非整数サンプルシフトのフレーム分析を用いた非同期録音の同期化*

○宮部滋樹 (筑波大), 小野順貴 (NII/総研大), 牧野昭二 (筑波大)

1 はじめに

複数の録音機器を用いてアレー信号処理を行う分散型マイクロホンアレーでは，チャネル間のサンプリング周波数ミスマッチから生じる時刻ドリフトの補償が必須である。我々は，短時間フレーム内におけるドリフトがフレーム分析長よりも十分小さいと見なして無視し，ドリフトをフレームの分析時刻を線形にずらす効果と捉えたモデル化により，ドリフトの最尤推定に基づくブラインドな推定と，時間周波数領域の線形位相による補償を提案した。しかし録音時間が長くサンプリング周波数ミスマッチが大きいデータに対しては，1つのフレームが分析する時刻のチャネル間の差異がフレーム長に対して十分小さくならないために，線形位相によるフレーム内の波形シフトだけでは十分な同期の補償ができない。本稿では，フレーム内の波形シフトではなくフレームの分析時刻を合わせることにより，長時間録音でも高精度にチャネル間同期を取る手法を提案する。

2 問題設定

2つのマイクロホンの位置の音圧 $\mathrm{x}_1(t), \mathrm{x}_2(t)$ (t は連続時刻) が別々の AD 変換器でサンプリングされて離散信号 $x_1(n), x_2(n)$ が得られているものとする (n は離散時刻)。第1チャネルのサンプリング周波数が f_s，第2チャネルのサンプリング周波数がサンプリング周波数ミスマッチ ϵ ($|\epsilon| \ll 1$) を持つ $(1+\epsilon) f_\mathrm{s}$ とすると，それぞれのチャネルの連続信号と離散信号の関係は

$$x_1(n) = \mathrm{x}_1\left(\frac{n}{f_\mathrm{s}}\right), \quad n = 0, \ldots, N_1 - 1 \tag{1}$$

$$x_2(n) = \mathrm{x}_1\left(\frac{n}{(1+\epsilon) f_\mathrm{s}} + T_{21}\right), \quad n = 0, \ldots, N_2 - 1 \tag{2}$$

のように表される。ここで t の原点は第1チャネルの録音開始時刻，T_{21} は第2チャネルの録音開始時刻，N_1, N_2 はそれぞれのチャネルの離散信号のサンプル数とする。ここで，同時刻を参照する第1，2チャネルそれぞれの離散時刻を n_1, n_2 とすると，これら2つの離散時刻の関係は以下のように表される。

$$n_2 = \phi_{21}(n_1) \tag{3}$$

$$\phi_{21}(n) = (1+\epsilon)(n - D_{21}) \tag{4}$$

$$D_{21} = f_\mathrm{s} T_{21} \tag{5}$$

そのため整数値の n_1 に対しては n_2 は一般に非整数となり，$x_2(n)$ の同期を補償した信号 $\hat{x}_2(n) = x(\phi(n))$ を正確に求めるためには無限長の sinc 関数畳込みが必要となる。そのため同期の推定と補償には何らかの近似による簡略化が必要となる。

サンプリング周波数ミスマッチ ϵ と録音開始オフセット D_{21} の推定は，アレー信号処理に必要な精度で求める必要がある。チャネル間の時刻差が線形に変化するドリフトは，到来時間差が音源ごとに固有であることを仮定したアレー信号処理を破綻させてしまうため，サンプリング周波数ミスマッチ ϵ の推定は精密に行う必要がある。録音開始時刻オフセット D_{21} については，ブラインド音源分離や SNR 最大化ビームフォーマのような，音源方位を明示的にパラメタとして与えないアレー信号処理では必ずしも必要はなく，フレーム分析がチャネル間でほぼ同じ時間区間を参照する精度で十分である。そのため数サンプル程度の誤差は許容される。そこで本稿では ϵ の高精度な推定と D_{21} の大まかな推定を求め，それらを補償するチャネル間の同期について議論する。

3 時間周波数領域における近似的同期

アレー信号処理は一般に時間周波数領域で行われるため，本稿では同期信号を時間周波数領域で求める問題について議論する。

3.1 短時間フレームにおける同期モデル

まずフレーム分析区間のような短時間では，サンプリング周波数ミスマッチによるドリフトが無視できることを示す。第一チャネルにおける離散時刻 $(n_1 + m)$ と同じ連続時刻を参照する第二チャネルの離散時刻 $n_2 = \phi_{21}(n_1)$ は，式 (5) より以下のように与えられる。

$$\begin{aligned}\phi_{21}(n_1 + m) &= (1+\epsilon)((n_1 + m) - D_{21}) \\ &= \phi_{21}(n_1) + (1+\epsilon) m\end{aligned} \tag{6}$$

