非同期録音信号の線形位相補償によるブラインド同期と音源分離への応用*

◎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] は，会議録音の音声強調のために参加者が持ち寄った携帯電話やボイスレコーダなどの複数の携帯型録音機器の同時録音を用いるというもので，専用の大規模な録音装置ではなく一般的な機器による安価で柔軟な構成を行えることが利点である．しかし，マイクロホン素子の配置が未知となることや [2, 3]，各チャネルごとの録音が同期していないことなど [1, 4]，通常のマイクロホンアレー信号処理では扱われない問題を解決する必要がある．最も重要な問題の一つに，各録音装置が別々の A/D 変換器を使用しているためにチャネル間のサンプリング周波数が違うものになることが挙げられ，適切な補正を施さなければアレー信号処理の性能は大幅に劣化してしまう．

本稿では，ブラインド音源分離 [5] の前処理のための，サンプリング周波数のチャネル間のミスマッチを推定し補正する手法について述べる．チャネル間のサンプリング周波数のミスマッチは，時刻の単位のずれのために信号間の時間差がドリフトしていくような効果を生じ，位相差が等速に変化して音源位置が疑似的に変化してしまうため，各音源は移動せず固有の位相差を持つという大多数の音源分離手法の仮定を破綻させてしまう．そこで，全ての音源は移動せず定常であると仮定し，音源の移動をなくす補償の推定によりミスマッチの推定を行う．音源分離手法の多くは時間周波数領域で信号分析を行うため，ミスマッチの補償は時間周波数領域で行う．これにより，一般的なサンプリング周波数の補正に必要な sinc 関数による離散信号の補間を省いた簡便なものとなる．

2 サンプリング周波数ミスマッチの線形位相補償

2.1 ミスマッチの時間領域モデル

本稿では同じ公証サンプリング周波数の録音機器が使用されるか，あるいは同じになるように観測信号がリサンプリングされていると仮定し，それでもわずかに残る機器ごとのミスマッチを扱う．一般的な録音装置の公称サンプリング周波数からの誤差は 10 ppm ($=10^{-5}$) の数倍以内であるが，これは 1 秒間に 10 μs オーダーの誤差に相当し，サンプル幅より小さい時間差を分析する音源分離を破綻させるのに十分な大きさのドリフトを引き起こす．なお本稿での議論は 2 チャネルの間のサンプリング周波数のミスマッチに限定するが，全てのチャネルのサンプリング周波数を特定の 1 つのチャネルに合わせるなどにより，容易に 3 チャネル以上に拡張することができる．

いま，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 $x_1(n_1)$, $x_2(n_2)$ (n_1, n_2 はサンプル番号）が得られたとする．ここで $x_1(n_1)$ のサンプリング周波数は f_s, $x_2(n_2)$ のサンプリング周波数は未知のミスマッチ ϵ により表される $(1+\epsilon) f_\mathrm{s}$ であるとする．このこのとき離散信号と連続信号の関係は以下のように表される．

$$x_1(n_1) = \mathrm{x}_1\left(\frac{n_1}{f_\mathrm{s}}\right) \tag{1}$$

$$x_2(n_2) = \mathrm{x}_2\left(\frac{n_2}{(1+\epsilon) f_\mathrm{s}} + \Delta T_{21}\right) \tag{2}$$

ここで t の時間原点を $x_1(n_1)$ の録音開始時刻とし，ΔT_{21} は $x_1(n_1)$ に対する $x_2(n_1)$ の録音開始時刻の遅れを表す．n_1 が参照する時刻は n_2 で表すと

