非同期録音ブラインド同期のための線形位相補償の効率的最尤解探索*

◎宮部滋樹 (筑波大), 小野順貴 (NII), 牧野昭二 (筑波大)

1 はじめに

非同期マイクロホンアレー [1, 2] は，複数の独立した録音機器を用いた非同期の多チャネル録音でアレー信号処理を行う枠組で，多チャネルの音声強調や音響モニタリングなどを，携帯電話やボイスレコーダといった汎用的な録音装置を組み合わせた簡易かつ柔軟な機器構成で行える利点がある．しかし，各チャネルの録音が同期していないために録音開始時刻やサンプリング周波数の不一致が起こり [1, 3]，アレー信号処理の性能が大幅に劣化するという問題がある．

我々は前回の発表 [4] において，ブラインド音源分離 (BSS) [5] の前処理のために，サンプリング周波数のチャネル間ミスマッチを短時間フーリエ変換 (STFT) 領域で最尤法により推定し補償する方法を提案した．しかし最尤推定が解析的に解くことができないものとなるため，計算量の大きい離散値全探索を行う必要があった．本稿では，最適値付近では局所的に短報的になるという経験的に得られた性質を利用し，荒い離散値全探索による探索範囲の絞り込みと，黄金分割探索による詳細な探索を併用した，効率的な最尤解探索法について議論する．

2 時間周波数領域におけるサンプリング周波数ミスマッチの最尤推定

本節では [4] で提案したサンプリング周波数ミスマッチ補償のための最尤推定について述べる．ここでは 2 素子マイクロホンアレーについて議論するが，3 素子以上への拡張は，基準チャネルを定めて基準以外のチャネルごとに補償することにより容易に行える．

2.1 サンプリング周波数ミスマッチの時間領域表現

同時刻における 2 つのマイクロホンの連続信号 $\mathrm{x}_1(t), \mathrm{x}_2(t)$（$t$ は連続時間）が別々の A/D 変換器でサンプリングされて離散信号 $x_1(n), x_2(n)$（n はサンプル番号）が得られたとする．ここで $x_1(n)$ のサンプリング周波数は f_s，$x_2(n)$ のサンプリング周波数は未知の無次元数 ϵ（$|\epsilon| \ll 1$）により表される $(1+\epsilon) f_\mathrm{s}$ であるとする．このとき各チャネルの離散信号と連続信号の関係は以下のように表される．

$$x_1(n) = \mathrm{x}_1\left(\frac{n}{f_\mathrm{s}}\right) \tag{1}$$

$$x_2(n) = \mathrm{x}_2\left(\frac{n}{(1+\epsilon) f_\mathrm{s}} + T_{21}\right) \tag{2}$$

ここで連続時刻 t の原点は $x_1(n)$ の録音開始時刻とし，T_{21} は $x_2(n)$ の録音開始時刻とする．従って同じ時刻 t を参照する第 i チャネル（$i = 1, 2$）の離散時刻 n_i の関係は

$$n_2 = (1+\epsilon) n_1 - f_\mathrm{s} T_{21} \tag{3}$$

となり，長い時間が経過して n_1, n_2 が大きくなるのに比例して参照する時刻の差が拡大する．

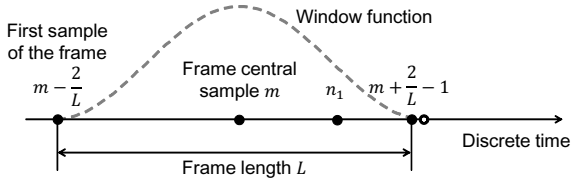

Fig. 1 Relation between the frame central sample m of a frame and the sample n_1 inside the frame.

