

調波・非調波音源モデルを用いたマイク数以上の音源分離

平澤恭治, 安良岡直希, 高橋徹, 尾形哲也, 奥乃博 (京都大学大学院 情報学研究科)

1. はじめに

我々の生活環境のように多数の音源が存在する状況でも、人間の指示を正確に聞きとれるロボット聴覚機能を実現するために、我々は音源分離技術の研究を行っている。また、実環境にはエアコンなどの定常音・ロボット自体の動作音・他話者の声などといった多数の音源が存在していることから、マイク数より多くの音源が存在する劣決定状況での音源分離が本質的に重要である。このような劣決定音源分離技術を実現することで、パーティー会場のようにマイク数よりはるかに多くの音源が存在する環境でも動作可能な、環境内の音源数に非依存のシステムが構築可能と期待される。

我々が開発中の劣決定同時発話分離システム (図 1) における主たる目標は、以下の 3 点である。

1. 他話者からの漏れノイズが少ない
2. 基本周波数の連続的な変化に対応可能
3. 残響に強い

ここで 1 つ目の要求について補足する。音源分離を行う以上、元の音に近い分離音を出力することが目的であるが、目的話者以外の音声 (漏れノイズ) を取り除くことと、目的話者の音声を取り除きすぎないことはトレードオフの関係にある。これに対して我々は、多少目的音声の一部成分が欠落しようとも、できるだけ他話者の音声の漏れが少ない分離音を出力すべきという立場をとる。これは、音成分の一部欠落はポストフィルターなどで補償することが可能だが、他話者の漏れノイズはフィルターで除去するのが難しいからである。

劣決定音源分離の従来研究は、いくつかのグループに分類できる。1 つ目は混合の過程を既知とした手法で、L1-norm 最小化法 [1] などがある。2 つ目は、混合の過程を明示的に必要としない手法で、Sawada らのクラスタリング・時間周波数マスク法 [2] などがある。3 つ目は、分離結果と混合過程を共に推定する手法であり、Ozerov らによる音源を NMF でモデル化する手法 [3] などがある。この NMF で音源をモデル化する方法を以下では NMF-SM (Source Model) と呼ぶ。

本稿では、各時刻の音源のスペクトルを GMM を用いてモデル化する手法を提案する。これを以下 GMM-SM と呼ぶ (図 2)。具体的には、調波部分をモデル化する等間隔の鋭い GMM と、非調波部分をモデル化する鈍い GMM を使用することで、調波音だけでなく無声子音のような非調波音も合わせてモデル化する。GMM-SM を用いることで NMF-SM に比べて基本周波数の変動を正確にモデル化でき、かつ、モデルの形状を限定することで他話者からの漏れノイズの削減を実現する。以下、補助関数法を用いて解析的にパラメータの更新則を導出し、最後に実験により提案手法が漏れノイズの少ない音源分離を実現することを示す。