ここで $|m\epsilon| \ll 1$ となる条件では

$$\phi_{21}(n + m) \approx \phi_{21}(n) + m \tag{7}$$

と近似することができる。これは，時刻 n から時刻 $n+m$ の間で起こるドリフトによる波形の伸縮を無視することに相当する。このようなモデル従い，チャネル間で同期したフレーム分析は，単にフレーム中心時刻が同じ連続時刻を参照する $n_i, i = 1, 2$ となる以下もので近似できる。

$$x_i^{\mathrm{fr}}(l, n_i) = w(l)\, x_i\left(l + n_i - \frac{L}{2}\right) \tag{8}$$

ここで l はフレーム内のサンプル番号，$w(l)$ は完全再合成が可能な窓関数，L はフレーム長を表す。録音機器間のサンプリング周波数のミスマッチは 10^{-5} オーダーに収まるのが一般的であり，またアレー信号処理における典型的なフレーム長は 0.1 s オーダーであるため，時刻の誤差の最悪値 $|\epsilon L|/2$ は典型的には

*Synchronization of asynchronous recording using frame analysis with non-integer sample shift. by MIYABE, Shigeki (University of Tsukuba), ONO, Nobutaka (Institute of Informatics/the Graduate University for Advanced Studies), MAKINO, Shoji (University of Tsukuba)

0.1 µs オーダーになる。また，このような時刻の誤差があるサンプルはフレームの始端と終端の付近であり，一般的な窓関数 $w(l)$ の選択では小さな値が掛けられてその誤差は抑圧される。

式 (8) の第 2 チャネルを第 1 チャネルに同期させたフレーム分析 $x_2^{\mathrm{fr}}(l, n_2)$ を，完全再合成が可能な n_1 のフレームシフトに対して求め，$x_1^{\mathrm{fr}}(l, n_1)$ を $x_1(n)$ に戻すのと同じフレームシフトのオーバーラップ加算を行うことにより，第 2 チャネルを第 1 チャネルに同期させた長時間信号 $\hat{x}_2(n)$ を得ることができる。しかし前述のように整数値の第 1 チャネルの離散時刻 n_1 と同時刻を参照する第 2 チャネルの離散時刻 n_2 は一般に非整数であり，非整数の n_2 に対する式 (8) のフレーム分析を求める問題は単純ではない。次節ではこれを近似的に達成する効率的な手法について述べる。

3.2 時間周波数領域での同期

式 (8) で表される，第一チャネルの時刻 n_1 を中心とするフレーム分析に同期した，非整数の第二チャネルの非整数の時刻 n_2 を中心とするフレーム分析 $x_i^{\mathrm{fr}}(l, n_i)$ を単純化して近似的に求めるために，我々はまず n_2 に最も近い整数時刻を中心とするフレーム分析を行い，次に線形位相を用いたフレーム内の波形シフトにより，小数点以下の時刻補償を行う。まず n_2 に最も近い整数 $\lfloor n_2 \rceil$ を中心とするフレーム分析 $x_2^{\mathrm{fr}}(l, \lfloor n_2 \rceil)$ を求める。

$$x_2^{\mathrm{fr}}(l, \lfloor n_2 \rceil) = w(l)\, x_2\left(l + \lfloor n_2 \rceil - \frac{L}{2}\right) \tag{9}$$

ここで $\lfloor \cdot \rceil$ は最も近い整数への丸めを表す。次に離散フーリエ変換を用いて時間周波数領域に写像する。

$$X_2(k, \lfloor n_2 \rceil) = \sum_{l=0}^{L-1} x_2^{\mathrm{fr}}(l, n) \exp\left(-\frac{2\pi \jmath k l}{L}\right) \tag{10}$$

