SN比最大化ビームフォーマを用いたオンライン会議音声強調*

◎小笠原基 (NTT/名大), 石塚健太郎，荒木章子，藤本雅清，中谷智広，大塚和弘 (NTT)

1 はじめに

近年，多人数での会議や会話状況において，画像や音響信号を収録し，それらを統合することにより話者ダイアライゼーション (「いつ，誰が話したか」の推定) を行ったり，話者間のインタラクションを自動分析するといった研究が盛んである．例えば [1] では会議の分析結果から，キーワードによる会議場面検索を行える．我々はこれまで，マルチモーダル会話シーン分析システムを構築した [2]．このシステムでは，実時間会議分析に加え，会話場面とその分析の結果を三次元的に可視化し，会議をプレイバックすることが可能である．本稿の目的は会議をプレイバックする際に各人物に対応した強調音声の再生を可能にすることである．音声強調や，音源分離に関する研究はこれまでにも多数行われてきているが，会議状況のように，話者交代やオーバーラップ話者の変化が生じるような状況において各個人の音声を強調しようという取り組みはあまりない．我々はこれまで，観測信号に対してバッチ処理により音声強調を行うことを検討してきた [3]．本稿では，オンライン処理での音声強調手法を提案し，評価，システム実装を行ったので報告する．本手法は話者ダイアライゼーションの結果を利用し，SN 比最大化ビームフォーマ形成に必要な相関行列を適応的に算出するものである．シミュレーションデータに対する実験の結果，従来法に比べて提案法は良好な強調結果を得た．また実環境で動作するシステムにおいても，各話者の音声を良好に強調できることを確認した．なおシステム全体の詳細については，文献 [4] を参照していただきたい．

2 会議状況の定式化と話者ダイアライゼーション方法

本稿のタスクである会議状況の定式化を行う．N 人の音声信号 $\mathrm{s}_1, \ldots, \mathrm{s}_N$ を M 個のマイクロホンで受音する状況を考える．観測信号 $\mathrm{x}(t)$ は以下のように表される．

$$\mathrm{x}_j(t) = \sum_{k=1}^{N} \sum_{l} \mathrm{h}_{jk}(l) s_n(t-l) + \mathrm{n}_j(t), j = 1, \ldots, M \tag{1}$$

ここで $\mathrm{h}_{jk}(l)$ は音源 k からマイクロホン j へのインパルス応答であり，n_j はマイクロホン j で観測される雑音である．会議状況では音声信号は断続的に発声され，PC やプロジェクター等の雑音も存在する．また部屋の残響の影響も無視できない．さらに話者数は未知であり，音源数 > マイクロホン数となる Undetermined な問題となりうる状況である．このような問題設定において，観測信号 $\mathrm{x}_1, \ldots, \mathrm{x}_M$ のみから各話者の音声強調を行う．

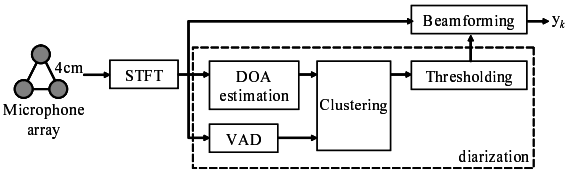

Fig. 1 話者ダイアライゼーション手法

これまで我々は，発話区間検出器 (voice activity detector :VAD) と音声到来方向 (direction of arrival :DOA) 推定を用いて話者ダイアライゼーションを行う手法を提案した [5]．この手法は，まず発話区間検出器 (VAD)[6] を用いて観測信号から雑音区間を除き，発話区間の観測信号に対して DOA 推定を行い，DOA をクラスタリングする．各クラスタが各発話者に対応し，この結果より発話者のダイアライゼーションを行うことができる (図 1)．

3 提案法

本稿では，話者ダイアライゼーションの結果を利用し，各話者に対応する音声を強調する．観測信号の時間周波数表現である $\mathbf{x}(f,\tau)$ に，話者 k の強調フィルタ $\mathbf{w}_k(f)$ を施すことにより強調音声 $y_k(f,\tau)$ を得る．

$$y_k(f,\tau) = \mathbf{w}_k^H(f)\mathbf{x}(f,\tau) \tag{2}$$

本稿ではこの強調フィルタ $\mathbf{w}(f)$ をオンラインで形成する手法を提案する．

3.1 DOA 推定手法および音声強調手法の比較

オンライン音声強調を行うための強調手法として，SN 比最大化ビームフォーマ (maxSNR beamformer)[7] と適応ビームフォーマ (adaptive beamformer)[8] の 2 種類の方法を比較検討する．また DOA 推定手法に関しても，GCC-PHAT 法 [9]，MUSIC 法 [10]，全ての時間周波数スロットで DOA を行う，TFDOA[11] の 3 種類の方法で比較検討を行う．

