Vol.2016-SLP-110 No.6
2016/2/6

粒子フィルタとガウス過程回帰による シングルチャネル音源分離

博多屋 涼[1]　篠崎 隆宏[1,a)]　郡山 知樹[1]

概要：電話や音声認識などの音声アプリケーションを利用する際には雑音による性能の低下が問題となる．雑音低減手法の一つであるシングルチャネル音源分離はマルチチャンネル音源分離と比べて音声と雑音を一本のマイクロホンで分離できることから，実用化されれば高い利便性が期待される．しかし，シングルチャネル音源分離では音声の到来方向に関する情報が使用できないため，高精度な音声の分離が難しい．そこで我々は，音声に内在する制約を効果的にモデル化し音源分離に役立てる方法として，粒子フィルタとガウス過程回帰を用いた手法を提案する．連続値ベクトルによる音声のコンパクトな表現を隠れ状態とし，それをもとに雑音重畳音声の尤もらしさを評価することがアイデアである．すなわち，音声の時間方向の変化はマルコフチェインによりモデル化され，スペクトルの変動は粒子により表現される音声の状態を入力とするガウス過程回帰によりモデル化される．具体的には，状態特徴量としてケプストラム，F0，非周期性指標を用い，観測特徴量として対数パワースペクトルを用いて検討を行った．AURORA2 を用いた実験を行い，雑音重畳音声と比べ雑音除去後の音声のケプストラム歪みが小さくなることを確認した．

キーワード：シングルチャネル音源分離，粒子フィルタ，ガウス過程回帰

1. はじめに

情報技術の発達により，人々がスマートフォンやタブレットなどのモバイル端末を使用する機会が増加している．それに伴い端末に内蔵されている電話や音声認識などの音声アプリケーションも屋内，屋外を問わずあらゆるシーンで手軽に利用できるようになった．しかしこれらのアプリケーションは周囲の雑音により性能が低下してしまうという問題があるため，利用時には雑音を低減することが重要となる．

雑音を低減する手法の一つであるシングルチャネル音源分離は，一本のマイクロホンのみを用いて音声と雑音を分離する手法である．マイクの数が一本で済むため利便性が高いという特長があるが，マルチチャンネル音源分離と比べ音源の位置情報を利用することが出来ないため高性能の音源分離を達成することが難しい．シングルチャネル音源分離の代表的な手法としてはスペクトルサブトラクション法が知られており，これは周波数領域において雑音重畳音声のパワースペクトルから別途推定した雑音のパワースペクトルを減算することで音声と雑音を分離するものであるが，非定常な雑音に対して十分な効果が得られないという問題点がある [1]．非定常な雑音にも頑健な音源分離を行うためには音声の性質を分離のプロセスに活用することが重要であり，そのための手段として音声の統計的なモデルを用いる方法が考えられる．

我々はこれまでに統計モデルを用いた手法として，ボルツマンマシンと MCMC（マルコフ連鎖モンテカルロ）サンプリングによるシングルチャネル音源分離法を提案し，音声と雑音が分離可能であることを示した [2]．しかしこの手法はボルツマンマシンに時間方向の結合を持たせたにもかかわらず音声のコンテキスト情報のモデル化が不十分であり，期待されるよりも分離精度が低いという問題があった．

本研究ではより高精度なシングルチャネル音源分離を実現するため，粒子フィルタとガウス過程回帰を用いて音声に内在する制約を効果的にモデル化し音源分離を行う手法を提案する．これは雑音重畳音声から抽出した特徴量が観測として与えられたとき，隠れ状態である音声のコンパクトな表現を元に雑音重畳音声の尤度を評価することで音声の特徴量を推定するというアイデアに基づいている．音声の時間的・周波数的連続性はそれぞれマルコフチェインとガウス過程回帰を用いて表現される．推定した状態特徴量から音声を再合成することで雑音の分離された音声を得る．

本論文の構成は以下に示す通りである．第 2 章では音声

[1] 東京工業大学
Tokyo Institute of Technology
[a)] www.ts.ip.titech.ac.jp

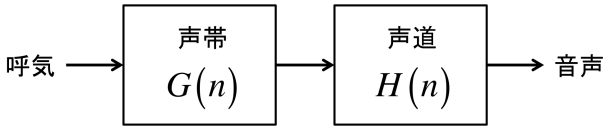

図 1 音声の生成過程

の生成モデルであるソース・フィルタモデルについて説明する．第 3 章では提案法で用いる統計モデルとして，粒子フィルタとガウス過程回帰についての基本原理を示す．第 4 章では提案法による音源分離の手順について解説する．第 5 章では提案法を用いた音源分離実験の結果を示し，その内容について考察を行う．第 6 章で本論文のまとめと今後の課題について述べ，結論とする．