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

サンプル目に相当し，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していくことになる．

2.2 ミスマッチの時間周波数領域モデリング

短時間のフレーム内のドリフトは無視できるものと考え，フレームの時刻のドリフトを補償する問題を議論する．まず式 (2) の時間原点の遅れ ΔT_{21} を補償する．サンプリング周波数のミスマッチは無視して n_1 と n_2 を同質のものとみなし，長時間の相関を用いて大まかな補正 τ_{21} を求め，これが最大となるよう $x_2(n_2)$ をシフトする．

$$\tau_{21} = \arg\max_m \sum_n x_1(n)\, x_2(n-m) \tag{4}$$

$$x_2(n_2) \leftarrow x(n_2 - \tau_{21}) \tag{5}$$

ここではこの補正誤差はフレーム幅よりも十分に小さくブラインド音源分離では問題とならないと考える．

次に，同様に n_1, n_2 を同質のものとみなして $x_j(n_j)$, $j=1,2$ のフレーム分析

$$x_j^{\mathrm{fr}}(l,r) = w(l)\, x_j(l+rM) \tag{6}$$

を得る．ここで l, r, M は全て整数で，それぞれ L をフレーム長として $l=0,\ldots,L-1$ はフレーム内のサンプル番号，フレーム数を R として $r=0,\ldots,R-1$ はフレーム番号，M はフレームシフト長を表し，また $w(l)$ は再合成が可能な窓関数とする．そして $x_j^{\mathrm{fr}}(l,r)$ の高速フーリエ変換により時間周波数領域信号 $X_j(k,r)$ ($k=-L/2+1,\ldots,L/2$ は周波数番号）を得る．ここで録音開始時刻の影響については補正されたので無視して考えると，$x_2^{\mathrm{fr}}(l,r)$ を $x_1^{\mathrm{fr}}(l,r)$ と同じ時刻を参照させるためには，式 (3) より $x_2^{\mathrm{fr}}(l,r)$ に $\epsilon r M$ サンプルに相当する遅延を与える必要があるので，この遅延に相当する円状シフトを施してミスマッチを補償した時間周波数領域信号 $\hat{X}_2(k,r;\epsilon)$ を求める．

$$\hat{X}_2(k,r;\epsilon) = X_2(k,r) \exp\left(-\frac{2\pi \jmath \epsilon r M k}{L}\right) \tag{7}$$

ここで $\jmath = \sqrt{-1}$ である．以下では最尤推定による ϵ の推定方法について述べる．

---

*Blind synchronization of unsynchronized recording signal by linear phase compensation and its application to source separation. by MIYABE, Shigeki (University of Tsukuba), ONO, Nobutaka (National Institute of Informatics), MAKINO, Shoji (University of Tsukuba)

Table 1　Experimental conditions

Source separation method	Auxiliary function based IVA [6]
Rverberation time	T_{60} of 130 ms
Frame length L	2,048 samples
Frame shift M	1,024 samples
Source distance	1.5 m
Source directions	[-50°, 30°], [-60°, -10°]
Microphone spacing	2 cm
Candidates of discrete search of ϵ	10,000 samples from $\left[-2^{-10}, 2^{-10}\right]$

2.3　ミスマッチの推定

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

$$\hat{\mathbf{X}}(k,r;\epsilon) = \left[X_1(k,r), \hat{X}_2(k,r;\epsilon)\right]^{\mathsf{T}} \tag{8}$$

は離散周波数 k 毎に定常であると仮定できるため，この仮定に基づいた最尤推定により ϵ を求める．$\hat{\mathbf{X}}(k,r;\epsilon)$ の分布を零平均，共分散行列 $\mathbf{V}(k)$ の多変量複素正規分布とおいた場合の対数尤度は

$$J(\mathbf{V},\epsilon) = \sum_{k,r}\Bigg(-\log 2\pi^2 - \log\left|\det \mathbf{V}(k)\right| - \hat{\mathbf{X}}(k,r;\epsilon)^{\mathsf{H}} \mathbf{V}(k)^{-1} \hat{\mathbf{X}}(k,r;\epsilon)\Bigg) \tag{9}$$

と表せる．ここで $\{\cdot\}^{\mathsf{H}}$ は複素共役転置を表し，$\mathbf{V}$ は $\mathbf{V}(k)$ の集合 $\{\mathbf{V}(k)\,|k = -L/2+1,\ldots,L/2\}$ とする．共分散行列 $\mathbf{V}(k)$ は未知であるため，$\hat{\mathbf{X}}(k,r;\epsilon)$ を用いた標本推定 $\hat{\mathbf{V}}(k;\epsilon)$ で置き換える．

$$\hat{\mathbf{V}}(k;\epsilon) = \frac{1}{R}\sum_{r=0}^{R-1} \hat{\mathbf{X}}(k,r;\epsilon)\,\hat{\mathbf{X}}(k,r;\epsilon)^{\mathsf{H}} \tag{10}$$

