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
EA2016-137,SIP2016-192,SP2016-132 (2017-03)

構音障害者のためのDurationを含んだ統計的話者変換

相原　龍　滝口　哲也　有木　康雄

神戸大学大学院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
E-mail: aihara@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

あらまし　音声に含まれる音韻性を維持しつつ，話者性を変換する技術に声質変換がある．近年，この技術を福祉分野に応用する研究が進められており，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する．アテトーゼ型脳性麻痺による構音障害者の多くは，運動障害の結果として構音と同時に手足にも障害があり，彼らの不安定な発話を聞き取りやすく変換する技術が求めれている．健常者の声質変換で一般的な混合正規分布モデル (GMM) や非負値行列因子分解 (NMF) による変換が障害者にも応用されていたが，スペクトル特徴量を中心とした変換が多かった．構音障害者の場合，健常者と比較してDurationが長くなる傾向があると同時に，発話リズムの乱れによって構音が不安定になる例があり，この現象が発話を聞き取りにくくする原因となっていた．そこで，本研究では，障害者発話のDurationを健常者発話のDurationへと変換することで，聞き取りやすさの向上を目指す．障害者発話は音素前後関係によってDurationが様々に変化するため，単純な線形変換では対応が困難であると考えられる．また，声質変換は明示的な音声認識を行わないため，音声合成モデルで用いられるDurationモデルの構築は難しい．そこで本研究では，従来の統計的声質変換モデルで変換可能なフレームベースのDuration特徴量を提案する．障害者と健常者の間のパラレル発話間のマッチング距離をDuration特徴量とし，入力話者スペクトル特徴量から出力話者のDurationを推定することにより，フレームベースでDurationを変換することができる．本稿では，声質変換において最も一般的なGMMに基づく最尤変換を用いる．変換したDurationを健常者のものと比較することで，手法の有効性を確認した．
キーワード　声質変換，障害者支援，発話長，構音障害

Statistical Voice Conversion Including Duration for Dysarthric Speech

Ryo AIHARA, Tetsuya TAKIGUCHI, and Yasuo ARIKI

Graduate School of System Informatics, 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 movements of such speakers are limited by their athetoid symptoms, and their utterances are often unstable or unclear, which makes it difficult for them to communicate. Most of conventional VC methods convert spectral features and duration is not converted. In a case of dysarthric speech, its duration tends to be longer than non-dysarthric speech. Therefore, duration conversion is needed to convert dysarthric speech to non-dysarthric speech. In a case of F0 conversion, linear conversion is often adopted. However, it does not work well in case of duration conversion because the duration of dysarthric speech will be different depends on phoneme and the physical condition of the speaker. In a text-to-speech system, the duration is modeled by hidden Markov Model (HMM). However, in the case of VC, we cannot use HMM because phoneme labels of an input utterance are not given. This paper proposes duration conversion for dysarthric speech which converts dysarthric duration to non-dysarthric duration. Frame-wise duration features are proposed and it is converted by Gaussian mixture model (GMM) which is the most approved model in VC.
Key words　voice conversion, assistive technology, duration, dysathric speech

1. はじめに

近年，情報技術の福祉分野への応用が進んでいる．例えば，画像認識技術の応用による手話認識 [1]〜[3]，文章読み上げシステム [4]〜[6]，ウェアラブル音声合成システム [7] など，その応用領域は幅広い．本研究では，脳性麻痺による構音障害者に焦点をあて，彼らの不安定な発話音声を，聞き取りやすく変換することを目指す．

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 [8]．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，彼らのコミュニケーション支援システムとして，ハンズフリーなデバイスシステムが求められている．

文献 [9] において，我々は DCT にかわる，構音障害者の発話変動にロバストな特徴量抽出方法として主成分分析 (Principal Component Analysis : PCA) に基づく特徴量抽出を提案した．さらに，文献 [10] において我々は，構音障害者の音声認識率を改善するため，Multiple Acoustic Frames (MAF) を提案した．これらの研究の成果にも関わらず，構音障害者の音声認識率は健常者の認識率と比較して低い．健常者音声から構築した不特定話者モデルを用いた，構音障害者の音声認識率は 3.5%であり，この結果は構音障害者の発話スタイルが健常者と大きく異なっていることを示している．

