修士論文

出力特徴量の状態識別と
長時間特徴量を用いた
区分的線形変換に基づく声質変換

2014 年 2 月 6 日

指導教員　広瀬 啓吉 教授

東京大学大学院 情報理工学系研究科
電子情報学専攻

48-126403　池島 純

内容梗概

声質変換とは、任意の文の入力音声に対して、所望の声質へと変換する技術のことである。声質変換は、話者変換をはじめとして、電話音声などの帯域拡張、雑音除去などの様々な目的で研究されているが、本稿では話者変換を目的とする。声質変換では音声信号を直接変換するのではなく、高さ、強さなどに相当するパラメータである特徴量を音声から抽出し、それを変換することで実現する。しかし、話者性などの非言語情報や雑音、残響などの影響で話者入力特徴量と出力特徴量の関係は非線形になる。この非線形な関数をモデル化する手法の一つとして区分的線形変換が知られている。区分的線形変換では、狭い領域において非線形変換を線形変換で近似可能であるとする局所線形性を仮定した手法であり、以下の二つのステップに分けられる。一つ目は特徴量をいくつかの領域に分割する機能であり、二つ目はそれぞれの領域ごとに線形変換を行う機能である。従来手法では、領域分割は入力話者の特徴量空間または、入出力特徴量を結合した特徴量空間内で行われている。しかし、これらの手法は入力特徴量空間と出力特徴量空間のクラスタリングによる偏りが同一であることが暗に仮定されていた。ここで、特に入出力特徴量空間全体の空間分割が大きく異なるような場合には、出力特徴量空間の領域分割がより重要になると考えられる。

そこで本論文では出力話者の特徴量空間を分割を行う手法を提案する。また、声質変換でも音声認識の時と同様に長時間特徴量は大事であると考えられるので、識別と変換の過程において時間的に連続なフレームの特徴量を特徴量ベクトルに加えた。提案手法の有効性を確かめるために、客観評価実験と主観評価実験を行った。客観評価実験では、提案手法は従来手法よりよい結果を示した。主観評価実験では、音声の自然性と話者性を評価し、従来手法よりわずかに良くなったものの、有意な差は得られなかった。

目次

図目次

表目次

第1章

はじめに

1.1　研究の背景

声質変換は、任意の文に対して入力音声の声質を所望の声質へと変換する技術である。特に入力発声の話者性を制御する話者変換はテキスト音声合成への応用を視野に数多く研究されている [1]-[3]。

たとえ同一の言語内容を表す音声であっても、音声に含まれる話者性などの非言語情報や、雑音や残響等の影響によって表出する特徴量は大きく異なる。一般にこれらの要因による特徴量の変化をモデル化する際、変化の前後は非線形な変換関係によって記述される。声質変換や雑音抑圧等の分野では、事前に入出力の対応のとれたデータ（パラレルデータ）を用いてこの非線形な変換関係を統計的にモデル化する統計的特徴量変換が広く用いられている。非線形変換のモデル化にはニューラルネットワークを用いた直接的なモデル化の他 [2] [3]、特に音声特徴量空間の局所線形性に着眼した手法が数多く研究されている [4] [5]。後者では、非線形変換が区分された領域毎の線形変換の重ね合わせによって表現されると仮定し、ベクトル量子化や混合正規分布 (Gaussian Mixture Model; GMM) による特徴量空間のクラスタリングとこれらのクラスタに対応する線形変換の推定という二つの機能によって構成される。特に GMM を用いた区分的線形変換は声質変換や雑音抑圧など、多岐にわたる音声情報処理に応用されている [4] [5] [6]。本稿では、以降この技術を GMM マッピングと呼び、GMM マッピングにおける特徴量空間のクラスタリング及び変換の推定の両面から議論する。

声質変換における GMM マッピングの具体的な実装法として入力特徴量と出力特徴量を結合した結合ベクトルの GMM を用いる手法が広く利用されている [4] [1]。この手法では、結合ベクトルの空間を GMM を用いてクラスタリングする。ここで GMM 中の各正規分布に対応する領域に一つの線形変換が対応する。変換時には入力特徴量が与えられた場合の各要素分布に対する事後確率を求め、各要素分布に対応する線形変換の、事後確率を重みとする重み付き和として変換を実現する。

一方、雑音抑圧応用における GMM マッピングの実装法としては、SPLICE（Stereo-based Piecewise LInear Compensation for Environments）と呼ばれる手法が広く利用されている [5]。SPLICE では，入力特徴量で学習した GMM を用いて GMM マッピングを行う。さらに近年、SPLICE の GMM の要素正規分布に対する事後確率の計算を、変換対象となるクリーン音声 GMM における要素分布の識別問題として取り扱う手法が提案された [7]。これにより、従来の SPLICE と比較して雑音環境下における音声認識精度が向上したこと確認されている。

本稿では、[7] で提案された GMM マッピング技術の効果を声質変換において検証する。声質変換においては従来は入力特徴量や、入力特徴量と出力特徴量を結合した結合ベクトルに基づいて GMM による特徴量クラスタリングを行う手法が一般的であった。しかし、これらの手法には入力特徴量空間と出力特徴量空間のクラスタリングによる偏りが同一であるという仮定が暗に存在し、変換時には入力特徴量空間における特徴量空間分割の情報をそのまま用いている。一方、変換を出力特徴量の生成とみなすと、特に入出力特徴量空間全体の空間分割が大きく異なるような場合には、出力特徴量空間の領域分割がより重要

になると考えられる。

また、本研究では声質変換における長時間特徴量の利用も合わせて検討する。音声認識においてフレーム特徴量に対しても、200ms 程度の範囲の長時間情報を利用する事の有効性が示されている [8]。声質変換においても劣化特徴量の変換等において、10 フレーム前後の特徴量を連結し次元圧縮した特徴量（セグメント特徴量）が利用されているが、本研究では特徴量空間分割および変換の推定の両面において長時間特徴量を利用する効果を検証した。

1.2　本論文の構成

本論文は全 6 章から構成される。まず 2 章では音声情報処理の基礎について述べる。3 章では声質変換の分野で広く用いられている技術と、従来の声質変換方法について紹介を行う。4 章では、提案手法について述べ、5 章で客観評価実験と主観評価実験を行い提案手法の有効性を検証した。最後に 6 章で、本論文のまとめについて述べる。

第2章

音声工学に関する技術

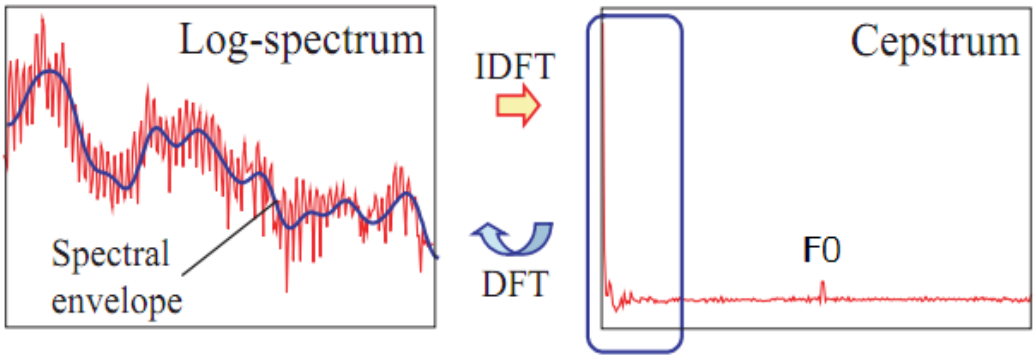

図 2.1: パワースペクトルとケプストラム

音声には、音韻情報の他にも話者情報や発声者の感情や意図など、様々な情報が含まれている。人間は音声の高さ、大きさ、長さ、声色などを変化させることによってそれらの情報を伝えている。

音声工学の主な目的は、人間が行っている音声認識や音声生成などを計算機に行わせることである。計算機においても音声波形を直接扱うのは困難なため、音声波形から声の高さなどの特徴に相当する値である特徴量を抽出しそれを操作することによってこれを実現する。

本章では音声から特徴量を抽出するための前提条件となっているソースフィルタモデルについて説明し、代表的な特徴量について説明する。そして音声特徴量のモデル化する際に広く用いられている GMM について説明する。

2.1　ソースフィルタモデル

音声は、声帯の震えや狭めで生じた乱流などで音声の元となる声帯音源が生じ、それが声帯を通ることによって周波数が加工されたものである。音源成分は、主にアクセントやイントネーションなどといった韻律と関係しており、声帯形状によって決定される周波数特性は音韻情報と関わりが深い。ソースフィルタモデルはこのような人間の音声発声過程をモデル化したものであり、音声 $s(t)$ を声帯振動の駆動音源波 $e(t)$ と、声道のインパルス応答 $h(t)$ の畳み込みで表現する。

$$s(t) = \int e(t) \oplus h(t) \mathrm{d}t \tag{2.1}$$

ただし $\oplus$ は畳み込みである。これをフーリエ変換し、パワーをとると以下の式になる。

$$|S(\omega)| = |E(\omega)| + |H(\omega)| \tag{2.2}$$

$S(\omega)$、$E(\omega)$、$h(\omega)$ はそれぞれ $s(t)$、$e(t)$、$h(t)$ をフーリエ変換したものである。図 2.1 の左側は有声音声の波形に対して、短時間フーリエ変換を行いパワーをとったパワースペクトルである。赤い線であらわされる細かい波が $|E(\omega)|$ に対応する部分であり、有声音の場

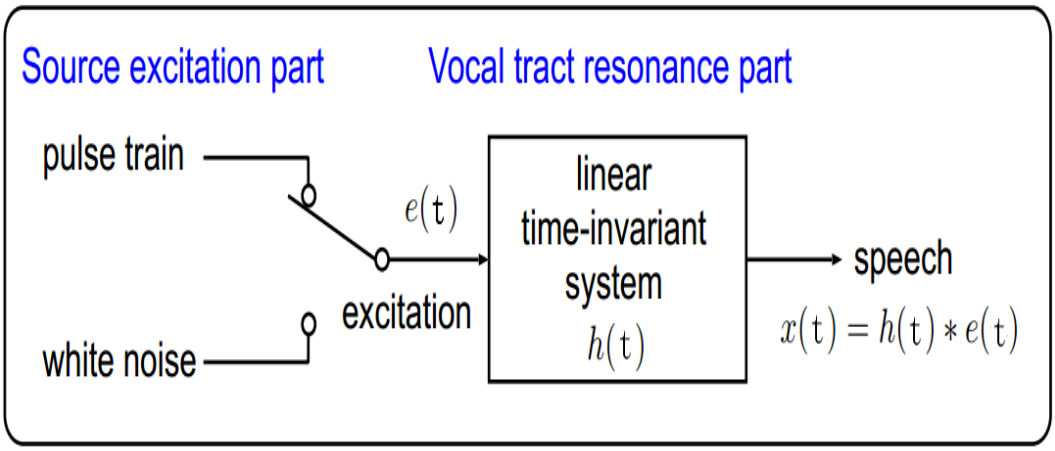

図 2.2: ソースフィルタモデル

合は図のように周期的になっている。この周期性はおよそ音の高さに相当する成分であり、声帯の振動の周期の逆数を基本周波数と呼ぶ。ただし無声音の場合は声帯振動が存在しないので基本周波数は定義されない。音声合成などでは、有声音の音源波形はインパルス列や三角波などで近似され、無声音では白色雑音などで近似されることが多い。図 2.1 の左側青い線で示されているスペクトルの包絡線が $|E(\omega)|$ に対応する成分であり、音韻などの情報と関係している。以上の事を踏まえ、ソースフィルタモデルに基いた音声の生成過程の例を図 2.2 に示す。

このように、音声を声帯に起因する成分と声道の周波数特性に起因する成分の 2 つの独立した成分に分けて考えることによって、それぞれの成分に対応する特徴量の抽出やモデル化などが可能となる。

2.2　音声工学で用いられる特徴量

前節で説明した通り、音声はソースフィルタモデルによって独立した 2 つの成分に分けられる。本章ではまず、音源に起因するスペクトルの鋸状の波に対応する特徴量について説明し、声道形状に起因するスペクトル包絡に対応する特徴量について説明する。

2.2.1　基本周波数

前章でも説明したが、声帯に起因する成分に対応する特徴量は、基本周波数である。基本周波数の定義は、声帯の振動周期に逆数をとったものであり、声帯の振動を伴わない無声音などの区間では定義されない。基本周波数はおよそ音の高さに相当する特徴量である。音の高さはアクセント、イントネーションや感情、発話スタイルなどの情報と深く関係しており、これらの情報は比較的長時間に渡って表れるので超文節的特徴と呼ばれる。

基本周波数の抽出方法として代表的なものに、音声波形の周期性を自己相関関数を利用して求める手法 [9] や、図 2.1 の右側のように、パワースペクトルを逆フーリエ変換を行っ

