雑音環境下における非負値行列因子分解を用いた声質変換
Voice Conversion based on Non-negative Matrix Factarization in Noisy Environments

○藤井貴生，相原龍，高島遼一，滝口哲也，有木康雄 (神戸大)
Takao Fujii, Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki (Kobe univ.)

Abstract This paper presents a voice conversion technique for noisy environments. We prepared parallel exemplars that consist of the source and target exemplars, which have the same texts uttered by the source and target speakers. The input source signal is decomposed into the source exemplars, noise exemplars obtained from the input signal, and their weights. Then, the converted signal is obtained by calculating the linear combination of the target exemplars and the weights which are calculated using the source exemplars. In the proposed method, a GMM-based conversion method is also applied to the feature vectors generated by the sparse coding in order to compensate a mismatch between the weights of source and target exemplars. The effectiveness of this method was confirmed by comparing its effectiveness with that of a conventional method.

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．Helnder 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．

しかし，これらの声質変換の従来手法のほとんどは学習・テストデータともにクリーン音声を用いることが前提となっており，雑音の重畳した入力音声に関する評価はされていない．入力音声に重畳した雑音は変換音声を生成する際の妨げとなり，その結果として変換される音声にも悪い影響が出ることは避けられない．よって雑音環境下を考慮した声質変換の手法の検討が必要であると言える．

近年，信号処理の分野において Sparse Coding によるアプローチが注目されており，音声信号処理の分野では Non-negative Matrix Factorization (非負値行列因子分解，NMF) [12] が音源分離や雑音抑圧などに特に用いられている [13, 14]．Sparse Coding によるアプローチでは，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．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，クリーン音声から構築する辞書とノイズ辞書のスパースな表現にし，クリーン音声の辞書に対する重みを音声認識における Hidden Markov Model (HMM) の尤度として用いることで，雑音にロバストな音声認識を行う手法を提案している．

本稿では，雑音環境下に強い Sparse Coding による声質変換の手法を提案する．ここでは入力話者と出力話者それぞれの同一発話内容の音声の特徴量をサンプルとするパラレル辞書を構築する．更に，入力音声の発話前後の非音声区間から雑音辞書を構築し，入力として与えられる雑音重畳音声を入力音声辞書と雑音辞書のスパースな表現にする．この入力音声と辞書から推定される重み行列のうち，音声辞書に関する重みのみを取り出し，出力話者の音声サンプルから構築した出力音声辞書との線形結合をとる．更に本手法では，より出力話者への音声へと近似させるため，ここで得られる特徴量に対して GMM 変換を適用することで出力話者の変換音声とする．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．

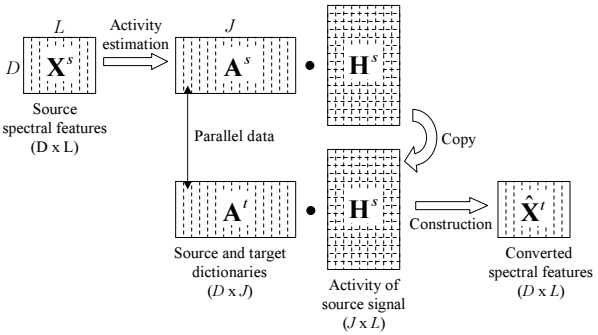

図 1: Sparse Coding による声質変換の概要

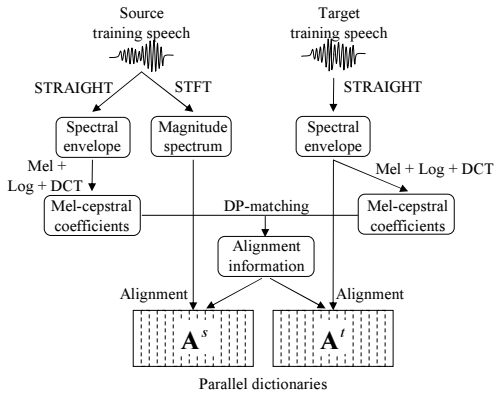

図 2: パラレル辞書の構築

2 提案手法

2.1 Sparse Coding による声質変換

本章では Sparse Coding による声質変換について述べる [16]．本手法では入力音声の特徴量は学習サンプルと重み行列の線形結合で表現される．

$$\mathbf{x}_l \approx \sum_{j=1}^{J} \mathbf{a}_j h_{j,l} = \mathbf{A}\mathbf{h}_l \qquad (1)$$