現在，音声信号処理分野で研究されているハンズフリーな音声デバイスシステムに，声質変換がある．声質変換は，音声に含まれている音韻性を維持しながら話者性を変換する技術であり，話者変換をタスクとして研究が進められてきた [11]．また，この技術は感情変換 [12] や帯域幅拡張 [13] などに応用されてきた．

なかでも混合正規分布モデル (Gaussian Mixture Model : GMM) を用いた手法 [11], [14] 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や最尤推定によって推定する．中村ら [15] 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．文献 [16] 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，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．

しかしながら，これまでの声質変換技術はスペクトル変換に着目したものがほとんどであり，発話長，Duration を変換したものは少なかった．構音障害者の発話の場合，健常者と比較して Duration が長くなる上，アテトーゼ現象によって発話リズムが崩れる例が指摘されている [17]．特に文章単位の長い発話においては，Duration が聞き取りやすさに与える影響が大きいと考えれる．声質変換において，F0 は線形変換を用いることが多かったが，Duration は前後の音素関係によって多様に変化するため，線形変換を用いることは困難である．音声合成においては，Duration は隠れマルコフモデル (Hidden Markov Model: HMM) によってモデル化されることが多いが，入力音声の音素ラベルが与えられない声質変換での使用は困難である．文献 [18] のように声質変換に Duration モデルを導入したものもあるが，これも音素認識を前提としており，認識が困難な構音障害者発話への応用は難しい．

本研究では障害者発話の Duration を健常者発話の Duration へと変換することで，聞き取りやすさの向上を目指す．障害者発話は音素のつながりによって，Duration が様々に変化するため，単純な線形変換では対応が困難であると考えられる．そこで，本研究では従来の統計的声質変換モデルをベースに，パラレルデータ間のアライメント情報に基づいたフレームベースの Duration 特徴量を提案する．障害者と健常者の間のパラレル発話間のマッチング距離を Duration 特徴量とすることで，入力話者に対する出力話者の Duration を，入力話者のフレームベースで記述することができる．入力話者のスペクトル特徴量と，求められた Duration 特徴量を結合しモデル学習することで，音素ラベルが与えられない声質変換のようなタスクにおいても Duration 変換が可能である．また，Duration は前後の音素関係によって変化すると考えられるため，入力特徴量として複数フレームを考慮した長距離特徴量を用いる．

以下，第 2 章で提案手法の概要を述べ，第 3 章で特徴量変換手法について説明する．第 4 章で評価実験を行い，第 5 章で本稿をまとめる．

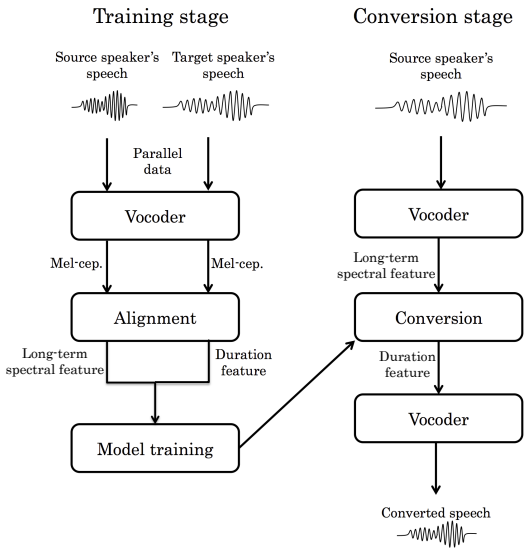

図 1　Duration 変換の概要

Fig. 1　Flow of the proposed duration conversion.