この尤度最大化は解析的に解くことができないため，$J\left(\hat{\mathbf{V}}(k;\epsilon),\epsilon\right)$ を最大化する ϵ の散値全探索により ϵ を推定する．

3　実験

提案手法の有効性を検証するため，混合音声のサンプリング周波数を人工的に変えた実験により，ミスマッチ ϵ の推定精度と音源分離性能を評価する．

使用した観測信号は 2 話者発話の 2 チャネルの畳込み混合で，実測したインパルス応答を使用し，SNR が 20 dB となるよう白色雑音を重畳した．音声は ATR データベースの 4 人の話者の単語発話を話者ごとに繋いで作成した．サンプリング周波数を人工的に変更したデータの作成には MATLAB の resample コマンドを用いた．変更したサンプリング周波数は，元のデータの 16,000 Hz に対して $16,000 \pm 1$, $16,000 \pm 2$ Hz の 4 種類で，$16,000 \pm 1$ kHz のものは $\epsilon = 62.5$ ppm に相当し，録音装置間の誤差として十分現実的なものである．その他の実験条件を Table 1 に示す．

まずサンプリング周波数の補正精度を評価した．用いた音声の長さは 3, 5, 10 秒の 3 種類で，ミスマッチ ϵ の推定の二乗平均平方根誤差 (RMSE) を Table 2 に示す．観測データが長いほど誤差は小さくなるが，3 秒の観測信号でも 3 秒の観測信号でも，サンプリング周波数ミスマッチを元のものから 1 桁以上補償できていることがわかる．

Table 2　Estimation errors for sampling frequency mismatch

Signal length [s]	3	5	10
RMSE [ppm]	3.48	1.46	0.27

Table 3　Source separation accuracy

	SDR [dB]	SIR [dB]
No mismatch	7.98	14.81
Unprocessed	0.33	1.35
Ideal compensation	8.10	14.74
Proposed method	7.96	14.48

最後に 10 秒の観測信号の音源分離性能の比較により，提案手法の効果を評価する．サンプリング周波数を変更しなかった場合の分離精度（no mismatch），サンプリング周波数のミスマッチに対して補償を行わなかった場合の分離精度（unprocessed），真のサンプリング周波数のミスマッチを与えた場合の線形位相補正による分離精度（ideal compensation）と提案手法によるサンプリング周波数の修正と補正による分離精度（proposed）を比較した．評価尺度には信号対歪比（SDR）と信号対混信比（SIR）を用いた．結果を Table 3 に示す．No mismatch がゼロに近い低い評価値であることから，このサンプリング周波数ミスマッチは補償なしでは音源分離が破綻してしまう厳しい条件であるということが確認できる．また idel compensation が no mismatch とほぼ同等の性能を示していることから，時間周波数領域の線形位相補償が有効であるということがわかる．また自動推定を行う proposed method と ideal compensation の性能差が小さいことから，最尤推定によるサンプリング周波数のミスマッチの推定の有効性が確認できる．

4　おわりに

本稿では，非同期マイクロホンアレーを用いたブラインド音源分離のための，サンプリング周波数ミスマッチの自動推定と線形位相補償による同期補正を提案した．実験により，提案手法はサンプリング周波数のミスマッチを高い精度推定精度で推定でき，かつ音源分離精度をサンプリング周波数のミスマッチがない場合と遜色のないレベルまで回復できるということを確認した．

参考文献

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

[2] Ono *et al.*, *Proc. WASPAA*, 161-164, 2009.

[3] Hasegawa *et al.*, *Proc. LVA/ICA*, 57-64, 2010.

[4] Robledo *et al.*, *Proc. WASPAA*, 34-37, 2007.

[5] Makino *et al.*, "Blind Speech Separation," Springer, 2007.

[6] Ono, *Proc. WASPAA*, 16-19, 2011.