2. 音声の生成過程

音声の生成過程は図 1 に示すように声帯の振動による音源（ソース）の生成と声道による共振（調音フィルタ）に分けて考えることが出来る．このような音響モデルをソース・フィルタモデルと呼び，音声合成や音声分析の分野で広く利用されている [3]．ソース・フィルタモデルでは，音声から抽出された基本周波数（F0），スペクトルに関する情報，非周期性指標などの音声パラメータを用いることで音声を再合成することが出来る．つまり音声の特徴を上記のパラメータの組み合わせとしてコンパクトに表現することが可能である．

本研究では音声の特徴量抽出および再合成に音声分析合成システムの WORLD[*1]を用いた [4]．WORLD は Vocoder[5] の考えに基づき，音声から F0，スペクトル包絡，非周期性指標の 3 つのパラメータを推定および推定したパラメータからの音声合成を行うことができる．

3. 提案法で用いる統計モデル

本章では提案法で用いる統計モデルとして，粒子フィルタおよびガウス過程回帰についての説明を行う．

3.1 粒子フィルタ

粒子フィルタは時系列データを処理する逐次的なベイズ推定法の一種であり，宇宙船のターゲット・トラッキングや動画中のビジュアル・トラッキングなどに応用されている [6]．同じ時系列データを扱うカルマンフィルタと比べ，非線形・非ガウス型の状態空間モデルにおける状態の推定に適用が可能であるという特徴がある．粒子フィルタでは状態の確率分布を重み付けされた多数の粒子で近似し，粒子をモデルに従って推移させることで分布の更新を行う [7]．図 2 は粒子フィルタの概念図である．粒子フィルタによる状態空間モデルは，時刻 t における状態ベクトルを $\mathbf{x}_t$，観測ベクトルを $\mathbf{y}_t$ とすると以下の式で定義される [8]．

[*1] http://ml.cs.yamanashi.ac.jp/world/index.html

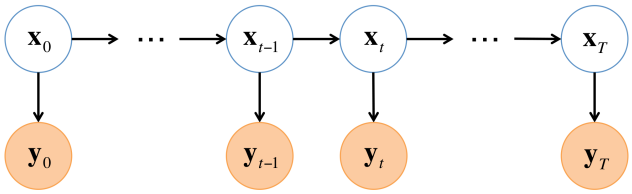

図 2 粒子フィルタの概念図

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}, \boldsymbol{\epsilon}_t) \tag{1}$$

$$\mathbf{y}_t = g(\mathbf{x}_t, \boldsymbol{\delta}_t) \tag{2}$$

ここで f は状態遷移関数，g は観測関数であり，$\boldsymbol{\epsilon}_t$ と $\boldsymbol{\delta}_t$ はそれぞれシステムノイズと観測ノイズを表す．粒子フィルタはこの状態空間モデルにより与えられた観測に対する状態を逐次的に決定するため，状態間および状態と観測間の関係を適切に表現する関数を用いる必要がある．

3.2 ガウス過程回帰

ガウス過程回帰は D 次元の入力ベクトル $\mathbf{x}$ とそれに対応する出力値 y をセットとする N 組の学習データ $\Theta = (\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N) = (\mathbf{X}, \mathbf{Y})$ および未知の入力 $\mathbf{x}_{N+1}$ が与えられたとき，出力の予測分布 $p(y_{N+1}|\mathbf{x}_{N+1}, \Theta)$ を求めることで y_{N+1} を推定するための手法である [9]．ガウス過程回帰は入力ベクトル $\mathbf{x}$ を高次の特徴空間に写像することで非線形な回帰を行う．$\mathbf{x}_m$ と $\mathbf{x}_n$ の高次特徴空間内における内積をカーネル関数 $k(\mathbf{x}_m, \mathbf{x}_n)$ を用いて表すと，出力 y_{N+1} の予測分布は以下の式で与えられる．

$$p(y_{N+1}|\mathbf{x}_{N+1}, \Theta) = \mathcal{N}(y_{N+1}|\mu_{N+1}, \sigma^2_{N+1}) \tag{3}$$

$$\mu_{N+1} = \mathbf{k}^{\mathrm{T}}_{N+1}(\mathbf{K} + \delta^2\mathbf{I})^{-1}\mathbf{y} \tag{4}$$

$$\sigma^2_{N+1} = k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) - \mathbf{k}^{\mathrm{T}}_{N+1}(\mathbf{K} + \delta^2\mathbf{I})^{-1}\mathbf{k}_{N+1} \tag{5}$$

