アクティビティマッピングによる非負値行列因子分解を用いた声質変換*

☆相原龍, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

近年，スパース表現は音声信号処理分野で大きな注目を集めている．一般的に入力信号を少量の基底ベクトルの線形結合で表現するアプローチをスパース信号分解と呼ぶ．なかでも非負値行列因子分解（Non-negative Matrix Factorization : NMF）[1] は文書データ解析 [2]，画像解析 [3]，遺伝子解析など幅広い分野で応用されており，音声信号処理においては雑音除去や音声強調においてその効果を発揮している．例えば音源分離に用いる場合，まず学習サンプルや基底を音源毎にグループ（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 [4] 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) の尤度算出に用いることで，雑音にロバストな音声認識を行う手法を提案している．

我々は，雑音にロバストな声質変換手法として NMF を用いた声質変換を提案してきた [5]．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り，音韻情報を維持しつつ話者情報を変換する “話者変換” [6] を目的として広く研究されてきたが，近年では，音声合成や音声認識における話者性の制御 [7] に用いられている他，感情情報を変換する “感情変換” [8, 9]，失われた話者情報を復元する “発話支援” [10] など多岐にわたって応用されている．声質変換の実用化を考えた場合，背景雑音は避けられない問題であり，NMF 声質変換は雑音除去を声質変換と一体で行うことにより一定の成果をあげることができた．

従来の声質変換で一般的な手法は混合正規分布モデル (Gaussian Mixture Model : GMM) を用いた手法 [6] であった．GMM 声質変換では，基本的には変換関数を目標話者と入力話者のスペクトル包絡の期待値によって表現し，変数をパラレルな学習データから最小二乗法で推定する．これまで，過学習と変換スペクトルの過剰な平滑化が問題点としてあげられていた．戸田ら [11] は従来の GMM を用いた声質変換法に動的特徴と Global Variance を導入することでより自然な音声として変換する手法を提案している．Helander ら [12] は従来手法における過適合の問題を回避するため，Partial Least Squares(PLS) 回帰分析を用いる手法を提案している．NMF 声質変換は GMM 声質変換と異なり，統計的モデルを用いない exemplar-based であるため，過学習の問題を避けることができ，GMM 声質変換比較して自然な音声が得られるのではないかと期待されてきた．

本論文では，声質変換においてもっとも一般的な，音声スペクトルを特徴量とした話者変換をタスクとし，NMF を用いた声質変換手法の精度を向上させるため，アクティビティマッピングを提案する．これまで，NMF 声質変換では，入力スペクトルから推定された基底の係数（アクティビティ）を出力話者辞書のパラレルな基底の係数と同一視して変換を行っていた．すなわち，同一発話内容であれば，話者にかかわらず選ばれる基底は同じであるという仮定していた．実際には，基底には多様なスペクトルが含まれるため，同一発話内容であっても選ばれる基底は話者によって異なり，話者間の “アクティビティ不一致問題” が発生していた．本論文では，この問題を解消するため，入力話者のアクティビティを出力話者のものに変換するマッピング行列を導入し，これを推定するためのアクティビティ適応型 NMF を提案する．

以下，第 2 章でこれまでの NMF による声質変換手法を述べ，第 3 章で本稿の提案手法を説明する．第 4 章で従来の GMM・NMF による声質変換手法と比較し，第 5 章で本稿をまとめる．

2 NMF による声質変換

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 はフレーム数を表す．

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，$\hat{\mathbf{V}}^t$ は変換音声，$\mathbf{H}^s$ 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．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て辞書化したものである．入力音声を入力話者辞書のスパース表現にし，得られたアクティビティ行列と出力話者辞書の内積をとることで，出力話者の音声へと変換する．

---

*Exemplar-based Voice Conversion Based on Activity-adaptive Non-negative Matrix Factorization by Ryo AIHARA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

