

辞書選択型非負値行列因子分解による構音障害者の声質変換

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: [†]aihara@me.cs.scitec.kobe-u.ac.jp, ^{††}{takigu,ariki}@kobe-u.ac.jp

あらまし 本研究では、アテトーゼ型脳性麻痺による構音障害者を対象とし、筋肉の不随意運動を原因とする障害者の不安定な発話を聞き取りやすく変換することを目指す。従来の声質変換手法で最も一般的なのは、混合正規分布モデル (GMM) を用いた統計的手法であった。この手法は主に話者変換を目的として研究されてきたため、GMM 声質変換を構音障害者の発話音声に適用し健常者の音声に変換した場合、障害者の話者性は別人のものに置き換わってしまう。「自分らしい声で話したい」という障害者のニーズに答えるため、本研究では従来の統計的モデルによる声質変換とは異なる、非負値行列因子分解 (NMF) を用いた Exemplar-based 声質変換を用いて、話者性を維持しつつ聞き取りやすい音声に変換する。これまで NMF 声質変換では、入力音声フレームと、辞書から選ばれる基底の音素が必ずしも一致しないという問題があった。本研究では、この問題を解決するため、NMF を用いて音素カテゴリ認識を行い、音素カテゴリに分割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。提案手法の有効性を評価するため、従来の GMM 声質変換、NMF 声質変換との比較実験を行った。

キーワード 声質変換, 構音障害者, 障害者支援, 非負値行列因子分解, 辞書選択

Voice Conversion for Articulation Disorders Using Dictionary Selective Non-negative Matrix Factorization

Ryo AIHARA[†], Tetsuya TAKIGUCHI^{††}, Yasuo ARIKI^{††}, and

[†] 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: [†]aihara@me.cs.scitec.kobe-u.ac.jp, ^{††}{takigu,ariki}@kobe-u.ac.jp

Abstract We present in this paper a voice conversion (VC) method for a person with an articulation disorder resulting from athetoid cerebral palsy. The movement of such speakers is limited by their athetoid symptoms, and their consonants are often unstable or unclear, which makes it difficult for them to communicate. In our previous method, exemplar-based spectral conversion using Non-negative Matrix Factorization (NMF) was applied to a voice with an articulation disorder. To preserve the speaker's individuality, we used a combined dictionary that is constructed from the source speaker's vowels and target speaker's consonants. However, this exemplar-based approach needs to hold all the training exemplars (frames), and it may cause mismatching of phonemes between input signals and selected exemplars. In this paper, in order to reduce the mismatching of phoneme alignment, we propose a phoneme-categorized sub-dictionary and a dictionary selection method using NMF. By using the sub-dictionary, the performance of VC is improved compared to a conventional NMF-based VC. The effectiveness of this method was confirmed by comparing its effectiveness with that of a conventional Gaussian Mixture Model (GMM)-based method and a conventional NMF-based method.

Key words Voice Conversion, Articulation Disorders, Assistive Technology, Non-negative Matrix Factorization, Sub-dictionary

1. はじめに

現在、日本だけでも約3万4千人の言語・聴覚障害者がおり、言語障害の原因の一つとして脳性麻痺をあげることができる。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、そのほかの神経障害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。また、脳性麻痺は脳の損傷部分によって痙直型（大脳皮質）、アテトーゼ型（中脳もしくは脳基底核）、失調型（小脳）、混合型（脳の広範囲）に分類される[1]。それぞれ、痙直型は正常な筋の伸張反射が過度になる、アテトーゼ型はアテトーゼと呼ばれる筋肉の不随意運動を伴う、失調型は協調運動の障害が現れ、混合型はそれぞれの症状が混合して現れるというような症状が見られる。

本論文では、アテトーゼ型の脳性麻痺による構音障害者を対象としている。アテトーゼ型は、脳性麻痺患者の約20%に発生する。筋肉の随意運動や姿勢の調整を行っている大脳基底核（大脳皮質、視床や脳幹を結び付けている神経核の集まり）に損傷を受けたことにより、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。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、この運動障害の一つとして、正しく構音できない場合がある。症状は軽度から重度まで様々であり、知能障害を合併していないケースや比較的知能障害の程度が軽いケースも多いのが特徴である。また、アテトーゼ型脳性麻痺による構音障害者の多くは、身体が不自由であるため、手話や文章読み上げ装置を使うことは困難である。そのため、構音障害者のための声質変換には十分なニーズがあり、研究の必要性があるといえる。これまで、声質変換技術は話者変換[2]、感情変換[3]、あるいは無喉頭音声変換[4]に応用されてきたが、アテトーゼ型の構音障害者への応用は研究が進んでいない。