SN 比最大化ビームフォーマは出力信号 $y_k(f,t)$ 中の話者 k の区間信号と，雑音および他話者区間の信号のパワーの比を最大化するビームフォーマとして設計される．最大の SN 比 $\lambda(f)$ は，式 (3) で与えられる一般化固有値問題における最大固有値に対応する．

$$\mathbf{R}_{\mathbf{T}k}(f)\mathbf{w}_k(f) = \lambda(f)\mathbf{R}_{\mathbf{I}k}(f)\mathbf{w}_k(f) \tag{3}$$

ここでフレーム τ がクラスタ k に分類された時 $C(\tau) = k$ と書くことにすると，

$$\mathbf{R}_{\mathbf{T}k}(f) = E[\mathbf{x}(f,\tau)\mathbf{x}^H(f,\tau)]_{C(\tau)=k} \tag{4}$$

$$\mathbf{R}_{\mathbf{I}k}(f) = E[\mathbf{x}(f,\tau)\mathbf{x}^H(f,\tau)]_{C(\tau)\neq k} \tag{5}$$

---

*Online speech enhancement in a meeting situation with maximum SNR beamformers by OGASAWARA Motoki(NTT/Nagoya Univ.), ISHIZUKA Kentaro, ARAKI Shoko, FUJIMOTO Masakiyo, NAKATANI Tomohiro and OTSUKA Kazuhiro(NTT.)

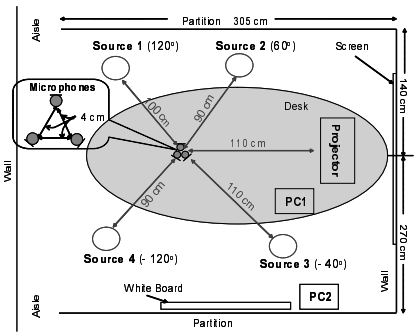

Fig. 2　実験環境 (残響時間 350ms.)

である．そして式 (3) 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 $\mathbf{w}_k(f) = \mathbf{e}(f)$ に対応する．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，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を最小にする補正フィルタにてビームフォーマ $\mathbf{w}_k(f)$ を補正する [3].

適応ビームフォーマは雑音の方向に適応的に死角を形成するものであり，以下の式で表される．

$$\mathbf{w}_k(f) = \frac{[\mathbf{R}_{\mathbf{I}k}(f)]^{-1}\mathbf{v}_k(f)}{\mathbf{v}_k^H(f)[\mathbf{R}_{\mathbf{I}k}(f)]^{-1}\mathbf{v}_k(f)} \tag{6}$$

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，本稿では DOA のクラスタリング結果で得られたセントロイドから $\mathbf{v}_k(f)$ を推定した．

GCC-PHAT 法は以下の式 (7) により，全てのマイクペア jj' に関して音声の到来時間差（time differences of arrival: TDOA) $q'_{jj'}(\tau)$ を求める手法である．

$$q'_{jj'}(\tau) = \mathrm{argmax}_{q'} \sum_f \frac{x_j(f,\tau)x^*_{j'}(f,\tau)}{|x_j(f,\tau)x^*_{j'}(f,\tau)|} e^{\jmath 2\pi f q'} \tag{7}$$

全てのマイクペアから得られた TDOA 値を並べたベクトル $\mathbf{q}'(\tau)$ と，マイク座標を表す行列 $\mathbf{D}$ より，DOA 値 θ を推定する．

MUSIC 法は次式で表される MUSIC スペクトルの値を最大にするステアリングベクトル $\mathbf{v}(\theta, f)$ の方向 θ を各周波数ごとに求める手法である．

$$P(\theta, f, \tau) = \frac{\mathbf{v}^H(\theta, f)\mathbf{v}(\theta, f)}{\mathbf{v}^H(\theta, f)\mathbf{R}_n\mathbf{R}_n^H\mathbf{v}(\theta, f)} \tag{8}$$

