Vol.2018-MUS-119 No.10
Vol.2018-SLP-122 No.10
2018/6/17

実環境音響信号処理における収音技術

小泉 悠馬[1,a)]

概要：雑音下で観測された信号から目的音を強調するため収音技術は，実環境で音響信号処理技術を頑健に動作させるために必要不可欠である．特に，音響信号処理技術の実用化では，アプリケーションによって雑音抑圧量や計算コストなど性能要件が異なり，また目的音と雑音を区別できる特徴も異なる．ゆえに実環境で頑健に音源強調を動かすためには，(i) 目的音と雑音の違いに適切な仮定をたて，(ii) 後段の音響信号処理の精度を高めるための必要十分な性能要件を明確にして，適切な音源強調法を利用することが重要である．本稿では，著者の所属する研究所で，近年の音声・音響信号処理の実用化に向けて研究開発した収音技術を概説する．特に，目的音と雑音のどのような違いに着目するかに重点を置き，方向，位置，音色の違いを利用した収音技術を紹介する．

Sound-Source Enhancement Techniques in Real-Environments

Yuma Koizumi[1,a)]

Abstract: Sound-source enhancement is a signal processing technique to enhance the target source from the observation signals under a noisy condition, and indispensable as a front-end processing for robust acoustic signal processing in a real environment. Especially, in a practical application of acoustic signal processing, performance requirements of sound-source enhancement depends on the application. Therefore, to implement an application of acoustic signal processing, it is important to (i) make an appropriate assumption for the difference between the target and noise, and (ii) clarify the sufficient performance requirements such as noise reduction level and calculation cost. This paper introduces sound-source enhancement methods developed for practical application of acoustic signal processing. In particular, by focusing on the difference between the characteristics of target and noise, we introduce three patterns of sound-source enhancement strategy using the difference in direction, position, and spectrum.

1. はじめに

現在，様々な音声・音響信号処理技術が実用化され，我々の生活に欠かせないものとなっている．例えば，音声通信技術 [1, 2] は遠隔地とのコミュニケーションに欠かせない手段であるし，近年では “しゃべってコンシェル [3]” などの音声アシスタント技術や，Dolby Atmos [4] や多チャンネル符号化技術 [5–7] などの高臨場音響技術，機械の故障を検知する異常音検知技術 [8, 9] なども実用化されている．

実環境でマイクロホンを用いて音を観測すると，信号処理をしたい目的音の他に，周囲の人の声，音楽，機器騒音などの雑音が混入する．雑音は信号処理の性能を劣化させる要因の一つのため，実環境で音響信号処理技術を利用するためには，マイクロホンで観測した信号から所望の目的音を抽出する収音技術 [10–13] が必要不可欠である．

本稿では，著者の所属する NTT 研究所において，近年の音声・音響信号処理の実用化に向けて研究開発した収音技術を，ユースケースを交えながら紹介する．なお，技術の詳細は適宜引用する文献に任せ，本稿では各問題の難しさと，何故その技術を利用したかに重点をおいて説明する．また収音技術の小分類には，信号処理の性能を高めるためのハードウェア設計技術 [14–16] や，観測音を有限個の音源に分類する “音源分離 (source separation)”，目的音を強調する “音源強調 (source enhancement)”，雑音の特性に着目して雑音を抑圧する “雑音抑圧 (noise reduction)” など様々だが，本稿ではソフトウェア技術の紹介に絞り，また用語の統一のために収音技術を “音源強調” と呼ぶ．

[1] NTT メディアインテリジェンス研究所
3-9-11, Midori-cho, Musashino-shi, Tokyo 180–8585, Japan
[a)] koizumi.yuma@lab.ntt.co.jp

Vol.2018-MUS-119 No.10
Vol.2018-SLP-122 No.10
2018/6/17

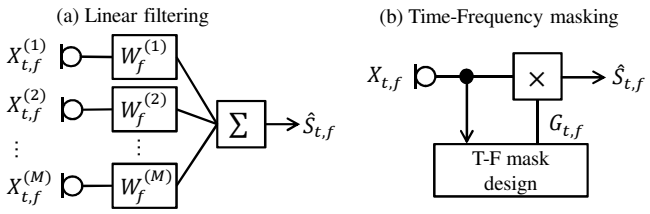

図 1 フィルタリングの一般形．(a) 線形フィルタリング (b) 時間周波数マスキング．

2. 音源強調の問題設定

M 本のマイクロホンを利用して，目的音 $S_{t,f}^{(1)} \in \mathbb{C}$ を収音することを考える．いま，m 番目のマイクロホンの観測 $X_{t,f}^{(m)} \in \mathbb{C}$ を以下のように記述する．

$$X_{t,f}^{(m)} = H_f^{(m,1)} S_{t,f}^{(1)} + \sum_{k=2}^{K} H_f^{(m,k)} S_{t,f}^{(k)} \tag{1}$$

ここで t と f はそれぞれ時間と周波数のインデックス，$S_{t,f}^{(k)} \in \mathbb{C}, (k = 2, ..., K)$ は雑音，$H_f^{(m,k)}$ は k 番目の音源から m 番目のマイクロホンまでの伝達特性である．以降の議論では，特にことわりのない限り，表記の簡単のために式 (1) の第二項を $N_{t,f}^{(m)}$ とまとめて以下のように表記する．

$$X_{t,f}^{(m)} = H_f^{(m,1)} S_{t,f}^{(1)} + N_{t,f}^{(m)} \tag{2}$$

式 (2) をベクトル形式で記述すると以下のようになる．

$$\boldsymbol{x}_{t,f} = \boldsymbol{h}_f S_{t,f}^{(1)} + \boldsymbol{n}_{t,f} \tag{3}$$

ここで $\boldsymbol{x}_{t,f} := (X_{t,f}^{(1)}, ..., X_{t,f}^{(M)})$，$\boldsymbol{n}_{t,f} := (N_{t,f}^{(1)}, ..., N_{t,f}^{(M)})$，$\boldsymbol{h}_f := (H_f^{(1,1)}, ..., H_f^{(M,1)})$ である．