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[2]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。基本的には、変換関数を目標話者と入力話者のスペクトル包絡の期待値によって表現し、変数をパラレルな学習データから最小二乗法で推定する。中村ら[4]はGMMを用いた声質変換手法を喉頭摘出者に応用し、電気式人工喉頭を用いた発話を自然性の高い音声へと変換した。しかしながら、GMMによる声質変換は主に話者変換をタスクとして研究されてきたため、この手法を構音障害者に適用し、健常者の声質へと変換した場合、音声は聞き取りやすく変換されると考えられるが、構音障害者音声の話者性は完全に別の健常者の話者性へと置き換えられてしまう。構音障害者のなかには「自分らしい声で話したい」というニーズがあり、障害者の話者性を維持した声質変換が求められている。

文献[5]において、我々はアテトーゼ型脳性麻痺による構音

障害者のための、話者性を維持した声質変換を提案した。この手法では、従来の声質変換手法で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（入力辞書）と出力話者の音声辞書（出力辞書）からなる同一発話内容のパラレル辞書を構築する。変換の際は非負値行列因子分解 (Non-negative Matrix Factorization: NMF) [6] を用いて、入力音声を入力辞書のスパース表現にする。そして得られた入力辞書内のサンプル毎の重み係数（アクティビティ）に基づいて、出力辞書内のサンプルを線形結合することで、出力話者の音声スペクトルへと変換する。この手法では、アテトーゼ型脳性麻痺による構音障害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、入力辞書に障害者発話、出力辞書に障害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、障害者の話者性を維持した変換を実現した。

さらに文献[7]において我々は、NMFを用いた声質変換手法の精度を向上させるため、辞書選択手法の導入を提案した。これまで、我々の手法では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ており、辞書のサイズが膨大となっていた。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が必ずしも一致しないといった問題があった。そこでこの手法においては、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った。この手法は、声質変換の一般的なタスクである健常者の話者変換において、従来の声質変換手法と比較して自然性の高い音声に変換できることが示された。

本研究では、アテトーゼ型構音障害者を対象として、辞書選択を用いた NMF 声質変換による話者性を維持した声質変換を提案する。出力話者のカテゴリ辞書のうち、子音に関するカテゴリ辞書のみに健常者のスペクトルを用い、母音に関するカテゴリ辞書に障害者のスペクトルを用いることで、障害者の話者性を維持した声質変換を行う。

以下、第2章で従来の NMF 声質変換手法を説明しその問題点を明らかにする。第3章で本稿の提案手法を述べるた後、第4章で従来の GMM・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

2.1 概要

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。

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

$\mathbf{x}_l$ 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

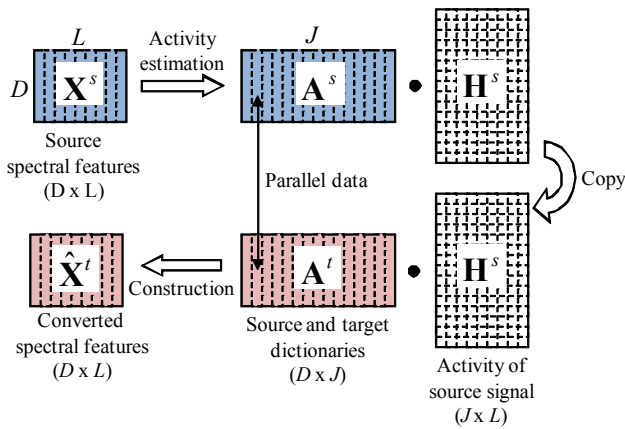


図 1 NMF を用いた声質変換の概要

Fig. 1 Basic approach of NMF-based voice conversion

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。

$$\mathbf{X} \approx \mathbf{A}\mathbf{H} \quad (2)$$

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

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辞書 $\mathbf{A}^s$ と出力話者辞書 $\mathbf{A}^t$ 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。Fig. 2 は 2 人の話者による発話単語 “ikiol” を入力信号としたときの、NMF で推定したアクティビティ行列である。ただし、簡易な例を示すため、辞書行列は 2 話者間でアライメントがとられた 1 単語のみから構成されている。Fig. 2 を見ると、高い重みを持つ基底の位置が類似していることがわかる。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 [8]。

