一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2014-116(2014-12)

雑音環境下における特徴量重み付き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††

† 神戸大学システム情報学研究科　〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
†† 神戸大学自然科学系先端融合研究環　〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: †{makka,aihara}@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし　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，発話支援など様々なタスクへの応用が期待されている．従来の代表的な手法である GMM を用いた統計的なアプローチ等は，あくまでクリーン音声を用いた評価を行っており，雑音環境下を考慮した定式化はされていない．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）による声質変換を扱う．我々はこれまで，この NMF に基づいた，音声だけではなく唇画像情報を用いたマルチモーダルな声質変換手法を提案してきた．入力話者の特徴量として，音声と画像情報を用いることで変換精度，及びノイズロバスト性の向上を目指した．本論文では，さらに特徴量重みを導入し，新たにコスト関数を提案した．実験結果より，音声情報のみを用いた NMF による声質変換，及び GMM による声質変換よりも提案手法が精度の良い変換が行える事を示す．
キーワード　声質変換，マルチモーダル，画像特徴量，NMF，雑音環境下

Multimodal Voice Conversion using Weighted Features in Noisy Environments

Kenta MASAKA†, Ryo AIHARA†, Tetsuya TAKIGUCHI††, and Yasuo ARIKI††

† Graduate School of System Informatics, Kobe University
1-1 Rokkodai, Nada, Kobe 657-8501, Japan
†† Organization of Advanced Science and Technology, Kobe University
1-1 Rokkodai, Nada, Kobe 657-8501, Japan
E-mail: †{makka,aihara}@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract　Voice conversion is a technique for converting specific information in speech while maintaining the other information, such as linguistic information. This technique has been applied to various tasks, for example, there are speaker conversion, emotion conversion and speaking assistance, etc. The GMM-based method is conventional VC method and widely used. In noisy environments, the GMM-based method cannot convert the speech well, because this method cannot model the noisy signal well. Therefore, we have been researched about a noise-robust VC method using Non Negative Matrix Factorization (NMF). In this paper, we propose a multimodal VC method that improves the noise robustness of our previous exemplar-based VC method. Furthermore, we introduce the combination weight between audio and visual features and formulate a new cost function in order to estimate the audio-visual exemplars. By using the joint audio-visual features as source features, the VC performance is improved compared to a previous audio-input exemplar-based VC method. The effectiveness of this method was confirmed by comparing it with that of the conventional audio input NMF-based method and the conventional GMM-based method.
Key words　voice conversion, multimodal, image features, non-negative matrix factorization, noisy environments

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．本研究における手法は，従来の NMF による声質変換では用いられていない唇画像特徴を組み込みんだ，マルチモーダルな変換手法 [7] となっている．さらに特徴量ごとの重みを導入し，新たな更新式を定義することでさらなる変換精度の向上を目指した．

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．Helander ら [9] 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．またこれらの手法では，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 [10] や Eigen-Voice GMM (EV-GMM) [11], [12] などが提案されている．

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．そのため，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 [13] 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 [14] 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 [15] 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 [16] 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 [17] 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．

本稿では，雑音環境下に強い NMF 基づく声質変換に唇画像特徴を組み込んだ手法を提案する．ここで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．更に本手法では，入力話者の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．また，音声と画像特徴量に重みを導入し，最適な重みを選ぶことで単に画像を導入したときよりも，さらに良い精度で変換が行えることを示す．

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・音声のみ NMF による声質変換手法と比較し，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．

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

$\mathbf{x}_l$ 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

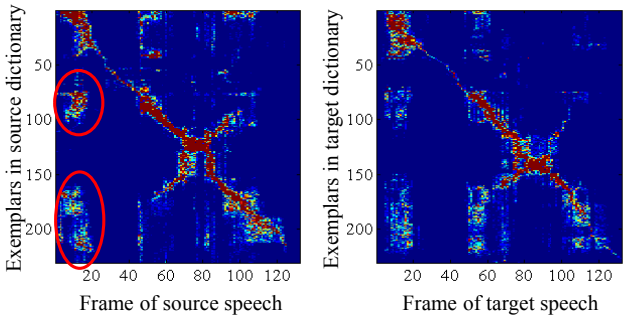

図 1　入力（左）・出力（右）話者の重み行列

Fig. 1　Activity matrices of the source signal (left) and target signal (right)

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．

$$\mathbf{X} \approx \mathbf{AH} \tag{2}$$

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

ここで L はフレーム数を表す．本手法の概要を図 2 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．

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．これによってより雑音に頑健な変換となる．本手法では，さらに音声，画像の特徴量ごとに重みを導入し，新たな更新式を定義した．

