非同期マイクロホンアレーの符号化録音における ビットレートと同期性能の関係*

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

1 はじめに

非同期マイクロホンアレーは，複数の独立した録音機器による同時録音を用いてアレー信号処理に用いる枠組で [1]，参加者が持ち寄った携帯電話やボイスレコーダ，ビデオカメラなどの，複数の機器の録音を集めて音声強調を行うなどの応用が期待されている。この枠組では，個々の各録音機器の A/D 変換器が同期を取らずに独立に動作するため，クロック周波数の個体差に起因するサンプリング周波数のミスマッチにより，チャネル間の時間差が時刻によって変化する，ドリフトと呼ばれる現象を引き起こす。そのため，同期を補償する何らかの処理を行わなければ，アレー信号処理の音源固有の位相差の分析を破綻させてしまう。この問題を解決するため，我々はこれまでの発表において，チャネル間のサンプリング周波数ミスマッチを，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手法を提案してきている [2]。

記憶装置への保存と通信の効率化のため，携帯型録音機器の多くは録音データを MP3 などの圧縮符号化に対応しており，また録音機器によっては圧縮した形式でのデータ出力しかサポートしていないものもある。そのため，符号化歪に対する頑健性は非同期マイクロホンアレーの信号処理の重要な課題である。本稿では，最尤法によるサンプリング周波数ミスマッチの符号化歪に対する頑健性評価について報告する。

2 最尤法による STFT 領域の同期補償

2.1 ミスマッチの時間領域モデル

いま，同時刻における 2 つのマイクロホンの連続信号 $x_1(t), x_2(t)$（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 は離散時刻）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パラメタ 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．

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

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

ここで t の時間原点を $x_1[n]$ の録音開始時刻とし，T_{21} は $x_1[n]$ に対する $x_2[n]$ の録音開始時刻の遅れを表す．これら 2 つの信号は同じ公称サンプリング周波数の録音機器で観測されるか，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ことを想定し，$|\epsilon| \ll 1$ であると仮定する．チャネル 1 とチャネル 2 の，同じ連続時刻 t を参照する離散時刻をそれぞれ n_1, n_2 とすると，これらは以下のように線形関数で表される関係にある。

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

したがって，n_1 と n_2 は長い時間が経過するとかけ離れた値となる。

2.2 時間周波数領域におけるミスマッチの推定と補償

フレーム内における時間差のドリフトは，1 フレームの時間区間が短いために十分小さく無視できるものとみなし，フレームで切り出される時刻がチャネル間でドリフトしていくのを補償する問題について議論する．また，録音開始時刻の差は相関を用いて容易に推定・補償することができるため，ここでは無視して $T_{21} = 0$ とみなした議論を行う。まず両チャネル $x_i[n], i = 1, 2$ のフレーム分析 $x_i^{\mathrm{fr}}[l, r]$ を求める。

$$x_i^{\mathrm{fr}}[l, r] = w[l]\, x_i[l + rM] \tag{4}$$

ここで，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 はフレームシフトのサンプル数を表し，また $w(l)$ は再合成が可能な窓関数とする．そして $x_i^{\mathrm{fr}}(l, r)$ の高速フーリエ変換により STFT 領域信号 $X_i[k, r]$（$k = -L/2+1, \ldots, L/2$ は周波数番号）を得る．両チャネルのフレーム $x_i^{\mathrm{fr}}[l, r], i = 1, 2$ に同じ連続時刻を参照させるためには，式 (3) より，第 2 チャネルのフレーム内の信号 $x_2^{\mathrm{fr}}(l, r)$ 全体に $\epsilon r M$ サンプルの遅延を与える必要がある。この遅延付与を以下のように線形位相で与えることにより，ドリフトを補償した STFT 領域信号 $\hat{X}_2[k, r; \epsilon]$ を求めることができる．

$$\hat{X}_2[k, r; \epsilon] = X_2[k, r] \exp\left(-\frac{2\pi \jmath \epsilon r M k}{L}\right) \tag{5}$$