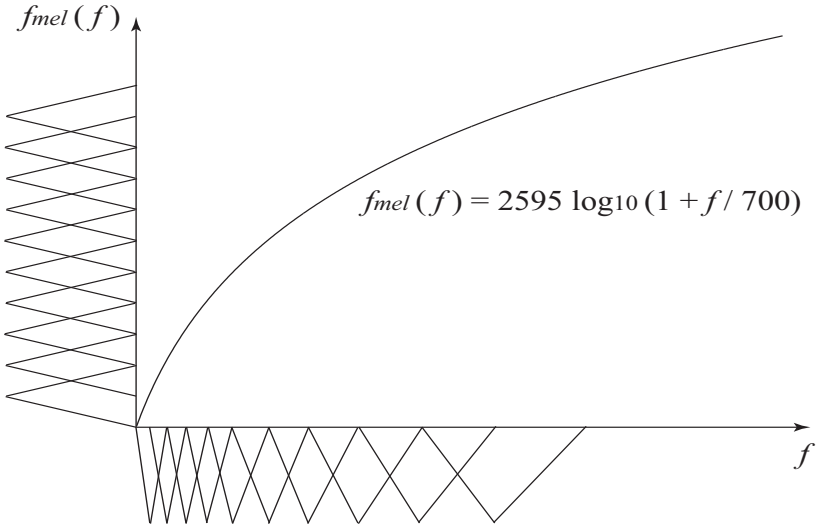

図 2.3: MFCC に用いるフィルタバンク

たケフレンシー軸で、ピークを求める手法 [10] がある。

また発話スタイルなど比較的長時間に存在する情報を扱い易くするために、フレームごとに捉えるのではなくそのパターンをモデル化したものがある。基本周波数パターンのモデルとしては、基本周波数パターン生成過程モデル [11] の他、ペンタモデル [12] や声調言語の音声をモデル化した Tone nucleus model [13] や Tonal Tilt Model [14] などがある。

2.2.2　スペクトル包絡特徴量

スペクトル包絡を表す特徴量は声色を表し、音韻性などの情報と関係が深い。スペクトル包絡を表す特徴量はより少ないパラメータでスペクトル包絡を表現する事が求められ、現在まで様々な特徴量が提案されている。

代表的な特徴量として、図 2.1 の右側に示したようにケフレンシー軸の低次元部分を用いるケプストラムが知られている。また、ケプストラムを改良し、人間の感覚を考慮したメル周波数を利用するものがある。メル周波数は人間が知覚する音の高さとメル周波数が比例するよう設定した。周波数 f とメル周波数 f_{mel} は以下の式で近似できると知られている。

$$f_{\mathrm{mel}}(f) = 2595 \log(1 + \frac{f}{700}) \tag{2.3}$$

この式から人間の知覚はおよそ対数的であるといえる。従って低い周波数は敏感であるので、重要度が高い。

このメル周波数を利用した特徴量として MFCC(Mel Frequency Cepstral coefficient) や MGC(Mel genaralized cepstram) [15] がある。MFCC はパワースペクトルに図 2.3 に示すメル周波数で等分されるようにフィルタバンクをかけ、逆コサイン変換を行った係数である。MFCC は音声認識では適しているが、スペクトル包絡の復元が難しいため音声を再合成するタスクには適さない。

音声合成に用いられる代表的な特徴量としては、MGC(mel generalized cepstrum) がある。まず、MGC の元となっている全極型モデルと極零型のモデルについて説明する。全極型では声道特性 H(z) は以下の式で表される。

$$H(z) = \frac{1}{1 - \sum_m c_{\gamma(m)} z^{-m}} \tag{2.4}$$

全極モデルは人間が知覚する上で重要とされるスペクトル包絡のピークを低次元でも上手く表現できるが、解が保証されない。逆に極零型では声道特性は以下の式で表され、

$$H(z) = \exp \sum_m c_{\gamma(m)} z^{-m} \tag{2.5}$$

解は保証されるものの、比較的なだらかな曲線になってしまうため、ピークを表現するためには次元数を増やす必要がある。この 2 つのモデルを統合したモデルの係数 $c_{\alpha.\gamma}$ が MGC である。フィルタは以下の式で表現する。

$$H(z) = (1 + \sum_m c_{\alpha,\gamma}(m) z_\alpha^{-1})^{\frac{1}{\gamma}} \tag{2.6}$$

$$z_\alpha^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z_\alpha^{-1}} \tag{2.7}$$

γ が 0 の場合が全極型に相当し、 γ が -1 の場合が極零型に相当する。今回使用した 16 kHz の音声の場合は $\alpha = 0.42$ 程度の時 z_α^{-1} がメル尺度に対応する。

2.3　Gaussian Mixture Model

自然界や人間社会の事象を記述するために用いられるモデルとしてガウス分布が知られている。特徴量 $\boldsymbol{x}$ の分布がガウス分布に従う場合、式 (2.8) で表す事が出来る。

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}\sqrt{|\boldsymbol{\Sigma}|}} \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})) \tag{2.8}$$

ただし、D は特徴量の次元であり、$\boldsymbol{\mu}$、$\boldsymbol{\Sigma}$ はそれぞれ正規分布の平均ベクトルと分散共分散行列である。音声特徴量のモデル化を考えると、音声には音韻性に代表される多数の情報が存在するために、単一のガウス分布では表現しきれない。そこで音声特徴量のモデル化にはガウス分布の足し合わせである GMM を用いることが多い。GMM は以下の式で表すことができる。

$$P(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \tag{2.9}$$

ただし、 k はガウス分布のインデックスで、$\boldsymbol{\mu_k}$、$\boldsymbol{\Sigma_k}$ はそれぞれ k 番目のインデックスのガウス分布の平均ベクトルと分散共分散行列、π_k は k 番目のガウス分布の重みであり、確

率の和が 1 にするために以下の制約がある。

$$\sum_{k=1}^{K} \pi_k = 1 \tag{2.10}$$

次にフレームインデックスを $i = [1, \ldots, I]$ としたとき、特徴量系列 $\boldsymbol{x} = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_T}]$ から GMM を学習する方法を説明する。尤度最大化基準で学習することが一般的であるが、 GMM は隠れ変数が存在するため、解析的に解けるアルゴリズムは存在しない。そこで逐次的に求めていく EM アルゴリズムを適用する [16]。学習するパラメータ集合を $\lambda = \{\boldsymbol{\Sigma}_k, \boldsymbol{\mu}_k, \pi_k\}$ とすると、対数尤度最大化基準で以下のように解く。

$$\begin{aligned}
\hat{\lambda} &= \underset{\lambda}{\operatorname{argmax}} \sum_{t=1}^{T} \log P(\boldsymbol{x_t}|\lambda) \\
&= \underset{\lambda}{\operatorname{argmax}} \sum_{t=1}^{T} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_t, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) && (2.11) \\
&\geq \underset{\lambda}{\operatorname{argmax}} \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_k(\boldsymbol{x}_t) \log \pi_k \mathcal{N}(\boldsymbol{x}_t, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) && (2.12)
\end{aligned}$$

式 (2.15) の不等式は Jensen の不等式を用いており。等号が成立するのは

$$\begin{aligned}
\gamma_k(\boldsymbol{x}_t) &= P(k|\boldsymbol{x}_t, \lambda) && (2.13) \\
&= \frac{\pi_k \mathcal{N}(\boldsymbol{x}_t, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_t, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})} && (2.14)
\end{aligned}$$

のときである。式 (2.11) を解析的に解くことが不可能である。EM アルゴリズムでは、式 (2.15) の argmax 以下を増大させていく事によって間接的に式 (2.11) を最大化することを目指す。最大化を行う式 (2.15) の argmax 以下は Q 関数と呼ぶ。

$$Q(\hat{\lambda}|\lambda) = \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_k(\boldsymbol{x}_t) \log \pi_k \mathcal{N}(\boldsymbol{x}_t, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \tag{2.15}$$

Q 関数の最大化は、Q 関数の左辺である $\gamma_k(\boldsymbol{x}_t)$ を計算する E ステップと、$\gamma_k(\boldsymbol{x}_t)$ を固定値と考えて Q 関数を最大化していく M ステップを交互に行うことによって求める。E ステップの $\gamma_k(\boldsymbol{x}_t)$ の更新式は式 (2.14) であり、M ステップのパラメータの最大化は、式 (2.15) の Q 関数を微分することによって求められる。(π_k は拘束条件である式 (2.10) が存在するためラグランジュ未定乗数法を用いる)。計算された更新式は以下になる。

$$\begin{aligned}
\pi_k &= \frac{\gamma_k}{\sum_{t=1}^{T} \gamma_k} && (2.16) \\
\boldsymbol{\mu_k} &= \frac{\sum_{t=1}^{T} \gamma_k \boldsymbol{x_t}}{\sum_{t=1}^{T} \gamma_k \boldsymbol{x}} && (2.17) \\
\boldsymbol{\Sigma_k} &= \frac{\sum_{t=1}^{T} \gamma_k (\boldsymbol{x_t} - \boldsymbol{\mu_k})(\boldsymbol{x_t} - \boldsymbol{\mu_k})^\top}{\sum_{t=1}^{T} \gamma_k} && (2.18)
\end{aligned}$$

パラメータの更新は誤差基準やイテレーションの回数など基準を収束条件として決めて終わらせる。

EM アルゴリズムは逐次的にパラメータを求めていくため局所解に陥ってしまう可能性がある。そのため、学習結果のモデルは初期値に大きく依存する。初期値の例としては、ランダムスタートの他、K-means の結果を利用する方法、混合数を段階的に増やしていく方法などがある。GMM は簡単なモデルの足し合わせで表現できるために、比較的計算コストが少なく、拡張性が高いなどの理由で音声工学で広く用いられている

第3章

声質変換に関する技術

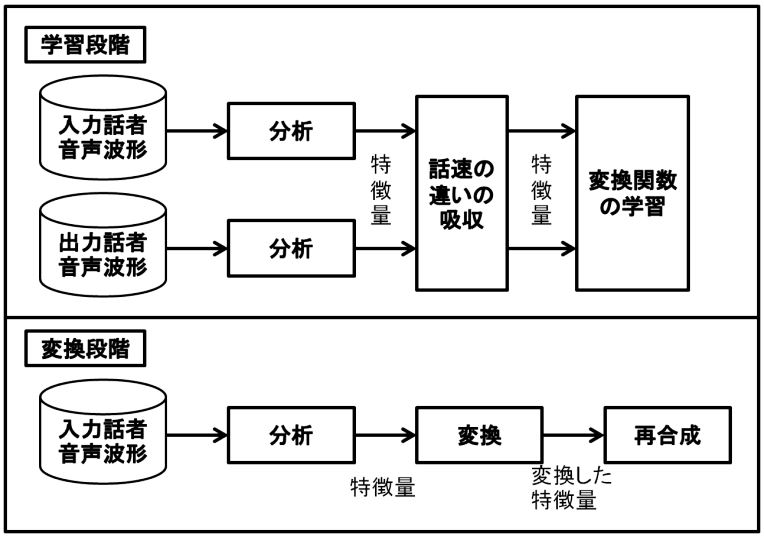

図 3.1: 一般的な声質変換手法の流れ

3.1　はじめに

本章では声質変換に関する技術について説明する。まず、統計的声質変換の大まかな流れを図 3.1 に示す。声質変換は大きく分けると、学習段階と変換段階に分けられる。学習段階では事前に用意した学習音声を用いて変換モデルを学習する。変換段階では学習したモデルを用いて変換を行う。学習に利用する音声コーパスとしては、入力話者の発声した音声と、それと同一文を読み上げた出力話者の発声した音声のパラレル音声を用いることが多い。同一発話内容の制約はしばしば達成が困難な場合もあるため、それを必要としない研究も存在するが [17]、本稿ではパラレル音声の存在を前提とする。

学習段階ではまず、音声から特徴量を抽出する。音声の特徴の中で話者性に最も影響を与える特徴は声色であると考えらえれるため、音声変換では声色の変換を扱う研究が多い。学習用の入力音声と出力音声から抽出した入力特徴量と出力特徴量を学習に用いるが、そのままでは入力音声と出力音声の話速の違いが原因で入出力の対応関係が上手く学習できない。したがって、話速の違いを吸収し、フレームごとに対応する入出力特徴量のペアを推定してから学習を行う。変換時は、音声から特徴量を抽出し、学習段階で求めた変換モデルのパラメータを用いて変換を行う。そして変換によって得られた特徴量を再合成することによって変換された音声を得る。

声質変換の流れは以上である。次節からはまず話速の違いの吸収方法について軽く説明してから、次に変換モデルとして広く利用されている区分的線形変換について説明し、以後の章では区分的線形変換を用いた具体的な手法について述べていく。

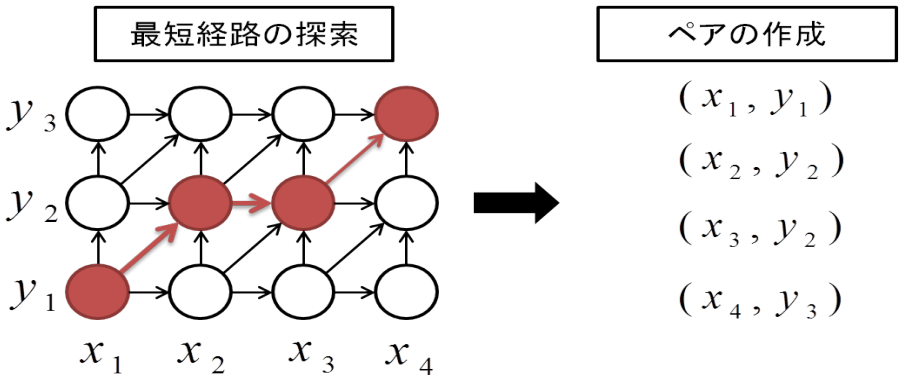

図 3.2: DP マッチング

3.2　話速の違いの吸収方法

入力音声と出力音声では話速が異なっているため、そのままでは入力特徴量と出力特徴量の対応関係が不明であり、変換モデルの学習が困難である。そのため、入力音声の話速と出力話者の話速の違いを吸収する必要がある。

