Multipliers by Ryo AIHARA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

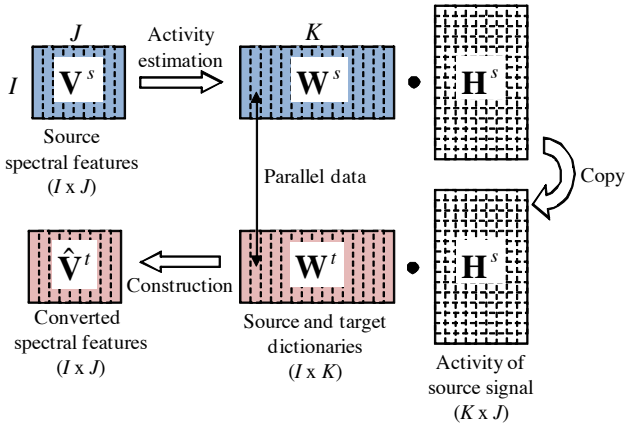

Fig. 1　Basic approach of NMF-based voice conversion

特徴量を使用しており，結果としてメモリ使用量が多くなるという問題が指摘されている．さらに，NMF 声質変換においてはパラレル学習データ全てがパラレル辞書として用いられる．アクティビティは多くの基底をもつ辞書から推定されるため計算コストが高くなる．これらの点から，NMF 声質変換はリアルタイムでの実現が困難であるという問題がある．

さらに文献 [6] においては，アライメントのずれが引き起こす NMF 声質変換の精度劣化が指摘されている．文献 [7] では，この問題を解決するべくアクティビティマッピング手法が提案されているが，学習データに加えて適応データを必要とするため，実用性に乏しいという問題点があった．

2.2　補助関数法を用いた Semi-NMF

2.2.1　定式化

Semi-NMF のコスト関数は以下のように定義できる．

$$d_F(\mathbf{V}, \mathbf{WH}) + \lambda||\mathbf{H}||_1 \quad s.t. \quad \mathbf{H} \geq 0 \tag{3}$$

ここで，第 1 項は $\mathbf{V}$ と $\mathbf{WH}$ の間のフロベニウスノルムであり，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．NMF と比較すると，Semi-NMF においては $\mathbf{W}$ の非負制約をはずしたため，コスト関数がフロベニウスノルムに限定される．

コスト関数は以下の更新式を繰り返し適用することで最小化される．

$$\mathbf{H} \leftarrow (-\lambda\mathbf{H}^{\mathsf{T}} + (\mathbf{H}^{\mathsf{T}}.*\sqrt{\lambda^2 + 16(\mathbf{A}.*\mathbf{B})})) ./(4\mathbf{A}) \tag{4}$$

$$\mathbf{A} = (\mathbf{V}^{\mathsf{T}}\mathbf{W})^{-} + (\mathbf{H}^{\mathsf{T}}(\mathbf{W}^{\mathsf{T}}\mathbf{W})^{+}) \tag{5}$$

$$\mathbf{B} = (\mathbf{V}^{\mathsf{T}}\mathbf{W})^{+} + (\mathbf{H}^{\mathsf{T}}(\mathbf{W}^{\mathsf{T}}\mathbf{W})^{-}) \tag{6}$$

ここで，正部と負部を分けるため，$\mathbf{X}^{+} = (|\mathbf{X}| + \mathbf{X})/2, \mathbf{X}^{-} = (|\mathbf{X}| - \mathbf{X})/2$ のように定める．

2.2.2　問題点

Semi-NMF は負値を含む特徴量を分解することができるため，Mel-cepstrum や動的特徴量を用いることができ，NMF 比較して，メモリ使用量を削減することができる．しかしながら，式 (6) には $\mathbf{W}^{\mathsf{T}}\mathbf{W}$ の項がある．従来のパラレル辞書のような基底数の多い辞書行列を用いた場合，この項の計算に膨大なメモリを使用し，計算コストが高くなる問題がある．従って，Semi-NMF を用いた声質変換を考える場合，コンパクトなパラレル辞書の学習が必要となる．さらに，補助関数法による最適化は収束速度が遅いという問題もある．

3　Semi-Non-negative Matrix Factorization を用いた声質変換

3.1　Basic Idea

前章の問題点を解決するため，本研究では ADMM に基づく Semi-NMF を用いた声質変換法を提案する．まず，Semi-NMF を用いたパラレル制約付き辞書学習で入力と出力のパラレル辞書を学習する．パラレル制約は，従来の NMF 声質変換で問題となっていたアクティビティのずれを解消する手法である．さらに，コンパクトな辞書行列によって計算コストの削減が期待できる．