音源強調は，観測信号から目的音の推定値 $\hat{S}_{t,f} \in \mathbb{C}$ を出力する信号処理のことを指す．実環境において利用される音源強調のほとんどは，フィルタリングに基づく手法である．音源強調におけるフィルタリングは，ビームフォーミング [10, 12] や独立成分分析 [17] などに代表される線形フィルタリング（図 1 (a)）と，ウィナーフィルタ [18] などに代表される時間周波数マスキング（図 1 (b)）に大別できる．周波数領域では，線形フィルタリングは観測信号の複素線形結合として記述できる．

$$\hat{S}_{t,f} = \sum_{m=1}^{M} W_f^{(m)*} X_{t,f}^{(m)} \tag{4}$$

ここで $W_f^{(m)} \in \mathbb{C}$ は線形フィルタの係数，上付きの $*$ は複素共役を表す．一方で時間周波数マスキングは，単一チャネル（*i.e.* $M = 1$）の観測信号のゲイン調整として以下のように記述される．

$$\hat{S}_{t,f} = G_{t,f} X_{t,f}^{(1)} \tag{5}$$

ここで $G_{t,f}$ は時間周波数マスクである．多くの場合，時間周波数マスクは $0 \le G_{t,f} \le 1$ に制限される．

では，フィルタ係数である $W_f^{(m)}$ や $G_{t,f}$ はどのように設計すべきだろうか．残念ながら，著者の知る限り，全ての環境かつ全ての目的音に安定かつ有効に動作する音源強調法は存在しない．実応用によって，求められる/求められない性能要件は異なるし，また各音源強調法には雑音抑圧量，目的音の歪みやすさ，計算コストなど様々なトレードオフが存在する．さらに，方向，位置，音色など目的音と雑音を区別できる特徴は解きたい問題に依存する．ゆえに実環境で頑健に音源強調を動かすためには，(i) 目的音と雑音の違いに適切な仮定をたて，(ii) 設置位置や各種コストの制限など，後段の音情報処理技術の精度を高めるための必要十分な性能要件を明確し，適切な手法を選択 · 利用することが重要である．

以降の章では，目的音と雑音の方向（3 章），位置（4 章），音色（5 章）の違いに着目した音源強調法を紹介する．また，音色の違いに着目した音源強調の近年の発展として，深層学習に基づく手法を 6 章で紹介する．

3. 音源方向に着目した音源強調

実環境で広く用いられる代表的な音源強調法は，目的音と雑音の方向の違いに着目するものである．例えば，小さな会議室向けの電話会議装置などでは，テーブルの中心に配置した装置を取り囲むように発話者が座ることが仮定できる．また，音声対話ロボットでは発話者はロボットの正面に立つことが仮定できる．これらの状況では，マイクロホンから見た音波の到来方向が目的音と雑音で異なるため，密配置した複数のマイクロホン（マイクロホンアレー）の観測信号間では振幅や位相に微小な差が生じる．この差を手掛かりに線形フィルタを設計する手法が，音源方向に着目した音源強調である．

代表的なフィルタ設計法には，音波の空間伝達の物理現象からフィルタ設計するビームフォーミング [10, 12] や，分離後の信号の統計的性質からフィルタ設計する独立成分分析 [17]，独立ベクトル分析 [19, 20]，独立低ランク行列分析 [21] などがある．前者は物理モデルから決定論的にフィルタが求まるため，フィルタは事前に設計でき，低演算量でのリアルタイム処理ができる．一方，後者はある時間区間の統計的性質から繰り返し最適化によってフィルタを求めるため，分離精度は高いもののバッチ処理が必要である．このような性質から，実応用では前者の方法でフィルタ設計することが多い．

ビームフォーマの設計法には様々な方法が提案されているが，その詳細は専門書 [10, 12] や各文献に譲り，ここでは一例として最尤法に基づく設計法を載せる．

$$\boldsymbol{w}_f^{\mathrm{ML}} = \frac{\Phi_{f,N}^{-1} \boldsymbol{h}_f}{\boldsymbol{h}_f^{\mathsf{H}} \Phi_{f,N}^{-1} \boldsymbol{h}_f} \tag{6}$$

ただし $\boldsymbol{w}_f := (W_f^{(1)}, ..., W_f^{(M)})$，$\Phi_{f,N} = \mathbb{E}[\boldsymbol{n}_{t,f} \boldsymbol{n}_{t,f}^{\mathsf{H}}]_t$ であり，

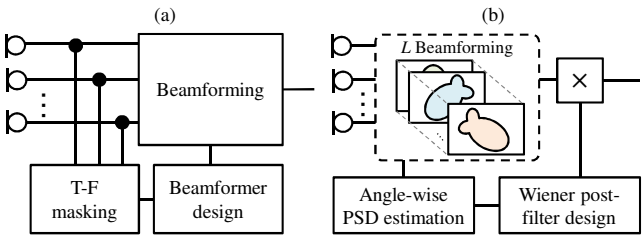

図 2 ビームフォーミングの性能を向上させるための手法例．(a) 時間周波数マスクを利用したビームフォーマ設計，(b) 時間周波数マスクを利用したポストフィルタリング．

$\mathbb{E}[\cdot]_x$ は x に関する期待値演算，上付きの H は複素共役転置である．式 (6) から分かるよう，ビームフォーマの設計には，目的音までからマイクロホンの伝達関数や空間相関行列などの音源の空間情報が既知である必要がある．しかし，これら空間情報は利用環境によって異なり，ほとんどの実応用では未知である．そこで最も簡便には，伝達関数を方向毎の時間遅延だけを表現するステアリングベクトルで代用したり，雑音の空間相関行列を単位行列で代用したりする．しかしこの方法では，音源強調の性能が劣化することが知られている[*1]．この問題を解決する手法として，次節以降では，ビームフォーマの設計制度を向上させる方法（図 2 (a)）と，後段に時間周波数マスクを組み合わせる方法（図 2 (b)）を紹介する．