ここで $\jmath = \sqrt{-1}$，$k = -L/2, \ldots, L/2 - 1$ は離散周波数インデックスを表す。ただしこの離散フーリエ変換は実際の処理では高速フーリエ変換で置き換える。そして n_2 の丸め $\lfloor n_2 \rceil$ による，小数点以下の時間シフトの遅延 $n_2 - \lfloor n_2 \rceil$ を線形位相によりフレーム内で補償し，第 1 チャネルの短離散フーリエ変換 $X_1(k, n_1)$ と同期した第 2 チャネルの短時間フーリエ変換 $\hat{X}_2(k, n_2)$ を得る。

$$\hat{X}_2(k, n_2) = X_2(k, \lfloor n_2 \rceil) \exp\left(\frac{2\pi \jmath k\,(n_2 - \lfloor n_2 \rceil)}{L}\right) \tag{11}$$

アレー信号処理は一般に時間周波数領域で行われるが，以上のような手順で同期した短時間フーリエ変換 $X_1(k, n_1), \hat{X}_2(k, n_2)$ を用いることによりドリフトの影響を取り除いたアレー信号処理を行うことができる。また，$\hat{X}_2(k, n_2)$ を逆離散フーリエ変換したのちにオーバーラップ加算を施すことにより，$x_2(n)$ と時間同期した第 2 チャネルの時間領域信号が得られる。

4 時間領域での大まかな同期

前節では同期の演算をサンプリング周波数ミスマッチ ϵ と録音開始時刻オフセット D_{21} が与えられたという条件で議論したが，これらは一般に未知であり，同期のためには推定する必要がある。我々は [2] において，これらを時間周波数領域で最尤規範の線形位相補償の最適化により求める手法を提案したが，この手法ではドリフトで補償されるべきフレーム内の時間シフトが，全てのフレームを通してフレーム幅よりも十分小さいことを仮定しており，非常に長い録音データや大きなサンプリング周波数ミスマッチのもとでは，十分な推定と同期を得ることができない。本節では，最尤法によるサンプリング周波数ミスマッチ推定と線形位相による補償の前処理として，同期を大まかに推定・補償した時間周波数分析を得る手法について議論する。

4.1 2 つの同時刻ペアを用いた同期情報の同定

おおまかな推定の議論に進む前に，サンプリング周波数ミスマッチ ϵ と録音開始時刻オフセット D_{21} が同定される条件について述べる。同じ連続時刻を参照する各チャネルの離散時刻の 2 つのペア $\{n_{\mathrm{A}1}, n_{\mathrm{A}2}\}$ と $\{n_{\mathrm{B}1}, n_{\mathrm{B}2}\}$ $(n_{\mathrm{A}i} < n_{\mathrm{B}i}, i = 1, 2)$ が与えられたとすると，それらは式 (5) より以下の関係を満たす。

$$n_{\mathrm{A}2} = (1 + \epsilon)\,(n_{\mathrm{A}1} - D_{21}) \tag{12}$$

$$n_{\mathrm{B}2} = (1 + \epsilon)\,(n_{\mathrm{B}1} - D_{21}) \tag{13}$$

この連立方程式の解として，ϵ と D_{21} が以下のように同定される。

$$\epsilon = \frac{n_{\mathrm{B}2} - n_{\mathrm{A}2}}{n_{\mathrm{B}1} - n_{\mathrm{A}1}} - 1 \tag{14}$$

$$D_{21} = \frac{n_{\mathrm{A}1} n_{\mathrm{B}2} - n_{\mathrm{A}2} n_{\mathrm{B}1}}{n_{\mathrm{B}2} - n_{\mathrm{A}1}} \tag{15}$$

従って，同時刻を参照する時刻ペアが $\{n_{\mathrm{A}1}, n_{\mathrm{A}2}\}$ と $\{n_{\mathrm{B}1}, n_{\mathrm{B}2}\}$ のように 2 つ得られれば，ϵ と D_{21} の同定が可能だということがわかる。このような同時刻を参照する離散時刻ペアを正確に推定することは難しいが，粗い推定が得られれば ϵ と D_{21} の粗い推定を求めることができ，粗く同期を取ったフレーム分析を得ることができる。

