Microphone array signal processing via convex optimization

凸最適化を用いたアレイ信号処理

5116E010-8　立川 智哉　　指導教員　及川 靖広　教授

TACHIKAWA Tomoya　　Prof. OIKAWA Yasuhiro

概要：マイクロホンアレイは様々なデバイスで使用されており，アレイを用いた様々な信号処理の研究が盛んに行われている．本研究では，アレイで観測した信号から個々の音源方向の推定と，音源分離を同時に行う手法を提案する．未知の混合系の候補として物理モデルに基づく空間辞書を用いて，観測信号をモデル化する．それから，2 種類の音源のスパース性を仮定し，その仮定を満たす適切な凸最適化問題として定式化する．シミュレーション実験により，他手法に比べて高精度で分離することが可能で，初期値に頑健な手法であることを確認した．

キーワード：グループスパース，$\ell_{2,1}$ 混合ノルム，$\ell_{1,2}$ 混合ノルム，平面波，近接分離

Keywords: Group sparsity, $\ell_{2,1}$-norm, $\ell_{1,2}$-norm, plane wave, primal-dual splitting

1. Introduction

A lot of signal processing methods utilizing a microphone array have been investigated for many techniques including the 3D sound source localization [1, 2] and source separation [3]. In underdetermined source separation problem that the number of sources is more than the number of microphones, many methods usually heavily rely on some sort of sparsity assumptions. Although a lot of algorithms are shown to be effective in many articles, many of these methods have the common drawback, i.e., performance of these methods depend on the initialization. That is, one usually obtains a different result as an initial value of the algorithm changes. Therefore, there is a risk of obtaining a terrible result. This dependence on initial values is basically owing to non-convexity of the problem formulation. Although there exist some research with convex formulations, the convexity is realized by the knowledge of the mixing filters. That is, these convex methods have to estimate the mixing filter first, and then the sources are separated through convex optimization. Therefore, the initialization issue remains in the mixing filter estimation part.

To circumvent such initialization dependency, an underdetermined source separation method formulated as a convex optimization problem is proposed. By approximating the mixing process with a predetermined spatial dictionary, the proposed method can simultaneously treat the mixing process and avoid the preceding estimation of mixing filters. Two penalty functions, $\ell_{1,2}$- and $\ell_{2,1}$-norm, corresponding to disjointness in time-frequency domain and sparseness of direction-of-arrival (DOA) are utilized to enhance the performance.

2. Proposed method

One of the difficulties of a source separation problem is the multiplicative nature of unknown variables. This form causes not only the scale indeterminacy but also local-minimum traps due to the non-convexity. Therefore, avoiding the multiplication of unknowns is desirable, which might be achieved by determining one of them in advance. To reduce the difficulty, a predetermined dictionary is adopted in the observation model.

Let K elements of a dictionary be chosen for the mixing matrix. Then, signals of N sources obtained from M microphones are represented by

$$\mathbf{y}(t,f) \simeq \mathbf{A}(f)\mathbf{s}(t,f), \tag{1}$$

where t is index of time frame, f is frequency index, $\mathbf{y}(t,f) = [y_1(t,f), \cdots, y_M(t,f)]^{\mathrm{T}}$ is an observed signal vector from microphones, $\mathbf{A}(f)$ is an $M \times K$ dictionary matrix consisting of the predetermined elements, and $\mathbf{s}(t,f) = [s_1(t,f), \cdots, s_K(t,f)]^{\mathrm{T}}$ is the corresponding coefficients. By fixing the matrix $\mathbf{A}(f)$ in advance, the unknown variables becomes $\mathbf{s}(t,f)$ only, which does not involve the difficulty associated with the multiplication. For the dictionary elements, while any other spatial model such as point sources [1, 2] can be utilized, the plane waves are considered in this paper for simplicity. That is, the dictionary matrix is constructed in the following manner:

$$\mathbf{A}(f) = [\mathbf{a}_{\theta_1}(f), \cdots, \mathbf{a}_{\theta_K}(f)], \tag{2}$$

where $\mathbf{a}_{\theta_k}(f)$ is a steering vector associated with direction θ_k,

$$\mathbf{a}_{\theta_k}(f) = [\exp(-j\tfrac{\omega_f}{c}\mathbf{u}_k^{\mathrm{T}}\mathbf{p}_1), \cdots, \exp(-j\tfrac{\omega_f}{c}\mathbf{u}_k^{\mathrm{T}}\mathbf{p}_M)]^{\mathrm{T}}, \tag{3}$$

c is the speed of sound, ω_f is angular frequency, $\mathbf{u}_k$ is the unit vector corresponding to the direction θ_k, $\mathbf{p}_m$ is the position of m-th microphone, and $j = \sqrt{-1}$. This choice of dictionary enables us to explicitly consider the direction of sound sources as schematically illustrated in Fig. 1. By considering sufficiently many directions (sufficiently large K), each sound arrived from some direction can be approximately represented by an element with similar direction. Since information of both sound sources and their directions are handled by the coefficients $\mathbf{s}(t,f)$, simultaneous estimation of DOA and sound source signals should be accomplished if appropriate assumptions are imposed on $\mathbf{s}(t,f)$.

