マルチモーダル声質変換を用いた脳性麻痺構音障がい者のコミュニケーション支援

代表研究者	滝口　哲也	神戸大学　都市安全研究センター　准教授
共同研究者	有木　康雄	神戸大学　都市安全研究センター　教授

1 はじめに

脳性麻痺とは，筋肉の動きをつかさどる脳の部分が受けた損傷が原因で筋肉の制御ができなくなり，けいれんや麻痺，そのほかの神経障がいが起こる症状のことである．それらの原因は多様であり，出生前・出生時・出生直後の脳への酸素供給，出生前の胎内感染，妊娠中毒症，分娩時の外傷，仮死状態，未熟出生，出生後の脳を覆う組織の炎症や外傷性損傷などがあげられている[1].

脳性麻痺は脳の損傷部分によって 4 つの種類に分類され，そのなかでも，脳性麻痺のアテトーゼ型患者はアテトーゼと呼ばれる筋肉が不随に動き正常に制御できない症状が現れる．この症状はとくに意図的な動作を行う場合や，緊張状態にある時に見られ，その運動障がいの一つとして，正しく構音できない場合がある．アテトーゼ症状は軽度から重度まで様々であり，さらに知能障がいを合併していないケースや比較的知能障がいの程度が軽いケースも多いのが特徴であることから，アテトーゼ型脳性麻痺による構音障がい者を対象とした発話支援システムが求められている．我々はこれまでに脳性麻痺構音障がい者のためのコミュニケーション支援として，声質変換，音声認識の研究[例えば 2, 3, 4]を行っている．これまでは声のみに注目していたが，本研究では声だけでなく唇映像にも注目した声質変換法を検討する．

2 声質変換

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に用いられている他，感情情報を変換する感情変換，失われた話者情報を復元する発話支援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) による声質変換を扱う．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声質変換に組み込むことで，変換精度の向上を目指す.

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)を用いた手法はその精度の良さと汎用性から広く用いられており，多くの改良が続けられている．戸田らは従来の 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や Eigen-Voice GMM (EV-GMM)などが提案されている.

しかし,これらの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．実際の環境では，周囲の背景雑音が音声に重畳するため，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てしまう．また脳性麻痺者のように運動麻痺なども伴う場合は，移動も容易ではないため，使用環境について条件を課すことは難しい．よって実際の生活環境下（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 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（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，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．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．

本研究では，実環境下（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．

3 非負値行列因子分解による声質変換

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．

$$\boldsymbol{x}_l \approx \sum_j a_j h_{j,l} = A\boldsymbol{h}_l$$

$\boldsymbol{x}_l$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．$\boldsymbol{a}_j$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 $\boldsymbol{a}_j$とする．基底を並べた行列 A は辞書と呼び，重みを並べたベクトル $\boldsymbol{h}_l$はアクティビティと呼ぶ．このアクティビティベクトル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．

