

独立成分分析を用いた音源数推定法*

○澤田 宏, 向井 良, 荒木 章子, 牧野 昭二

(日本電信電話株式会社, NTT コミュニケーション科学基礎研究所)

1 はじめに

空間内に存在する音源の個数を推定することは、ブラインド音源分離などの応用に有益な情報を与える。従来の推定法としては、空間相関行列の支配的な固有値の数を音源数として推定する方法がある[1]。しかし残響が無視できない実環境では、どこまでが支配的な固有値かを判別することが難しい。この残響の問題に対処するために、近年、それら固有値の分布をSVMなどの識別器に学習させる方法が提案された[2]。

本稿では、まず、固有値による方法を実環境に適用した際の問題点を指摘する。次に、固有値分析の代わりに独立成分分析を用いるという異なるアプローチによりそれらの問題に対処する。最後に実験結果により、既存の方法と比較して提案法がどのような特長を持つかを示す。

2 畳み込み混合と周波数領域での近似

N 個の音源 $s_k(t)$ が実環境で混合され、 M 個のマイクでそれぞれ観測信号

$$x_j(t) = \sum_{k=1}^N \sum_{l=0}^{P-1} h_{jk}(l) s_k(t-l) + n_j(t) \quad (1)$$

が得られたとする。ここで、 $h_{jk}(l)$ は音源 k からマイク j へのインパルス応答、 P はインパルス応答の持続時間、 $n_j(t)$ は各マイクでのノイズを表す。本稿では、このような畳み込み混合に対して、観測信号 $x_1(t), \dots, x_M(t)$ のみから音源の数 N を推定することを目的とする。ただし、マイクの数 M は音源の数と同等かそれ以上、すなわち $N \leq M$ を仮定する。

畳み込み混合の結果である観測信号 $x_j(t)$ に対しては、 L 点の短時間離散フーリエ変換 (STFT) を適用して周波数 f 毎の時系列 $X_j(f, \tau)$ を求めることが有効である。式 (1) で示される時間領域での畳み込み混合が、

$$X_j(f, \tau) = \sum_{k=1}^N H_{jk}(f) S_k(f, \tau) + N_j(f, \tau) \quad (2)$$

と各周波数での単純混合にある程度近似できるからである。ここで、 $H_{jk}(f)$ は音源 s_k からマイク x_j までの周波数応答、 $S_k(f, \tau)$ と $N_j(f, \tau)$ はそれぞれ音源 $s_k(t)$ およびノイズ $n_j(t)$ に STFT を施したものである。

3 固有値による音源数推定とその問題点

式 (2) の近似が十分に満たされているとして、空間相関行列 $\mathbf{R}(f) = \langle \mathbf{X}(f, \tau) \mathbf{X}(f, \tau)^H \rangle_\tau$, $\mathbf{X}(f, \tau) = [X_1(f, \tau), \dots, X_M(f, \tau)]^T$ の固有値により音源数を推

*Estimating the Number of Sources using Independent Component Analysis, by H. Sawada, R. Mukai, S. Araki, S. Makino (NTT Communication Science Labs., NTT Corporation)

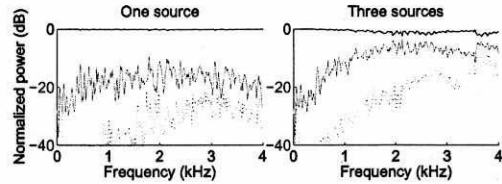


図 1: 各周波数での正規化された固有値

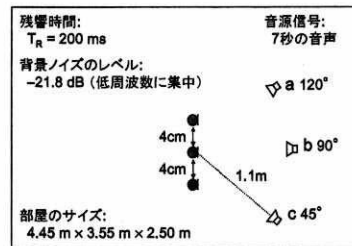


図 2: 部屋の特徴とスピーカ/マイクの配置

定する方法が広く知られている。 M 個の固有値のうち、支配的な値を持つものの個数が音源の数 N に対応し、残りの $M - N$ 個の小さな固有値の大きさがノイズ $N_j(f, \tau)$ のパワー σ^2 と等しくなる。

