スパース表現を用いた雑音環境下の声質変換*

☆高島遼一，滝口哲也，有木康雄 (神戸大)

1 はじめに

声質変換は，入力した音声を音韻情報などは保ったまま，話者性のような特定の情報のみを変換する技術であり，話者変換や感情変換 [1, 2]，発話支援 [3] など様々なタスクへの応用が期待されている．これまで声質変換のための統計的手法が多く提案されているが [4, 5, 6]，中でも Gaussian Mixture Model (GMM) を用いた手法 [6] が広く用いられており，多くの改良がされ続けている．

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な音声として変換する手法を提案している．Helnder ら [8] 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．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．

しかしながら，現在提案されている声質変換の手法はクリーン音声を用いて評価が行われており，雑音環境下を考慮した定式化はされていない．もし入力音声に雑音が重畳していた場合，その雑音は出力音声にも重畳するだけでなく，雑音によって誤った特徴変換が行われ，声質変換の性能自体も下がってしまうと考えられる．従って，雑音の影響を考慮した声質変換の手法が重要であると言える．

近年，信号処理の分野においてスパース表現に基づく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]．スパース表現のアプローチでは，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．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，スパース表現に基づく雑音重畳音声のための声質変換の手法を提案する．本手法では，従来の GMM ベースの手法において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と出力話者の音声辞書からなる同一発話内容のパラレル辞書を構築する．変換の際は，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，入力音声を入力音声辞書と雑音辞書のスパース表現にする．このときに得られる重み（アクティビティ）行列のうち，入力音声辞書に関する重みのみを取り出し，この重み係数に基づいて出力音声辞書内のサンプルを線形結合することで，出力話者の音声スペクトルへと変換する．実験では雑音重畳音声を用いて従来の GMM を用いた手法と比較を行い，提案手法の優位性を示す．

2 提案手法

2.1 スパース表現に基づく声質変換

スパース表現のアプローチでは，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．

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

$\mathbf{x}_l$ は観測信号の 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．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学習サンプルのみで表現されることになる．

本手法では，入力音声を出力音声へ変換するために，パラレル辞書と呼ばれる入力音声辞書と出力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に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，入力話者と出力話者の学習サンプルをそれぞれ辞書化したものである．

図 1 は入力話者と出力話者がそれぞれ “ikioi” と発話した音声に対して，それらの音声のフレーム間対応を取ったものを辞書にし，NMF によりアクティビティ行列を推定したものである．このとき，入力話者の音声には入力話者の辞書を，出力話者の音声には出力話者の辞書を用いて，それぞれのアクティビティ行列を求めている．また，入力/出力音声の特徴量及び辞書内のサンプルは STRAIGHT 分析 [16] によって得られる平滑化スペクトル（STRAIGHT スペクトル）である．この実験では入力/出力音声と辞書が同じ単語であるため，得られるアクティビティ行列は対角線上に高いエネルギーを持つ．また，図中の赤線で

*Exemplar-Based Voice Conversion in Noisy Environments. by Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki (Kobe univ.)

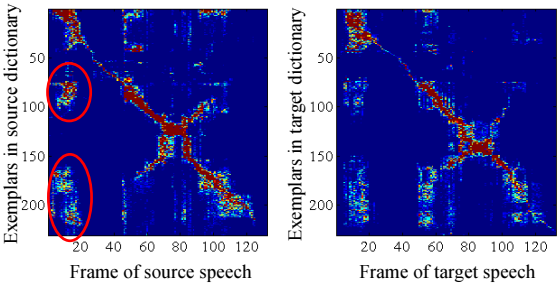

Fig. 1　入力信号のアクティビティ行列（左図）と出力信号のアクティビティ行列

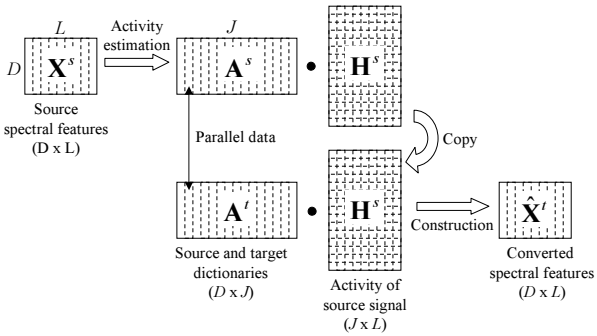

Fig. 2　スパース表現に基づく声質変換の概要

囲まれた箇所のような，対角線上から離れた場所にエネルギーが現れているのは，単語中に ‘i’ が 3 回含まれているため，辞書内の ‘i’ に対応する区間も 3 箇所存在するためである．