ここで，$\mathbf{K}$ は $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ を要素とする行列，$\mathbf{k}_{N+1}$ は事前に与えられている入力ベクトル $\mathbf{x}_i$ と $\mathbf{x}_{N+1}$ とのカーネル関数の値を第 i 行に持つベクトル，δ^2 は出力に加わる観測ノイズの分散値を表す．また，$^{\mathrm{T}}$ は行列の転置を表し，$\mathbf{I}$ は単位行列である．

4. 提案法による音源分離

提案法では音声の時間変化をモデル化するために，図 2 の粒子フィルタにおける状態遷移関数として条件付きガウス分布による遷移確率 $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ を用い，観測関数として状態から観測の非線形な変換をモデル化するためにガウス過程回帰による観測確率 $p(\mathbf{y}_t|\mathbf{x}_t)$ を用いる．状態特徴量には F0，ケプストラム，非周期性指標を，観測特徴量には対数パワースペクトルを用い，雑音重畳音声の観測特徴量が与えられた際にその観測を尤も良く説明するような音声の状態特徴量を推定する．推定した状態特徴量から音声を再合成することで雑音が分離された音声を得る．

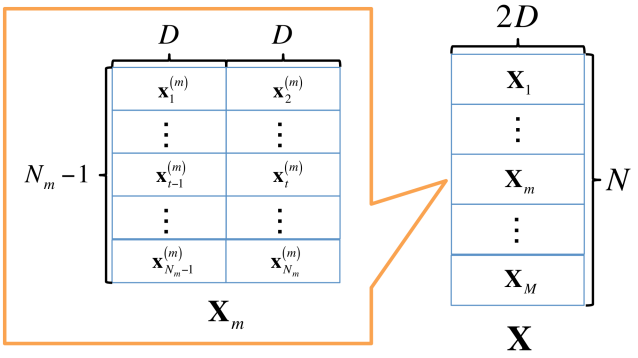

図 3 学習データの結合

4.1 遷移確率の導出

前状態から現状態への遷移確率 $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ を求めるため，まず隣接する時間フレームの同時確率 $p(\mathbf{x}_t, \mathbf{x}_{t-1})$ を導出する [10]．学習データから抽出した D 次元の特徴量を以下の手順で結合する．学習データとして M 文章の音声を用い，学習データの m $(1 \leq m \leq M)$ 番目の文章が N_m 個の時間フレームに分割されているとする．文章 m の t $(1 \leq t \leq N_m)$ フレームにおける特徴ベクトルを $\mathbf{x}_t^{(m)}$ で表すと，発話単位で $\mathbf{x}_t^{(m)}$ を隣接する時間フレームで結合し $N_{m-1} \times 2D$ の行列 $\mathbf{X}_m$ を得る．

$$\mathbf{X}_m = \begin{pmatrix} \mathbf{x}_1^{(m)} & \dots & \mathbf{x}_{N_m-1}^{(m)} \\ \mathbf{x}_2^{(m)} & \dots & \mathbf{x}_{N_m}^{(m)} \end{pmatrix}^{\mathrm{T}} \tag{6}$$

さらにこの行列を発話間で結合し，図 3 に示すような $2D$ 次元の特徴ベクトル $\mathbf{x}_n^{'}$ を成分とする $N \times 2D$ の行列 $\mathbf{X}$ を得る．ただし，$N = \sum_{m=1}^{M}(N_m - 1)$ である．

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1^T & \dots & \mathbf{X}_M^{\mathrm{T}} \end{pmatrix}^{\mathrm{T}} = \begin{pmatrix} \mathbf{x}_1^{'} & \dots & \mathbf{x}_N^{'} \end{pmatrix}^{\mathrm{T}} \tag{7}$$

求めた $\mathbf{X}$ に対し，平均ベクトル $\boldsymbol{\mu}_{\mathbf{X}}$ および共分散行列 $\boldsymbol{\Sigma}_{\mathbf{X}}$ を以下の式で計算する．

$$\boldsymbol{\mu}_{\mathbf{X}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n^{'} \tag{8}$$

$$\boldsymbol{\Sigma}_{\mathbf{X}} = \frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n^{'} - \boldsymbol{\mu}_{\mathbf{X}}\right)\left(\mathbf{x}_n^{'} - \boldsymbol{\mu}_{\mathbf{X}}\right)^{\mathrm{T}} \tag{9}$$

