話者適応を用いたNMFによる雑音環境下の声質変換*

☆藤井貴生，相原龍，中鹿亘，滝口哲也，有木康雄 (神戸大)

1 はじめに

声質変換は，入力された音声の言語情報を保ったまま，話者性や感情といった特定の情報のみを変換する技術である．応用例としては話者変換や感情変換 [1, 2] をはじめとし，発話支援 [3] など多岐に渡る．これまで様々な声質変換の手法が提案されており，中でも Gaussian Mixture Model (GMM) を用いた手法 [4] に代表されるような統計的アプローチに基づく手法 [5, 6] が広く用いられている．

戸田ら [7] 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性の高い音声として変換する手法を提案している．Helander ら [8] は Partial Least Squares (PLS) 回帰分析を用いることにより，従来手法における過適合の問題を回避するための手法を提案している．また従来手法では，入力話者と出力話者が同じテキストを発話して得られるパラレルデータが必要であるが，このパラレルデータを使用せずに声質変換を行うために，GMM の話者適応を行う手法 [9] や Eigen-Voice GMM (EV-GMM)[10, 11] などが提案されている．

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 [12]．スパース表現に基づくアプローチ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 [13, 14]．このアプローチでは，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．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 [15] 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．

本研究では，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)[16] を用いる．我々の提案している声質変換手法では，従来の声質変換手法でも用いられていたパラレルデータから，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．従来の声質変換のように統計的モデルを用いない Exemplar-based な手法であるため，過学習がおこりにくく，自然性の高い音声へと変換可能であると考えられる．

本稿では，話者適応を用いた 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，入力話者と出力話者の同一発話内容のパラレルデータを用いることが前提となっていた．つまり，対応する任意の話者の大量のデータをあらかじめ用意しておかなければならないという問題点があった．そこで，出力話者の少量の音声データのみを辞書適応に用いることで，入力話者辞書から出力話者辞書を生成する手法を提案する．評価実験では，雑音重畳音声に対して，提案手法の有効性を示す．

2 雑音環境下の声質変換

2.1 NMF による声質変換

スパースコーディングの考え方において，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．

$$\mathbf{x}_l \approx \sum_{j=1}^{J} \mathbf{b}_j h_{j,l} = \mathbf{B}\mathbf{h}_l \tag{1}$$

$\mathbf{x}_l$ 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．$\mathbf{b}_j \in \mathbb{R}_+^D$ は j 番目の学習サンプル，あるいは基底を表し，$h_{j,l} \in \mathbb{R}_+$ 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 $\mathbf{b}_j$ とする．基底を並べた行列 $\mathbf{B} = [\mathbf{b}_1 \dots \mathbf{b}_J]$ 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．

$$\mathbf{X} \approx \mathbf{B}\mathbf{H} \tag{2}$$

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

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．この手法では，パラレル辞書と呼ばれる入力話者音声辞書と出力話者音声辞書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動的計画法 (DP) を適用することでフレーム間の対応を取った後，入力

*Voice Conversion based on NMF using Speaker Adaptation in Noisy Environments, by Takao Fujii, Ryo Aihara, Toru Nakashika, Tetsuya Takiguchi, Yasuo Ariki (Kobe univ.)

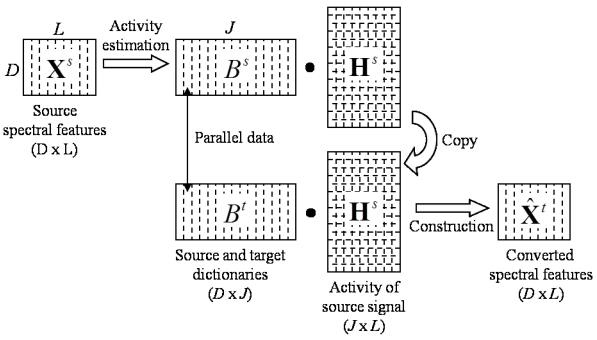

Fig. 1　NMF による声質変換の概要

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．

2.2　雑音重畳音声からのアクティビティ行列の推定

入力話者の辞書に付随する雑音辞書は，雑音の重畳した入力音声の非音声区間のフレームから構築される．NMF による雑音除去手法において，観測信号の l 番目のフレームは，クリーン音声から構築した辞書とノイズ辞書の非負の線形結合により近似される．

$$\begin{aligned}\mathbf{x}_l &= \mathbf{x}_l^s + \mathbf{x}_l^n \\ &\approx \sum_{j=1}^{J} \mathbf{b}_j^s h_{j,l}^s + \sum_{k=1}^{K} \mathbf{b}_k^n h_{k,l}^n \\ &= [\mathbf{B}^s \mathbf{B}^n] \begin{bmatrix} \mathbf{h}_l^s \\ \mathbf{h}_l^n \end{bmatrix} \quad s.t. \quad \mathbf{h}_l^s, \mathbf{h}_l^n \geq 0 \\ &= \mathbf{B}\mathbf{h}_l \quad s.t. \quad \mathbf{h}_l \geq 0 \end{aligned} \tag{4}$$