2. 劣決定音源分離

2.1 問題設定

まず、本稿で用いる主な変数の定義を表 1 に示す。なお、本稿では時間周波数領域にて処理を行うため、多くの変数は複素数となる。これらを用いて本研究で扱う

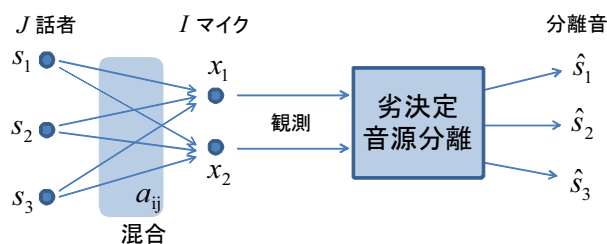


図 1 劣決定同時発話分離システム

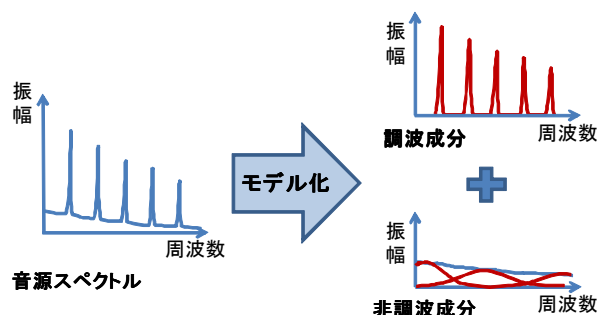


図 2 2種類のガウシアンによる音源モデル

問題は以下のように表される。

入力	I マイクで観測した J 音源の混合音 $x_{i,fn}$
出力	J 音源の分離音 $\hat{s}_{j,fn}$
仮定	劣決定状況 ($J \geq I$)、混合は線形時不変 残響はフレーム長に対して充分短い

2.2 音源モデルとコスト関数

本稿では各時刻での音源のスペクトルを、等間隔の鋭いガウシアンからなる調波部分と、鈍いガウシアンからなる非調波部分に分けてモデル化する。これは図 2 のように表され、等間隔の鋭いガウシアンは McAulay らによる正弦波重畳モデル [4] に相当している。なお、図 2 ではスペクトルの振幅を示しているが、実際の計算には複素スペクトルを用いており、位相が考慮される点に注意しておく必要がある。

具体的には、音源モデルは以下の式で表される。

$$\hat{s}_{j,fn} = \sum_{m_h} p_{j,n,m_h}^H g_{j,fn,m_h}^H + \sum_{m_n} p_{j,n,m_n}^N g_{f,m_n}^N \phi_{j,fn}^N \quad (1)$$

ここで、調波の各倍音のピークを表す p_{j,n,m_h}^H は複素数であるのに対し、非調波のピークを表す p_{j,n,m_n}^N は実数であり、別途各時間周波数毎に独立な位相 $\phi_{j,fn}^N$ を付与している。これは、調波を示す鋭いガウシアンは単一の正弦波と窓関数の積の周波数表現なので共通の位相を持つと想定されるのに対し、非調波を示す鈍いガウシアンはそのような規則性を持たないためである。なお、

表1 変数の定義

Indices	
i, I	マイク番号, マイク数
j, J	音源番号, 音源数
f, F	周波数ビン番号, 周波数ビン数
n, N	フレーム番号, フレーム数
m_h, M_H	調波倍音番号, 調波倍音数
m_n, M_N	非調波倍音番号, 非調波倍音数
t, T	調波/非調波を示す記号 ($T = \{H, N\}, t \in T$)
Signals	
$s_{j,fn}$	真の音源の音 ($\in \mathbb{C}$)
$\hat{s}_{j,fn}$	推定された音源の音 ($\in \mathbb{C}$)
$x_{i,fn}$	観測音 ($\in \mathbb{C}$)
$\hat{x}_{i,fn}$	パラメータから推定された観測音 ($\in \mathbb{C}$)
Parameters	
$a_{ij,f}$	混合行列の要素 ($\in \mathbb{C}$)
$F_{0,j,n}^H$	基本周波数
p_{j,n,m_h}^H	調波ガウシアン of 複素振幅 ($\in \mathbb{C}$)
p_{j,n,m_n}^N	非調波ガウシアン of 振幅
$\phi_{j,fn}^N$	非調波成分の位相 ($\in \mathbb{C}$)
Others	
F_0^N	非調波ガウシアン of 中心 (定数)

調波用・非調波用のガウシアンは次の様に定義される。

$$g_{j,fn,m_h}^H = \exp\left(-\frac{(f - m_h F_{0,j,n}^H)^2}{2\sigma_H^2}\right) \quad (2)$$

$$g_{f,m_n}^N = \exp\left(-\frac{(f - m_n F_0^N)^2}{2\sigma_N^2}\right) \quad (3)$$

