パラメトリックピアソン分布を用いた周波数領域ブラインド音源分離*

○ 加藤比呂子(NTT), 永原裕一(明大), 荒木章子(NTT), 澤田宏(NTT), 牧野昭二(NTT)

1 はじめに

残響のある部屋で混合された音声に対するブラインド音源分離手法（Blind Source Separation 以下 BSS）としては、周波数帯域ごとに独立成分分析（Independent Component Analysis 以下 ICA）を適用する技術がこれまで提案されていた(例えば[1]等)。分離行列を勾配法で学習するが、その際、周波数領域における音源信号の分布をモデル化したスコア関数を用いる。音声信号に対する ICA の場合、分布のモデルとして super-Gaussian distribution が適用され、スコア関数として tanh 関数が幅広く用いられている。最近では、tanh よりも柔軟な generalized Gaussian distribution（以下 GGD）を適用するアプローチもある[2]。いずれにせよこれら従来手法では、全ての周波数で一様に同じスコア関数を適用している。

音声信号を短時間フーリエ変換（以下 STFT）で周波数領域での信号に変換してみると、周波数ビンごとの信号の分布は必ずしも一様でないことがわかる。その形状は中心が歪み裾の重い非ガウス的なものが様々存在し、ピアソン分布系に属する分布形状と類似している。実際、時間領域における信号分離のための ICA では、すでにピアソン分布の適用が提案されていて、非線形スコア関数と比べ分離精度が上がる結果が報告されている[3]。しかし、残響が畳み込まれた音を分離する場合には時間領域での処理は効率的ではない。

そこで本稿では、各周波数領域でのスコア関数をその系列に適合したピアソン分布でモデル化し分離精度を上げる手法を提案する。分布の同定に関しては、これまでモーメントに依存したパラメータの取り扱いが困難であったが、[4]により導入されたピアソン分布タイプの分類手法と、モーメントと分布パラメータの変換式を適用することでこの点が解消された。

2 音源分離システム

Fig.1 に音源分離のブロック図を示す。音源 i の信号 $s_i(n)$ と部屋の応答（残響等）が畳み込まれた信号が混合したものが、マイクロホン j で観測される信号は以下のようにモデル化される：

$$x_j(n)=\sum_{i=1}^{N}\sum_{p=1}^{P}h_{ji}(p)s_i(n-p+1) \qquad j=1,\cdots,M \quad (1)$$

$h_{ji}(n)$: 音源iからマイクロホンjへのPタップインパルス応答

尚、Fig.1 では音源数 $N=2$、マイク数 $M=2$。信号は畳み込み混合のため STFT を用いて周波数領域に変換する（Fig.1:周波数領域変換部）。すなわち(1)式は

$$X_j(f,m)=\sum_{i=1}^{N}H_{ji}(f)S_i(f,m)$$

f : 周波数、m : フレーム数

行列表現では、$\mathbf{X}(f,m)=\mathbf{H}(f)\mathbf{S}(f,m)$ と表される。次に ICA による分離処理部において、各周波数の分離プロセスは

$$\mathbf{Y}(f,m)=\mathbf{W}(f)\mathbf{X}(f,m) \quad (2)$$

となり、$\mathbf{W}(f)$ の推定は ICA を用いて更新式

$$\mathbf{W}_{i+1}(f)=\mathbf{W}_i(f)+\eta[\mathbf{I}-\langle\Phi(\mathbf{Y}(f,m))\mathbf{Y}^H(f,m)\rangle]\cdot\mathbf{W}_i(f) \quad (3)$$

$\langle x(f,m)\rangle$: m についての平均操作、
η : 学習ステップサイズ、i : 更新回数

により行う（Fig.1:独立性判定部）。$\Phi(\cdot)$ はスコア関数で、[1]では

$$\Phi(y)=\phi(|y|)\exp(j\angle y) \quad (4)$$

y : 複素数値、$|\cdot|$: 絶対値、$\angle$: 偏角

としている。最終的な分離信号 $y_k(n)$ は(2)により得られた各周波数における分離信号 $Y_k(f,m)$ を短時間離散逆フーリエ変換を用いて時間領域の信号に戻すことで得られる。