2.2 時間周波数領域におけるサンプリング周波数ミスマッチのモデル化

アレー信号処理の多くは STFT 領域で行われるため，サンプリング周波数ミスマッチ補償は STFT 領域での信号表現がよく近似する方法であれば十分であると考える．そのためにまず STFT のフレーム長を L，チャネル 1 のあるフレームの中心サンプルを m として，$m - L/2 \le n_1 \le m + L/2 - 1$ という 1 フレーム内での時間の対応関係を考える（Fig. 1）．

式 (3) の関係から，サンプル m の近傍の n_1 には

$$\begin{aligned} n_2 &= (1+\epsilon)(n_1 - m) + (1+\epsilon) m - f_\mathrm{s} T_{21} \\ \leftrightharpoons (n_2 - m) &= (1+\epsilon)(n_1 - m) + \epsilon m - f_\mathrm{s} T_{21} \end{aligned} \tag{4}$$

という関係が成り立ち，$(n_2 - m)$ と $(n_1 - m)$ の対応関係は，m とともに ϵm だけ拡大していくことがわかる．$|\epsilon(n_1 - m)|$ は第 2.3 節の処理によって 0 に近い値となるため，ここでは $\epsilon(n_1 - m)$ を無視すると，

$$\begin{aligned} (n_2 - m) &\approx (n_1 - m) + \epsilon m - f_\mathrm{s} T_{21} \\ &= (n_1 - m) + \tau_{21}(m; \epsilon) \end{aligned} \tag{5}$$

$$\tau_{21}(m; \epsilon) = \epsilon(m - M) \tag{6}$$

$$M = \frac{f_\mathrm{s} T_{21}}{\epsilon} \tag{7}$$

のように，フレーム内で時間差が n_1 に依存せず一定と仮定したモデルを得る．ここで $\tau_{21}(m; \epsilon)$ はフレームの時刻オフセットであり，M は両チャネルが同じ連続時刻を指して $n_1 = M$ のとき $n_2 = M$ となる離散時刻で，以下ではオフセット原点と呼ぶこととする．

もし全てのフレームの中心サンプル m に対して

$$|\tau_{21}(m; \epsilon)| = |\epsilon(m - M)| \ll L \tag{8}$$

---

*Efficient Maximum Likelihood Solution Search of Linear Phase Compensation for Blind Synchronization of Asynchronized Recording. by MIYABE, Shigeki (University of Tsukuba), ONO, Nobutaka (NII), MAKINO, Shoji (University of Tsukuba)

が満たされるとき，オフセット $\tau_{21}(m;\epsilon)$ の補償はフレーム内の時間シフトとみなすことができ，これはSTFT 領域では線形位相により近似的に補償することができる．観測信号 $x_1(n), x_2(n)$ を一律でフレーム分析に施して得られる第 m サンプルを中心とするフレーム波形の短時間フーリエ変換 $X_i(k,m), i = 1,2, -L/2 < k \leq L/2$ を以下のように表される．

$$X_i(k,m) = \sum_{l=0}^{L-1} w(l)\, x_i\left(l + m - \frac{L}{2}\right) \exp\left(-\frac{2\pi \jmath k l}{L}\right) \tag{9}$$

ここで $w(l)$ は長さ L の再合成可能な窓関数，$k = -L/2+1, \ldots, L/2$ は離散周波数インデックス，$\jmath = \sqrt{-1}$ を表す．第 2 チャネルの信号 $X(k,m)$ にオフセット $\tau_{21}(m;\epsilon)$ を補償する線形位相を与えることにより，補償した信号 $\hat{X}_2(k,m)$ は以下のように得られる．

$$\begin{aligned}\hat{X}_2(k,m) &= X_2(k,m) \exp\left(\frac{2\pi \jmath k \tau_{21}(m;\epsilon)}{L}\right) \\ &= X_2(k,m) \exp\left(\frac{2\pi \jmath k \epsilon (m-M)}{L}\right)\end{aligned} \tag{10}$$

2.3 観測信号全体の相互相関を用いた録音開始時刻ずれの大まかな補償

式 (2) に現れる録音開始時刻の差 T_{21} は，観測信号 $x_1(n)$ と $x_2(n)$ のみを用いて推定することは難しい．しかしながら，BSS のような音源方位情報を明に利用しないクラスのアレー信号処理では，小さなチャネル間の固定時間差は許容され，正確な録音開始時刻差 T_{21} の推定は不要となる．そのためここでは，第 2.2 節で述べた線形位相補償モデルが成り立つための条件である，$|\epsilon(n_1 - m)|$ が全てのフレーム中心サンプル m おいて無視できるようにするための，大まかな録音時刻差 T_{21} の補償を行う．

