非負値行列因子分解を用いたマルチモーダル声質変換における画像特徴量の検討 *

☆羅里奈, 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交換する技術である．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, 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．従来，声質変換においては混合正規分布を中心とした統計的な手法が多く提案されてきた [5].

近年では，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は Non-negative Matrix Factorization (非負値行列因子分解：NMF) [6] が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 [7, 8]．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．これまで我々は，このアプローチを応用した NMF 声質変換を提案してきた [9]．この手法では，統計的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．他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．

我々は，NMF 声質変換のノイズロバスト性向上を目標として，マルチモーダル声質変換を提案してきた [10]．この手法では，背景雑音にロバストな唇画像特徴量を音声特徴量と合わせて用いた声質変換を行い，その有効性を示した．この手法では画像特徴量として DCT (離散コサイン変換：Discrete Cosine Transform) を用いており，雑音環境下では一定の成果を上げたが，クリーンな環境下では画像特徴量が変換精度向上に寄与しないという問題があった．

近年, 畳み込みニューラルネット (Convolution Neural Network ：CNN) が注目を集めており，音声・画像信号処理において広く用いられている [11]．CNN は多層パーセプトロン (Multi perceptron：MLP) の前段に畳み込み層とプーリング層をもっており，画像特徴量から識別的な特徴量を学習するのに適していると考えられる．また，出力層付近ではユニット数を少なくしたボトルネック層を学習しておき，そのボトルネック特徴量を用いることで音声認識精度を向上させたという事例も報告されている [12].

本稿では，CNN を用いて抽出したボトルネック特徴量を画像特徴量としたマルチモーダル声質変換を提案し，従来の画像特徴量と比較する．識別的特徴量であるボトルネック特徴量により，クリーン環境下において画像特徴量が変換精度向上に寄与することが期待される．

以降，2 章では，声質変換に用いる画像特徴量について述べ，3 章では，NMF を用いた声質変換手法について説明する．4 章では，評価実験とその結果を示し，5 章で本稿をまとめる．

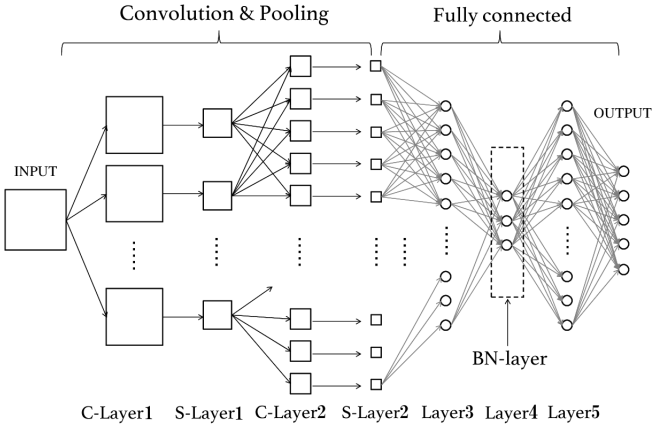

Fig. 1　Convolutive Bottoleneck Network.

*Image Feature for Multimodal Voice Conversion using Non-negative Matrix Factorization. by Rina Ra, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki (Kobe univ.)

2　畳み込みニューラルネットのボトルネック特徴量

2.1　Convolutive Bottleneck Network

提案手法では，Fig.1 に示すようにボトルネックの構造を持つ CNN(以下 ，CBN) を考える．入力層からの数層は，フィルタの畳み込みとプーリングをこの順 で何度か繰り返す構造をとる．つまりフィルタ出力層，プーリング層の 2 層を，プーリング層を次の層の入力層とする形で積み重ねる．出力層は識別対象のクラス数と同じサイズを持つ一次元ベクトルであり，そこに至る何層かは畳み込み・プーリングを挟まない全結合の MLP とする．提案手法では，MLP を 3 層に設計し，中間層のユニット数を少なく抑える (ボトルネック) 構成をとっている．ボトルネック特徴量はボトルネック層のニューロンの線形和で構成され，少ないユニットで多くの情報を表現しているため，入力層と出力層を結び付けるための重要な情報が集約されていると考えられる．

2.2　ボトルネック特徴量の抽出

本節では，CNN のアーキテクチャ，ボトルネック特徴量抽出について述べる．提案手法である CNN を用いたボトルネック特徴量抽出のフローチャートを Fig. 2 に示す．

はじめに，入力話者の発話時の唇領域画像列を用いて，CNN の学習を行う．CNN の入力層 (in) には，スプライン補間によって音声と同期のとれた唇領域画像列データに対して差分を取ったものを用いる．差分画像を用いることで，画像の動的な情報を入力することができると考えられる．出力層 (out) の各ユニットには，入力層の画像データに対する音素ラベル (例えば，音素/i/のメルマップであれば，/i/に対応するユニットだけが値 1，他のユニットが値 0 になる) を割り当てる．ここで，音素ラベルはデータベースで与えられたラベルデータを使用した．CNN は，ランダムな初期値から学習を開始し，確率的勾配降下法 (Stochastic Gradient Descent：SGD) を用いた誤差逆伝搬により，結合パラメータを修正する．