話速の違いを吸収するための手法として、DP(Dynamic Programing) マッチングが広く用いられている。DP マッチングは動的計画法を用いて入力特徴量系列と出力特徴量系列の距離が最小となるようにペアを作っていく。$i = [1, \ldots, I]$ を入力特徴量のインデックス、$j = [1, \ldots, J]$ を出力特徴量のインデックスとしたとき、まず入力特徴量 $\boldsymbol{x}_i$ と出力特徴量 $\boldsymbol{y}_j$ の距離 $d(i, j)$ を計算する。本論文では距離尺度として、ケプストラム空間でのユークリッド距離を用いている。$g(i, j)$ を $(\boldsymbol{x}_1, \boldsymbol{y}_1)$ から $(\boldsymbol{x}_i, \boldsymbol{y}_j)$ までの特徴量系列の最短距離とするとき、$g(i, j)$ を動的計画法で計算する。ただし、極端なフレーム間の遷移を避けるため、以下では隣接するフレームへの遷移しか許さない経路制限を設ける。このとき、$g(i, j)$ は以下の式で計算できる。

$$g(i,j) = \min \begin{cases} g(i-1,j) + d(i-1,j) \\ g(i,j-1) + d(i,j-1) \\ g(i-1,j-1) + 2d(i-1,j-1) \end{cases} \tag{3.1}$$

また、$g(i, j)$ を求める際にどのフレームペアから計算されたかを記憶していく。これを繰り返し行い、$g(I, J)$ まで求める。$g(I, J)$ が $(\boldsymbol{x}_1, \boldsymbol{y}_1)$ から $(\boldsymbol{x}_I, \boldsymbol{y}_J)$ までの最短距離であり、そこに到達するまでの経路上にある $\boldsymbol{x}_i$、$\boldsymbol{y}_j$ をペアとする。DP マッチングを行う事により話速の違う特徴量から、音韻性の揃った特徴量系列のペアを得ることが可能である。また、経路制限を変えることによって、入力特徴量だけを伸縮するなども可能である。

3.3　区分的線形変換

線形変換は $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ で表されるものであり、入力と出力の関係は変換特徴量が 1 次元ならば直線、 2 次元ならば平面、3 次元以上ならば超平面となる。線形変換は非常に単

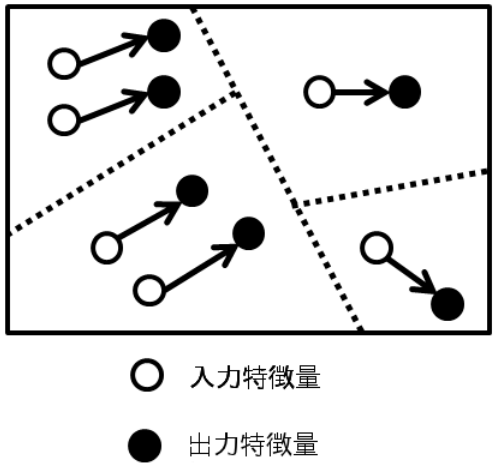

図 3.3: 区分的線形変換

純で計算コストが非常に低い反面、曲線などの非線形な関係を記述できない。そこで、非線形な関係を記述するための手法として、複数の線形変換を用いて非線形変換を表現する区分的線形変換がある。区分的線形変換はその前提として、狭い領域では非線形変換は線形変換で近似できるという局所線形性を仮定している。そして、区分的線形変換は以下の 2 つの機能で行われる。

- 特徴量空間をいつくかの領域へと分割する
- 分割された領域ごとに線形変換を行う

これを式で記述すると式 (3.2) となる。

$$\hat{\boldsymbol{y}} = \sum_{k=1}^{K} \gamma_k(\boldsymbol{x}) E_k(\boldsymbol{x}) \tag{3.2}$$

ただし、K は分割した領域の数であり、$\boldsymbol{x}$、$\hat{\boldsymbol{y}}$ はそれぞれ入力特徴量と推定された出力特徴量である。また、$E_k(\boldsymbol{x_t})$、$\gamma_k(\boldsymbol{x_t})$ は k 番目のインデクスの領域の線形変換とその重みであり、以下の条件を満たす。

$$\sum_{k=1}^{K} \gamma_k(\boldsymbol{x}) = 1 \tag{3.3}$$

区分的線形変換は非常に簡単なモデルの足し合わせであり、比較的計算コストが少ない。次節からは、この考えを声質変換に応用した具体的な手法をいくつか紹介する。

3.4　ベクトル量子化を用いた手法

声質変換の研究が始まったのは 1980 年代後半であるが、その頃に現在の研究の基礎となっている区分的線形変換を用いた手法がある。それはベクトル量子化 (Vector Quantization;

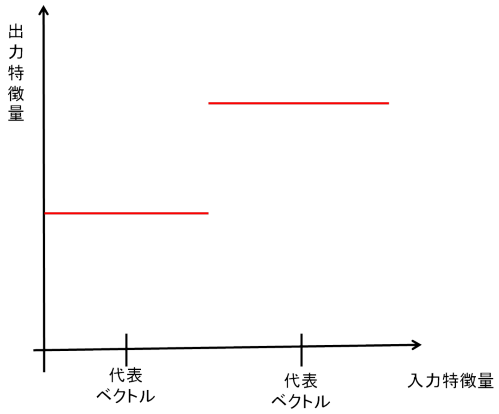

図 3.4: VQ を用いた変換の様子

VQ）を用いた手法である [18]。ベクトル量子化とは、特徴量を k クラスに分け、クラスに所属する特徴量をそのクラスの代表値に置き換える事である。ベクトル量子化を行うアルゴリズムとして K-means などがある。

以下で変換方法について詳しく説明する。まず、話速の吸収を行う前の入力特徴量と出力特徴量に対してそれぞれベクトル量子化を行う。ベクトル量子化した特徴量に対して、話速の違いを吸収するため DP マッチングを行い、その結果を用いてクラス間の対応付けのヒストグラムを作成する。そのヒストグラムから重み w_k を決定し、以下の式で入力特徴量の代表ベクトル $\boldsymbol{r}_{\boldsymbol{k}}^{(\boldsymbol{x})}$ に対応する変換後の特徴量 $\hat{\boldsymbol{y}}_k$ を決定する。

$$\hat{\boldsymbol{y}}_k = \sum_{k'=1}^{K} w_k(\boldsymbol{r}_{\boldsymbol{k}}^{(\boldsymbol{x})})\boldsymbol{r}_{k'}^{(y)} \tag{3.4}$$

ただし、$\boldsymbol{v}_k^{(\boldsymbol{y})}$ はそれぞれ出力特徴量の代表値である。入力特徴量の代表値 $\boldsymbol{r}_k^{(x)}$ と $\hat{\boldsymbol{y}}_k$ の関係をマッピングコードブックとして記録する。

変換時には、入力特徴量をベクトル量子化し、マッピングコードブックから対応する変換特徴量を取り出す。区分的線形変換の式 (3.2) に当てはめると、$\boldsymbol{\gamma}_k(x)$ は $\boldsymbol{x}$ がクラス k に所属する場合は 1 を返し違うクラスに所属される場合は 0 を返す関数であり、$\boldsymbol{E}_k(x)$ は $\hat{\boldsymbol{y}}_k$ を返す定数関数である。

この手法による変換の様子を図 3.4 に示した。マッピングコードブックに記憶されている特徴量しか出力できないため、変換の精度が悪く、また出力される特徴量は不連続になってしまう。したがってこれを改良するために、$\gamma(\boldsymbol{x})$ や $E(\boldsymbol{x})$ を連続的にする手法が提案されたが [19] [20]、特に 1990 年代後半に提案された GMM を用いた手法 [21] [4] は高品質な音声が合成できるため広く用いられるようになった。

3.5　GMM を用いた入力特徴量空間の領域分割に基づく手法

ベクトル量子化に基づく手法では、領域分割による線形変換の重みは 0 と 1 の 2 値としている。しかし、線形変換の重みを 2 値で扱うと、領域の境で不連続になり、音質の劣化

につながる。そこで、GMM を用いる事によって領域の境でも重みを連続的に変化させる手法が提案された。本節では GMM を用いた領域分割を行う手法のうち、最初に提案された入力特徴量による GMM を用いる手法（以下入力特徴量ベースの手法とする）を紹介する。この手法の変換は、以下の式で表わされる。

$$\hat{\boldsymbol{y}} = \sum_{k=1}^{K} \gamma_k(\boldsymbol{x})(\boldsymbol{A}_k \boldsymbol{x}_t + \boldsymbol{b}_k) \tag{3.5}$$

これは $E(\boldsymbol{x}) = \boldsymbol{A}_k \boldsymbol{x}_t + \boldsymbol{b}_k$ とすれば式 (3.2) と等しくなる。計算の簡略化のために、$E(\boldsymbol{x})$ の項を以下の式で表現する。

$$E_k(\boldsymbol{x}_t) = \tilde{\boldsymbol{A}}_k \tilde{\boldsymbol{x}}_t \tag{3.6}$$

$$\tilde{\boldsymbol{A}}_k = \begin{bmatrix} \boldsymbol{b}_k & \boldsymbol{A}_k \end{bmatrix} \tag{3.7}$$

$$\tilde{\boldsymbol{x}}_t = \begin{bmatrix} 1 \\ \boldsymbol{x}_t \end{bmatrix} \tag{3.8}$$

以下でパラメータの学習方法について説明する。フレームインデックスを $t = [1, \ldots, T]$ とし、入力特徴量を $\boldsymbol{x}_t$、出力特徴量を $\boldsymbol{y}_t$ とおく。入力特徴量の分布を GMM と仮定し、以下のようにおく。

$$P(\boldsymbol{x}_t) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_k^{(\boldsymbol{x})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{x})}) \tag{3.9}$$

ただし、π_k、$\boldsymbol{\mu}_k$、$\boldsymbol{\Sigma}_k$ は、k 番目の正規分布の重み、平均ベクトル、分散共分散行列で予め学習データを用いて学習しておく。GMM のパラメタは $\lambda^{(\boldsymbol{x})} = \{\pi_k, \boldsymbol{\mu}_k^{(\boldsymbol{x})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{x})}\}_{k=1,\ldots,K}$ とする。

この手法では、入力話者の特徴量 $\boldsymbol{x}_t$ が与えられた時のインデクス k の事後確率を線形変換の重みとする。

$$\gamma_k^{(\boldsymbol{x})}(\boldsymbol{x}_t) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_k^{(\boldsymbol{x})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{x})})}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_k^{(\boldsymbol{x})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{x})})} \tag{3.10}$$

線形変換のパラメタ $\theta^{(\boldsymbol{x})} = \{\boldsymbol{A}_k\}_{k=1,\ldots,L}$ はパラレルデータを用いて最小二乗誤差基準を基に学習する。最小二乗誤差基準は以下の式になる。

$$\hat{\boldsymbol{A}}_k = \underset{\boldsymbol{A}_k}{\operatorname{argmin}} \sum_{t=1}^{T} ||\boldsymbol{y}_t - \sum_{k=1}^{K} \gamma_k(\boldsymbol{x}_t) \boldsymbol{A}_k \tilde{\boldsymbol{x}}_t||^2 \tag{3.11}$$

この式は解析的に解くことができないので、近似した以下の式を解く。

$$\hat{\boldsymbol{A}}_k = \underset{\boldsymbol{A}_k}{\operatorname{argmin}} \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_k(\boldsymbol{x}_t) ||\boldsymbol{y}_t - \boldsymbol{A}_k \tilde{\boldsymbol{x}}_t||^2 \tag{3.12}$$

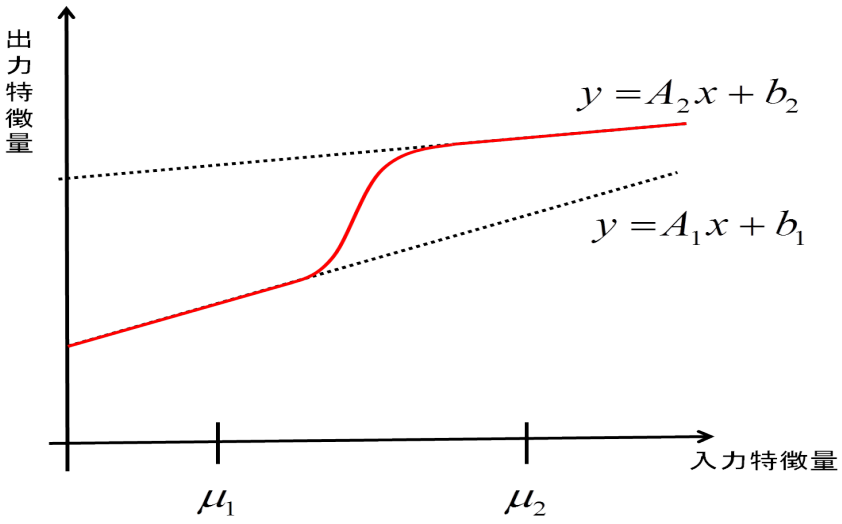

図 3.5: 入力特徴量に基づく手法の変換の様子

これは $\boldsymbol{A}_k$ で微分することによって簡単に計算でき、以下の式になる。

$$\hat{\boldsymbol{A}}_k = (\sum_{t=1}^{T} \gamma_k^{(\boldsymbol{x})}(\boldsymbol{x}_t)\boldsymbol{y}_t\tilde{\boldsymbol{x}}_t^\top)(\sum_{t=1}^{T} \gamma_k^{(\boldsymbol{x})}(\boldsymbol{x}_t)\tilde{\boldsymbol{x}}_t\tilde{\boldsymbol{x}}_t^\top)^{-1} \tag{3.13}$$

