適応型Restricted Boltzmann Machineを用いたパラレルデータフリーな任意話者声質変換*

◎中鹿亘，滝口哲也，有木康雄 (神戸大)

1 はじめに

音声信号処理の分野の中でも，声質変換技術（入力話者音声の音韻情報を保存したまま，話者性に関する情報のみを出力話者のものへ変換させる技術）が，様々なタスク [1, 2] への応用が可能であることから近年盛んに研究されている．これまでの声質変換法として，GMM (Gaussian Mixture Model) を用いた手法 [3] が最も広く用いられており，様々な改良がなされてきた [4, 5].

しかしながら，ほとんどの従来手法ではモデルの学習時にパラレルデータ（入力話者と出力話者の，同一発話内容による音声対）を必要とし，パラレルデータの作成には様々な制限が課せられる．第一に，発話データは同一の発話内容でないといけないという制限があるため，選択（または作成）できる学習データセットの自由度は低い．第二に，フレーム単位で入出力音声の同期を取る必要があるため，動的計画法などを用いてアライメントを取るが，完全にフレームの同期が取れている保証がない，伸縮の際に音声に変換が加わっているなどの問題がある．また，学習を行っていない話者対に対して，既存の変換モデルを利用できない．

入出力話者間のパラレルデータを必要としない，若しくは少量のパラレルデータを用いて，話者性を柔軟に制御するアプローチもいくつか提案されている [6, 7, 8, 9]．例えば文献 [6] では，参照話者のパラレルデータを用いて二話者間の関係性を GMM でモデル化しておき，入力話者（もしくは出力話者）を参照話者の特徴空間へ射影する行列を求めるため，入力話者-出力話者間のパラレルデータは必要としない（しかしながら，参照話者の間でパラレルデータを必要とする）．また，文献 [8] では，予め複数の話者によるパラレルデータを用いて固有声（Eigenvoice）を作成し，入力話者から固有声，固有声から出力話者へマッピングすることで多対多声質変換を実現している（このアプローチでも，固有声作成時に複数話者のパラレルデータを用意する必要がある）．

本研究では，確率モデルの一つである restricted Boltzmann machine (RBM) [10] を拡張したモデル（adaptive restricted Boltzmann machine; ARBM）を用いて，入力話者-出力話者間のパラレルデータだけではなく，参照話者間のパラレルデータさえも必要としない任意話者声質変換法を提案する．本研究で提案する適応型 RBM は，複数の話者が混在する音声データから，話者に依存しない情報と話者に依存した情報に分離しながら，潜在的な特徴を抽出する確率モデルである．このモデルは可視素子層と隠れ素子層からなる無向グラフで表現され，同層素子間の結合はなく，異層素子間のみ話者に依存した強度（重み）で結合が存在する．さらに，この重みは話者依存項と話者非依存項で表現され，複数の話者が混在した音声データ（パラレルである必要はない）を用いて，それぞれが教師なし学習で同時に推定される．結果として，話者依存重みと話者非依存重みに分離しながら潜在特徴（隠れ素子）を得ることができる．任意話者声質変換を行う際，まず，複数の話者（参照話者）のデータを用いて，上記のように話者依存重みと話者非依存重みを同時推定する．次に，入力話者と出力話者の少量データ（適応データ）を用いて，話者非依存重みを固定しながらそれぞれの話者依存重みを推定する．そして，変換したい音声から，入力話者の話者依存重み，話者非依存重みを用いて潜在特徴を推定し，その後，出力話者の話者依存重み，話者非依存重みを用いて音響特徴ベクトルを逆推定することで変換音声を得る．

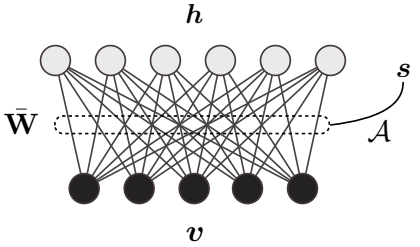

Fig. 1 Graphical representation of an ARBM.