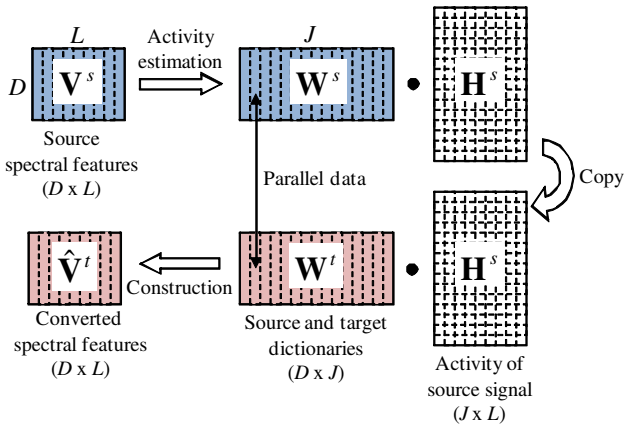

Fig. 1　Basic approach of NMF-based voice conversion

本手法では，アクティビティ行列の推定にスパースコーディングの代表的手法である NMF を用いる [13]. NMF のコスト関数は，$\mathbf{V}^s$，$\mathbf{W}^s$，$\mathbf{H}^s$ を用いて以下のような式で表せる．

$$d(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \tag{4}$$

ここで，第 1 項は $\mathbf{V}^s$ と $\mathbf{W}^s\mathbf{H}^s$ の間の Kullback-Leibler(KL) 距離であり，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．このコスト関数は Jensen の不等式を用いることで，繰り返し適用を用いて最小化できる．

$$\begin{aligned}\mathbf{H}^s_{n+1} &= \mathbf{H}^s_n .* (\mathbf{W}^{s\mathsf{T}}(\mathbf{V}^s ./ (\mathbf{W}^s\mathbf{H}^s))) \\ &\quad ./ (\mathbf{H}^{s\mathsf{T}}\mathbf{1}^{D\times L} + \lambda \mathbf{1}^{1\times L})\end{aligned} \tag{5}$$

変換音声 $\hat{\mathbf{V}}^t$ は出力話者辞書行列と推定されたアクティビティの内積をとることで得られる．

$$\hat{\mathbf{V}}^t = \mathbf{W}^t\mathbf{H}^s \tag{6}$$

3　アクティビティマッピングによる声質変換

3.1　概要

前章で述べた NMF を用いた声質変換法では，入力話者の辞書行列から推定したアクティビティをパラレルな出力話者の辞書行列と内積をとることで変換していた．これは，“仮に入力話者の音声と，それと同一発話の出力話者の音声をそれぞれ入力辞書と出力辞書のスパース表現にした場合，それぞれから得られるアクティビティ行列は互いに類似している” という仮定に基づくものであった．

Fig. 2 に 2 人の日本語話者によって発話されたスペクトル “あこがれる”から推定したアクティビティを示す．発話スペクトルは DTW でアライメントがとられたもので，推定に用いた辞書は，パラレルな 1,000 基底から構成されているものを用いた．Fig. 2 からわかるように，同一発話スペクトルのアクティビティには話者によって差異があることがわかる．このアクティビティの差は話者性の違いと考えることができ，アクティビティを変換することにより声質変換の精度向上が期待できる．

本研究ではアクティビティ適用 NMF を提案し，入力話者アクティビティを出力話者アクティビティへと変換する．Fig. 1 で示されていた NMF 声質変換の概要は Fig. 3 のように変化する．第 1 段階では，入力話者スペクトル $\mathbf{V}^s$ が学習された入力話者辞書行列 $\mathbf{W}^s$ の線形結合で表現される．このとき，基底の結合重みがアクティビティ $\mathbf{H}^s$ として NMF を用いて推定される．第 2 段階では，入力話者アクティビティ $\mathbf{H}^s$ が．学習された変換行列 $\mathbf{A}$ によって出力話者アクティビティ $\mathbf{H}^t$ へと変換される．第 3 段階では，変換されたアクティビティ $\mathbf{H}^t$ と出力話者辞書行列 $\mathbf{W}^t$ によって変換スペクトル $\hat{\mathbf{V}}^t$ が得られる．

