スパースな混合モデルに基づく雑音・残響環境下の劣決定BSS

Underdetermined BSS in Noisy/Echoic Environment Based on Sparse Mixture Model

和泉洋介 † 小野順貴 † 嵯峨山茂樹 †
Yosuke Izumi Nobutaka Ono Shigeki Sagayama

東京大学大学院 情報理工学系研究科 †
Graduate School of Information Science and Technology, The Univ. of Tokyo

1 はじめに

複数の音声信号が混合した観測信号だけから個々の信号を分離するブラインド音源分離 (BSS) は、混合過程の情報を必要としないため実環境に幅広く適用できる有用な技術として活発な研究が行われている [1]。特に音源数がマイクロフォン数より大きい劣決定の場合には音源信号のスパース性を積極的に用いる時間周波数マスキングが有効な手法である [2, 3]。しかし従来提案されてきた時間周波数マスクの設計法では、クリーンな環境では良い分離性能を示すものの、雑音・残響が存在する環境下では著しく性能が落ちることが課題となっていた。本稿では特に観測信号が 2ch の場合に議論を絞り、雑音・残響のモデルに基づき、尤度を基準にした最適な分離マスクを EM アルゴリズムにより設計する手法を提案し、実験結果と共に報告する。

2 音声のスパース性に基づく劣決定 BSS

音声信号の有意なエネルギーは時間周波数空間の疎な領域に分布しているため、同時に K 個の音声を観測しても各時間周波数成分において個々の音声の成分が重なることはほとんど起こらない。そこで観測信号 $M_L(\tau,\omega), M_R(\tau,\omega)$ の各成分には $i(\tau,\omega)$ 番目の音声信号 $S_i(\tau,\omega)$ だけが寄与すると仮定すると混合モデルは、

$$\begin{pmatrix} M_L(\tau,\omega) \\ M_R(\tau,\omega) \end{pmatrix} = \begin{pmatrix} 1 \\ e^{j\omega\delta_i} \end{pmatrix} S_i(\tau,\omega) + \begin{pmatrix} N_L(\tau,\omega) \\ N_R(\tau,\omega) \end{pmatrix}$$

と表せる。ここで δ_i は S_i が平面波伝播で到来したときに生じる時間差、N_L, N_R は残響を含む雑音を表す。以後は表記の簡単のため、左辺を $\boldsymbol{M}$、右辺第 1 項のベクトルを $\boldsymbol{b}_i$、雑音項を $\boldsymbol{N}$ とし、(τ,ω) を省略して表記する。

このモデルのもと、k 番目の音源の分離信号 $\hat{S}_k$ を得るには、各時間周波数成分に寄与している音源のインデックス $i(\tau,\omega)$ が k と等しい成分だけを抜き出すようなマスキング処理を施せばよい。雑音が十分小さい場合には図 2 左に示すように、M_L, M_R の比から推定された時間差 δ_i は真の時間差を中心に分布し、δ の空間でクラスタリングすることで適切な時間周波数マスクが設計できる。しかし、雑音・残響環境下で雑音パワーが大きくなると時間差 δ_i は図 2 右のように大きなばらつきを生じ、クラスタリングが困難になる問題があった。

3 雑音・残響環境下の観測モデル

従来法における分離マスク設計は、各時間周波数成分で複数の音声が同時に active になることはないというスパース性の仮定に基づいている。このとき実際の観測信号には、いずれの音源も non-active で、雑音のみの成分も含まれるはずである。雑音パワーが十分小さい環境では、これらをクラスタリングの過程でいずれかの音源に帰属させても、音源定位や分離信号に与える影響は小さい。しかし雑音パワーが大きい環境では、雑音のみの成分を強引にいずれかの音源に帰属させることは、クラスタリングの精度低下を招くと考えられる。我々はこの問題を回避するために、こうした雑音のみの成分を集めるための雑音クラスを導入する。このとき観測モデルは、

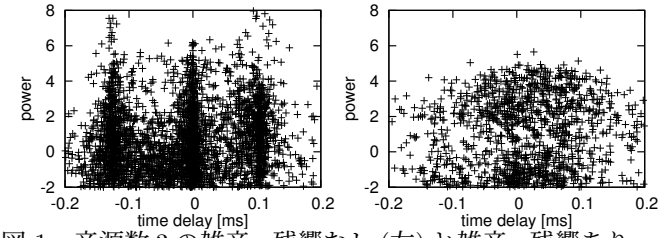

図 1 音源数 3 の雑音・残響なし (左) と雑音・残響あり (右) の場合の δ のプロット図

$$\boldsymbol{M} = \begin{cases} \boldsymbol{b}_i S_i + \boldsymbol{N} & (i \text{ 番目の音源のみが active}) \\ \boldsymbol{N} & (\text{すべての音源が non-active}) \end{cases}$$