3.1 時間周波数マスクを利用したビームフォーマ設計

音声認識などの識別タスクでは，音源強調により目的音に歪みが生じると認識精度が低下することが知られている．そのため，比較的歪みの生じにくい線形フィルタリングを利用して音源強調したい．本節では，線形フィルタの精度を向上させるために，時間周波数マスクを利用して目的音や雑音の空間情報を推定する手法を概説する（図 2 (a)）．

空間相関行列の定義から，観測信号から目的音や雑音の複素スペクトルが推定できれば，空間相関行列は推定可能である．そこで荒木らは，観測信号から目的音区間検出を行い，目的音区間/雑音区間から目的音/雑音の空間相関行列を推定する手法を提案した [23]．この改良として，目的音や雑音を強調する時間周波数マスクを推定し，以下のように空間相関行列を推定する手法も提案されている [22,24]．

$$\hat{\Phi}_{f,S} = \frac{1}{T}\sum_{t=1}\mathbf{x}_{t,f}\mathbf{x}_{t,f}^{\mathsf{H}} - \Phi_{f,N} \tag{7}$$

$$\hat{\Phi}_{f,N} = \frac{1}{\sum_{t=1}^{T}(1-G_{t,f})}\sum_{t=1}(1-G_{t,f})\mathbf{x}_{t,f}\mathbf{x}_{t,f}^{\mathsf{H}} \tag{8}$$

この方法で最小分散無歪みビームフォーマ [25] を求める手法は，音声認識の国際技術評価コンペティションである CHiME-3 [26] において最高精度を達成したシステムの音源強調処理として利用されており [27,28]，実環境音声認識でもフロントエンドとして頑健に動作することが知られている [29]．また近年では，時間周波数マスクの推定に DNN (deep neural network) を利用してビームフォーマを設計し，音声認識率を最大化するように時間周波数マスク推定用の DNN を学習する方式も検討されている [30]．

[*1] 時間差のみを考慮する遅延和ビームフォーマと等価になる [12]．

3.2 時間周波数マスクを利用したポストフィルタリング

通話などの用途では，発話内容を人間が聞き取ることが目的であるため，目的音を多少歪ませてでも大きく雑音を抑圧したい．この達成のために，図 2 (b) のように，線形フィルタの後段で時間周波数マスキングを利用する手法が古くから検討されている [31–33]．本節ではこの枠組みの一例として，局所 PSD (power spectral density) 推定法 [33] を簡単に紹介する．なお，厳密な理論の議論やアプリケーション例は先行文献 [33,34] を参照されたい．

以降の説明では，式 (1) による観測信号の定義を利用する．今，異なる方向に対して焦点を形成した L 個のビームフォーマ $W_f^{(l,m)}$ を観測信号に適用する．各音源が互いに無相関であると仮定すると l 番目のビームフォーマの出力 $Y_{t,f}^{(l)}$ のパワースペクトルは近似的に以下のように記述できる．

$$\underbrace{\begin{bmatrix}|Y_{t,f}^{(1)}|^2\\ \vdots \\ |Y_{t,f}^{(L)}|^2\end{bmatrix}}_{\boldsymbol{\Psi}_{Y,t,f}} \approx \underbrace{\begin{bmatrix}|D_f^{(1,1)}|^2 & \cdots & |D_f^{(1,K)}|^2\\ \vdots & \ddots & \vdots \\ |D_f^{(L,1)}|^2 & \cdots & |D_f^{(L,K)}|^2\end{bmatrix}}_{\boldsymbol{D}_f}\underbrace{\begin{bmatrix}|S_{t,f}^{(1)}|^2\\ \vdots \\ |S_{t,f}^{(K)}|^2\end{bmatrix}}_{\boldsymbol{\Psi}_{S,t,f}} \tag{9}$$

ここで，$D_f^{l,k} = \sum_{m=1}^{M} W_f^{(l,m)*} H_f^{(m,k)}$ である．すると，各音源のパワースペクトルは，L 個のビームフォーマの出力と $\boldsymbol{D}_f$ の一般化逆行列 $\boldsymbol{D}_f^{+}$ を利用して以下のように記述できる．

$$\boldsymbol{\Psi}_{S,t,f} = \boldsymbol{D}_f^{+}\boldsymbol{\Psi}_{Y,t,f} \tag{10}$$

これにより各音源のパワースペクトルを得られたので，以下のウィナーフィルタを設計し，目的音方向のビームフォーマの出力に式 (5) の形で乗ずることで出力音を得る．

$$G_{t,f} = \frac{|S_{t,f}^{(1)}|^2}{|S_{t,f}^{(1)}|^2 + \sum_{k=2}^{K}|S_{t,f}^{(k)}|^2} \tag{11}$$

なお，$H_f^{(m,k)}$ が未知の場合は，ステアリングベクトルを代用して $D_f^{l,k}$ を求める．また，$\boldsymbol{D}_f^{+}$ を安定に解くための条件として，$\boldsymbol{D}_f$ が M 行列となるように L 個のビームフォーマを設計する必要がある [35]．

4. 音源位置に着目した音源強調

前章では，小さい会議室などの小規模な空間での音源強調について議論した．本章では，より大規模な空間での音源強調について議論する．例えば図 3 (a) のような縦長の会議室で，長机に向かい合わせで目的話者（目的音）と妨害話者（雑音）が座っている状況を想定してほしい．一般に，目的音と雑音の距離が同じの場合，マイクロホンから