ここで $\mathbf{R}_n$ は観測信号の相関行列の雑音部分空間に属する固有ベクトルを並べた行列である．各周波数ごとに算出した推定値 θ の最頻値を時刻 τ における DOA 推定値とする．

TFDOA は DOA を全ての時間周波数スロットで計算する手法であり [11]，まず TDOA を時間周波数スロット毎に計算する．

$$q'_{jj'}(f,\tau) = \frac{1}{2\pi f} \arg\left[x_j(f,\tau)x^*_{j'}(f,\tau)\right] \tag{9}$$

次に TFDOA 値を GCC-PHAT と同様に求める．

以上の (ビームフォーマ 2 種類 {maxSNR,adaptive}) × (DOA3 種類 {GCC-PHAT,MUSIC,TFDOA}) の計 6 種類で性能評価実験 (実験 1) を行った．シミュレーションデータとして，話者 3 人が断続的に発話する状況をインパルス応答を用いて作成した．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 512 である．マイクロホン数は 3 本であり，その配置位置と話者の位置は図 2 に示す通りである．評価値には signal-to-interference plus noise-ratio (SINR) と signal-to-distortion-ratio (SDR) を用い，30 秒のシミュレーションデータ 10 通りでの平均値を算出した．実験結果を表 1 に示す．

Table 1　実験 1:ビームフォーマと DOA の性能比較 (SINR[dB] and SDR[dB])

	GCC-PHAT		MUSIC		TFDOA	
	SINR	SDR	SINR	SDR	SINR	SDR
maxSNR	9.0	13.1	8.4	12.3	9.3	13.7
adaptive	7.4	7.4	7.6	7.3	7.3	7.4

この結果より，DOA 手法にはさほど差異はないが，ビームフォーマに関しては SN 比最大化ビームフォーマが適応ビームフォーマに勝っているということが言える．これは適応ビームフォーマは，ステアリングベクトル推定の精度に性能が大きく左右されるため，ステアリングベクトルの推定精度が下がる雑音下や残響下ではあまり効果的ではないということが原因であると考えられる．一方 SN 比最大化ビームフォーマは，音源方向推定の精度が高くなくても，話者ダイアライゼーションが正確に行われれば高い性能が得られる．以上の結果より，本稿では以降，DOA 手法として演算量が比較的少ない GCC-PHAT を用い，強調方法には SN 比最大化ビームフォーマを用いたオンライン音声強調を行うことを検討する．

3.2　オンライン音声強調

従来までの音声強調は，収録したデータ全体に対して処理を行うバッチ処理が多く，マイクロホン数が話者数よりも多い場合は良好に動作する．しかし会議中にマイクロホン数よりも話者数が増えた場合には一般的に性能が劣化する．これは，SN 比最大化ビームフォーマはマイクロホン数 −1 つの死角を形成するという動作をするため，話者数がマイクロホン数以上になると雑音や他話者の方向にうまく死角を形成することができなくなるためである．このような状況に対応するため，データを例えば数秒 (5 秒) 程度のブロックに区切り，各ブロック毎にビームフォーマを形成する手法が考えられる．これは各ブロック内では話者数がマイクロホン数より少なくなることを期待するものであるが，数秒のような短いブロックではビームフォーマの推定精度が落ち，高い性能を得ることはできない．そこで本稿では，式 (3) における $\mathbf{R}_{\mathbf{T}k}(f)$, $\mathbf{R}_{\mathbf{I}k}(f)$ をオンラインで適応的に算出し，音声強調の性能を向上させる手法を提案する．これは，ブロック処理において現時点でのブロックのデータのみを用いて相関行列を算出するのではなく，過去のブロックのデータを適切に用いることでビームフォーマ形成のためのサンプル数を増やし，フィルタの精度を上げるものである．これを以後，適応的相関行列算出法と呼ぶ．また音声強調を行いたい目的話者が発話していないが，他話者が発話しているためにそれらの抑圧が必要となるブロックにおける処理手法も提案する．