図より，入力話者音声と出力話者音声それぞれから得られるアクティビティ行列は，似た位置に高いエネルギーを持っていることが分かる．この理由から，入力話者の辞書と出力話者の辞書が同一発話でフレーム対応が取れている，つまりパラレル辞書になっているとき，入力音声から得られるアクティビティ行列は，出力音声のアクティビティ行列に代用可能であると考えられる．この考えに基づくと，図 2 のように入力音声のアクティビティ行列と出力音声辞書の内積から，出力音声を生成することが可能となる．図中の D, L, J はそれぞれ次元数，フレーム数，辞書内のサンプルの数を表す．

2.2　パラレル辞書の構築

前節の予備実験では簡単の為，辞書内のサンプルは入力辞書と出力辞書のどちらも STRAIGHT スペクトルにより表現されていた．実際の声質変換の実験においても，入力音声がクリーン音声であれば，両辞書を STRAIGHT スペクトルで表現しても問題なく動作した．しかし，入力音声に雑音重畳音声を考慮した場合，音声信号の分析合成ツールである STRAIGHT では，雑音信号を上手く表現できないという問題がある．そのため，出力信号の辞書のみ STRAIGHT スペクトルで表現しておき，入力信号及びその辞書は短時間フーリエ変換 (STFT) によって得られる通常の振幅スペクトルを用いて表現する．

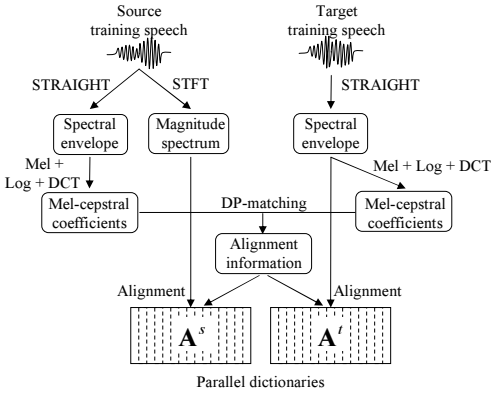

Fig. 3　パラレル辞書の構築

図 3 にパラレル辞書の構築手順を示す．入力話者と出力話者の教師データは同一発話の音声である．出力話者の教師データは STRAIGHT 分析を行い，STRAIGHT スペクトルを辞書内のサンプルとして用いる．さらに STRAIGHT スペクトルからメルケプストラムを計算しておき，これをフレーム同期を取る際の DP マッチングに用いる．一方入力話者の教師データにおいては，辞書内の各サンプルは STFT によって得られる振幅スペクトルにより表現されており，STRAIGHT 分析は DP マッチングを行う際のメルケプストラムの計算のみに用いられる．

声質変換の際には，入力音声には STFT と STRAIGHT 分析の両方が行われる．その後，振幅スペクトルはノイズ辞書の構築とアクティビティ行列の推定に用いる．STRAIGHT スペクトルは変換には用いないが，STRAIGHT 分析によって得られる F0 と非周期成分は変換音声の合成に用いる．

2.3　雑音重畳音声からのアクティビティ行列の推定

声質変換を行う際には，雑音の重畳した入力音声の発話前後の非音声区間のフレームを取り出して雑音辞書を構築する．スパース表現に基づく雑音除去手法において，観測信号の l 番目のフレームは，クリーン音声辞書と雑音辞書の非負の線形結合により近似される．

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

$\mathbf{x}_l^s$ 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．

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

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．

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

$\mathbf{1}$ は全ての要素が 1 の行列である．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．スパース制約付き NMF では以下のコスト関数を最小とするように $\mathbf{H}$ を推定する．

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

第一項は $\mathbf{X}$ と $\mathbf{AH}$ 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 0.1 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．(5) 式を最小とする $\mathbf{H}$ は，以下の更新ルールを繰り返し行うことにより，推定される．

