複素 NMF を用いた声質変換の検討 *

☆李権俊, 相原龍, 滝口哲也, 有木康雄 (神戸大学)

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, 中でも混合正規分布モデル (Gaussian Mixture Model: GMM) を用いた手法 [4] が広く用いられており, 多くの改良がされ続けている.

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 [5]. 近年スパース表現に基づく手法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 [6, 7]. スパース表現の考え方においては, 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ごとにグループ (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 [8] 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.

我々の提案している声質変換手法では, Non-negative Matrix Factorization (NMF) [9] を用いてきた. この手法では, 入力話者の音声辞書 (入力話者辞書) と出力話者の音声辞書 (出力話者辞書) からなる同一発話内容のパラレル辞書を構築する. 変換時には辞書を固定し, 入力音声を NMF によって入力辞書に含まれる少量の基底からなるスパース表現にする. 得られた入力辞書の基底毎の重み係数 (アクティビティ) に基づいて, 入力話者辞書の基底を出力辞書内の基底と置き換え, 線形結合することで, 出力話者の音声へと変換する.

NMF 声質変換では特徴量として振幅スペクトルが用いられてきたが, 振幅スペクトルは非加法的であるため, 入力信号を振幅スペクトル基底の線形和でモデル化する NMF では誤差が生じることが考えられる. さらに従来の声質変換では, 位相情報を考慮した変換がなされていない. これらの問題を解決するために本研究では, 振幅スペクトルではなく複素スペクトルでのモデル化を行った声質変換についての検討を行う. 具体的には亀岡ら [10] が提案している複素 NMF, Ahuja ら [11] が提案している複素スペクトルを非負値行列に変形して, 分解する手法の 2 つを用いて声質変換手法を実装し, 比較検討を行う.

2 NMF を用いた声質変換

スパース表現の考え方において，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.

$$\mathbf{v}_l \approx \sum_{j=1}^{J} \mathbf{w}_j h_{j,l} = \mathbf{W}\mathbf{h}_l \tag{1}$$

$\mathbf{v}_l$ 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. $\mathbf{w}_j$ 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 $\mathbf{w}_j$ とする. 基底を並べた行列 $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_J]$ 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.

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \tag{2}$$

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

ここで L はフレーム数を表し, 本手法において，$\mathbf{W}$ は学習データで固定される.

本手法の概要を Fig. 1 に示す. $\mathbf{V}^s$ は入力話者スペクトル，$\mathbf{W}^s$ は入力話者辞書，$\mathbf{W}^t$ は出力話者辞書，$\hat{\mathbf{V}}^t$ は変換されたスペクトル，$\mathbf{H}^s$ は入力話者スペクトルから推定されるアクティビティを表す. D，J はそれぞれスペクトルの次元数，辞書の基底数である. この手法では，パラレル辞書と呼ばれる入力話者辞書 $\mathbf{W}^s$ と出力話者辞書 $\mathbf{W}^t$ 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 Dynamic Time Warping (DTW) 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たものである.

---

* Voice conversion using complex NMF, by Konjun I, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki (Kobe univ.)

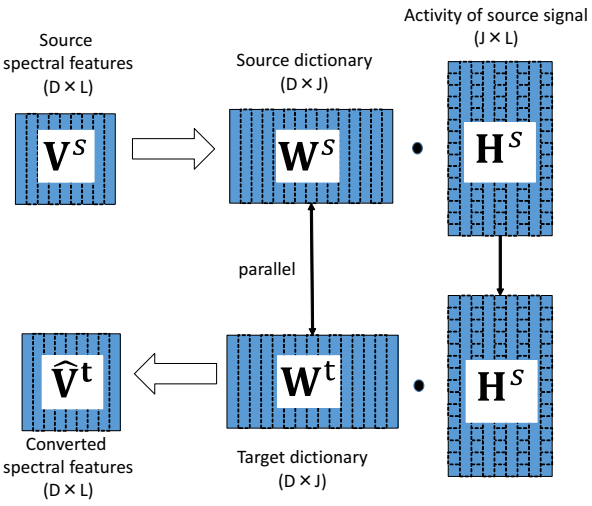

Fig. 1　Basic approach of exemplar-based voice conversion

入力スペクトル $\mathbf{V}^s$ は NMF によって $\mathbf{W}^s$ と $\mathbf{H}^s$ の積に分解される．本手法では，「パラレル辞書で推定したパラレルな発話のアクティビティは置き換え可能である」と仮定している．従って，変換スペクトル $\hat{\mathbf{V}}^t$ は，$\mathbf{W}^t$ と推定した $\mathbf{H}^s$ の積によって得られる．

