基本周波数軌跡の同時生成モデル化に基づく韻律変換 *

☆石原達馬 [1]，吉里幸太 [1]，亀岡弘和 [1,2]
([1] 東大院・情報理工, [2]NTT CS 研)

1 はじめに

音声の韻律情報は感情，個人性，意図などの非言語・パラ言語情報を含んでおり，重要な音声特徴量の一つである．従って韻律情報を，言語情報を保存したまま操作する技術は，音声に込められた感情を操作する感情変換や，より忠実に対象の特徴を再現する話者変換など，様々なアプリケーションへの応用が期待できる．基本周波数（F_0）軌跡は主要な韻律特徴量であるため，上記のような技術を実現するための取り組みとして我々は F_0 軌跡を操作する新しいフレームワークの構築を試みてきた．

類似の問題を扱った手法として，統計的声質変換 [1] が挙げられる．統計的声質変換は発話内容を保存したままその他の情報を変換する手法であり，個人性の変換や帯域拡張などに用いられている．この手法は音韻特徴量ペアを生成する確率モデルをベースに，変換規則の学習と変換処理を確率モデルのパラメータ推定として定式化した手法であると捉えることができる．本稿ではこの考え方を骨子とし，F_0 軌跡の生成過程を考えることで，韻律特徴量の変換規則の学習と変換アルゴリズムを確率的同時生成モデルの考え方に従って導出する．

2 本研究のアプローチ

2.1 韻律変換の問題設定

音声の F_0 軌跡は大きく分けて 2 つの成分 [2] からなる．緩やかな時間変化の成分であるフレーズ成分と，急激な時間変化の成分であるアクセント成分である．この内，アクセント成分は更に言語内容によって定まる高低アクセントの離散的な情報と，個人性や感情を反映した連続的な情報（タイミング・プロミネンス）とに分けられる [3]．よって F_0 軌跡の変換を扱う場合，これらの情報を適切に分離し，個人性や感情を反映した情報のみを置き換える事が必要となる．

音韻情報の言語情報を保存したまま個人性を操作する手法である統計的声質変換においても類似した課題が発生する．これらの手法において，各時刻の音韻情報の同時生成モデルを混合ガウスモデルで表すことで，パラレルコーパスからそのモデルを学習することが可能となった．F_0 軌跡の変換でも同様に，F_0 軌跡の同時生成モデルを考えることができれば，大量のパラレルデータから変換則を学習することができると考えられる．ここで音韻情報と韻律情報の大きな違いは，韻律情報は長時間の軌跡全体にわたって現れるということである．従ってフレームベースの瞬間的な生成モデルを考えていた音韻情報の変換とは異なり，軌跡全体の生成モデルを考える必要があるのが，音韻情報の変換とは大きく異なる点である．では F_0 軌跡全体の生成モデルはどのように定式化できるだろうか．

F_0 軌跡の生成過程は様々な要因が関係する複雑なものであるが，大きく分けて 3 つの層に分けて考えることができる．アクセントの高低パターンなど言語的な情報に関係する層，タイミングやプロミネンスのなど個人性・感情・意図などに関係する層，そして物理的な生成プロセスの層である．生成過程をこのように捉えた場合，F_0 軌跡を変換する問題は，言語層や物理層を固定したまま個人・感情層の情報を特定のものに置き換える問題であると言い換えることができる．

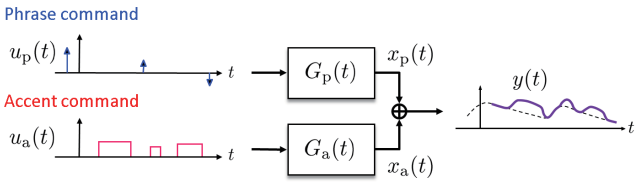

Fig. 1 藤崎モデル [2]．フレーズ指令（上段左）はパルス列，アクセント指令（下段左）は矩形パルス列であり，それぞれ独立に発話・個人によらない 2 次の臨界制動系フィルタを通して足しあわされる．

2.2 同時生成モデル化

F_0 軌跡の生成過程が先に述べた 3 つの層からなるという考えのもとで，F_0 軌跡ペアの同時生成モデルは，F_0 軌跡ペア間で共通の言語情報にそれぞれ独立に個人性・感情層の情報が付与され，それぞれ独立に物理的生成過程により観測される F_0 軌跡が生成される，というように考えることができる．このように考えることで，変換のためのモデル学習と変換則は，ともにモデルの未知パラメータの推定問題として定式化することができる．つまり，モデル学習は所与の F_0 軌跡ペアのコーパスを最もよく説明するようなパラメータを推定する問題となり，F_0 軌跡の変換は不完全な観測データから隠れパラメータを推定する問題となる．

