スパースパラレル学習を用いたマルチモーダル声質変換*

☆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]，失われた話者情報を復元する “発話支援” [4] など多岐にわたって応用されている．我々はこれまで，唇画像特徴量を用いたマルチモーダルな声質変換法を提案し，その有効性を示してきた [5]．本稿ではスパースパラレル学習を用いたマルチモーダル声質変換手法を提案し，従来手法に比べて精度の高い変換を目指した．

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 [6] 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．しかし，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．

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 [7] 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．

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．この入力音声と辞書から推定される重み行列と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ことにより変換音声を得る．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．さらにこの手法は雑音環境下においても有効である．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．推定された重み行列のうち，クリーン部分のみを取り出し出力話者の音声辞書と掛け合わせることで，雑音を除去することができる．

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クリーン環境下，雑音環境下ともに認識・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 [8] 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 [9] 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 [10] 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．

これまでの NMF によるマルチモーダル声質変換法 [5] では，非負値制約のために画像特徴量が負値にならないように底上げをしなければならなかった．また，膨大な基底数から成る辞書行列を用いて変換しなければならず，画像特徴量の冗長な情報までもが辞書行列に含まれていた．本手法は，これらの問題点を解消するために，マルチモーダルなスパースパラレル学習を提案する．画像特徴量においては，非負値制約を取り除くことで特徴量の底上げを不要にした．さらに，パラレル制約を加えた辞書学習を行うことで，コンパクトな辞書を推定し変換精度の向上を目指した．

以下，第 2 章でこれまでのスパース表現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 NMF による声質変換手法に加え，マルチモーダル声質変換と本手法を比較し評価する．第 5 章で本稿をまとめる．

---

*Multimodal Voice Conversion using Sparse-Parallel Training. by Kenta Masaka, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

2　スパース表現を用いた声質変換

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．

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

$\mathbf{x}_l$ 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．$\mathbf{a}_j$ 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．

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

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

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．

3　スパースパラレル学習を用いた声質変換

これまでの NMF によるマルチモーダル声質変換法では，非負値制約のために画像特徴量が負値にならないように底上げをしなければならなかった．さらに学習データ全てを用いて辞書行列を構築し変換するため，画像特徴量の冗長な情報までもが辞書行列に含まれていた．そこで，本手法は画像特徴量の非負値制約

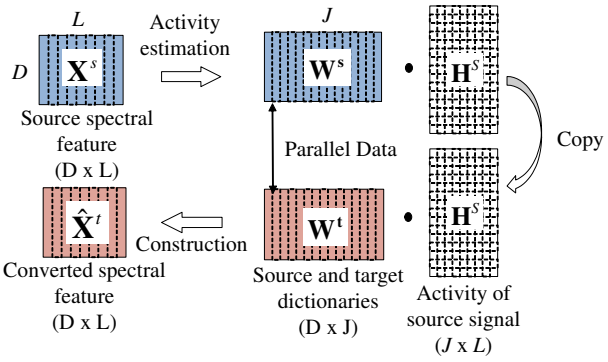

Fig. 1　Basic approach of NMF-based voice conversion

を取り除き，DCT などの負値を含む特徴量を用いた変換手法となっている．これにより，画像特徴量を抽出したそのままの状態で変換に用いることができる．また，辞書学習を用いることにより少量の基底数で変換を行えるため，画像特徴量の冗長性が解消され，変換精度が向上すると期待できる．

3.1　学習用パラレルデータ

Fig. 2 は学習用パラレルデータの構成法を示したものである．学習用データとして，入力･出力話者の同一発話によるパラレルデータを用いることにする．入力話者音声の特徴量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を，出力話者音声に関しては STRAIGHT 分析によって得られるスペクトルを用いる．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．入力話者の画像特徴量として 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 200 次元を画像特徴量とする．こうして得られたパラレルデータ $\mathbf{V}^{sa}$, $\mathbf{V}^{sv}$, $\mathbf{V}^{ta}$ を学習用入力データとして用いる．