2 適応型 RBM

本稿で定義する適応型 RBM は，Fig. 1 のように，従来の RBM [10] で見られた可視素子と隠れ素子だけでなく，識別素子 $\boldsymbol{s} = [s_1, \cdots, s_S]^T, s_k \in \{0, 1\}$ が加わったモデルである（S は識別素子の数とする）．例えば声質変換において，入力 $\boldsymbol{v}$ が話者 k の発話であることを示す場合，$s_k = 1, \forall s_{k'} = 0\ (k' \neq k)$ とな

*Parallel-dictionary-free voice conversion using adaptive restricted Boltzmann machine. by Toru NAKASHIKA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

る．このモデルでは，可視素子と隠れ素子の間には識別素子 $\boldsymbol{s}$ で制御される重みの結合が存在する．この結合重み $\mathbf{W}(\boldsymbol{s})$ を以下のように定義する．

$$\mathbf{W}(\boldsymbol{s}) = \mathcal{A} \otimes_3 \boldsymbol{s}\bar{\mathbf{W}} + \mathcal{B} \otimes_3 \boldsymbol{s} \tag{1}$$

ただし，$\mathcal{A}$ と $\mathcal{B}$ はいずれも，不特定重み行列 $\bar{\mathbf{W}} \in \mathbb{R}^{I\times J}$ を特定化（適応）するための 3 階のテンソルパラメータ（$\mathcal{A} \in \mathbb{R}^{I\times I\times S}, \mathcal{B} \in \mathbb{R}^{I\times J\times S}$）である．また，$\mathcal{X} \otimes_d \boldsymbol{y}$ はモード d を展開した 3 階テンソル $\mathcal{X}$ の各行列とベクトル $\boldsymbol{y}$ の内積をとる演算子を表す．声質変換の場合，$\bar{\mathbf{W}}$ が不特定話者による結合重み，$\mathbf{A}_{:,:,k}$ と $\mathbf{B}_{:,:,k}$ が話者 k の適応行列及びバイアス行列を表す（ただし $\mathbf{A}_{:,:,k}$ は 3 階テンソル $\mathcal{A}$ のモード 3 の第 k 行列を表す）．

適応型 RBM では，式 (1) で定義した $\mathbf{W}(\boldsymbol{s})$ を用いて，可視素子 $\boldsymbol{v}$，隠れ素子 $\boldsymbol{h}$，識別素子 $\boldsymbol{s}$ の同時確率 $p(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s})$ を以下のように定義する．

$$p(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) = \frac{1}{Z_{\mathrm{A}}} e^{-E_{\mathrm{A}}(\boldsymbol{v},\boldsymbol{h},\boldsymbol{s})} \tag{2}$$

$$E_{\mathrm{A}}(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) = \left\| \frac{\boldsymbol{v} - \boldsymbol{b}}{2\boldsymbol{\sigma}} \right\|^2 - \boldsymbol{c}^T\boldsymbol{h} - \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2}\right)^T \mathbf{W}(\boldsymbol{s})\boldsymbol{h} \tag{3}$$

$$Z_{\mathrm{A}} = \sum_{\boldsymbol{v},\boldsymbol{h},\boldsymbol{s}} e^{-E_{\mathrm{A}}(\boldsymbol{v},\boldsymbol{h},\boldsymbol{s})} \tag{4}$$

これらの定義により，条件付き確率 $p(\boldsymbol{h}|\boldsymbol{v}, \boldsymbol{s})$，$p(\boldsymbol{v}|\boldsymbol{h}, \boldsymbol{s})$ は以下のように計算できる．

$$p(h_j = 1|\boldsymbol{v}, \boldsymbol{s}) = \mathcal{S}(c_j + (\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2})^T \mathbf{W}(\boldsymbol{s})_{:j}) \tag{5}$$

$$p(v_i = v|\boldsymbol{h}, \boldsymbol{s}) = \mathcal{N}(v|b_i + \mathbf{W}(\boldsymbol{s})_{i:}\boldsymbol{h}, \sigma_i^2) \tag{6}$$