変換時には，ADMM ベースの Semi-NMF を用いて，入力スペクトルは学習された入力辞書基底の線形結合で表現される．ADMM を用いた Semi-NMF を用いることで，補助関数法を用いた Semi-NMF と比較して，よりスパースで精度の高いアクティビティが得られる．

3.2　辞書学習

基底数の少ない，コンパクトな辞書を学習するため，入力話者と出力話者のパラレル辞書は，パラレル制約付き Semi-NMF によって推定される．

目的関数を下記のように定義する．

$$\begin{aligned} min \quad & d_F(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + d_F(\mathbf{V}^t, \mathbf{W}^t\mathbf{H}^t) \\ & + \frac{\epsilon}{2}||\mathbf{H}^s - \mathbf{H}^t||_F^2 + \lambda||\mathbf{H}^s||_1 + \lambda||\mathbf{H}^t||_1 \\ sub\ to \quad & \mathbf{H}^s = \mathbf{H}^s_+, \mathbf{H}^s_+ \geq 0, \mathbf{H}^t = \mathbf{H}^t_+, \mathbf{H}^t_+ \geq 0 \end{aligned} \tag{7}$$

ここで，$\mathbf{V}^s$，$\mathbf{V}^t$，$\mathbf{W}^s$，$\mathbf{W}^t$，$\mathbf{H}^s$，$\mathbf{H}^t$ はそれぞれ，入力話者と出力話者のパラレル学習データ，推定する入力話者と出力話者のパラレル辞書行列，入力と出力のアクティビティを表す．パラレル学習データは DTW でアライメントを取られたものを用いる．ϵ と λ はそれぞれ，パラレル制約重みとスパース制約重みを表す．式 (7) のラグランジアンは下記のようになる．

$$\begin{aligned} L_\rho(\mathbf{W}^s, \mathbf{H}^s, \mathbf{W}^t, \mathbf{H}^t, \mathbf{H}^s_+, \mathbf{H}^t_+) = \\ & d_F(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + d_F(\mathbf{V}^t, \mathbf{W}^t\mathbf{H}^t) \\ & + \frac{\epsilon}{2}||\mathbf{H}^s - \mathbf{H}^t||_F^2 \\ & + \lambda||\mathbf{H}^s||_1 + \lambda||\mathbf{H}^t||_1 \\ & + \langle \alpha_{\mathbf{H}^s}, \mathbf{H}^s - \mathbf{H}^s_+ \rangle + \langle \alpha_{\mathbf{H}^t}, \mathbf{H}^t - \mathbf{H}^t_+ \rangle \\ & + \frac{\rho}{2}||\mathbf{H}^s - \mathbf{H}^s_+||_F^2 + \frac{\rho}{2}||\mathbf{H}^t - \mathbf{H}^t_+||_F^2 \end{aligned} \tag{8}$$

ここで，ρ は収束率を調性するチューニングパラメータである．式 (8) は，Table 1 に示されたアルゴリズムを適用することで最適化される．

Table 1　Algorithm of Dictionary Learning

Input $\mathbf{V}^s, \mathbf{V}^t$
Initialize $\mathbf{W}^s, \mathbf{H}^s, \mathbf{W}^t, \mathbf{H}^t, \mathbf{H}^s_+, \mathbf{H}^t_+, \alpha_{\mathbf{H}^s}, \alpha_{\mathbf{H}^t}$
Repeat
$\mathbf{W}^s \leftarrow (\mathbf{V}^s(\mathbf{H}^{s\mathsf{T}}))/(\mathbf{H}^s(\mathbf{H}^{s\mathsf{T}}))$
$\mathbf{W}^t \leftarrow (\mathbf{V}^t(\mathbf{H}^{t\mathsf{T}}))/(\mathbf{H}^t(\mathbf{H}^{t\mathsf{T}}))$
$\mathbf{H}^s \leftarrow (2\mathbf{W}^{s\mathsf{T}}\mathbf{W}^s + (\rho+\epsilon)\mathbf{I})$
$\backslash(2\mathbf{W}^{s\mathsf{T}}\mathbf{W}^s - \alpha_{\mathbf{H}^s} + \rho\mathbf{H}^s_+ + \epsilon\mathbf{H}^t - \lambda)$
$\mathbf{H}^t \leftarrow (2\mathbf{W}^{t\mathsf{T}}\mathbf{W}^t + (\rho+\epsilon)\mathbf{I})$
$\backslash(2\mathbf{W}^{t\mathsf{T}}\mathbf{W}^t - \alpha_{\mathbf{H}^t} + \rho\mathbf{H}^t_+ + \epsilon\mathbf{H}^s - \lambda)$
$\mathbf{H}^s_+ \leftarrow max(\mathbf{H}^s + \frac{1}{\rho}\alpha_{\mathbf{H}^s}, 0)$
$\mathbf{H}^t_+ \leftarrow max(\mathbf{H}^t + \frac{1}{\rho}\alpha_{\mathbf{H}^t}, 0)$
$\alpha_{\mathbf{H}^s} \leftarrow \alpha_{\mathbf{H}^s} + \rho(\mathbf{H}^s - \mathbf{H}^s_+)$
$\alpha_{\mathbf{H}^t} \leftarrow \alpha_{\mathbf{H}^t} + \rho(\mathbf{H}^t - \mathbf{H}^t_+)$
Until convergence **return** $\mathbf{W}^s, \mathbf{H}^s_+, \mathbf{W}^t, \mathbf{H}^t_+$