以降，各層でどのような生成モデルを考えるべきかについて詳細に述べる．

3 基本周波数軌跡の同時生成モデルの定式化

3.1 言語情報を背後に持つ F_0 軌跡の生成モデル

提案生成過程の物理層のモデルの候補となる F_0 軌跡の物理的生成モデルとして，藤崎による F_0 軌跡の生成過程モデル [2]（以下藤崎モデル）が広く知られている．藤崎モデルでは，対数 F_0 軌跡 $y(t)$ が以下のように 3 つの成分の和で表されると仮定する．

$$y(t) = x_p(t) + x_a(t) + x_b. \tag{1}$$

ここで，t は時間，$x_p(t)$ はフレーズ成分，$x_a(t)$ はアクセント成分，x_b はベースライン成分と呼ばれる，時間によらない定数である．さらにフレーズ成分，アクセント成分はそれぞれ，フレーズ指令，アクセント指令と呼ばれる信号の 2 次のフィルタの出力であると仮定される．

$$x_p(t) = G_p(t) * u_p(t) \tag{2}$$

$$G_p(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & (t \geq 0) \\ 0 & (t < 0) \end{cases} \tag{3}$$

$$x_a(t) = G_a(t) * u_a(t) \tag{4}$$

$$G_a(t) = \begin{cases} \beta^2 t e^{-\beta t} & (t \geq 0) \\ 0 & (t < 0) \end{cases} \tag{5}$$

*Prosody Conversion Based on Joint Generative Model of F_0 Contours . by ISHIHARA Tatsuma, YOSHIZATO Kota, KAMEOKA Hirokazu (The University of Tokyo)

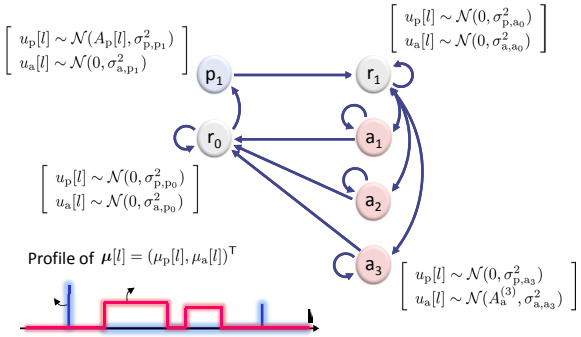

Fig. 2　フレーズ・アクセント指令列の状態遷移モデル [6, 7]．状態 r_0 において $\mu_p[t]$ と $\mu_a[t]$ はゼロである．状態 p_1 において $\mu_p[t]$ は非負値 $A_p[t]$ をとることができ，$\mu_a[t]$ はゼロである．状態 p_1 において自己遷移は禁止される．状態 r_1 において $\mu_p[t]$ と $\mu_a[t]$ はまたゼロのみに制限される．この状態は $\mu_p[t]$ がパルス列になることを保証するものである．状態 r_0 は状態 $a_1, \ldots, a_N$ へのみ遷移することができ，これらの状態において $\mu_a[t]$ はそれぞれ異なる値 $A_a^{(n)}$ をとることができるが，$\mu_p[t]$ はゼロに制限される．直接 a_n から $a_{n'}$ へ r_1 を通らずに遷移することは禁止される．これは $\mu_a[t]$ が矩形パルス列であることを保証するためのものである．

ここで $u_p(t)$ はフレーズ指令と呼ばれるデルタ列であり，$u_a(t)$ はアクセント指令と呼ばれる矩形パルス列である．これらのうち非ゼロの値をとるのは各時刻で高々1つである．α，β はそれぞれ2次フィルタの応答の速さを表す角周波数であり，個人や発話によらずおおよそ $\alpha = 3$ rad/s，$\beta = 20$ rad/s 程度の値をとることが知られている．