4.2 相関を用いた同時刻ペアの推定

同じ連続時刻を参照する離散時刻のペア 2 つ $n_{\mathrm{A}i}, n_{\mathrm{B}i}, i = 1, 2$ の推定について議論する。基本的な考え方は，チャネル間で同時刻付近に相当する信号は相関が高くなるため，短時間の相関を基にして時刻の対応関係を推定するということである。一般に観測信号は到来時間差の違う複数の音源の混合であり，この到来時間差の違いのせいで相関では完全な時刻ペアの推定を得ることができない。しかし，マイクロホン同士が近接して置かれている場合にはこの時間差の違いは小さなものとなり，式 (14)，(15) において $n_{\mathrm{B}i} \ll n_{\mathrm{B}i}, i = 1, 2$ であればこの誤差の推定への影響を相対的に小さくすることができる。相関の評価する短時間区間には，録音の開始・終了付近でパワーが閾値を超える区間を用いる。ここでチャネル 1 とチャネル 2 のどちらが先に録音を開始・終了するかは未知であるため，短時間区間は録音・開始付近の両方において両方のチャネルから選び出し，反対側のチャネルにより相関の高い区間が見つかる方を採用する。処理の模式図を Fig. 1 に示す。

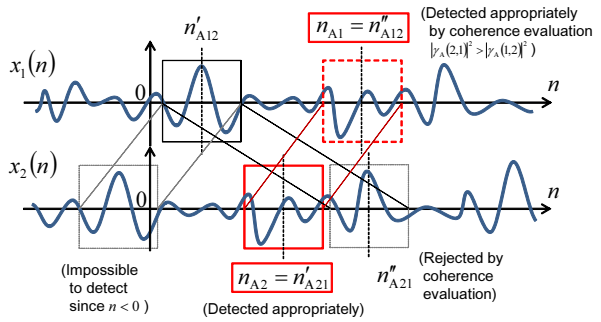

Fig. 1 Estimation of $n_{A1} and n_{A2}$ correspond to the same continuous time near the beginning of the asynchronous recording.

まず，録音開始付近の時刻 $n_{Ai}, i = 1, 2$ の候補として，それぞれのチャネルにおいて，時刻 $N/2$ 以降で最初に振幅が閾値 δ を超えるサンプル $n'_{Aij}, (i, j) = (1, 2), (2, 1)$ をみつける。

$$n'_{Aij} = \arg\min_n \{|x_i(n)| > \delta\} + N/2, \quad n \geq N/2 \tag{16}$$

ここで N は相関評価に用いる区間長を表す。録音時刻終了付近でも同様に時刻 $n_{Bi}, i = 1, 2$ の候補として，それぞれのチャネルから時刻 $N_i - N/2$ より前の振幅が閾値 δ を超えるサンプルをみつける。

$$n'_{Bij} = \arg\max_n \{|x_i(n)| > \delta\} + N/2, \quad n < N_i - N/2 \tag{17}$$

次に，$h = \mathrm{A}, \mathrm{B}, (i, j) = (1, 2), (2, 1)$ の組み合わせのそれぞれについて，選択された区間との相関が高くなるサンプル n''_{hij} を求める。

$$n''_{hij} = \arg\max_{n''} \sum_{n=-L/2}^{N/2-1} x_i\left(n + n'_{hij}\right) x_j\left(n + n''\right) \tag{18}$$

この相関の計算は直接求めると計算量が大きくなるが，等価な処理は区間 $x_i\left(n + n'_{hij}\right)$ の時刻を符号反転した高速フーリエ変換に基づくオーバーラップ加算畳込みによって効率的に達成することができる。

続いて，$h = \mathrm{A}, \mathrm{B}$ のそれぞれについて，相関が最大となる反対側のチャネルの短時間区間との相関関数が大きい方をみつける。これは 2 つのインデックスペア $(i, j) = (1, 2), (2, 1)$ のうちの，以下で与えられるコヒーレンス $|\gamma_h(i, j)|^2$ が大きくなる方を選ぶ問題として定式化される。

$$(i^*_h, j^*_h) = \arg\max_{(i,j)=(1,2),(2,1)} |\gamma_h(i, j)|^2$$

$$|\gamma_h(i, j)|^2 = \frac{\left|\sum_{n=-\frac{L}{2}}^{N-1} x_i\left(n + n'_{hji}\right) x_j\left(n + n''_{hij}\right)\right|^2}{\sum_{n=-\frac{L}{2}}^{N-1} \left|x_i\left(n + n'_{hji}\right)\right|^2 \sum_{n=-\frac{L}{2}}^{N-1} \left|x_j\left(n + n''_{hij}\right)\right|^2} \tag{19}$$