$\mathbf{x}_l^s$ と $\mathbf{x}_l^n$ はそれぞれ入力話者のクリーン音声の振幅スペクトル，雑音の振幅スペクトルを表す．$\mathbf{B}^s, \mathbf{B}^n, \mathbf{h}_l^s, \mathbf{h}_l^n$ は入力話者の辞書，雑音の辞書，そして l フレームにおけるそれぞれのアクティビティを表す．(4) 式を時間-周波数のスペクトログラムで表現すると，以下の通りになる．

$$\begin{aligned}\mathbf{X} &\approx [\mathbf{B}^s \mathbf{B}^n] \begin{bmatrix} \mathbf{H}^s \\ \mathbf{H}^n \end{bmatrix} \quad s.t. \quad \mathbf{H}^s, \mathbf{H}^n \geq 0 \\ &= \mathbf{B}\mathbf{H} \quad s.t. \quad \mathbf{H} \geq 0. \end{aligned} \tag{5}$$

本手法ではスペクトルの形状のみを考慮するため，まず $\mathbf{X}$，$\mathbf{B}^s$ 及び $\mathbf{B}^n$ について，フレーム毎，あるいは辞書内のサンプル毎に，各周波数ビンの振幅の総和で正規化を行う．クリーン音声と雑音のアクティビティが並んだ行列 $\mathbf{H}$ はスパース制約付き NMF[15] により推定される．

$$\begin{aligned}\mathbf{M} &= \mathbf{1}^{(D\times D)}\mathbf{X} \\ \mathbf{X} &\leftarrow \mathbf{X}./\mathbf{M} \\ \mathbf{B} &\leftarrow \mathbf{B}./(\mathbf{1}^{(D\times D)}\mathbf{B}) \end{aligned} \tag{6}$$

$\mathbf{1}$ は全ての要素が 1 の行列である．スパース制約付き NMF において $\mathbf{H}$ を推定するためにコスト関数が設定されている．コスト関数の式と $\mathbf{H}$ の更新式は以下のようになる．

$$d(\mathbf{X}, \mathbf{BH}) + ||(\lambda \mathbf{1}^{(1\times L)}) .* \mathbf{H}||_1 \quad s.t. \quad \mathbf{H} \geq 0. \tag{7}$$

$$\begin{aligned}\mathbf{H}_{n+1} &= \mathbf{H}_n .* (\mathbf{B}^T(\mathbf{X}./(\mathbf{BH}))) \\ &\quad ./(\mathbf{1}^{((J+K)\times L)} + \lambda \mathbf{1}^{(1\times L)}). \end{aligned} \tag{8}$$

(7) 式を最小にするように $\mathbf{H}$ が推定される．第一項は $\mathbf{X}$ と $\mathbf{BH}$ の Kullback-Leibler divergence である．第二項は $\mathbf{H}$ をスパースにするための L1 ノルム正則化項である．スパース制約の重みは $\lambda^T = [\lambda_1 \ldots \lambda_J \ldots \lambda_{J+K}]$ を調節することで，辞書内のサンプル毎に定義することができる．本稿ではクリーン音声辞書に関する制約重み $[\lambda_1 \ldots \lambda_J]$ を 0.2 に，雑音辞書に関する制約重み $[\lambda_{J+1} \ldots \lambda_{J+K}]$ を 0 に設定した．

3　話者適応を用いた声質変換

我々がこれまで提案してきた NMF による声質変換法では，入力話者と出力話者の同一発話内容のパラレルデータを用いることが前提となっていた．つまり，対応する話者の大量の音声データをあらかじめ用意しておかなければならないという問題点があった．本稿では出力話者の少量の音声データを辞書適応に用いることで，入力話者の辞書から出力話者の辞書を作成する手法を提案する．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．適応データである出力話者の STRAIGHT[17] スペクトル $\mathbf{X}^t$ は，適応行列 $\mathbf{A}$，入力話者辞書 $\mathbf{B}^s$ 及び入力信号のアクティビティ行列 $\mathbf{H}^s$ の線形結合によって表現される．

$$\mathbf{X}^t \approx \mathbf{A}\mathbf{B}^s\mathbf{H}^s \tag{9}$$

ここで，入力話者辞書 $\mathbf{B}^s$ は入力話者の音声から抽出した STRAIGHT スペクトルを並べたものである．アクティビティ行列は入力話者辞書からどの基底を選択するかという指標になる行列であるので，本稿では適応に用いる出力話者と同じ発話内容である入力話者音声から推定した重み行列を用いている．ここで得られた適応行列 $\mathbf{A}$ と入力話者辞書 $\mathbf{B}^s$ の線形結合によって出力話者辞書 $\hat{\mathbf{B}}^t$ が生成される．

$$\hat{\mathbf{B}}^t = \mathbf{A}\mathbf{B}^s \tag{10}$$

適応行列 A は，下記評価関数に基づき求められる．

$$\mathbf{A} = \arg\min_{\mathbf{A}} D(\mathbf{X}^t | \mathbf{A}\mathbf{B}^s\mathbf{H}^s) \tag{11}$$