藤崎モデルは決定論的なモデルであり，モデル学習を適用できるようにするために確率モデル化が必要である．藤崎モデルをベースに F_0 軌跡の確率的生成過程を記述したモデル [6, 7] を我々は提案してきた．上述の藤崎モデルにおいて，フレーズ指令，アクセント指令はそれぞれデルタ列，矩形パルス列であり，さらにこれらは互いに重ならないという仮定が置かれる．我々はこれらの制約を満たすような指令列をうまく確率モデルの形として記述するために，フレーズ指令 $u_p[t]$，アクセント指令 $u_a[t]$ のペア $\boldsymbol{u}[t] = (u_p[t], u_a[t])^T$ を，HMM の出力として表現するモデルを考案した．各状態の出力分布を正規分布とした場合，出力系列 $\{\boldsymbol{u}[t]\}_{t=1}^T$ は

$$\boldsymbol{u}[t] \sim \mathcal{N}\left(\boldsymbol{u}[t]; \boldsymbol{c}_{s[t]}, \boldsymbol{\Upsilon}_{s[t]}\right) \tag{6}$$

に従う．ここで $s[t]$ は時刻 k における状態を表す．すなわち，式 (6) は平均 $\boldsymbol{\mu}[t] = (\mu_p[t], \mu_a[t])^T = \boldsymbol{c}_{s[t]}$ と分散 $\boldsymbol{\Sigma}[t] = \boldsymbol{\Upsilon}_{s[t]}$ が状態遷移の結果として時間とともに変化することを意味する．以上の HMM の構成は以下となる．

出力系列: $\{\boldsymbol{u}[t]\}_{t=1}^T$
状態系列: $\{s[t]\}_{t=1}^T$
出力確率分布: $\mathrm{P}(\boldsymbol{u}[t]|s[t]) = \mathcal{N}(\boldsymbol{u}[t]; \boldsymbol{c}_{s[t]}, \boldsymbol{\Upsilon}_{s[t]})$
平均値の系列: $\boldsymbol{\mu}[t] = (\mu_p[t], \mu_a[t])^T = \boldsymbol{c}_{s[t]}$
遷移確率: $\phi_{i',i} = \mathrm{P}(s[t] = i | s[t-1] = i')$

上記の HMM から出力された指令関数 $u_p[t]$, $u_a[t]$ にそれぞれ異なるフィルタ $G_p[t]$ と $G_a[t]$ が畳み込まれたものがフレーズ成分とアクセント成分

$$x_p[t] = u_p[t] * G_p[t] \tag{7}$$
$$x_a[t] = u_a[t] * G_a[t] \tag{8}$$

となる．ただし，$*$ は離散時間 k に関する畳込みを表す．また，$G_p[t]$ と $G_a[t]$ はそれぞれ $G_p(t)$ と $G_a(t)$ を離散時間表現である．以上より，F_0 軌跡の離散時間表現 $x[t]$ は

$$x[t] = x_p[t] + x_a[t] + x_b \tag{9}$$

となる．x_b はベースライン成分を表す．

無声区間においては F_0 は観測されないことがあったり，観測されていたとしても信頼できない場合が多い．また，F_0 抽出において推定誤りが生じる場合もある．そこで観測 F_0 軌跡 $y[t]$ を，上述の F_0 軌跡モデル $x[t]$ とノイズ $x_n[t] \sim \mathcal{N}\left(0, \upsilon_n[t]^2\right)$ との和として表すことで，観測 F_0 系列の不確実性を分散 $\upsilon_n^2[t]$ の設定を通して組み込むことができる．よって，観測 F_0 系列 $y[t]$ は

$$y[t] = x[t] + x_n[t] \tag{10}$$

と表される．ここで，$x_n[t]$ を周辺化すると，$\boldsymbol{u} = \{\boldsymbol{u}[t]\}_{t=1}^T$ が与えられたもとでの $\boldsymbol{y} = \{y[t]\}_{t=1}^T$ の条件つき確率密度関数 $P(\boldsymbol{y}|\boldsymbol{u})$ は

$$P(\boldsymbol{y}|\boldsymbol{u}) = \prod_{t=1}^{T} \mathcal{N}(y[t]; x[t], \upsilon_n^2[t])$$
$$x[t] = G_p[t] * u_p[t] + G_a[t] * u_a[t] + u_b \tag{11}$$