音声強調処理全体の構成を図 3 に示し，ビームフォーマの形成に必要な係数を求めるフロー図を図 4 に

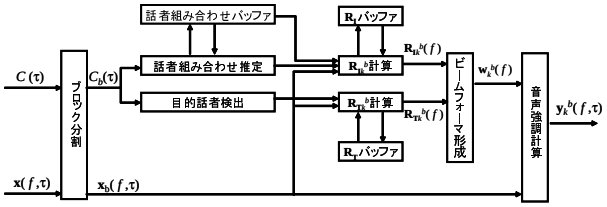

Fig. 3　音声強調処理全体の構成図

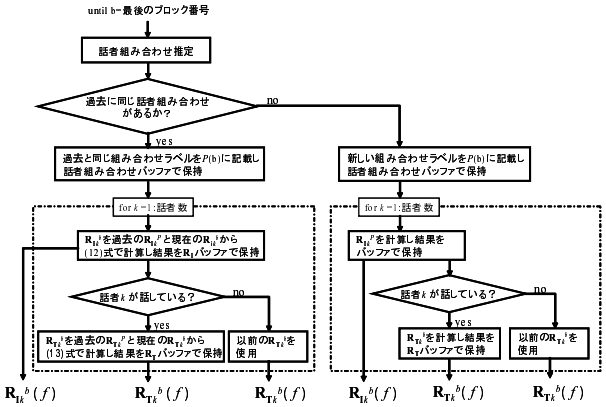

Fig. 4　ビームフォーマ形成に必要な係数を求めるためのフロー図

示す．また，図 5 に話者ダイアライゼーションの結果をブロック分割した例を示し，これらを用いて提案手法の処理を説明する．観測信号 $\mathbf{x}(f,\tau)$ とクラスタ情報 $C(\tau)$ はまず数秒ごとのブロックに分割され，話者組み合わせ推定と，目的話者検出を行う．ブロック分割された観測信号とクラスタ情報をそれぞれ $\mathbf{x}_b(f,\tau)$, $C_b(\tau)$ とする．以降の処理は全てのブロック b，全ての周波数 f において行う．まず話者組み合わせ推定を，クラスタ情報をもとに行い，ブロック b で発話している話者組み合わせ $P(b)$ を推定する．例えば図 5 の $b=1$ では話者 1,2,4 の組み合わせである．そして同じ話者組み合わせが過去にあったかを判定し，過去にあれば過去と同じ話者組み合わせラベルを $P(b)$ に記載し，なければ新しい話者組み合わせラベルを $P(b)$ に記載する．以降の処理は話者 k 毎に行う．まず同じ話者組み合わせが過去にない場合は，$\mathbf{R}_{\mathbf{T}k}$ と $\mathbf{R}_{\mathbf{I}k}$ をブロック内で式 (4)，(5) を用いて計算する．またそれぞれの結果をバッファに保持する．次に，同じ話者組み合わせが過去にあった場合は，まず目的話者 k 以外の観測信号の相関行列 $\mathbf{R}_{\mathbf{I}k}^b$ を式 (5) で計算する．そして以下の式で，$\mathbf{R}_{\mathbf{I}k}^b$ を更新する．

$$\mathbf{R}_{\mathbf{I}k}{}^b(f) \leftarrow \mathbf{R}_{\mathbf{I}k}{}^b(f) + \alpha \mathbf{R}_{\mathbf{I}k}{}^P(f) \qquad (10)$$

ここで α は忘却係数であり，$\mathbf{R}_{\mathbf{I}k}^P(f)$ はこれまでの同じ話者組み合わせブロックで計算された相関行列であり，$\mathbf{R}_{\mathbf{I}k}^P(f) \leftarrow \mathbf{R}_{\mathbf{I}k}^b(f)$ により算出する．その後，ブロック b で目的話者 k が発話しているかどうかを検出する．話者 k が発話している場合は $\mathbf{R}_{\mathbf{T}k}^b$ を式 (11) で計算する．

$$\mathbf{R}_{\mathbf{T}k}{}^b(f) \leftarrow \mathbf{R}_{\mathbf{T}k}{}^b(f) + \alpha \mathbf{R}_{\mathbf{T}k}{}^P(f) \qquad (11)$$

ここで $\mathbf{R}_{\mathbf{T}k}^P(f)$ は，同じ話者組み合わせのブロックの相関行列を用いて $\mathbf{R}_{\mathbf{T}k}^P(f) \leftarrow \mathbf{R}_{\mathbf{T}k}^b(f)$ により算出