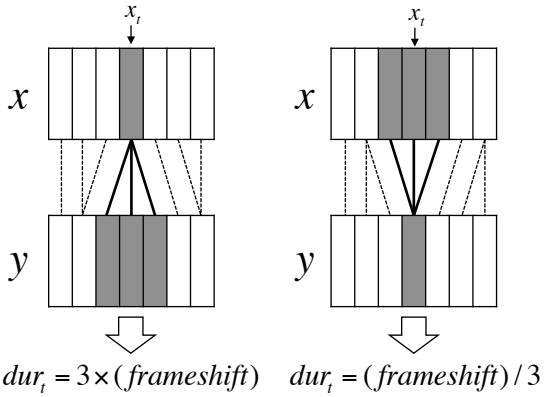

図 2　Duration 特徴量の概要

Fig. 2　Flow of the duration feature extraction.

2. Duration 変換とその特徴量抽出

図 1 に提案手法の概要を示す．本手法には，学習と変換の 2 つの段階が存在する．学習段階では，学習データとして入力話者と出力話者の間の同一発話内容から成るパラレルデータを用いる．まず，パラレルデータから vocoder を用いて特徴量抽出を行い，得られたスペクトル包絡からメルケプストラムを求める．続いて，求められたパラレルデータのメルケプストラムを用いて，データ間のアライメントを取る．得られたアライメント情報から，入力話者と出力話者の Duration の関係性を表した特徴量を求め，これを出力特徴量とする．入力特徴量としては，入力話者のメルケプストラムから求められた長距離特徴量を用いる．入力特徴量と出力特徴量を結合し，モデル学習を行う．

変換段階では，入力話者のテスト発話から Vocoder を用いて特徴量抽出を行う．特徴量で得られたスペクトル包絡からメルケプストラムを求め，変換モデルに入力する．変換によって得られた Duration 特徴量と入力された発話の特徴量から，Vocoder を用いて波形生成を行う．

2.1 Duration 特徴量

本章では，フレームベースの Duration 特徴量について述べる．ここでは，入力話者の発話に対する出力話者の Duration を，入力話者の発話フレームベースで求めることを目的とする．まず，入力話者の特徴量系列 $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_{T_x}]$ と同一発話内容の出力話者特徴量系列 $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_t, \ldots, \mathbf{y}_{T_y}]$ の間のアライメントを取る．ここで，T_x, T_y はそれぞれの特徴量系列のフレーム数を表す．アライメント情報の取得には Dynamic Programming (DP) による伸縮マッチングを用いる．DP マッチングは，下記のコスト関数を最小化する経路を求める最適化問題である．

$$F = \min \frac{\sum w(k) d(\mathbf{x}_k, \mathbf{y}_k)}{\sum w(k)} \tag{1}$$

ここで，F は $(x, y) = (1, 1)$ から (T_x, T_y) へ至る経路を表す．d は 2 つのベクトル間のユークリッド距離，w は傾斜重みを表す．本稿では，整合窓は用いず，声質変換で一般的に用いられる対称形の経路を用いる．

続いて，アライメントの取れたパラレル発話から，入力話者の Duration 特徴量を求める．図 2 にその概要を示す．入力話者特徴量の t 番目のフレームを $\mathbf{x}_t$ としたとき，対応するフレームの Duration 特徴量を dur_t とする．任意のフレーム $\mathbf{x}_t$ に対して，アライメントの取られ方は図 2 のような 2 つのパターンに分けられる．まず，図の左側のように入力話者特徴量の 1 つのフレーム $\mathbf{x}_t$ に対して出力話者特徴量の複数のフレームが結びついている場合を考える．出力話者特徴量に合わせて入力話者特徴量を引き伸ばした場合，$\mathbf{x}_t$ のフレームシフトは，それに対して結びついた出力話者特徴量のフレーム数分引き伸ばされることになり，この仮想的なフレームシフトを $\mathbf{x}_t$ に対応する Duration 特徴とする．次に，図の右側のように，出力話者特徴量の 1 つのフレームに対して，$\mathbf{x}_t$ を含む複数フレームの入力話者特徴量が結びついている場合を考える．出力話者特徴量に合わせて入力話者特徴量を場合，$\mathbf{x}_t$ のフレームシフトは，出力話者特徴量と結びついた入力話者特徴量のフレーム数分押し縮められることになり，この仮想的なフレームシフトを $\mathbf{x}_t$ に対応する Duration 特徴とする．