2.2 Combined Dictionary による話者性の維持

パラレルな辞書を構成するため、障害者と健常者による複数の同一内容発話を用意する。同一内容発話から抽出されたスペクトル包絡は、DP マッチングによってアライメントをとる。アクティビティを正確に推定するため、前後数フレームをまと

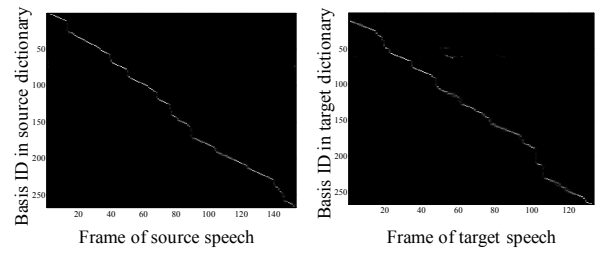


図 2 パラレル発話から推定されたアクティビティ行列

Fig. 2 Activity matrices for parallel utterances

めて 1 本のベクトルにしたセグメント特徴量を求めて入力特徴量とする。

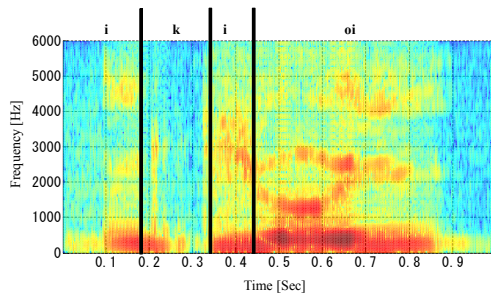
Fig. 3(a) に構音障害者による発話スペクトル、(b) に健常者による発話スペクトルを示す。障害者の発話例では、子音 ‘k’ に当たる部分のパワーが健常者と比較して弱くなっていることがわかる。その他の母音部分に関しては、障害者と健常者の間において、スペクトルの違いはあまり見られない。以上より、アテトーゼ型脳性麻痺による構音障害者の発話が聞き取りにくい原因は、子音にあると考えることができる。そこで、障害者の話者性を維持するため、アライメントをとったパラレルなスペクトル包絡から健常者の子音と障害者の母音を組み合わせる出力特徴量し、この出力特徴量を出力辞書に用いる。これらの処理を全ての同一内容発話について行い、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。本論文ではこのような健常者の子音と障害者の母音から構成される出力辞書行列を Combined Dictionary と呼ぶ。

Fig. 1 において、出力話者辞書 $\mathbf{A}^t$ を Combined Dictionary におきかえる。入力された障害者スペクトルは、障害者の発話スペクトルから構成された入力話者辞書 $\mathbf{A}^s$ の基底の線形結合とその重みで表現される。重み行列 $\mathbf{H}^s$ は健常者の子音と障害者の母音から構成される出力辞書行列 $\mathbf{A}^t$ と掛け合わされる。このとき、入力された障害者スペクトルは障害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 $\hat{\mathbf{X}}^t$ は、入力スペクトル $\mathbf{X}^s$ 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。

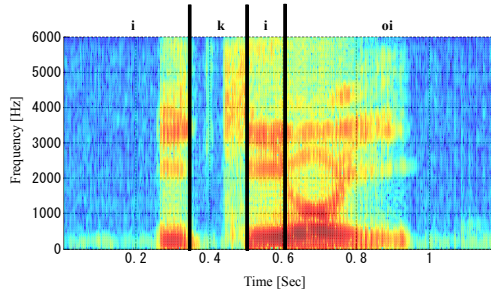
2.3 問題点

Fig. 3(c) に Combined-dictionary を用いた障害者音声の NMF 声質変換結果を示す。子音 ‘k’ に当たる部分のパワーが強調され、健常者の子音に近くなっていることがわかる。しかしながら、変換スペクトルには、子音 ‘k’ にづく母音 ‘i’、‘o’ の高域に、障害者スペクトルにも健常者スペクトルにもみられなかったパワーが出現している。

