時間周波数スイッチングビームフォーマと時間周波数マスキングによる劣決定音声強調*

◎山岡洸瑛（筑波大），小野順貴（首都大），牧野昭二，山田武志（筑波大）

1 はじめに

十分な数のマイクロフォンが使用できる時，ビームフォーマによる音声強調やブラインド音源分離 [1] は効果的な音声強調を達成する．音声強調を事前に適用することで音声認識性能が向上することが報告されており（例えば [2]），そういった音声アプリケーションの前処理として音声強調は重要なタスクである．ところが，マイクロフォンアレイに基づく音声強調性能はマイクロフォンの数 M に依存する．一般に 1 個の目的音源と $N-1$ 個の干渉音源が存在する時，音源数 N と同数以上のマイクロフォンが必要となる（$M \geq N$）．一方で近年普及している IC レコーダなどの小型機器は高々 2 個のマイクロフォンを持つことが多く，そういった小型機器でも効果的に音声強調を達成する手法の開発が求められている．従来の時間周波数マスキング [3–5] やマルチチャネル Wiener フィルタ [6] などの手法は劣決定条件下（$M < N$）においても音声強調が可能である．しかし，これらの手法はミュージカルノイズのような人工ノイズを生じさせやすく，音声認識などの後段のアプリケーションに好ましくない．

そこで本研究では，線形信号処理の拡張として，複数のビームフォーマを組み合わせることにより少ない歪みで高い音声強調性能を示す音声強調手法を提案している．N 個の音源と 2 個のマイクロフォンが存在する時，従来の線形ビームフォーマはただ 1 つの干渉音源のみを抑圧し，残りの干渉音源は抑圧されない．しかし，$N-1$ 個の干渉音源それぞれを抑圧するような $N-1$ 個のビームフォーマが構成できれば，それらを組み合わせることで音声強調性能を向上させることができる．

[7] においては，オーディオズームのための異なる指向性を有する複数ビームフォーマの組み合わせが提案された．一方で，我々は同一の目的音源を強調し，異なる干渉音源を抑圧する複数のビームフォーマを組み合わせる．正方形マイクロフォンアレイを用いた複数の固定ビームフォーマの周波数方向の組み合わせと Wiener フィルタによる音声強調手法が提案されているが [8,9]，目的音声に歪みを生じやすいことが知られている．また，ロボットの機械の駆動音（モータなど）の抑圧を目的とし，時間周波数点毎に最適な雑音共分散行列をクラスタリングにより選択し，ビームフォーミングを行う手法が提案されている [10]．この手法は，ロボットの駆動音の種類が限られているという仮定のもと，事前に雑音をクラスタリングする必要がある．一方，本研究では劣決定条件下における音声強調のために，複数の適応ビーフォーマを組み合わせる．

これまでに，複数ビームフォーマの組み合わせ方法として，複数ビームフォーマ出力の積の累乗根をとる複素値相乗平均（Complex-valued geometric mean; CGM）と，出力の最小絶対値をとる最小値選択（Minimum value selection; MIN）の 2 つの方法を提案してきた．本報告では，高い音声強調を示す最小値選択，すわなち時間周波数スイッチング（Time-frequency-bin-wise switching; TFS）ビームフォーマと，時間周波数マスキングの利用によるその拡張を提案し，提案手法の性能評価を行った．

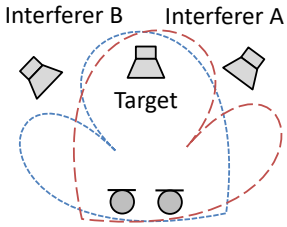


Fig. 1: Combination of two beamformers with a spatial null for each interferer

2 従来の線形ビームフォーマ

多くの音声強調手法において，マイクロフォン観測は短時間フーリエ変換により時間周波数領域で表される．ここで，$x_i(\omega, t)$ を 周波数 ω，t 番目の時間フレームにおける i 番目のマイクロフォン観測とする．