以上のように，DP マッチングに基づいて，入力話者特徴量に対する出力話者の Duration が, 入力話者のフレームベースで得られる．Duration の急激な変化によって，不自然に変換されることを防ぐため，実装上はこうして得られた Duration 特徴に対して平均フィルターをかけたものを用い，その窓幅は実験的に最適なものを選択する．

2.2 長距離特徴量

従来の声質変換手法では，スペクトル特徴量にシングルフレームとその動的特徴量を連結したものを用いることが一般的であった．しかしながら，Duration の変換を考えた場合，Duration は前後の音素の影響を大きく受けると考えられるため，従来のような短時間情報のみを含んだ特徴量では不十分であると考えられる．そこで本稿では，入力特徴量として，メル

ケプストラムの複数のフレームを考慮した長時間特徴量を用いる．

図 3 に長距離特徴量の概要を示す．入力話者特徴量を $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{T_x}\}$ とし，その次元数を d_x とする．前後 L フレームをまとめたセグメント特徴量を求めると，その次元数は $d_x(2L-1)$ となる．セグメント特徴量に対して主成分分析 (Principal Component Analysis: PCA) を適用することで，D_x 次元の長距離特徴量 $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_{T_x}\}$ が得られる．

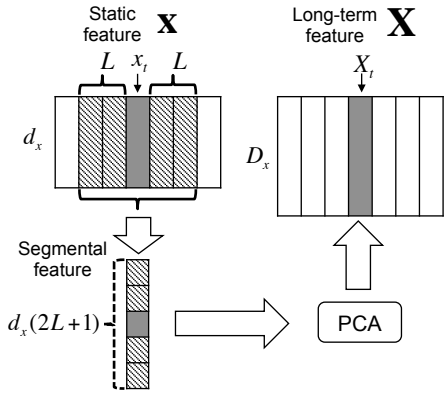

図 3　長距離特徴量の概要

Fig. 3　Flow of the construction of long-term image features.

3. GMM に基づく最尤特徴量変換

本稿では，特徴量変換手法として，声質変換で最も一般的な GMM に基づく最尤特徴量変換 [14] を用いる．入力話者のスペクトル特徴量と，入力話者に対応する出力話者の Duration 特徴量を結合して GMM を学習することで，特徴量空間は教師なしで音響クラスに分離され，その音響クラスごとに Duration 変換が可能となる．

入力特徴量と出力特徴量の結合確率分布は複数の多次元正規分布モデル $\mathcal{N}(.; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ によって表される．入力特徴量を $\mathbf{X}$，出力特徴量を $\mathbf{Y}$ とすると，特徴量結合ベクトル $\mathbf{Z} = [\mathbf{X}^\mathsf{T}\mathbf{Y}^\mathsf{T}]^\mathsf{T}$ の確率分布 $p(\mathbf{Z})$ は下記のようにモデル化される．

$$p(\mathbf{Z}|\boldsymbol{\Theta}^{(z)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\mathbf{Z}; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}), \tag{2}$$

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \tag{3}$$

$\boldsymbol{\mu}_m^{(x)}$，$\boldsymbol{\Sigma}_m^{(xx)}$ 並びに $\boldsymbol{\mu}_m^{(y)}$，$\boldsymbol{\Sigma}_m^{(yy)}$ はそれぞれ入力特徴量・出力特徴量の平均ベクトル，分散共分散行列を表す．α_m は m 番目の正規分布に対する重みであり，分布の総数が M で表される．$\boldsymbol{\Theta}^z$ はパラメータの集合であり，全ての m に対する α_m, $\boldsymbol{\mu}_m^{(x)}$, $\boldsymbol{\mu}_m^{(y)}$, $\boldsymbol{\Sigma}_m^{(xx)}$, $\boldsymbol{\Sigma}_m^{(yy)}$, and $\boldsymbol{\Sigma}_m^{(xy)}$ が含まれる．これらのパラメータは EM アルゴリズムで推定される．