次に，学習したネットワークを用いて特徴量抽出を行う．学習データと同様に，テストデータの唇領域画像を生成し，学習した CNN への入力とする．その後，畳み込みフィルタとプーリングによって入力データの局所的特徴を捉えて，後部の MLP 層によって音素ラベルへと非線形に変換する．入力データの情報はボトルネック層上に集約されているため，提案手法では，このボトルネック特徴量を用いて声質変換を行う．

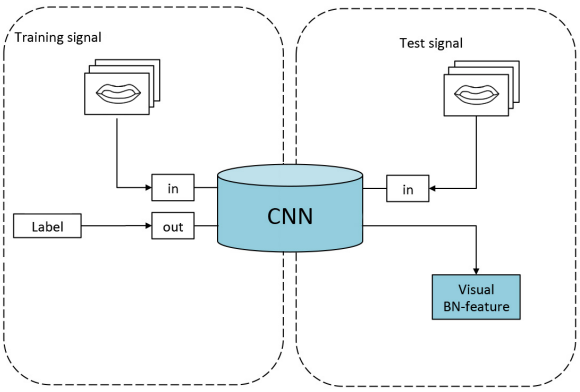

Fig. 2　Flowchart of the bottleneck feature extraction using CBN.

3　NMF を用いたマルチモーダル声質変換

まず, 辞書の構成方法について述べる．Fig.3 は本手法の流れを示す．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 $\mathbf{W}^{sa}$，$\mathbf{W}^{ta}$ をそれぞれ求める．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 $\mathbf{W}^{sv}$ の構築には，第 2 章で述べたように，音声と同期の取れた画像配列に対して差分画像を取り，それを入力とした CNN によるボトルネック特徴量を用いる．このとき，画像特徴量は正の値のみを持つ．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 $\mathbf{W}^{s} = [\mathbf{W}^{sa}; \mathbf{W}^{sv}]$ とし，変換に用いる．

次に，特徴量の重みについて述べる．本手法では，NMF を用いた声質変換を行う．アクティビティ $\mathbf{h}$ を推定する上で，画像特徴量に対して，重み α を導入する．この重み α を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 $\mathbf{h}$ は以下のコスト関数を最小にすることで推定される [13].

$$d(\mathbf{X}^{sa}, [\mathbf{W}^{sa}\ \mathbf{W}^{na}]\mathbf{h}) + \alpha d(\mathbf{X}^{sv}, [\mathbf{W}^{sv}\ \mathbf{W}^{nv}]\mathbf{h}) + \lambda||\mathbf{h}||_1 \qquad s.t. \quad \mathbf{h} \geq 0 \tag{1}$$

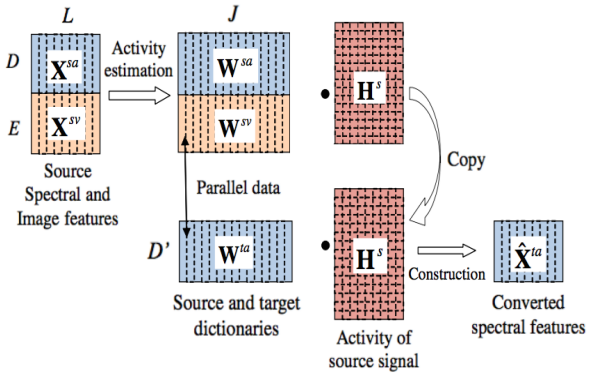

Fig. 3　Flow of multimodal voice conversion

第 1 項と第 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 $\mathbf{h}^s$ をスパースにするための L1 ノルム正規化項である．本稿では画像辞書に関する制約重み $[\lambda_1 \dots \lambda_J]$ を 0.1 に設定した．(1) 式を最小化するよう，以下の更新式に従いアクティビティ行列 $\mathbf{h}^s$ が推定される．

$$\mathbf{h}_{n+1} = \mathbf{h}_n . * ((\mathbf{W}^{sa})^T(\mathbf{X}^{sa}./(\mathbf{W}^{sa}\mathbf{h}^s)) \quad (2)$$
$$+ \alpha(\mathbf{W}^{sv})^T(\mathbf{X}^{sv}./(\mathbf{W}^{sv}\mathbf{h}^s)))$$
$$./(1+\alpha+\lambda)\mathbf{1}^{(J\times L)} \quad (3)$$

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

$$\hat{\mathbf{X}}^{ta} = \mathbf{W}^{ta}\mathbf{h}^s \quad (4)$$

4　評価実験

4.1　実験条件

実験データとして，男性 2 名のパラレルな連続文章発話を用いた．このうち，男性 1 名は M2TINIT [14] に収録されている話者であり，他 1 名は ATR 研究用日本語音声データベースセット [15] に含まれる話者を用いる．サンプリング周波数は 12kHz，フレームシフトは 5ms とした．以上のデータベースを用い，closed 実験を行った．closed 実験では，連続文章発話 50 文をパラレル辞書構築に用いた．辞書に含まれる 10 文を，テストデータとして用いた．

画像特徴量は従来手法で用いた DCT と，提案手法である CBN によるボトルネック特徴量を用いた．

Table 1　Size of each feature map of CNN. $(k, i \times j)$ indicates that the layer has k maps of size $i \times j$.