$\mathbf{x}_l$ は入力音声の特徴量の l 番目のフレームを表す．$\mathbf{a}_j$ 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．本手法においては $\mathbf{a}_j$ は学習サンプルを表す．学習サンプルを並べた行列 $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_J]$ は本稿では “辞書” と呼び，重みを並べたベクトル $\mathbf{h}_l = [h_{1,l} \dots h_{J,l}]^T$ を “アクティビティ” と呼ぶことにする．

本手法では入力話者の音声を出力話者の音声へと変換するため，パラレルな入力話者辞書と出力話者辞書を用意する．入力話者と出力話者の同一発話内容の音声の特徴量を動的計画法 (DP) によってフレーム間同期を取ったパラレルデータを入力話者，出力話者それぞれのサンプルとして辞書を構築する．

図 1 に Sparse Coding による声質変換の概要を示す．入力音声と入力話者の辞書からアクティビティ行列を推定する．このアクティビティ行列は辞書内のサンプルに対する重みを表す行列である．入力話者の辞書と出力話者の辞書がパラレルとなっており，入力話者の辞書と推定されたアクティビティ行列の内積で入力音声を表現できるので，そのアクティビティ行列と出力話者の辞書によって出力話者の音声が表現できる．

2.2 パラレル辞書の構築

パラレル辞書の構築には入力話者音声と出力話者音声それぞれの特徴量が必要となる．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．従って，入力話者音声から構築する辞書内のサンプルは短時間フーリエ変換 (STFT) によって計算される振幅スペクトルとし，出力話者音声の辞書に関しては STRAIGHT 分析によって得られるスペクトルをサンプルとする．

図 2 にパラレル辞書の構築手順を示す．学習データとなる入力話者音声と出力話者音声は同一発話内容のものである．入力話者の学習データとして，辞書内のサンプルには STFT によって計算される振幅スペクトルを用いる．出力話者に関しては STRAIGHT 分析によって得られるスペクトルを辞書のサンプルとして用い，学習データとする．入力話者，出力話者ともに STRAIGHT 分析によって得られるメルケプストラムを用いて，フレーム間同期を取るための DP マッチングを行い，パラレルデータを作成する．

声質変換を行う際には，入力音声に対して STFT と STRAIGHT 分析の両方を行う．入力話者の辞書内のサンプルは STFT によって得られた振幅スペクトルであるため，入力音声の振幅スペクトルからノイズ辞書の構築とアクティビティ行列の推定が行われる．変換音声の合成の際に基本周波数と非周期成分は入力音声のものを用いるため，STRAIGHT でこれらを分析する．

2.3 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．Sparse Coding 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．

$$\begin{aligned}\mathbf{x}_l &= \mathbf{x}_l^s + \mathbf{x}_l^n \\ &\approx \sum_{j=1}^{J} \mathbf{a}_j^s h_{j,l}^s + \sum_{k=1}^{K} \mathbf{a}_k^n h_{k,l}^n \\ &= [\mathbf{A}^s \mathbf{A}^n] \begin{bmatrix} \mathbf{h}_l^s \\ \mathbf{h}_l^n \end{bmatrix} \quad s.t. \quad \mathbf{h}_l^s, \mathbf{h}_l^n \geq 0 \\ &= \mathbf{A}\mathbf{h}_l \quad s.t. \quad \mathbf{h}_l \geq 0 \end{aligned} \tag{2}$$

$\mathbf{x}_l^s$ 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．$\mathbf{A}^s$, $\mathbf{A}^n$, $\mathbf{h}_l^s$, $\mathbf{h}_l^n$ 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．(2) 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．

$$\begin{aligned}\mathbf{X} &\approx [\mathbf{A}^s \mathbf{A}^n] \begin{bmatrix} \mathbf{H}^s \\ \mathbf{H}^n \end{bmatrix} \quad s.t. \quad \mathbf{H}^s, \mathbf{H}^n \geq 0 \\ &= \mathbf{A}\mathbf{H} \quad s.t. \quad \mathbf{H} \geq 0. \end{aligned} \tag{3}$$