ここで $\jmath = \sqrt{-1}$ である．

次にサンプリング周波数ミスマッチ ϵ の推定方法を述べる。観測される全ての音源は移動せず，かつ振幅が定常である仮定すると，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 2 チャネル観測信号

$$\hat{\mathbf{X}}[k, r] = \left[X_1[k, r], \hat{X}_2[k, r; \epsilon]\right]^{\mathsf{T}} \tag{6}$$

は離散周波数 k 毎に定常であると仮定できる。ただし $\{\cdot\}^{\mathsf{T}}$ は行列の転置を表す。正確な補償により定常性を十分に回復した信号 $\hat{\mathbf{X}}[k, r; \epsilon]$ が多変量正規分布に従うと仮定すると，定数項を除いた対数尤度関数は以下のようになる [2]。

$$J(\epsilon) = \sum_k -\log\det \sum_{r=0}^{R-1} \hat{\mathbf{X}}[k, r; \epsilon]\, \hat{\mathbf{X}}[k, r; \epsilon]^{\mathsf{H}} \tag{7}$$

---

* Relation between synchronization performance and bit rate of audio coding in asynchronous microphone array. by Shigeki MIYABE (University of Tsukuba), Nobutaka ONO (National Institute of Informatics / The Graduate University for Advanced Studies), Shoji MAKINO (University of Tsukuba), Yu TAKAHASHI (Yamaha)

ここで $\{\cdot\}^{\mathsf{H}}$ は複素共役転置，det は行列式を表す。この対数尤度を最大化する ϵ を探索することにより，サンプリング周波数ミスマッチの推定が達成される。

補償後の信号処理を行うためには，短い信号であれば，最適化の際に得られる最尤な ϵ による補償後 STFT 信号 $\hat{\mathbf{X}}[k,r;\epsilon]$ をそのまま用いることができ，時間領域信号が必要な場合にはこれを逆 STFT 分析することにより求めることができる。また長い信号の同期の場合には誤差が無視できなくなるが，その場合の補償方法として非整数サンプルシフトのフレーム分析 [3] を提案している。

2.3 符号化の影響

符号化歪の最尤推定への影響について議論する。最尤推定の性質を分析するために，式 (7) の行列式を展開して以下のように変形する。

$$
\begin{aligned}
J(\epsilon) \propto \sum_k -\log\Bigg(& E\left[|X_1[k,r]|^2\right] E\left[|X_2[k,r;\epsilon]|^2\right] \\
& - \left|E\left[X_1[k,r]\, X_2[k,r;\epsilon]^*\right]\right|^2 \Bigg) \\
\propto \sum_k -\log & \left(1 - \frac{\left|E\left[X_1[k,r]\, X_2[k,r;\epsilon]^*\right]\right|^2}{E\left[|X_1[k,r]|^2\right] E\left[|X_2[k,r]|^2\right]}\right)
\end{aligned}
\tag{8}
$$

ただし $\{\cdot\}^*$ は複素共役を，$E[\cdot]$ はフレーム平均 $\sum_{r=0}^{R-1}\{\cdot\}/R$ を表す。ここでは $E[|X_i[k,r]|], i = 1,2$ が ϵ に依存しないことと，式 (5) から自明な $|\hat{X}_2[k,r;\epsilon]| = |X_2[k,r]|$ という関係を用いた。式 (8) の最終行の対数の中の第 2 項は，$X_1[k,r]$ と $\hat{X}_2[k,r;\epsilon]$ のコヒーレンスであることがわかる。したがって，尤度関数 $J(\epsilon)$ は線形位相補償によるコヒーレンスの向上を評価していることになる。移動しない少数の音源が存在する帯域では，ドリフトがなければコヒーレンスが高くなるため，ドリフトを補償して位相差の変化を止めることによりコヒーレンスが大きく回復する。反対に，最適な位相補償を行ってもコヒーレンスが小さくなる帯域は，ドリフトの有無でコヒーレンスがほとんど変化しないため，推定にはあまり寄与しない。