Input	C1	P1	
1, 12×16	8, 8×12	8, 4×8	
M1	M2	M3	Output
108	30	108	133

CBN の学習には，入力話者の連続発話文 150 文を用い，変換時は，NMF の入力として音声の振幅スペクトルと CBN を結合したものを特徴量として用いる．音声特徴量は，唇動画像収録と同時に収録した音声を用いる．音声特徴量として振幅スペクトル 257 次元を用いた．画像データ構築に用いた動画のフレームレートは 29.97fps である．唇画像のサイズは 100 ×140 pixel であり，従来手法ではこの唇画像に DCT をかけたものを用いた．提案手法では，12 ×16 pixel に切り出して前後 1 フレームとの差分画像としたものが CNN の入力となっている．音声フレームと画像フレームの動機をとるために，従来手法では，画像列に対してスプライン補間を行い，提案手法では，CBN で得たボトルネック特徴量を用いて画像辞書を構築した．画像に対する重み α は 0.5 である．NMF の更新回数は 300 回とした．画像特徴量は学習した CNN で求められた 30 次元のボトルネック特徴量である．出力話者の音声特徴量は，男性話者音声から抽出した STRAIGHT スペクトル 513 次元を用いる．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　ネットワークのサイズ

本稿では，Fig. 1 に示すように，畳み込み層とプーリング層からなる CNN と，ボトルネック層を含む 3 層の MLP とが階層的に接続されたネットワークを考える．CNN の各層における特徴マップのサイズには Table 1 の値を用いた．畳み込みフィルタの数とサイズ，及びプーリングサイズは，これらの値から一意に決定される．なお，画像 CNN ともに，MLP の各層 (ボトルネック層を除く) のユニット数は 108, ボトルネック層のユニット数は 30, 出力層のユニット数は 133 としている．

4.3　実験結果

本手法における目標音声と変換音声の Mel-CD(Mel-cepstral Distortion) を Table 2 に示す．

Table 2　Mel-cepstral distortion of each method

source	DCT	CBN
4.0139	2.329	3.054

Mel-CD は以下の式で表される．

$$\text{Mel-CD[dB]} = 10/\ln 10 \sqrt{2\sum_{d=1}^{10}(mc_d^t - \hat{mc}_d^t)^2} \quad (5)$$

mc_d^t と $\hat{mc}_d^t$ は目標音声と変換音声のメルケプストラムの d 次元目の係数である．source は入力音声とターゲット音声の Mel-CD を表す．表を見ると，従来手法の DCT を用いたほうが良い結果となった．変換した音声を聞いたところ二つの音声に大きな差は感じられなかった．歪が大きくなった原因としては，CBN を学習する際のデータの数が，各パラメータ・出力に対して不十分であったことが考えられる．close 実験での CBN の認識率は 53% であり，出力の次元数やパラメータを調整し十分なデータを用意して CBN を学習する必要がある．

5　おわりに

本研究では CBN のよる唇画像ボトルネック特徴量を用いたマルチモーダル声質変換の手法を提案した．結果，CNN に関して十分なデータを用いて学習し最適な特徴量を得る必要があるとわかった，

今後は，差分画像をとる前後のフレーム数や，出力次元数を検討する．差分画像に対して，認識の精度に重要な影響を与えるとされる動的特徴を用いた画像特徴量を考える．更に，雑音環境下で提案手法の有効性を示していく．

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. Marcon, “Spectral voice conversion for text-to-speech synthesis,” 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] 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.

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

[7] 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.

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

[9] 相原龍，中鹿亘，滝口哲也，有木康雄，“辞書選択に基づく非負値行列因子分解による声質変換” 日本音響学会 2013 秋季研究会, 3-7-6, pp. 1473-1476, 2013.

[10] K. Masaka, R. Aihara,T. Takiguchi, and Y. Ariki, “Multimodal voice conversion using Non-negative matrix Factorization in noisy environments,” ICASSP2014, pp. 1561-1565, 2014.

[11] Y. Lecun and Y. Bengio , “Convolutional networks for images,speech, and time-series,” in Arbib, M. A. (Eds), The Handbook of Brain Theory and Neural Networks, MIT Press, 1995.

[12] Y. Takashima, Y. Kakihara *et al* “Audio-Visual Speech Recognition Using Convolutive Bottleneck Networks for a Person with Severe Hearing Loss,” IPSJ Trans. pp. 64-68, 2015.

[13] K. Masaka, R. Aihara, Tetsuya Takiguchi and Yasuo Ariki, “Multimodal exemplar-based voice conversion using lip features in noisy environments,” Interspeech, pp. 1159-1163, 2014.

[14] 小林隆夫，徳田恵一，“http://m2tinit.ics.nitech.ac.jp” 2013.

[15] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol. 9, pp. 357-363, 1990.