ここで，d $(1 \leq d \leq 2D)$ は特徴量の次元を表す．上記の $\boldsymbol{\mu}_{\mathbf{X}}$，$\boldsymbol{\Sigma}_{\mathbf{X}}$ が同時確率 $p(\mathbf{x}_t, \mathbf{x}_{t-1})$ の平均および共分散となる．

次に，求めた平均ベクトル $\boldsymbol{\mu}_{\mathbf{X}}$ と共分散行列 $\boldsymbol{\Sigma}_{\mathbf{X}}$ を以下のように分割する．

$$\boldsymbol{\mu}_{\mathbf{X}} = \begin{pmatrix} \boldsymbol{\mu}_a & \boldsymbol{\mu}_b \end{pmatrix} \tag{10}$$

$$\boldsymbol{\Sigma}_{\mathbf{X}} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \tag{11}$$

平均ベクトルは $\boldsymbol{\mu}_a$，$\boldsymbol{\mu}_b$ の次元数がともに D，共分散行列は分割した 4 つの行列のサイズが全て $D \times D$ となるように分割する．これらの分割した平均ベクトルと共分散行列を用いると，条件付き確率 $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ の平均 $\boldsymbol{\mu}_{t|t-1}$ および共分散 $\boldsymbol{\Sigma}_{t|t-1}$ は以下の式で求められる [11]．

$$\boldsymbol{\mu}_{t|t-1} = \boldsymbol{\mu}_b + \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}(\mathbf{x}_{t-1} - \boldsymbol{\mu}_a) \tag{12}$$

$$\boldsymbol{\Sigma}_{t|t-1} = \boldsymbol{\Sigma}_{bb} - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab} \tag{13}$$

式（12）および（13）を用いると，遷移確率 $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ は以下の式で定義される．

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}) \tag{14}$$

$$= \frac{1}{\sqrt{(2\pi)^D \left|\boldsymbol{\Sigma}_{t|t-1}\right|}}$$

$$\exp\left(-\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_{t|t-1})^{\mathrm{T}}\boldsymbol{\Sigma}_{t|t-1}^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_{t|t-1})\right)$$

4.2 観測確率の導出

提案法では粒子フィルタの状態が D_x 次元の状態ベクトル $\mathbf{x}$，観測が D_y 次元の観測ベクトル $\mathbf{y}$ であるため，$\mathbf{y}$ の各次元 d に対し学習データ $\Theta^d = (\mathbf{X}, \mathbf{Y}^d)$ を用いて観測確率 $p(y_{N+1}^d|\mathbf{x}_{N+1}, \Theta^d)$ を独立に求めることを考える．ここで，ガウス過程回帰を用いる際に音声の調波構造をより明示的にモデルへと反映させるため，状態ベクトル $\mathbf{x}$ における F0 の値から励起信号スペクトルを計算して用いる．今，m 番目の学習データにおける基本周波数の値を $f_0^{(m)}$ とすると，励起信号 $\mathbf{exc}_m$ は以下の式で表される [12]．

$$exc_m[n] = \sum_{u=1}^{U} \sin\left(2\pi f_0^{(m)} un + \phi[m,u]\right) + w[n] \tag{15}$$

ただし，$n = [1, ..., 512]$ であり，U は 4kHz 以下の倍音数，$\phi[m,u]$ は u 番目の倍音の位相を表す．また，$w[n]$ は白色雑音であり，$f_0^{(m)} = 0$ となる無声区間において付加する．励起信号スペクトル $\mathbf{EXC}_m$ は式（15）により得られた励起信号をフーリエ変換し（記号 $DFT\{\}$ で表す）絶対値を取ることで求められる．

$$EXC_m[d] = |DFT\{exc_m[n]\}| \tag{16}$$

上記の励起信号スペクトルの d 次元目とケプストラム **CEP**，非周期性指標 **AP** を用いると，ガウス過程回帰の入力ベクトルは $\mathbf{x}^d = (EXC[d], \mathbf{CEP}, \mathbf{AP})$ と表される．つまり $\mathbf{X}^d = (\mathbf{x}_1^d, ..., \mathbf{x}_N^d)$ とすると学習データ $\hat{\Theta}^d = (\mathbf{X}^d, \mathbf{Y}^d)$ を用いて，観測確率 $p(y_{N+1}^d|x_{N+1}^d, \hat{\Theta}^d)$ が計算される．モデルを学習する際に学習データの数が多くなると学習にかかる時間が膨大となるため，学習データの中からランダムに n 個のサンプルを選び代表点とする FITC 近似を用いた [13]．

4.3 粒子フィルタによる状態推定

まず，時刻 $t = 0$ における D 次元の粒子 $\mathbf{x}_0^{(n)}$ $(1 \leq n \leq$

N) の値を決定する．次に，$t = 1, ..., T$ で以下の処理を繰り返す [14].

