6P-2

L1ノルム最小化による劣決定音源分離のための線形計画と二次錐計画の比較評価

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　知能情報学専攻

1. はじめに

我々はコンピュータによる音環境理解を実現するため, 多数の音源を分離認識する技術の研究を進めてきた. 人間が生活する空間には多数の音源が存在するため, 音源数がマイク数を上回る “劣決定状況” での音源分離が本質的に重要である.

劣決定音源分離手法の 1 つに, 各音源の振幅がラプラス分布に従うという仮定を置くもの [1] があり, この統計的仮定に基づく最尤推定問題は L1 ノルム最小化問題と定式化できる. L1 ノルムの最小化問題は, 時間領域では線形計画問題 (LP) となり容易に計算できる一方, 時間周波数領域では二次錐計画問題 (SOCP) の一種となる [2]. 通常の SOCP ソルバーを用いると計算量が莫大となるので, 高速化のために実数の場合と同じ式により近似を行なうことが多かったが, 理論的裏付けのある厳密解との誤差が分離性能を低下させている可能性があった.

本稿では, LP による近似解と SOCP による厳密解の比較を行なう. また, 上記 SOCP の高速解法として, 補助関数法 [3] を用いた導出方法を提案し, 一般の SOCP ソルバーより約 11 倍高速に厳密解が求まる事を示す.

2. L1 ノルム最小化による劣決定音源分離

2.1 音声の混合の定式化・問題設定

まず本稿で用いる変数の定義と, 音声の混合の定式化を行なう. 本稿では観測音は時不変な混合過程に従うと仮定する. また, 音声のスパース性を向上させるために時間周波数領域上で分離を行なうことを考えると, 音声の混合過程は以下のように書ける.

$$\boldsymbol{x}_{fn} = \boldsymbol{A}_f \boldsymbol{s}_{fn} \tag{1}$$

ここで f は周波数ビン, n は時間フレームを表す. また, $\boldsymbol{s}$ は各話者の音声, $\boldsymbol{x}$ は各マイクでの観測であり, $\boldsymbol{A}$ は時不変な混合行列を示している. なお本稿では各時間周波数領域で独立に音源分離を行なうため, 以下の式では f と n を省略する.

本稿で取り上げる問題設定は次のように表される.

入力 I 個のマイクでの観測音 $\boldsymbol{x} = [x_1, ..., x_I]^T$
出力 J 人の話者それぞれの分離音 $\hat{\boldsymbol{s}} = [\hat{s}_1, ..., \hat{s}_J]^T$
仮定 混合行列 $\boldsymbol{A} = \{a_{ij}\}$ は既知, 劣決定 ($I < J$)

2.2 L1 ノルム最小化による分離

音声は優ガウス性をもつという性質から, 各音源の振幅が独立かつ同一の一般化ガウス分布 ($0 < p < 2$) に従うと仮定すると, 上記問題の最尤推定は次の式の最小化問題で表現できる.

$$C = \sum_j |s_j|^p \quad \text{subject to} \quad \boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} \tag{2}$$

Comparison Between Linear Programming and Second Order Cone Programming for Under-Determined Sound Source Separation based on L1-norm Minimization: *Yasuharu Hirasawa, Naoki Yasuraoka, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

ここで p は一般化ガウス分布の形状を示すパラメータである. 本稿では以下明示するまで $p = 1$ (ラプラス分布) として議論を進める.

$\boldsymbol{s}, \boldsymbol{x}, \boldsymbol{A}$ が全て実数の場合には, 式 (2) の最小化問題は線形計画問題 (LP) となる. この解は高々マイク数 I 個の非零要素しか持たず, 少なくとも $J - I$ 個の要素はその推定値が零となることが知られている. このため混合行列 $\boldsymbol{A}$ の部分行列の逆行列を用いることで, 式 (2) を最小化する $\boldsymbol{s}$ の候補を高速に求めることができる.

一方本稿のように時間周波数領域での信号を対象にすると, 各変数は複素数となる. この場合には上記問題は二次錐計画問題 (SOCP) となり, 計算量は劇的に増加する. また, 実数と同様に LP を用いることで, ある程度の近似解が得られるという知見 [2] があるため, 厳密解を用いての研究はあまり盛んではなかった.

3. 補助関数法による解法

3.1 更新式の導出

本稿では複素数の L1 ノルム最小化問題の解を求めるため, 補助関数法を用いたアプローチを提案する. ここで補助関数法とは, 目的関数 Q に対して以下の性質を満たす補助関数 Q^+ を定義し, その上で変数 θ と補助変数 ϕ を交互に更新していく, という手法である.