変換時には，入力特徴量 $\mathbf{X}$ が与えられたときの $\mathbf{Y}$ の確率分布を考える．すなわち，

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Theta}^{(z)}) \simeq p(\hat{\mathbf{m}}|\mathbf{X}, \boldsymbol{\Theta}^{(z)}) p(\mathbf{Y}|\mathbf{X}, \hat{\mathbf{m}}, \boldsymbol{\Theta}^{(z)}) \tag{4}$$

ただし，

$$p(m_t|\mathbf{X}_t, \boldsymbol{\Theta}^{(z)}) = \frac{\alpha_m \mathcal{N}(\mathbf{X}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{n=1}^{M} \alpha_n \mathcal{N}(\mathbf{X}_t; \boldsymbol{\mu}_n^{(x)}, \boldsymbol{\Sigma}_n^{(xx)})} \tag{5}$$

$$p(\mathbf{Y}_t|\mathbf{X}_t, m_t, \boldsymbol{\Theta}^{(z)}) = \mathcal{N}(\mathbf{Y}_t; \mathbf{E}_{m,t}^{(y|x)}, \mathbf{D}_m^{(y|x)}). \tag{6}$$

$\hat{\mathbf{m}}$ は準最適な混合系列であり，次のように表される．

$$\hat{\mathbf{m}} = \arg\max P(\mathbf{m}|\mathbf{X}, \boldsymbol{\Theta}^{(z)}). \tag{7}$$

式 (4) の対数尤度関数は次のように表される．

$$\begin{aligned} &\log p(\mathbf{Y}|\mathbf{X}, \hat{\mathbf{m}}, \boldsymbol{\Theta}^{(z)}) \\ &= -\frac{1}{2}\mathbf{Y}^\mathsf{T}\mathbf{D}_{\hat{\mathbf{m}}}^{(y|x)^{-1}}\mathbf{Y} + \mathbf{Y}^\mathsf{T}\mathbf{D}_{\hat{\mathbf{m}}}^{(y|x)^{-1}}\mathbf{E}_{\hat{\mathbf{m}}}^{(y|x)} + K \end{aligned} \tag{8}$$

ただし，

$$\mathbf{E}_{\hat{\mathbf{m}}}^{(y|x)} = [\mathbf{E}_{\hat{m}_1,1}^{(y|x)}, \mathbf{E}_{\hat{m}_2,2}^{(y|x)}, \cdots, \mathbf{E}_{\hat{m}_T,T}^{(y|x)}] \tag{9}$$

$$\mathbf{D}_{\hat{\mathbf{m}}}^{(y|x)} = diag[\mathbf{D}_{\hat{m}_1,1}^{(y|x)}, \mathbf{D}_{\hat{m}_2,2}^{(y|x)}, \cdots, \mathbf{D}_{\hat{m}_T,T}^{(y|x)}]. \tag{10}$$

ここで，静的特徴量と動的特徴量の関係を表す行列 $\mathbf{W}$ を用いると，求めたい特徴量系列 $\hat{\mathbf{y}}$ は

$$\hat{\mathbf{y}} = (\mathbf{W}^\mathsf{T}\mathbf{D}_{\hat{\mathbf{m}}}^{(y|x)^{-1}}\mathbf{W})^{-1}\mathbf{W}^\mathsf{T}\mathbf{D}_{\hat{\mathbf{m}}}^{(y|x)^{-1}}\mathbf{E}_{\hat{\mathbf{m}}}^{(y|x)}. \tag{11}$$

のように表される．

一般的な声質変換では，パラメータ数の削減のため，分散共分散行列に対角行列を用いることが多かった．本研究では，異なる次元数かつ異なる性質の特徴量変換を行うため，非対角の分散共分散行列を用いる．

4. 評 価 実 験

4.1 実 験 条 件

入力話者として構音障害者の音声として使用するため，男性のアテトーゼ型構音障害者 1 名による 100 文を収録した．発話内容は ATR 音素バランス文 [19] を用いた．50 文を学習データ，残りの 50 文をテストデータとした．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．障害者音声は，HMM の強制アライメントをベースに，手動でラベルの修正を行った．