本手法の概要を図 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

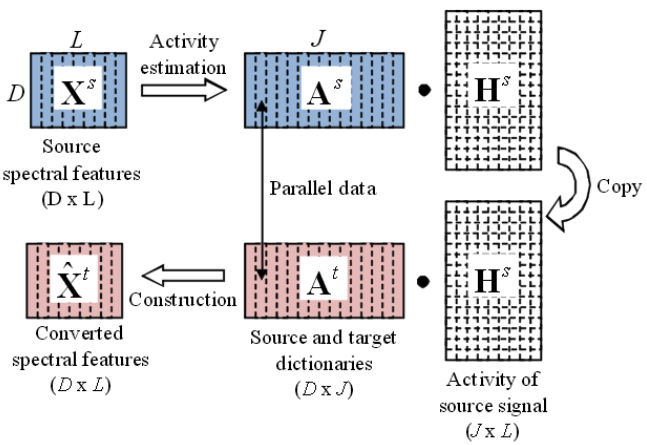


図 1. NMF による声質変換（シングルモーダル）

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．

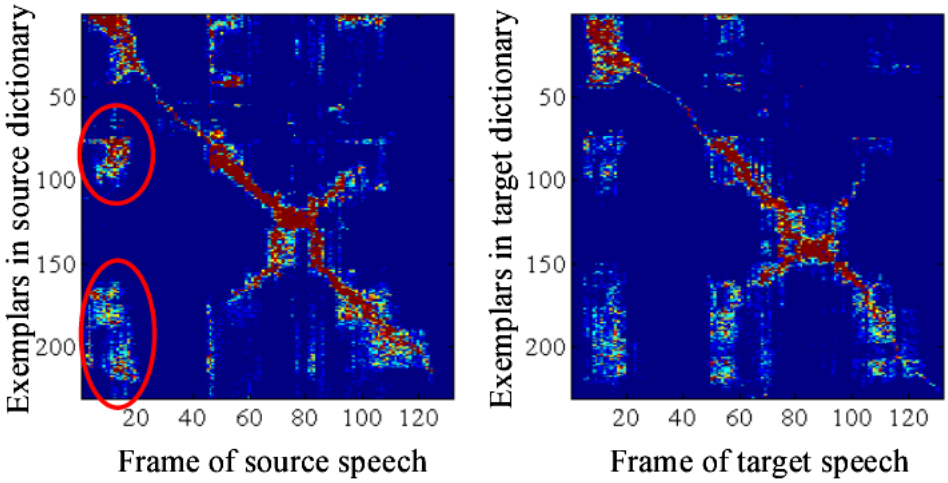


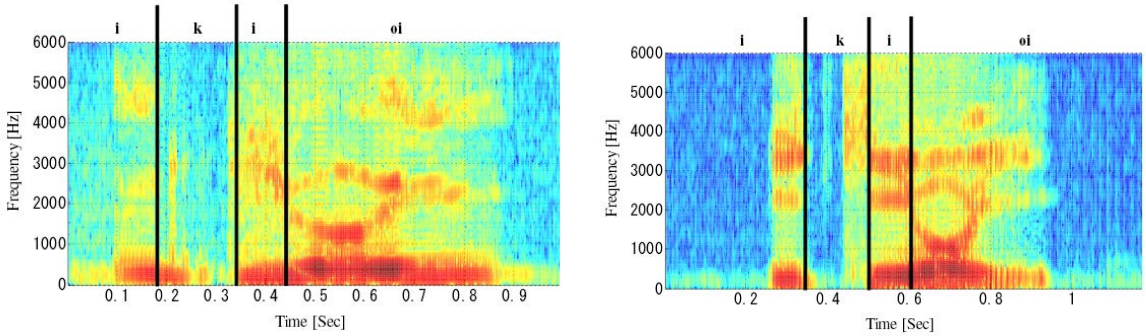
図 2. 入力信号のアクティビティ行列と出力信号のアクティビティ行列

図 2 は入力話者と出力話者がそれぞれ/ikioi/と発話した音声に対して,それらの音声のフレーム間対応を取ったものを辞書にし，NMF によりアクティビティ行列を推定したものである．このとき，入力話者の音声には入力話者の辞書を，出力話者の音声には出力話者の辞書を用いて，それぞれのアクティビティ行列を求めている．また，入力/出力音声の特徴量及び辞書内のサンプルは STRAIGHT 分析によって得られる平滑化スペクトル（STRAIGHT スペクトル）である．この実験では入力/出力音声と辞書が同じ単語であるため，得られるアクティビティ行列は対角線上に高いエネルギーを持つ．また，図中の赤線囲まれた箇所のような，対角線上から離れた場所にエネルギーが現れているのは,単語中に“i”が 3 回含まれているため,辞書内の“i”に対応する区間も 3 箇所存在するためである．

図より，入力話者音声と出力話者音声それぞれから得られるアクティビティ行列は，似た位置に高いエネルギーを持っていることが分かる．この理由から，入力話者の辞書と出力話者の辞書が同一発話でフレーム対応が取れている，つまりパラレル辞書になっているとき，入力音声から得られるアクティビティ行列は，出力音声のアクティビティ行列に代用可能であると考えられる．この考えに基づくと，図 1 のように入力音声のアクティビティ行列と出力音声辞書の内積から，出力音声を生成することが可能となる．