3.2　スパースパラレル学習

Fig. 3 にスパースパラレル学習の概要を示す．$\mathbf{V}^{sa}$, $\mathbf{V}^{sv}$, $\mathbf{V}^{ta}$ はそれぞれパラレルな入力話者の学習用音声と学習用画像，出力話者の学習用音声となっている．$\mathbf{W}^{sa}$, $\mathbf{W}^{sv}$, $\mathbf{W}^{ta}$ はそれぞれ推定する入力話者の音声辞書と画像辞書，出力話者の音声辞書となっている．$\mathbf{H}^s$, $\mathbf{H}^t$ は入力話者，出力話者それぞれに対するアクティビティを表す．D^{sa} と D^{ta} はそれぞれ入力音声と出力音声の次元数を表す．学習用のデータはすべてフレーム毎に正規化されているものとする．

各話者のパラレルデータ $\mathbf{V}^{sa}$, $\mathbf{V}^{sv}$, $\mathbf{V}^{ta}$ をもとに，以下のコスト関数を最小にするように $\mathbf{W}^{sa}$, $\mathbf{W}^{sv}$,

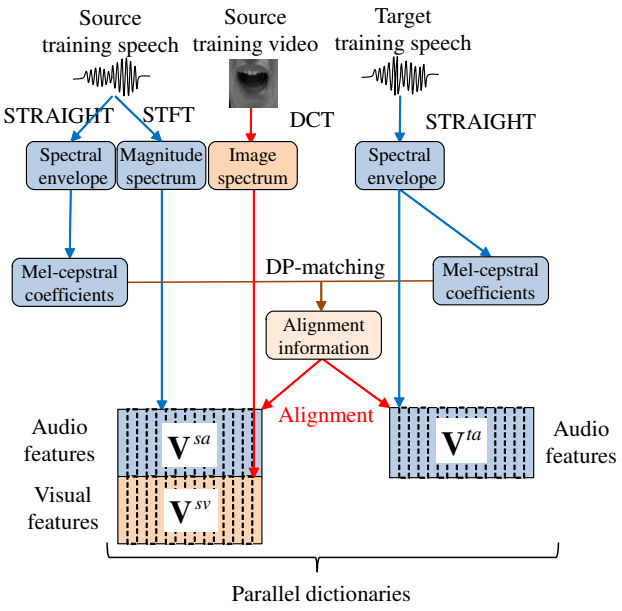

Fig. 2　Multimodal dictionary construction

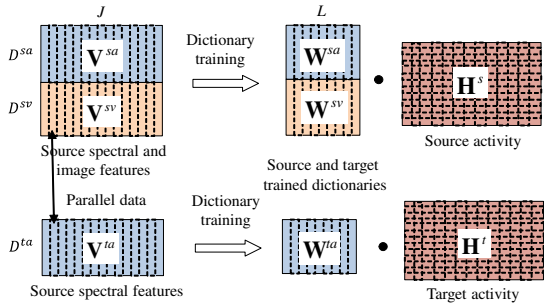

Fig. 3　Flow of sparse-parallel training

$\mathbf{W}^{ta}$ を求める．

$$\begin{aligned}\min\quad & d_{KL}(\mathbf{V}^{sa},\mathbf{W}^{sa}\mathbf{H}^{s})\\ & +(\psi/2)d_{F}(\mathbf{V}^{sv},\mathbf{W}^{sv}\mathbf{H}^{s})\\ & +d_{KL}(\mathbf{V}^{ta},\mathbf{W}^{ta}\mathbf{H}^{t})+(\epsilon/2)||\mathbf{H}^{s}-\mathbf{H}^{t}||_{F}\\ & +\lambda||\mathbf{H}^{s}||_{1}+\lambda||\mathbf{H}^{t}||_{1}\\ s.t\quad & \mathbf{W}^{sa}\geq 0,\mathbf{H}^{s}\geq 0,\mathbf{W}^{ta}\geq 0,\mathbf{H}^{t}\geq 0\end{aligned}\tag{4}$$

d_{KL} は Kullback-Leibler (KL) divergence，d_F Frobenius norm を表す．ψ は画像特徴量の重みを調整するパラメータである．第 4 項は $\mathbf{H}^s$ と $\mathbf{H}^t$ をパラレルにするためのパラレル制約項であり，ϵ によって調整される．第 5 項と第 6 項は $\mathbf{H}^s$, $\mathbf{H}^t$ をそれぞれスパースにするための L1 ノルム正規化項である．