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

2.4 変換音声の生成

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，変換音声の生成を行う．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，変換音声のスペクトルが生成される．

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

入力音声及びその辞書のスペクトル情報は振幅スペクトルで表現されているが，出力話者の辞書は STRAIGHT スペクトルにより構築されているため，上式により得られる変換音声のスペクトルは STRAIGHT スペクトルにより表現される．従って，得られた STRAIGHT スペクトルから STRAIGHT 合成ツールにより変換音声を合成することが可能となる．本稿では，このとき 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．

Table 1　変換前の入力音声と各手法における変換音声のメルケプストラム歪み

	Original source	Conventional	Proposed
Mel-CD [dB]	6.70	4.74	3.97

3 評価実験

3.1 変換音声の生成

本実験では雑音重畳音声を用いて従来の GMM を用いた手法 [6] と比較を行った．ATR 研究用日本語音声データベースより，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 8kHz である．

216 単語からパラレルデータを作成し，本手法におけるパラレル辞書の構築，従来法における GMM の学習に用いた．各話者の辞書に含まれるサンプルの数は 57,033 である．評価用の雑音重畳入力音声は，クリーン音声 25 文に雑音信号を加算することで作成した．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使用した．平均 SNR は 24dB である．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．

入力話者の辞書の構築，及び入力音声に用いる振幅スペクトルの次元数は 256，出力話者の辞書の構築，及び出力音声の生成に用いる STRAIGHT スペクトルの次元数は 512 である．アクティビティ行列の推定値の更新回数は 500 とした．従来手法では STRAIGHT スペクトルから計算される線形ケプストラム 40 次元を特徴量として用いた．

3.2 実験結果

変換前の入力音声と，各手法における変換音声それぞれのメルケプストラム歪みを表 1 に示す．表より，提案手法が従来手法よりもケプストラム歪みが低く，より出力話者の音声に近くなるような変換が行えていることが分かる．

次に，変換された音声の自然性，話者性に関して，7 人の被験者による主観評価実験を行った．自然性の評価については，各手法によって変換された音声を聴き比べて，どちらが自然な音声に聞こえるかを選択する（一対比較法）．話者性の評価については，出力話者の音声を聞いた後，各変換音声を聴き比べて，どちらが出力話者の音声に似ているかを選択する（XAB 法）．

図 4 にそれぞれの主観評価結果を示す．縦軸はそ

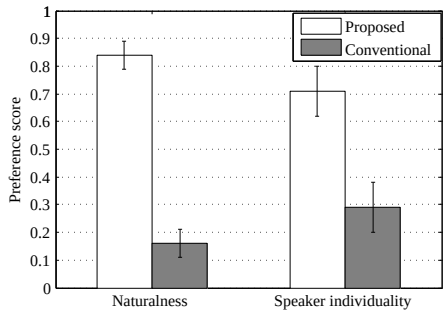

Fig. 4　自然性及び話者性に関する主観評価結果

れぞれの手法が被験者に選択された割合を表し，エラーバーは 95% 信頼区間を表す．どちらの評価基準に関しても提案手法の方が高いスコアが得られており，特に自然性において両手法の差が出ている．これは，従来手法ではノイズによって誤った変換が行われたことで，自然性が損なわれたためであると考えられる．

4　おわりに

本稿では，入出力話者のパラレルデータから構築したパラレル辞書と入力音声から構築した雑音辞書を用いて，雑音が重畳した入力音声を入力話者辞書と雑音辞書のスパース表現にし，入力話者辞書のアクティビティ行列に基づいて出力話者辞書内のサンプルを線形結合することで，出力話者の音声へ変換する手法を提案した．雑音重畳音声を用いて従来の GMM を用いた手法と比較実験を行った結果，特に自然性において提案手法が優位であることが示された．

今後は低 SNR の雑音環境での評価を行い，またセグメント特徴を導入することで動的変化を考慮した変換手法についてもを検討を行う．本手法はアクティビティ行列を推定するため，計算コストが増大になってしまうという問題がある．そのため，より少ないサイズの辞書で変換を行う手法を検討する必要がある．さらに，本手法はパラレルデータを用いるため，1 対 1 の話者変換にしか適用できない．そのため，少量のパラレルデータで行える手法についても検討する．

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

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

[6] 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.

[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] 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.