サンプリング周波数のミスマッチが十分小さく $|\epsilon| \ll 1$ ととなる場合は，サンプリング周波数のミスマッチの影響を受けていても $x_1(n)$ と $x_2(n)$ の相関は高いと仮定できる．そこで以下のように相関を最大にする $x_2(n)$ の遅延量 δ_{12} を求める．

$$\delta_{12} = \underset{-N_2<\delta<N_1}{\arg\max} \sum_{n=\max(0,\delta)}^{\min(N_1,N_2+\delta)-1} x_1(n)\, x_2(n-\delta) \tag{11}$$

ここで $N_i, i = 1,2$ は $x_i(n)$ のサンプル数を表す．そして $x_2(n)$ を δ_{12} だけ遅延させて

$$x_2(n) \leftarrow x_2(n - \delta_{12}) \tag{12}$$

とすることにより，$x_1(n_1)$ と $x_2(n_2)$ のサンプリング周波数ミスマッチ原点を信号のオーバーラップするサンプル区間の中央付近に移動する．また式 (10) 中のオフセット原点 M には，以下のようにオーバーラップの中央付近のサンプル番号を推定として与える．

$$M \leftarrow \left\lfloor \frac{\min(N_1 - \delta_{12}, N_2) - \max(0, \delta_{21}) - 1}{2} \right\rfloor \tag{13}$$

ここで $\lfloor \cdot \rfloor$ は床関数を表す．この全サンプルのシフトとオフセット原点の推定により，式 (8), (10) の $\epsilon(m-M)$ は全ての m において可能な限り 0 に近い値となる．

2.4 サンプリング周波数ミスマッチ推定を評価する尤度の定式化

観測されるすべての音源は定常かつ位置の移動が無いと仮定すると，正確な ϵ の推定を用いてサンプリング周波数のミスマッチを補償した観測信号

$$\hat{\mathbf{X}}(k,m;\epsilon) = \left[X_1(k,m), \hat{X}_2(k,m;\epsilon)\right]^{\mathsf{T}} \tag{14}$$

は離散周波数 k 毎に定常となる．正確な ϵ の推定による補償で定常性を回復した観測信号 $\hat{\mathbf{X}}(k,m;\epsilon)$ が共分散行列 $\mathbf{V}(k)$ の零平均多変量複素正規分布に従うと仮定すると，その対数尤度は

$$\begin{aligned}J(\mathbf{V},\epsilon) = \sum_{k,m} \Big(&- \hat{\mathbf{X}}(k,m;\epsilon)^{\mathsf{H}} \mathbf{V}(k)^{-1} \hat{\mathbf{X}}(k,m;\epsilon) \\ &- \log 2\pi^2 - \log\det \mathbf{V}(k) \Big)\end{aligned} \tag{15}$$

と表せる．ここで $\{\cdot\}^{\mathsf{H}}$ は複素共役転置を表し，また共分散行列 $\mathbf{V}(k)$ は尤度関数を決定するパラメタであり，以下のように $\hat{\mathbf{X}}(k,m;\epsilon)$ を用いた標本推定により推定する．

$$\mathbf{V}(k) \leftarrow \frac{1}{|\forall m|} \sum_{\forall m} \hat{\mathbf{X}}(k,m;\epsilon)\, \hat{\mathbf{X}}(k,m;\epsilon)^{\mathsf{H}} \tag{16}$$

これを式 (15) に代入すると第 1 項が定数となるため、対数尤度関数は定数項を除いて以下のように単純化される。

$$J(\epsilon) = -\sum_{k} \log\det \sum_{m} \hat{\mathbf{X}}(k,m;\epsilon)\, \hat{\mathbf{X}}(k,m;\epsilon)^{\mathsf{H}} \tag{17}$$