簡単のため 2 マイクの場合を考えると，線形ビームフォーマは一般に以下の式で与えられる．

$$y(\omega,t) = \mathbf{w}^{\mathrm{H}}(\omega)\mathbf{x}(\omega,t) \quad (1)$$

$$\mathbf{x}(\omega,t) = [x_1(\omega,t), x_2(\omega,t)]^{\mathrm{T}} \quad (2)$$

$$\mathbf{w}(\omega) = [w_1(\omega), w_2(\omega)]^{\mathrm{T}} \quad (3)$$

ここで $y(\omega,t)$ はビームフォーマの出力，$\mathbf{w}(\omega)$ は構成された空間フィルタ，$\{\cdot\}^{\mathrm{T}}$ は転置，$\{\cdot\}^{\mathrm{H}}$ は複素共役転置を示す．$\mathbf{w}(\omega)$ の設計には maximum signal-to-noise ratio (MaxSNR) ビームフォーマ [11,12] や minimum variance distortionless response (MVDR) ビームフォーマ [11,13] などが提案されている．しかし，一般に M 個のマイクロフォンでは $M-1$ 個の干渉音源のみ抑圧が可能である．従って，線形音声強調は，音源数よりもマイクロフォンの数が少ない場合には（劣決定条件，$M < N$）性能が劣化する．

3 複数ビームフォーマの組み合わせによる音声強調

簡単のため，目的音源と干渉音源 A, B からなる 3 音源（以下 tgt, $\mathrm{i_A}$, $\mathrm{i_B}$ とする）を 2 つのマイクロフォンで抑圧することを考える．この状況下では，2 つの

*Underdetermined speech enhancement with time-frequency-bin-wise switching beamformer and time-frequency masking. by Kouei YAMAOKA (University of Tsukuba), Nobutaka ONO (Tokyo Metropolitan University), Shoji MAKINO, Takeshi YAMADA (University of Tsukuba)

Table 1: Dominant sound(s) in each signal

$\mathbf{x}$	y_{A}	y_{B}	y_{CGM}	y_{MIN}
tgt	tgt	tgt	tgt	tgt
$\mathrm{i_A}$	0 *	$\mathrm{i_A}$	0 *	0 *
$\mathrm{i_B}$	$\mathrm{i_B}$	0 *	0 *	0 *
tgt, $\mathrm{i_A}$	tgt	tgt, $\mathrm{i_A}$	tgt, $\mathrm{i_A}$	tgt
tgt, $\mathrm{i_B}$	tgt, $\mathrm{i_B}$	tgt	tgt, $\mathrm{i_B}$	tgt
$\mathrm{i_A}$, $\mathrm{i_B}$	$\mathrm{i_B}$	$\mathrm{i_A}$	$\mathrm{i_A}$, $\mathrm{i_B}$	$\mathrm{i_A}$ or $\mathrm{i_B}$
tgt, $\mathrm{i_A}$, $\mathrm{i_B}$	tgt, $\mathrm{i_B}$	tgt, $\mathrm{i_A}$	tgt, $\mathrm{i_A}$, $\mathrm{i_B}$	tgt, $\mathrm{i_A}$ or tgt, $\mathrm{i_B}$

* suppressed

干渉音源を同時に抑圧する空間フィルタは構成できない．ここで，もし目的音源と干渉音源 A のみが観測されたならば，干渉音源 A を抑圧するビームフォーマ A を従来のビームフォーマを用いて構成することができる．同様に干渉音源 B を抑圧するビームフォーマ B も構成することができる（図 1）．これらのビームフォーマを用いることで，3 音源からなる観測信号 $\mathbf{x}$ を用いて以下の 2 出力 y_{A}，y_{B} を得る．

$$y_{\mathrm{A}}(\omega,t) = \mathbf{w}_{\mathrm{A}}^{\mathrm{H}}(\omega)\mathbf{x}(\omega,t) \quad (4)$$
$$y_{\mathrm{B}}(\omega,t) = \mathbf{w}_{\mathrm{B}}^{\mathrm{H}}(\omega)\mathbf{x}(\omega,t) \quad (5)$$

ここで $\mathbf{w}_{\mathrm{A}}$，$\mathbf{w}_{\mathrm{B}}$ はそれぞれビームフォーマ A，B の空間フィルタである．