学習データから音声分析合成 WORLD [20] を用いてスペクトル包絡，基本周波数，非周期成分を抽出する．抽出したスペクトル特徴量から 24 次元のメルケプストラムを計算し，DP マッチングの特徴量として用いる．Duration 変換時には，入力特徴量として入力話者のメルケプストラムから求めた長時間特徴量を，出力特徴量として Duration 特徴量とその δ 特徴量を結合したものを用いる．WORLD は分析窓のフレームシフトが可変であるため，推定された Duration 特徴量をを用いて，入力話者のスペクトル包絡，基本周波数，非周期成分が再合成される．本稿では，Duration に着目するため，スペクトル包

絡，基本周波数，非周期成分は入力話者のものを用いる．

客観評価指標として，式 (1) で示される，入力話者あるいは変換音声と出力話者の間のメルケプストラムによる DP 距離を用いた．また，音素ラベルごとに，健常者，障害者，変換後の Duration を求め，その差を比較した．

4.2 実験結果

図 5 に長距離特徴量の窓幅による DP 距離の変化を示す．Source は変換前の障害者音声と目標となる健常者音声間の DP 距離示し，L は図 3 における窓幅を表す．図より，Duration 変換によって，無変換の場合と比較して DP 距離が削減できており，提案手法の有効性が確認できる．$L = 5$ の場合が最も適切に変換できることがわかる．

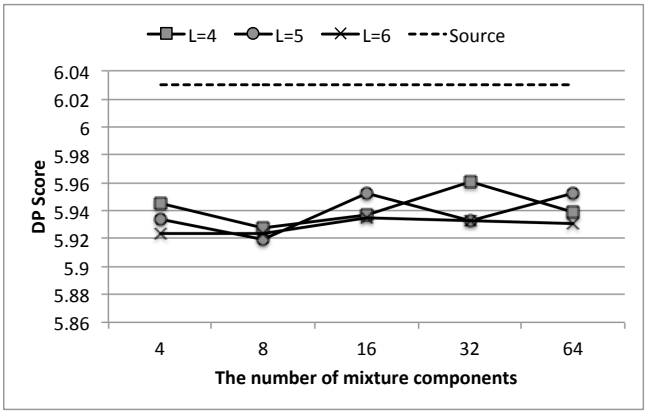

図 4　長距離特徴量の窓幅による DP 距離の変化

Fig. 4　DP score as a function of the number of mixture components using long-term features.

図 5 に，平均フィルタの窓幅による DP 距離の変化を示す．M がその窓幅を表す．長距離特徴量は $L = 5$ とした．図より，$M = 6$ の場合が最も適切であることがわかる．

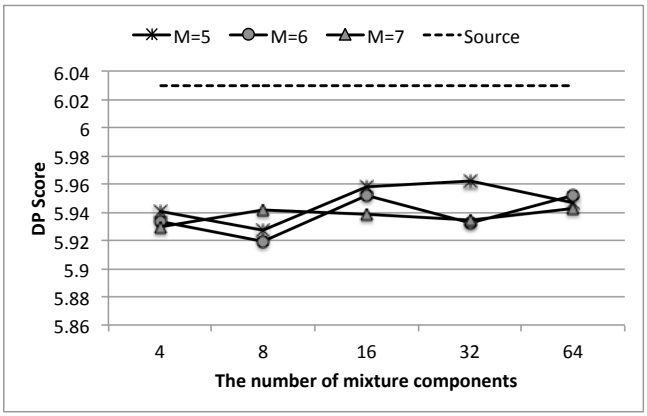

図 5　平均フィルタの窓幅による DP 距離の変化

Fig. 5　DP score as a function of the number of mixture components using mean filters.