3.1　辞書構成法

図 3 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

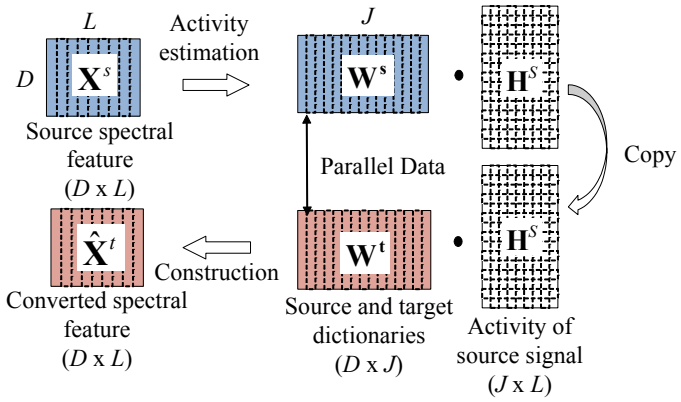

図 2　NMF に基づく声質変換

Fig. 2　Basic approach of NMF-based voice conversion

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た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 50 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．

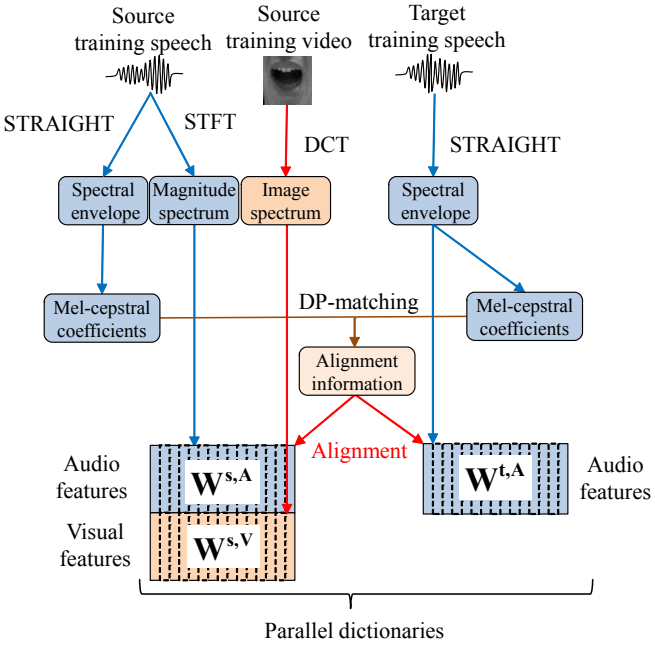

図 3　音声・画像辞書構築方法

Fig. 3　Multimodal dictionary construction

3.2 変換手法

提案手法の概略を図 4 に示す．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．

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

$\mathbf{x}^s$ 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 α と β 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．

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

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．

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

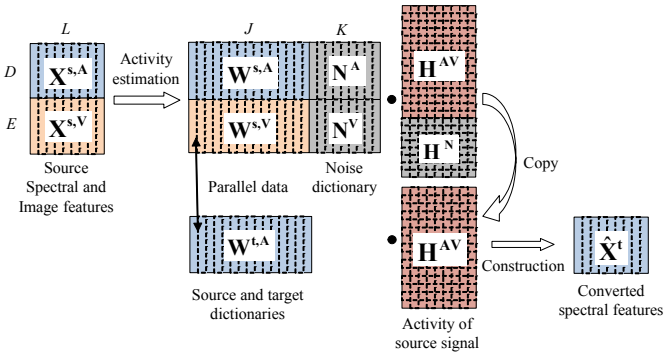

図 4　マルチモーダル声質変換
Fig. 4　Flow chart of multimodal conversion

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，音声特徴のみを用いた NMF による手法を比較手法として実験を行った．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．収録した音声の波形と唇の画像を図 5，図 6 に示す．CENSREC-1-AV 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 5ms とした．連続数字発話 40 文からパラレルデータを作成し，NMF におけるパラレル辞書の構築，従来手法における GMM の学習にそれぞれ用いた．桁数ごとの文章の内訳は 1 桁が 10 文，2 桁が 6 文，3 桁が 6 文，4 桁が 8 文，5 桁が 6 文，7 桁が 4 文である．入力話者の音声辞書には，振幅スペクトル 256 次元，出力話者の音声辞書には STRAIGHT スペクトル 513 次元を用いた．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 MFCC 24 次元を特徴量とした．混合数は 32 となっている．テストデータとなる連続数字発話として，辞書作成時に使用した文章とは別の 2 桁から 7 桁の文章 10 文を用い，それぞれに雑音信号を加算した．桁数ごとの文章の内訳は 2 桁が 3 文，3 桁が 1 文，4 桁が 2 文，5 桁が 2 文，7 桁が 2 文である．雑音信号は ホワイトノイズ，及び CENSREC-1-C データベース [18] に含まれる車内，空港，食堂内，地下鉄で収録されたそれぞれ音声の無音声部分の雑音を用いた．雑音信号の SNR は 0，10，20 dB においてそれぞれ評価した．アクティビティ行列の推定値の更新回数は 300 とした．動画のフレームレートは 29.97 fps で，画像のサイズは 130 × 80 となっている．画像の特徴量として DCT を用いている．DCT された画像から低次 50 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．音声フレームと画像フレームの同期を取るために画像特徴量に対してスプライン補間を行い，セグメント特徴量を導入し，前後 2 フレーム分，計 250 次元を画像特徴として画像辞書を構築した．また，画像と音声の重みについては音声に対する重み α は 1 に固定し，画像に対する重み β を 1 から 10 の間で変化させて評価を行った．