4 話者性を維持した NMF 声質変換

構音障がい者の聞き取りにくい発話を聞き取りやすい声に変換するには，話者変換を用いて障がい者の発話音声を健常者の発話へと変換することも考えられるが，構音障がい者のなかには「自分らしい声で話したい」というニーズがあり，障がい者の話者性を維持した声質変換が求められている．文献[4]において，我々は非負値行列因子分解(Non-negative Matrix Factorization: NMF)を用いた構音障がい者のための，話者性を維持した声質変換を提案した．この手法では，アテトーゼ型脳性麻痺による構音障がい者の発話特徴である子音が不安定になりやすいという性質を利用し，入力辞書に障がい者発話，出力辞書に障がい者の母音と健常者の子音とを組み合わせた Combined-dictionary を用いることで，障がい者の話者性を維持した変換を実現した．



(a) 構音障がい者の音声スペクトル　　(b) 健常者の音声スペクトル

図 3.　音声スペクトル（/i k i oi/）

図 3(a)に構音障がい者による発話スペクトル，(b)に健常者による発話スペクトルを示す．障がい者の発話例では，子音/k/に当たる部分のパワーが健常者と比較して弱くなっていることがわかる．その他の母音部分に関しては，障がい者と健常者の間において，スペクトルの違いはあまり見られない．以上より，アテトーゼ型脳性麻痺による構音障がい者の発話が聞き取りにくい原因は，子音にあると考えることができる

そこで，障がい者の話者性を維持するため，アライメントをとったパラレルなスペクトル包絡から健常者の子音と障がい者の母音を組み合わせて出力特徴量し，この出力特徴量を出力辞書に用いる．これらの処理を全ての同一内容発話について行い，抽出した特徴量を入力・出力それぞれについて水平に結合することで辞書行列とする．このような健常者の子音と障がい者の母音から構成される出力辞書行列を Combined Dictionary と呼ぶ.

図 1 において,出力話者辞書 A^t を Combined Dictionary におきかえる.入力された障がい者スペクトルは,障がい者の発話スペクトルから構成された入力話者辞書の基底の線形結合とその重みで表現される．重み行列 H^s は健常者の子音と障がい者の母音から構成される出力辞書行列 A^t と掛け合わされる．このとき，入力された障がい者スペクトルは障がい者の母音基底と健常者の子音基底の線形結合で表現され，出力スペクトルは，入力スペクトルのうち，子音フレームのみが健常者のものに変換されることになる．本実験では，マルチモーダルの効果を確認するため，3 章で説明した NMF 声質変換を用いることにした.

5 辞書選択による NMF 声質変換

我々はこれまでに NMF 声質変換の改良版を提案している.その一つに辞書選択による NMF 声質変換がある.パラレルデータの全フレームをそのまま辞書の基底として用いた場合,辞書のサイズが膨大となってしまう.そのため,入力音声のフレームと,入力話者辞書から選ばれる基底の音素の一致精度が劣化する場合がある.そこで，入力・出力話者辞書を音素カテゴリに分けた副辞書を作成し，NMF を用いて音素カテゴリ認識を行った後，選択した副辞書上でマッピングを行うことで声質変換を行うことが考えられる[7].

以下簡単に述べると，まず障がい者の話者性を維持するため，音素カテゴリに分けられた副辞書のうち，母音の出力話者副辞書は，障がい者の発話スペクトルから構成される入力話者副辞書と同じものを用いる．さらに，それぞれの副辞書を表現する代表基底を集め，副辞書を選択するためのカテゴライズ辞書を作成する．それぞれの代表基底は，副辞書において仮定した Gaussian Mixture Model (GMM)の平均ベクトルから構成される．正規分布の混合数は，対応する副辞書の基底数に応じて定めるため，副辞書を代表する基底の数も副辞書ごとに異なる[7]．今回，辞書選択 NMF を用いた実験は行われていないが，今後検討していく予定である.