聴覚特性を用いて帯域ごとの量子化精度を動的に変更する MP3 などの符号化は，主に高域の位相をランダムに変化させることが知られている。そのため同期を回復しても高域のコヒーレンスが高くならず，推定に寄与する帯域が減少するために推定精度は低下する。しかしこれは推定を破綻させるような重大な副作用をもたらすものではなく，サンプリング周波数ミスマッチの最尤推定は符号化歪に対してある程度頑健であると考えられる。

3 実験

サンプリング周波数ミスマッチの推定精度への符号化歪の影響を調査するため，実収録のインパルス応答を用いて会議録音を想定した 2 名の音声混合の観測信号を作成し，人工的なサンプリング周波数ミスマッチと符号化歪を与えて推定精度を評価した。評価に用いる音源は，ATR 音声コーパス [4] の男女 2 人ずつの 20 秒の音声で，4 人から 2 人を選ぶ全ての組み合わせを評価した。観測信号はサンプリング周波数 48 kHz で作成し，第 2 チャネルをリサンプリングして $\pm 63, \pm 125$ ppm のサンプリング周波数ミスマッチを人工的に与えた。第 2 チャネルのみまたは両チャネルを，32, 48, 64, 80, 96, 128, 160 kbps の 7 種類となるビットレートの MP3 および MPEG-4 AAC により，チャネル単位で符号化・復号して符号化歪を与え，非圧縮の PCM とともに評価した。評価指標はパラメタ ϵ の推定の平均二乗誤差 (RMSE) である。

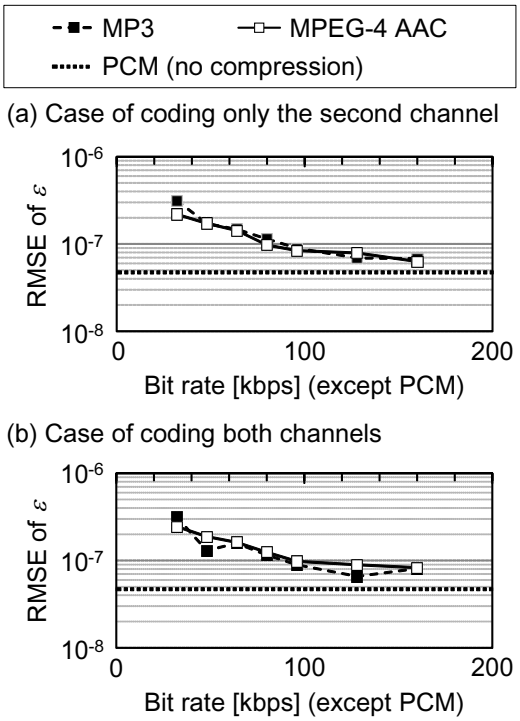

Fig. 1 Comparison of estimation accuracy for different coding conditions.

実験の結果を図 1 に示す。MP3 と MPEG-4 AAC の結果，および符号化するチャネルが片側チャネルの場合と両チャネルの場合に大きな違いはなく，全体として RMSE はビットレートに対してほぼ単調減少の傾向にある。高いビットレートでは精度の低下はわずかであり，非常に低いビットレートでも推定の破綻が起こることはない。また最も低いビットレートの 32 kbps でも PCM の 10 倍以下の誤差であり，これはアレー信号処理の前処理として補償を行うのに十分な推定精度である [2]。以上より，サンプリング周波数ミスマッチの最尤推定が符号化歪に頑健であることが確認された。

4 おわりに

本稿では，非同期マイクロホンアレーのためのサンプリング周波数ミスマッチ推定の，符号化歪に対する頑健性を評価した。MP3 と MPEG-4 AAC による符号化を評価した実験により，推定精度の低下は小さく符号化歪に頑健であることが確認された。

参考文献

[1] Liu, Proc. IWAENC, 2008.
[2] Miyabe *et al.*, Proc. ICASSP, 674-678, 2013.
[3] Miyabe *et al.*, Proc. WASPAA, 2013.
[4] Kurematsu *et al.*, Speech Communication, 9(4), 357-363, 1990.