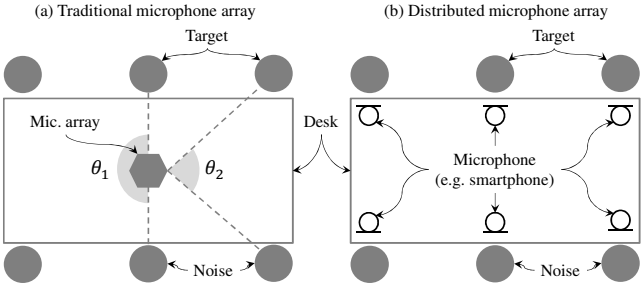

図 3 (a) 密配置したマイクロホンアレーと (b) 分散マイクロホンアレーのマイク配置の違い．

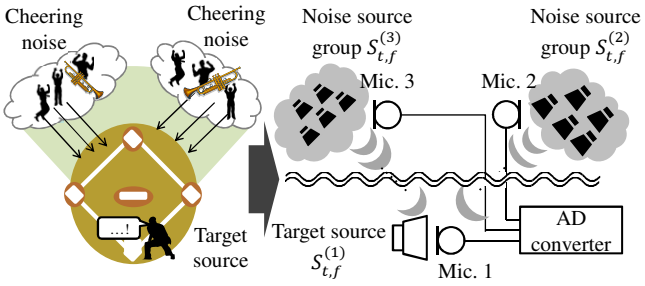

図 4 野球場における分散マイクロホンアレー．ホームベース付近のパラボラマイクと，外野スタンドの収音マイクを利用する．

音源までの距離が遠くなるほど，マイクロホンからみた音源同士の見込み角は小さくなる (*i.e.* $\theta_1 > \theta_2$ が成り立つ)．ゆえに，音源間の距離は同じであったとしても，密配置したマイクロホンアレーの近傍で向かい合わせた場合より，長机の端で向かい合わせた場合の方が音源強調が難しくなる．これが，大規模な空間で遠方に存在する目的音の強調を困難にする理由の一つである．

遠方の目的音を強調する方法として，チャネル数を増やしたり，特殊なハードウェアを利用したりすることで指向性の鋭いビームフォーマを設計する方法が知られている [14, 15]．しかしこの方法は，装置が大型化したり，高価な A/D 変換機が必要になったりするため，適用できる実環境は限られる．

一方近年では，従来のように複数のマイクロホンを密配置するのではなく，図 3 (b) のように，各音源の近くにマイクロホンを分散配置して連携させる “分散マイクロホンアレー” に関する研究が行われている [36–44]．音量は距離の二乗に比例して減衰するため，各音源は近傍に配置されたマイクロホンには大きく収音され，遠方に配置されたマイクロホンには小さく収音される．ゆえに，各マイクロホン間の音量差の情報を利用することで，音源の位置を利用した音源強調が可能になる．

分散マイクロホンアレーには，大きく分けて 2 つの解くべき問題が存在する．一つ目は，スマートホンなどを同期録音されていないマイクロホンとして利用した場合，サンプリング周期が完全一致しない場合がある．ゆえに線形フィルタリングなどの位相スペクトルに依存する音源強調を行うためには，これによって生じる時間ドリフトを推定 · 補正する必要がある [38, 39]．二つ目は，音源から各マイクロホンまでの伝達関数（もしくはそのゲイン）の推定である [40–43]．本稿ではスペースの関係上，二つ目の伝達関数推定に焦点をあてた手法を概説する．

4.1 瞬時混合を仮定した時間周波数マスク設計

本節では，伝達関数推定の手法として，Kako らが提案した方法を紹介する [41]．この方法では，マイクロホンとしてスマートホンを利用し，各スマートホン間のパワースペクトルのゲイン差から時間周波数マスクを推定して音源強調する．今，各音源は無相関であると仮定し，M 個のスマートホンで観測した信号 $X^{(m)}_{t,f}$ にパワースペクトル領域で加法性が成り立つとする．すると，式 (1) は以下のように記述できる．

$$|X^{(m)}_{t,f}|^2 \approx a^{(m,1)}_f \cdot |S^{(1)}_{t,f}|^2 + \sum_{k=2}^{K} a^{(m,k)}_f \cdot |S^{(k)}_{t,f}|^2 \qquad (12)$$

ここで $a^{(m,k)}_f$ は伝達特性のゲイン値 $|H^{(m,k)}_f|^2$ であり，“感度” と呼ぶ．式 (12) は，以下のような行列形式に書き換えることができる．

$$\underbrace{\begin{bmatrix} |X^{(1)}_{t,f}|^2 \\ \vdots \\ |X^{(M)}_{t,f}|^2 \end{bmatrix}}_{\boldsymbol{\Psi}_{X,t,f}} \approx \underbrace{\begin{bmatrix} a^{(1,1)}_f & \cdots & a^{(1,K)}_f \\ \vdots & \ddots & \vdots \\ a^{(M,1)}_f & \cdots & a^{(M,K)}_f \end{bmatrix}}_{\boldsymbol{A}_f} \underbrace{\begin{bmatrix} |S^{(1)}_{t,f}|^2 \\ \vdots \\ |S^{(K)}_{t,f}|^2 \end{bmatrix}}_{\boldsymbol{\Psi}_{S,t,f}} \qquad (13)$$

上記の式も，式 (9) と同様に，感度行列 $\boldsymbol{A}_f$ が求まれば，擬似逆行列から $\boldsymbol{\Psi}_{S,t,f}$ を求めることができる．ここで $\boldsymbol{A}_f$ の各列が，各音源から各マイクロホンへの感度の比率を表すことに着目すると，話者が一人の区間で各マイクロホン間のゲイン比を求めることで $\boldsymbol{A}_f$ を推定することができる．そこでこの手法では，観測信号から発話者が一人の区間を識別し，その区間から感度行列を推定している．

4.2 広域に分散配置したマイクロホンを利用した時間周波数マスク設計

分散マイクロホンアレーを用いた多くの手法では，各音源の瞬時混合を仮定している．これは，音源からマイクロホンまでの到達時間と残響時間が，短時間フーリエ変換（STFT: short-time Fourier transform）長よりも短いことを仮定している．しかし，より大規模な空間では，この仮定は成り立たないことがある．例えば図 4 のような野球場では，目的音はバッティング音や審判の声などの競技音であり，主な雑音は外野スタンドの応援団である．そして，バックネット裏にあるマイクロホンと，外野スタンドにあるマイクロホンを連携させて歓声を抑圧しようとしたとき，外野スタンドからバックネットまでは，100 m 以上の距離があるため，瞬時混合は成り立たない．