最終的に，2 つの同時刻ペア $\{n_{A1}, n_{A2}\}, \{n_{B1}, n_{B2}\}$ は相関関数が高いものとして以下のように与えられる。

$$n_{hi^*_h} = n'_{hj^*_h i^*_h} \tag{20}$$

$$n_{hj^*_h} = n''_{hi^*_h j^*_h} \tag{21}$$

これらの推定を式 (14)，(15) に代入することにより，ϵ と D_{21} の荒い推定が得られ，これらを 3 節の処理に用いることにより，荒く最適化した時間周波数分析が得られる。

5 サンプリング周波数ミスマッチの詳細な推定

荒く最適化されたフレーム分析の時間周波数領域において，サンプリング周波数ミスマッチの詳細な推定を行う。未知パラメタ D_{21} と ϵ のうち，D_{21} については厳密に推定する必要はなく，前節の推定で十分な制度であると考えられる。そこで，サンプリング周波数ミスマッチの荒い推定 ϵ に以下のような修正を加えるパラメタを，録音開始時刻オフセットに極力修正を加えない形で求める。

$$\begin{aligned} n_2 &= \phi'_{21}(n_1; \epsilon') \\ &= (1 + \epsilon)\left((1 + \epsilon')(n_1 - M) + M - D_{21}\right) \end{aligned} \tag{22}$$

ここで M はフレーム分析が適用される区間の中心時刻を表す。

サンプリング周波数ミスマッチの荒い推定 ϵ は既に得られているため，その修正パラメタ ϵ' の最適値はは典型的に小さなものとなる。そのため，これを用いた同期の補正 $\hat{X}'_2(k, \phi'_{21}(n_1; \epsilon'))$ は線形位相により十分高い精度で得られる。

$$\begin{aligned} &\hat{X}'_2(k, \phi'_{21}(n_1; \epsilon')) \\ &= \hat{X}_2(k, \phi_{21}(n_1)) \exp\left(\frac{2\pi \jmath k \epsilon'(n_1 - M)}{L}\right) \end{aligned} \tag{23}$$

このパラメタ ϵ' の推定は時間周波数領域の 2 チャネル信号

$$\hat{\mathbf{X}}'(k, n_1; \epsilon') = \left[X_1(k, n_1), \hat{X}'_2(k, \phi'_{21}(n_1; \epsilon'))\right]^{\mathsf{T}} \tag{24}$$

(T は転置）を用いて表される以下の対数尤度関数 $L(\epsilon)$ の最大化により求めることができる。

$$L(\epsilon') = -\sum_k \log\det \sum_{n_1} \hat{\mathbf{X}}'(k, n_1; \epsilon') \hat{\mathbf{X}}'(k, n_1; \epsilon')^{\mathsf{H}} \tag{25}$$

ここで H は複素共役転置を表す。この回の探索は黄金分割探索により効率的に行うことができる [2]。

ここで ϵ' を求めることは，ϵ と D_{21} に以下のような修正を加えることに相当する。

$$\epsilon \leftarrow (1 + \epsilon)(1 + \epsilon') - 1 \tag{26}$$

$$D_{21} \leftarrow \frac{\epsilon' M + D_{21}}{1 + \epsilon'} \tag{27}$$

この修正は小さなものであるため，時間周波数領域の同期としては ϵ' を直接用いて位相補償を行う式 (23) でも十分であるが，第 3 節の処理にこれらの修正パラメタを代入してより精度を高めた分析を得ることも可能である。

Table 1　Experimental conditions.

Signal length [s]	3, 5, 10, 20, 30, 60, 120, 300 and 600
Reverberation time	T_{60} of 130 ms
Frame length L	4,096 samples
Frame shift	2,048 samples
Source distances	1.5 m
Source directions	[-50°, 30°], [-60°, -10°]
Microphone spacing	2 cm
N	16,000 samples
Candidates of discretized search [2]	10 samples from $[-2\times10^{-4}, 2\times10^{-4}]$

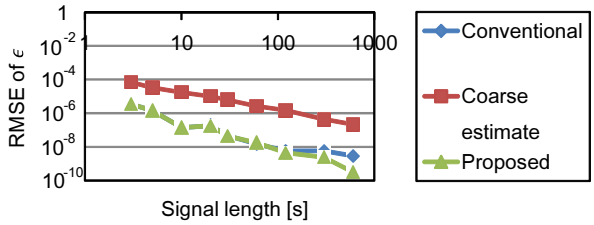

Fig. 2　Root mean squared errors (RMSEs) of estimations of sampling frequency mismatches ϵ.