となる．(6) より，状態系列 $\boldsymbol{s} = \{s[t]\}_{t=1}^T$ が与えられたもとでの $\{\boldsymbol{u}[t]\}_{t=1}^T$ の条件つき確率密度関数 $P(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\theta})$ は $P(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\theta}) = \prod_{t=1}^T \mathcal{N}(\boldsymbol{u}[t]; \boldsymbol{c}_{s[t]}[t], \boldsymbol{\Upsilon}_{s[t]})$ で与えられる．ここで，$\boldsymbol{\theta}$ は出力分布の平均と分散の系列を表す．状態系列 $\boldsymbol{s}$ の確率分布 $P(\boldsymbol{s})$ は HMM におけるマルコフ性の仮定より，遷移確率の積 $P(\boldsymbol{s}) = \phi_{s[1]} \prod_{t=2}^T \phi_{s[t],s[t-1]}$ で与えられる．

指令列関数が提案した3層の過程のうち，どの層の情報を含んでいるかを考えると，藤崎モデルの指令列は言語情報とよく対応する [2] ことから，指令列関数は言語層の情報を含んでいる．一方で，指令列関

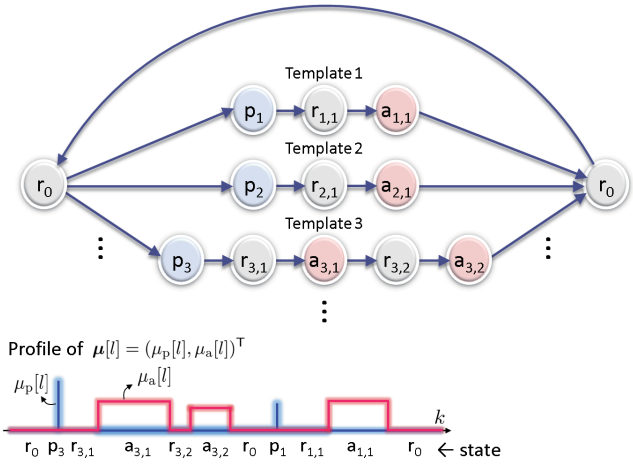

Fig. 3　ピッチパターンテンプレートの語彙モデルに基づくフレーズ・アクセント指令列の状態遷移トポロジー

数の位置や大きさはそれぞれタイミングとプロミネンスに対応するため，指令列関数には 2 つの層の情報が混在して含まれていると考えられる．この 2 つの情報を分離する手法を考えるため，以下で日本語のもつ性質について考察する．

通常の発話では，様々なイントネーション型が現れる．とはいえ，イントネーション型の種類には限りがある．これは日本語の場合ピッチアクセントは高い-低いの 2 値で表され，1 アクセント句に含まれるモーラ数には限りがあるためである．例えば，「あらゆる現実を」と「明日は輪講だ」のアクセントパターンは同一であるため，イントネーションはほとんど同一である．以上の観察から，F_0 軌跡がある仮想的な辞書から生起したものであると仮定することは自然であると考えられる．このような語彙構造を持つ生成モデルは，これまで述べた F_0 軌跡の生成モデルの状態遷移図を修正することで表現できる．F_0 軌跡の統計的語彙モデル [4, 5] は，このような語彙構造を隠れマルコフモデルを用いて表現し，これまでの確率モデルと統合したものである．このような表現により確率モデルを記述したとき，使用されたテンプレート番号が言語情報を，テンプレートごとの指令列の強度が個人性・感情の情報を反映していると解釈できる．

以上の確率モデルをもとに，F_0 軌跡 $\boldsymbol{y}^{(\mathrm{A})}, \boldsymbol{y}^{(\mathrm{B})}$ の同時確率密度を以下で導出する．$\boldsymbol{s}$ を Fig. 3 の状態遷移図に従う隠れマルコフモデルの状態系列とする．

$$P(\boldsymbol{s}) = \phi_{s[1]} \prod_{t=2}^{T} \phi_{s[t],s[t-1]} \tag{12}$$

この $\boldsymbol{s}$ が軌跡ペアの共通の言語情報を表すことは前述のとおりであり，それぞれの軌跡に独立にプロミネンスが付加されることは指令列 $\boldsymbol{u}^{(\mathrm{A})}, \boldsymbol{u}^{(\mathrm{B})}$ が HMM から独立に出力されることにより表現できる．

$$\mathrm{P}\Big(u^{(\mathrm{A})}[t]|s[t]\Big) = \mathcal{N}\Big(u^{(\mathrm{A})}[t]; \mu^{(\mathrm{A})}_{s[t]}, \sigma^{(\mathrm{A})}_{s[t]}\Big) \tag{13}$$

$$\mathrm{P}\Big(u^{(\mathrm{B})}[t]|s[t]\Big) = \mathcal{N}\Big(u^{(\mathrm{B})}[t]; \mu^{(\mathrm{B})}_{s[t]}, \sigma^{(\mathrm{B})}_{s[t]}\Big) \tag{14}$$