4.3.1 予測

各粒子 $\mathbf{x}_{t-1}^{(n)}$ に対し，条件付きガウス分布より次の時刻における粒子の候補 $\hat{\mathbf{x}}_t^{(n)}$ を予測する．

$$p(\hat{\mathbf{x}}_t^{(n)}|\mathbf{x}_{t-1}^{(n)}) = \mathcal{N}(\hat{\mathbf{x}}_t^{(n)}|\boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}) \tag{17}$$

4.3.2 重みの更新

ガウス過程回帰を用いて $\hat{\mathbf{x}}_t^{(n)}$ に対する予測分布の平均 $\hat{\mu}_t^{(n)}$ と分散 $\hat{\sigma}_t^{(n)}$ を推定し，与えられた観測 $\mathbf{y}_t$ に基づく $\hat{\mathbf{x}}_t^{(n)}$ の尤度 $p(\mathbf{y}_t|\hat{\mathbf{x}}_t^{(n)})$ を計算する．

$$p(\mathbf{y}_t|\hat{\mathbf{x}}_t^{(n)}) = \mathcal{N}(\mathbf{y}_t|\hat{\mu}_t^{(n)}, \hat{\sigma}_t^{(n)}) \tag{18}$$

式（17）および（18）を用いて，各粒子の重み $w_t^{(n)}$ を以下の式で求める．

$$w_t^{(n)} = \frac{p(\mathbf{y}_t|\hat{\mathbf{x}}_t^{(n)})p(\hat{\mathbf{x}}_t^{(n)}|\mathbf{x}_{t-1}^{(n)})^{\gamma}}{\sum_{n=1}^{N} p(\mathbf{y}_t|\hat{\mathbf{x}}_t^{(n)})p(\hat{\mathbf{x}}_t^{(n)}|\mathbf{x}_{t-1}^{(n)})^{\gamma}} \tag{19}$$

ここで，γ は遷移確率の重みを表すパラメータである．

4.3.3 状態の推定

現在の状態 $\mathbf{x}_t$ を全ての粒子の重み付き平均により求める．

$$\mathbf{x}_t = \mathbf{w}_t \mathbf{x}_t^{(n)^{\mathrm{T}}} \tag{20}$$

4.3.4 リサンプリング

現時刻における全ての粒子 $\hat{\mathbf{x}}_t$ から重み w_t に比例した確率で N 個の粒子を選び直し，新たな粒子 $\mathbf{x}_t^{(n)}$ を得る．

4.4 音声の再合成

粒子フィルタの動作により得られた各時刻における状態ベクトルを用いて音声を再合成する．本研究では WORLD でスペクトル包絡を生成する過程における音声スペクトルの特徴量をケプストラムとして用いた．また，非周期性指標は抽出した値の低次成分を用い，音声を再合成する際に高次の成分については推定した低次成分の値を用い算出した．

5. 実験

モデルの学習および評価には AURORA-2 データベース[*2] に収録されているサンプリング周波数 8kHz の音声および雑音を用いた．学習データとして（A）男性話者一人による発話 100 文章（B）男性話者一人による発話 450 文章（C）男性話者一人・女性話者一人による発話各 450 文章の計 900 文章の 3 種類を用いた．評価データは学習データと同一の男性話者による学習データに含まれない 20 文章とした．雑音は街中の雑音である “Babble”，および車の運転音である “Car” の 2 種類を用い，音声と雑音の SNR が 0 になるように雑音を重複させた．粒子フィルタの状態特徴量としては 1 次元の F0，13 次元のケプストラム，3 次元の非周期性指標を用いた．ケプストラムの 1 次元目は音声のパワーである．観測特徴量としては 257 次元の対数パワースペクトルを用いた．特徴量は平均 0，分散 1 となるように正規化を行った．音声の分析は窓幅が 32ms のハニング窓を 10ms のフレーム周期でシフトして行った．フーリエ変換の次数は 512 とした．粒子フィルタにおける粒子の個数は 1000 とした．粒子の初期値は各次元について独立に平均 0，分散 1 のガウス分布を用いてランダムに値を定めた．

