# ハイスピードカメラ画像を用いたマルチモーダルNMF声質変換*

☆真坂健太 相原龍, 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．音韻情報を維持しつつ話者情報を変換する “話者変換” [1] 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 [2] に用いられている他，感情情報を変換する “感情変換” [3, 4]，失われた話者情報を復元する “発話支援” [5] など多岐にわたって応用されている．本研究では，雑音環境下での声質変換など，これまでになかったタスクに対応可能な非負値行列因子分解（Non-negative Matrix Factorization : NMF）による声質変換 [6] を扱う．我々はこれまで，唇画像特徴量を用いた声質変換法を提案してきた [7]．本稿ではハイスピードカメラ画像を用いることで，唇のより微細な動きを正確に捉える特徴量を抽出する．その特徴量をもとに，一般的なカメラで撮られた画像から得られた特徴量を用いた変換より精度の良い変換を目指した．

従来，声質変換においては統計的な手法が多く提案されてきた．なかでも混合正規分布モデル (Gaussian Mixture Model : GMM) を用いた手法 [1] はその精度のよさと汎用性から広く用いられており，多くの改良がされ続けられている．戸田ら [8] 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．

GMM を含む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．

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．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．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．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 [9] 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 NMF [10] 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．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．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．

また，音声だけでなくその他のセンサーも用いたマルチモーダルな手法が認識・変換においてよりよい結果をもたらすと言われている．駒井ら [11] は，雑音環境下において AAM を用いた発話認識手法を提案している．音声情報のみを用いた結果よりも画像情報を取り入れたことでその有効性が示されている．Bateson ら [12] は顔にモーションセンサーを付け，特徴量を取り出し顔モデルを作成する手法を提案している．四倉ら [13] らはハイスピードカメラを用いて，顔の筋肉が動く順番から表情合成する手法を提案している．

本手法では，入力話者の唇画像特徴から得られた唇画像辞書を導入することで変換精度をより向上させる．更に本手法ではハイスピードカメラ画像から得られた画像特徴量を用いることで，唇のより微細な動きを捉え，一般的なカメラ画像を用いた変換よりも精度の良い変換を目指した．

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に加

*Multimodal Voice Conversion Based on Non-negative Matrix Factorization Using High Speed Image. by Kenta Masaka, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

え，一般カメラ画像を用いた変換とハイスピードカメラ画像を用いた変換を比較した．第 5 章で本稿をまとめる．

## 2 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．

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

$\mathbf{x}_l$ 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．$\mathbf{a}_j$ 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．本手法では学習サンプルそのものを基底 $\mathbf{a}_j$ とする．基底を並べた行列 $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_J]$ 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．

$$\mathbf{X} \approx \mathbf{A}\mathbf{H} \quad (2)$$

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

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．

このとき，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仮定できる．このことから，辞書行列がパラレルであれば，入力話者の辞書行列を用いて推定された入力特徴量のアクティビティは出力特徴量のアクティビティとして置き換え可能であると考えられる．以上の仮定に基づき，入力音声は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．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．

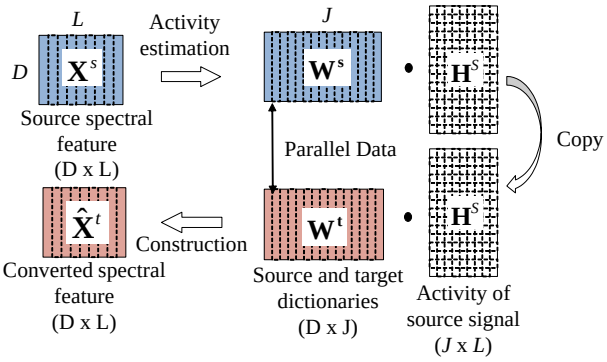


Fig. 1 Basic approach of NMF-based voice conversion

## 3 唇画像特徴を用いた NMF 声質変換

これまでの NMF による声質変換法では，音声特徴のみを用いた変換手法となっていた．本稿では，唇画像特徴を組み込んだマルチモーダルな声質変換手法となっている．更に本手法に用いる画像特徴量をハイスピードカメラ画像から得ることで，唇の微細な動きを捉えることができる．