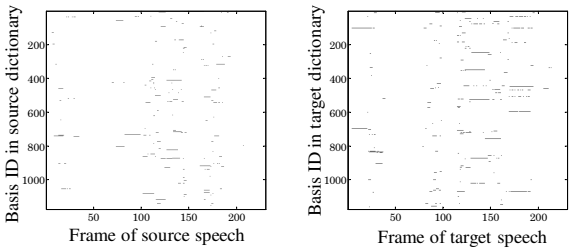

Fig. 2　Activity matrices for parallel utterances

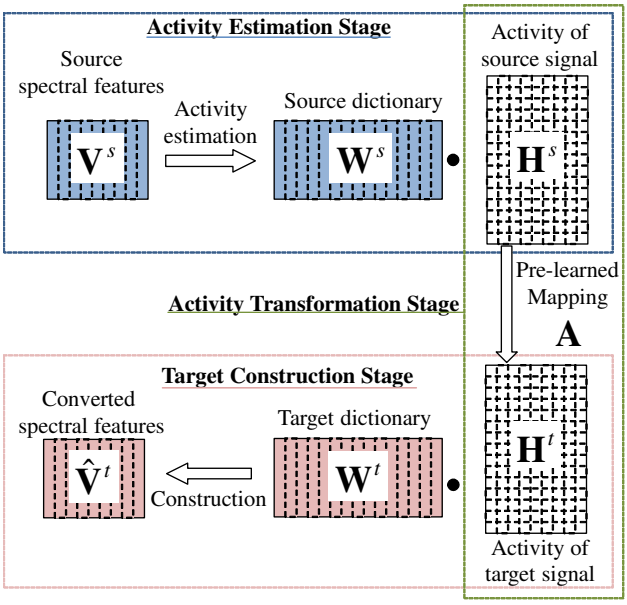

Fig. 3　Flow chart of Voice Conversion Using Activity-mapping

3.2　アクティビティ適応型 NMF

従来の NMF 声質変換にアクティビティ適応を導入する．事前にアクティビティ変換行列 $\mathbf{A}$ を入力話者，出力話者のパラレルな適応データを用いて，推定する．アクティビティ適応型 NMF のコスト関数は以下のように定義できる．

$$d(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 + d(\mathbf{V}^t, \mathbf{W}^t\mathbf{A}\mathbf{H}^s) + \lambda||\mathbf{A}\mathbf{H}^s||_1 \tag{7}$$

第 1 項，第 2 項は式 (4) と同様である．第 3 項は，$\mathbf{V}^t$ と $\mathbf{W}^t\mathbf{A}\mathbf{H}^t$ の間の KL 距離であり，第 4 項は L1 ノルム制約項である．

ここで，$\mathbf{H}^s$ は，変換時に入力話者アクティビティをより正確に推定するため，式（7）の第 1 項，第 2 項からのみ推定し，従来の NMF 声質変換と同様の更新式で求める．アクティビティ行列 $\mathbf{A}$ は，式 (7) に対して，Jensen の不等式を用いることで以下のように求められる．

$$\mathbf{A}_{n+1} = \mathbf{A}_n . / ((\mathbf{W}^{t\mathsf{T}} \mathbf{1}^{(I\times J)}) . * (\mathbf{1}^{(J\times L)} \mathbf{H}^{s\mathsf{T}})) \\ . * (\mathbf{W}^{t\mathsf{T}} (\mathbf{V}^t . / (\mathbf{W}^t \mathbf{A}_n \mathbf{H}^s)) \mathbf{H}^{s\mathsf{T}}) \quad (8)$$

変換時には，入力スペクトル $\mathbf{V}^s$ に対して，式 (5) を用いて入力話者アクティビティ $\mathbf{H}^s$ を推定する．適応時に推定しておいたアクティビティ変換行列 $\mathbf{A}$ を用いて，出力話者アクティビティ $\mathbf{H}^t$ が得られ，変換スペクトルは以下の式で求まる．

$$\hat{\mathbf{V}}^t = \mathbf{W}^t \mathbf{H}^t = \mathbf{W}^t \mathbf{A} \mathbf{H}^s \quad (9)$$

3.3 辞書分割・選択

第 2 章で述べた従来の NMF 声質変換では，学習データ全てをひとつの辞書行列のペアとして用いていた．しかしながら，アクティビティ適応を導入する際，基底数の多い辞書行列ペアのアクティビティを，ひとつの変換行列で変換することは困難である．したがって，本論文ではパラレルデータをいくつかのクラスタに分類し，アクティビティ適応をクラスタ毎に適用する．つまり，アクティビティ適応行列はクラスタの数だけ存在する．

