

# 話者分類とSN比最大化ビームフォーマに基づく会議音声強調\*

荒木 章子, 澤田 宏, 牧野 昭二 (NTT 研究所)

## 1 はじめに

近年会議録プロジェクトが国内外で検討されているが [1, 2]、話者音声のオーバーラップを積極的に扱う例はまだ少ない。一方、音声のオーバーラップを解決すべき課題とする音源分離 (e.g., [3]) では、全ての音源が区間中話し続けているシナリオが主で、話者の交代等についてはあまり議論がなされてこなかった。

本稿では、話者が交代で発話するが自然なオーバーラップも生じる、という会議環境における話者音声強調手法を提案する。

## 2 問題設定

$N$  人の話者音声  $M$  個のマイクにて観測されるとする。マイク  $j$  による観測信号  $x_j$  は、 $x_j(t) = \sum_{k=1}^N \sum_{l=1}^{\infty} h_{jk}(l) s_k(t-l+1) + n_j(t)$  ( $j = 1, \dots, M$ ) とモデル化される。ここで  $s_k$  は話者  $k$  の音声であり、各話者は断続的な発話をする。人数  $N$  は未知とする。また  $h_{jk}$  は話者  $k$  とマイク  $j$  間のインパルス応答、 $n_j(t)$  はマイク  $j$  におけるノイズである。本稿の目的は、観測信号  $x_1, \dots, x_M$  から、それぞれの話者音声を強調することである。

本稿では、時間領域での観測信号  $x_j(t)$  ( $j = 1, \dots, M$ ) に短時間フーリエ変換 (STFT) を適用し、時間周波数領域にて信号を  $x_j(f, t) \approx \sum_{k=1}^N h_{jk}(f) s_k(f, t) + n_j(f, t)$  として扱う。また  $\mathbf{x}(f, t) = [x_1(f, t), \dots, x_M(f, t)]^T$  を観測信号ベクトルとする。

## 3 提案手法

本稿では、話者  $k$  毎に SN 比最大化ビームフォーマ  $\mathbf{w}_k(f)$  [4] を構成し、

$$y_k(f, t) = \mathbf{w}_k^H(f) \mathbf{x}(f, t) \quad (1)$$

にて強調音声を得る方法を提案する。SN 比最大化ビームフォーマは、出力信号  $y_k(f, t)$  中の目的話者  $k$  信号 (信号成分) とノイズおよび他話者信号 (ノイズ成分) とのパワー比を最大化するよう求められる。SN 比最大化ビームフォーマは、適応ビームフォーマ (e.g., [1]) におけるステアリングベクトルを必要としないため、残響下でもロバストな動作が期待される一方、信号成分およびノイズ成分の推定を要する。本章では、信号成分およびノイズ成分の推定方法と、SN 比最大化ビームフォーマの構成方法を順に説明する。

[STEP1] ノイズ区間推定: まず、誰も発話していないノイズ区間  $\mathcal{P}_N$  を推定する。これは、後の話者分類で、ノイズによる誤分類が生じることを避けるためである。ここでは [5] による音声区間検出を用いた。

具体的には

$$\Lambda(t) = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \{\gamma(f, t) - \log \gamma(f, t) - 1\} \quad (2)$$

をフレーム毎に計算し、この値  $\Lambda(t)$  が閾値  $\eta$  より小さければノイズ区間  $\mathcal{P}_N$ 、大きければ音声区間  $\mathcal{P}_S$  と判定する。ここで  $\mathcal{F}$  は考慮する周波数の集合、 $|\mathcal{F}|$  はその集合の要素数、 $\gamma(f, t) = \frac{|x_j(f, t)|^2}{\sigma(f)}$  は事後 SN 比、 $\sigma(f)$  はノイズパワーの推定値である。

[STEP2] 話者分類: 次に、音声区間  $\mathcal{P}_S$  における観測信号ベクトルを、各話者の発話区間に分類する。本稿では、マイク間の信号の到来時間差をフレーム毎に推定し、これを分類する。

マイク  $j$  とマイク  $j'$  に関する信号の到来時間差  $\tau_{jj'}$  は、GCC-PHAT [6] により以下のように推定できる。

$$\tau_{jj'}(t) = \operatorname{argmax}_{\tau} \sum_f \frac{x_j(f, t) x_{j'}^*(f, t)}{|x_j(f, t) x_{j'}^*(f, t)|} e^{j2\pi f \tau} \quad (3)$$

これをある基準マイク  $j'$  とその他のマイク  $j$  について求め、それらを並べた縦ベクトルを  $\boldsymbol{\tau}(t)$  とする。