1. $Q(\theta) \leq Q^+(\theta, \phi)$
2. 各 θ について上記等号を成立させる ϕ が存在し, その値を解析的に導出できる
3. 各 ϕ について $Q^+(\theta, \phi)$ を最小化する θ の値を解析的に導出できる

この条件の下で, (1) まず ϕ を更新し, 上記 1 の等式を成立させる. (2) 次に θ を更新し, $Q^+(\theta, \phi)$ を最小化する, という 2 ステップを繰り返すことで, 目的関数 $Q(\theta)$ が広義単調減少し, 収束することが示せる.

本稿では, 亀岡らが提案した手法 [3] を複素数に対して適用し, 式 (2) に対し以下のような補助関数を考える.

$$C^+ = \sum_j \left(\frac{|s_j|^2}{2\gamma_j} + \frac{\gamma_j}{2} \right) \quad \text{subject to} \quad \boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} \tag{3}$$

ここで正の数 γ_j は補助変数であり, 不等式

$$|s_j| \leq \frac{|s_j|^2}{2\gamma_j} + \frac{\gamma_j}{2} \tag{4}$$

を用いた (等号成立は $\gamma_j = |s_j|$).

式 (3) には等号による制約条件が残っているので, これに Lagrange の未定乗数 λ_i を導入すると, 次式を得る.

$$L = \sum_j \left(\frac{|s_j|^2}{2\gamma_j} + \frac{\gamma_j}{2} \right) + \sum_i \lambda_i \left(x_i^* - \sum_j a_{ij}^* s_j^* \right) \tag{5}$$

未定乗数法を用いて式 (5) を解くと,

$$s_j = \gamma_j \sum_i 2\lambda_i a_{ij}^* \tag{6}$$

表 1: LP / SOCP solver と本補助関数法で解ける問題

手法 ＼ 変数	p=1		0<p<2
	実数のみ	複素数	複素数
LP solver	○	×	×
SOCP solver	○	○	×
本補助関数法	○	○	○

表 2: 実験条件

音源	JNAS 男女 200 文
STFT フレーム長	1024 点 (64ms)
STFT シフト幅	256 点 (16ms)
話者位置, 間隔	マイクから 1m, 60 度間隔
言語モデル	統計モデル, 語彙数 21k
音声特徴量	MFCC 25 次元 (12+Δ12+ΔPow)

表 3: 平均 SNR(dB) と 平均 ASR 正解率 (%)

	平均 SNR			平均 ASR 正解率		
解の種類	Left	Center	Right	Left	Center	Right
$p = 1$ の近似解	6.1	5.5	5.7	68.6	64.3	66.1
$p = 1$ の厳密解	6.2	5.7	5.8	68.9	65.6	66.9
$p = 0.5$ の極小解	6.2	5.6	5.8	70.9	64.3	68.0

表 4: 分離所要時間と Real Time Factor

手法	時間 (s)	RTF
LP による近似	15.6	0.013
SOCP solver	18500	15.2
本補助関数法	1670	1.37

$$\begin{bmatrix} 2\lambda_1 \\ \vdots \\ 2\lambda_I \end{bmatrix} = \begin{bmatrix} \sum_j \gamma_j a_{1j} a_{1j}^* & \cdots & \sum_j \gamma_j a_{1j} a_{Ij}^* \\ \vdots & \ddots & \vdots \\ \sum_j \gamma_j a_{Ij} a_{1j}^* & \cdots & \sum_j \gamma_j a_{Ij} a_{Ij}^* \end{bmatrix}^{-1} \begin{bmatrix} x_1 \\ \vdots \\ x_I \end{bmatrix} \quad (7)$$

という更新式が得られる.

以上をまとめると, 補助関数法を用いた解の導出は次のように行なわれる.

1. 適当な非零値により s_j を初期化する
2. $\gamma_j = |s_j|$ により補助変数を更新する
3. 式 (7) により λ_i を更新する
4. 式 (6) により s_j を更新する
5. 精度が充分でなければ手順 2 に戻る

なお実装上の注意として, γ_j が 0 になると式 (4) の右辺が定義できずに更新が正しく行なわれなくなるので, 適当な微小値 ϵ を用いたフロアリングが必要になる.