ここで、非調波用のガウシアンは音源 j や時刻 n に依存せず、常に同じ形状のものを使用する。

この音源モデルと、混合が線形時不変であるという仮定より、観測音の推定値を以下の様に計算できる。

$$\hat{x}_{i,fn} = \sum_j a_{ij,f} \hat{s}_{j,fn} \quad (4)$$

このままでは $a_{ij,f}$ と p_{j,n,m_h}^H の間、並びに $a_{ij,f}$ と p_{j,n,m_n}^N と $\phi_{j,fn}^N$ の間にスケーリングの任意性が存在するので、 $a_{ij,f}$ と $\phi_{j,fn}^N$ の大きさを以下の様に制限する。

$$\sum_i |a_{ij,f}| = 1, \quad |\phi_{j,fn}^N| = 1 \quad (5)$$

次に、コスト関数をパラメータから計算される観測音の推定値と、実際の観測値との二乗誤差で定義する。

$$C = \sum_{ifn} |x_{i,fn} - \hat{x}_{i,fn}|^2 \quad (6)$$

$\hat{x}_{i,fn}$ の中にはガウシアンを含む複数の項の和が含まれており、単純な偏微分によるパラメータ更新式の導出は困難であるため、次節の補助関数法を利用する。

2.3 補助関数法

本稿では補助関数法 [5] を用いて、パラメータの更新式を解析的に導出する。補助関数法のアイデアは、元のコスト関数 $C(\theta)$ と下限が一致する補助関数 $C^+(\theta, \psi)$ を導入し、その上で更新式を求める、というものである。

具体的には、以下の性質を満たす関数をコスト関数 $C(\theta)$ の補助関数という。

$$1. C(\theta) = \min_{\psi} C^+(\theta, \psi)$$

ここで、 ψ は補助変数と呼ばれる。また、補助関数法で利用するためには以下の性質を満たす必要がある。

$$2. \psi_{new} = \operatorname{argmin}_{\psi} C^+(\theta, \psi) \text{ が解析的に求まる}$$

$$3. \theta_{new} = \operatorname{argmin}_{\theta} C^+(\theta, \psi) \text{ が解析的に求まる}$$

これらの性質を用いて、

- 性質 2 を用いて ψ を更新する
- 性質 3 を用いて θ を更新する

というように変数を更新すると、元のコスト関数の値が広義単調減少することが示される [6]。

2.4 更新式の導出

式 (6) に対し補助関数法を用いて、各パラメータの更新式を導出する。補助関数は Kameoka ら [6] により提案されたものをベースとし、本稿では観測が多チャンネルで、非調波部分もモデル化したものを扱う。使用する補助関数の詳細については上記の論文を参照されたい。

まず式 (6) に式 (1) と式 (4) を代入して展開すると、

$$C = \sum_{ifn} \left| x_{i,fn} - \sum_{jTm_t} a_{ij,f} p_{j,n,m_t}^T g_{j,fn,m_t}^T \phi_{j,fn}^T \right|^2 \quad (7)$$

となる。ここで $T \in \{H, N\}$ は調波・非調波を選択する変数であり、これに応じて m_t が m_h か m_n の一方を指すものとする。また、式の簡単化のために導入した $\phi_{j,fn}^H, g_{j,fn,m_n}^N$ は $\phi_{j,fn}^H = 1, g_{j,fn,m_n}^N = g_{f,m_n}^N$ である。これに 1 つ目の補助関数を導入すると、

$$C^+ = \sum_{ijfnTm_t} \frac{\left| \bar{\alpha}_{ij,fn,m_t}^T - a_{ij,f} p_{j,n,m_t}^T g_{j,fn,m_t}^T \phi_{j,fn}^T \right|^2}{\beta_{ij,fn,m_t}^T} \quad (8)$$

とかける。ここで

$$\bar{\alpha}_{ij,fn,m_t}^T = \alpha_{ij,fn,m_t}^T x_{i,fn} \quad (\in \mathbb{C}) \quad (9)$$

であり、右辺に出てきた $\alpha_{ij,fn,m_t}^T (\in \mathbb{C})$ は補助変数である。 β_{ij,fn,m_t}^T はそのパラメータで、

$$\sum_{jTm_t} \beta_{ij,fn,m_t}^T = 1, \quad 0 < \beta_{ij,fn,m_t}^T \in \mathbb{R} \quad (10)$$