$\boldsymbol{y}^{(\mathrm{A})}, \boldsymbol{y}^{(\mathrm{B})}$ を F_0 軌跡のペアとすると，これらは指令列 $\boldsymbol{u}^{(\mathrm{A})}, \boldsymbol{u}^{(\mathrm{B})}$ が与えられたもとで以下のように生成される．

$$\mathrm{P}\Big(y^{(\mathrm{A})}[t]|x^{(\mathrm{A})}[t]\Big) = \mathcal{N}\Big(y^{(\mathrm{A})}[t]; x^{(\mathrm{A})}[t], \upsilon^{(\mathrm{A})2}_{\mathrm{n}}[t]\Big) \tag{15}$$

$$x^{(\mathrm{A})}[t] = x^{(\mathrm{A})}_{\mathrm{p}}[t] + x^{(\mathrm{A})}_{\mathrm{a}}[t] + x^{(\mathrm{A})}_{\mathrm{b}} \tag{16}$$

$$x^{(\mathrm{A})}_{\mathrm{p}}[t] = u^{(\mathrm{A})}_{\mathrm{p}}[t] * G_{\mathrm{p}}[t] \tag{17}$$

$$x^{(\mathrm{A})}_{\mathrm{a}}[t] = u^{(\mathrm{A})}_{\mathrm{a}}[t] * G_{\mathrm{a}}[t] \tag{18}$$

以上をまとめるとモデルパラメータを θ として $\boldsymbol{y}^{(\mathrm{A})}, \boldsymbol{y}^{(\mathrm{B})}$ の同時確率密度は以下のように

$$\begin{aligned}
&\mathrm{P}\Big(\boldsymbol{y}^{(\mathrm{A})}, \boldsymbol{y}^{(\mathrm{B})}|\theta\Big) \\
&= \int \mathrm{P}\Big(\boldsymbol{y}^{(\mathrm{A})}|\boldsymbol{u}^{(\mathrm{A})}\Big)\mathrm{P}\Big(\boldsymbol{y}^{(\mathrm{B})}|\boldsymbol{u}^{(\mathrm{B})}\Big) \\
&\quad \mathrm{P}\Big(\boldsymbol{u}^{(\mathrm{A})}|\theta\Big)\mathrm{P}\Big(\boldsymbol{u}^{(\mathrm{B})}|\theta\Big)\mathrm{d}\boldsymbol{u}^{(\mathrm{A})}\mathrm{d}\boldsymbol{u}^{(\mathrm{B})} \\
&= \sum_{\boldsymbol{s}} \int \mathrm{P}\Big(\boldsymbol{y}^{(\mathrm{A})}|\boldsymbol{u}^{(\mathrm{A})}\Big)\mathrm{P}\Big(\boldsymbol{y}^{(\mathrm{B})}|\boldsymbol{u}^{(\mathrm{B})}\Big) \\
&\quad \mathrm{P}\Big(\boldsymbol{u}^{(\mathrm{A})}|\boldsymbol{s},\theta\Big)\mathrm{P}\Big(\boldsymbol{u}^{(\mathrm{B})}|\boldsymbol{s},\theta\Big)\mathrm{P}(\boldsymbol{s}|\theta)\mathrm{d}\boldsymbol{u}^{(\mathrm{A})}\mathrm{d}\boldsymbol{u}^{(\mathrm{B})}
\end{aligned} \tag{19}$$

と表せる．

3.2 パラメータ推定アルゴリズム

パラメータ θ を大量のデータから学習によって推定することを考える．解くべき問題は学習データ $\boldsymbol{y}^{(\mathrm{A})}, \boldsymbol{y}^{(\mathrm{B})}$ が与えられたもとで θ を最大化することである．すなわち，

$$\begin{aligned}
&\operatorname*{argmax}_{\theta} \sum_{\boldsymbol{s}} \int \mathrm{P}\Big(\boldsymbol{y}^{(\mathrm{A})}|\boldsymbol{u}^{(\mathrm{A})}\Big)\mathrm{P}\Big(\boldsymbol{y}^{(\mathrm{B})}|\boldsymbol{u}^{(\mathrm{B})}\Big) \\
&\quad \mathrm{P}\Big(\boldsymbol{u}^{(\mathrm{A})}|\boldsymbol{s},\theta\Big)\mathrm{P}\Big(\boldsymbol{u}^{(\mathrm{B})}|\boldsymbol{s},\theta\Big)\mathrm{P}(\boldsymbol{s}|\theta)\mathrm{d}\boldsymbol{u}^{(\mathrm{A})}\mathrm{d}\boldsymbol{u}^{(\mathrm{B})}
\end{aligned} \tag{20}$$