本論文では，k 近傍法のコスト関数をユークリッド距離から KL 距離に置き換えたものを用いて，パラレルデータをクラスタリングする．入力スペクトルベクトル $\mathbf{v}_l = [\mathbf{v}_l^{s\mathsf{T}}, \mathbf{v}_l^{t\mathsf{T}}]^\mathsf{T}$ は次のようなコスト関数によってクラスタリングされる．

$$Dis = \sum_{l=1}^{L} d(\mathbf{v}_l, \mathbf{m}_{c_l}) \quad (10)$$

$\mathbf{v}_l$，$\mathbf{m}_{c_l}$，L はそれぞれ l 番目の入力ベクトル，c_l-番目のクラスタ，フレーム数を表す．c_l は l 番目のフレームが属するクラスタを表し，以下のように決定される．

$$c_l = \arg\min_k d(\mathbf{v}_l, \mathbf{m}_k) \quad (11)$$

ここで，K はクラスタ数を表す．

変換時は NMF を用いて，入力スペクトルベクトルに対して次式を用いて辞書を選択する．

$$c_l = \arg\min_k d(\mathbf{v}_l^s, \mathbf{W}_k^s \mathbf{h}_{kl}^s) + \lambda ||\mathbf{h}_{kl}^s||_1 \\ s.t. \quad \mathbf{h}_{kl}^s \geq 0 \quad (12)$$

提案手法のアルゴリズムを Table 1 に示す．

Table 1　Algorithm of Activity-mapping VC

Initialize for estimating activity-mapping
Set source and target exemplars to $\mathbf{V}^s$ and $\mathbf{V}^t$.
Set source and target dictionaries to $\mathbf{W}^s$ and $\mathbf{W}^t$.
$\mathbf{A}$ and $\mathbf{H}^s$ are initialized with a random matrix.
Clustering
Jointed $\mathbf{V}^s$ and $\mathbf{V}^t$ are clustered by (10).
For each iteration
For each cluster
- Optimize $\mathbf{H}^s$ by (5)
- Optimize $\mathbf{A}$ by (8)

Initialize for Conversion
Set input spectra $\mathbf{V}^s$, mapping matrix $\mathbf{A}$
source dictionary $\mathbf{W}^s$, target dictionary $\mathbf{W}^t$.
Clustering
Cluster the input spectrum $\mathbf{v}_l^s$ by (12).
For each iteration
For each cluster
- Optimize $\mathbf{H}^s$ by (5)
- Construct $\hat{\mathbf{V}}^t$ by (9)

4 評価実験

4.1 実験条件

本実験では ATR 研究用日本語音声データベース [14] より，男性話者 1 名の音声を入力話者音声に，女性話者 1 名の音声を出力話者音声として用いた．サンプリング周波数は 12kHz である．音素バランス 50 文を学習データとし，従来手法における GMM の学習ならびに従来の NMF による手法，提案手法における辞書の構築にそれぞれ用いた．提案手法のアクティビティ適応には，学習に含まれない 50 文を用いた．評価には学習・適応に含まれない 50 文を用いた．サンプルベースによる手法及び提案手法では STRAIGHT スペクトル [15] と前後 2 フレームを含む 2,565 次元特徴量とした．それぞれの手法において NMF の更新回数は 500 とした．GMM を用いた従来手法では，STRAIGHT スペクトルから計算された MFCC+ΔMFCC+ΔΔMFCC の 60 次元を特徴量とした．GMM の混合数は 96 である．本稿では，提案手法・従来手法ともに F0 情報は従来の単回帰分析により変換し，非周期成分は変換せず入力音声のものをそのまま用いている．

提案手法の有効性を確かめるため，客観評価を行った．客観評価は STRAIGHT スペクトル 516 次元を特徴量とし，式 (13) で表される SDIR (Spectral Distortion Improvement Ratio) [dB] によって各手法を比較した．

$$SDIR[dB] = 10 \log_{10} \frac{\sum_d |\mathbf{V}^t(d) - \mathbf{V}^s(d)|^2}{\sum_d |\mathbf{V}^t(d) - \hat{\mathbf{V}}^t(d)|^2} \quad (13)$$

ただし，$\mathbf{V}^s$，$\mathbf{V}^t$，$\hat{\mathbf{V}}^t$ はそれぞれ入力話者のスペクトル，出力話者のスペクトル，変換後のスペクトルを表す．