次に、各フレーム  $t \in \mathcal{P}_S$  における到来時間差  $\boldsymbol{\tau}(t)$  を、オンラインクラスタリング [7] にて話者別に分類する。オンラインクラスタリングは、クラスタリングを 1 クラスタから始め、既存のクラスタのセントロイドからある閾値以上離れたデータが観測された時に、そのデータをセントロイドとした新たなクラスタを生成する手法であり、発話者数が未知である場合にも対応することができる。各クラスタが各話者に対応している。以降フレーム  $t$  がクラスタ  $k$  に分類された時  $C(t) = k$  と書くことにする。

[STEP3] 音声強調: 最後に、検出された各話者  $k$  毎に SN 比最大化ビームフォーマ  $\mathbf{w}_k(f)$  を構成する。SN 比最大化ビームフォーマは前述の通り、出力信号  $y_k(f, t)$  中の話者  $k$  の区間信号と、ノイズおよび他話者区間の信号のパワーの比

$$\lambda(f) = \frac{\sum_{C(t)=k} |y_k(f, t)|^2}{\sum_{C(t) \neq k} |y_k(f, t)|^2} = \frac{\mathbf{w}_k^H(f) \mathbf{R}_T^k(f) \mathbf{w}_k(f)}{\mathbf{w}_k^H(f) \mathbf{R}_I^k(f) \mathbf{w}_k(f)} \quad (4)$$

を最大化するビームフォーマとして設計される。ここで  $\mathbf{R}_T^k(f) = \sum_{C(t)=k} \mathbf{x}(f, t) \mathbf{x}^H(f, t)$ 、 $\mathbf{R}_I^k(f) = \sum_{C(t) \neq k} \mathbf{x}(f, t) \mathbf{x}^H(f, t)$  である。

式 (4) を  $\mathbf{w}_k(f)$  で微分し 0 とおくと、

$$\mathbf{R}_T^k(f) \mathbf{w}_k(f) = \lambda(f) \mathbf{R}_I^k(f) \mathbf{w}_k(f) \quad (5)$$

という関係が得られる。最大の SN 比  $\lambda(f)$  は、(5) で与えられた一般化固有値問題における最大固有値となり、その最大固有値に対応する固有ベクトル  $\mathbf{e}(f)$  が話者  $k$  に関する SN 比最大化ビームフォーマの係数

\*Speech enhancement in a meeting situation with speaker clustering and maximum SNR beamformers by ARAKI Shoko, SAWADA Hiroshi, and MAKINO Shoji (NTT Communication Science Laboratories, NTT Corporation).

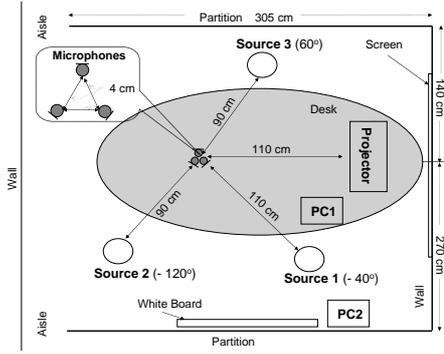


Fig. 1 実験環境 (残響時間 350 ms)

を与える。

$$\mathbf{w}_k(f) = \mathbf{e}(f). \quad (6)$$

尚、SN 比最大化ビームフォーマは、ゲインに関して不定性を持つため、これをそのまま音声信号のような広帯域信号に適用すると、出力が  $\mathbf{w}_k(f)$  の周波数特性により歪む [8]。これをここでは、観測信号とビームフォーマ  $\mathbf{w}_k(f)$  の出力信号との誤差

$$\begin{aligned} \mathcal{G}(\mathbf{a}(f)) &= \sum_t \|\mathbf{x}(f, t) - \mathbf{a}(f)y_k(f, t)\|^2 \\ &= \sum_t \|\mathbf{x}(f, t) - \mathbf{a}(f)\mathbf{w}_k^H(f)\mathbf{x}(f, t)\|^2 \end{aligned}$$

を最小にする補正フィルタ

$$\mathbf{a}(f) = \frac{\sum_t y_k^*(f, t)\mathbf{x}(f, t)}{\sum_t |y_k(f, t)|^2} = \frac{\mathbf{R}_x(f)\mathbf{w}_k(f)}{\mathbf{w}_k^H(f)\mathbf{R}_x(f)\mathbf{w}_k(f)} \quad (7)$$