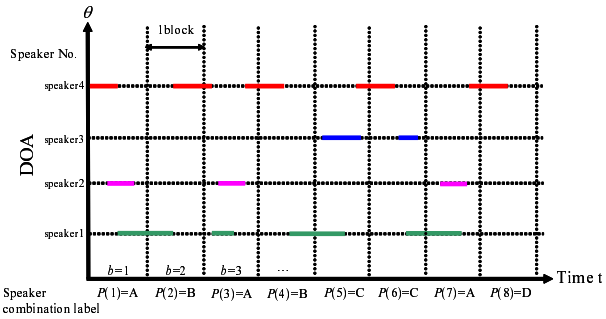

Fig. 5　話者ダイアライゼーションの結果をブロック分割した例

Table 2　実験 2:提案手法の評価実験の結果

	SINR[dB]	SDR[dB]
バッチ処理	5.1	14.1
ブロック処理 (各ブロックごとの処理)	6.5	6.7
提案手法 (適応的相関行列算出法)	7.3	7.7

する．また発話していない場合はこれまでで発話していたブロックの中で一番近いブロックの相関行列を用いる．なお，得られた $\mathbf{R}_{\mathbf{I}k}^b(f)$，$\mathbf{R}_{\mathbf{T}k}^b(f)$ ともにフレーム数で割るという正規化処理を行う．

ここまでの処理を図 5 で話者 2 の強調音声を出力することを例にあげて説明する．図 5 のブロック $P(4)$ =B では話者 2 は発話していない．しかし他話者が発話しているため，話者 2 に関する出力を得るには話者 2 以外の発話を抑圧する必要がある．そこで $\mathbf{R}_{\mathbf{I}2}^b$ を，$P(4)$ =B のブロックで得たものと，過去の同じ話者組み合わせである $P(2)$ =B ブロックで既に計算されている $\mathbf{R}_{\mathbf{I}2}^P(f)$ を用いて，式 (10) により計算する．また $\mathbf{R}_{\mathbf{T}2}^b(f)$ は直近の $P(3)$ =A ブロックで得られたものを用いる．以上の処理で得られた相関行列を用いて，式 (3) の一般化固有値問題を解き，ブロック $b=4$ についての SN 比最大化ビームフォーマを形成する．

4　実験と結果

提案手法の有用性を示すための評価実験 (実験 2) を行った．シミュレーションデータとして，話者 4 人が断続的に発話し，オーバーラップや話者交代の頻度が様々であるシミュレーションデータを 5 通り作成した．マイクロホン配置や話者の位置は図 2 と同様である．また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 512 であり，忘却係数 $\alpha=1$ とした．ブロック長は 150 フレーム (4.8 秒) とした．評価値には SINR と SDR を用いた．また比較のために，バッチ処理と，ブロック処理 (相関行列を各ブロック内のみで算出) でも求めた．実験結果を表 2 に示す．バッチ処理は，話者が 4 人であるために Undetermined な問題となり，うまく死角が形成できないために SINR は上がらなかったと考えられる．また提案手法は SINR，SDR 共にブロック長を固定した時よりも良い値となっており，適応的に相関行列を算出した方が良好な強調音声が得られることが示

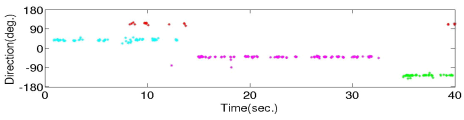

Fig. 6　話者ダイアライゼーションの結果 (実環境)

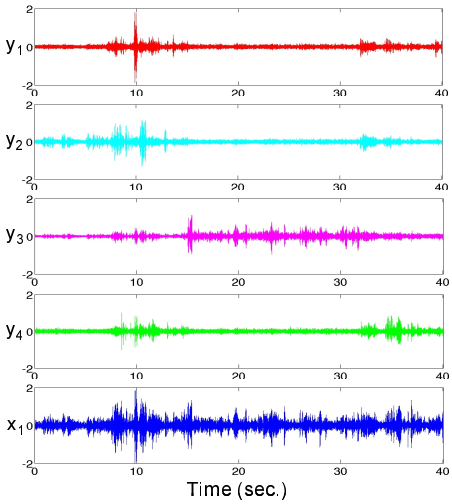

Fig. 7　強調した波形の例 (y1〜y4 が強調信号，x1 は観測信号)