3　Joint-NMF による声質変換

複素スペクトルを非負値に変形して, NMF を用いて分解する手法を [11] らが提案している．この手法では複素行列 $\mathbf{Y}$ を複素行列 $\mathbf{W}$ と実数値行列 $\mathbf{H}$ に分解することを考える．このとき, $\mathbf{Y}$ を以下のように非負値を用いて表す．

$$\mathbf{Y} = \mathbf{Y}_{+r} - \mathbf{Y}_{-r} + j(\mathbf{Y}_{+i} - \mathbf{Y}_{-i}) \tag{4}$$

$$\begin{aligned} \mathbf{Y}_{+r} &= max(0, real(\mathbf{Y})) \\ \mathbf{Y}_{-r} &= -min(0, real(\mathbf{Y})) \\ \mathbf{Y}_{+i} &= max(0, imag(\mathbf{Y})) \\ \mathbf{Y}_{+r} &= -min(0, imag(\mathbf{Y})) \end{aligned} \tag{5}$$

$\mathbf{W}$, $\mathbf{H}$ についても以下のように表す．

$$\mathbf{W} = \mathbf{W}_{+r} - \mathbf{W}_{-r} + j(\mathbf{W}_{+i} - \mathbf{W}_{-i}) \tag{6}$$

$$\mathbf{H} = \mathbf{H}_{+} - \mathbf{H}_{-} \tag{7}$$

これらを用いて, 以下のように $\mathbf{Y}$ を分解する．

$$\begin{aligned} \mathbf{Y}_{+r} &= \mathbf{W}_{+r}\mathbf{H}_{+} + \mathbf{W}_{-r}\mathbf{H}_{-} \\ \mathbf{Y}_{-r} &= \mathbf{W}_{+r}\mathbf{H}_{-} + \mathbf{W}_{-r}\mathbf{H}_{+} \\ \mathbf{Y}_{+i} &= \mathbf{W}_{+i}\mathbf{H}_{+} + \mathbf{W}_{-i}\mathbf{H}_{-} \\ \mathbf{Y}_{-i} &= \mathbf{W}_{+i}\mathbf{H}_{-} + \mathbf{W}_{-i}\mathbf{H}_{+} \end{aligned} \tag{8}$$

式 (8) を以下のようにまとめて表す．

$$\begin{pmatrix} \mathbf{Y}_{+r} & \mathbf{Y}_{-r} \\ \mathbf{Y}_{+i} & \mathbf{Y}_{-i} \end{pmatrix} = \begin{pmatrix} \mathbf{W}_{+r} & \mathbf{W}_{-r} \\ \mathbf{W}_{+i} & \mathbf{W}_{-i} \end{pmatrix} \begin{pmatrix} \mathbf{H}_{+} & \mathbf{H}_{-} \\ \mathbf{H}_{-} & \mathbf{H}_{+} \end{pmatrix} \tag{9}$$

式 (9) を用いて, NMF 声質変換と同様にアクティビティ $\mathbf{H}$ を求め, 出力話者辞書との積をとることで入力音声を変換する．

4　複素 NMF による声質変換

4.1　複素 NMF

音声信号として, 時間周波数スペクトルを入力し, NMF を用いて分解することを考えると, 時間周波数スペクトルは振幅スペクトルであるため, 周波数スペクトルを基底とすると, 加法性が成立しない．これに対し複素 NMF の考え方では, 加法性が成り立つ複素スペクトルでのモデル化を行う．複素 NMF では音声の複素スペクトルが以下のように分解表現される．

$$F_{x,t} = \textstyle\sum_k W_{k,x} H_{k,t} e^{j\phi_{k,x,t}} \tag{10}$$

$W_{k,x}$ は, 時間に依存しないグローバルに決定される基底となり, $H_{k,t}$ は k 番目のスペクトルのアクティビティ係数で, 位相スペクトル $e^{j\phi_{k,x,t}}$ とともに時間ごとに変化する．ここでスケールの任意性を除くため,

$$\sum_x W_{k,x} = 1 \tag{11}$$

とする．観測信号の複素スペクトルを $\mathbf{Y}$ とすると, $\mathbf{W}$, $\mathbf{H}$, $e^{j\phi}$ を求める問題は,

$$\begin{aligned} f(\mathbf{W}, \mathbf{H}, \phi) &= \sum_{x,t} |Y_{x,t} - \sum_k W_{k,x} H_{k,t} e^{j\phi_{k,x,t}}|^2 \\ &\quad + 2\lambda \sum_{k,t} |H_{k,t}|^p \end{aligned} \tag{12}$$