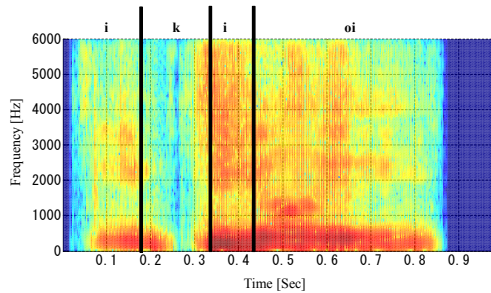
この原因として考えられるのは、辞書行列の基底数が多すぎることである。この手法では、学習サンプル全てを基底としてパラレルな辞書を構成していた。その結果、辞書に含まれる基底数が膨大になり、入力信号が値の小さなアクティビティから成る多数の基底の線形結合で表現されてしまう。障害者の母音部分であっても、多数の基底の線形結合で表現された結果、変換音声の母音部に障害者の母音と健常者の子音が混在してし



(a) Spoken by a person with an articulation disorder



(b) Spoken by a physically unimpaired person



(c) Converted by NMF-based VC

図3 障害者・健常者・変換後のスペクトル “ikioi”

Fig. 3 Examples of spectrogram /i k i oi

まっていると考えられる。Fig. 3(c)の母音‘i’、‘o’の誤変換は、健常者の子音と障害者の子音が混ざってしまったために発生したと見ることができる。さらに、線形結合の基底の数が増加すると、Fig. 2で示した平行な発話のアクティビティの形状が類似するという仮定が成り立たなくなり、このことが変換音声全体の劣化につながっていると考えられる。

そこで、本稿では入力信号を表現する基底数を限定するため、音素カテゴリごとに副辞書を作成し、限定された基底内でスパースコーディングを行う、辞書選択によるNMF声質変換を提案する。

3. 辞書選択によるNMF声質変換

3.1 辞書構成法

Fig. 4は副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障害者の発話スペクトルから構成される入力話者辞書 $\mathbf{A}^s$ と、健常者の発話スペクトルから構成される出力話者辞書行列 $\mathbf{A}^t$ を求める。次に平行な辞書行列は、Table 1に示す音素カテゴリに従って、 K 個の副辞書に分けられる。

障害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、障害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リ化辞書 Θ 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。

$$p(\mathbf{x}_n^{(k)}) = \sum_{m=1}^{M_k} \alpha_m^{(k)} N(\mathbf{x}_n^{(k)}; \boldsymbol{\mu}_m^{(k)}, \boldsymbol{\Sigma}_m^{(k)}) \quad (4)$$

ここで、 M_k 、 $\alpha_m^{(k)}$ 、 $\boldsymbol{\mu}_m^{(k)}$ 、 $\boldsymbol{\Sigma}_m^{(k)}$ はそれぞれ k 番目の副辞書における正規分布の混合数、 m 番目の正規分布の混合重み、平均、分散である。それぞれのパラメータはEMアルゴリズムを用いて推定する。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。

カテゴリ化辞書の基底は、それぞれの副辞書と対応し、下のよう表現される。

$$\Phi_k^s = [\mathbf{x}_1^{(k)}, \dots, \mathbf{x}_{N_k}^{(k)}] \quad (5)$$

$$\boldsymbol{\theta}_k = [\boldsymbol{\mu}_1^{(k)}, \dots, \boldsymbol{\mu}_{M_k}^{(k)}] \quad (6)$$