Vol.2018-MUS-119 No.10
Vol.2018-SLP-122 No.10
2018/6/17

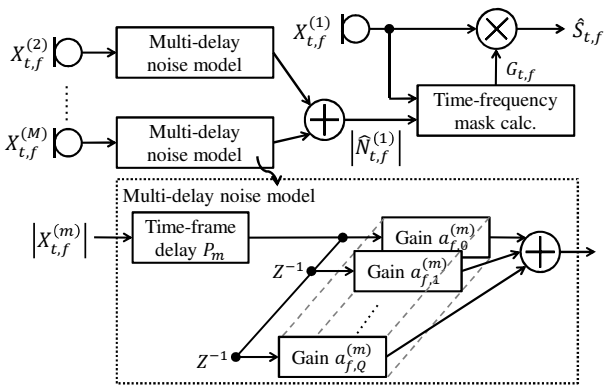

図 5 多重遅延ノイズモデル

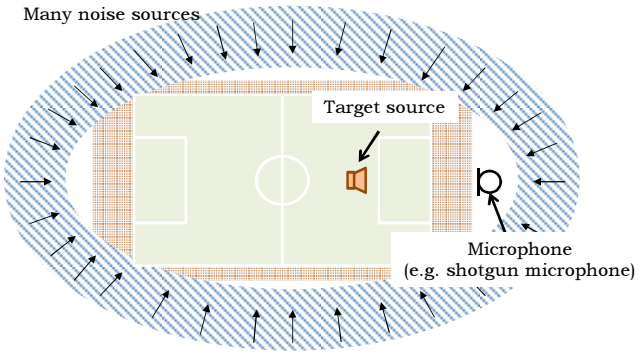

図 6 サッカー場での収音例．全方位から歓声雑音が到来し，かつ競技音の近くにマイクロホンが配置できない．

そこで我々は，多重遅延分割フィルタ [45] の考え方を応用し，到達遅延をフレームシフト P_m，残響を振幅領域の周波数領域での畳み込み係数（伝達ゲイン）$a_{f,q}^{(m)}$ で表現する多重遅延ノイズモデル [43] を提案した（図 5）．

$$|\hat{N}_{t,f}^{(1)}| \approx \sum_{m=2}^{M} \sum_{q=0}^{Q} a_{f,q}^{(m)} |X_{t-P_m-q,f}^{(m)}| \qquad (14)$$

ここで到達遅延 P_m はマイクロホン間の距離から概算が可能である．また，伝達ゲインは非負の実数かつ時間方向に指数減衰することが知られている [46]．そこで我々は，これらの物理的な事前知識を事前分布においた事後確率最大化推定でパラメータ推定する方式を提案している [43]．

5. 音色に着目した音源強調

前章までは，音源の空間的な差異を利用した音源強調を紹介した．本章では，音源の信号的な差異を利用した音源強調を紹介する．臨場感の高いスポーツ中継を実現するために，サッカー場のボールのキック音などの競技音をクリアに強調したい．もしこのような技術が実現し，各競技音をオブジェクトとして保持できれば，多数のスピーカを利用して音源ごとに定位を操作することで，競技場に潜り込んだような音響空間を生成できるだろう[*2]．しかし，例えばサッカー場などでは，図 6 のように全方位から歓声雑音が到来し，かつ目的音付近にマイクロホンを配置することが困難なので，音源の空間的な情報を利用して音源強調することはほとんど困難である．

ところで，キック音やタックル音など競技音は突発音が多いが，歓声などの競技場における雑音は定常的なものが多い．こういった，スペクトル形状やスペクトルの時間変化など，音色の情報を元に時間周波数マスクを設計する研究もまた，古くから取り組まれている．古くは，観測信号に含まれる定常的な音を雑音として推定するスペクトルサブトラクション法 [47] が有名である．より時間変化が複雑な雑音を分離するために，非負値行列因子分解 [48] を応用して観測振幅スペクトログラムを高々数個のスペクトルテンプレートに分解し，目的音を表現するスペクトルテンプレートだけを利用して信号を復元する音源強調法が検討されている [49]．

一方で，スポーツの生中継などの放送用途ではリアルタイム性が要求されるため，ルールベースで時間周波数マスクを設計する方法も検討されている [50, 51]．我々は，人間の耳では知覚できないほどのごく短い遅延時間で，突発的な競技音を強調する方式を研究している [52]．この方式では，48kHz サンプリングの信号に対し，STFT 長 128 点，シフト長 32 点 で突発音を検知し，スペクトル形状から突発音がキック音か否かを識別し，キック音であればそれを強調する時間周波数マスクを設計している．

上記の手法は，観測信号から時間周波数マスクを推定する回帰関数のルールベース設計と捉えることもできる．回帰関数をより簡便に設計するには，統計的機械学習に基づくアプローチが一般的である．我々は，より簡便に多数の種類の競技音を強調するために，統計的機械学習に基づき回帰関数を設計する研究 [53,54] や，回帰関数に入力すべき音響特徴量の性質を明らかにする研究 [55] も行っている．

6. 深層学習を利用した音源強調

音色に着目した音源強調の近年の発展として，深層学習 [56] の適用が挙げられる [57]．前章で述べた通り，統計的機械学習に基づく音色に着目した音源強調は，観測信号から時間周波数マスクを推定する回帰関数の最適化問題とも捉えることができる．この回帰関数として，ニューラルネットワークを利用する取り組みは古くから行われてきた [58–60]．しかし，ニューラルネットワークで複雑な回帰関数を学習するためには大規模なデータセットや膨大な計算時間が必要であり，実用には至らなかった．近年の計算機や最適化アルゴリズムの発達に伴い上述の問題が解決され，Narayanan ら [61] によって音声認識向けの音源強調として利用されたことを皮切りに，DNN を利用した音源強調が広く取り組まれるようになった [62–68]．