本手法ではスペクトルの形状のみを考慮するため，まず $\mathbf{X}$，$\mathbf{A}^s$ 及び $\mathbf{A}^n$ 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．

$$\begin{aligned}\mathbf{M} &= \mathbf{1}^{(D\times D)}\mathbf{X} \\ \mathbf{X} &\leftarrow \mathbf{X}./\mathbf{M} \\ \mathbf{A} &\leftarrow \mathbf{A}./(\mathbf{1}^{(D\times D)}\mathbf{A}) \end{aligned} \tag{4}$$

$\mathbf{1}$ 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．

$$d(\mathbf{X}, \mathbf{A}\mathbf{H}) + ||(\lambda \mathbf{1}^{(1\times L)}) .* \mathbf{H}||_1 \quad s.t. \quad \mathbf{H} \geq 0. \tag{5}$$

$$\begin{aligned}\mathbf{H}_{n+1} = \mathbf{H}_n .* (\mathbf{A}^T(\mathbf{X}./(\mathbf{A}\mathbf{H}))) \\ ./(\mathbf{1}^{((J+K)\times L)} + \lambda \mathbf{1}^{(1\times L)}).\end{aligned} \tag{6}$$

(5) 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 $\mathbf{A}\mathbf{H}$ 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．

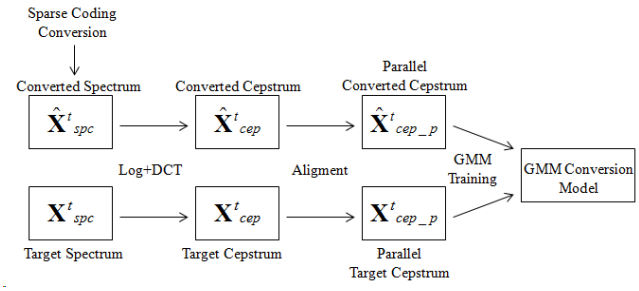

図 3: 変換後の特徴量を用いた GMM 学習の流れ

2.4 Sparse Coding による音声特徴量の生成

推定されたアクティビティ行列 $\mathbf{H}$ から，入力話者辞書に関するアクティビティ$\mathbf{H}^s$ のみを取り出し，これと出力話者の辞書を用いることで，ノイズ除去されたクリーンなスペクトルを得る．このとき，出力話者の辞書も入力話者の辞書と同様に，振幅の総和で正規化しておく．

$$\mathbf{A}^t \leftarrow \mathbf{A}^t./(\mathbf{1}^{(D\times D)}\mathbf{A}^t) \tag{7}$$

次に，正規化された出力話者辞書と $\mathbf{H}^s$ の内積を取り，(4) であらかじめ計算しておいた入力音声の振幅をかけることで，Sparse Coding 変換後のスペクトルを得る．

$$\hat{\mathbf{X}}^t = (\mathbf{A}^t\mathbf{H}^s) .* \mathbf{M} \tag{8}$$

入力とする特徴量と入力話者の辞書のサンプルには振幅スペクトルが用いられているが，出力話者の辞書は STRAIGHT スペクトルをサンプルとして構築されているため，上式により得られる Sparse Coding 変換後のスペクトルは STRAIGHT スペクトルにより表現される．

2.5 GMM を用いた補正

本稿では変換音声のスペクトルをより出力話者のものへと近似させるために，Sparse Coding 変換によって生成されたスペクトルに対して更に，GMM に基づく変換手法を適用する．GMM の学習データには Sparse Coding により声質変換を行ったスペクトルと，出力話者音声のスペクトルをそれぞれケプストラムへと変換し，パラレルデータとする．学習は従来手法である GMM に基づく声質変換と同様の手順で行われる [4]．変換の際には辞書とアクティビティの内積から生成されるスペクトルからケプストラムを算出する．その後 GMM 変換を適用し，より出力話者の声質に近い音声特徴量を得る．