適応型 RBM のパラメータ $\Theta_{\mathrm{A}} = \{\bar{\mathbf{W}}, \mathcal{A}, \mathcal{B}, \boldsymbol{b}, \boldsymbol{\sigma}, \boldsymbol{c}\}$ は，N 個の学習データ $\{\boldsymbol{v}_n, \boldsymbol{s}_n\}_{n=1}^N$ を用いて，対数尤度 $\mathcal{L}_{\mathrm{A}} = \log \prod_n p(\boldsymbol{v}_n, \boldsymbol{s}_n) = \log \prod_n \sum_{\boldsymbol{h}} p(\boldsymbol{v}_n, \boldsymbol{h}_n, \boldsymbol{s}_n)$ を最大化するように推定される．この対数尤度を $\bar{\mathbf{W}}$，$\mathcal{A}$，$\mathcal{B}$ の要素（$\bar{W}_{ij}$，$A_{ii'k}$，B_{ijk}）で偏微分したものは，それぞれ

$$\frac{\partial \mathcal{L}_{\mathrm{A}}}{\partial \bar{W}_{ij}} = \langle \sum_{l,k} \frac{A_{lik} v_l h_j s_k}{\sigma_l^2} \rangle_{data} - \langle \sum_{l,k} \frac{A_{lik} v_l h_j s_k}{\sigma_l^2} \rangle_{model} \tag{7}$$

$$\frac{\partial \mathcal{L}_{\mathrm{A}}}{\partial A_{ii'k}} = \langle \sum_{m} \frac{W_{i'm} v_i h_m s_k}{\sigma_i^2} \rangle_{data} - \langle \sum_{m} \frac{W_{i'm} v_i h_m s_k}{\sigma_i^2} \rangle_{model} \tag{8}$$

$$\frac{\partial \mathcal{L}_{\mathrm{A}}}{\partial B_{ijk}} = \langle v_i h_j s_k \rangle_{data} - \langle v_i h_j s_k \rangle_{model}, \tag{9}$$

と計算できる．ただし，$\langle\cdot\rangle_{\mathrm{data}}$ と $\langle\cdot\rangle_{\mathrm{model}}$ はそれぞれ，観測データ，モデルデータの期待値を表す．他のパラメータ $\boldsymbol{b}$，$\boldsymbol{\sigma}$，$\boldsymbol{c}$ に関しては，従来の RBM の更新式 [10] と同様である．適応型 RBM においても CD（Contrastive Divergence）法を適用することができるため，各偏微分値の第二項 $\langle\cdot\rangle_{\mathrm{model}}$ を観測データの再構築値 $\langle\cdot\rangle_{\mathrm{recon}}$ として計算することで効率よくパラメータを推定することができる．

3 声質変換への応用

適応型 RBM を声質変換へ応用する場合，Fig. 2 のようにまず複数（S 人）の参照話者によるデータを用いて適応型 RBM の各パラメータを同時推定する (Step 1)．次に，$\bar{\mathbf{W}}$ など話者に依存しないパラメータを固定して，適応データを用いて入力話者と出力話者の適応パラメータ $\mathcal{A}' \ni \{\mathbf{A}_s = \mathbf{A}_{::(S+1)}, \mathbf{A}_t = \mathbf{A}_{::(S+2)}\}$，$\mathcal{B}' \ni \{\mathbf{B}_s = \mathbf{B}_{::(S+1)}, \mathbf{B}_t = \mathbf{B}_{::(S+2)}\}$ を式 (8)(9) より推定する (Step 2)．そして，入力話者の変換したい音声のフレーム音響特徴量 $\boldsymbol{v}_s$ から，次式のように潜在特徴量（隠れ素子）を推定する (Step 3)．

$$\hat{\boldsymbol{h}} = \underset{\boldsymbol{h}}{\mathrm{argmax}}\, p(\boldsymbol{h}|\boldsymbol{v}_s, \boldsymbol{s}_s) = \mathcal{S}(\boldsymbol{c} + (\frac{\boldsymbol{v}_s}{\boldsymbol{\sigma}^2})^{\mathrm{T}} \mathbf{W}(\boldsymbol{s}_s)) \tag{10}$$

ただし，$\boldsymbol{s}_s$ は第 $S+1$ 要素のみ 1，他を 0 とするベクトルとする．また，同時に変数 $\boldsymbol{s}$ の長さを $S+2$ へ拡張し，$\mathcal{A}, \mathcal{B}$ をモード 3 に沿ってそれぞれ $\mathcal{A}', \mathcal{B}'$ を追加するものとする．式 (10) を書き直すと，

$$\hat{\boldsymbol{h}} = \mathcal{S}(\boldsymbol{c} + (\frac{\boldsymbol{v}_s}{\boldsymbol{\sigma}^2})^{\mathrm{T}} (\mathbf{A}_s\bar{\mathbf{W}} + \mathbf{B}_s)) \tag{11}$$