The proposed method is formulated as the following convex optimization problem:

$$\min_{\mathbf{s}} \ \|\mathbf{A}\mathbf{s} - \mathbf{y}\|_2^2 + \alpha \|\mathbf{s}\|_{2,1} + \beta \|\mathbf{s}\|_{1,2}^2, \tag{4}$$

where the first term corresponds to Eq. (1) in a matrix form containing all time-frequency bins and $\alpha, \beta \geqq 0$

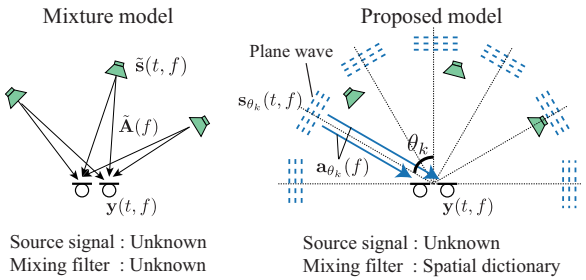

Fig. 1 Illustration of the mixing model and proposed one.

are regularization parameters. Each of the second and third terms correspond to two sparsity assumptions: sources exist at a few directions (spatial sparsity), and spectrogram of each source is sparse (disjointness). The second term is the $\ell_{2,1}$-norm concentrates the energy into a few directions so that the coefficients corresponding to DOA of the sources have higher energy than the other directions. Therefore, this penalty imposes the assumption of spatial sparsity of sound sources [1, 2]. The third term is the squared $\ell_{1,2}$-norm promotes sparsity such that a few elements have energy at each time-frequency bin. That is, this mixed norm promotes disjointness to each time-frequency bin. Since Eq. (4) is convex, the proposed formulation is robust against a choice of an initial value because the global solution of convex optimization problem is unique.

3. Experiments

Performance of the proposed method was evaluated by applying it to a publicly available audio source separation task. Table. 1 is the experimental condition. The proposed method was compared with a minimum variance distortionless response (MVDR) beamformer and multichannel NMF (MNMF). Separation performance was evaluated by Signal-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR), and Source-to-Artifacts Ratio (SAR).

In the proposed method, estimated DOA and separated signals are obtained simultaneously. Fig. 3. shows magnitude for each direction $(\sum_{t,f}|s_k(t,f)|^2)^{\frac{1}{2}}$ indexed by k. From the figure, it can be seen that both female and male data resulted in four peaks (red circles) which appeared near the true directions (doted lines) as expected. Separation performance was evaluated for the directions of these four peaks (red circles).

Fig. 3. shows female/male speech separation results, where each color represents each source, and the error bars of MNMF and the proposed method represent standard deviations within ten trials with random initialization. The separation performance of the proposed method can be obtained the separation score for all sources with comparing other methods. Although MNMF has dependence on initial values because the difficulty arising from the multiplication caused a lot of local minimum traps, the proposed method was not affected by initial values. Therefore, the proposed method does not depend on a choice of initial values thanks to the convexity of the proposed formulation.

4. Conclusion

In this paper, an underdetermined source separation method based on convex optimization is proposed.

Table 1 Experimental condition.

Number of sources	$N = 4$
Number of microphones	$M = 2$
FFT length	256 ms
Window shift	64 ms
Reverberation time	130 ms
Number of dictionary elements (directions)	$K = 37$ (5 deg. interval)

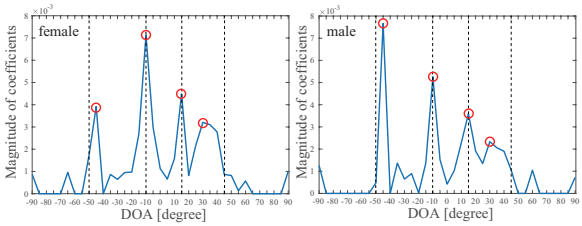

Fig. 2 Magnitude of estimated coefficients for each direction. The black doted lines are the true directions. The directions of the peaks were chosen for evaluation.

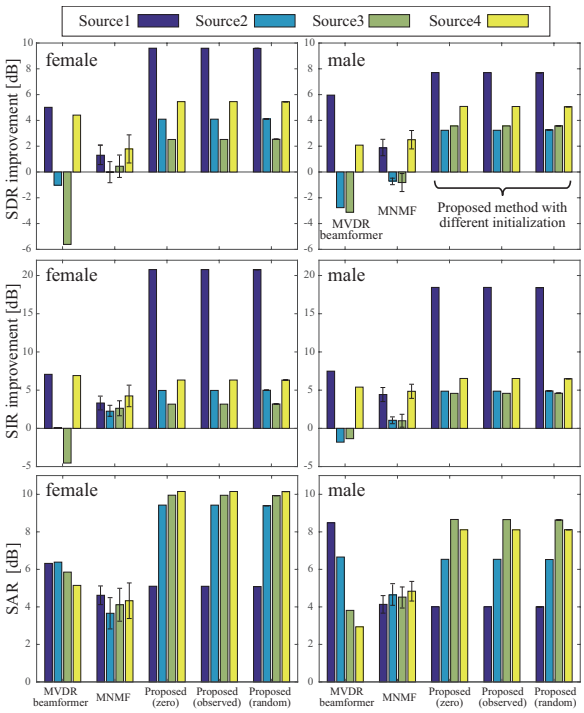

Fig. 3 Separation results. Error bars represent their standard deviations obtained from 10 random initial values.

The predetermined spatial dictionary and two sparsity assumptions allow simultaneous estimation of DOA and source signals, while initialization issue is totally eliminated by the convex formulation as shown in the experimental section.

References

[1] T. Tachikawa, K. Yatabe and Y. Oikawa, "Coherence-adjusted monopole dictionary and convex clustering for 3D localization of mixed near-eld and far-eld sources," in IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), pp.3191-3195, Mar. 2017.

[2] T. Tachikawa, K. Yatabe, Y. Ikeda and Y. Oikawa, "Sound source localization based on sparse estimation and convex clustering," 5th Joint Meet. Acoust. Soc. Am. Acoust. Soc. Jpn., Hawaii, Nov.Dec. 2016.

[3] 立川智哉, 矢田部浩平, 及川靖広, "点音源モデルを用いたスパース推定による音源分離," 日本音響学会講演論文集, pp.535-536, Sep. 2017.