$\mathbf{x}$，y_{A}，y_{B} における支配的な音源を表 1 の 1–3 列に示した．$\mathbf{x}$ の全ての時間周波数点において支配的な音源は 1 列目に示した 7 パターンとなる．ここで音源が存在しないケースは自明であるため考慮しない．y_{A} 及び y_{B} の列に着目すれば，目的音源のみが支配的である場合，2 つのビームフォーマは共に目的音を出力する（2 行目参照）．干渉音源 A のみが支配的な場合，ビームフォーマ A は抑圧された信号を出力するが，ビームフォーマ B は干渉音源 A に対する制約を持たないため，何らかの影響が及ぼされた干渉音 A を出力する．提案法では，これら複数のビームフォーマを組み合わせることで音声強調を行う．

3.1 時間周波数スイッチングビームフォーマ

干渉音源 A (B) が到来した時，ビームフォーマ A (B) の出力はビームフォーマ B (A) よりも小さくなる．従って以下のように，時間周波数点毎に最小の振幅値を持つ出力を選択することで，効果的な音声強調を達成する．

$$y_{\mathrm{MIN}}(\omega,t) = \begin{cases} y_{\mathrm{A}}(\omega,t) & \text{if } |y_{\mathrm{A}}(\omega,t)| \le |y_{\mathrm{B}}(\omega,t)| \\ y_{\mathrm{B}}(\omega,t) & \text{otherwise} \end{cases} \quad (6)$$

ここで目的音源の振幅が，目的音源と 1 つの干渉信号からなる信号の振幅よりも小さいと仮定する．これは，音源の統計的独立性を考慮することで有効な仮定であると言える．この仮定により，目的音源と 1 つの干渉音源からなる時間周波数点においても干渉音源を抑圧することができる（表 1 の 5,6 行目参照）．しかし，干渉音源 A，B が同時に存在する時間周波数点では，出力として干渉音 B もしくは A のどちらかを選択する必要がある（7 行目参照）．その出力は干渉音源 A，B の混合よりも小さくなるが，片方の音源は依然として抑圧されない．

最小値選択と時間周波数マスキングは似た点がある．時間周波数マスキングは，各時間周波数点の信号が目的音源かどうかを決定するマスクを作成する．従って，W-disjoint orthogonality (W-DO) の仮定が必要となる．一方で最小値選択は，どちらのビームフォーマがより良く干渉音源を抑圧するかを選択する．従って，最小値選択によるビームフォーマの組み合わせでは，目的音源と 1 つの干渉音源が存在する時間周波数点でも，すなわち W-DO が成り立たなくとも抑圧が可能である．以上より，最小値選択は従来の W-DO の仮定を必要とする時間周波数マスキングの拡張であると言える．また，時間周波数点毎にビームフォーマを切り替えていることから，最小値選択による組み合わせを改めて時間周波数スイッチングビームフォーマと呼ぶ．

3.2 時間周波数マスキングを用いた時間周波数スイッチングビームフォーマの拡張

最小値選択による音声強調はシンプルな組み合わせ方法でありながら，高い音声強調性能を示す．しかし，表 1 の 7 行目のように，複数の雑音が同時に存在する時間周波数点においては，全てを抑圧することができない．もしも，そのような時間周波数点に目的音源が存在しないのであれば，時間周波数マスキング同様に抑圧すべきである．

$$y_{\mathrm{DOA}}(\omega,t) = M(\omega,t)y_{\mathrm{MIN}}(\omega,t) \quad (7)$$

ここで，$M(\omega,t)$ はソフトマスクである．最小値選択の出力に時間周波数マスキングを適用することで，表 1 の 7 行目も抑圧が可能となる．なお，8 行目の抑圧にはより高度なソフトマスクの構成が必要となる．

$M(\omega,t)$ の構成のため，本稿では direction of arrival (DOA) 推定による音源のアクティビティ推定を行った．まず，[14] の手法を 1 マイクペアのみで行う．これにより時間周波数点毎に DOA 推定値が得られる．この推定は W-DO の仮定下で有効に働く．次に周波数ビン方向に平均を取る形で，時間フレーム毎の音源のアクティビティ推定を行う．最終的に，W-DO が十分に成り立っている時間周波数点では時間周波数点毎の，成り立っていない点では時間フレーム毎の音源アクティビティ推定に基づいてマスクの構成を行う．マスク構成の詳細は [15] を参照されたい．