が得られ，話者に依存しない項 $\bar{\mathbf{W}}$ を入力話者に適応させた結合重みを用いて潜在特徴量を推定していることになる．また式 (11) は，一度適応型 RBM の学習が終われば $\hat{\boldsymbol{h}}$ は変数 $\boldsymbol{v}$ の関数となるので，$\hat{\boldsymbol{h}}$ は話者に依存しない潜在特徴量であることを示唆している．すなわち，話者性は $\boldsymbol{s}$ のみで制御され，$\hat{\boldsymbol{h}}$ は話者に依存しない音韻に近い情報を表すと考えられる．したがって，出力話者の話者性を持つ音声を得たい場合，音韻情報 $\hat{\boldsymbol{h}}$ から，$\boldsymbol{s}_t$（第 S+2 要素のみ 1，他が 0 となるベクトル）を用いて音響特徴量を復元すればよい．すなわち，出力話者の変換先のフレーム特徴量 $\boldsymbol{v}_t$ を以下のように計算する (Step 4)．

$$\begin{aligned} \boldsymbol{v}_t &= \underset{\boldsymbol{v}}{\mathrm{argmax}}\, p(\boldsymbol{v}|\hat{\boldsymbol{h}}, \boldsymbol{s}_t) \\ &= \boldsymbol{b} + (\mathbf{A}_t\bar{\mathbf{W}} + \mathbf{B}_t)^{\mathrm{T}}\hat{\boldsymbol{h}} \end{aligned} \tag{12}$$

これは，入力話者音声から得られた音韻情報を基に，話者非依存項を出力話者に適応した基底を用いて，出力話者の音響特徴量を生成していることを表している．また，式 (11)(12) にもあるように，入力話者の音響特徴量 $\boldsymbol{v}_s$ を出力話者の音響特徴量 $\boldsymbol{v}_t$ へ変換す

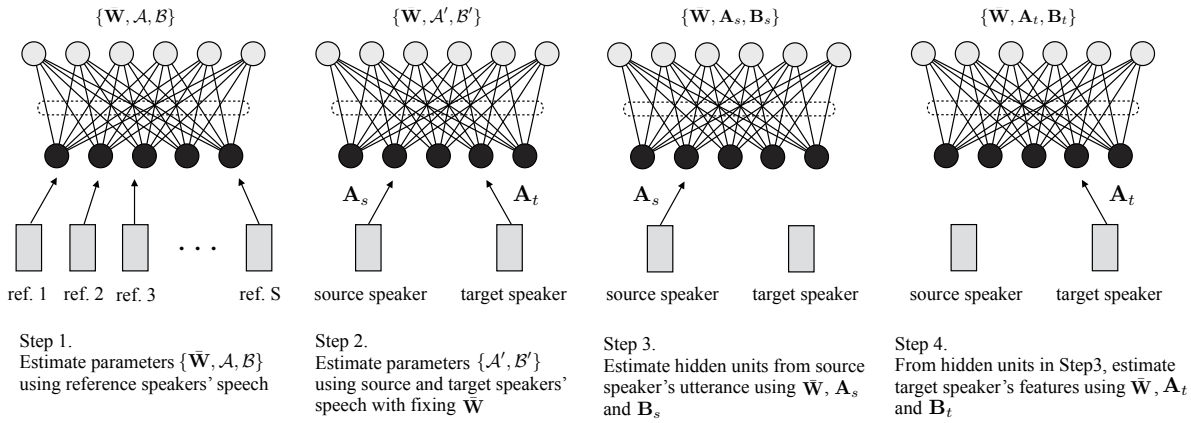

Fig. 2　Procedure of voice conversion using an ARBM.

Table 1　Performance of our method (SDIR [dB]).