を満たす任意の値である。なおこの時

$$\alpha_{ij,fn,m_t}^T = \frac{1}{x_{i,fn}} \left\{ a_{ij,f} p_{j,n,m_t}^T g_{j,fn,m_t}^T \phi_{j,fn}^T + \beta_{ij,fn,m_t}^T (x_{i,fn} - \hat{x}_{i,fn}) \right\} \quad (11)$$

とすると、補助関数の性質 1 の等号が成立する。

2.4.1 混合行列・振幅の更新式の導出

式 (8) を偏微分し、各パラメータの更新式を導出する。実変数と複素変数に気をつけながら $\frac{\partial C^+}{\partial a_{ij,f}} = 0, \frac{\partial C^+}{\partial p_{j,n,m_h}^*} = 0, \frac{\partial C^+}{\partial p_{j,n,m_n}^N} = 0$ を解いて整理すると、以下の更新式が得られる。なお記号 * は複素共役を示す。

$$a_{ij,f} = \frac{\sum_{nTm_t} \frac{\bar{\alpha}_{ij,fn,m_t}^T p_{j,n,m_t}^{T*} g_{j,fn,m_t}^T \phi_{j,fn}^{T*}}{\beta_{ij,fn,m_t}^T}}{\sum_{nTm_t} \frac{|p_{j,n,m_t}^T|^2 g_{j,fn,m_t}^T}{\beta_{ij,fn,m_t}^T}} \quad (12)$$

$$p_{j,n,m_h}^H = \frac{\sum_{if} \frac{\bar{\alpha}_{ij,f,n,m_h}^H a_{ij,f}^* g_{j,f,n,m_h}^H}{\beta_{ij,f,n,m_h}^H}}{\sum_{if} \frac{|a_{ij,f}|^2 g_{j,f,n,m_h}^H}{\beta_{ij,f,n,m_h}^H}} \quad (13)$$

$$p_{j,n,m_n}^N = \frac{\sum_{if} \frac{\Re \left[\bar{\alpha}_{ij,f,n,m_n}^N a_{ij,f}^* g_{f,m_n}^N \phi_{j,f,n}^{N*} \right]}{\beta_{ij,f,n,m_n}^N}}{\sum_{if} \frac{|a_{ij,f}|^2 g_{f,m_n}^N}{\beta_{ij,f,n,m_n}^N}} \quad (14)$$

2.4.2 位相の更新式の導出

式(8)中の絶対値二乗を展開すると、式(5)の制約より $\phi_{j,f,n}^N$ が残るのは以下の項だけとなる。

$$-2\Re \left[\sum_{im_n} \frac{\bar{\alpha}_{ij,f,n,m_n}^N a_{ij,f}^* p_{j,n,m_n}^N g_{f,m_n}^N \phi_{j,f,n}^{N*}}{\beta_{ij,f,n,m_n}^N} \right] \quad (15)$$

これを最大化するには $\phi_{j,f,n}^N$ の位相を調整して、 $\Re[\dots]$ 内部を非負実数にしてやれば良い。 $|\phi_{j,f,n}^{N*}| = 1$ から次のような更新式が導ける。

$$\phi_{j,f,n}^N = \frac{\sum_{im_n} \frac{\bar{\alpha}_{ij,f,n,m_n}^N a_{ij,f}^* p_{j,n,m_n}^N g_{f,m_n}^N}{\beta_{ij,f,n,m_n}^N}}{\left| \sum_{im_n} \frac{\bar{\alpha}_{ij,f,n,m_n}^N a_{ij,f}^* p_{j,n,m_n}^N g_{f,m_n}^N}{\beta_{ij,f,n,m_n}^N} \right|} \quad (16)$$