6 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．本研究では，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を提案する．これによってより雑音に頑健な変換となる．

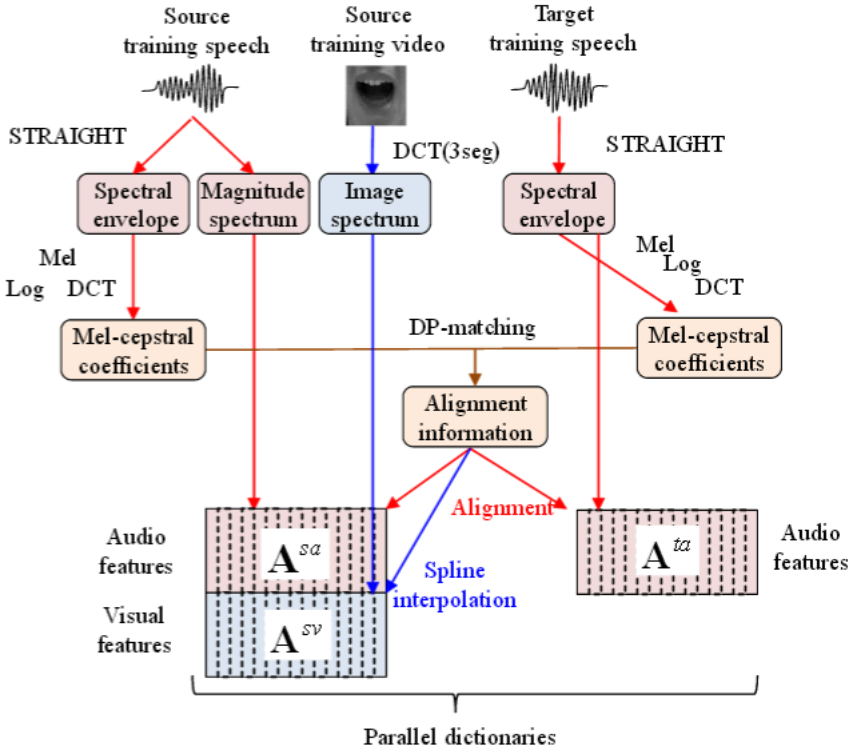


図 4. マルチモーダル辞書の構築

6-1 辞書構成法

図 4 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辞書 A^{sa} と出力話者辞書 A^{ta} を求める．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 [5]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 A^{sv} に関して，フレームごとに得られる画像から特徴量を取り出し並べ，スプライン補間を行い音声フレームとの同期を取っ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[6]を行い，低次 40 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 A^{s} とし，変換に用いるものとする．

6-2 変換手法

本研究で用いる変換手法は 2 段階に分かれている．1 段階目は従来の雑音除去 NMF と同様，入力音声からノイズを除去する．入力話者の音声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，クリーン音声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．

$$x = x^a + x^n = \begin{bmatrix} A^{sa} & A^{sn} \end{bmatrix} \begin{bmatrix} h^a \\ h^n \end{bmatrix} = A^{sa} h$$

x^a と x^n はそれぞれ入力話者のクリーン音声の振幅スペクトル，雑音の振幅スペクトルを表す．A^{sa}，A^{sn}，h^a，h^n は入力話者の音声辞書，雑音の辞書，クリーン音声，雑音に対するそれぞれのアクティビティを表す．本手法ではスペクトル形状のみを考慮するため，スペクトル，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.スパース制約付き NMF においてh を推定するためにコスト関数が以下のように設定されている．

$$d(x, A^{sa}h) + \left\| 1^{(J\times 1)} \lambda^T Ih \right\|_1$$

第一項は x と Ah の Kullback-Leibler divergence である．第二項は h をスパースにするための L1 ノルム正則化項である．1 はすべての要素が 1 の行列，I は単位行列を表す．λ 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．上式を最小にするアクティビティ行列 h を求めればよい．