6　評価実験

提案したブラインド同期の分散型マイクロホンアレーを用いたブラインド音源分離への寄与を評価するため，2 つのマイクロホンによる 2 話者の混合の観測に人工的なサンプリング周波数のミスマッチを与えて同期補償と音源分離を評価する。

観測信号は実測したインパルス応答を日本語の単語発話を接続した音声に畳込むことにより作成した。話者数は 4 人で，全ての 12 組み合わせの平均を評価した。サンプリング周波数は，元々は 16,000 Hz で同期がとれていたものに対して，±0.5, ±1, ±1.5 Hz のミスマッチを与えて非同期録音を模擬した。これらのミスマッチの大きさは実際の機器間のミスマッチとして現実的な大きさである。人工的なサンプリング周波数ミスマッチにはポリフェーズフィルタによるリサンプリングを用いた。ブラインド音源分離の手法としては補助関数法独立ベクトル分析（AuxIVA）[3] を用いた。それ以外の条件は Table 1 に示す。

サンプリング周波数ミスマッチ ϵ の推定精度を平均二乗誤差平方根 (RMSE) による評価を Fig. 2 に示す。我々が以前に提案した手法 [2] および第 4 節の荒い推定の精度と提案手法を比較する。どの手法も信号長に応じて推定精度が向上し，すべての信号長において提案手法により最も高精度な推定が得られている。

音源分離性能の比較を Fig. 3 に示す。評価尺度には Signal-to-distortion ratio (SDR) [4] を用いた。No mismatch と示される折れ線はサンプリング周波数のミスマッチがない条件での音源分離性能を表し，言うまでもなくどのサンプリング周波数のミスマッチの補償よりも高い性能を示す。それに対して Unprocessed はサンプリング周波数ミスマッチを補償しないまま音源分離を行った結果で，SDR が 0 付近になりほとんど分離ができていない。そのためこの条件では同期の補償が不可欠であるということがわかる。Upper limit of synchronization は非整数フレーム分析による同期補償に真のパラメタを与えた場合で，提案手法の性能限界を表す。Conventional は我々が以前に提案した従来手法で，信号長が 60 秒を超えたあたりから急激に性能が劣化していることがわかる。提案手法は長時間の録音に対しても頑健な同期の補償を達成し，Upper limit と同等な精度まで分離性能を回復させている。以上より，提案手法の有効性が確認された。

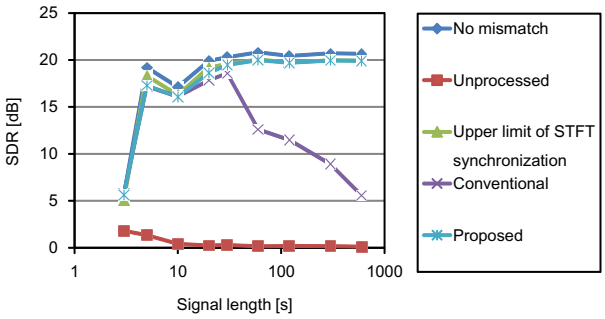

Fig. 3　Source separation performance with SDR measure.

7　おわりに

本稿では，非整数サンプルシフトのフレーム分析による，長時間録音にも頑健な非同期録音のブラインド同期手法を提案した。サンプリング周波数のミスマッチにより生じるチャネル間の時刻のドリフトを，短時間での小さな影響を無視して分析フレームの中心時刻が線形に変化するとみなし，フレームの中心時刻を補償することによりチャネル間の同期を回復させた。このような分析ではフレーム中心時刻は非整数サンプルとなるため，整数に丸めた時間領域の窓分析と，小数点以下の時刻シフトを線形位相で与えることにより非整数サンプルシフトのフレーム分析を達成した。また，時間領域の相関によりパラメタを荒く推定して非整数サンプルシフトのフレーム分析を行い，より詳細な補償を最尤な線形位相を与えることにより，効率的で高精度な同期を行った。評価実験により，提案手法は長時間録音でも頑健に同期を行い，アレー信号処理の性能を回復させることを確認した。

参考文献

[1] Liu *et al.*, *Proc. IWAENC*, 2008.

[2] Miyabe *et al.*, *Proc. ICASSP*, pp.674-678, 2013.

[3] Ono, *Proc. WASPAA*, pp.189-192, 2011.

[4] Vincent *et al.*, *Proc. ICA*, pp. 552–559, 2007.