ガウス過程回帰にはガウス過程のパッケージである pyGPs[*3] を用いて行った．ガウス過程回帰におけるカーネル関数として以下のものを用いた．

$$k(\mathbf{x}, \mathbf{x}^{'}) = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}^{'}||^2}{2l^2}\right) \tag{21}$$

ここで l はカーネル関数のハイパーパラメータであり，学習データを用いて最適な値を定めた．FITC 近似における近似点の個数は 100 とした．状態の推定時に励起信号を作成する際の位相は観測として与えられる雑音重畳音声の位相を用いた．

5.1 評価尺度

雑音重畳音声と雑音除去後の音声を比較するため，以下に示す 2 種類の評価尺度を用いた．

5.1.1 ケプストラム歪み（CD）

ケプストラム歪み（Cepstral Distortion）は原音声と目標音声間におけるケプストラム値の差を表し，以下の式を用いて計算する [15].

$$CD = \frac{\alpha}{T'} \sum_{t \in notSIL} \sqrt{\sum_{d=1}^{D} \left(cep_{org}^{(d)} - cep_{target}^{(d)}\right)^2} \tag{22}$$

$$\alpha = \frac{10\sqrt{2}}{\ln 10} = 6.14185 \tag{23}$$

ここで D はケプストラムの次元数であり，$cep_{org}^{(d)}$ と $cep_{target}^{(d)}$ はそれぞれ原音声および目標音声のケプストラムにおける d 次元目を表す．ただし $0 \leq d \leq D$ とし，$d = 0$ はパワー項である．また，$T^{'}$ は原音声における無音区間を除いた総フレーム数を表す．本実験では原音声における無音区間を除いた時間フレームに対し，パワー項を除いたケプストラムの値について CD の値を求めた．

5.1.2 F0 歪み（F_0D）

F0 歪み（F0 Distortion）は原音声と目標音声間における F0 値の差を表し，以下の式で定義する．

[*2] http://aurora.hsnr.de/aurora-2.html

[*3] http://www-ai.cs.uni-dortmund.de/weblab/static/api_docs/pyGPs/index.html

表 1 ケプストラム歪み（CD）の評価

Noise	Smooth	Noisy	Train A	Train B	Train C
Babble	OFF	8.81	8.50	8.09	8.27
	ON	8.81	7.63	7.49	7.73
Car	OFF	9.14	8.82	8.42	8.18
	ON	9.14	7.59	7.47	7.77

表 2 F0 歪み（F_0D）の評価

Noise	Smooth	Train A	Train B	Train C
Babble	OFF	103	101	86.5
	ON	75.0	71.1	75.8
Car	OFF	88.2	88.1	82.1
	ON	77.3	72.0	73.8

$$F_0D = \sqrt{\frac{1}{T'}\sum_{t\in voiced}\left(F0^{(t)}_{org} - F0^{(t)}_{target}\right)^2} \qquad (24)$$

T' は原音声における有声音の総フレーム数を表す．本実験では原音声における有声音の時間フレームに対し F_0D の値を求めた．

5.2 F0 の平滑化

音声のモデル化に際して，無声音と有声音のフレームで F0 の値に大きな差異が生じるため，モデル化が正しく行われない可能性がある．そこで，無声音のフレーム n における F0 の値 $\mathrm{F0}_{(n)}$ を一つ前のフレームにおける F0 の値 $\mathrm{F0}_{(n-1)}$ に置き換える処理を行う．これを F0 の平滑化と定義する．

$$if \quad \mathrm{F0}_{(n)} = 0 \quad then \quad \mathrm{F0}_{(n)} = \mathrm{F0}_{(n-1)} \quad \forall n \qquad (25)$$

5.3 実験結果

表 1 は “Babble” と “Car” の 2 種類の雑音に対し，(A) ～ (C) の 3 種類の学習データにより学習したモデルを用いて雑音を分離した音声におけるケプストラム歪み（CD）の値を示したものである．実験は粒子フィルタの遷移確率の重み γ を 1.0 として行った．雑音を分離する前の CD 値は “Babble” で 8.81，“Car” で 9.14 であった．