(4) 式を最小にするように更新するための辞書行列の更新式は，補助関数法を用いて求められる．

以下の更新式に従い繰り返し更新することで，それぞれの学習辞書行列が求められる．

$$\begin{aligned}\mathbf{W}^{sa} &\leftarrow \mathbf{W}^{sa}.*((\mathbf{V}^{sa}./(\mathbf{W}^{sa}\mathbf{H}^{s}))\mathbf{H}^{s^{\mathrm{T}}})\\ &\quad ./(\mathbf{1}^{(D^{sa}\times L)}\mathbf{H}^{s^{\mathrm{T}}}) &(5)\\ \mathbf{W}^{sv} &\leftarrow (\mathbf{V}^{sv}\mathbf{H}^{s^{\mathrm{T}}})/(\mathbf{H}^{s}\mathbf{H}^{s^{\mathrm{T}}}) &(6)\\ \mathbf{W}^{ta} &\leftarrow \mathbf{W}^{ta}.*((\mathbf{V}^{ta}./(\mathbf{W}^{ta}\mathbf{H}^{t}))\mathbf{H}^{t^{\mathrm{T}}})\\ &\quad ./(\mathbf{1}^{(D^{ta}\times L)}\mathbf{H}^{t^{\mathrm{T}}}) &(7)\end{aligned}$$

$\mathbf{H}^s$, $\mathbf{H}^t$ についても同様に以下の更新式に従い繰り返し更新することで求められる．

$$\begin{aligned}\mathbf{H}^{s} &\leftarrow (-\mathbf{Q}+\sqrt{\mathbf{Q}^{2}+4(\mathbf{P}.*\mathbf{R})})./(2\mathbf{P}) &(8)\\ \mathbf{P} &= \psi((\mathbf{W}\mathbf{W}^{+}\mathbf{H}^{s})./\mathbf{H}^{s})+\epsilon &(9)\\ \mathbf{Q} &= \mathbf{W}^{sa^{\mathrm{T}}}\mathbf{1}^{(D^{sa}\times J)}-\psi\mathbf{W}^{sv^{\mathrm{T}}}\mathbf{V}^{sv}\\ &\quad -\epsilon\mathbf{H}^{t}+\lambda &(10)\\ \mathbf{R} &= \mathbf{H}^{s}.*(\mathbf{W}^{sa^{\mathrm{T}}}(\mathbf{V}^{sa}./(\mathbf{W}^{sa}\mathbf{H}^{s})))\\ &\quad +\psi((\mathbf{W}\mathbf{W}^{-}\mathbf{H}^{s}).*\mathbf{H}^{s}) &(11)\end{aligned}$$

$$\begin{aligned}\mathbf{H}^{t} &\leftarrow (1/(2\epsilon))(-\mathbf{B}+\sqrt{\mathbf{B}^{2}+4\epsilon\mathbf{S}}) &(12)\\ \mathbf{S} &= \mathbf{H}^{t}.*(\mathbf{W}^{ta^{\mathrm{T}}}(\mathbf{V}^{ta}./(\mathbf{W}^{ta}\mathbf{H}^{t}))) &(13)\\ \mathbf{B} &= -\epsilon\mathbf{H}^{s}+\mathbf{W}^{ta^{\mathrm{T}}}\mathbf{1}^{(D^{ta}\times J)}+\lambda &(14)\end{aligned}$$

ここで正部と負部を分けるため，$\mathbf{X}^{+}=(|\mathbf{X}|+\mathbf{X})./2$, $\mathbf{X}^{-}=(|\mathbf{X}|-\mathbf{X})./2$ のように定める．また，$\mathbf{W}\mathbf{W}^{+}=\mathbf{W}^{+^{\mathrm{T}}}\mathbf{W}^{+}$，$\mathbf{W}\mathbf{W}^{-}=\mathbf{W}^{-^{\mathrm{T}}}\mathbf{W}^{-}$ である．