# of hidden units	128	192	256	512
female-to-female	7.18	7.26	7.30	7.14
female-to-male	7.64	7.81	7.81	7.82
male-to-female	7.50	7.54	7.61	7.48
male-to-male	7.86	8.00	8.03	8.06
avg.	7.54	7.65	**7.69**	7.63

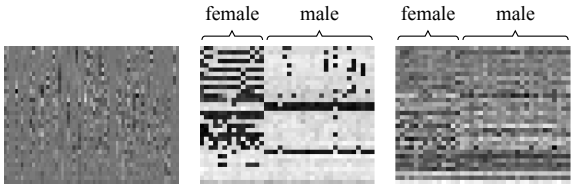

Fig. 3　Left to right: estimated $\bar{\mathbf{W}}$, $\mathcal{A}$, and $\mathcal{B}$.

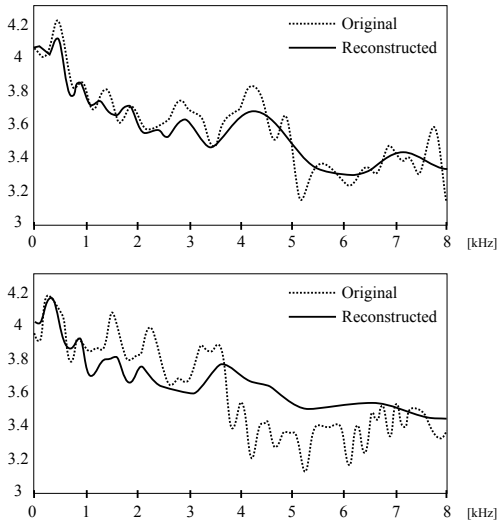

Fig. 4　Log-spectrum from a source speaker and the reconstructed spectrum (above), and log-spectrum from a target speaker and the converted spectrum from the source speech to the target speech (below).

る際，$\hat{\boldsymbol{h}}$ の推定に非線形関数を用いているため，提案法は非線形変換ベースの声質変換だと言える．

なお，現実の音声データを使って適応型 RBM を学習する場合，話者は豊富に存在するが，それぞれの発話データは少ないといったケースがある．この場合，$\bar{\mathbf{W}}$ の推定に用いられるデータは十分存在するが，適応パラメータ $\mathcal{A}, \mathcal{B}$ を推定するためのデータが少量となるため，誤推定もしくは過学習の要因となる．そこで本稿で述べる実験では，$\mathbf{A}_{::k}$ を対角行列，$\mathbf{B}_{::k}$ を各列が等しい行列で近似することでパラメータ数を抑える．

4　評価実験

4.1　実験条件

本実験では，英語圏の複数の話者による音声が含まれたコーパスである TIMIT[1] を用いて，提案手法である適応型 RBM を用いた声質変換の精度を調べた．このコーパスから，話者非依存パラメータの推定のために，参照話者として 38 名（内女性 14 名男性 24 名）を選んだ．各話者からは，5 文の発話データを学習に用いている（学習に用いた総フレーム数はおよそ 27 万）．提案手法を評価するために，女性 4 名，男性 4 名の音声を用いて入力話者・出力話者のペア（計 28 ペア）を作成し，異性間及び同性間の声質変換の性能比較を行った．このとき，入力・出力話者のパラレルデータ（同一発話内容による，学習データには含まれない 2 文のデータから動的計画法によって作成）を用いて SDIR (spectral distortion improvement ratio) による評価をおこなっている．音響特徴量として，STRAIGHT スペクトルから計算された 32 次元の MFCC を用いた．適応型 RBM における学習率，バッチサイズ，繰り返し回数はそれぞ

[1] https://catalog.ldc.upenn.edu/LDC93S1

れ 0.005，50，500 とした．隠れ素子数を 128，192，256，512 と変えて比較を行った．

4.2 実験結果と考察

提案手法による声質変換の結果を Table 1 に示す．例えば female-to-female では評価用の女性 4 名の音声を，それぞれ他の女性 3 名へ変換し，全フレームの SDIR の平均をとったものを表す．“avg.” は全組み合わせの平均値である．Table 1 から，一部を除いて隠れ素子数が増加すれば変換精度が向上していることが分かる．隠れ素子数が 512 と 256 の結果を比較すると，512 の場合は男性への変換（female-to-male，male-to-male）で優っているが，女性への変換（female-to-female，male-to-female）で精度が下がってしまい，結果として全平均の SDIR 値が低くなってしまっている．この理由として，パラメータ数の増加に伴い，モデルが過学習しているためだと考えられる（男性と女性の話者数は 14 対 28 であり，隠れ素子数 512 のモデルでは変換音声が男性側へ強く反応していることからも過学習が窺える）．