2.4.3 基本周波数の更新式の導出

同様に式(8)中の絶対値二乗を展開すると、 $F_{0,j,n}^H$ に依存する g_{j,f,n,m_h}^H が残るのは以下の2項となる。

$$\sum_{ifm_h} \frac{|a_{ij,f} p_{j,n,m_h}^H|^2}{\beta_{ij,f,n,m_h}^H} g_{j,f,n,m_h}^H \quad (17)$$

$$-2 \sum_{ifm_h} \frac{\Re \left[\bar{\alpha}_{ij,f,n,m_h}^H a_{ij,f}^* p_{j,n,m_h}^{H*} \right]}{\beta_{ij,f,n,m_h}^H} g_{j,f,n,m_h}^H \quad (18)$$

ここでガウシアン g_{j,f,n,m_h}^H の局所性より、式(17)中の本来 f に依存する $a_{ij,f}$ を定数と見なすことができる。このとき、 β_{ij,f,n,m_h}^H を f に依存しないように定めておくと、 $\bar{\alpha}_{ij,f,n,m_h}^H$ も f に依存しないようになるため、

$$\sum_f g_{j,f,n,m_h}^H \approx \sigma^H \sqrt{\pi} \quad (19)$$

というガウス積分により式(17)全体が定数となる。

これより $F_{0,j,n}^H$ の更新の際には式(18)のみを考慮すればよく、以降ではこの式(18)を C_F とおく。この時不等式

$$-e^{-x} \leq e^{-\gamma} (x - \gamma - 1) \quad (20)$$

の右辺を左辺に対する補助関数と見ると、次の補助関数が定義できる。

$$C_F^+ = 2 \sum_{ifm_h} \frac{\Re \left[\bar{\alpha}_{ij,f,n,m_h}^H a_{ij,f}^* p_{j,n,m_h}^{H*} \right]}{\beta_{ij,f,n,m_h}^H} \times e^{-\gamma_{ij,f,n,m_h}^H} \left(\frac{(f - m_h F_{0,j,n}^H)^2}{2\sigma^{H^2}} - \gamma_{ij,f,n,m_h}^H - 1 \right) \quad (21)$$

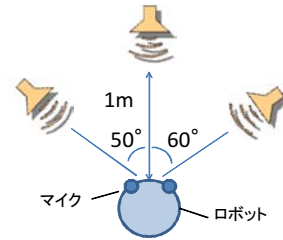


図3 話者とマイクの配置

この時 $\gamma_{ij,f,n,m_h}^H (\in \mathbb{R})$ は補助変数で、

$$\gamma_{ij,f,n,m_h}^H = \frac{(f - m_h F_{0,j,n}^H)^2}{2\sigma^{H^2}} \quad (22)$$

とすることで補助関数の性質1の等号が成立する。最後にこれを偏微分して $\frac{\partial C_F^+}{\partial F_{0,j,n}^H} = 0$ を解くと、以下の更新式が得られる。

$$F_{0,j,n}^H = \frac{\sum_{ifm_h} \frac{\Re \left[\bar{\alpha}_{ij,f,n,m_h}^H a_{ij,f}^* p_{j,n,m_h}^{H*} \right]}{\beta_{ij,f,n,m_h}^H} e^{-\gamma_{ij,f,n,m_h}^H} f m_h}{\sum_{ifm_h} \frac{\Re \left[\bar{\alpha}_{ij,f,n,m_h}^H a_{ij,f}^* p_{j,n,m_h}^{H*} \right]}{\beta_{ij,f,n,m_h}^H} e^{-\gamma_{ij,f,n,m_h}^H} m_h^2} \quad (23)$$

2.5 パラメータの更新

以上の更新式を用いて、例えば次のような順序でのパラメータ更新が可能となる。

1. 式(11)を用いて α_{ij,f,n,m_t}^T を更新
2. 式(13)を用いて p_{j,n,m_h}^H を更新
3. 式(14)を用いて p_{j,n,m_n}^N を更新
4. 式(16)を用いて $\phi_{j,f,n}^N$ を更新
5. 式(12)を用いて $a_{ij,f}$ を更新
6. 式(22)を用いて γ_{ij,f,n,m_h}^H を更新
7. 式(23)を用いて $F_{0,j,n}^H$ を更新