ここで，$\boldsymbol{u}^{(\mathrm{A})}, \boldsymbol{u}^{(\mathrm{B})}$ は各時刻で非負であるという制約がある．この制約を扱うため，目的関数の $\boldsymbol{u}^{(\mathrm{A})}, \boldsymbol{u}^{(\mathrm{B})}$ に関する積分を最大値で近似する．このとき，解くべき問題は以下のとおりである．

$$\begin{aligned}
&\operatorname*{argmax}_{\theta, \boldsymbol{u}^{(\mathrm{A})}, \boldsymbol{u}^{(\mathrm{B})}} \mathrm{P}\Big(\boldsymbol{y}^{(\mathrm{A})}|\boldsymbol{u}^{(\mathrm{A})}\Big)\mathrm{P}\Big(\boldsymbol{y}^{(\mathrm{B})}|\boldsymbol{u}^{(\mathrm{B})}\Big) \\
&\quad \sum_{\boldsymbol{s}} \mathrm{P}\Big(\boldsymbol{u}^{(\mathrm{A})}|\boldsymbol{s},\theta\Big)\mathrm{P}\Big(\boldsymbol{u}^{(\mathrm{B})}|\boldsymbol{s},\theta\Big)\mathrm{P}(\boldsymbol{s}|\theta)
\end{aligned} \tag{21}$$

この最適化問題を解析的に解くことは困難であるが，以下のように EM アルゴリズムと補助関数法に基づく反復法により局所最適解を求めることができる．目的関数の対数を $L\big(\theta, \boldsymbol{u}^{(\mathrm{A})}, \boldsymbol{u}^{(\mathrm{B})}\big)$ と置く．

$$\begin{aligned}
&L\Big(\theta, \boldsymbol{u}^{(\mathrm{A})}, \boldsymbol{u}^{(\mathrm{B})}\Big) \\
&= \log \mathrm{P}\Big(\boldsymbol{y}^{(\mathrm{A})}|\boldsymbol{u}^{(\mathrm{A})}\Big) + \log \mathrm{P}\Big(\boldsymbol{y}^{(\mathrm{B})}|\boldsymbol{u}^{(\mathrm{B})}\Big) \\
&+ \log \sum_{\boldsymbol{s}} \mathrm{P}\Big(\boldsymbol{u}^{(\mathrm{A})}|\boldsymbol{s},\theta\Big)\mathrm{P}\Big(\boldsymbol{u}^{(\mathrm{B})}|\boldsymbol{s},\theta\Big)\mathrm{P}(\boldsymbol{s}|\theta)
\end{aligned} \tag{22}$$

第 1 項は Jensen の不等式

$$-\left(\sum_{\tau,i} x_i[\tau]\right)^2 \geq -\sum_{\tau,i} \frac{x_i[\tau]^2}{c_i[\tau]} \tag{23}$$

$$\sum_{\tau,i} c_i[\tau] = 1, c_i[\tau] \geq 0 \tag{24}$$

より，

$$\begin{aligned}
&\log \mathrm{P}\Big(\boldsymbol{y}^{(\mathrm{A})}|\boldsymbol{u}^{(\mathrm{A})}\Big) \qquad (25) \\
&\overset{c}{=} -\sum_{t=1}^{T} \frac{\Big(y^{(\mathrm{A})}[t] - \sum_{\tau,i}\Big(G_i[\tau]u^{(\mathrm{A})}_i[t-\tau]\Big)\Big)^2}{2\upsilon^{(\mathrm{A})}_{\mathrm{n}}[t]^2} \\
&\geq -\sum_{t=1}^{T} \frac{\Big(\sum_{\tau,i} \frac{w^{(\mathrm{A})}_i[\tau]^2}{c^{(\mathrm{A})}_i[\tau]} + 2y^{(\mathrm{A})}[t]w^{(\mathrm{A})}_i[\tau] + y^{(\mathrm{A})}[t]^2\Big)}{2\upsilon^{(\mathrm{A})}_{\mathrm{n}}[t]^2}
\end{aligned}$$