本論文における D は Kullback-Leibler divergence を表す．また，適応行列 $\mathbf{A}$ は，文献 [18] で音源分離における基底の適応行列の推定として提案されている方法と同様にして，以下の式により推定される．

$$\begin{aligned}\mathbf{A} &\leftarrow \mathbf{A} .* (\mathbf{X}^t./(\mathbf{A}(\mathbf{B}^s\mathbf{H}^s)))(\mathbf{B}^s\mathbf{H}^s)^T \\ &\quad ./\mathbf{1}(\mathbf{B}^s\mathbf{H}^s)^T \end{aligned} \tag{12}$$

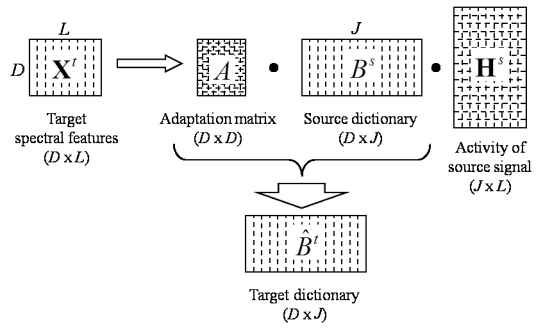

Fig. 2　話者適応によるパラレル辞書作成

4　評価実験

4.1　実験条件

本実験ではテストデータの入力音声に雑音重畳音声を用いた．従来の GMM を用いた手法と，話者適応を用いない従来の NMF による変換手法と比較を行った．ATR 研究用日本語音声データベースから，男性話者と女性話者それぞれ 2 名ずつの音声データを使用した．入力話者と出力話者の組み合わせは男性から女性，男性 1 から男性 2 及び女性 1 から女性 2 の 3 パターンを用意した．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は入力話者と出力話者の同一発話内容の 10 単語，25 単語，50 単語，216 単語からそれぞれ作成したパラレルデータを用いた．また提案手法では，216 単語で構築した入力話者辞書に対して，出力話者音声のうち 10 単語，25 単語，50 単語を用いて話者適応を行い，それぞれ出力話者辞書を作成した．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ことでパラレルデータとして用いた．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ラムで，次元数は 40 である．GMM の混合数は 64 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 単語に雑音信号を加算したものを用いた．雑音信号は CENSREC-1-C データベースにて食堂内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．雑音辞書は評価音声毎に発話の前後区間から構築しており，雑音辞書に含まれるサンプル数は平均 104 である．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を，出力音声の生成及び出力話者辞書の構築には 512 次元の STRAIGHT スペクトルを用いた．これは，本稿の入力音声には雑音が重畳しており，音声信号の分析合成ツールである STRAIGHT[17] ではそ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．アクティビティ行列の推定の更新回数は 300 回とした．

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．512 次元の STRAIGHT スペクトルを特徴量とし，式 (13) で表される Normalized Spectrum Distortion(NSD)[19] によって各手法との比較を行った．

$$NSD = \sqrt{\frac{||S^Y - S^{\hat{X}}||^2}{||S^Y - S^X||^2}} \tag{13}$$

ただし，S^X，S^Y，$S^{\hat{X}}$ 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

提案手法による変換と 2 つの比較手法による変換によって出力された 50 単語の音声それぞれに対して算出した NSD を Fig. 3，Fig. 4，Fig. 5 に示す．図より，本提案である話者適応を用いた NMF 変換における NSD が全体的に小さくなっていることが分かる．これは，話者適応を用いた手法では目標話者辞書は入力話者辞書を元に作られているが，話者性の部分が大きく適応により変換されたと考えられる．