変換時には、式 (3.10) により、$\gamma_k^{(x)}(\boldsymbol{x}_t)$ を計算し、式 (3.6) から $E_k(\boldsymbol{x}_t)$ が計算でき、変換を行えばよい。

GMM の分散共分散行列 $\Sigma_k^{(\boldsymbol{x})}$ の次元数は対称行列であるので、入力特徴量の次元数 D_x とすると、$D_x(D_x+1)/2$ 次元になる。パラメータ数が多いので計算コストが高くなり、過学習が起こりやすくなる。そこで、次元間に相関が無いと仮定し、分散共分散行列 $\Sigma_k^{(\boldsymbol{x})}$ の対角成分以外はゼロにする方法もある。この仮定を置いて対角成分だけを学習するとパラメータ数は D_x 個となり大幅に削減出来る。

入力特徴量ベースの手法は、入力特徴量空間の領域分割を行って変換を行っている。また、領域分割によって得られる重み $\gamma_k(\boldsymbol{x}_t)$ を GMM を用いて計算することによって、入力特徴量が連続であれば出力される値も連続であることが保証される。また、線形変換の部分も分散を考慮した線形変換になっており、平均値だけを返す前節の手法よりもよい近似となっている。この変換の様子を図 3.5 に示した。領域の境界周辺は、両方の領域の線形変換の影響を受けた変換結果となる。

3.6　結合ベクトル空間の領域分割を用いた手法

前節では、入力特徴量空間を領域分割していたが、本節では、入力特徴量ベクトルと出力特徴量ベクトルを連結した結合ベクトル空間の領域分割によって変換する手法（以下結合ベクトルベースの手法）をについて述べる。

3.6.1　結合確率密度関数

フレームインデックスを $t = [1, \ldots, T]$ としたとき、入力特徴量を $\boldsymbol{x}_t$、出力特徴量を $\boldsymbol{y}_t$、そして結合ベクトルを $\boldsymbol{z}_t = [\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]^\top$、とおく。結合ベクトルの分布が GMM に従うと仮定し、以下の式でおく。

$$P(\boldsymbol{z}_t) = \sum_{k=1}^{K} \pi_k^{(\boldsymbol{z})} \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_k^{(\boldsymbol{z})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{z})}) \tag{3.14}$$

ただし、K は混合数を表す。また $\pi_k^{(z)}$、$\boldsymbol{\mu}_k^{(\boldsymbol{z})}$、$\boldsymbol{\Sigma}_k^{(\boldsymbol{z})}$ は、k 番目の正規分布の重み、平均ベクトルと分散共分散行列である。これらのパラメタは、予めパラレルデータを用いて学習される。$\boldsymbol{\mu}_k^{(\boldsymbol{z})}$ と $\boldsymbol{\Sigma}_k^{(\boldsymbol{z})}$ は、以下のように $\boldsymbol{x}$ と $\boldsymbol{y}$ に関連した要素に分解することができる。

$$\boldsymbol{\mu}_k^{(\boldsymbol{z})} = \begin{bmatrix} \boldsymbol{\mu}_k^{(\boldsymbol{x})} \\ \boldsymbol{\mu}_k^{(\boldsymbol{y})} \end{bmatrix} \tag{3.15}$$

$$\boldsymbol{\Sigma}_k^{(\boldsymbol{z})} = \begin{bmatrix} \boldsymbol{\Sigma}_k^{(\boldsymbol{xx})} & \boldsymbol{\Sigma}_k^{(\boldsymbol{xy})} \\ \boldsymbol{\Sigma}_k^{(\boldsymbol{yx})} & \boldsymbol{\Sigma}_k^{(\boldsymbol{yy})} \end{bmatrix} \tag{3.16}$$

前節同様、分散共分散行列のパラメータを全て学習する方法と、対角成分だけを学習する方法がある。対角成分だけを学習する方法では、$\Sigma_k^{\boldsymbol{z}}$ の対角成分ではなく、要素である $\Sigma_k^{\boldsymbol{xx}}$、$\Sigma_k^{\boldsymbol{xy}}$、$\Sigma_k^{\boldsymbol{yx}}$、$\Sigma_k^{\boldsymbol{yy}}$ を対角とする必要がある。

以上のように学習した結合ベクトルの GMM のパラメタ $\lambda^{(\boldsymbol{z})} = \{\pi_k^{(z)}, \boldsymbol{\mu}_k^{(\boldsymbol{z})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{z})}\}_{k=1\ldots M}$ を用いると、$\boldsymbol{x}_t$ が与えられた時の $\boldsymbol{y}_t$ の条件付き確率密度分布は、以下のようにして計算できる。

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t; \lambda^{(\boldsymbol{z})}) = \sum_{k=1}^{K} \gamma_{k,t}^{(\boldsymbol{z})} \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{E}_{k,t}, \boldsymbol{D}_k) \tag{3.17}$$

$$\gamma_{k,t}^{(\boldsymbol{z})} = \frac{\pi_k^{(z)} \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_k^{(\boldsymbol{x})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{xx})})}{\sum_{k=1}^{K} \pi_k^{(z)} \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_k^{(\boldsymbol{x})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{xx})})} \tag{3.18}$$

$$\boldsymbol{E}_{k,t} = \boldsymbol{\mu}_k^{(\boldsymbol{y})} + \boldsymbol{\Sigma}_k^{(\boldsymbol{yx})} \boldsymbol{\Sigma}_k^{(\boldsymbol{xx})^{-1}} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_k^{(\boldsymbol{x})} \right) \tag{3.19}$$

$$\boldsymbol{D}_k = \boldsymbol{\Sigma}_k^{(\boldsymbol{yy})} - \boldsymbol{\Sigma}_k^{(\boldsymbol{yx})} \boldsymbol{\Sigma}_k^{(\boldsymbol{xx})^{-1}} \boldsymbol{\Sigma}_k^{(\boldsymbol{xy})} \tag{3.20}$$

この条件付き確率を用いて変換を行う。変換基準として最小二乗誤差基準を用いる手法と尤度最大化基準で求める手法がある。

3.6.2　最小二乗誤差基準による変換手法

変換関数を最小二乗誤差基準で行う手法を紹介する [4]。最小二乗基準では、前項の条件付き確率の期待値を計算すればよい。

$$\hat{\boldsymbol{y}} = \int P(\boldsymbol{y_t}|\boldsymbol{x_t};\lambda^{(x)})\boldsymbol{y_t}d\boldsymbol{y_t} \tag{3.21}$$

$$= \sum_k \gamma_k^{(z)}(\boldsymbol{x}_t)\boldsymbol{E_{k,t}} \tag{3.22}$$

式 (3.22) の 2 つ目の項は線形変換であり、1 つ目の項がその重みとなっているので、区分的線形変換を行っていることに相当する。

3.6.3　尤度最大化基準による変換手法

変換を尤度最大化基準で行う手法を紹介する [22]。尤度最大化では解が解析的に解けないため、 EM アルゴリズムを適用する。対数尤度の最大化基準は以下の式になる。

$$\hat{\boldsymbol{y}}_{\boldsymbol{t}} = \underset{\boldsymbol{y_t}}{\operatorname{argmax}} \sum_{t=1}^{T} \log P(\boldsymbol{y_t}|\boldsymbol{x_t};\lambda^{(\boldsymbol{z})}) \tag{3.23}$$

この式から Q 関数を求めると以下の式になる。

$$Q(\hat{\boldsymbol{y}}_t|\boldsymbol{y}_t) = \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k^{(\boldsymbol{z'})}(\boldsymbol{z})(\boldsymbol{x_t}) \log P(\boldsymbol{y_t}|\boldsymbol{x_t},k;\lambda^{(z)}) \tag{3.24}$$

ただし、$\gamma_k^{(z')}(\boldsymbol{z})$ は以下の式で与えられる。

$$\gamma_k^{(z')}(\boldsymbol{z}_t) = P(k|\boldsymbol{x}_t,\hat{\boldsymbol{y}}_t;\lambda^{(\boldsymbol{z})}) \tag{3.25}$$

$$= \frac{\pi_k N(\boldsymbol{z}_t;\boldsymbol{\mu_k^{(z)}},\boldsymbol{\Sigma}_k^{(z)})}{\sum_{k=1}^{K}\pi_k N(\boldsymbol{z}_t;\boldsymbol{\mu_k^{(z)}},\boldsymbol{\Sigma}_k^{(z)})} \tag{3.26}$$

$\boldsymbol{y}_t$ で微分するために Q 関数を展開すると以下のようになる。

$$Q(\hat{\boldsymbol{y}}_t|\boldsymbol{y}_t) = \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k^{(z')}(\boldsymbol{x_t}) \log P(\boldsymbol{y_t}|\boldsymbol{x_t},k;\lambda) \tag{3.27}$$

$$= \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k^{(z')}(\boldsymbol{x_t}) \log \mathcal{N}(\boldsymbol{y_t};\boldsymbol{E_{k,t}},\boldsymbol{D}_{k,t}) \tag{3.28}$$

$$= \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k^{(z')}(\boldsymbol{x_t})(-\frac{1}{2}\log(\boldsymbol{D}_k) - \frac{1}{2}(\boldsymbol{y_t}-\boldsymbol{E}_{k,t})^\top \boldsymbol{D}_k^{-1}(\boldsymbol{y_t}-\boldsymbol{E}_{k,t})) + const \tag{3.29}$$

ただし const は $\boldsymbol{y}_t$ に依存しない項である。これは上に凸の関数なのでの最大値は、$\boldsymbol{y}$ で微分すればよい。

$$\boldsymbol{y_t} = (\sum_{k=1}^{K}\gamma_k(\boldsymbol{x_t})\boldsymbol{D}_k^{-1})^{-1}\sum_{k=1}^{K}\gamma_k\boldsymbol{x_t}\boldsymbol{D}_k^{-1}\boldsymbol{E}_{k,t} \tag{3.30}$$

以上より、式 (3.26) と式 (3.30) を交互に更新していくことによって、出力特徴量の推定値を得る。

3.6.4　最小二乗誤差基準と最尤推定基準の違い

ここで、前項の最尤推定による手法において、結合ベクトルで学習した GMM を構成する正規分布の分散が全て共通であるという仮定を付け加える。分散が共通であるので、$\boldsymbol{D}_k^{-1}$ が k に依存しない。したがって式 (3.29) は以下のようになる。

$$Q(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \boldsymbol{D}^{-1} \sum_{k=1}^{K} \gamma_k(\boldsymbol{x_t})((\boldsymbol{y_t} - \boldsymbol{E_k})^\top (\boldsymbol{y_t} - \boldsymbol{E_k})) + const \tag{3.31}$$

これを $\boldsymbol{y_t}$ で微分すると式 (3.22) と同様の結果が得られる。したがって、最小二乗誤差基準は、最尤推定基準の手法に結合ベクトルの GMM を構成する分散共分散行列の違いだけである。前者は全てのガウス分布で分散共分散行列が共通であり、後者はガウス分布ごとに分散共分散行列が違っている。変換結果は、最小二乗誤差基準による手法よりも、最尤推定による手法の方が良い結果となる [22]。

3.6.5　入力特徴量ベースの手法と結合ベクトルベースの手法の違い

最小二乗誤差基準に基づく結合ベクトルベースの手法と入力特徴量ベースの手法は、線形変換の部分は計算していくとほぼ等価となる。この 2 つの手法の大きく異なる点は、領域分割をどの特徴量空間で行っているかである。入力特徴量ベースでは、入力特徴量空間の領域分割を行っているため、線形変換の重みは入力特徴量のみに依存している。

$$\gamma_k(\boldsymbol{x}_t) = (k|\boldsymbol{x}) \tag{3.32}$$

結合ベクトルベースでは、結合ベクトル空間の領域分割を行っているため、線形変換の重みは入力特徴量と出力特徴量に依存している。

$$\gamma_k(\boldsymbol{x}_t) = (k|\boldsymbol{x}, \boldsymbol{y}) \tag{3.33}$$

これが、入力特徴量ベースの手法と結合ベクトルベースの手法の最大の違いである。

3.7　系列単位の変換手法

前節まで紹介した GMM を用いた変換手法は全てフレーム単位でパラメータを変換している。これらの手法では入力特徴量が連続であるときは、出力特徴量が連続であることが保証されている。しかし実際の音声特徴量は、雑音や特徴量の抽出ミスなどの要因で、入力特徴量が連続であるべき部分が不連続になっている事がある。したがって変換特徴量が不連続になるため、それが合成音声の不自然さの要因となる。出力特徴量の不連続性を解消するために、戸田らは前後のフレームとのパラメータの変動を動的特徴量を制約として加えることに

より、連続的な音声を合成する手法を提案した [1]。動的特徴量としては、$\Delta \boldsymbol{x}_t = \boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1}$ などの一次微分に相当する特徴量や、一次微分に加えて $\Delta\Delta \boldsymbol{x}_t = \boldsymbol{x}_{t+2} - 2\boldsymbol{x}_t + \boldsymbol{x}_{t-2}$ などの２次微分に相当する特徴量を利用する方法がある。動的特徴量を考慮する場合、フレームごとの変換ができないので、系列単位で変換を行うことになる。

フレームインデックスを $t = [1, \ldots, T]$ としたとき、動的特徴量を考慮した入力特徴量を $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta \boldsymbol{x}_t^\top]^\top$、入力特徴量系列を $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T]$、出力特徴量を $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta \boldsymbol{y}_t^\top]^\top$、出力特徴量系列を $\boldsymbol{Y} = [\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_T]$ とする。そして結合ベクトルを $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$、その時系列を $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_T]$ とおく。