$$\Theta = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K] \quad (7)$$

ここで、 Φ_k^s 、 N_k 、 Θ はそれぞれ、 k 番目の副辞書、 k 番目の副辞書の基底数、カテゴリ化辞書を表す。

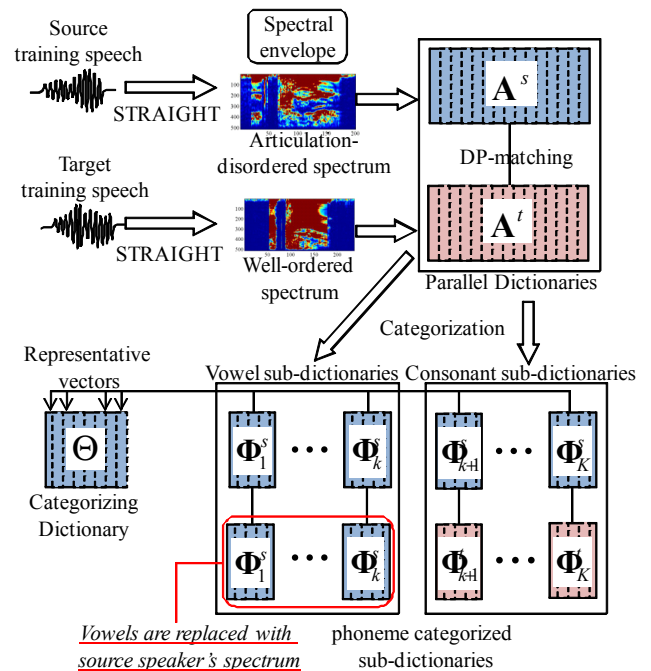


図4 副辞書の構成法

Fig. 4 Making sub dictionary

3.2 辞書選択・変換法

Fig. 5は変換手法の流れを示したものである。変換する入力信号 $\mathbf{X}^s$ はカテゴリ化辞書 Θ とアクティビティ $\mathbf{H}_\Theta^s$ によって、カテゴリ化辞書の基底とその重みの線形和で表される。

$$\mathbf{X}^s \approx \Theta \mathbf{H}_\Theta^s \quad s.t. \quad \mathbf{H}_\Theta^s \geq 0 \quad (8)$$

$$\mathbf{X}^s = [\mathbf{x}_1^s, \dots, \mathbf{x}_L^s] \quad (9)$$

$$\mathbf{H}^s_{\Theta} = [\mathbf{h}^s_{\Theta_1}, \dots, \mathbf{h}^s_{\Theta_L}] \quad (10)$$

$$\mathbf{h}^s_{\Theta_l} = [\mathbf{h}^s_{\Theta_{1l}}, \dots, \mathbf{h}^s_{\Theta_{Kl}}]^T \quad (11)$$

$$\mathbf{h}^s_{\Theta_{kl}} = [h^s_{\Theta_{1l}}, \dots, h^s_{\Theta_{Mkl}}]^T \quad (12)$$

つづいて入力信号のフレーム $\mathbf{x}_l^s$ は、対応する入力話者副辞書 Φ_k^s が以下のように選ばれ、その副辞書内の基底の線形和で表現される。

$$\hat{k} = \arg \max_k \mathbf{1}^{1 \times M_k} \mathbf{h}^s_{\Theta_{kl}} = \arg \max_k \sum_{m=1}^{M_k} h^s_{\Theta_{ml}} \quad (13)$$

$$\mathbf{x}_l = \Phi_{\hat{k}}^s \mathbf{h}_{\hat{k},l} \quad (14)$$

選ばれた入力話者副辞書 $\Phi_{\hat{k}}^s$ が子音副辞書の場合、推定されたアクティビティと出力話者副辞書 $\Phi_{\hat{k}}^t$ により、変換された特徴量は以下のように表現できる。

$$\hat{\mathbf{y}}_l = \Phi_{\hat{k}}^t \mathbf{h}_{\hat{k},l} \quad (15)$$

一方、選ばれた入力話者副辞書 $\Phi_{\hat{k}}^s$ が母音副辞書の場合は、入力話者副辞書をそのまま用いる。

$$\hat{\mathbf{y}}_l = \Phi_{\hat{k}}^s \mathbf{h}_{\hat{k},l} \quad (16)$$

表 1 Sub-dictionary categories

	Category	phoneme
vowels	a	a
	e	e
	i	i
	o	o
	u	u
consonants	plosives	p, t, k, b, d, g
	fricatives	s, h, j, z,
	nasals	m, n, N
	liquid	r