### 3.1 辞書構成法

Fig. 2 は音声辞書，画像辞書の構成法を示したものである．従来の NMF による声質変換と同様にして各話者の同一発話によるパラレルデータから入力話者と出力話者の音声辞書 $\mathbf{W}^{s,A}$，$\mathbf{W}^{t,A}$ をそれぞれ求める．本稿のテストデータ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 ではその雑音を表現するのが難し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．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．画像辞書 $\mathbf{W}^{s,V}$ の構築には，動画から抽出したフレーム画像を用いる．それらのフレーム画像から得られら特徴量を並べたものを画像辞書とする．画像の特徴量として DCT（Discrete Cosine Transform）を用いる．DCT された画像からジグザグスキャンを行い，低次 100 次元を負値を取らないように底上げしたものを画像特徴量とする．この画像辞書と音声辞書を結合したものを音声画像結合辞書 $\mathbf{W}^s$ とし，変換に用いるものとする．

### 3.2 変換手法

入力話者の辞書に付随する雑音辞書は，雑音の重畳したテストデータの非音声区間のフレームから構

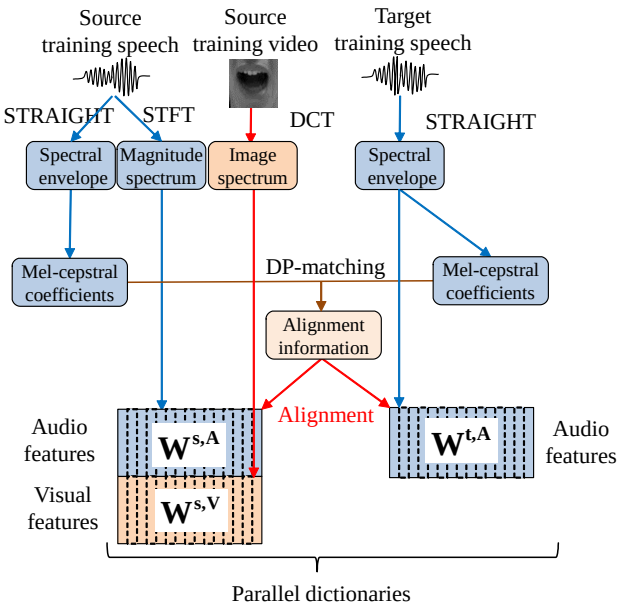


Fig. 2　Multimodal dictionary construction

築される．NMF による雑音除去手法において，観測信号のあるフレームは，クリーン信号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．

$$\begin{aligned}\mathbf{x} &= \mathbf{x}^s + \mathbf{x}^n \\ &\approx \sum_{j=1}^{J} \mathbf{w}_j^s h_j^{av} + \sum_{k=1}^{K} \mathbf{w}_k^n h_k^n \\ &= [\mathbf{W}^s\ \mathbf{N}] \begin{bmatrix} \mathbf{h}^{av} \\ \mathbf{h}^n \end{bmatrix} \quad s.t. \quad \mathbf{h}^{av}, \mathbf{h}^n \geq 0 \\ &= \mathbf{W}\mathbf{h} \quad s.t. \quad \mathbf{h} \geq 0 \end{aligned} \tag{4}$$