表 1 に音素ごとの Duration 変換結果を示す．$s/t-1$，$c/t-1$ はそれぞれ，障害者音声と健常者音声の Duration 比から 1 を引いたもの，変換音声と健常者音声の Duration 比から 1 を引いたものを表し，絶対値が 0 に近いほど正しく変換できていることを示す．表より，ほとんど全ての音素について適切に変換できていることがわかり，特に母音については健常者の Duration に大きく近づいていることがわかる．

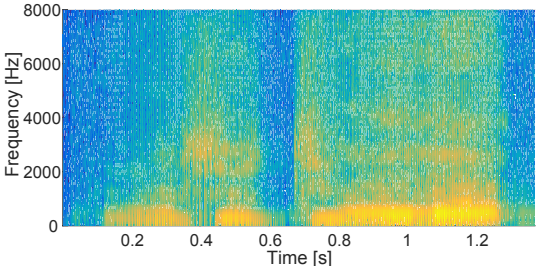

図 6　構音障害者の発話スペクトル例 //rusuchuno

Fig. 6　An example of dysarthric speech spectrogram.

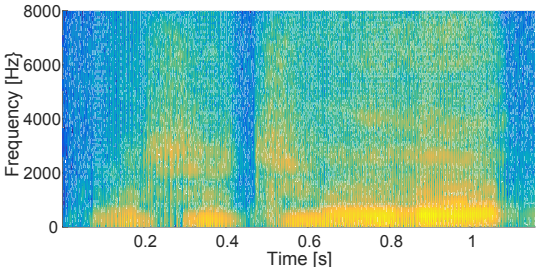

図 7　Duration 変換を行った発話スペクトル例 //rusuchuno

Fig. 7　An example of duration converted spectrogram.

図 6 に構音障害者の発話スペクトル例，図 7 に Duration 変換を行った発話スペクトル例を示す．発話は全体的に短くなっており，特に障害者に見られる “rusu” と “chuno” の間の不自然なポーズが，変換によって短くなっていることがわかる．

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 Duration 変換を行った．従来一般的な GMM による特徴量変換で用いることができるフレームベース Duration 特徴量を提案し，客観評価によってその有効性を確認した．入力話者のスペクトル特徴量と，対応する出力話者の Duration を結合してモデル学習することにより，音響クラスごとに Duration が表現でき，声質変換のような明示的な音素ラベルを推定しないタスクにおいても Duration 変換が可能になった．今後は，提案した Duration 変換を行った上で，スペクトル変換や F0 変換を行い，障害者音声を健常者音声へと変換することを目指す．また，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] J. Lin, W. Ying, and T.S. Huang, “Capturing human hand motion in image sequences,” in Proc. IEEE Motion and Video Computing Workshop, pp.99–104, 2002.

[2] T. Starner, J. Weaver, and A. Pentland, “Real-time American sign language recognition using desk and wearable computer based video,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.20, no.12, pp.1371–1375, 1998.

[3] G. Fang, W. Gao, and D. Zhao, “Large vocabulary sign language recognition based on hierarchical decision trees,”

表 1　音素ごとの Duration 変換結果

Table 1　Result of duration conversion.

Phn.	Src. [ms]	Trg. [ms]	Cnv. [ms]	s/t-1	c/t-1
N	2837	1396	2025	1.03	**0.45**
Q	6153	3333	3734	0.85	**0.12**
a	33662	29125	30893	0.16	**0.06**
aN	6284	5093	5472	0.23	**0.07**
aa	1043	775	835	0.35	**0.08**
ai	7749	5737	6874	0.35	**0.20**
ao	1042	770	854	0.35	**0.11**
b	3761	2191	2586	0.72	**0.18**
by	162	100	155	0.62	**0.55**
ch	5038	3612	4458	0.39	**0.23**
d	4733	3391	3581	0.40	**0.06**
e	15128	13314	14501	0.14	**0.09**
eN	6264	4559	5544	0.37	**0.22**
ee	585	290	559	1.02	**0.93**
ei	4012	3326	3432	0.21	**0.03**
f	2002	1137	1770	0.76	**0.56**
g	5385	4269	4166	0.26	**-0.02**
gy	352	290	256	0.21	**-0.12**
h	7447	6066	6933	0.23	**0.14**
hy	620	580	545	0.07	**-0.06**
i	21773	15563	18492	0.40	**0.19**
iN	5490	3407	4189	0.61	**0.23**
ii	2700	1295	1450	1.08	**0.12**
j	5076	4124	4527	0.23	**0.10**
k	16556	11121	13470	0.49	**0.21**
ky	1410	1378	1142	0.02	**-0.17**
m	10088	4834	7674	1.09	**0.59**
n	13621	6543	11231	1.08	**0.72**
ny	893	505	732	0.77	**0.45**
o	29634	22522	22884	0.32	**0.02**
oN	3290	2620	2343	0.26	**-0.11**
oo	2062	1383	1234	0.49	**-0.11**
ou	13840	8731	9201	0.59	**0.05**
p	1768	864	1257	1.05	**0.45**
r	10316	2793	8110	2.69	**1.90**
ry	483	313	454	0.54	**0.45**
s	9103	6234	8150	0.46	**0.31**
sh	10012	7014	8507	0.43	**0.21**
t	14052	9457	11583	0.49	**0.22**
ts	3584	2658	3167	0.35	**0.19**
u	18364	14352	14211	0.28	**-0.01**
ui	1870	1106	1417	0.69	**0.28**
uu	4353	2696	3509	0.61	**0.30**
w	1948	808	1354	1.41	**0.68**
y	4381	3063	4062	0.43	**0.33**
z	3028	1755	2474	0.73	**0.41**