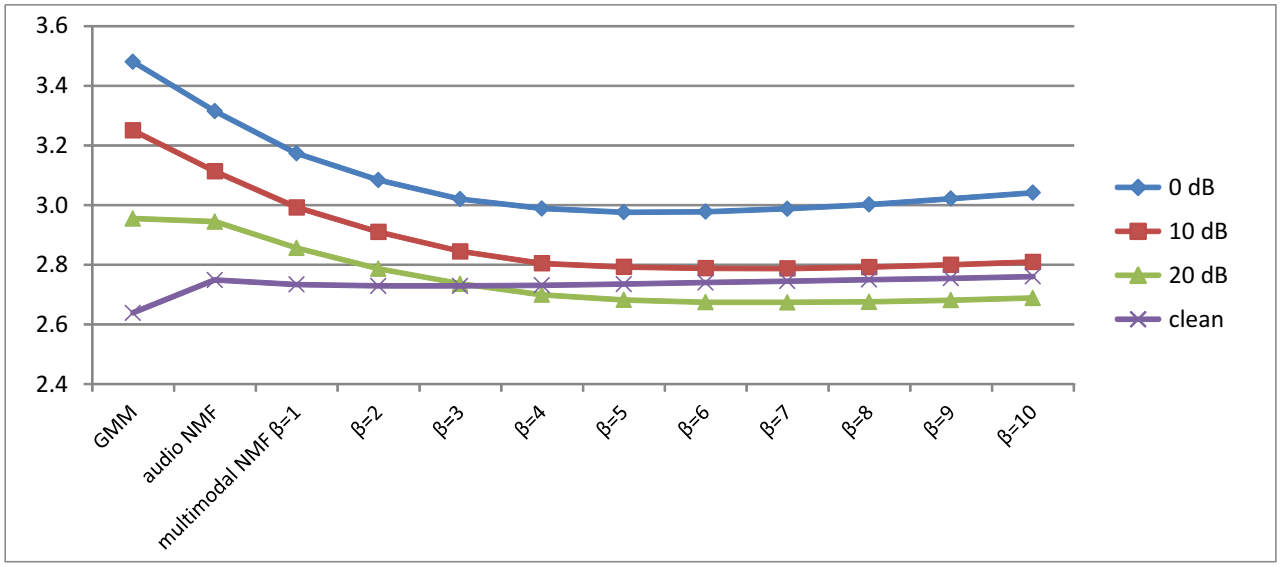

図 7　Mel-cepstrum distortion (ホワイトノイズ環境下，β は画像の重み)

Fig. 7　Mel-cepstrum distortion in white noise (β is the weight of image feature)

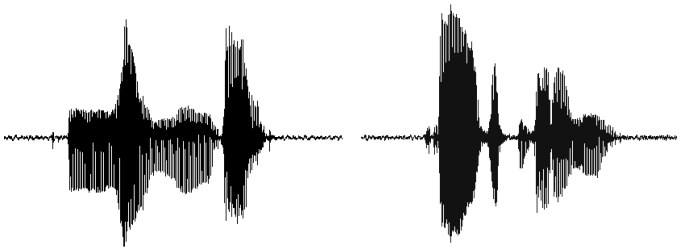

図 5　音声波形図

Fig. 5　Audio waves

図 6　唇　画　像

Fig. 6　Lip images

4.2　実験結果・考察

出力話者音声と，各手法における変換音声の MCD（Mel-cepstrum Distortion）を図 7 に示す．MCD は以下の式で表される．

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

mc_d^t と $\hat{mc}_d^t$ はターゲットと変換音声の d 次元目の係数である．図 7 より，雑音環境下において，提案手法の結果 (multimodal NMF) が GMM，音声のみの NMF による変換に比べて良くなっていることがわかる．また，重み β を導入することにより，最適な画像特徴量の重みを選ぶことができ，重みを導入しない場合（β=1）に比べて良い結果を得られることがわかった．SNR=0dB において，マルチモーダル NMF の一番良い MCD の値は，$\beta = 5$ のとき 2.976 であり，このときの音声のみの NMF との差は 0.338 であった．また，SNR=20dB においては，$\beta = 7$ のとき 2.674 であり，音声のみの NMF との差は 0.271 であった．以上の結果より，雑音が大きい環境の方がひずみの差が大きいため，音声のみ NMF に比べより雑音の大きい環境において提案手法が有効であることが示された．クリーン環境下においても，音声のみの NMF 変換に比べ，画像を入れた変換がわずかに有効であることがわかる．