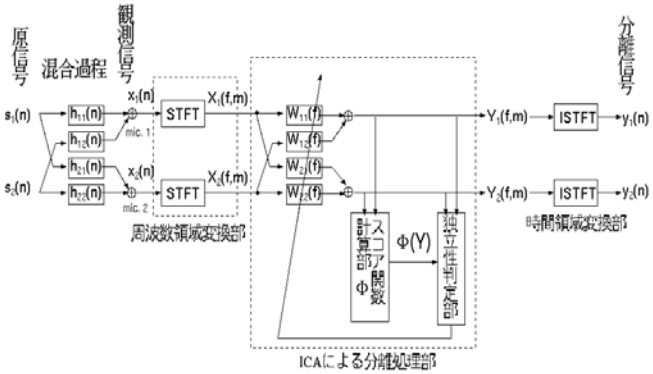

Fig. 1 音源分離ブロック図

3 ピアソン分布の適用

上述したスコア関数には、従来 $\phi(y)=\tanh(gy)$ が幅広く用いられている。この非線形関数が、全

*Frequency-domain speech blind source separation based on the parametric Pearson distribution, by KATO, Hiroko (NTT), NAGAHARA Yuichi (Meiji Univ.), ARAKI, Shoko (NTT), SAWADA Hiroshi (NTT) and MAKINO Shoji(NTT).

ての周波数ビンに対して一様に適用されている。しかし、実際は音声の STFT の系列はビンごとに分布形状が一様ではない。

そこで本稿は系列の分布にピアソン分布の適用を提案する。ピアソン分布系は微分可能な分布で、データの分布 $f(\cdot)$ の微分方程式は、b_0, b_1, c_0, c_1, c_2 を係数にもつ多項式で次のように定義される：

$$-\frac{f'(y)}{f(y)} = \frac{b_0 + b_1 |y|}{c_0 + c_1 |y| + c_2 |y|^2} e^{j\angle y} \quad (y: \text{複素数}) \quad (5)$$

分布のパラメータはデータのモーメントで決まる。この(5)式は、ICA のスコア関数 $\phi(y)$ の部分に対応する。すなわち、データから多項式(5)の係数を何らかの方法で推定すれば分布形状に見合ったスコア関数を求めることができる。また、ピアソン分布系は多数の分布族から成り立っているが、[4]では、データの歪度 Skew と尖度 Kurt からピアソン分布のタイプを直接定めるパラメータ κ を導入した：

$$\kappa = \frac{(\text{Skew})^2 (\text{Kurt}+3)^2}{4(2\times\text{Kurt}-3\times(\text{Skew})^2 - 6)(4\times\text{Kurt} - 3\times(\text{Skew})^2)}$$

ある話者の音声データを STFT 変換した系列に対して各周波数で κ を算出すると Fig.2 のようになる。男声、女声問わず同じような傾向が得られる。

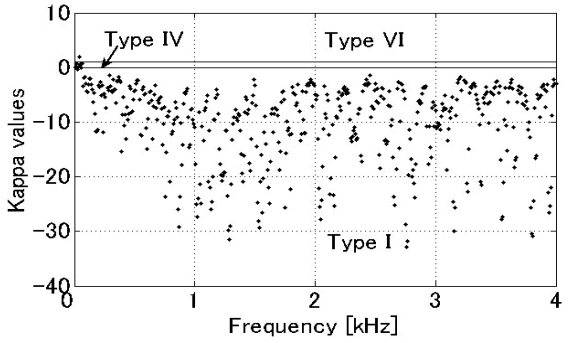

Fig. 2. 各周波数ビンにおける κ 値

κ は、とりうる範囲が $0>\kappa$、$0<\kappa<1$、$1<\kappa$ と分けられ、その範囲はピアソン分布タイプ I,IV,VI に対応する[4]。それらの分布のパラメータを直接適用したスコア関数は、