$\mathbf{x}^s$ と $\mathbf{x}^n$ はそれぞれ入力話者のクリーン信号，雑音信号を表す．$\mathbf{x}^s$ は入力音声 $\mathbf{x}^{s,A}$ と入力画像 $\mathbf{x}^{s,V}$ の結合ベクトルとなっている．$\mathbf{W}^s$, $\mathbf{N}$, $\mathbf{h}^{av}$, $\mathbf{h}^n$ は入力話者の辞書，雑音の辞書，それぞれに対するアクティビティを表す．$\mathbf{W}^s$, $\mathbf{N}$ はそれぞれ，音声辞書 $\mathbf{W}^{s,A}$ と画像辞書 $\mathbf{W}^{s,V}$，音声雑音辞書 $\mathbf{N}^A$ と画像雑音辞書 $\mathbf{N}^V$ の結合行列となっている．入力信号，辞書はすべてフレーム毎に正規化されているものとする．本手法では入力信号，辞書に含まれる音声と画像に対して，重み $\alpha$ と $\beta$ を導入する．この重みは SNR に応じて調整することで，最適なアクティビティを推定することができる．このとき，音声と画像を結合した特徴量から得られる $\mathbf{h}^{av}$ は以下のコスト関数を最小にすることで推定される．

$$\begin{aligned}&\alpha d(\mathbf{x}^{s,A}, [\mathbf{W}^{s,A}\ \mathbf{N}^A]\mathbf{h}) + \beta d(\mathbf{x}^{s,V}, [\mathbf{W}^{s,V}\ \mathbf{N}^V]\mathbf{h}) \\ &+ ||\lambda . * \mathbf{h}||_1 \quad s.t. \quad \mathbf{h} \geq 0 \end{aligned} \tag{5}$$

第 1 項と 2 項はそれぞれ音声側，画像側の Kullback-Leibler (KL) divergence である．第 3 項は $\mathbf{h}^{av}$ をスパースにするための L1 ノルム正規化項である．$\lambda^T = [\lambda_1 \ldots \lambda_J \ldots \lambda_{J+K}]$ 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画像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) 式を最小にするように以下の更新式に従いアクティビティ行列 $\mathbf{h}^{av}$ が推定される．

$$h_j \leftarrow h_j \frac{\sum_d^D \mathbf{f}_d + \sum_e^E \mathbf{g}_e}{\alpha + \beta + \lambda_j} \tag{6}$$

$$\mathbf{f}_d = \alpha W_{d,j}^{s,A} \alpha x_d^A / (\alpha \mathbf{W}^{s,A} \mathbf{h}^{av})_d \tag{7}$$

$$\mathbf{g}_e = \beta W_{e,j}^{s,V} \beta x_e^V / (\beta \mathbf{W}^{s,V} \mathbf{h}^{av})_e \tag{8}$$

$D$ と $E$ はそれぞれ音声と画像特徴量の次元数を表す．推定されたアクティビティ行列と出力話者辞書 $\mathbf{W}^{t,A}$ の内積を取り，NMF 変換後のスペクトル $\hat{\mathbf{x}}^t$ を得る．

$$\hat{\mathbf{x}}^t = \mathbf{W}^{t,A}\mathbf{h}^{av} \tag{9}$$

## 4　評価実験

### 4.1　実験条件

本実験では従来の GMM を用いた手法と，音声特徴量のみを用いた NMF に基づく手法に加え，一般カメラを用いた画像特徴量とハイスピードカメラを用いた画像特徴量を用いた変換手法を比較した．入力話者として被験者 1 名から収録した音声を入力話者音声とした．ATR 研究用日本語音声データベースより，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，フレームシフトは 1ms とした．音声辞書の構築には入力話者と出力話者の同一発話の 20 単語から作成したパラレルデータを用いた．GMM の学習に用いるパラレルデータとして，辞書構築時に使用した同一発話から得られたケプストラム 24 次元を特徴量とした．テストデータとなる単語それぞれに雑音信号を加算した．雑音信号は CENSREC-1-C データベースに含まれる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 SNR は 0 dB とした．今回の実験において，ハイスピードカメラと一般カメラを用いて収録を行った．実験に用いたカメラはそれぞれ MEMRECAM GX-1, SONY HDR-CX590 となっている．フレームレートはそれぞれ 4,000fps，30fps である．変換に用いたハイスピードカメラ画像は音声側 (1000fps) に合わせるために間引いた後，特徴量を取得した．一般カメラ画像は音声側に合わせるために特徴量を取得した後にスプライン補間を用いた．また，画像と音声の重みについては音声に対する重み $\alpha$ は 1 に固定し，画像に対する重み $\beta$ は最適なものを選びアクティビティを推定した．