サンプリング周波数ミスマッチ ϵ の補償が不正確であれば，時刻のドリフトによる定常性の低下により対数尤度 $J(\epsilon)$ が小さくなるため，$J(\epsilon)$ の最大化により ϵ を推定することができる．しかしこの尤度最大化は解析的に解くことができないため，次節では効率的な ϵ の最尤解の探索方法について議論する．

3 黄金分割法による最尤解の効率的探索

前節まででは [4] で提案したサンプリング周波数のミスマッチ ϵ を評価するための対数尤度関数 $J(\epsilon)$ を議論したが，この対数尤度関数を最大化する ϵ は解析的に求めることができない．推定するパラメタは ϵ のみであり，その最適化には離散値全探索を行うことも考えられるが，一つの ϵ の評価のために全帯域での共分散行列とその逆行列の計算が必要であるため，高い解像度の離散値全探索を行うためには計算量が膨大になる．本節ではこの最尤推定問題の効率的な解探索法について議論する．

この最尤推定問題で求めるべきパラメタは ϵ のみであるため，一次元の最適化問題の代表的な解法で

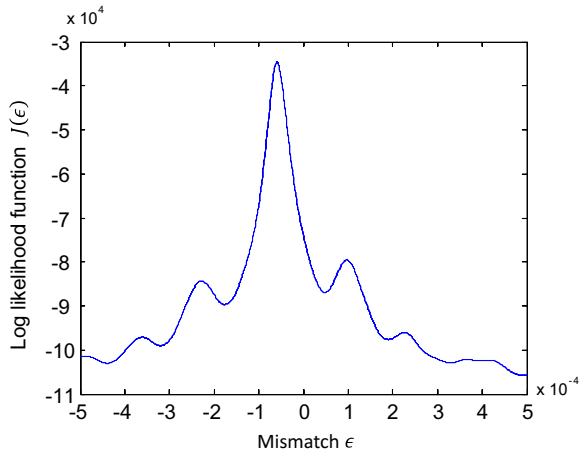

Fig. 2 An example of the log likelihood function $J(\epsilon)$. We can see that the function is locally convex in the neighbor of its maximum.

ある黄金分割探索 [6] を用いることが考えられる．黄金分割探索は関数の最大値または最小値を繰り返し探索により探索範囲を狭めながら求める手法で，関数が局所的に単峰的な範囲では最適解に一意に収束する．図 2 に示す例のように，対数尤度関数 $J(\epsilon)$ は最大値周辺では単峰性関数となることが経験的にわかっているため，適切に探索範囲を絞り込むことで黄金分割探索の利用が可能となる．

黄金分割探索の前にまず刻みの荒い離散値全探索により探索範囲を絞り込む．サンプリング周波数ミスマッチ ϵ を範囲 $[-E, E]$ で等間隔に D 分割した

$$\epsilon_d = -E + \frac{2dE}{D-1}, \quad d = 0, \ldots, D-1 \tag{18}$$

について，すべての $J(\epsilon_d)$ を比較して最大値を与える分割番号 d^* を求める．

$$d^* = \underset{d=0,\ldots,D-1}{\arg\max} \; J(\epsilon_d) \tag{19}$$

この離散値全探索の範囲 $[-E, E]$ は，機器間のサンプリング周波数ミスマッチとして妥当な範囲に設定すればよい．一般的な機器のサンプリング周波数ミスマッチは 10^{-5} オーダーであるといわれているため，E は 10^{-4} またはその数倍が妥当といえる．また，探索範囲の分割数 D は，音声を観測する場合なら 10 から 100 程度で適切に単峰な範囲を絞り込むことができる．

次に，$[\epsilon_{d^*-1}, \epsilon_{d^*+1}]$ を探索範囲とした ϵ についての $J(\epsilon)$ の黄金分割探索を以下のように行う．