を最小化する最適化問題となる．ここで式の 2 項目はスパース制約項であり, λ は任意の定数, p は $0 < p < 2$ を満たす定数である．この最適化問題を補助関数法を用いて解くと以下の更新式が得られる．

$$W_{k,x} = \frac{\sum_t \frac{H_{k,t}|\bar{Y}_{k,x,t}|}{\beta_{k,x,t}}}{\sum_t \frac{{H_{k,t}}^2}{\beta_{k,x,t}}} \tag{13}$$

$$H_{k,t} = \frac{\sum_x \frac{W_{k,x}|\bar{Y}_{k,x,t}|}{\beta_{k,x,t}}}{\sum_x \frac{{W_{k,x}}^2}{\beta_{k,x,t}} + \lambda p |\bar{H}_{k,t}|^{p-2}} \tag{14}$$

$$e^{j\phi_{k,x,t}} = \frac{\bar{Y}_{k,x,t}}{|\bar{Y}_{k,x,t}|} \tag{15}$$

4.2 位相情報の適応

複素 NMF を用いた声質変換においても, NMF 声質変換と同様に, 入力話者辞書 $\mathbf{W}^s$ と出力話者辞書 $\mathbf{W}^t$ を, 入力話者と出力話者の同一発話の振幅スペクトルで固定し, アクティビティを入れ替えることで変換音声を合成することを考える. しかし, アクティビティを入れ替えただけでは位相情報に含まれる話者性を変換することはできないため, 入力話者の位相を出力話者の位相に適応させることを考える. 適応データとして, 入力話者と出力話者の同一内容の発話 $\mathbf{Y}^{sa}$, $\mathbf{Y}^{ta}$ を用意する. 入力話者の適応データの複素スペクトル $\mathbf{Y}^{sa}$ と入力話者辞書 $\mathbf{W}^s$ を用いて以下の式を最小化するアクティビティ $\mathbf{H}^{sa}$ と位相 $e^{j\phi^{sa}}$ を推定する.

$$\sum_{x,t}|Y^{sa}_{x,t}-\sum_k W^s_{k,x}H^{sa}_{k,t}e^{j\phi^{sa}_{k,x,t}}|^2+2\lambda\sum_{k,t}|H^{sa}_{k,t}|^p \tag{16}$$

ここで, 入力話者辞書は入力話者の音声から抽出した振幅スペクトルを並べたものである. 適応データである出力話者の複素スペクトル $\mathbf{Y}^{ta}$ は, $\mathbf{Y}^{sa}$ のパラレルデータなので , $\mathbf{Y}^{ta}$ は出力話者辞書 $\mathbf{W}^t$, 位相 $e^{j\phi^{ta}}$, 推定されたアクティビティ $\mathbf{H}^{sa}$ を用いて以下のように表せる.

$$Y^{ta}_{x,t}=\sum_k W^t_{k,x}H^{sa}_{k,t}e^{j\phi^{ta}_{k,x,t}} \tag{17}$$

ここで $e^{j\phi^{ta}}$ が適応行列 A と $e^{j\phi^{sa}}$ の積によって表現できると考えると

$$Y^{ta}_{x,t}=\sum_k W^t_{k,x}H^{sa}_{k,t}e^{jA_k\phi^{sa}_{k,x,t}} \tag{18}$$

となる. 適応行列 $\mathbf{A}$ は以下の式により求める.

$$e^{jA_k\phi^{sa}_{k,x,t}}=\frac{\bar{Y}_{k,x,t}}{|\bar{Y}_{k,x,t}|} \tag{19}$$

変換する音声のアクティビティ$\mathbf{H}^s$ と位相 $e^{j\phi^s}$ を推定し, 以下の式のように計算することで, 変換音声の複素スペクトル $\mathbf{Y}^t$ を得る.

$$Y^t_{x,t}=\sum_k W^t_{k,x}H^s_{k,t}e^{jA_k\phi^s_{k,x,t}} \tag{20}$$

5 評価実験

5.1 実験条件

ATR 研究用日本語音声データベースセット [12] を用いて話者変換を行い，提案手法である複素 NMF を用いた声質変換 (CNMF), Ahuja らの手法を用いた声質変換 (JNMF) と，従来の NMF 声質変換，及び GMM 声質変換との比較を行った．入力話者は男性, 出力話者は女性, サンプリング周波数は 8kHz とした. パラレル辞書の構築に 50 単語を使用した．従来の NMF 声質変換の特徴量には STRAIGHT [13] で計算されたスペクトル包絡を用い, 提案手法の特徴量には短時間フーリエ変換によって計算された複素スペクトルを用いた. GMM に基づく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メルケプストラムをフレーム間同期を取る事で作った 50 単語のパラレルデータを用いた．メルケプストラムは STRAIGHT スペクトルから計算される線形ケプストトラムで, 次元数は 24 である. GMM の混合数は 40 とした. 従来の NMF 声質変換及び GMM 声質変換では, F0 をパラレルデータを用いた単回帰分析によって変換している. 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ない 50 単語を用いた.