2 段階目は，1 段階目で得られた振幅スペクトルに画像特徴量を結合する．これを入力として，音声画像結合辞書 A^s からアクティビティ行列を推定する．ここで，入力，辞書にそれぞれ音声と画像の重み，α と β を掛けることにする．このとき，音声と画像を結合した特徴量から得られる新たな h^{av} は以下のコスト関数を最小にすることで推定できる．

$$\alpha d(x^a, A^{sa}h^{av}) + \beta d(x^v, A^{sv}h^{av}) + \left\| 1^{(J\times 1)} \lambda^T Ih^{av} \right\|_1$$

7 評価実験

7-1 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．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．また，今回はハイスピードカメラによる結果も報告するため，入力話者に健常者の発話を用いることにした（ハイスピードカメラの収録には多くの時間がかかるため，被験者の体力等を考慮して，今回は健常者の発話を用いることにした)．

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は 1 に固定し，画像に対する重みを 1 から 10 の間で変化させて評価を行った．

7-2 実験結果

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 5 に示す．

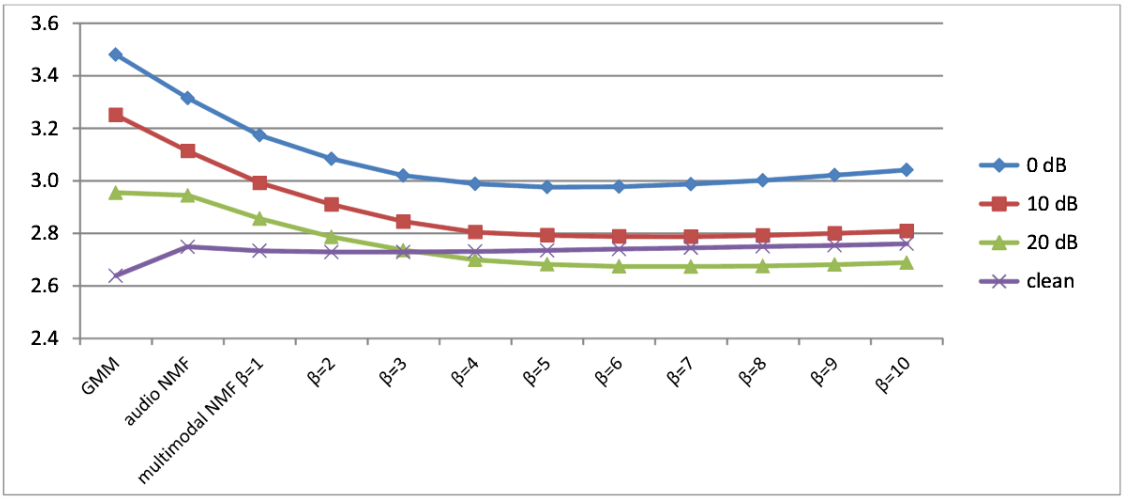


図 5. 声質変換音声のメルケプストラム歪（ホワイトノイズ環境下）

実験結果より雑音環境下において，提案手法の結果(multimodalNMF)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 = 0 dB において，マルチモーダル NMF の一番良い MCD の値は，β= 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 = 20 dB においては，β= 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，画像の重みは 5 となっている．

表 1. 各雑音下におけるメルケプストラム歪

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，画像の重みは 5 となっている．

$$MelCDR = \frac{\sqrt{\sum_{d=1}^{24}\left(mc_d^t - mc_d^s\right)^2}}{\sqrt{\sum_{d=1}^{24}\left(mc_d^t - \hat{mc}_d^t\right)^2}}$$

ここで mc_d^s は入力音声の d 次元目の係数，mc_d^t は出力音声の d 次元目の係数である．

表 2. 母音と子音におけるメルケプストラム歪比

	Audio NMF	Multimodal NMF
Vowel	1.508	1.620
Consonant	1.477	1.676

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．

7-3 ハイスピードカメラ画像を用いた声質変換

通常のカメラでは，1 秒間あたり 30 枚の画像が録画される．一方，今回の実験で使用している音声では 1 秒間あたり 8,000 の値が保存されている．よって，画像のサンプル間隔は約 33 ミリ秒，音声のサンプル間隔は 0.1 ミリ秒となる．画像サンプル数が音声に対して非常に少ないため，スプライン補間を行い，(見かけ上) 同じサンプル数にすることになる．