4 評価実験

4.1 実験条件

提案手法の有効性を確認するため評価実験を行った．データベースとして community-based Signal Separation Evaluation Campaign (SiSEC) の underdetermined-speech and music mixtures task で提供されている dev1 の 3 話者のデータセットを利用した．データセットにはそれぞれ男性 3 名，女性 3 名の混合音が含まれており，各話者を目的音源として計 6 通りの音声強調を行い，その平均を結果として示す．その他の実験条件は表 2 に示す．

実験では，MVDR ビームフォーマ [11, 13] を従来のビームフォーマとして利用した．ビームフォーマの事前情報として目的音源区間と非目的音区間を与えた．MVDR ビームフォーマにおいては目的音源区間の空間相関行列に対して固有値分解を行い，最大固有値に対応する固有ベクトルを伝達関数の推定値として用いた．また，非目的音区間からは雑音共分散行列計算し，事前情報として用いた．

比較のための従来法として，MVDR ビームフォーマを単体で用いた劣決定音声強調である MVDR，また，2 チャネルの時間周波数マスキングとして degenerate unmixing estimation technique (DUET) [4] を用いた．更に，spatial subtraction array (SSA) [16] を 2 チャネルで実行した場合の性能も示す．ここで参照パスの推定（雑音の推定）には null beamformer や独立成分分析に基づく手法が提案されているが，本稿では MaxSNR ビームフォーマ [11,12] を用いた．なお，SSA で用いるパラメータは [16] と同一とした．また，位相の補償には Delay and sum ビームフォーマの出力値が用いられているが，本稿では MaxSNR ビームフォーマの出力値を用いている．上記に加え，以前提案したヴァーチャルマイクロホン（VM） [17] を用いた MaxSNR ビームフォーマも評価した．この手法では，実マイクロホンとヴァーチャルマイクロホンの両方を用いることで，劣決定条件を回避することができる．ヴァーチャルマイクロホン合成のパラメータとして $\alpha = 0.5$，$\beta = 2$ を用いた．

提案手法として，MVDR ビームフォーマを用いた時間周波数スイッチングビームフォーマと，時間周波数マスキングを用いたその拡張を検討し，以下ではそれぞれを TFS, TFS+TFM と省略する．各手法の計算のため，それぞれ干渉音源 A，B を抑圧するビームフォーマ A，B を事前に構成した．そのため，目的音源区間と 2 つの雑音それぞれの雑音源区間が必要となる．DOA 推定のパラメータは [15] と同一とした．

本手法の有効性の調査のため，表 1 の 1 列目に示す 7 パターンの音源の組み合わせに対して実験を行った．ここで，全ての音源は音声であるためスパースである．従って各時間周波数点においては，複数の音源からなる区間であっても常に同時に存在するとは限らない．評価尺度としては signal-to-distortion ratio (SDR)，signal-to-interference ratio (SIR)，signal-to-artifacts ratio (SAR) [18] を用いた．実験結果は 6 名の話者それぞれを目的話者として，3 話者が同時に存在する区間に対して音声強調を行った結果を平均して示す．

4.2 結果と考察

実験結果を図 2 に示した．従来の単一ビームフォーマはただ 1 つの妨害音のみ抑圧できるため，音声強調性能は低いが，SAR に示されるように人工的な雑音は生じにくい．提案手法は全ての評価尺度に置いて高い性能を示し，単一ビームフォーマだけでなく従来の時間周波数マスキング手法である DUET も上回る結果となった．特に TFS を時間周波数マスキングを用いて拡張することで，SIR に大きな向上が見られる．以上より，提案法は高い音声強調性能を示すと言える．SSA に関しては，特に音声認識の前段の処理として開発されており，位相情報を正しく保持しない．そのため全ての評価が低くなっている．なお，SSA が提案されている [16] などにおいては音声認識の単語誤り率で評価されている．