$$\begin{aligned}
\text{I} &: \frac{-(p+q-2)|y| + (p+q-2)a + (p-1)b}{|y|^2 - (2a+b)|y| + a(a+b)} e^{j\angle y} \\
\text{IV} &: \frac{2b|y| - 2b\mu - 2b\tau\delta}{|y|^2 - 2\mu|y| - \mu^2 + \tau^2} e^{j\angle y} \\
\text{VI} &: \frac{(c+1)|y| - (c+1)a - (\beta-1)\alpha}{|y|^2 - (2a-\alpha)|y| + a(a-\alpha)} e^{j\angle y}
\end{aligned} \quad (6)$$

となる。分布の各パラメータはデータのモーメントから直接求めることができる[4]。そこで、周波数領域 BSS に対し以下の２つの手法を提案する：

手法１. (5)式のパラメータをグリッドサーチで求める。具体的には、評価基準に(3)式中の $\mathbf{I} - \langle \Phi(\mathbf{Y})\mathbf{Y}^H \rangle$ の非対角要素を最小にするパラメータを用いる。

手法 2. 各周波数で歪度、尖度を算出し、κ をもとめ分布タイプを特定し、(6)式の対応する式に代入してスコア関数を求める。手法１よりも計算量が減る。

4 実験と結果

4.1 実験条件

2 マイク 2 音源の場合について実験をおこなった。マイクロホンアレイに対する話者の位置は、右を 0 度、正面を 90 度としたときにそれぞれ 30 度と 150 度とした。残響時間 130ms の部屋で実測したインパルス応答を用い、(1)式に従って畳み込み、加算を行い、マイクの観測信号を模擬した。サンプリング周波数は 8kHz、STFT のフレームサイズは 512 である。

4.2 結果

提案する手法に対し、従来法として、全ての周波数で tanh と、[2]により提案された generalized Gaussian distribution (以下 GGD)を用いた。それぞれの分離精度の良さに対する評価には Signal-to-interference-ratio (以下 SIR)を用いた。Fig.3 は原音声に女声 2 名を用いた結果である。周波数領域全体に同じ非線形関数を適用するよりも、提案したピアソン分布による周波数ごとの系列のモデル化が高い分離精度を示した。手法 2 の方が計算時間は短いが、手法 1 より分離精度は劣る。音声の組み合わせを変えた場合でも同じ傾向が得られた。

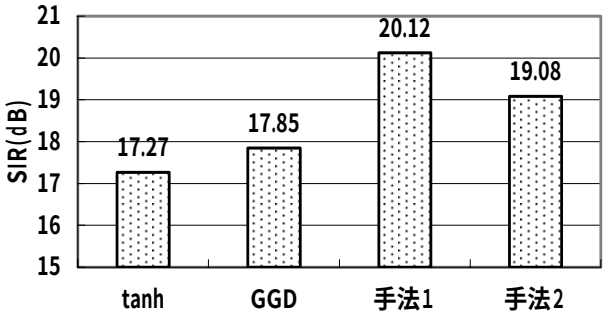

Fig. 3. 結果 (原音声：女声 2 名)

参考文献

[1] Sawada *et al.*, IEICE Trans., Fund., E86-A, 3, 590-596, 2003.

[2] Kokkinakis *et al.*, ICA 2004, LNCS 3195, Springer-Verlag Berlin Heidelberg, 486-493, 2003.

[3] Karvanen *et al.*, Proc. Sec. Int. Work. ICA. BSS, 585-590, 2000.

[4] Nagahara, J. Time. Ser. Anal, 24, 6, 721-738, 2003.