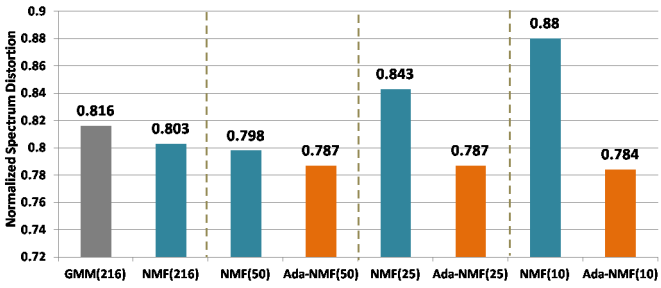

Fig. 3　男性から女性への変換時の NSD

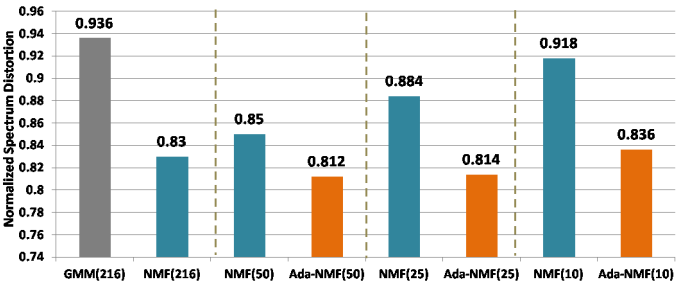

Fig. 4　男性から男性への変換時の NSD

5　おわりに

本稿では，雑音重畳音声に対して，入力話者辞書から推定されたアクティビティ行列と出力話者辞書の内積から得られたスペクトルから音声を再合成する NMF を用いた声質変換を行った．実験結果より，話者適応を用いることで入力話者と出力話者それぞれの同一発話内容の音声から作成する少量のパラレルデータのみから声質変換を行う本手法の有効性が示

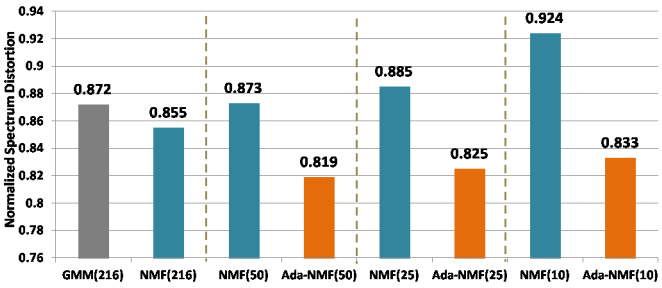

Fig. 5　女性から女性への変換時の NSD

された．今後は NMF を用いた変換手法における辞書の構築において，音素ごとのクラスタリングを行うことによって，より入力系列を表現するのに適した基底を選択できる手法の検討を進めていく．

参考文献

[1] Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, “GMM-based Voice Conversion Applied to Emotional Speech Synthesis,” IEEE Trans. Seech 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, T. Toda, H. Saruwatari, and K. Shikano, “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” Speech Communication, Vol. 54, No. 1, pp. 134–146, 2012.

[4] Y. Stylianou, O. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech and Audio Processing, Vol. 6, No. 2, pp. 131–142, 1998.

[5] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” in Proc. ICASSP, pp. 655–658, 1988.

[6] H. Valbret, E. Moulines and J. P. Tubach, “Voice transformation using PSOLA technique,” Speech Communication, Vol. 11, No. 2-3, pp. 175–187, 1992.

[7] T. Toda, A. Black, and K. Tokuda, “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.

[8] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, “Voice conversion using partial least squares regression,” IEEE Trans. Audo, Speech, Lang. Process., Vol. 18, No. 5, pp. 912–921, 2010.

[9] C. H. Lee and C. H. Wu, “MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training,” in Proc. INTERSPEECH, pp. 2254–2257, 2006.

[10] T. Toda, Y. Ohtani, and K. Shikano, “Eigenvoice conversion based on Gaussian mixture model,” in Proc. INTERSPEECH, pp. 2446–2449, 2006.

[11] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, “One-to-many voice conversion based on tensor representation of speaker space,” in Proc. INTERSPEECH, pp. 653–656, 2011.

[12] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, “Exemplar-Based Voice Conversion in Noisy Environment,” IEEE Workshop on Spoken Language Technology (SLT2012), pp. 313-317, 2012.

[13] 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.

[14] M. N. Schmidt and R. K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” in Proc. INTERSPEECH, pp. 2614-2617, 2006.

[15] J. F. Gemmeke, T. Viratnen, and A. Hurmalainen, “Exemplar-Based Sparse Representations for Noise Robust Automatic Speech Recognition,” IEEE Trans. Audio, Speech, Lang. Process., Vol. 19, Issue 7, pp. 2067–2080, 2011.

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

[17] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, Vol.27, pp. 187–207, 1999.

[18] Emad M. Grais, and Hakan Erdogan, “Adaptation of speaker-specific bases in non-negative matrix factorization for single channel speech-music separation,” in Proc. INTERSPEECH, pp. 569–572, 2011.

[19] T. En-Najjary, O. Rosec, and T. Chonavel, “A voice conversion method based on joint pitch and spectral envelope transformation,” in Proc. ICSLP, pp. 199-203, 2004.