Table 2: 実験条件

マイクロフォンの数	2
マイクロフォンの間隔	5 cm
音源到来方向	40°, 80°, and 105°
残響時間	250 ms
サンプリング周波数	16 kHz
FFT フレーム長	2048 samples
FFT フレームシフト幅	512 samples
学習区間長	5 s
テスト区間長	35 s (5 s × 7)

図 3(a) に時間周波数スイッチングビームフォーマ（TFS）において，時間周波数点毎に選択されたビームフォーマの例を示す．これにより，各時間周波数点毎に 2 つのビームフォーマが頻繁に切り替わっている事がわかる．従って，2 つのビームフォーマ出力の位相もしくはゲインが異なる場合，時間周波数マスキングのように，ビームフォーマの切り替えによって歪みが生じてしまう．しかし，MVDR ビームフォーマは目的音源方向に対する制約をもつため，2 つのビームフォーマが位相とゲインの意味で全く同一の目的音声を出力する．従って切り替えによる歪みは生じず，高い SDR, SAR を示す．

図 3(b) では選択されたビームフォーマに加えて，時間周波数マスキングが適用された時間周波数点も示している（図中の赤点）．DOA 推定による拡張の最大の利点は，ソフトマスク適用による SIR の向上である．図中の i_A, i_B の区間に着目すると，低域ではビームフォーマ A(緑点) が，高域ではビームフォーマ B(青点) が選択されているが，大部分はマスキングされており，TFS では抑圧できない点においても効果的に雑音を抑圧している．従って，DOA 推定に基づく拡張は，雑音抑圧性能の向上に有効だと言える．

5 まとめ

本稿では，新たな劣決定音声強調手法として時間周波数スイッチングビームフォーマを提案し，ステレオマイクを用いた場合の性能を評価した．これは，事前に構成した複数のビームフォーマから，時間周波数点毎に最適なビームフォーマを選択する手法であり，特に MVDR ビームフォーマを用いることでひずみの少ない音声強調を達成する．また，本手法は時間周波数マスキングと併用することで，更に雑音抑圧性能を向上させることができる．時間周波数マスキングは，時間周波数点毎，及び時間フレーム事の音源アクティビティ推定に基づき構成した．両者の手法は共に W-DO の成立を必要とせず，従来の時間周波数マスキングの拡張であると言える．

本稿では，提案手法の有効性を確認するため，2 マイク 3 音源の環境（SiSEC）における提案法と従来法の

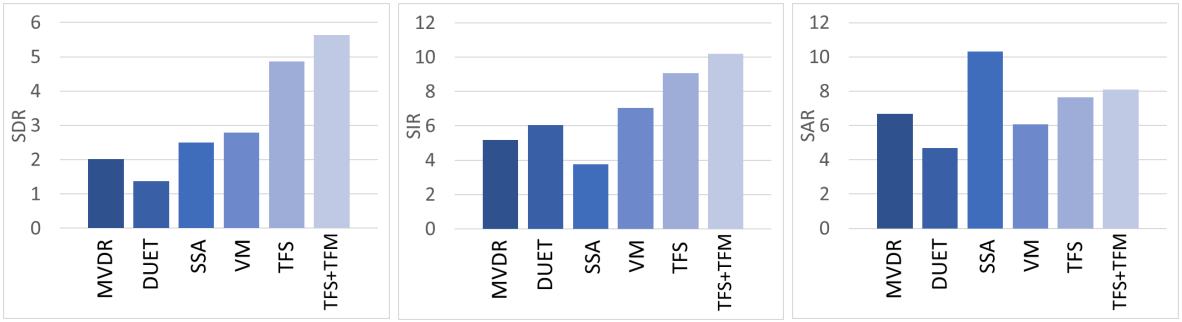
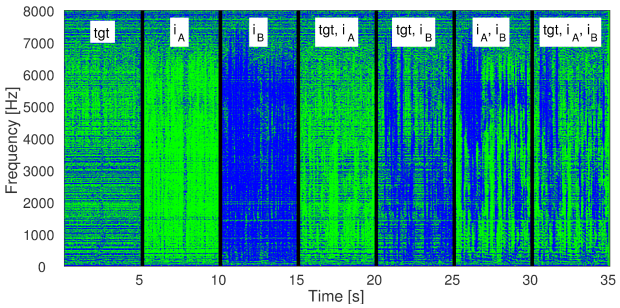
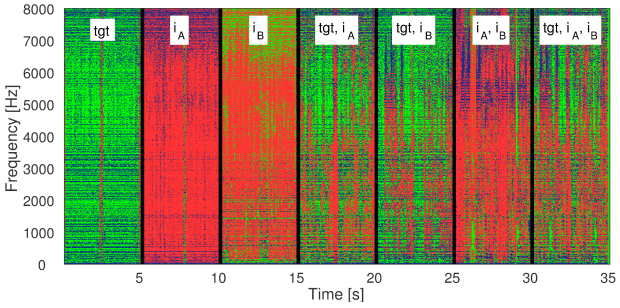