Step 1: 探索範囲 $[a, b]$ の初期値を $a = \epsilon_{d^*-1}$，$b = \epsilon_{d^*+1}$ と定める．

Step 2: $p = b - \varphi(b-a)$, $q = a + \varphi(b-a)$ として $J(p)$，$J(q)$ を求める．ここで φ は以下で与えられる．

$$\varphi = \frac{\sqrt{5}-1}{2} \tag{20}$$

Step 3: $J(p) \leq J(q)$ なら

$$a = p, \quad p = q, \quad q = a + \varphi(b-a) \tag{21}$$

とし，そうでなければ以下のように更新する．

$$b = q, \quad q = p, \quad p = b - \varphi(b-a) \tag{22}$$

Step 4: $(b-a)$ が十分に小さくなければ step 2 に戻り，十分に小さければ以下のようにとして ϵ の最尤推定値を求め，終了する．

$$\epsilon = \frac{a+b}{2} \tag{23}$$

黄金分割探索は探索範囲を繰り返し毎に $\varphi \approx 0.62$ 倍に狭めることができるため，離散値全探索で P 回の関数評価が必要な P 点等分割の精度の探索は，$\frac{\log \frac{1}{P}}{\log \varphi} \approx 2.08 \log P$ 回の関数評価で達成することができる．

4 実験

提案手法の性能と非同期アレー信号処理への有効性を検証するため，複数話者の音声の混合のマイクロホンアレーによる観測信号に人工的にサンプリング周波数のミスマッチを与え，提案手法にるサンプリング周波数ミスマッチ補償の計算効率，補償精度とブラインド音源分離の性能への寄与を評価する．

4.1 実験条件

使用した観測信号は，2 人の話者による発話に実測したインパルス応答を畳み込んで混合したマイクロホンアレー観測信号で，片方のチャネルのサンプリング周波数を 100 タップのポリフェーズフィルタを用いて人為的に変更してサンプリング周波数のミスマッチを模擬した．変更前のサンプリング周波数は 16,000 Hz で，サンプリング周波数の変更は $16,000 \pm 0.5$，$16,000 \pm 1$，16,000±1.5 Hz の 6 種類とした．これらはそれぞれ ±31.25，±62.5，±93.75 ppm に相当し，別々の A/D 変換器を用いた場合のサンプリング周波数のミスマッチとして現実的な大きさのものである．音源分離評価のための分離手法には補助関数法独立ベクトル分析 [7] を用いた．その他の実験条件を Table 1 に示す．

4.2 計算量の評価

Table 2 に，10 秒の観測信号を提案手法の $E = 5 \times 10^{-4}, D = 10$ とした分割の荒い離散値全探索の範囲絞り込みの後に，絞り込まれた範囲を P 点の解候補に等分割して比較する離散値全探索と，この P 点解候補探索と同等の精度の黄金分割探索の計算量を比較する．表より，提案手法は高い解像度でも計算量をあまり増やすことなく探索を行えていることがわかる．

以下の実験では，$E = 5 \times 10^{-4}, D = 10$ の離散値全探索による探索範囲絞り込みと，Step 4 における収束条件を $(b-a) < 10^{-9}$ とした，$P = 10^5$ 等分割の解候補探索に相当する黄金分割探索用いて評価を行う．

Table 1　Experimental conditions

Speech signal	Concatenated word uttrances by 4 speakers
Signal length	3, 5, 10, 20, 30 seconds
Rverberation time	T_{60} of 130 ms
Frame length L	4,096 samples
Frame shift M	2,048 samples
Source distance	1.5 m
Source directions	[-50°, 30°], [-60°, -10°]
Microphone spacing	2 cm
CPU	Opteron 2.8 GHz

Table 2　Comparison of computation time in second

P	10^2	10^3	10^4	10^5
Discrete search	10.17	85.52	339.70	3337
Golden section	2.52	2.96	3.31	3.74

4.3　サンプリング周波数ミスマッチの推定精度

複数の長さのデータに対するサンプリング周波数の補正精度を平均二乗誤差 (RMSE) で評価した結果を図 3 に示す．最も短い 3 秒の観測信号でも RMSE は元のサンプリング周波数ミスマッチ ϵ の 10 分の 1 以下に収束し，データが増えるにつれて急速に小さくなる．従って観測信号の定常性を仮定した尤度がサンプリング周波数のミスマッチの評価尺度として有効であるということがわかる．