提案法によって，実際に推定されたパラメータ $\bar{\mathbf{W}}$，$\mathcal{A}$ および $\mathcal{B}$ の一部を Fig. 3 に示す．$\mathcal{A}$ に関しては，対角行列として近似した $\mathbf{A}_{::k}$ の対角成分を列ベクトルとして話者ごとに並べた行列を示しており（$\mathcal{B}$ も同様に話者ごとに並べた列ベクトルを示している），左 14 列ベクトルは女性話者，右 24 列ベクトルは男性話者に相当する．この図から分かるように，$\mathcal{A}$（または $\mathcal{B}$）の各々の列ベクトルは同性間で類似性が高く，異性間で類似性が低いベクトルとなっている．これは，音声を聴いて話者の違いを認識する際，個人の差異よりも性別の違いをより大きく感じ取っているという直感と一致する．

最後に，提案手法によって女性話者音声（コーパスでは FCJF0）を男性話者音声（MWAR0）へ変換した例を Fig. 4 に示す．この例では，FCJF0 のある時刻における対数スペクトル（図上段点線）から MFCC を計算し，FCJF0 の適応型 RBM によって $\hat{\boldsymbol{h}}$ を推定した後，MWAR0 の適応パラメータを用いて変換された音響特徴量を対数スペクトルへ復元した（図下段実線）．参考として，$\hat{\boldsymbol{h}}$ の推定後 FCJF0 の適応パラメータによって復元したスペクトル（図上段実線），目標となる MWAR0 のスペクトル（図下段点線）を載せている．この図より，FCJF0 の音声から FCJF0 の音声へ再構築したスペクトルのみならず，別の話者である MWAR0 へ変換した音声スペクトルにおいても，低域におけるスペクトルピークの周波数（フォルマント）がおおよそ目標と一致するなど，その話者の特徴を捉えていることが分かる．高周波数域に関してはいずれも目標と大きく異なっているが，MFCC からスペクトルを復元しているため，高域における情報が損失してしまうことに起因する．パラレルデータを学習時に一切使用せず，かつ FCJF0 から MWAR0 への変換モデルを学習していないにも関わらず FCJF0 から MWAR0 へ変換できていることは提案手法の大きな利点であると言える．

5 おわりに

本研究では，潜在的な特徴量を抽出する RBM を拡張して，話者に依存する項と依存しない項に分離してモデル化することで学習時にパラレルデータを必要としない任意話者声質変換手法を提案した．本研究で提案する RBM の拡張モデル（適応型 RBM）は声質変換のみならず，音声の感情付与や物体認識など，様々なタスクへの応用が考えられる．また，このモデルにおいて識別素子 $\boldsymbol{s}$ を推定することで，例えば話者認識へ応用することも可能であると考えられる．音韻情報と話者情報が混在した音声からそれぞれを分離し，話者性を制御できる点が適応型 RBM の強みであり，今後は適応型 RBM を用いた話者認識と音声認識の同時推定法について検討していきたい．

参考文献

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

[2] K. Nakamura et al., “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” Speech Commun., vol. 54, no. 1, pp. 134–146, 2012.

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

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

[5] D. Saito et al., “Application of matrix variate Gaussian mixture model to statistical voice conversion,” Interspeech, pp. 2504–2508, 2014.

[6] A. Mouchtaris et al., “Nonparallel training for voice conversion based on a parameter adaptation approach,” IEEE Trans. Audio, Speech, and Lang. Processs, vol. 14, no. 3, pp. 952–963, 2006.

[7] C. H. Lee and C. H. Wu, “Map-based adaptation for speech conversion using adaptation data selection and non-parallel training,” Interspeech, pp. 2254–2257, 2006.

[8] T. Toda et al., “Eigenvoice conversion based on Gaussian mixture model,” Interspeech, pp. 2446–2449, 2006.

[9] D. Saito et al., “One-to-many voice conversion based on tensor representation of speaker space,” Interspeech, pp. 653–656, 2011.

[10] K. Cho et al., “Improved learning of Gaussian-Bernoulli restricted Boltzmann machines,” ICANN, pp. 10–17, 2011.