2.6 変換音声の生成

Sparse Coding 変換後に GMM に基づく変換を加えて生成されたケプストラムに対して逆変換を行うことで

STRAIGHT スペクトルを再生成する．これは STRAIGHT 合成ツールにより変換音声を合成するためである．本稿では，音声合成に必要である基本周波数は従来の単回帰分析により変換を行い，非周期成分は入力音声から抽出されたものを直接用いている．

3 評価実験

3.1 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 NMF での変換後に GMM 変換を加えない手法を比較手法として実験を行った．ATR 研究用日本語音声データベースから，入力話者音声は男性話者，出力話者音声は女性話者とした．サンプリング周波数は 8kHz とした．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 216 単語から作成したパラレルデータを用いた．各話者のパラレル辞書内のサンプル数は 57,033 である．

比較手法である GMM による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る入力話者音声の 50 単語と，辞書内に含まれない 25 文を用いた．テストデータとなる単語，文章のセットそれぞれに雑音信号を加算した．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の数は平均 104 である．テスト時の入力音声および入力話者のパラレル辞書の構築には 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のパラレル辞書の構築には 512 次元の STRAIGHT スペクトルを用いた．アクティビティ行列の推定の更新回数は 500 回とした．

提案手法における GMM 変換の学習には，216 単語の雑音重畳音声に対して Sparse Coding 変換を行ったものを用いた．

3.2 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，25 文章の音声それぞれに対してケプストラム分析を行い，それらと目標音声となる出力話者の発話した同一発話とのケプストラム距離 (CEP) を図 4，図 5 に示す．

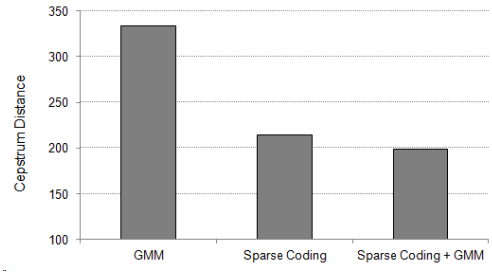

図 4: テストデータ 50 単語における目標音声と声質変換音声のケプストラム距離

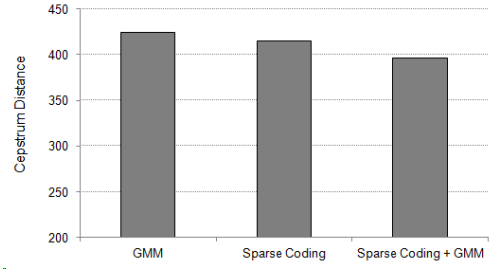

図 5: テストデータ 25 文章における目標音声と声質変換音声のケプストラム距離

図より，NMF による変換後に更に GMM 変換を加えた本手法が最も目標音声に近くなっていることが分かる．単語を用いたテストにおいて，GMM のみを用いた手法により変換された音声が最も目標音声との差が大きくなった．これはモデル構築時に用いた学習データがクリーン音声であるため，入力音声に重畳した雑音によって誤ったパラメータ変換が行われてしまったためであると考えられる．文を用いたテストでは，GMM のみを用いた手法と Sparse Coding を用いた 2 つの手法との差が小さくなっている．これは，GMM に基づく手法が統計的アプローチであるため，学習に用いていない音声にも対応できるということを示している．

4 考察と今後の課題

本稿では，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したパラレルデータから各話者の辞書を構築し，雑音重畳音声に対して，入力話者の辞書から推定されたアクティビティ行列と出力話者の辞書の内積から得られたスペクトルに対して更に，目標となる出力話者の音声のスペクトルへの GMM を用いた変換を行った．実験結果より，Sparse Coding により得られるスペ

クトルに GMM 変換を加えた場合が最も目標音声に近くなっていることから，本提案が従来手法より有効であることが示された．

今後はセグメント特徴を導入する動的変化を考慮した変換手法を検討する．また，本手法においては出力話者の辞書に対するアクティビティ行列は入力音声と入力話者の辞書から推定されたものを用いているが，出力話者の音声と出力話者の辞書から推定されるアクティビティ行列とは相違が生じる可能性がある．よって，入力から推定されるアクティビティ行列に対して変換を行ったものを出力話者の辞書との内積に用いる手法についても今後検討していく．

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

[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] 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.

[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.