3.2 L_p ノルムへの拡張

以上の導出では一般化ガウス分布のパラメータを $p = 1$ とした更新式を導出したが, 同様の補助関数を用いて $0 < p < 2$ の任意のパラメータに対する導出が可能となる (表 1). これは補助関数として二次関数を用いているためで, 副次的にラプラス分布から一般化ガウス分布への拡張が行なわれたことになる. 音声の分布は $p = 0.5$ 付近に従うという知見 [4] もあり, 有意義な拡張が行なわれていると言える. なお, $0 < p < 2$ に対する更新式は, 式 (6) と式 (7) の γ_j を γ_j^{2-p}/p で置き換えたものである.

ここで注意するべきことは, $p < 1$ の場合には問題の凸性が失われ, 非凸問題となっていることである. 補助関数法は目的関数を広義単調減少させていく手法であるため, 非凸最適化では局所的最適解への収束が起こり得る. なお本稿の実験では約 83%が大域的最適解へ収束し, おおむね良好な結果が得られることが示されている.

4. 評価実験と考察

2 つの目的 (1) 各手法の分離精度の音声認識率の観点からの評価 と, (2) 各手法の高速性の評価 のために, 2 マイクで 3 話者の混合音声を分離する実験を行なった. ここで混合音声は無響室のインパルス応答を用いて合成したものを用いた. また SOCP ソルバーは商用ソフトウェア mosek を使用し, その他の手法は C++言語で実装した. 使用した計算機の CPU は Intel Core i7 2.93GHz, RAM は 8GB であり, その他実験条件の詳細は表 2 の通りである.

LP による近似解と SOCP による厳密解について, 平均 SNR と平均音声認識 (ASR) 正解率の観点からの結果を, 表 3 に示す. 表より, 平均 SNR・平均 ASR 正解率ともに差は小さく, 近似による性能の低下はほぼ無視できることが分かった. これは SDR などの尺度で行なわれた従来の結果 [4] とも一致している. 実際に厳密解の値を見てみると, 58%の時間周波数では分離結果のうち 1 話者のパワーが 0 値となっており, 近似解と同じ出力が得られていた. これはスパースな解を出力しやすいという L1 ノルム最小化問題の特性が有効なためと考えられる. また, $p = 0.5$ の分離では中央話者の認識率が低下し, 左右話者の認識率が向上する結果が得られた.

次に, 補助関数法による更新手法の高速性を, 計 1218 秒の混合音の分離時間から評価した. なお実験環境では SOCP ソルバーの出力精度が 40dB 程度であったので, 補助関数法も 40dB の精度が得られた時点を反復の終了条件とした. 表 4 に示した実験結果から, 本補助関数法の求解速度が SOCP ソルバーのそれより約 11 倍であることが分かる. LP による近似に比べれば依然 100 分の 1 の速度だが, Real Time Factor が実時間動作を示す 1 に近づいたことからも, マルチコアによる並列化を用いれば実時間動作は充分可能であると期待できる.

5. 終わりに

本稿では L1 ノルム最小化による劣決定音源分離に関して, 線形計画 (LP) による近似解と二次錘計画 (SOCP) による厳密解の比較を平均 SNR と平均 ASR 正解率の観点から行なった. また, 補助関数法を用いて式 (2) の最小化問題の解を計算する手法を提案し, その有効性と高速性を示した. 今後の課題として, 提案した更新式を用いた L_p ノルム最小化による音源分離手法の詳細な解析, 及び分離手法の拡張などが考えられる.

謝辞 本研究の一部は科研費 (S), 特定領域, GCOE, 日仏研究交流の支援を受けた.

参考文献

[1] P.Bofill, and M.Zibulevsky: 'Underdetermined blind source separation using sparse representations', *Signal Processing*, Vol.81 No.11, pp.2353–2362 (2001).

[2] S.Winter, H.Sawada, and S.Makino: 'On real and complex valued L1-norm minimization for overcomplete blind source separation', *In Proc. of WASPAA 2005*, pp.86–89 (2005).

[3] 亀岡弘和, 鎌本優, 原田登, 守谷健弘: 予測誤差の Golomb-Rice 符号量を最小化する線形予測分析, 電子情報通信学会論文誌, Vol.J91-A, No. 11, pp.1017-1025 (2008).

[4] E.Vincent: 'Complex Nonconvex l_p Norm Minimization for Underdetermined Source Separation', *In Proc. of ICA 2007*, pp.430–437 (2007).