in Proc. 5th International Conference on Multimodal Interfaces, pp.125–131, 2003.

[4] N. Ezaki, M. Bulacu, and L. Schomaker, “Text detection from natural scene images: Towards a system for visually impaired persons,” in Proc. ICPR, pp.683–686, 2004.

[5] M.K. Bashar, T. Matsumoto, Y. Takeuchi, H. Kudo, and N. Ohnishi, “Unsupervised texture segmentation via wavelet-based locally orderless images (wlois) and SOM,” in Proc. 6th IASTED International Conference COMPUTER GRAPHICS AND IMAGING, pp.279–284, 2003.

[6] V. Wu, R. Manmatha, and E.M. Riseman, “Textfinder: an automatic system to detect and recognize text in images,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.21, no.11, pp.1224–1229, 1999.

[7] K. Yabu, T. Ifukube, and S. Aomura, “A basic design of wearable speech synthesizer for voice disorders [japanese],” EIC Technical Report (Institute of Electronics, Information and Communication Engineers), vol.105, no.686, pp.59–64, 2006.

[8] S.T. Canale and W.C. Campbell, “Campbell’s operative orthopaedics,” Technical report, Mosby-Year Book, 2002.

[9] H. Matsumasa, T. Takiguchi, Y. Ariki, I. Li, and T. Nakabayachi, “Integration of metamodel and acoustic model for dysarthric speech recognition,” Journal of Multimedia, vol.4, no.4, pp.254–261, 2009.

[10] C. Miyamoto, Y. Komai, T. Takiguchi, Y. Ariki, and I. Li, “Multimodal speech recognition of a person with articulation disorders using AAM and MAF,” in Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP’10), pp.517–520, 2010.

[11] Y. Stylianou, O. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech and Audio Processing, vol.6, no.2, pp.131–142, 1998.

[12] C. Veaux and X. Robet, “Intonation conversion from neutral to expressive speech,” in Proc. Interspeech, pp.2765–2768, 2011.

[13] Y. Ohtani, M. Tamura, M. Morita, and M. Akamine, “GMM-based bandwidth extension using sub-band basis spectrum model,” in Proc. Interspeech, pp.2489–2493, 2014.

[14] 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.

[15] 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.

[16] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, “Individuality-preserving voice conversion for articulation disorders based on Non-negative Matrix Factorization,” in Proc. ICASSP, pp.8037–8040, 2013.

[17] F. Rudzicz, “Adjusting dysarthric speech signals to be more intelligible,” Computer Speech and Language, vol.27, no.6, pp.1163–1177, 2014.

[18] R. Srikanth, B. Bajibabu, and K. Prahallad, “Duration modelling in voice conversion using artificial neural networks,” in Proc. IWSSIP, pp.556–559, 2012.

[19] 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.

[20] M. Morise, F. Yokomori, and K. Ozawa, “World: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE transactions on information and systems, vol.E99-D, no.7, pp.1877–1884, 2016.