2.3節で述べた通り、パラメータを更新する前に、その計算に使用する補助変数を更新しておく必要がある。一方、その順序さえ守ればパラメータ間の更新順序は任意である。実験的には、 p_{j,n,m_h}^H と p_{j,n,m_n}^N は他のパラメータの変更の影響を受けやすいので、他より頻繁に更新することが望ましい。

3. 実験

3.1 実験環境

提案手法の性能を確認するために、2マイクを用いて図3に示す配置で3話者の同時発話を観測する状況をシミュレートし、その分離結果を確認した。合成にはロボット頭部のマイクで観測した無響室のインパルス応答を使用し、音源にはASJ/JNASデータベースから無作為に抽出した男女の音声を利用した。この時、サンプリング周波数は16kHzで、短時間フーリエ変換のフレーム長は1024点(64ms)、シフト幅は256点(16ms)とした。評価尺度として全体の分離性能を示すSignal to Distortion Ratio (SDR)と、個々の分離性能を示すImage to Spatial distortion Ration (ISR), Source to Interference Ratio (SIR), Source to Artifacts Ratio (SAR)の計4尺度[7]を用いた。1章で述べた通り、我々は他話者からの漏れノイズの少ない音源分離の実

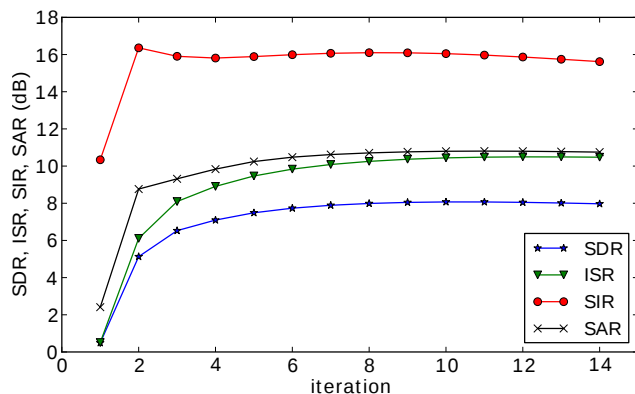


図4 イテレーションごとの3話者の平均評価値の変化

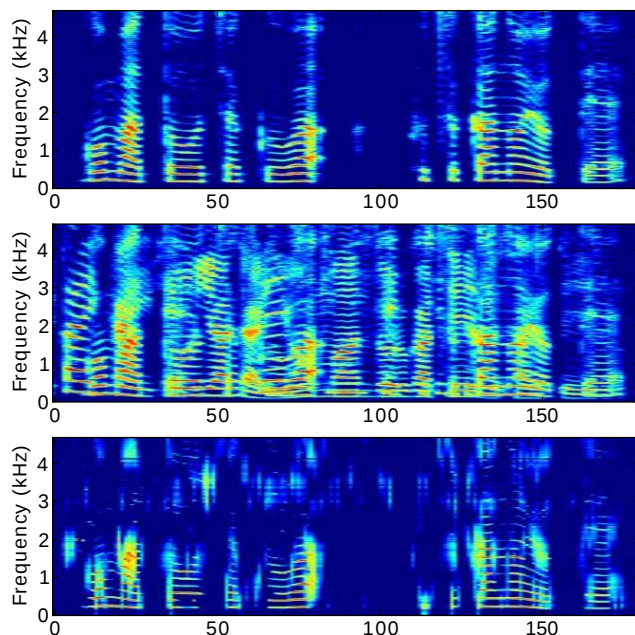


図5 (上) 元音源 (中) 混合音 (下) 分離結果

現を目指しているの、主に他話者からの漏れノイズの少なさを示す指標である SIR と、全体の分離性能を示す指標である SDR に着目すると良い。

3.2 実験結果

まず初めに、各話者の基本周波数を既知とし、基本周波数を正しく初期化した際の分離性能の確認を行った。混合行列 $\{a_{ij,f}\}$ の初期値は話者方向から大まかに推定した値とし、それ以外のパラメータは0や1などの一定値で初期化した。イテレーション毎の SDR, ISR, SIR, SAR の変化を図4に、音源、混合音、分離結果の一例を図5に示す。図4からは、各尺度の値が8イテレーション程度で収束している様子が分かる。また、図5からは、低周波部分を中心によく分離できているが、高周波部分には誤推定が生じやすくなっていることも分かる。これは、高周波部分にはパワーの強い調波構造が少なく、混合行列 $\{a_{ij,f}\}$ の推定が比較的困難になっているためだと考えられる。

また、各話者の基本周波数の真値を与えず、以下の方法で推定した場合の分離性能の確認も行った。この時基本周波数の初期値は各フレーム、各話者ごとに

1. 基本周波数の候補を 10Hz ごとに複数用意
2. 式 (11)(13)(22)(23) を用いてパラメータを反復更新
3. 最も誤差の少ない状態に収束した候補の、最終的

表2 最終的な分離性能 (dB)

	基本周波数	SDR	ISR	SIR	SAR
NMF-SM	-	6.4	12.1	9.9	11.4
HSS-SM	既知	8.0	10.4	16.1	10.7
HSS-SM	未知	6.8	9.3	14.3	8.8

な基本周波数の値を採用

4. 基本周波数のない部分や、明らかな誤収束を除去という手順を踏むことで推定した。なおこれは naive な基本的手法であり、改善の余地は大きい。