学習データに対し F0 の平滑化（Smooth）を行った場合（ON）と行わなかった場合（OFF）のそれぞれについて CD を求めた．まず平滑化を行わなかった場合について雑音分離前と分離後の CD 値を比較すると，全ての学習条件において雑音分離後における CD の値が小さくなっていることから，提案法により原音声に近いケプストラムが得られたことがわかる．最も値が改善したのは（C）の学習データを用いて “Car” の雑音を分離した場合で，その値は 0.96 小さくなっていた．次に F0 の平滑化を行った場合の CD について見ると，雑音分離前と比べ分離後の値が最も小さくなったのは（B）の学習データを用いて “Car” の雑音を分離した場合で，CD の値は 1.67 小さくなっていた．F0 の平滑化を行わなかった場合よりも CD の値が減少したことから，平滑化が音声のモデル化に対し有効であると考えられる．雑音の種類における CD の値に注目すると，どちらの雑音においても分離後の CD の値が減少していた．また平滑化を行わなかった場合は雑音分離前の CD と比較した際，学習データ（A），（B）でそれぞれ 0.3，0.7 程度の値の改善が見られたが，平滑化を行った後の CD は “Babble” と “Car” でほぼ同じ値となったことから，“Car” の雑音に対してより値が改善していることがわかる．

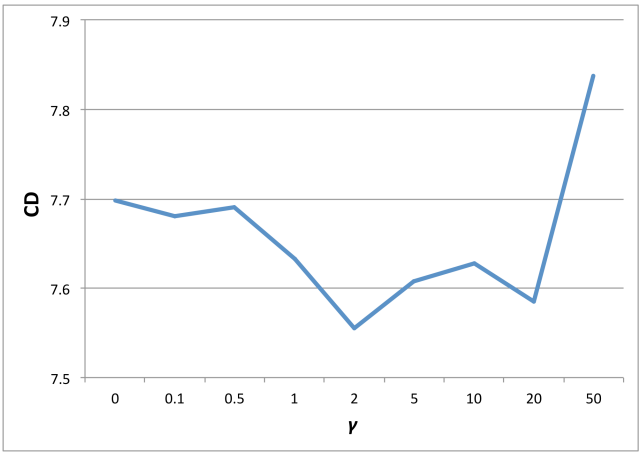

図 4 遷移確率の重み γ と CD の変化

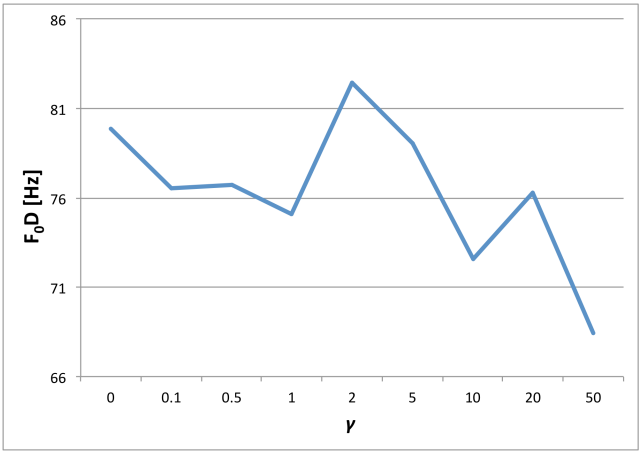

図 5 遷移確率の重み γ と F_0D の変化

表 2 は上記のケプストラム歪みと同様の条件で求めた F0 歪み（F_0D）の値を示している．F_0D の値が極めて大きくなっていることから，F0 の推定が不十分であると考えられる．平滑化を行わなかった場合に比べ，平滑化を行った場合の F_0D はどの学習条件および雑音条件においても減少していたことから，平滑化処理はケプストラムだけでなく F0 の推定にも有効であると考えられる．

図 4 および図 5 は式（19）における遷移確率の重み γ と CD，F_0D の値との関係を示している．モデルは学習データ（A）を平滑化を行い学習したものを用いた．また，評価データに重畳させる雑音は “Babble” とした．図 4 を見ると $\gamma = 2$ で CD の値が最小となった．これはケプストラ

ム項に関してのモデル化が時間方向の推移についてもスペクトルの変動についても良く行われていることを表していると考えられる．一方，図 5 では $\gamma = 50$ のときに F_0D が最も良い値を示した．F_0D の値は γ の値を大きくしたときに減少傾向にあるため，ガウス過程回帰によるモデルの調整が不十分な可能性がある．