こうして学習された辞書 $\mathbf{W}^{sa}$, $\mathbf{W}^{sv}$, $\mathbf{W}^{ta}$ をそれぞれ変換の辞書行列として用いることにする．

3.3　変換方法

$\mathbf{V}^{sa}$, $\mathbf{V}^{sv}$, をそれぞれ変換用入力音声，入力画像とする．$\mathbf{W}^{sa}$, $\mathbf{W}^{sv}$, $\mathbf{W}^{ta}$ はそれぞれ学習した入力音声辞書，入力画像辞書，出力音声辞書である．変換に用いられるアクティビティ行列 $\hat{\mathbf{H}}^s$ は以下のコスト関数を最小にすることで推定される．

$$\begin{aligned}&d_{KL}(\mathbf{V}^{sa},\mathbf{W}^{sa}\hat{\mathbf{H}}^{s})+(\psi/2)d_{F}(\mathbf{V}^{sv},\mathbf{W}^{sv}\hat{\mathbf{H}}^{s})\\ &\qquad +\lambda||\hat{\mathbf{H}}^{s}||_{1}\end{aligned}\tag{15}$$

推定されたアクティビティ行列と (4) 式で推定した出力話者辞書 $\mathbf{W}^{ta}$ の内積を取り，変換後のスペクトル $\hat{\mathbf{V}}^{ta}$ を得る．

$$\hat{\mathbf{V}}^{ta}=\mathbf{W}^{ta}\hat{\mathbf{H}}^{s}\tag{16}$$

4　評価実験

4.1　実験条件

本実験では クリーン環境下で従来の音声特徴量のみを用いた NMF に基づく手法，マルチモーダル声質

変換手法とスパースパラレル学習を用いた変換手法を比較した．本稿では，入力話者の発話映像として男性被験者 1 名から CENSREC-1-AV データベースに含まれる数字発話 65 文を収録した．収録した 65 発話のうち 50 発話を学習データとし， 学習に含まない 15 発話をテストデータとした．学習用データの基底数の総数は 12,870 である．学習される辞書行列の基底数は 200 である．音声特徴量は，唇動画収録と同時に収録した音声を用いる．特徴量として振幅スペクトル 257 次元を用いた．サンプリング周波数は 8 kHz，フレームシフトは 5ms である．画像特徴量は，唇領域を抽出した後 DCT を行って得た低周波成分 200 次元を用いた．出力話者の音声特徴量は，入力と同じデータベースに含まれる女性話者音声から抽出した STRAIGHT スペクトル 513 次元を用いる．GMM の学習に用いるパラレルデータとして，学習時に使用した同一発話から得られたメルケプストラム 24 次元を特徴量とした．また，コスト関数における各パラメータについては実験的に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

本手法における目標音声と生成音声の Mel-CD（Mel-cepstrum Distortion）を Table 1 に示す．Mel-CD は以下の式で表される．

$$\text{Mel-CD[dB]} = 10/\ln 10 \sqrt{2\sum_{d=1}^{24}(mc_d^t - \hat{mc}_d^t)^2} \quad (17)$$

mc_d^t と $\hat{mc}_d^t$ は目標音声と生成音声の MFCC の d 次元目の係数である．横軸の source は入力音声とターゲット音声の Mel-CD を表す．audio NMF， multi NMF は従来手法，sparse-parallel は提案手法による変換音声とターゲット音声との Mel-CD を表す．Table 1 より，スパースパラレル学習を用いた変換手法が，他の 2 つの従来手法に比べて精度が上がっている．これは学習によって画像特徴量の冗長性が解消されたからだと考えられる．

Table 1 Mel-cepstral distortion of each method

source	audio NMF	multi NMF	sparse-train
4.37	3.24	3.23	3.17

5 おわりに

本稿では，これまで提案してきたマルチモーダル声質変換手法にスパースパラレル学習を導入した手法を提案した．従来手法と比べ，学習したコンパクトな辞書で変換した精度が上がっていることから，画像の冗長性を解消した変換が行えることがわかった．今後は本手法を雑音環境下に適応し，実験・評価をしていく．また新たな画像特徴量の導入なども検討していく．

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

[4] 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.

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

[6] 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.

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

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

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

[10] 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.