以上の実験を行い、基本周波数が既知の場合と未知の場合、ならびに音源モデルとして GMM-SM ではなく NMF-SM を用いた Ozerov らの従来手法 [3] について、最終的な分離結果を表2にまとめた。表2より、ISR と SAR の尺度では NMF-SM が GMM-SM に勝る一方、漏れノイズの小ささを示す SIR と全体の分離性能を示す SDR の尺度では GMM-SM が NMF-SM に勝っていることが確認できる。結論として、本稿が提案する GMM-SM による音源分離が、漏れノイズの少ない分離音の出力を実現することができたと考えられる。

4. おわりに

本稿では、音源数がマイク数以上である劣決定状況における、同時発話を正しく分離するために、調波用 GMM と非調波用 GMM からなる音源モデルを提案し、そのパラメータ更新式を導出した。実験により、従来提案されていた NMF による音源モデルと比較して分離結果の漏れノイズが少なく、総合的な指標でも高性能な音源分離が実現できたことを確認した。今後は基本周波数のより良い初期化手法や、削りすぎた音声成分を回復するポストフィルター、今回対応できなかったフレームをまたぐ残響への対処などについて検討していきたいと考えている。

謝辞 本研究の一部は科研費基盤 (S)、JST-ANR BINAAHR, GCOE の支援を受けた。

参考文献

- [1] P. Bofill *et al.*, Underdetermined blind source separation using sparse representations. *Signal processing*, 81(11), pp. 2353–2362, 2001.
- [2] H. Sawada *et al.*, Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Trans. on ASLP*, 19(3), pp. 516–527, 2011.
- [3] A. Ozerov *et al.*, Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Trans. on ASLP*, 18(3), pp. 550–563, 2010.
- [4] R.J. McAulay *et al.*, Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. on ASSP*, 34(4), pp. 744–754, 1986.
- [5] D.D. Lee *et al.*, Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, Vol. 13, pp. 556–562, 2001.
- [6] H. Kameoka *et al.*, Auxiliary function approach to parameter estimation of constrained sinusoidal model for monaural speech separation. In *Proc. of ICASSP 2008*, pp. 29–32, 2008.
- [7] E. Vincent *et al.*, First stereo audio source separation evaluation campaign: data, algorithms and results. *Independent Component Analysis and Signal Separation*, pp. 552–559, 2007.