本章では，図 7 に示すような，$M = 1$ の DNN 音源強調を

[*2] この概念は Object-base audio [5] として知られており，Dolby Atomos [4] など映画の音響技術として利用されている．

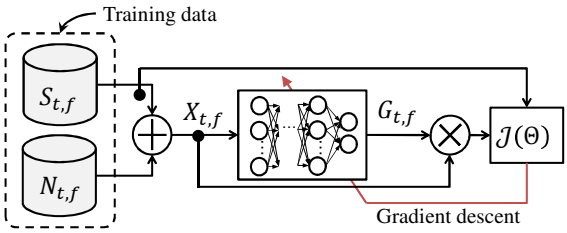

図 7 一般的な DNN 音源強調の学習フロー．目的音と雑音の学習データを用いて，式 (2) で観測信号をシミュレーションし，目的音と出力音の間で計算される目的関数を最小化/最大化するように DNN を更新する．

概説する．また，DNN を用いて推定する変数は様々なものがあるが，本章では時間周波数マスクを $\boldsymbol{G}_t = (G_{t,1}, ..., G_{t,F})$ と並べたベクトルを推定する手法を紹介する [62–64].

$$\boldsymbol{G}_t = \mathcal{M}(\boldsymbol{\zeta}_t|\Theta) \tag{15}$$

ここで $\boldsymbol{\zeta}_t$ は学習データから抽出されたフレーム t の音響特徴量[*3]，$\mathcal{M}$ は DNN や LSTM (Long short-term memory) で実装される回帰関数，Θ はそのパラメータである．Θ の学習には，誤差逆伝搬法 [69] を利用して求めた勾配を利用した勾配法が利用される．

$$\Theta \leftarrow \Theta - \lambda \nabla_\Theta \mathcal{J}(\Theta) \tag{16}$$

ここで $\mathcal{J}(\Theta)$ は目的関数であり，Θ に関して微分が容易な関数が用いられる．現在，広く利用されている目的関数が，目的音と出力音の複素平面上の二乗誤差の最小化である [62] .

$$\mathcal{J}^{\text{PSM}}(\Theta) = \sum_{t=1}^{T} \|\mathbf{S}_t - \mathcal{M}(\mathbf{x}_t|\Theta) \odot \mathbf{X}_t\|_2^2 \tag{17}$$

ただし $\mathbf{S}_t = (S_{t,1}^{(1)}, ..., S_{t,F}^{(1)})$，$\mathbf{X}_t = (X_{t,1}, ..., X_{t,F})$ であり，$\odot$ はアダマール積である．この他にも微分可能な目的関数として，対数尤度 [70]，Kullback-Leibler 情報量や板倉齊藤距離 [71]，クラスタリングコスト [72] や，敵対的生成ネットワークによる識別コスト [73] など様々なものが提案されている．

しかし，本来，音情報処理で求められる仕様の全てが微分可能な目的関数で記述できるわけではない．例えば，音声対話ロボットでは対話満足度を上げることが求められることもあるだろうし，通話では音声品質を向上させることが求められることもあるだろう．ゆえに，（DNN を利用した）音源強調はアプリケーションの仕様を目的関数として最適化されるべきである．しかし，主観評価のような曖昧な指標は微分可能な形で記述することが困難なため，誤差逆伝搬法を利用した最適化の枠組みに組み込むことが困難であった．次節では，著者らが取り組んだ，出力音の音質を向上させるよう DNN を学習する手法を概説する．

[*3] 多くの場合，前後数フレームを連結した対数振幅スペクトルが利用される．

6.1 聴感評点を最大化する DNN 音源強調関数の学習法

微分不可能な目的関数を利用して DNN を学習する方式の一つに，強化学習 [74, 75] などで利用される方策勾配法 [76] がある．この方法は，膨大な回数の試行を元に，目的関数の値を上昇させた行動を出力するように DNN を学習する方式である．近年では，囲碁の勝敗のような，目的関数を決定論的に記述できないタスクを行う DNN の学習に利用されている [75]．この考えを応用し，主観品質を最大化するように，方策勾配法で DNN を学習することで出力音の音質を向上させる DNN 学習法が確立できるのではないかと考えた．しかし，方策勾配法による学習では膨大な回数の評価を行う必要があり，聴取実験をそのまま評価関数とするのは現実的ではない．そこで我々は，PESQ (perceptual evaluation of speech quality) [77] や STOI (short-time intelligibility measure) [78] などの音質の主観評価を模擬した定量評価指標（聴感評点）を最大化するように DNN を学習する方式を提案した [79–81].

今，観測信号と DNN 音源強調の出力音をそれぞれ $\hat{\boldsymbol{S}}, \boldsymbol{X}$ とし，そこから求まる聴感評点を $\mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X})$ とする．そして DNN は，観測信号を得た元で聴感評点を最大化する出力音の従う確率分布 $p(\hat{\boldsymbol{S}}_t|\boldsymbol{X}_t, \Theta)$ を推定する関数として利用する．本稿では，式 (17) の二乗誤差最小化が，分散を単位行列とした複素ガウス分布における出力音の最尤推定と等価であることから，これを拡張し以下の複素ガウス分布を利用する．

$$p(\hat{\boldsymbol{S}}_t|\boldsymbol{X}_t, \Theta) = \prod_{f=1}^{F} \frac{1}{2\pi\sigma_{t,f}^2} \exp\left\{-\frac{\left|\hat{\boldsymbol{S}}_{t,f} - G_{t,f}X_{t,f}\right|^2}{2\sigma_{t,f}^2}\right\}. \tag{18}$$

つまり DNN は，平均パラメータである時間周波数マスク $G_{t,f}$ と，分散パラメータ $\sigma_{t,f}^2$ を推定する関数として実装する．そして目的関数を $\mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X})$ の期待値の最大化として以下のように定義する．

$$\mathcal{J}(\Theta) = \int p(\boldsymbol{X}) \int \mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X}) p(\hat{\boldsymbol{S}}|\boldsymbol{X}, \Theta) d\hat{\boldsymbol{S}} d\boldsymbol{X} \tag{19}$$

