2 値マスクと独立ベクトル分析を併用したセミブラインド音源分離

SEMI-BLIND SOURCE SEPARATION USING BINARY MASK AND INDEPENDENT VECTOR ANALYSIS

太刀岡勇気　成田知宏　石井純
Yuuki Tachioka　Tomohiro Narita　Jun Ishii

三菱電機株式会社 情報技術総合研究所
Information Technology R&D Center, Mitsubishi Electric Corporation

1　まえがき

複数話者による発話の同時認識は，単一システムの同時使用を可能にする．認識の前段の音源分離には，音源到来方向などの物理的な情報に基づく方法が一般的であるが誤差の影響を受けやすい．一方で，統計的な独立性を元に分離を行うブラインド音源分離法には，誤差を補正する効果がある．本報では，両者を組み合わせることで，音源分離の頑健性を向上させることを目的とする．

2　時間・周波数 2 値マスク (binM)

時間フレーム $t(1 \leq t \leq T)$，周波数ビン ω の時間差 $\tau(\omega,t)$ は $x_2(\omega,t)/x_1(\omega,t) = Ae^{\jmath\omega\tau(\omega,t)}$ で表される．ここで，$\jmath$ は虚数単位，$A(>0)$ は実数，x_1,x_2 はマイク 1,2 の観測信号の短時間フーリエ変換であり，まとめて $\boldsymbol{x}(\omega,t) = (x_1(\omega,t), x_2(\omega,t))^\top$ で表す ($\top$ は転置)．通常，到来方向 θ に対して，マスク $W(\omega,t) = (\boldsymbol{w}_1(\omega,t), \boldsymbol{w}_2(\omega,t))^h$ は (1) のように設定される (h は Hermite 転置)[1]．k はマイクの ID である．

$$\boldsymbol{w}_k(\omega,t) = \begin{cases} \epsilon\boldsymbol{e}_k & \text{if} |c/l_m \sin^{-1}\tau_{\omega,t} - \theta| > \theta_c, \\ \boldsymbol{e}_k & \text{if} |c/l_m \sin^{-1}\tau_{\omega,t} - \theta| \leq \theta_c, \end{cases} \quad (1)$$

$\boldsymbol{e}_k$ は単位ベクトル (k 番目の要素が 1) である．ϵ は小さい定数，θ_c は許容誤差，c は音速，l_m はマイク間隔である．$\boldsymbol{y}(\omega,t) = W(\omega,t)\boldsymbol{x}(\omega,t)$ より分離信号 $\boldsymbol{y}$ を得る．

3　補助関数に基づく独立ベクトル分析 (IVA)

音源の独立性に基づく手法は，一般的に周波数ビンごとに音声を分離するため (例えば独立成分分析)，分離話者間の混同が起こる．IVA では，周波数ビンをまたがる目的関数 (2) を最小化することでこの問題を回避し，時不変の分離行列 $W(\omega)$ の組 $\boldsymbol{W}$ を決定する．($r_{k,t}$ は (3) の L_2 ノルム，E は時間に関する期待値)

$$J(\boldsymbol{W}) = \sum_k E[r_{k,t}] - \sum_\omega \log|\det W(\omega)|. \quad (2)$$

補助関数で J の上限を抑えることで最適化を進める方法が提案されている [2]．補助変数の更新 ((3)) 後に分離行列の更新 ((4),(5)) を行う手順を繰り返す．

$$r_{k,t} = L_2\left[\boldsymbol{w}_k^h(\omega)\boldsymbol{x}(\omega,t)\right], V_k(\omega) = E\left[\frac{\boldsymbol{x}(\omega,t)\boldsymbol{x}^h(\omega,t)}{r_{k,t}}\right]. \quad (3)$$

$$\boldsymbol{w}_k(\omega) \leftarrow (W(\omega)V_k(\omega))^{-1}\boldsymbol{e}_k, \quad (4)$$

$$\boldsymbol{w}_k(\omega) \leftarrow \boldsymbol{w}_k(\omega)/\sqrt{\boldsymbol{w}_k^h(\omega)V_k(\omega)\boldsymbol{w}_k(\omega)}. \quad (5)$$

4　提案法

空間折り返し歪の生じない $f_c = c/(2l_m)$ 以下の周波数帯域は，binM により分離する．次に IVA で f_c 以上の帯域を分離するが，その際に全 ω に関して補助変数と分離行列を更新することで，分離話者の同一性を担保する．ただし f_c 以下は既に分離されているので，式 (4) の更新は不要で，単に $\boldsymbol{w}_k(\omega) = \boldsymbol{e}_k$ とする．

5　音声認識実験

提案法の有効性を検証するために，音声認識実験を行った．RWCP DB 中の残響時間が 300 ms の E2A を対象とし，線アレイのうち 2 マイクを使った．l_m は 2.85, 5.7, 37.5 cm の 3 通りとした．到来方向は既知とし，(10,170), (30,130), (30,70), (70,130), (70,90)° の 5 通りの条件の単語正解精度を平均した．マイク中心と音源の距離は 2 m である．JEIDA-JCSD(B-set)(100 地名) 中，30 地名分を ID が話者ごとに 1 つずつずれるように混合音声を作成した．話者は男女 5 名ずつで，計 20 通りとした．

表 1 に binM，IVA と提案法の反復回数と単語正解精度の関係を示す．binM は，$l_m = 2.85$[cm] の場合には性能が高いが，マイク間隔が大きくなるとエリアシングの影響で性能が低下する．IVA の性能は，マイク間隔にそれほど依存しないものの，マイク間隔が狭い場合に binM に劣る．提案法は，$l_m = 2.85$[cm] の場合には binM と，$l_m = 37.5$[cm] の場合には IVA とほぼ同等の性能を達成し，$l_m = 5.7$[cm] の場合には最も性能が高くなっている．

6　まとめ

提案法は，2 値マスクと独立ベクトル分析を併用することで音源分離の頑健性が向上し，それらの上限と同等の性能を達成した．

表 1　2 値マスク (binM)，独立ベクトル分析 (IVA)，提案法 (prop) の反復回数と単語正解精度 [%] の関係．

	$l_m = 2.85$[cm]			$l_m = 5.7$[cm]			$l_m = 37.5$[cm]		
iter	binM	IVA	prop	binM	IVA	prop	binM	IVA	prop
5	**84.8**	61.1	84.2	76.4	60.9	78.2	37.0	57.6	56.8
10	-	69.1	84.3	-	69.3	79.1	-	64.0	61.8
15	-	72.6	84.3	-	72.5	79.0	-	66.8	64.4
20	-	74.1	84.4	-	73.5	**78.9**	-	**68.0**	65.3

参考文献

[1] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," IEEE Trans. Audio, Speech, Language Process., **19**, 516-527, 2011.

[2] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in Proc. WASPAA, 189-192, 2011.