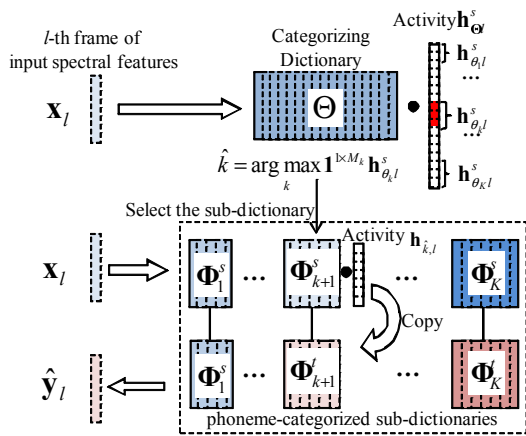


図 5 音素カテゴリ副辞書を用いた NMF 声質変換

Fig. 5 NMF-based voice conversion using categorized dictionary

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、以前に提案したパラレルデータ全てを辞書に用いる手法と比較を行った。入力話者として障害者の音声として使用するため、男性のアテローゼ型構音障害者 1 名による 432 発話を収録した。発話内容は ATR 音素バランス単語 B セット [9] から 216 語を用いた。対となる健常者音声は、ATR 音声データベースに収録されている男性話者のものを使用した。それぞれの音声のサンプリング周波数は 16kHz、フレームシフトは 5ms である。対となったパラレルデータのうち、216 発話を学習に、残りの 216 発話をテストに用いた。入力特徴量、出力特徴量の次元数はそれぞれ 2565 次元と 513 次元である。パラレルデータ間の時間的なゆらぎを解消するため、STRAIGHT スペクトル [10] から求めたメルケプストラム係数を用いて DP マッチングを行った。

GMM を用いた従来手法では、STRAIGHT スペクトルから計算された mfcc+ Δ mfcc+ $\Delta\Delta$ mfcc の 64 次元を特徴量とした。GMM の混合数は 64 である。なお、本実験では F_0 は変換せず、障害者のものをそのまま用いた。

結果評価のために、成人男女 10 名による、聴取実験を行った。評価項目は、聞き取りやすさ (listening intelligibility)、話者性 (similarity)、自然性 (naturalness) の 4 項目とした。「聞き取りやすさ」との評価にはテストデータから構音障害者が発話しにくい 22 単語を選び、提案手法と従来手法それぞれで変換した音声と無変換の障害者音声を評価した。評価基準は MOS 評価基準に基づく主観評価 (5:とてもよい, 4:よい, 3:ふつう, 2:わるい, 1:とてもわるい) とした。「話者性」「自然性」の評価には、テストデータから 50 単語をランダムに選び、提案手法と従来手法それぞれで変換した。「話者性」の評価では、無変換の障害者の音声を聴いた後、各変換音声を聴き比べてどちらが障害者の性質に似ているかを選択する XAB 法とした。「自然性」の評価では、各変換音声を聴き比べてどちらが障害者の性質に似ているかを選択する 1 対 1 の対比較法としたいずれの評価項目も、静かな部屋においてヘッドホンを用いた両耳聴取を行った。

4.2 実験結果・考察

Fig. 6(a) に提案手法で変換したスペクトルの例、(b) に従来の GMM 声質変換で変換したスペクトルの例を示す。Fig. 3(c) と比較すると、母音部の誤変換がなくなっていることがわかる。Fig. 6(b) と比較しても、母音部が明瞭に変換できていることが確認できる。

「聞き取りやすさ」の主観評価結果を Fig. 7 に示す。提案手法は、無変換の障害者音声と比較して、聞き取りやすさを向上させている一方、従来の GMM による声質変換は無変換音声と比較して劣化している。これは、変換ノイズによるものと考えられる。提案手法も変換ノイズを発生させるものの、GMM に基づくものよりは少ない変換ノイズとなっている。

「話者性」の主観評価結果を Fig. 7 に示す。提案手法は、従来の GMM、NMF による手法と比較して話者性が維持できて

いる。従来の NMF による手法も GMM による手法と比較して話者性を維持している。「自然性」の主観評価結果を Fig. 9 に示す。提案手法は、従来の GMM, NMF による手法と比較して、高い自然性を示している。