一つの子音の継続時間長がせいぜい 50 ミリ秒程度であるため，通常のカメラでは子音に対しては 1 枚しか録画されないことになる (音声は十分のサンプル数が録音されている)．そこで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．今回の実験ではハイスピードカメラを用いて 1 秒間あたり 4,000 枚の録画を行ってみた．各手法における変換音声の SDIR (Spectral Distortion Improvement Ratio) を図 5，6 に示す．SDIR は以下の式で表される．

$$SDIR[dB] = 10\log\frac{\sum|X^t(d) - X^s(d)|^2}{\sum|X^t(d) - \hat{X}^t(d)|^2}$$

ここで，X^s，X^t，$\hat{X}^t$ はそれぞれ入力話者音声スペクトル，出力話者音声スペクトル，変換後の音声スペクトルを表す．

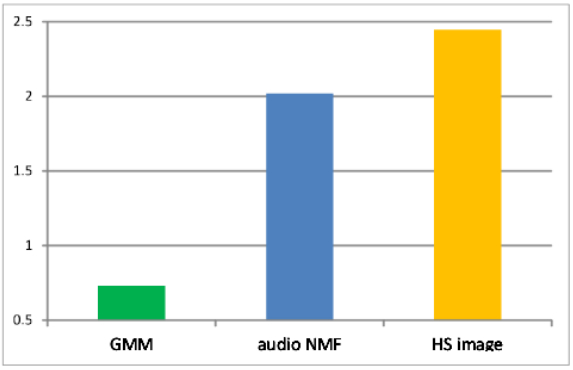


図 6. ハイスピードカメラ画像を用いた結果

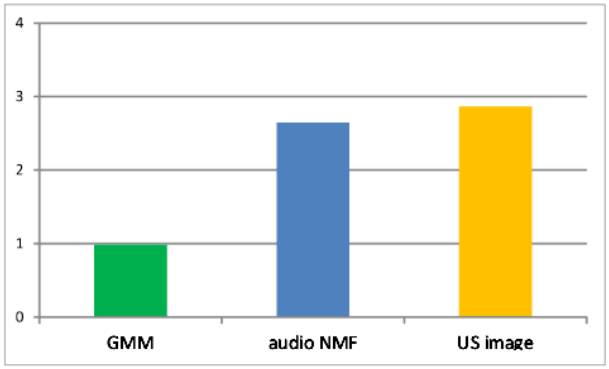


図 7. 一般カメラ画像を用いた結果

図 6, 7 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．実験結果より，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．

8 おわりに

本研究では，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，構音障がい者発話の評価，より精度の高い変換手法の改良を進めていく．

【参考文献】

1. S. T. Canale and W. C. Campbell, “Campbell’s operative orthopaedics,” Tech. Rep., Mosby Year Book, 2002.
2. Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki, “Individuality-preserving Voice Conversion for Articulation Disorders Using Dictionary Selective Non-negative Matrix Factorization,” Workshop on Speech and Language Processing for Assistive Technologies, pp. 29-37, June 2014.
3. Toshiya Yoshioka, Tetsuya Takiguchi, and Yasuo Ariki, “Robust Feature Extraction to Utterance Fluctuation of Articulation Disorders Based on Random Projection,” Workshop on Speech and Language Processing for Assistive Technologies, pp. 129-133, 2013.
4. Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, “Exemplar-based Individuality-Preserving Voice Conversion for Articulation Disorders in Noisy Environments,” INTERSPEECH, pp. 3638-3641, 2013.
5. 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.
6. J. Almajai and B. Milner, “Using audio-visual features for robust voice activity detection in clean and noisy speech," EUSIPCO, 2008.
7. 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 年秋季研究発表会論文集, 3-7-6, pp. 1473-1476, 2013.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
非負値行列因子分解に基づく唇動画像からの音声生成	日本音響学会2015年春季研究発表会講演論文集	2015 年 3 月