により上限関数が設計できる．ここで $\overset{c}{=}$ は定数項を除き等しいことを表す．ただし，$w^{(\mathrm{A})}_i[\tau] = G_i[\tau]u^{(\mathrm{A})}_i[t-\tau]$ と置いた．等号は

$$c^{(\mathrm{A})}_i[\tau] = \frac{w^{(\mathrm{A})}_i[\tau]}{\sum_{\tau,i} w^{(\mathrm{A})}_i[\tau]} \tag{26}$$

のとき成り立つ．第 2 項も同様に上限関数を設計できる．第 3 項は Baum-Welch アルゴリズムにより

更新則を導出できる．以上をまとめると，上限関数 $L^{+}\left(\theta, \boldsymbol{u}^{(\mathrm{A})}, \boldsymbol{u}^{(\mathrm{B})}, \boldsymbol{c}\right)$ は

$$
\begin{aligned}
& L^{+}\left(\theta, \boldsymbol{u}^{(\mathrm{A})}, \boldsymbol{u}^{(\mathrm{B})}, \boldsymbol{c}, \boldsymbol{\gamma}\right) \\
= & \\
& -\sum_{t=1}^{T} \frac{\left(\sum_{\tau, i} \frac{w_i^{(\mathrm{A})}[\tau]^2}{c_i^{(\mathrm{A})}[\tau, t]} + 2 y^{(\mathrm{A})}[t] w_i^{(\mathrm{A})}[\tau]\right)}{2 v_{\mathrm{n}}^{(\mathrm{A})}[t]^2} \\
& -\sum_{t=1}^{T} \frac{\left(\sum_{\tau, i} \frac{w_i^{(\mathrm{A})}[\tau]^2}{c_i^{(\mathrm{A})}[\tau, t]} + 2 y^{(\mathrm{A})}[t] w_i^{(\mathrm{A})}[\tau]\right)}{2 v_{\mathrm{n}}^{(\mathrm{B})}[t]^2} \\
& -\sum_{t=1}^{T} \sum_{i \in (p,a)} \frac{\left(u_i^{(\mathrm{A})}[t] - \sum_{s[t]=1}^{S} \gamma_{s[t]} \mu_{s[t]}^{(\mathrm{A})}\right)^2}{2\left(\sigma_i^{(\mathrm{A})}\right)^2} \\
& -\sum_{t=1}^{T} \sum_{i \in (p,a)} \frac{\left(u_i^{(\mathrm{B})}[t] - \sum_{s[t]=1}^{S} \gamma_{s[t]} \mu_{s[t]}^{(\mathrm{B})}\right)^2}{2\left(\sigma_i^{(\mathrm{B})}\right)^2} \\
& + C
\end{aligned}
\tag{27}
$$

と表せる．更新に関係の無い項は C と置いた．ここで，$\gamma_{s[t]}$ は時刻 t に状態 s にいる事後確率であり，Forward-Backward アルゴリズムにより効率的に計算される．補助変数 $\boldsymbol{c}, \boldsymbol{\gamma}$ とモデル変数 θ を交互に更新することで目的関数は単調増加するため，局所最適解に収束することが保証される．

3.3 提案モデルに基づく韻律変換

F_0 軌跡を変換する問題は，$\boldsymbol{y}^{(\mathrm{A})}$ が与えられたもとで $\hat{\boldsymbol{y}}^{(\mathrm{B})}$ を推定する問題となる．

$$
\begin{aligned}
\hat{\boldsymbol{y}}^{(\mathrm{B})} &= \underset{\boldsymbol{y}^{(\mathrm{B})}}{\operatorname{argmax}} \mathrm{P}\left(\boldsymbol{y}^{(\mathrm{B})} | \boldsymbol{y}^{(\mathrm{A})}, \theta\right) \\
&= \underset{\boldsymbol{y}^{(\mathrm{B})}}{\operatorname{argmax}} \int \sum_{\boldsymbol{s}} \mathrm{P}\left(\boldsymbol{y}^{(\mathrm{B})} | \boldsymbol{u}^{(\mathrm{B})}, \theta\right) \mathrm{P}\left(\boldsymbol{u}^{(\mathrm{B})} | \boldsymbol{s}, \theta\right) \\
&\quad \mathrm{P}\left(\boldsymbol{s} | \boldsymbol{u}^{(\mathrm{A})}, \theta\right) \mathrm{P}\left(\boldsymbol{u}^{(\mathrm{A})} | \boldsymbol{y}^{(\mathrm{A})}, \theta\right) \mathrm{d}\boldsymbol{u}^{(\mathrm{A})} \mathrm{d}\boldsymbol{u}^{(\mathrm{B})}
\end{aligned}
\tag{28}
$$