と表される。ただし、$\boldsymbol{N}$ として拡散性雑音を仮定し、

$$\begin{aligned} \boldsymbol{N} &\sim \mathcal{N}(0, V(\omega)) \\ V(\omega) &= \sigma^2(\omega) \begin{pmatrix} 1 & \mathrm{sinc}\,(\omega D/c) \\ \mathrm{sinc}\,(\omega D/c) & 1 \end{pmatrix} \end{aligned}$$

とする。ここで D はマイクロフォン間距離、c は音速を表す。

4 雑音・残響環境下の 2ch BSS への EM アルゴリズムの適用

前節の観測モデルに基づくと $\boldsymbol{M}$ が $i(\tau,\omega)$ 番目の音源から到来するときの尤度は

$$\begin{aligned} &p\left(\boldsymbol{M} \middle|\, \sigma(\omega), \delta_i, S_i\right) \\ &= \frac{1}{2\pi\sqrt{|V|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{M} - \boldsymbol{b}_i S_i\right)^h V^{-1} \left(\boldsymbol{M} - \boldsymbol{b}_i S_i\right)\right) \end{aligned}$$

と与えられる。雑音クラスに属する成分の場合は、$i = K+1$ 番目のクラスタに属する尤度として $S_i = 0$ とすれば同様に与えられる。ここで $i(\tau,\omega)$ が観測できない隠れ変数であることに注意すると、目的関数となる時間周波数平面全体の対数尤度は、i について周辺化し、

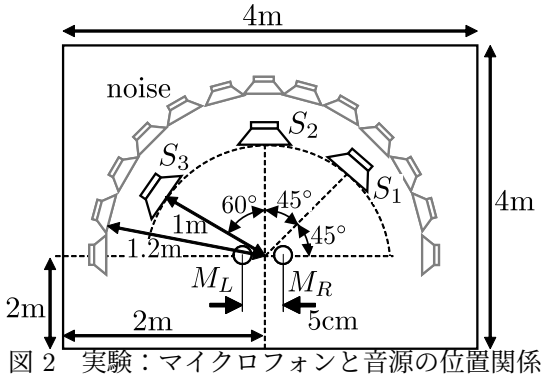

図 2　実験：マイクロフォンと音源の位置関係

$$LL = \sum_{\tau,\omega} \log \sum_{i=0}^{K+1} r_i p\left(\boldsymbol{M}|\ \sigma(\omega), \delta_i, S_i\right)$$

と表される。ここで r_i は i 番目の音源が active になる事前確率に相当する変数で $\sum_k r_k = 1$ を満たす。

我々のこのクラスタリングの問題は、混合ガウス分布モデル (GMM) のフィッティング問題と同型であり、GMM のパラメータ推定に広く用いられる EM アルゴリズムを適用することで効率的に未知パラメータ $\boldsymbol{\theta} = \{\sigma(\omega), \delta_i, S_i\}$ の最尤値を導出できる [5]。各パラメータの更新式は、

$$\begin{aligned}
m_{\tau,\omega,i} &\leftarrow \frac{r_i p\left(\boldsymbol{M}|\ \sigma(\omega), \delta_i, S_i\right)}{\sum_{i'} r_{i'} p\left(\boldsymbol{M}|\ \sigma(\omega), \delta_{i'} S_{i'}\right)} \\
Q(\boldsymbol{\theta}) &\leftarrow \sum_{\tau,\omega,i} m_{\tau,\omega,i} \log r_i\ p(\boldsymbol{M}|\ \sigma(\omega), \delta_i, S_i) \\
S_i &\leftarrow \left(\boldsymbol{b}_i^h V^{-1} \boldsymbol{M}\right) \Big/ \left(\boldsymbol{b}_i^h V^{-1} \boldsymbol{b}_i\right) \\
r_i &\leftarrow \sum_{\tau,\omega} m_{\tau,\omega,i} \Big/ \sum_{\tau,\omega,i'} m_{\tau,\omega,i'} \\
\sigma^2(\omega) &\leftarrow \frac{1}{2C} \sum_{\tau,i} \frac{m_{\tau,\omega,i} \left(\left(\boldsymbol{M} - \boldsymbol{b}_i S_i\right) V^{-1} \left(\boldsymbol{M} - \boldsymbol{b}_i S_i\right)\right)}{1 - \mathrm{sinc}^2\left(\omega D/c\right)} \\
\delta_i &\leftarrow \underset{\delta_i}{\mathrm{argmax}}\, Q(\boldsymbol{\theta}; \boldsymbol{\theta})
\end{aligned}$$