Fig. 2: Results of speech enhancement for the test period consisting of three speakers



(a)



(b)

Fig. 3: Selected beamformers in (a) TFS and (b) TFS+TFM for each test period. Green, beamformer A; blue, beamformer B; red, masked by (7)

性能評価を行った．実験により，時間周波数スイッチングビームフォーマが従来法と比較して高い音声強調性能を示した．また，時間周波数マスキングと併用することで雑音抑圧性能の向上に寄与することを示した．

謝辞

本研究は JSPS 科研費 16H01735, SECOM 科学技術振興財団, Tsukuba-DAAD Joint Research Program の助成を受けた．また，本研究に関して多くの助言をいただいた FAU Erlangen-Nürnberg の Andreas Brendel 氏，Michael Buerger 氏，Walter Kellermann 教授に謝意を表する．

参考文献

[1] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*, Springer, 2007.

[2] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 4, pp. 780–793, Apr. 2017.

[3] O. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[4] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. Springer, pp. 217–241, 2007.

[5] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516–527, 2010.

[6] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[7] N. Q. K. Duong, P. Berthet, S. Zabre, M. Kerdranvat, A. Ozerov, and L. Chevallier, "Audio zoom for smartphones based on multiple adaptive beamformers," *Proc. LVA/ICA*, pp. 121–130, Feb. 2017.

[8] S. Takada, S. Kanba, T. Ogawa, K. Akagiri, and T. Kobayashi, "Sound source separation using null-beamforming and spectral subtraction for mobile devices," *Proc. WASPAA*, pp. 30–33, 2007.

[9] T. Ogawa, S. Takada, K. Akagiri, and T. Kobayashi, "Speech enhancement using a square microphone array in the presence of directional and diffuse noise," *IEICE Trans. Fundamentals*, vol. E93-EA, no. 5, pp. 926–935, May 2010.

[10] M. Togami, T. Sumiyoshi, Y. Obuchi, Y. Kawaguchi, and H. Kokubo, "Beamforming array technique with clustered multichannel noise covariance matrix for mechanical noise reduction," *Proc. EUSIPCO*, Aug. 2010.

[11] H. L. Van Trees, *Optimum Array Processing*, John Wiley & Sons, 2002.

[12] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," *Proc. ICASSP*, vol. I, pp. 41–45, Apr. 2007.

[13] O. L. Frost Ⅲ, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[14] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 392–402, Feb. 2014.

[15] K. Yamaoka, A. Brendel, N. Ono, S. Makino, M. Buerger, T. Yamada, and W. Kellermann, "Time-frequency-bin-wise beamformer selection and masking for speech enhancement in underdetermined noisy scenarios," *Proc. EUSIPCO*, 2018.

[16] 高橋 祐, 猿渡洋, 鹿野清宏, "独立成分分析を導入した空間的サブトラクションアレーによるハンズフリー音声認識システムの開発," 電子情報通信学会誌 D, vol. J93-D, no. 3, pp. 312–325, 2010.

[17] H. Katahira, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Nonlinear speech enhancement by virtual increase of channels and maximum SNR beamformer," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–8, Jan. 2016.

[18] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.