簡単のため $\boldsymbol{u}$ に関する積分と $\boldsymbol{s}$ に関する総和を最大値で置き換えると，

$$
\begin{aligned}
\hat{\boldsymbol{y}}^{(\mathrm{B})} = & \underset{\boldsymbol{y}^{(\mathrm{B})}, \boldsymbol{u}^{(\mathrm{A})}, \boldsymbol{u}^{(\mathrm{B})}, \boldsymbol{s}}{\operatorname{argmax}} \mathrm{P}\left(\boldsymbol{y}^{(\mathrm{B})} | \boldsymbol{u}^{(\mathrm{B})}, \theta\right) \mathrm{P}\left(\boldsymbol{u}^{(\mathrm{B})} | \boldsymbol{s}, \theta\right) \\
& \mathrm{P}\left(\boldsymbol{s} | \boldsymbol{u}^{(\mathrm{A})}, \theta\right) \mathrm{P}\left(\boldsymbol{u}^{(\mathrm{A})} | \boldsymbol{y}^{(\mathrm{A})}, \theta\right)
\end{aligned}
\tag{29}
$$

これは $\boldsymbol{u}^{(\mathrm{A})}, \boldsymbol{s}, \boldsymbol{u}^{(\mathrm{B})}, \boldsymbol{y}^{(\mathrm{B})}$ の順に最大値を求めることに相当する．$\boldsymbol{u}^{(\mathrm{A})}, \boldsymbol{s}$ は我々が提案した手法により局所最適値を求めることができる．

4 提案手法の適用例

提案手法を用いてモデルを学習し，クローズドデータに対して変換処理を行った一例を Fig. 4 に示す．変換元の音声は男声，参照音声は女声である．参照話者の女性は変換元の男声よりアクセントをはっきり付ける話し方をする傾向があり，変換後の軌跡にもその特徴が現れている．このことは，提案法が発話の個人性を転写する能力をもつ可能性を示唆している．

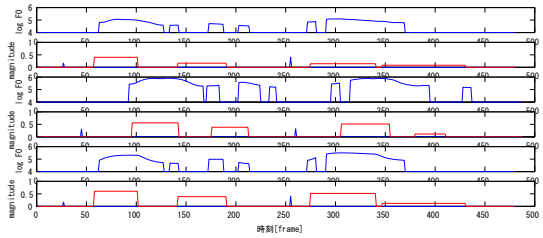

Fig. 4 提案手法の適用例（大声を出しすぎて，かすれ声になってしまう.） *1* 段目： 変換前の F_0 軌跡 *2* 段目： 変換前の F_0 軌跡の推定指令列 *3* 段目： 参照 F_0 軌跡 *4* 段目： 参照 F_0 軌跡の推定指令列 *5* 段目： 変換後の F_0 軌跡 *6* 段目： 変換された指令列

5 結論

F_0 軌跡の変換およびそのための学習則を，F_0 軌跡の同時生成モデルに基づき導出した．提案手法を実音声に適用した結果，提案手法が発話の個人性を転写する能力をもつ可能性を示した．

参考文献

[1] T. Toda et al., Audio, Speech, and Language Processing, IEEE Transactions on (Volume:15 , Issue: 8)

[2] H. Fujisaki, Raven Press, 1988.

[3] H. Fujisaki et al., Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference, Vol. 2, p.p. 634 637, 1996.

[4] 石原 他, 日本音響学会 2013 年春季研究発表会講演論文集, 1-7-9, Mar. 2013.

[5] T. Ishihara et al, in Proc. The 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), Aug. 2013.

[6] H. Kameoka et al, in *Proc. SAPA*, 2010, pp. 43–48.

[7] K. Yoshizato et al, in Proc. The 13th Annual Conference of the International Speech Communication Association (Interspeech 2012), Sep. 2012.