された．しかし，バッチ処理の値に比べると SINR は上昇しているが，SDR はこの値には達してない．追実験として音源数を 3 とし，4.8 秒のシミュレーションデータに対して同様のブロック処理を行う実験で SDR を算出した場合でも，SDR は同様に約 7dB となることを確認した．この結果から，Undetermined な問題であることが原因で SDR が低下したのではなく，単純にサンプル数が足らなかったと考えられる．提案手法はブロック長を固定した場合に比べると SDR は上昇しているため，よりサンプル数が増えれば，SDR はバッチ処理の値まで達するであろうと考えられる．150 フレーム (4.8 秒) という短い時間から良好な歪み補正フィルタ [3] を形成することが本手法の今後の課題である．

また，実環境で動作するシステムを構築し，話者 4 人での会議タスクを収録した．この会議タスクでの話者ダイアライゼーションの結果と強調音声の例を図 6，7 にそれぞれ示す．波形 (図 7) からわかるように，各話者の音声が強調されており，聴感上も，実環境でも良好に動作することが確認された．またこの強調音声を [2] のプレイバックシステムに組み込むことで，あとから会議を見直したときに，各個人の強調音声を再生することができる．マウス操作で，ある話者の映像をズームにする (図 8 右) と，それに合わせてその話者に対応する強調音声が出力されるといったように，画面の人物配置に応じて音場を再構成できる．このシステムにより，より没入感の高い会議プレイバックシステムを構築することができた．

5　まとめと今後の課題

本稿では，会議状況におけるオンラインで動作する音声強調システムを構築した．観測信号をブロック

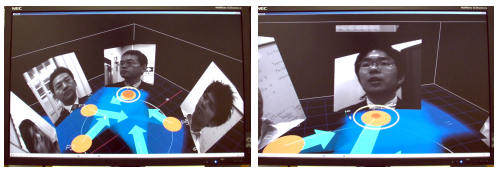

Fig. 8　会議プレイバックシステムで各人の強調音声を再生

ごとに区切り，話者ダイアライゼーション結果を利用して相関行列を適応的に算出する手法を提案した．評価実験の結果，バッチ処理やブロック処理に比べて SINR が上昇することを確認した．また実環境においても良好に動作することを確認した．今後の課題としては，本稿ではブロック長を 150 フレームに経験的に設定したが，これを適切に決定する方法を考える必要がある．

参考文献

[1] F. Asano et al., “Detection and separation of speech events in meeting recordings using a microphone array,” EURASIP Journal on Audio, Speech, and Music Processing, Volume 2007, Article ID 27616, 8 pages

[2] K. Otsuka et al., “A Realtime Multimodal System for Analyzing Group Meetings by Combining Face Pose tracking and Speaker Diarization,” *ICMI’08*, pp.257-264, 2008.

[3] 荒木 章子 他, “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,” 音講論 (春), pp.571-572, 2007.

[4] 石塚 健太郎 他, “音響情報と映像情報から得られる位置情報の統合による話者ダイアライゼーション,” 音講論 (春), 2009.

[5] 荒木 章子 他, “音声区間検出と方向情報を用いた会議音声話者識別システムとその評価,” 音講論 (春), pp.1-2, 2008.

[6] M. Fujimoto, et al., “A voice activity detection based on adaptive integration of multiple speech feature and signal decision scheme,” in *Proc. of ICASSP ’08*, 2008, pp. 4441–4444.

[7] H. L. Van Tree, ed., *Optimum Array Processing,* Wiley,2002.

[8] D. H. Johnson et al., *Array Signal Processing*, Prentice-Hall, Englewodd Cliffs, NJ, USA, 1993.

[9] C. H. Knapp et al., “The generalized correlation method for estimation of time delay,” *IEEE Trans. ASSP*, vol.24, no.4, pp. 320-327, 1976.

[10] R. O. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Transactions on Antennas and Propagation*, vol.34, no.3, pp.276-280, 1986.

[11] S. Araki et al., “DOA estimation for multiple sparse sources with normalized observation vector clustering,” in *Proc. of ICASSP’06*, 2006, vol. 5, pp. 33–36.