また，様々な種類の雑音環境下における提案手法の有効性を示す実験を行った．各雑音環境下における変換結果を 表. 1 に示す．本実験において各雑音の SNR は 10 ，画像の重み β は 5 となっている．

表 1　各雑音環境下における MCD

Table 1　MCD in each noisy environment

Noise	audio NMF	multimodal NMF
white	3.113	**2.788**
car	3.041	**2.896**
airport	2.978	**2.869**
restaurant	3.021	**2.893**
subway	3.064	**3.006**

表 1 より，すべての雑音環境下において，提案手法が有効であることわかる．ホワイトノイズのような定常な雑音環境下においては，画像を入れた変換の改善値が大きいが，空港や地下鉄などの非定常な雑音環境下においては改善値が小さくなっている．

さらに，改善部分の解析のため，子音と母音に分けて改善値を算出した．改善値は以下の式で表される MCD 比を用いて算

出した．本実験において雑音の SNR は 10 dB，画像の重み β は 5 となっている．

$$\text{Mel-CDR} = \frac{\sqrt{\sum_{d=1}^{24}(mc_d^t - mc_d^s)^2}}{\sqrt{\sum_{d=1}^{24}(mc_d^t - \hat{mc}_d^t)^2}} \tag{11}$$

その結果を図 2 に示す．

表 2　母音と子音における MCDR
Table 2　MCDR of vowel and consonant

phoneme	audio NMF	multimodal NMF
vowel	1.508	1.620
consonant	1.477	**1.676**

mc_s^t は入力音声の d 次元目の係数である．表 2 の結果より，子音，母音ともに画像を入れた変換が音声のみの結果に比べて良くなっている．また，音声のみの変換では母音の改善値が子音の改善値より大きいが，画像を入れた変換では子音の改善値の方が大きくなっている．これは音声のみの変換では劣化してしまっている子音部分を，画像を組み込むことによって補っているものだと考えられる．

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，画像特徴を導入した．これにより，音声特徴のみを用いた変換よりもさらに雑音に頑健な変換を行うことが可能となり変換精度が向上した．評価実験を行い，従来の統計的モデルを用いた声質変換法や音声特徴のみを用いた NMF よりも高い精度で変換できることを示した．さらに，音声と画像の特徴量ごとに重みを導入することにより，単に画像を入れた変換よりも良い変換結果を示すことができた．また，どの雑音環境下においても本提案手法が有効であることがわかった．今後は他の画像特徴量での比較や，より精度の高い変換手法の改良を進めていく．

文　献

[1] Y. Stylianou, O. Cappe, and E. Moilines, “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., 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, R. Takashima, T. Takiguchi, and Y. Ariki, “GMM-based emotional voice conversion using spectrum and prosody features,” American Journal of Signal Processing, vol.2, no.5, pp.134–138, 2012.

[5] 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.

[6] R. Takashima, T. Takiguchi, and Y. Ariki, “Exemplar-based voice conversion in noisy environment,” in SLT, pp.313–317, 2012.

[7] K. Masaka, R. Aihara, T. Takiguchi, and Y. Ariki, “Mutimodal voice conversion using non-negative matrix factorization in noisy environments,” in ICASSP 2014, pp.1561–1565, 2014.

[8] 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.

[9] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, “Voice conversion using partial least squares regression,” IEEE Trans. Audio, Speech, Lang. Process., vol. 18, Issue:5, pp.912–921, 2010.

[10] C.H. Lee and C.H. Wu, “Map-based adaptation for speech conversion using adaptation data selection and non-parallel training,” in Interspeech, pp.2254–2257, 2006.

[11] T. Toda, Y. Ohtani, and K. Shikano, “Eigenvoice conversion based on Gaussian mixture model,” in Interspeech, pp.2446–2449, 2006.

[12] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, “One-to-many voice conversion based on tensor representation of speaker space,” in Interspeech, pp.653–656, 2011.

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

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

[15] Y. Komai, N. Yang, T. Takiguchi, and Y. Ariki, “Robust aam-based audio-visual speech recognition against face direction changes,” ACM Multimedia, pp.1161–1164, 2012.

[16] E. Bateson, G. Munhall, M. Hirayama, Y. Lee, and D. Terzopoulos, The Dynamics of Audiovisual Behavior in Speech, Springer Berlin Heidelberg, 1996.

[17] 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.

[18] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda, and S. Nakamura, “CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments,” Acoustical Science and Technology, Vol. 30 (2009), No. 5, pp.363–371, 2009.