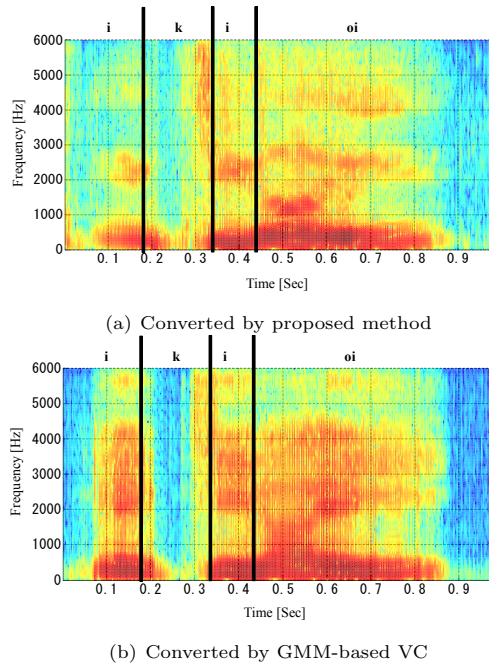


図 6 変換後のスペクトル “ikioi”
Fig. 6 Examples of converted spectrogram //i k i oi

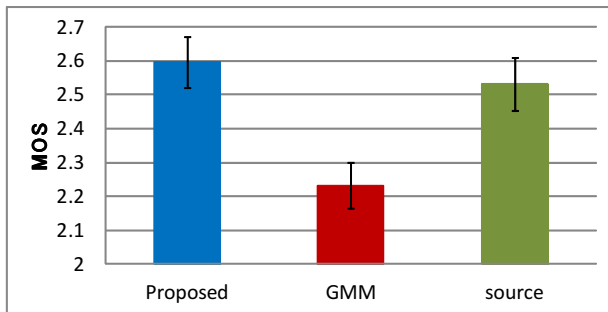


図 7 Results of MOS test on listening intelligibility

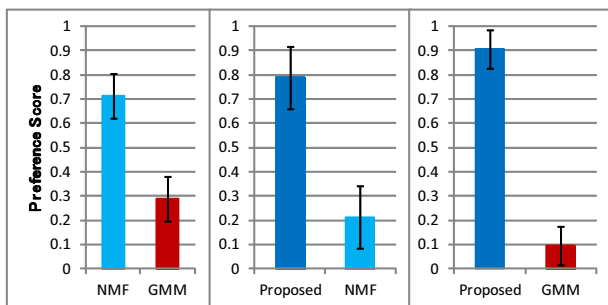


図 8 Preference scores for the individuality

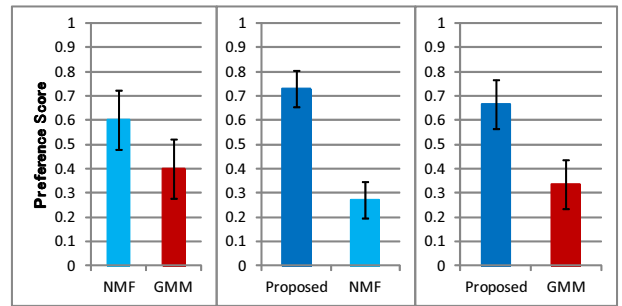


図 9 Preference scores for the naturalness

5. おわりに

本論文では、アテトーゼ型構音障害者を対象とした話者性を維持した声質変換技術を提案した。これまで提案してきた NMF に基づく声質変換に辞書選択を導入し、精度の向上を目指した。聴取実験によって、提案手法は構音障害者の話者性を維持しつつ聞き取りやすさを向上することを示した。さらに、従来手法と比較して、提案手法は自然性の高い音声で変換できることを示した。本実験では対象とした構音障害者は 1 名にとどまっているため、今後は話者数を増やして提案手法の有効性を確認する予定である。

文 献

- [1] S.T. Canale and W.C. Campbell, “Campbell’s operative orthopaedics,” Technical report, Mosby-Year Book, 2002.
- [2] 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.
- [3] C. Veaux and X. Robet, “Intonation conversion from neutral to expressive speech,” in Interspeech, pp.2765–2768, 2011.
- [4] 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.
- [5] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, “Individuality-preserving voice conversion for articulation disorders based on Non-negative Matrix Factorization,” in ICASSP, pp.8037–8040, 2013.
- [6] D.D. Lee and H.S. Seung, “Algorithms for non-negative matrix factorization,” Neural Information Processing System, pp.556–562, 2001.
- [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, 3-7-6, pp.1473–1476, 2013.
- [8] J.F. Gemmeke, T. Viratnen, and A. Hurmalainen, “Exemplar-based sparse representations for noise robust automatic speech recognition,” IEEE Trans. Audio, Speech and Language Processing, vol. 19, no. 7, pp.2067–2080, 2011.
- [9] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol.9, pp.357–363, 1990.
- [10] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequencybased F0 extraction: possible role of a repetitive structure in sounds,” Speech Communication, vol.27, no.3-4, pp.187–207, 1999.