まず、結合ベクトルの分布を GMM と仮定し、以下のようにおく。

$$P(\boldsymbol{Z}_t) = \sum_{k=1}^{K} \pi_k^{(\boldsymbol{Z})} \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\mu}_k^{(\boldsymbol{Z})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{Z})}) \tag{3.34}$$

ただし、K は混合数を表す。また $\pi_k^{(\boldsymbol{Z})}$、$\boldsymbol{\mu}_k^{(\boldsymbol{Z})}$、$\boldsymbol{\Sigma}_k^{(\boldsymbol{Z})}$ はそれぞれ k 番目の正規分布の重み、平均ベクトルと分散共分散行列である。これらのパラメータは、予めパラレルデータを用いて学習される。$\boldsymbol{\mu}_k^{(\boldsymbol{Z})}$ と $\boldsymbol{\Sigma}_k^{(\boldsymbol{Z})}$ は、以下のように $\boldsymbol{X}$ と $\boldsymbol{Y}$ に関連した要素に分解することができる。

$$\boldsymbol{\mu}_k^{(\boldsymbol{Z})} = \begin{bmatrix} \boldsymbol{\mu}_k^{(\boldsymbol{X})} \\ \boldsymbol{\mu}_k^{(\boldsymbol{Y})} \end{bmatrix} \tag{3.35}$$

$$\boldsymbol{\Sigma}_k^{(\boldsymbol{Z})} = \begin{bmatrix} \boldsymbol{\Sigma}_k^{(\boldsymbol{XX})} & \boldsymbol{\Sigma}_k^{(\boldsymbol{XY})} \\ \boldsymbol{\Sigma}_k^{(\boldsymbol{YX})} & \boldsymbol{\Sigma}_k^{(\boldsymbol{YY})} \end{bmatrix} \tag{3.36}$$

以上のように学習した結合ベクトルの GMM のパラメータ $\lambda^{(\boldsymbol{Z})} = \{\pi_k^{(\boldsymbol{Z})}, \boldsymbol{\mu}_k^{(\boldsymbol{Z})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{Z})}\}_{k=1 \ldots K}$ を用いると、$\boldsymbol{X}_t$ が与えられた時の $\boldsymbol{Y}_t$ の条件付き確率密度分布は、以下のようにして計算できる。

$$P(\boldsymbol{Y}_t | \boldsymbol{X}_t; \lambda^{(\boldsymbol{Z})}) = \sum_{k=1}^{K} \gamma_{k,t}^{(\boldsymbol{Z})} \mathcal{N}(\boldsymbol{Y}_t; \boldsymbol{E}_{k,t}^{(\boldsymbol{Z})}, \boldsymbol{D}_k^{(\boldsymbol{Z})}) \tag{3.37}$$

$$\gamma_{k,t}^{(\boldsymbol{Z})} = P(k | \boldsymbol{X}_t, \lambda^{(\boldsymbol{Z})}) \tag{3.38}$$

$$= \frac{\pi_k^{(\boldsymbol{Z})} \mathcal{N}(\boldsymbol{X}_t; \boldsymbol{\mu}_k^{(\boldsymbol{X})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{XX})})}{\sum_{k=1}^{K} \pi_k^{(\boldsymbol{Z})} \mathcal{N}(\boldsymbol{X}_t; \boldsymbol{\mu}_k^{(\boldsymbol{X})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{XX})})} \tag{3.39}$$

$$\boldsymbol{E}_{k,t}^{(\boldsymbol{Z})} = \boldsymbol{\mu}_k^{(\boldsymbol{Y})} + \boldsymbol{\Sigma}_k^{(\boldsymbol{YX})} \boldsymbol{\Sigma}_k^{(\boldsymbol{XX})^{-1}} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_k^{(\boldsymbol{X})} \right) \tag{3.40}$$

$$\boldsymbol{D}_k^{(\boldsymbol{Z})} = \boldsymbol{\Sigma}_k^{(\boldsymbol{YY})} - \boldsymbol{\Sigma}_k^{(\boldsymbol{YX})} \boldsymbol{\Sigma}_k^{(\boldsymbol{XX})^{-1}} \boldsymbol{\Sigma}_k^{(\boldsymbol{XY})} \tag{3.41}$$

声質変換は、上記の条件付き確率密度分布を、静的特徴量の時系列全体として最尤となるように変換を行う。ただし、出力話者の静的特徴量系列を $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T]$ とし静的特徴量系列から動的特徴量系列への変換行列を $\boldsymbol{W}$ とすると、

$$\boldsymbol{Y} = \boldsymbol{W} \boldsymbol{y} \tag{3.42}$$

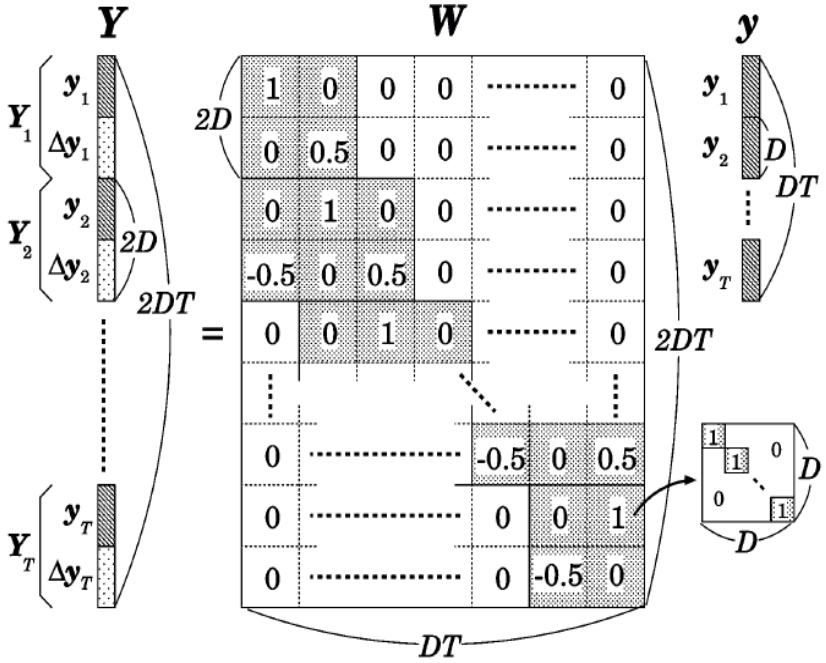

図 3.6: 静的特徴量から静的特徴量への変換行列（ [1] より引用）

の制約条件がある。今回用いた動的特徴量 $\Delta x_t = x_{t+1} - x_{t-1}$ では $\boldsymbol{W}$ は図 3.6 のようになる。この式 (3.42) の制約条件のもと、以下の対数尤度最大化基準で解く。

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} \sum_{t=1}^{T} \log P(\boldsymbol{Y}|\boldsymbol{X}; \lambda^{(\boldsymbol{Z})}) \tag{3.43}$$

これは EM アルゴリズムによって求めることができる。

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} \sum_{t=1}^{T} \log P(\boldsymbol{Y}|\boldsymbol{X}; \lambda^{(\boldsymbol{Z})}) \tag{3.44}$$

$$= \underset{\boldsymbol{y}}{\operatorname{argmax}} \sum_{t=1}^{T} \log \sum_{k=1}^{K} P(\boldsymbol{Y_t}, k|\boldsymbol{X_t}; \lambda^{(\boldsymbol{Z})}) P(k|\boldsymbol{X_t}; \lambda^{(\boldsymbol{Z})}) \tag{3.45}$$

この式から Q 関数を求めると以下の式になる。

$$Q(\hat{\boldsymbol{Y}}|\boldsymbol{Y}) = \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_k^{(Z')} \log P(\boldsymbol{Y_t}, k|\boldsymbol{X_t}; \lambda^{(\boldsymbol{Z})}) \tag{3.46}$$

ただし、$\gamma_{k,t}^{(\boldsymbol{Z'})}(\boldsymbol{Z}_t)$ は以下で表わされる。

$$\begin{aligned} \gamma_{k,t}^{(\boldsymbol{Z'})} &= P(k|\boldsymbol{X}_t, \boldsymbol{Y}_t, \lambda^{(\boldsymbol{Z})}) \\ &= \frac{\pi_k^{(\boldsymbol{Z})} \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_k^{(\boldsymbol{x})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{XX})})}{\sum_{k=1}^{K} \pi_k^{(\boldsymbol{Z})} \mathcal{N}(\boldsymbol{X}_t; \boldsymbol{\mu}_k^{(\boldsymbol{X})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{XX})})} \end{aligned} \tag{3.47}$$

これを $\boldsymbol{y}$ で微分するために展開していく。

$$
\begin{aligned}
Q(\hat{\boldsymbol{Y}}|\boldsymbol{Y}) &= \sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_k^{(\boldsymbol{Z'})}(\boldsymbol{Z}_t)\log P(\boldsymbol{Y}_t,k|\boldsymbol{X}_t;\lambda^{(\boldsymbol{Z})}) \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_k^{(\boldsymbol{Z'})}(\boldsymbol{Z}_t)\log \pi_k^{(\boldsymbol{Z})}\mathcal{N}(\boldsymbol{Y}_t;\boldsymbol{E}_{k,t}^{(\boldsymbol{Z})},\boldsymbol{D}_k^{(\boldsymbol{Z})}) \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_k^{(\boldsymbol{Z'})}(\boldsymbol{Z}_t)(-\frac{1}{2}(\boldsymbol{Y}_t{}^{\top}(\boldsymbol{D}_k^{(\boldsymbol{Z})})^{-1}\boldsymbol{Y}_t)+\boldsymbol{Y}_t{}^{\top}(\boldsymbol{D}_k^{(\boldsymbol{Z})})^{-1}\boldsymbol{E}_{k,t}^{(\boldsymbol{Z})})+\text{const} \\
&= \sum_{t=1}^{T}(-\frac{1}{2}(\boldsymbol{Y}_t{}^{\top}\overline{(\boldsymbol{D}_t^{(\boldsymbol{Z})})^{-1}}\boldsymbol{Y}_t)+\boldsymbol{Y}_t{}^{\top}\overline{(\boldsymbol{D}_t^{(\boldsymbol{Z})})^{-1}\boldsymbol{E}_t^{(\boldsymbol{Z})}})+\text{const} \\
&= -\frac{1}{2}(\boldsymbol{Y}_t{}^{\top}\overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}}\boldsymbol{Y}_t)+\boldsymbol{Y}_t{}^{\top}\overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}\boldsymbol{E}^{(\boldsymbol{Z})}}+\text{const} \qquad (3.48)
\end{aligned}
$$

ただし、const は $\boldsymbol{y}$ に依存しない項であり、$\gamma_{k,t}^{(\boldsymbol{Z})}$、$\overline{(\boldsymbol{D}_t^{(\boldsymbol{Z})})^{-1}}$、$\overline{(\boldsymbol{D}_t^{(\boldsymbol{Z})})^{-1}\boldsymbol{E}_t^{(\boldsymbol{Z})}}$、$\overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}}$、$\overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}E^{(\boldsymbol{Z})}}$ は以下の式で表わされる。

$$\overline{(\boldsymbol{D}_t^{(\boldsymbol{Z})})^{-1}} = \sum_{k=1}^{K}\gamma_k^{(\boldsymbol{Z'})}(\boldsymbol{Z}_t)(\boldsymbol{D}_k^{(\boldsymbol{Z})})^{-1} \qquad (3.49)$$

$$\overline{(\boldsymbol{D}_t^{(\boldsymbol{Z})})^{-1}\boldsymbol{E}_t^{(\boldsymbol{Z})}} = \sum_{k=1}^{K}\gamma_k^{(\boldsymbol{Z'})}(\boldsymbol{D}_k^{(\boldsymbol{Z})})^{-1}\boldsymbol{E}_{k,t}^{(\boldsymbol{Z})} \qquad (3.50)$$

$$\overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}} = \text{diag}[(\boldsymbol{D}_1^{(\boldsymbol{Z})})^{-1},(\boldsymbol{D}_2^{(\boldsymbol{Z})})^{-1},\ldots,(\boldsymbol{D}_K^{(\boldsymbol{Z})})^{-1}] \qquad (3.51)$$

$$\overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}\boldsymbol{E}^{(\boldsymbol{Z})}} = [\overline{(\boldsymbol{D}_1^{(\boldsymbol{Z})})^{-1}\boldsymbol{E}^{(\boldsymbol{Z})}}^{\top},\overline{(\boldsymbol{D}_2^{(\boldsymbol{Z})})^{-1}E^{(\boldsymbol{Z})}}^{\top},\ldots\overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}\boldsymbol{E}^{(\boldsymbol{Z})}}^{\top}]^{\top} \qquad (3.52)$$

ここで、制約条件である式 (3.42) を適用させることによって以下の式を得る。

$$Q(\hat{\boldsymbol{Y}}|\boldsymbol{Y}) = -\frac{1}{2}\boldsymbol{y}^{\top}\boldsymbol{W}^{\top}\overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}}\boldsymbol{W}\boldsymbol{y}+\boldsymbol{y}^{\top}\boldsymbol{W}^{\top}\overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}\boldsymbol{E}^{(\boldsymbol{Z})}}+\text{const} \qquad (3.53)$$

この式は $\boldsymbol{y}$ について上に凸な関数となるので Q 関数を $\boldsymbol{y}$ で微分すればよい。

$$\frac{\partial Q(\hat{\boldsymbol{Y}}|\boldsymbol{Y})}{\partial \boldsymbol{y_t}} = -\boldsymbol{W}^{\top}\overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}}\boldsymbol{W}\boldsymbol{y}+\boldsymbol{W}^{\top}\overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}\boldsymbol{E}^{(\boldsymbol{Z})}} \qquad (3.54)$$