4.4　音源分離への寄与

サンプリング周波数のミスマッチを提案手法により補償することで，音源分離の性能が回復することを確かめるための分離精度評価を行った．分離フィルタの学習には与えられた観測信号全体を用いた．評価尺度には，非目的成分の抑圧の強さを表す信号対干渉比 (signal-to-interference ratio; SDR) を用いた [8]．また，SDR を算出するための参照信号としては，サンプリング周波数の変更を施していないマイクロホンにおける各音源の音像を用いた．

実験結果を図 4 に示す．まず Unprocessed が非常に低い値を示していることから，この条件ではサンプリング周波数ミスマッチの補償をしなければ音源分離ができない厳しい条件であるということがわかる．手動でパラメタ ϵ の正しい値を与えた位相補償はサンプリング周波数ミスマッチがない場合よりも SDR が 2 dB 程度低いだけであり，BSS のためのサンプリング周波数ミスマッチ補償に STFT 領域における位相補償が有効であるということがわかる．また，サンプリング周波数ミスマッチをブラインドに推定して補償する提案手法は，手動でパラメタ ϵ を与えた場合とほとんど性能差がなく，提案手法の最尤推定は性能限界に近い高い精度でサンプリング周波数ミスマッチ ϵ を推定できているということを示している．以上より，音源分離のためのサンプリング周波数ミスマッチ補償として提案手法が有効であるということが確認された．

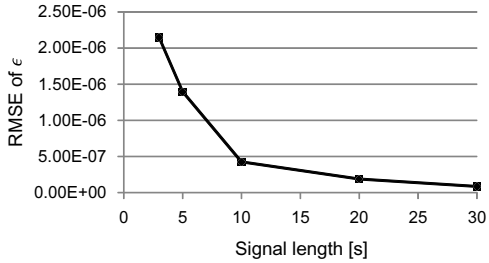

Fig. 3　Root mean squared error (RMSE) of estimated sampling frequency mismatch ϵ.

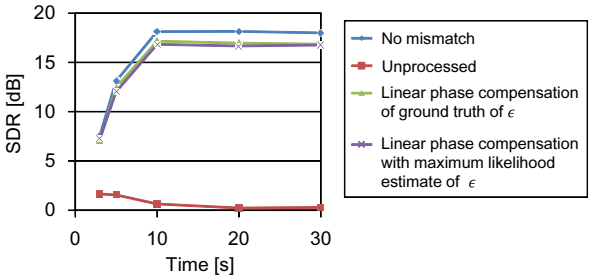

Fig. 4　Signal to distortion ratios (SDRs) of BSS results.

5　おわりに

本稿では，文献 [4] において提案した，非同期マイクロホンアレーのための観測チャネル間サンプリング周波数ミスマッチのブラインド補償を，効率的な計算で達成する手法について述べた．STFT 領域のサンプリング周波数ミスマッチの影響をフレームの位置に線形な位相シフトとモデル化し，観測される音源が定常で移動がないと仮定した最尤推定は，解が解析的に得られないために計算量の大きな離散値全探索が必要であった．そこで尤度関数が大域最適解周辺で単峰性を示すことに着目し，荒い離散値全探索による範囲の絞り込みと黄金分割探索による高速な推定を用いる効率的な探索法を提案した．黄金分割探索の計算効率を評価し，またサンプリング周波数ミスマッチ補償のブラインド音源分離への寄与を確認した．

参考文献

[1] Liu, Proc. IWAENC, 2008.
[2] Ono *et al.*, Proc. WASPAA, 161-164, 2009.
[3] Markovich-Golan *et al.*, Proc. IWAENC, 2012.
[4] 宮部ほか, 音講論 (秋), 689-670, 2005.
[5] Makino *et al.*, "Blind Speech Separation," Springer, 2007.
[6] Gill *et al.*, "Practical Optimization," Academic Press, 1981.
[7] Ono, Proc. WASPAA, 189-192, 2011.
[8] Vincent *et al.*, Proc. ICA, 552-559, 2007.