### 4.2 実験結果・考察

各手法における変換音声の SDIR（Spectral Distortion Improvement Ratio）を Fig. 3 と Fig. 4 に示す．SDIR は以下の式で表される．

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

ただし $X^s$,$X^t$,$\hat{X}^t$ 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．スペクトルはそれぞれ次元ごとに正規化されている．Fig. 3,Fig. 4 はそれぞれハイスピードカメラ，一般カメラと同時に収録した音声に対して評価している．Fig. 3,Fig. 4 より，一般カメラ画像を用いた変換より，ハイスピード画像を用いた変換のほうが精度向上率が高いことがわかる．これはハイスピード画像のほうが子音などの微細な動きを捉えられているからだと考えられる．

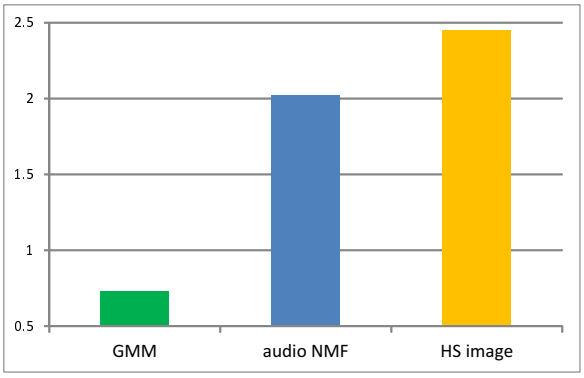


Fig. 3　Spectral Distortion Improvement Ratio calculated from converted speech (high speed)

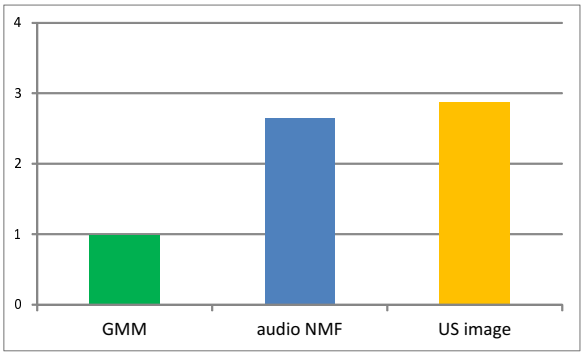


Fig. 4　Spectral Distortion Improvement Ratio calculated from converted speech (usual speed)

## 5　おわりに

本稿では，これまで提案してきた NMF に基づく声質変換法において，ハイスピードカメラ画像から得られた特徴量を導入した．これにより，一般カメラ画像から得られた特徴量を補間したものを用いるより，高速に撮影した画像を用いたほうが，より正確な変換が行えることがわかった．今後はハイスピードカメラで撮った画像をもとに，顔の部位によって微細な変動がないか考察していく必要がある．

## 参考文献


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

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

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

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

[5] 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.

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

[7] K. Masaka *et al.*, “Multimodal voice conversion using non-negative matrix factorization in noisy environments,” in ICASSP, 2014.

[8] 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.

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

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

[11] Y. Komai *et al.*, “Robust AAM-based audio-visual speech recognition against face direction changes,” ACM Multimedia, pp. 1161–1164, 2012.

[12] E. Bateson *et al.*, “The dynamics of audiovisual behavior in speech,” Speechreading by Humans and Machines, 1996.

[13] 四倉達夫, “高速度カメラによる動的な顔表情の分析および合成,” 電子情報通信学会, pp. 7–12, 2002.