以上より以下の更新式を得る。

$$\boldsymbol{y} = (\boldsymbol{W}^{\top}\overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}}\boldsymbol{W})^{-1}(\boldsymbol{W}^{\top}\overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}\boldsymbol{E}^{(\boldsymbol{Z})}}) \qquad (3.55)$$

動的特徴量を考慮することによって、出力特徴量が不連続になりにくくなり音質が向上した。

また、動的特徴量の制約は、入力特徴量ベースの手法にも適用可能である。入力特徴量ベースの手法は最小二乗誤差基準であったが、この手法に合わせて最尤推定で解くこととすると、式 (3.48) の $\gamma_k^{(Z')}\boldsymbol{X}_t) = P(k|\boldsymbol{x},\boldsymbol{y},\lambda)$ を入力特徴量のみに依存させ、$P(k|\boldsymbol{X},\lambda)$ に変更すればよい。これは動的特徴量を結合した入力特徴量の分布をモデル化した GMM を学習し、計算すれば実現できる。

第4章

提案手法

4.1　出力特徴量空間の領域分割に基づく手法

前章で述べた GMM を用いた区分的線形変換による手法では、領域分割を行う空間は結合ベクトルベースの手法では入力特徴量と出力特徴量の結合ベクトル空間であり、入力特徴量 ベースの手法では入力特徴量空間である。ここで、入出力特徴量空間全体の空間分割が大きく異なるような場合を考えると、変換時に入力特徴量における領域分割の情報を用いた際、出力特徴量空間での対応する分布における線形変換とのミスマッチが生じる可能性がある。この着眼点から本研究では、出力特徴量のみに基づく領域分割を検討した。

提案手法ではまず、出力話者の音声特徴量のみから GMM を学習する。そして、入力特徴量が与えられた際に、対応する出力特徴量がどの要素分布から生成されたかを入力特徴量から識別的に推定し、その結果を線形変換の重みとして用いることを考える。このように得られた事後確率を利用し、あとは入力特徴量に基づく方法と同様に θ を学習すれば、声質変換を実現することができる。

この考え方は既に、入力となるノイジー音声の特徴量から出力となるクリーン音声の特徴量を推定する雑音抑圧の分野で利用されており [7]、出力特徴量の GMM の要素正規分布に対する事後確率を、入力特徴量から識別して得られた事後確率を用いて区分的線形変換を行うことで、雑音抑圧の精度が向上することが示されている。本稿では、この考え方を、声質変換に導入したことになる。以下で具体的に提案手法を説明する。

提案手法ではまず、出力特徴量 $\boldsymbol{Y}_t$ の分布を GMM と仮定し、以下のようにおく。

$$P(\boldsymbol{Y}_t) = \sum_{n=1}^{N} \pi_n^{(y)} \mathcal{N}(\boldsymbol{Y}_t; \boldsymbol{\mu}_n^{(y)}, \boldsymbol{\Sigma}_n^{(y)}) \tag{4.1}$$

ただし、N は混合数を表す。$\pi_n^{(y)}$、$\boldsymbol{\mu}_n^{(y)}$、$\boldsymbol{\Sigma}_n^{(y)}$ は、n 番目の正規分布の重み、平均、分散で、予め学習データを用いて学習しておく。GMM のパラメータ集合は $\lambda^{(y)} = \{\pi_n^{(y)}, \boldsymbol{\mu}_n^{(y)}, \boldsymbol{\Sigma}_n^{(y)}\}_{n=1\ldots N}$ とおく。

次に、パラレルデータ $\{\boldsymbol{X}_t, \boldsymbol{Y}_t\}_{t=1\ldots T}$ の $\boldsymbol{Y}_t$ と上記の GMM を利用することにより、$\gamma_{n,t}^{(y)} = P(n|\boldsymbol{Y}_t)$ を得ることができる。これを用いて $\{\boldsymbol{X}_t, \{\gamma_{n,t}^{(y)}\}_{n=1\ldots N}\}_{t=1\ldots T}$ を得る。例えば $n_t^* = \mathrm{argmax}_n\, \gamma_{n,t}^{(y)}$ として、n_t^* を時刻 t におけるインデックスラベルとして $\{\boldsymbol{X}_t, n_t^*\}_{t=1\ldots T}$ を用意すれば、これは単純なラベル付きデータであるので、入力特徴量 $\boldsymbol{X}_t$ から n_t^* を推定する識別モデルを学習することができる。本稿では、具体的な識別モデルの実装として、[7] で利用されている、前後数フレームを連結した特徴量を LDA で次元圧縮した後、その空間で新しく GMM を学習して利用する方法を採用する。ただし LDA は以下で述べるように、連続値ラベルを用いる。前後数フレームを連結することで $\boldsymbol{X}_t$ 次元を拡張したものを $\boldsymbol{X}'_t$ とすると、まず、$\{\boldsymbol{X}'_t, \{\gamma_{n,t}^{(y)}\}_{n=1\ldots N}\}_{t=1\ldots T}$ を用いて、式 (19) の基準で $\{\gamma_{n,t}^{(y)}\}_{n=1\ldots N}$ に対応する特徴量のクラス内分散 $\boldsymbol{S}_B$ を小さく、クラス外分散 $\boldsymbol{S}_W$ を大きくするような $\boldsymbol{X}'_t$ に対する線形の次元圧縮行列 $\boldsymbol{L}$ を学習する。

$$\boldsymbol{L} = \underset{\boldsymbol{U}}{\mathrm{argmax}} \frac{\boldsymbol{U}\boldsymbol{S}_W\boldsymbol{U}^\top}{\boldsymbol{U}\boldsymbol{S}_B\boldsymbol{U}^\top} \tag{4.2}$$

ただし、

$$\boldsymbol{S}_B = \sum_{n=1}^{N}\sum_{t=1}^{T}\gamma_{n,t}^{(\boldsymbol{y})}(\boldsymbol{X}_t-\boldsymbol{\mu}_n^{(w)})(\boldsymbol{X}_t-\boldsymbol{\mu}_n^{(w)})^\top \tag{4.3}$$

$$\boldsymbol{S}_W = \sum_{n=1}^{N}\left(\sum_{t=1}^{T}\gamma_{n,t}^{(\boldsymbol{y})}\right)\left(\boldsymbol{\mu}_n^{(w)}-\frac{\sum_t \boldsymbol{X}_t}{T}\right)\left(\boldsymbol{\mu}_n^{(w)}-\frac{\sum_t \boldsymbol{X}_t}{T}\right)^\top$$

$$\boldsymbol{\mu}_n^{(w)} = \frac{1}{\sum_j \gamma_{n,t}^{(\boldsymbol{y})}}\sum_{t=1}^{T}\gamma_{n,t}^{(\boldsymbol{y})}\boldsymbol{X}_t \tag{4.4}$$

である。これは通常の LDA と同様に $\boldsymbol{S}_W^{-1}\boldsymbol{S}_B$ の固有ベクトルのを求めることで計算できる。次元圧縮後の特徴量を $\boldsymbol{V}_t = \boldsymbol{L}\boldsymbol{X}_t'$ とおく。次に、$\boldsymbol{V}_t$ の分布を GMM と仮定し、以下のように学習しておく。

$$P(\boldsymbol{V}_t) = \sum_{l=1}^{L}\pi_l^{(v)}\mathcal{N}(\boldsymbol{V}_t;\boldsymbol{\mu}_l^{(\boldsymbol{v})},\boldsymbol{\Sigma}_l^{(\boldsymbol{v})}) \tag{4.5}$$

GMM のパラメータは $\lambda^{(v)} = \{\pi_l^{(v)},\boldsymbol{\mu}_l^{(\boldsymbol{v})},\boldsymbol{\Sigma}_l^{(\boldsymbol{v})}\}_{l=1\ldots L}$ とおき、予め学習しておく。次に、SPLICE と同様の方法を用いて、入力特徴量 $\boldsymbol{X}_t$ が与えられた時の求めるべき出力特徴量 $\hat{\boldsymbol{Y}}_t$ の条件付き確率密度分布を、以下のように計算する。

$$P(\boldsymbol{Y}_t|\boldsymbol{X}_t;\lambda^{(v)},\theta^{(\boldsymbol{y})}) = \sum_{l=1}^{L}\gamma_{l,t}^{(v)}\mathcal{N}(\boldsymbol{Y}_t;\boldsymbol{\nu}_{l,t},\boldsymbol{\Gamma}_l) \tag{4.6}$$

$$\gamma_{l,t}^{(v)} = \frac{\pi_l^{(v)}\mathcal{N}(\boldsymbol{V}_t;\boldsymbol{\mu}_l^{(\boldsymbol{v})},\boldsymbol{\Sigma}_l^{(\boldsymbol{v})})}{\sum_{l=1}^{L}\pi_l^{(v)}\mathcal{N}(\boldsymbol{V}_t;\boldsymbol{\mu}_l^{(\boldsymbol{v})},\boldsymbol{\Sigma}_l^{(\boldsymbol{v})})} \tag{4.7}$$

$$\boldsymbol{\nu}_{l,t} = \boldsymbol{A}_l\boldsymbol{X}_t + \boldsymbol{b}_l \tag{4.8}$$

ただし、$\theta^{(\boldsymbol{v})}$ は $\boldsymbol{L}$ と $\lambda^{(v)}$ 以外の提案手法のパラメータで、$\theta^{(\boldsymbol{v})} = \{\boldsymbol{A}_l,\boldsymbol{b}_l,\boldsymbol{\Gamma}_l\}_{l=1\ldots L}$ である。$\theta^{(\boldsymbol{v})}$ の学習は、入力特徴量ベースの手法と同様に行うことができるが、ここでは [7] で導入されている、前後数フレームを連結した特徴量を入力特徴量とすることと、正則化を導入する。具体的には、LDA と同様に前後数フレームを連結することで $\boldsymbol{X}_t$ 次元を拡張したものを $\boldsymbol{X}_t'$ とすると、$\boldsymbol{A}_l$ と $\boldsymbol{b_l}$ の学習を、

$$\boldsymbol{A}_l,\boldsymbol{b}_l = \underset{\boldsymbol{A}_l,\boldsymbol{b}_l}{\operatorname{argmin}}\sum_{t=1}^{T}P(l|\boldsymbol{V}_t)||\boldsymbol{Y}_t-\boldsymbol{A}_l\boldsymbol{X}_t'-\boldsymbol{b}_l||^2+\frac{C}{2}||\boldsymbol{A_l}||^2 \tag{4.9}$$

として解く。ただし C は正則化の強さを調整するパラメータである。$\boldsymbol{\Gamma}_l$ は SPLICE と同様に学習する。$\boldsymbol{L}$、$\lambda^{(v)}$、$\theta^{(\boldsymbol{v})}$ が学習できれば、これまでの手法と同様に、式 (3.42) の条件のもと

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\,P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{L},\lambda^{(v)},\theta^{(\boldsymbol{v})}) \tag{4.10}$$

を解くことで、声質変換が実現できる。この式から Q 関数を求めると以下の式になる。

$$Q(\hat{\boldsymbol{Y}}|\boldsymbol{Y}) = \sum_{t=1}^{T}\sum_{l=1}^{L} \gamma_l^{(v)} \log P(\boldsymbol{Y_t}, k|\boldsymbol{V_t}; \lambda^{(\boldsymbol{Y}),\theta^{(\boldsymbol{Y})}}) \tag{4.11}$$

これを 3.42 の制約のもと Q 関数を最大化させる $\boldsymbol{y}$ を求める。これは、前章の動的特徴量を考慮した結合ベクトルベースの手法とほぼ同様にして計算でき、$\boldsymbol{y}$ の値は以下の式となる。

$$\boldsymbol{y} = (\boldsymbol{W}^\top \overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}} \boldsymbol{W})^{-1} (\boldsymbol{W}^\top \overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1} \boldsymbol{E}^{(\boldsymbol{Z})}}) \tag{4.12}$$

ただし、$\gamma_{k,t}^{(\boldsymbol{Z})}$、$\overline{(\boldsymbol{D}_t^{(\boldsymbol{Z})})^{-1}}$、$\overline{(\boldsymbol{D}_t^{(\boldsymbol{Z})})^{-1}\boldsymbol{E}_t^{(\boldsymbol{Z})}}$、$\overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}}$、$\overline{(\boldsymbol{D}^{(\boldsymbol{Z})})^{-1}E^{(\boldsymbol{Z})}}$ は以下の式で表わされる。

$$\overline{(\boldsymbol{D}_t^{(\boldsymbol{v})})^{-1}} = \sum_{l=1}^{L} \gamma_l^{(\boldsymbol{v})} (\boldsymbol{Z}_t)(\boldsymbol{D}_l^{(\boldsymbol{v})})^{-1} \tag{4.13}$$

$$\overline{(\boldsymbol{D}_t^{(\boldsymbol{v})})^{-1}\boldsymbol{E}_t^{(\boldsymbol{Y})}} = \sum_{l=1}^{L} \gamma_l^{(\boldsymbol{v})} (\boldsymbol{D}_l^{(\boldsymbol{v})})^{-1} \boldsymbol{E}_{l,t}^{(\boldsymbol{v})} \tag{4.14}$$

$$\overline{(\boldsymbol{D}^{(\boldsymbol{Y})})^{-1}} = \mathrm{diag}[(\boldsymbol{D}_1^{(\boldsymbol{v})})^{-1}, (\boldsymbol{D}_2^{(\boldsymbol{v})})^{-1}, \ldots, (\boldsymbol{D}_l^{(\boldsymbol{v})})^{-1}] \tag{4.15}$$

$$\overline{(\boldsymbol{D}^{(\boldsymbol{Y})})^{-1}\boldsymbol{E}^{(\boldsymbol{Y})}} = [\overline{(\boldsymbol{D}_1^{(\boldsymbol{v})})^{-1}\boldsymbol{E}_1^{(\boldsymbol{v})}}^\top, \overline{(\boldsymbol{D}_2^{(\boldsymbol{v})})^{-1}E_2^{(\boldsymbol{v})}}^\top, \ldots \overline{(\boldsymbol{D}_T^{(\boldsymbol{v})})^{-1}\boldsymbol{E}_T^{(\boldsymbol{v})}}^\top]^\top \tag{4.16}$$