にてビームフォーマ  $\mathbf{w}_k(f)$  を補正する。ここで  $\mathbf{R}_x(f) = \sum_t \mathbf{x}(f, t)\mathbf{x}^H(f, t)$  は観測信号の全時間区間における相関行列である。ビームフォーマの補正は、 $\mathbf{a}(f)$  のある任意の  $J$  番目の要素  $a_J(f)$  を用い

$$\mathbf{w}_k(f) \leftarrow a_J(f)\mathbf{w}_k(f) \quad (8)$$

により行う。

この補正されたビームフォーマを用い、(1)にて話者  $k$  に関する強調音声  $y_k(f, t)$  を得る。

## 4 実験と結果

図 1 に示す環境で実験を行った。混合信号は、図 2 に例示するタイミングで発話される英語音声および、図 1 の環境で計測したインパルス応答とノイズを用いて作成した。信号長は 30 秒間である。サンプリング周波数は 16kHz、STFT フレーム長  $L$  は 2048、フレームシフトは 256 である。性能を signal-to-interference plus noise-ratio (SINR) と signal-to-distortion-ratio (SDR) にて評価した。比較方法としては、適応ビームフォーマ

$$\mathbf{w}_k(f) = \frac{[\mathbf{R}_I^k(f)]^{-1}\mathbf{v}_k(f)}{\mathbf{v}_k^H(f)[\mathbf{R}_I^k(f)]^{-1}\mathbf{v}_k(f)}$$

を用いた。ここで  $\mathbf{v}_k(f)$  は話者  $k$  に関するステアリングベクトルであり、STEP2 で得られたセントロイドで与えられる到来時間差を用いて求めた。

強調音声の例を図 3 に、SINR と SDR の値を表 1 に示す。提案法では、比較法よりも高い性能にて、それぞれの話者音声を強調することができている。また、補正フィルタにより SDR 値も改善されたことが

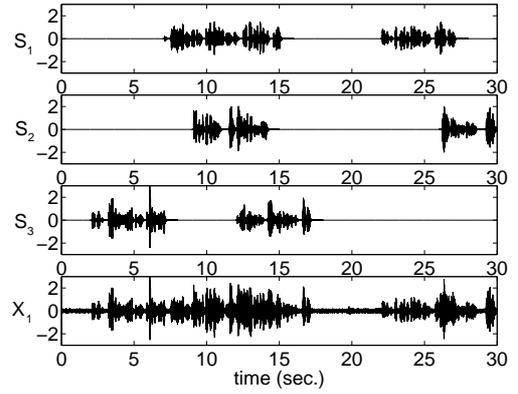


Fig. 2 原音声 ( $s_1 \sim s_3$ ) と混合音声 ( $x_1$ ) の例

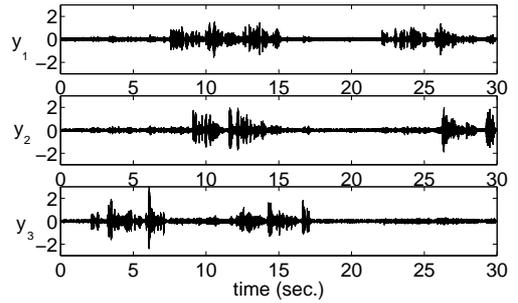


Fig. 3 強調音声の例

Table 1. 強調結果。SDR': 補正フィルタ (8) 無し。

	SINR[dB]	SDR[dB]	SDR'[dB]
比較法	7.0	11.5	-
提案法	9.0	12.6	3.8

分かる。

また我々は、図 1 と同じ環境で話者 4 名による会話を収録し、本手法を適用した。30 秒間の収録全体では 4 名全ての発話が含まれたため、マイク 3 個の本システムではあまり良い強調はできなかったが、収録音声を 5 秒ブロックに区切ることで、各ブロックではほぼ 3 話者以下の発話に限ることができ、良い強調音声を得ることができた。講演要旨に例を示している。

## 参考文献

- [1] 浅野他, “会議収録データにおける発話イベントの構造化と分離について”, 音講論 2006 年秋, pp. 29-30.
- [2] [http://www.nist.gov/speech/test\\_beds/mr\\_proj/](http://www.nist.gov/speech/test_beds/mr_proj/)
- [3] S. Makino, et al., “Blind source separation of convolutive mixtures of speech in frequency domain,” *IEICE Trans. Fundamentals*, vol. E88-A, no. 7, pp. 1640-1655, 2005.
- [4] H. L. Van Trees, ed., *Optimum Array Processing*, Wiley, 2002.
- [5] J. Sohn, et al., “A statistical model-based voice activity detection,” *IEEE SP Letters*, vol. 6, no. 1, pp. 1-3, 1999.
- [6] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. ASSP*, vol. 24, no. 4, pp. 320-327, 1976.
- [7] R. O. Duda, et al., *Pattern Classification*, 2nd ed., Wiley, 2000.
- [8] E. Warsitz and R. Haeb-Unbach, “Controlling speech distortion in adaptive frequency-domain principal eigenvector beamforming,” in *Proc. IWAENC 2006*, 2006.