この勾配は，log-derivative trick [76] を使うことで，以下のように記述できる．

$$\nabla_\Theta \mathcal{J}(\Theta) = \int p(\boldsymbol{X}) \int p(\hat{\boldsymbol{S}}|\boldsymbol{X}, \Theta) \mathcal{L}(\hat{\boldsymbol{S}}, \boldsymbol{X}, \Theta) d\hat{\boldsymbol{S}} d\boldsymbol{X} \tag{20}$$

$$\mathcal{L}(\hat{\boldsymbol{S}}, \boldsymbol{X}, \Theta) = \mathcal{B}(\hat{\boldsymbol{S}}, \boldsymbol{X}) \nabla_\Theta \ln p(\hat{\boldsymbol{S}}|\boldsymbol{X}, \Theta) \tag{21}$$

つまり目的関数の勾配は，聴感評点で重み付けられた観測音から推定された出力音の対数尤度の期待値で求められる．しかし式 (20) の期待値は解析的に求めることができないため，期待値をサンプリングで置き換え，以下のように勾配推定する．

$$\nabla_\Theta \mathcal{J}(\Theta) \approx \frac{1}{T} \sum_{t=1}^{T} \frac{1}{Q} \sum_{q=1}^{Q} \mathcal{L}(\hat{\boldsymbol{S}}_t^{(q)}, \boldsymbol{X}_t, \Theta) \tag{22}$$

$$\hat{\boldsymbol{S}}_t^{(q)} \sim p(\hat{\boldsymbol{S}}|\boldsymbol{X}_t, \Theta) \tag{23}$$

ここで Q は期待値を近似するためのサンプリングを実行する回数である．細かい実装テクニックには注意が必要なものの，上式によって求めた勾配を元に DNN を学習することで，聴感評点とそれに対応する音質が向上することがわかっている．詳細な実装法と実験結果は [81] を参照されたい．

7. おわりに

本稿では，著者の所属する NTT 研究所において，近年の音声・音響信号処理の実用化向けに研究開発した収音技術を，各問題の難しさと，何故その技術を利用したかに重点をおいて説明した．特に，目的音と雑音のどのような違いに着目するかに重点を置き，方向，位置，音色の違いを利用した収音技術を紹介した．本稿および本講演が，実環境で頑健に動作する音響信号処理アプリケーションの開発の一助になれば幸いである．

謝辞 本稿の執筆に関してご協力いただきました，NTT 研究所の原田 登博士，小林 和則博士，荒木 章子博士，齊藤 翔一郎氏，丹羽 健太博士，伊藤 弘章氏，川瀬 智子博士，オークランド大の日岡 祐輔准教授，電気通信大学の羽田 陽一教授に感謝いたします．また貴重な講演機会をいただきました，NTT メディアインテリジェンス研究所の小橋川 哲博士をはじめとする，音学シンポジウム 2018 実行委員会の皆さまに感謝いたします．

参考文献

[1] K. Kobayashi, *et al.*, “A hands-free unit with noise reduction by using adaptive beamformer,” *IEEE Trans. on Consumer Electronics*, 2008.
[2] Y. Hioka, *et al.*, “Angular region-wise speech enhancement for hands-free speakerphone,” *IEEE Trans. on Consumer Electronics*, 2012.
[3] 辻野ほか “実サービスにおける音声認識と自然言語 インタフェース技術,” 人工知能学会誌, 2013.
[4] C. Q. Robinson, *et al.*, “Scalable format and tools to extend the possibilities of cinema audio,” in *Proc. of SMPTE*, 2012.
[5] J. Engdegard, *et al.*, “Spatial audio object coding (SAOC) - The upcoming MPEG standard on parametric object based audio coding,” *AES 124th Convention*, 2008.
[6] J. Herre, *et al.*, “MPEG-H 3D Audio - The new standard for coding of immersive spatial audio,” *IEEE J. Sel. Top. Signal Process*, 2015.
[7] ISO/IEC 14496-3:2009, Information technology - Coding of audio-visual objects - Part 3: Audio (4th Edition), Subpart 11.
[8] Y. Koizumi, *et al.*, “Optimizing acoustic feature extractor for anomalous sound detection based on Neyman-Pearson lemma,” *in Proc.* of EUSIPCO, 2017.
[9] Y. Kawachi, *et al.*, “Complementary set variational autoencoder for supervised anomaly detection,” *in Proc.* of ICASSP, 2018.
[10] M. Brandstein, *et al.* Eds., “Microphone arrays,” Digital Signal Processing, Springer, 2001.
[11] J. Benesty, *et al.* Eds., “Speech enhancement,” Springer, 2005.
[12] 浅野 太, “音のアレイ信号処理 –音源の定位 · 追跡と分離–,” コロナ社, 2011.
[13] S. Makino Eds., “Audio source separation,” *Springer*, 2018.
[14] K. Niwa, *et al.,* “Diffused sensing for sharp directive beamforming,” *IEEE Trans. Audio, Speech and Language Processing*, 2013.
[15] K. Niwa, *et al.,* “Optimal microphone array observation for clear recording of distant sound sources,” *IEEE Trans. Audio, Speech and Language Processing*, 2016.
[16] Y. Koyano, *et al.,* ‘Infinite-dimensional SVD for revealing microphone array’ s characteristics,” *Applied Acoustics,* 2018.
[17] A. Hyvarinen, *et al.*, “詳解 独立成分分析–信号解析の新しい世界,” 東京電機大学出版局, 2005.
[18] Y. Ephraim, *et al.*, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Audio, Speech and Language Processing*, 1984.
[19] T. Kim, *et al.*, “Independent vector analysis: An extension of ICA to multivariate components,” in *Proc. of Independent Compon. Anal. Blind Source Separation*, 2006.
[20] A. Hiroe, “Solution of permutation problem in frequency domain ICA using multivariate probability density functions,” in *Proc. of Independent Compon. Anal. Blind Source Separation*, 2006.
[21] D. Kitamura, *et al.,* “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” *IEEE Trans. Audio, Speech and Language Processing*, 2016.
[22] J. Cermak, *et al.*, “Blind speech separation by combining beamformers and a time frequency binary mask,” in *Proc. of IWAENC*, 2006.
[23] S. Araki, *et al.*, “Blind speech separation in a meeting situation with maximum SNR beamformers,” in *Proc. of ICASSP*, 2007.
[24] M. Souden, *et al.*, “A multichannel MMSE-based framework for speech source separation and noise reduction,” *IEEE Trans. Audio, Speech and Language Processing*, 2013.
[25] M. Souden, *et al.*, “On optimal frequency-domain multichannel linear filtering for noise reduction,” *IEEE Trans. Audio, Speech and Language Processing*, 2010.
[26] “The 3rd CHiME Speech Separation and Recognition Challenge,” `http://spandh.dcs.shef.ac.uk/chime_challenge/chime2015/`
[27] T. Yoshika, *et al.*, “The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices,” in *Proc. of ASRU*, 2015.
[28] T. Higuchi, *et al.,* “Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR,” *IEEE Trans. Audio, Speech and Language Processing*, 2017.
[29] T. Yoshioka, *et al.,* “Multi-micrphone neural speech separation for far-field multi-talker speech recognition,” in *Proc. of ICASSP*, 2018.
[30] T. Ochiai, *et al.*, “Unified architecture for multichannel end-to-end speech recognition with neural beamforming,” *IEEE J. Sel. Topics Signal Processing*, 2017.
[31] R. Zelinski, “A microphone array with adaptive post-filtering for noise reduction in reverberant rooms,” in *Proc. of ICASSP*, 1988.
[32] I. A. MacCowan, *et al.*, “Microphne array post-filter based on noise field coherence,” *IEEE Trans. Audio, Speech and Language Processing*, 2003.
[33] Y. Hioka, *et al.*, “Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain,” *IEEE Trans. Audio, Speech and Language*