と表される。ただし $m_{\tau,\omega,i}$ は (τ, ω) 成分が i 番目の音源に帰属する期待値を、C は全フレーム数を表す。また、δ_i の更新は解析解が求まらないので適当に離散化し全探索することで更新した。各パラメータの推定値に基づき、分離信号は

$$E\left[S_i\right] = m_{\tau,\omega,i} \frac{\boldsymbol{b}_i^h V^{-1} \boldsymbol{M}}{\boldsymbol{b}_i^h V^{-1} \boldsymbol{b}_i}$$

としてソフトマスキングによる期待値として求められる。

5　定位・分離実験

提案法による音源分離実験を，図 5 のように 3 つの音源および 2 つのマイクロフォンとこれらを取り囲む雑音源を配置し，鏡像法 [4] による残響シミュレーションによって行った。雑音源には 500Hz と 2000Hz を中心周波数とするバンドノイズを使用し、拡散音場を模するためすべて同じパワーとした。分離性能の評価には分離の前後での元音声に対する S/N 比の改善値を用い、音声データは研究用連続音声データベース (ⓒ板橋秀一 [日本音響学会／編]1991Vol. 1-3) を使用した．サンプリング周期 16kHz，フレーム長は 2^{10} 点, シフトは 2^9 点，窓関数を Hamming 窓として，観測信号を短時間 Fourier 変換して時間周波数表現を得た。残響時間 170ms の場合の音源分離・定位結果を表 1,2 に示す。比較対象とした従来法は Yilmaz らの手法 [2] に基づく。また、雑音クラスを作らずに EM アルゴリズムを適用した分離マスクの結果も並記した。これらの結果から，従来法に比べて提案法は雑音・残響環境下でも音源定位の精度が良く、雑音クラスを追加した効果が認められる。また、σ の推定値は図 3 に示すように、付加したバンドノイズを推定できていることが確認できた。

表 1　音源定位結果の比較 (時間差 [μs])

手法	S_1	S_2	S_3
従来手法	−47	−40	−11
雑音クラス無し	−147	−103	134
提案手法	−106	−8	132
真の位置	−104	0	127

表 2　音源分離性能結果の比較 ([dB])

手法	S_1	S_2	S_3
従来手法	2.2	1.2	3.3
雑音クラス無し	6.2	−1.5	7.4
提案手法	6.7	3.1	7.8

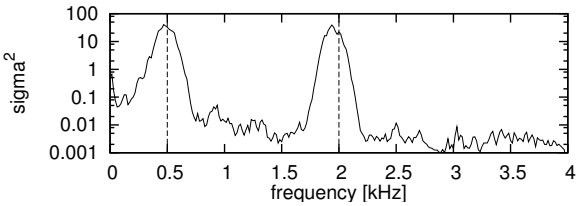

図 3　雑音分散 $\sigma(\omega)$ の推定値

6　結論

本稿ではスパースな混合モデルに基づく雑音・残響環境下の 2ch BSS の問題に対し雑音成分を雑音クラスタに振り分けることで、より頑健に分離マスクを設計する手法を提案した。シミュレーション実験において、雑音と残響を同時に付与し分離が困難な状況においても音源定位が精度よくできており、分離性能も従来法より優れていることを確認した。

参考文献

[1] J-F Cardoso, Proc. of the IEEE, Vol. 90, No. 8, pp. 2009-2026, 1998

[2] O. Yilmaz *et al.*, IEEE Trans. on Signal Processing, Vol. 52, No. 7, pp 1830-1847, 2004.

[3] H. Sawada *et al.*, IEEE Trans. Audio, Speech and Language Processing, vol. 14, no. 6, pp. 2165-2173, 2006.

[4] J. B. Allen et al., JASA, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[5] 小野他, 電子情報通信学会技術研究報告 (応用音響), Vol. 107, No. 240, pp.25-30, 2007.