【問題点 1: 残響の問題】しかし、残響の影響が比較的に長い場合は、式 (2) の近似が不十分となり*、支配的な固有値の数が実際の音源数よりも多くなることがある。図 1 の左側は、図 2 に示す条件で 1 音源だけが存在する場合に、各周波数で算出した固有値を、総和が 1 になるように正規化したものである。残響の影響により 2 個目の固有値が -20dB 程度になっている。ノイズのパワーを -20dB より十分大きいと見なす場合は良いが、 -20dB より小さいと見なした場合は 2 個の音源と推定してしまう。

【問題点 2: パワー推定の問題】次に、各音源のパワーが正しく推定されないことを指摘する。図 1 の右側は、図 2 に示す条件で 3 音源すべてが存在する場合に、各周波数で算出した固有値である。各音源のパワーは同等に設定したにもかかわらず、固有値は、2 番目、3 番目と順に小さくなっていく。この傾向は、マイク間の位相差が小さい低周波数ほど顕著になる。多くの周波数で 3 音源が存在すると推定するためには、ノイズのパワーを大きくとも -30dB 以下と設定しなければならない。しかし、そのように設定すると、左図の 1 音源の場合に 2 音源以上と推定されてしまう。

*残響の影響 (インパルス応答の持続時間 P) をすべてカバーする程に STFT のフレーム長 L を長くして、式 (2) の近似を満たすことも可能であるが、フレーム長 L を過度に長くすることは、種々の弊害を伴う。

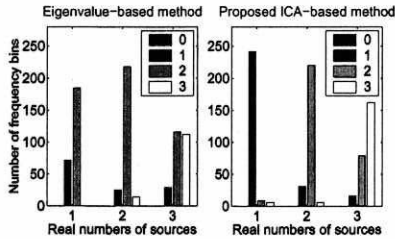


図 3: 音源数推定結果: 従来法 (左) と提案法 (右)

4 独立成分分析による音源数推定

提案法では、3 章で指摘した【問題点 2】をまずは解決することを意図して、独立成分分析 (ICA) [3] を用いる。以下の処理は周波数毎に行うが、簡単のため周波数を示す f は省略して表記する。観測信号ベクトル $\mathbf{X}(f)$ に ICA を適用すると、 $M \times M$ 行列の分離行列 $\mathbf{W}$ と分離信号ベクトル $\mathbf{Z}(f) = \mathbf{W}\mathbf{X}(f)$ が得られる。分離信号ベクトル $\mathbf{Z}(f)$ は M 個の要素を持つが、そのうち N 個は音源に対応し、残りの $M - N$ 個はノイズや残響成分に対応する。

しかしながら、ICA の解にはスケーリングの任意性があるため、 $\mathbf{Z}(f)$ の要素は音源やノイズのパワーを正しく反映していない可能性が高い。そこで、音源やノイズのパワーを回復するために、提案手法では、

$$\mathbf{W} \leftarrow \Lambda \mathbf{W}, \Lambda = \text{sqrt}(\text{diag}[(\mathbf{W}\mathbf{W}^H)^{-1}]) \quad (3)$$

によりスケーリングを決定する。ここで、 diag は対角成分以外を 0 にする操作、 sqrt は各要素の平方根を計算する操作である。証明は省略するが、式 (3) により混合行列 $\mathbf{W}^{-1}$ の列ベクトルのノルムが 1 となるため、スケーリングを決定した分離信号ベクトル $\mathbf{Y}(f) = \Lambda \mathbf{Z}(f)$ の要素 $Y_1(f), \dots, Y_M(f)$ は、

$$|Y_i(f)|^2 = \sum_{j=1}^M |H_{jk} S_k(f)|^2 \quad (4)$$

を満たすものとなる。すなわち、 $Y_i(f)$ のパワーと、それに対応する音源 $S_k(f)$ の全マイクでのパワーの総和が等しくなる。これにより問題点 2 が解決できる。