3.3　変換

コンパクトなパラレル辞書行列 $\mathbf{W}^s$，$\mathbf{W}^t$ が推定された後，入力スペクトル $\mathbf{V}^s$ は，ADMM に基づく Semi-NMF を用いて変換スペクトル $\hat{\mathbf{V}}^t$ へ変換される．目的関数は下記のように定める．

$$\begin{aligned} min \quad & d_F(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \\ sub\ to \quad & \mathbf{H}^s = \mathbf{H}^s_+, \mathbf{H}^s_+ \geq 0. \end{aligned} \tag{9}$$

式 (9) のラグランジアンは次のようになる．

$$\begin{aligned} &L_\rho(\mathbf{W}^s, \mathbf{H}^s, \mathbf{H}^s_+) = \\ &\quad d_F(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \\ &\quad + \langle \alpha_{\mathbf{H}^s}, \mathbf{H}^s - \mathbf{H}^s_+ \rangle + \frac{\rho}{2}||\mathbf{H}^s - \mathbf{H}^s_+||^2_F \end{aligned} \tag{10}$$

ここで，$\mathbf{W}^s$ は固定され，$\mathbf{H}^s$ は Table 2 に示されたアルゴリズムによって推定される．

Table 2　Algorithm of Conversion

Input $\mathbf{V}^s, \mathbf{W}^s$
Initialize $\mathbf{H}^s, \mathbf{H}^s_+, \alpha_{\mathbf{H}^s}$
Repeat
$\mathbf{H}^s \leftarrow (2\mathbf{W}^{s\mathsf{T}}\mathbf{W}^s + \rho\mathbf{I})$
$\backslash(2\mathbf{W}^{s\mathsf{T}}\mathbf{W}^s - \alpha_{\mathbf{H}^s} + \rho\mathbf{H}^s_+ - \lambda)$
$\mathbf{H}^s_+ \leftarrow max(\mathbf{H}^s + \frac{1}{\rho}\alpha_{\mathbf{H}^s}, 0)$
$\alpha_{\mathbf{H}^s} \leftarrow \alpha_{\mathbf{H}^s} + \rho(\mathbf{H}^s - \mathbf{H}^s_+)$
Until convergence **return** $\mathbf{H}^s_+$

推定されたアクティビティ $\mathbf{H}^s$ と辞書学習によって求められた出力辞書行列 $\mathbf{W}^t$ によって変換スペクトル $\hat{\mathbf{V}}^t$ は以下のように求められる．

$$\hat{\mathbf{V}}^t = \mathbf{W}^t\mathbf{H}^s \tag{11}$$

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 216 単語を学習データとし，音素バランス文 50 文をテストデータとして用いた．提案手法において，ρ, ϵ, λ はそれぞれ，1, 1, 0.1 とした．Semi-NMF の更新回数は辞書学習時には 50，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 GMM 声質変換において，音声分析合成手法 [8] 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．一方，NMF 声質変換では，STRAIGHT 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 の混合数は 128 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・考察

まず，ADMM ベースの Semi-NMF と，補助関数法を用いた Semi-NMF の間の収束速度を比較した．辞書の基底数は 1,000 とした．結果を Fig. 2 に示す．横軸は更新回数，縦軸は log スケールにおける目的関数の値である．図より，ADMM を用いた場合，ρ が小さくなるにしたがって収束率が向上し，従来の補助関数法に基づく Semi-NMF を上回っていることがわかる．

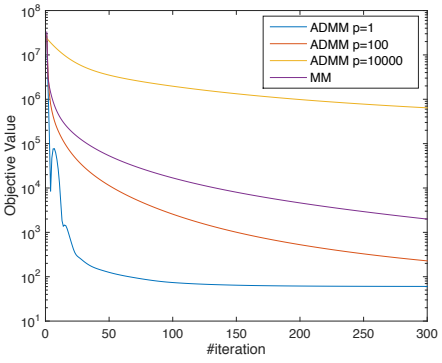

Fig. 2　Objective value as a function of iteration

客観評価指標として，Normalized Spectrum Distortion (NSD) を用いた．

$$NSD = \sqrt{||\mathbf{X}^t - \hat{\mathbf{X}}^t||^2 / ||\mathbf{X}^t - \mathbf{X}^s||^2} \tag{12}$$