提案手法の有効性を確かめるため，客観評価実験を行った. 客観評価は短時間フーリエ変換によって計算された, 振幅スペクトルとそのメルケプストラム 24 次元を用いて，式 (21) で表される NSD(Normalized Spectrum Distortion) と, 式 (22) で表されるメルケプストラム歪 (Melcepstrum distortion: MCD) [dB] とによって各手法を比較した.

$$NSD=\sqrt{\frac{||S^X-S^{\hat{X}}||^2}{||S^Y-S^X||^2}} \tag{21}$$

ここで, S^X，S^Y，$S^{\hat{X}}$ 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.

$$MCD=(10/\log 10)\sqrt{2\sum_d^{24}(mc_d^{conv}-mc_d^{tar})^2} \tag{22}$$

ここで，mc_d^{conv}，mc_d^{tar} は d 次元目の変換後のメルケプストラム，目標音声のメルケプストラムを表す.

5.2 実験結果・考察

Fig. 2 に NSD による比較, Fig. 3 に MCD による実験結果の比較を示す.

図より提案手法は NSD, MCD による比較では従来手法を上回る精度で変換できていることが確認できる．しかし, 合成された音声を聞き比べてみると, 提案手法で合成された音声は従来手法で合成された音声と比べて, 自然性に欠ける音声になってしまっていた. 理由としては, CNMF を用いた変換では, 位相情報の適応が不十分であること, JNMF を用いた変換では，複素成分のアクティビティの推定精度が不十分であることが考えられる.

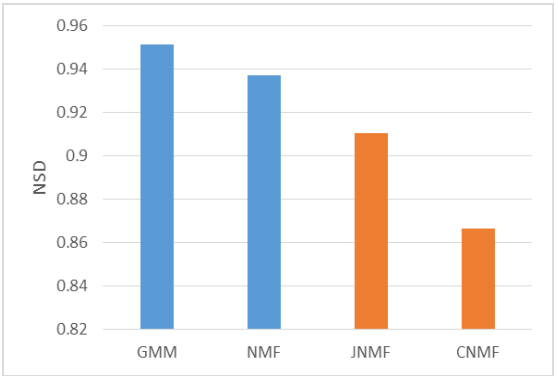

Fig. 2　NSD for converted voice

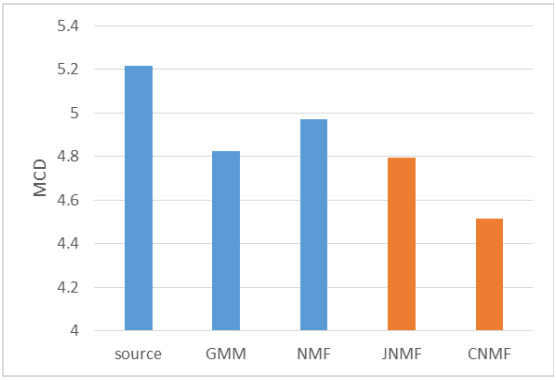

Fig. 3　MCD for converted voice

6　おわりに

本稿では複素スペクトルを用いた声質変換手法についての検討を行った．実験結果より，提案手法は客観評価では従来の NMF 声質変換を上回る変換精度を持っていることがわかる．しかし音声を聞き比べてみると提案手法はいずれも STRAIGHT を用いて合成される従来の NMF 声質変換と比べて，自然性で劣っている．位相スペクトルの適応精度が不十分であると考えられ，今後研究を進めていく．一方で STRAIGHT には雑音重畳音声の雑音を上手く表現できないという問題点があるので，提案手法は雑音重畳音声の変換にはより適しているとも考えられる．今後は雑音重畳音声の変換を中心に複素 NMF による声質変換について検討を進めて行く．

参考文献

[1] Y. Iwami *et al.*, “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 *et al.*, “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” Speech Communication, vol. 54, no. 1, pp. 134–146, 2012.

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

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

[6] 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.

[7] M. N. Schmidt and R. K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” in Proc. Interspeech, 2006.

[8] J. Gemmeke *et al.*, “Exemplar-based sparse representations for noise robust automatic speech recognition,” IEEE Trans. Audio, Speech, Lang. Process., pp. 2067–2080, 2011.

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

[10] H. Kameoka *et al.*, “Complex nmf: A new sparse representation for acoustic signals,” in Proc. ICASSP, p. 34373440, 2009.

[11] C. Ahuja *et al.*, “A complex matrix factorization approach to joint modeling of magnitude and phase for source separation,” arXiv, 2014.

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

[13] H. Kawahara *et al.*, “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, no. 3-4, pp. 187–207, 1999.