次に、3 章で指摘した【問題点 1】すなわち残響への対処を考える。例えば要素 $Y_i(f)$ が、ある音源 $Y_k(f)$ の残響を主たる成分とする場合、それらは類似する時間構造を持つはずである。その類似度は、平均が 0 になるように正規化したエンベロープ $v_i(f) = |Y_i(f)| - \langle |Y_i(f)| \rangle_f$ の適切な時間差 $\Delta\tau$ (例えば $L/2$ や $L/4$) による相関

$$\frac{\langle v_i(f) \cdot v_k(f - \Delta\tau) \rangle_f}{\sqrt{\langle v_i^2(f) \rangle_f} \cdot \sqrt{\langle v_k^2(f - \Delta\tau) \rangle_f}} \quad (5)$$

により計算できるため、この相関が大きければ残響成分として判別できる。

以上をまとめると、提案法は以下の手順となる。

1. ICA と式 (3) により、各 $i = 1, \dots, M$ に対して分離信号 $Y_i(f)$ を得て、その正規化パワー $NP_i = \langle |Y_i(f)|^2 \rangle_f / \sum_{k=1}^M \langle |Y_k(f)|^2 \rangle_f$ を計算する。
2. NP_i がある閾値 (例えば $0.01 = -20$ dB) より小さければ $Y_i(f)$ をノイズと見なす。
3. NP_i がある閾値 (例えば 0.2) より小さく、さらに式 (5) による相関の最大値がある閾値 (例えば 0.5) より大きければ $Y_i(f)$ を残響と見なす。
4. 上記以外ならば $Y_i(f)$ を音源としてカウントする。

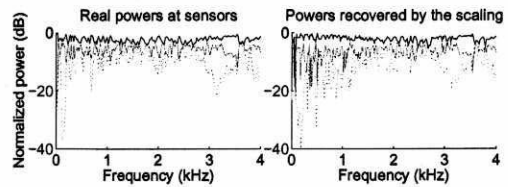


図 4: 3 音源の場合にマイクで観測された各音源のパワー (左) と提案法で推定した各音源のパワー (右)

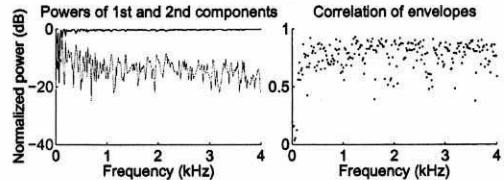


図 5: 1 音源の場合の 1 番目と 2 番目の分離信号のパワー (左) とそれらのエンベロープの相関 (右)

5 実験結果および考察

図 2 に示す条件で 1~3 個の音源を鳴らし、3 個のマイクでの観測信号を用いて音源数を推定した。固有値に基づく従来法と独立成分分析を用いた提案法の結果を図 3 に示す。横軸は真の音源数、縦軸は音源数がある値としてそれぞれ推定した周波数ビンの数を示す。従来法では、多くの周波数ビンで推定を誤っている。これは、残響の影響を考慮していないことや、図 1 に示したように個々の音源のパワーが正しく回復されていないことが原因である。一方、提案法によると、多くの周波数ビンで正しい推定値を得ている。

以下、提案法の特長を考察する。まず、各音源のパワーに関しては、図 4 を見ると、音源数を推定できる程度に正しくパワーが回復されていることがわかる。これは、ICA と式 (3) によるものである。次に、残響への対処に関して考察する。図 5 において、左側のパワーだけを見ると、2 番目の分離信号が音源なのか残響を含むノイズなのか判断し難い。しかし、右側に示す 1 番目と 2 番目の分離信号のエンベロープの相関を見ると、その値が十分に大きいため、2 番目の分離信号は 1 番目の音源の残響成分を多く含む成分であることがわかり、音源数へのカウントから排除できる。

6 まとめ

提案した音源数推定法は、独立成分分析およびパワーを回復するスケーリングによりマイクで観測された音源のパワーを正しく推定でき、さらにエンベロープの相関により残響成分を判別することを特長とする。

参考文献

- [1] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. 33, No. 2, pp. 387-392, April 1985.
- [2] 山本潔, W.F.G. van Rooijen, E.Y. Ling, 浅野太, 山田武志, 北脇信彦. SVM を用いた音源数推定法の音源分離システムへの応用. 音講論集, pp. 537-538, September 2002.
- [3] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. John Wiley & Sons, 2001.