結合ベクトルの手法では γ は $\boldsymbol{y}$ に依存し、同様に $\boldsymbol{y}$ は γ に依存するという関係があったため、交互にパラメータを更新していくことで Q 関数を解いていた。しかし、提案手法の場合は γ は $\boldsymbol{y}$ に依存しないので、パラメータが一意に定まり、逐次更新無しで変換が可能である。

第5章

実験

話者	msh, mht（男性）, fws（女性）
学習セット	ATR 音素バランス 503 文の A セット
テストセット	ATR 音素バランス 503 文の J セット
訓練セット	ATR 音素バランス 503 文の B,C セット
変換特徴量	メル一般化ケプストラムの 24 次元

表 5.1: 実験に使用した音声データ

提案手法と従来手法の比較するために、評価実験を行った。従来手法としては入力特徴量の GMM を用いた手法 [21] に動的特徴量を考慮したものと (以下入力特徴量ベースの手法とする)、動的特徴量を考慮している結合ベクトルの GMM を用いた手法（以下結合ベクトルベースの手法とする）[1] である。まず、予備実験として提案手法のハイパーパラメータを決定する実験を行った。ただし、これには学習、評価ともに後の評価実験で用いない訓練データセットを使用した。そして決定したパラメータを用いて客観評価実験と主観評価実験を行った。

5.1　予備実験

提案手法では、LDA と線形変換の入力特徴量として前後のフレームの特徴量と連結した拡大特徴量を用いる。その連結するフレーム数をそれぞれ F_{LDA},F_{LT} とすると、F_{LDA},F_{LT} と線形変換の正則化項のパラメータ C を事前に決定しておく必要がある。そのため、後の客観評価実験及び主観評価実験で用いる学習データとテストデータとは別に用意した訓練セットを用いてパラメータの決定を行った。本節ではその結果を一部記す。

5.1.1　実験条件

予備実験の実験条件は後に行う客観評価実験の実験条件と、使用する音声データセット以外はすべて同じとした。音声データセットの条件は表 5.1 の通りである。音声の分析再合成ツールとして STRAIGHT を用いた [23]。スペクトル特徴量としては、STRAIGHT 分析で得られたスペクトルから計算されたメル一般化ケプストラムの 1 次元から 24 次元までを使用した。GMM の分散共分散行列は特徴量の次元間の相関が無いものと仮定した。具体的には入力特徴量ベースの手法と提案手法では対角のみを用い、結合ベクトルの手法では分散共分散行列の要素である Σ_k^{XX}、Σ_k^{XY}、Σ_k^{YX}、Σ_k^{YY} を対角とした。学習音声は 50 文で行い、混合数は 64 にした。

5.1.2　客観評価基準

変換された音声と出力話者の音声との差を客観的に評価するために、変換特徴量と出力話者の自然音声から抽出した特徴量（以下目的特徴量とする）との差を評価する。ただし、変換ではフレーム数などは変えないため、話速の違いが残っている。そこで変換後の特徴

量と目的特徴量で、DP マッチングを行い話速の違いを吸収した後に評価を行う。客観評価尺度としては式 (5.1) で定義されるメルケプストラム歪み (Mel-Cepstram Distortion) で行った。

$$\mathrm{MCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}(mc_d^{(y)} - \hat{mc}_d^{(y)})^2} \tag{5.1}$$

ただし、$mc_d^{(\boldsymbol{y})}$、$hatmc_d^{(\boldsymbol{y})}$ はそれぞれ変換特徴量、目的特徴量のメルケプストラムの d 次元目であり、これは、変換特徴量が目的特徴量に近いほど値が小さくなり、特徴量が全く同じであればメルケプストラム歪みが 0 になる。予備実験ではこの客観評価尺度が最も小さくなるパラメータの決定を行う。

5.1.3　実験結果

正則化項のパラメータは 1、0.1、0、10 で動かした。正則化パラメータが 10 では値が大きすぎ、逆に 0 では拡大特徴量を用いるとすぐに過学習がおこり、歪みが大きくなった。比較的良い結果を示したのが、正則化項が 0.1 と 1 の場合で、正則化項の値が 0.1 である場合を図 5.1、正則化項の値が 1 の場合の結果を図 5.2 に示した。複数の折れ線は、F_{LDA} のパラメータを変更した場合である。グラフを見ると、F_{LDA} の値が変更されてもグラフの概形がほぼ同じである。したがって、F_{LDA} の値とその他のパラメータの間には相関が少ないことがわかる。また、正則化項大きくなると、F_{LT} の値が大きくなっても過学習が起こりにくくなっているのがわかる。そして、正則化項が 0.1 のときよりも 1 の時の方が極小値の歪みが小さくなっている。また LDA の入力を拡大特徴量にした場合と線形変換の入力に拡大特徴量にした場合を比較すると、LDA の拡大特徴量は 0.4 程度のケプストラム歪みの削減効果があったのにたいして、図にはのせてないが線形変換に正則化を行った場合は 1.0 程度のケプストラム歪み削減率があった。したがって、線形変換の方の効果が高いことがわかった。したがってこの実験条件の場合の最適な正則化パラメータは C は 1 、F_{LDA} は 7 、F_{LDA} は 9 となった。

5.2　客観評価実験

客観評価実験では同性への変換と異性 (男性から女性) への変換を行った。発声話者は、同性での変換は msh から mht への変換とし、異性への変換は msh から fws への変換とした。GMM の分散共分散行列は SPLICE ベースの手法と提案手法ではすべて対角とし、結合ベクトルベースの手法では、ブロック対角とした。提案手法の正則化パラメタおよび、LDA と変換の入力特徴量として前後何フレーム目までの特徴量を用いるかは予備実験で求めた値を適用した。

客観評価実験では、入力特徴量を変換したものと、自然音声から抽出した出力特徴量を比較した。学習音声の文数は 10 、20 、50 で変化させた。学習音声を一定にして混合数を変化させていき、歪みが最も小さかった値をその学習音声の文数の結果とした。客観評価

実験の結果を 図 5.3、図 5.4 に示す。同性への変換と異性への変換の両方で提案手法が最も歪みが少なくなり、提案手法の有効性が確認された。また、入力特徴量ベースの手法と結合ベクトルを用いた手法を比較すると、同性への変換は結合ベクトルの手法が良い結果となっているが、異性への変換では学習音声が 20 音声と 50 音声の時は入力特徴量ベースの手法の方が良い結果となった。これは結合ベクトルの手法では声質の近い人はうまく学習できるが、声質が大きく異なると 入出力の特徴量空間分割の違いが影響し、結合ベクトルを用いたGMMクラスタリングが適切に作用していないためと考えられる。これに対して入力特徴量ベースの手法や提案手法では、入力特徴量もしくは出力特徴量の一方のみを用いて GMM を学習しているので、双方の特徴量空間の分割の違いに影響されずに、GMMによるクラスタリングが実現していると考えられる。また、入力特徴量ベースの手法と提案手法はどちらも入力特徴量と出力特徴量の一方のみを用いているため、ほぼ同じ精度の GMM を学習できていると考えられるが提案手法の方が常に良い結果となった。

5.3　主観評価実験

主観評価実験では、変換した音声の話者性を評価する実験と、音声の自然性を評価する実験を行った。

5.3.1　実験条件

今回はスペクトル包絡特徴量だけを変換した。非周期性指標については全周波数において -60dB とした。F_{LDA} は 3 とし、F_{LT} は 9、C は 1 とした。号学習音声は 50 音声とし、混合数は結合ベクトルベースの手法は 256 とし、提案手法は 64 とした。話者性を評価する実験は RAB テストで行った。具体的な方法としては、まず正解音声である出力話者の音声の分析再合成音声を聞かせ、そのあとテスト音声である従来手法の音声と提案手法の音声を聞かせ、どちらが正解音声に近いか答えさせた。自然性を評価する実験では、テスト音声を聞かせ 5 段階で評価させた（ 5 が最も自然で、 1 が最も不自然とした)。テスト音声は ATR 音素バランス文の J セットからランダムに選んだ。評価音声数は話者性のテストは 20 文とした。自然性の評価では、出力話者の分析再合成音声、従来手法で変換した音声、提案手法で変換した音声を、それぞれ 10 文ずつ計 30 文を用いた。被験者は話者性の評価では 6 人、自然性の評価では 9 人で行った。F0 とパワーに関しては、次項で述べる方法を用いてターゲットの音声から抽出した特徴量を用いて合成した。

5.3.2　ターゲットの F0 とパワーの利用方法

話者情報は声色に相当するスペクトル包絡特徴量の変換をターゲットとしているが、F0 やパワーなどにも話者性は存在する。実用上は、これらの特徴量も分散を考慮した線形変換などで変換する必要があるが、変換誤差が生じてしまい、それが音質の劣化や話者性の再現の妨げとなってしまう。今回の主観評価実験では、声色に相当する特徴量の変換を評

価するのが目的であるため、F0 やパワーなどは出力話者の自然音声から抽出した特徴量を用いることにした。

変換した特徴量と出力特徴量は話速の違いからフレーム数に差異があるので、出力特徴量のフレーム数は固定で入力特徴量のフレーム数のみを伸縮させるような経路制限を付けた DP マッチングを行うことで、入力特徴量のフレーム数を変化させた。そして入力フレーム数を変化させた特徴量と、出力話者から抽出した F0 やパワーを用いることで、ターゲットの F0 やパワーを用いた変換音声を合成した。

5.3.3　主観評価実験結果

自然性の評価実験の結果を図 5.5、話者性の評価実験の結果を図 5.6 に示した。自然性評価では、 同性への変換、異性への変換のどちらも評価は高くなっているものの 95% 信頼区間での有意差は見られなかった。また、話者性の評価実験では、 95% 信頼区間での検定で、 男性間の変換では有意差があったが、男女間への変換では、有意差が見られなかった。異性への変換については、従来法、提案法ともに十分に話者情報の識別が可能なほどの変換が達成できなかった可能性があり、韻律的情報の変換も含めて、今後検討すべき課題と考えられる。