4.2 実験結果・考察

Fig. 4 に，各手法による SDIR を示す．なお，“NMF” は第 2 章で述べた従来の NMF を用いた手法，“sub-NMF” は提案手法のうち 3.3 章で述べた辞書分割・選択は行うがアクティビティ適応を行わずに変換した手法，“Act-NMF” は本論文の提案手法であ

るアクティビティ適応を辞書分割したうえで行った手法である．

Fig. 4 から，従来の NMF，GMM による手法を比較すると，GMM による手法の方が優れていることがわかる．提案手法は，これら 2 つの手法よりもすぐれた SDIR を示してしていることから，本論文において提案したアクティビティ適応は，NMF 声質変換において有効であることがわかる．しかしながら，辞書分割のみを行った手法は，従来の NMF 声質変換よりも低い値となっている．このことから，本論文で用いた辞書分割・選択法には改善の余地があることがわかる．より適切な辞書分割・選択が可能になれば，提案手法であるアクティビティ適応による声質変換の精度もさらに向上させられると考えられる．

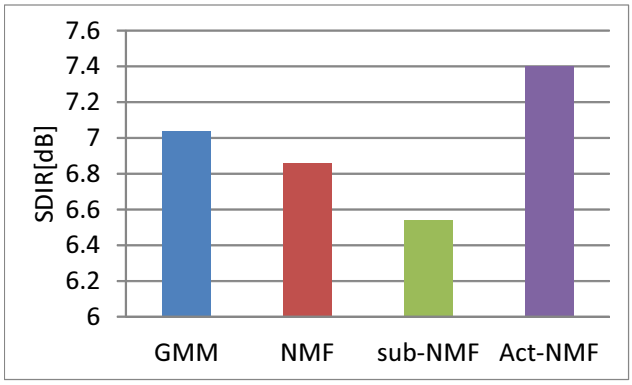

Fig. 4　SDIR of each method

5　おわりに

本論文では，これまで提案してきた NMF に基づく声質変換法において，アクティビティの不一致という問題に対処するため，入力話者と出力話者のスパース係数を変換する “アクティビティマッピング” を導入し，それを可能にするためのアクティビティ適応型 NMF を提案した．客観評価により，提案手法の有効性が示された．

従来の NMF 声質変換は 1 対のパラレル辞書行列を用いていたが，1 つのアクティビティ変換行列であらゆるスパース係数を変換することは困難であった．本論文では，パラレル辞書をいくつかの副辞書に分割するため k 近傍法をベースとした辞書分割をおこなった．しかしながら本論文で用いた辞書分割・選択法は従来の単一パラレル辞書を用いる手法と比較して変換精度が劣化している．文献 [16] でわれわれは NMF 声質変換における音素カテゴリ選択を提案している．今後は，この手法と合わせてさらによい辞書分割・選択法を研究する必要がある．

さらに，雑音辞書を結合することで雑音重畳音声の分離と声質変換を行う手法の性能を評価する予定である．また，構音障害者のための発話支援にも応用していく．

参考文献

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

[2] W. Xu *et al.*, “Document clustering based on non-negative matrix factorization,” Proc. 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 267–273, 2003.

[3] M. Rajapakse *et al.*, “Color channel encoding with nmf for face recognition,,” International Conference on Image Processing, pp. 2007–2010, 2004.

[4] J. F. Gemmeke and T. Virtanen, “Noise robust exemplar-based connected digit recognition,” in ICASSP, pp. 4546–4549, 2010.

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

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

[7] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” in ICASSP, vol. 1, pp. 285–288, 1998.

[8] C. Veaux and X. Robet, “Intonation conversion from neutral to expressive speech,” in Interspeech, pp. 2765–2768, 2011.

[9] R. Aihara *et al.*, “GMM-based emotional voice conversion using spectrum and prosody features,” American Journal of Signal Processing, vol. 2, no. 5, pp. 134–138, 2012.

[10] 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.

[11] 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.

[12] E. Helander *et al.*, “Voice conversion using partial least squares regression,” IEEE Trans. Audio, Speech, Lang. Process., vol. 18, Issue:5, pp. 912–921, 2010.

[13] J. F. Gemmeke *et al.*, “Exemplar-based sparse representations for noise robust automatic speech recognition,” IEEE Trans. Audio, Speech and Language Processing, vol. 19, no. 7, pp. 2067–2080, 2011.

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

[15] 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.

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