6. まとめ

粒子フィルタとガウス過程回帰を用いたシングルチャネル音源分離法について提案した．粒子フィルタの状態として音声の情報を F0，ケプストラム，非周期性指標の 3 つのパラメータでコンパクトに表現し，時間的および周波数的連続性を考慮しながら音声に内在する制約をモデル化することで音源分離を試みた．実験の結果から，提案法による音源分離によりケプストラム歪みの値が元の雑音重畳音声よりも小さくなることを確認した．また，F0 の平滑化を行うことで雑音の分離精度が向上することを示した．しかし F0 の値については雑音重畳音声よりも歪みが大きくなってしまい，推定精度が十分であるとは言えなかった，ガウス過程回帰による音声モデルの構築において F0 の代わりに励起信号スペクトルを用いたが，推定した F0 が原音声の F0 に対する倍音に当たるなどの問題点もあり，モデルの最適化について再検討する必要がある．今後の課題として F0 およびケプストラムの推定精度向上のためモデルの構造や学習に用いる特徴量の再考，パラメータの調整などが挙げられる．

謝辞 本研究は JSPS 科研費 26280055 の助成をうけたものです．

参考文献

[1] 大槻典行, 宮永喜一, “音声雑音除去に関する研究: ランニングスペクトルフィルタ（RSF）の効果,” 釧路工業高等専門学校要 37, pp.23-27, 2003 年 12 月.
[2] 博多屋涼, 篠崎隆宏, 小林隆夫, “ボルツマンマシンと MCMC サンプリングを用いた音声のシングルチャネル雑音除去,” 日本音響学会 2014 秋季研究発表会講演論文集, 1-R-1, pp.59-60.
[3] 徳田恵一, 大浦圭一郎, “自動学習により人間のように歌う音声合成システム-Sinsy-,” 音声言語情報処理（SLP）研究報告, 第 90 巻, 第 1 号, pp.1-6, 2012 年 1 月.
[4] 森勢将雅, 西浦敬信, 河原英紀, “高品質音声分析変換合成システム WORLD の提案と基礎的評価 ～基本周波数・スペクトル包絡制御が品質の知覚に与える影響～,” 日本音響学会聴覚研究会, vol. 41, no. 7, pp.555-560, Toyama, Oct. 1-2, 2011.
[5] H. Dudley, “Remaking speech,” J. Acoust. Soc. Am., vol. 11, pp.169-177, 1939.
[6] 生駒哲一 (2008).「逐次モンテカルロ法とパーティクルフィルタ」北川源四郎, 竹村彰通編（編）『「21 世紀の統計科学」第 III 巻』東京大学出版会.
[7] 北川源四郎, “モンテカルロ・フィルタおよび平滑化について,” 統計数理, 第 44 巻, 第 1 号, pp.31-48, 1996.
[8] 矢野浩一, “粒子フィルタの基礎と応用: フィルタ・平滑化・パラメータ推定,” 日本統計学会誌 第 44 巻, 第 1 号, pp.189-216, 2014 年 9 月.
[9] 奥村麻由, 槇原靖, 八木康史, “大規模歩容データベースを用いたガウス過程回帰による年齢推定の評価,” 電子情報通信学会技術研究報告 第 110 巻, 第 382 号, pp.183-190, 2011 年 1 月.
[10] T. Toda, A.W. Black, K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory.” IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, No. 8, pp.2222-2235, Nov. 2007.
[11] Christopher M. Bishop (2006). Pattern Recognition and Machine Learning. Springer.（C.M. ビショップ 元田浩・栗田多喜夫・樋口知之・松本裕治・村田昇監訳（訳）(2010) . 『パターン認識と機械学習 上』丸善出版
[12] M. Wohlmayr, M. Stark, and F. Pernkopf, “A probabilistic interaction model for multipitch tracking with factorial hidden Markov models,” IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 4, pp.799-810, May 2011.
[13] Naish-Guzman, A. and Holden, S, “The Generalized FITC Approximation,” in Advances in Neural Information Processing Systems 21, pp.534-542, Cambridge, MA, USA. The MIT Press.
[14] 島倉諭, 田崎勇一, 稲垣伸吉, 鈴木達也, “GMM を用いた複数予測モデル型パーティクルフィルタによる意図推定,” 日本機械学会 ロボティクス・メカトロニクス講演会講演概要集, 2A1-E08, 2010.
[15] J. Kominek, T. Schultz, and A. W. Black, “Synthesizer voice quality on new languages calibrated with mean Mel-Cepstral distortion,” in Proc. Inte. Workshop Spoken Lang. Technol. for Under-Resourced Lang. (SLTU), Hanoi, Vietnam, 2008.