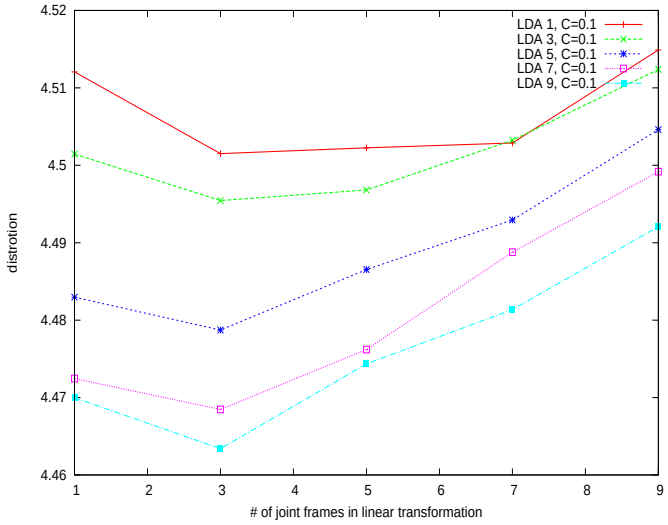

図 5.1: 正則化項のパラメータが 0.1 の場合の客観評価（複数の線は F_{LDA} の値を変化させたもの）

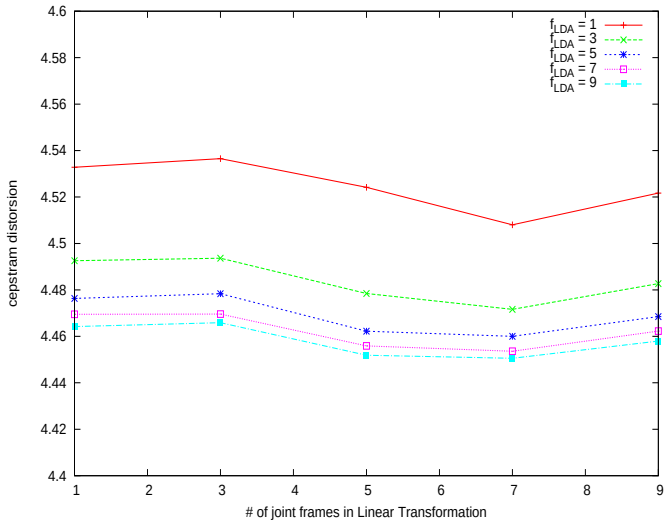

図 5.2: 正則化項のパラメータが 1 の場合の客観評価（複数の線は F_{LDA} の値を変化させたもの）

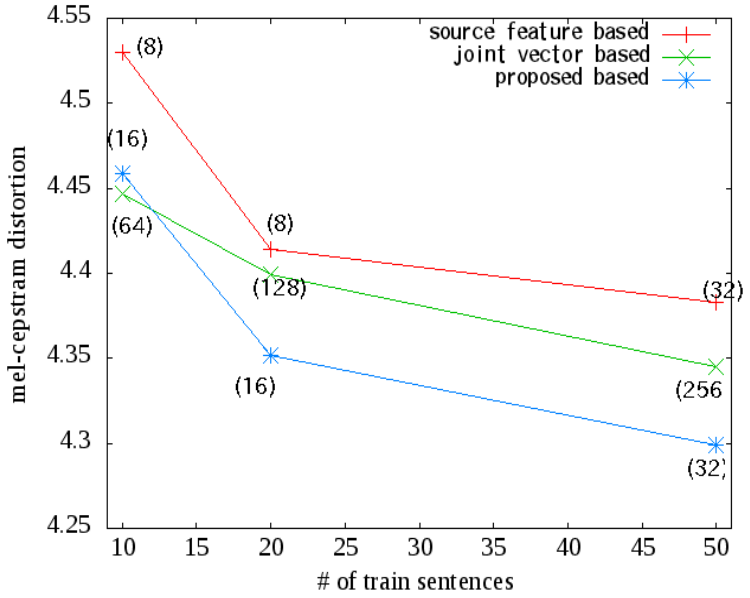

図 5.3: 客観評価実験結果（男性から男性）

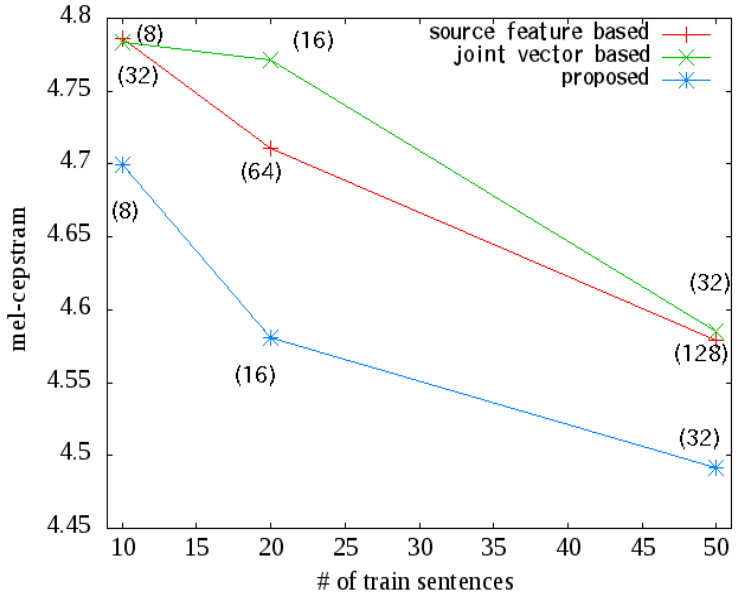

図 5.4: 客観評価実験結果（男性から女性）

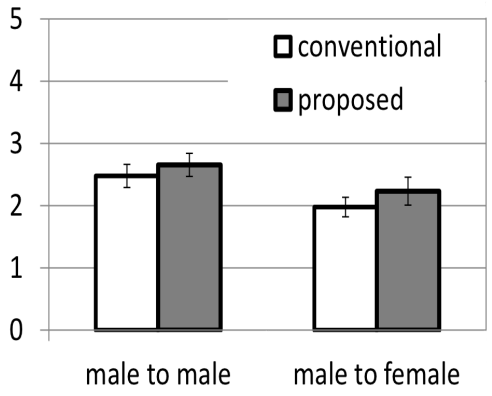

図 5.5: 自然性の主観評価実験結果

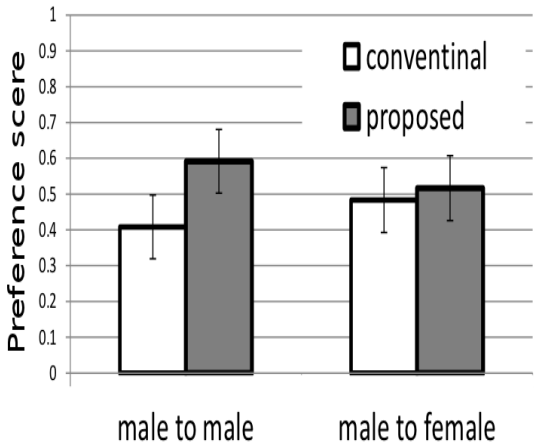

図 5.6: 話者性の主観評価実験結果

第6章

おわりに

本稿では、声質変換の精度向上を目的とし、出力特徴量の状態識別と長時間特徴量に基づく区分的線形変換に基づく手法を提案した。

第 2 章では音声工学一般で用いられている技術について説明した。音声工学では音声を直接扱うのではなく、特徴量を操作する。そこで、特徴量を抽出するためにソースフィルターモデルを仮定している。これは音声を音源成分と声道形状に起因する成分の 2 つの独立した成分に分けるものである。そして、それぞれの独立した成分に対応する特徴量として、基本周波数とスクトル包絡特徴量がある。音声認識や音声合成ではこれらの特徴量をモデル化する必要があるが、そのモデルとして用いられることが多い GMM について説明した。

第 3 章では、声質変換で利用されている技術について説明した。まず、変換モデルを学習するために入出力音声の話速の違いを吸収するためのアルゴリズムである DP マッチングについて説明したあと、代表的な変換モデルである区分的線形変換について述べた。区分的線形変換は領域分割と領域ごとの線形変換の 2 つの機能で構成されており、比較的計算コストが軽いというメリットがある。それら用いた具体的な声質変換を説明した。最初に考えられたベクトル量子化を用いた手法では、ベクトルを量子化しているため、出力特徴量が離散的になる問題があった。それを改善するために、GMM を用いる手法がほぼ同時期に 2 つ提案された。GMM を用いた手法では入力特徴量が連続的であれば出力特徴量が連続であることが保証されているが、雑音などが乗って入力特徴量が離散的になると出力特徴量も離散的になってしまう。そこで、前後のパラメータとの差分を考慮した GMM による区分的線形変換手法が提案された。これらの手法はすべて、領域分割を入力特徴量空間もしくは入出力特徴量を結合した特徴量の空間において行っており、これは入力特徴量と出力特徴量の分布の偏りが同一であることを仮定してる。しかし、入力特徴量と出力特徴量の分布が異なるときを考えると、ターゲットとなる出力特徴量の空間分割のほうがより重要であると考えられる。

第 4 章では出力特徴量の空間分割に基づく手法を提案した。変換時には出力特徴量がないので、空間分割に相当するものを識別によって求める。また、識別や変換を高精度にするため、長時間特徴量を利用した。

第 5 章では、評価実験を行った。主観評価実験では有意差はほとんど見られなかったものの客観評価実験では、従来手法より良い結果となり提案手法の有効性を確認した。

謝辞

3年間の研究生活にわたって、常日頃からご指導ご鞭撻を承りました指導教員の広瀬啓吉教授と峯松信明教授に大変感謝しております。また助教の齋藤大輔には研究に関係ない知識まで広く深く教えて頂き、誠に感謝しております。また、研究活動を支えて頂いた高橋登技官、秘書の池上恵氏、折茂結実子氏にも厚く御礼を申し上げます。また鈴木雅之氏には、研究を手とり足とり教えて頂き大変感謝しております。博士課程の柏木陽佑氏、橋本浩弥氏には時には深夜まで添削などをして頂き大変お世話になりました。また、同期や後輩に恵まれたおかげで楽しい研究生活になりました。感謝しています。広瀬・峯松研究室の方々は、皆優しい方達であり、素晴らしい研究室であると確信しています。最後に、自分を支えてくれた家族に感謝します。

参考文献

[1] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[2] M Narendranath, Hema A Murthy, S Rajendran, and B Yegnanarayana. Transformation of formants for voice conversion using artificial neural networks. *Speech communication*, Vol. 16, No. 2, pp. 207–216, 1995.

[3] Srinivas Desai, E Veera Raghavendra, B Yegnanarayana, Alan W Black, and Kishore Prahallad. Voice conversion using artificial neural networks. In *Proc. ICASSP*, pp. 3893–3896. IEEE, 2009.

[4] Alexander Kain and Michael W Macon. Spectral voice conversion for text-to-speech synthesis. In *Acoustics, Speech and Signal Processing*, Vol. 1, pp. 285–288. IEEE, 1998.

[5] Jasha Droppo, Li Deng, and Alex Acero. Evaluation of the splice algorithm on the aurora2 database. In *Proc. INTERSPEECH*, Vol. 1, pp. 217–220, 2001.

[6] グェンドゥックズイ, 鈴木雅之, 峯松信明, 広瀬啓吉. 識別的な区分的線形変換を用いた狭帯域音声の帯域拡張. 日本音響学会, 2013.

[7] Masayuki Suzuki, Takuya Yoshioka, Shinji Watanabe, Nobuaki Minematsu, and Keikichi Hirose. Mfcc enhancement using joint corrupted and noise feature space for highly non-stationary noise environments. In *Proc. ICASSP*, pp. 4109–4112. IEEE, 2012.

[8] Takashi Fukuda, Osamu Ichikawa, and Masafumi Nishimura. Short-and long-term dynamic features for robust speech recognition. In *Proc. INTERSPEECH*, pp. 2262–2265, 2008.

[9] M Ross, H Shaffer, Andrew Cohen, Richard Freudberg, and H Manley. Average magnitude difference function pitch extractor. *Acoustics, Speech and Signal Processing*, Vol. 22, No. 5, pp. 353–362, 1974.

[10] A Michael Noll. Short-time spectrum and “cepstrum” techniques for vocal-pitch detection. *the Jornal of the Acoustical Society of America*, Vol. 36, p. 296, 1964.

[11] Hiroya Fujisaki and Keikichi Hirose. Analysis of voice fundamental frequency contours for declarative sentences of japanese. *the Journal of the Acoustical Society of Japan*, Vol. 5, No. 4, pp. 233–242, 1984.

[12] Yi Xu. Transmitting tone and intonation simultaneously-the parallel encoding and target approximation (penta) model. In *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, 2004.

[13] Jinsong Zhang and Keikichi Hirose. Tone nucleus modeling for chinese lexical tone recognition. *Speech Communication*, Vol. 42, No. 3, pp. 447–466, 2004.

[14] Paul Taylor. Analysis and synthesis of intonation using the tilt model. *The Journal of the acoustical society of America*, Vol. 107, No. 3, pp. 1697–1714, 2000.

[15] Keiichi Tokuda, Takao Kobayashi, Takashi Masuko, and Satoshi Imai. Mel-generalized cepstral analysis-a unified approach to speech spectral estimation. In *ICSLP*, Vol. 94, pp. 18–22. Citeseer, 1994.

[16] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, pp. 1–38, 1977.

[17] Takashi Nose and Takao Kobayashi. Speaker-independent hmm-based voice conversion using adaptive quantization of the fundamental frequency. *Speech Communication*, Vol. 53, No. 7, pp. 973–985, 2011.

[18] Masanobu Abe, Satoshi Nakamura, Kiyohiro Shikano, and Hisao Kuwabara. Voice conversion through vector quantization. In *Proc. ICASSP*, pp. 655–658. IEEE, 1988.

[19] Hélène Valbret, Eric Moulines, and Jean-Pierre Tubach. Voice transformation using psola technique. *Speech Communication*, Vol. 11, No. 2, pp. 175–187, 1992.

[20] Hiroshi Matsumoto and Yasuki Yamashita. Unsupervised speaker adaptation from short utterances based on a minimized fuzzy objective function. *Journal of the Acoustical Society of Japan (E)*, Vol. 14, No. 5, pp. 353–361, 1993.

[21] Yannis Stylianou, Olivier Cappé, and Eric Moulines. Continuous probabilistic transform for voice conversion. *Speech and Audio Processing*, Vol. 6, No. 2, pp. 131–142, 1998.

[22] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model. *Speech Communication*, Vol. 50, No. 3, pp. 215–227, 2008.

[23] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné. 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, pp. 187–207, 1999.

発表文献

国際会議論文

[1] Keikichi Hirose, Hiroya Hashimoto, Jun Ikeshima and Nobuaki Minemats. Fundamental Frequency Contour Reshaping in HMM-based Speech Synthesis and Realization of Prosodic Focus Using Generation Process Model. Proc. Speech prosody, SS 1–3, 2012.

国内研究会論文

[2] 池島　純, 鈴木雅之, 齋藤大輔, 峯松信明, 広瀬啓吉. 出力特徴量の状態識別と長時間特徴量を用いた区分的線形変換による声質変換. 電子情報通信学会技術報告, SP–2013–56, pp. 19–24, 2013.

国内全国大会論文

[3] 池島純, 橋本浩弥, 広瀬啓吉, 峯松信明. 基本周波数パターン生成過程モデルを用いた音声合成の焦点制御の検討. 日本音響学会春季講演論文集, pp. 445–448, 2012.

[4] 池島純, 鈴木雅之, 峯松信明, 広瀬啓吉. ターゲット話者の特徴量状態識別に基づく区分的線形変換を用いた声質変換. 日本音響学会春季講演論文集, pp. 521–524, 2013.

[5] 岡安貴大, 池島純, 柏木陽佑, 鈴木雅之. 峯松信明, 広瀬啓吉, 実環境下における GMM を用いた統計的声質変換の検討. 日本音響学会春季講演論文集, pp. 525–528, 2013.

学位論文

[6] 池島純. 基本周波数パターン生成過程モデルを利用した HMM 音声の焦点制御, 東京大学工学部電子工学科卒業論文. 2012.