ここで，$\mathbf{X}^s, \mathbf{X}^t, \hat{\mathbf{X}}^t$ はそれぞれ入力スペクトル，出力スペクトル，変換スペクトルを表す．Fig. 3 と Table 3

にそれぞれの手法における NSD と計算時間を示す．提案手法において，辞書の基底数を 1,000 とした場合と 5,000 とした場合では，NSD にほとんど差がないことがわかる．提案手法は NMF 声質変換と比較して，わずかに NSD が大きくなっているが計算時間が大幅に削減できている．提案手法と GMM 声質変換の間の NSD はほとんど差がない．

Table 3　NSD and computational times of each method

	NSD	times [s]
GMM	1.66	2
NMF	1.54	916
Proposed(1,000)	1.69	12
Proposed(5,000)	1.70	310

主観評価として，15 人の日本語話者によるヘッドホンを用いた聴取実験を行った．客観評価の結果から，提案手法における辞書の基底数は 1,000 とした．

Fig. 3 の左側に，音質の評価結果を示す．評価基準として，Mean Opinion Score (MOS) による 5 段階評価（5：とても良い，4：良い，3：普通，2：悪い，1：とても悪い）を用いた．図より，提案手法，NMF 声質変換はそれぞれ GMM 声質変換よりも音質が優れていることがわかる．この結果は t 検定により有意傾向が示されている．

Fig. 3 の右側に話者性の評価結果を示す．評価基準は 5 段階評価（5：とても近い，4：近い，3：普通，2：遠い，1：とても遠い）を用いた．図より，3 つの手法の間には有意差がないことがわかる．

以上の結果から，提案手法は従来の NMF 声質変換法とほぼ同等の精度を得られ，計算コストを削減することができたことがわかる．

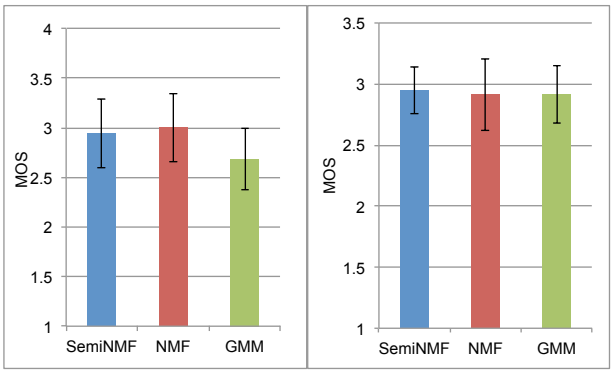

Fig. 3　MOS test on speech quality (left) and similarity (right)

5　おわりに

本稿では，ADMM に基づく最適化手法を採用した Semi-NMF による声質変換法を提案した．従来の NMF 声質変換の問題点であった計算コストとメモリ使用量を削減するため，NMF は Semi-NMF で置き換えられ，よりコンパクトなスペクトル特徴量を使用することが可能になった．さらに，従来の補助関数法に基づく Semi-NMF は収束速度に問題があったため，ADMM に基づく Semi-NMF を提案し，より少ない計算時間でスパースなアクティビティが得られるようになった．また，基底数の少ないパラレル辞書行列を求めるため，パラレル辞書学習法を提案した．評価実験により，提案手法は従来の NMF 声質変換とほぼ同程度の精度で変換が可能でありながら，計算時間を大幅に削減することが可能になった．この手法はトピックモデルや超解像など，他のタスクにも応用可能であると考えられる．

今後の課題として，依然として提案手法の計算コストが GMM 声質変換と比較して高いことがあげられる．また，今後は提案手法を構音障害者のための声質変換に応用していく予定である．

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] C. Ding *et al.*, “Convex and semi-nonnegative matrix factorization,” IEEE Trans. Pettern Analysis And Machine Intelligence, vol. 32, no. 1, pp. 45–55, 2010.

[5] D. L. Sun and C. Févotte, “Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence,” in Proc. ICASSP, pp. 6242–6246, 2014.

[6] R. Aihara *et al.*, “Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary,” in Proc. ICASSP, pp. 7944–7948, 2014.

[7] R. Aihara *et al.*, “Activity-mapping non-negative matrix factorization for exemplar-based voice conversion,” in Proc. ICASSP, pp. 4899–4903, 2015.

[8] 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.