Processing, 2013.
[34] 日岡ほか, “パワースペクトル密度推定に基づく実用的な音源信号分離,” 日本音響学会誌, 2017.
[35] K. Niwa, *et al.*, “PSD estimation in beamspace using property of M-matrix,” in *Proc. of IWAENC*, 2016.
[36] A. Pentland, “Smart rooms,” *Scientific American*, 1996.
[37] M. Coen, “Design principles for intelligent environments,” in *Proc. AAAI (5th NCAI)*, 1998.
[38] N. Ono, *et al.*, “Blind alignment of asynchronously recorded signals for distributed microphone array,” in *Proc.* WASPAA, 2009.
[39] S. Miyabe, *et al.*, “Optimizing frame analysis with non-integer shift for sampling mismatch compensation of long recording,” in *Proc.* WASPAA, 2013.
[40] H. Chiba, *et al.*, “Amplitude-based speech enhancement with non-negative matrix factorization for asynchronous distributed recording,” in *Proc.* IWAENC, 2014.
[41] T. Kako, *et al.*, “Wiener filter design by estimating sensitivities between distributed asynchronous microphones and sound sources,” in *Proc.* WASPAA, 2015.
[42] Y. Matsui, *et al.*, “Multiple far noise suppression in a real environment using transfer-function-gain NMF,” in *Proc.* EUSIPCO, 2017.
[43] Y. Koizumi, *et al.*, “Distant noise reduction based on multi-delay noise model using distributed microphone array,” in *Proc.* EUSIPCO, 2018 (*accepted*).
[44] K. Imoto, *et al.*, “Spatial cepstrum as a spatial feature using distributed microphone array for acoustic scene analysis,” *IEEE Trans. Audio, Speech and Language Processing*, 2017.
[45] J. S. Soo, *et al.*, “Multidelay block frequency domain adaptive filter,” IEEE Trans. ASSP, 1990.
[46] S. Makino, et al., “Exponentially weighted step-size projection algorithm for acoustic echo cancellers,” IEICE Trans. Fundamentals 1992.
[47] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Audio, Speech and Signal Processing*, 1979.
[48] D. D. Lee, *et al.*, “Learning the parts of objects by non-negative matrix factorization”, *Nature*, 1999.
[49] P. Smaragdis, *et al.*, “Non-negative matrix factorization for polyphonic music transcription,” in *Proc. WASPAA,* 2003.
[50] R. Oldfield, *et al.*, “Demo paper: Audio object extraction for live sports broadcast,” in *Proc.* ICMEW, 2013.
[51] R. Oldfield, *et al.*, “Object-based audio for interactive football broadcast,” *Multimedia Tools and Applications*, 2015.
[52] “Taget microphone,” `https://www.youtube.com/watch?v=GKmBAPkwYwA`
[53] Y. Koizumi, *et al.*, “Informative acoustic feature selection on microphone array Wiener filtering for collecting target source on sports ground,” in *Proc. WASPAA*, 2015.
[54] Y. Koizumi, *et al.*, “Integrated approach of feature extraction and sound source enhancement based on maximization of mutual information,” in *Proc. ICASSP*, 2016.
[55] Y. Koizumi, *et al.*, “Informative acoustic feature selection to maximize mutual information for collecting target sources,” *IEEE Trans. Audio, Speech and Language Processing*, 2017.
[56] Y. LeCun, *et al.*, “Deep learning,” *Nature*, 521, pp.436–444, 2015.
[57] D. L. Wang, *et al.*, “Supervised speech separation based on deep learning: An overview,” *arXiv preprint, arXiv:1708.07524,* 2017.
[58] F. Xie, *et al.*, “A family of MLP based nonlinear spectral estimators for noise reduction,” in *Proc. ICASSP*, 1994.
[59] M. Dahl, *et al.*, “A neural network trained microphone array system for noise reduction, *In IEEE Neural Networks for Signal Processing VI*, 1996.
[60] E. A. Wan, *et al.*, “Networks for speech enhancement,” *in Handbook of Neural Networks for Speech Processing, Ed. S. Katagiri*, 1998.
[61] A. Narayanan, *et al.*, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” in *Proc. ICASSP*, 2013.
[62] H. Erdogan, *et al.*, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in *Proc. ICASSP*, 2015.
[63] D. S. Williamson, *et al.*, “Time-frequency masking in the complex domain for speech dereverberation and denoising,” *IEEE/ACM Trans. Audio, Speech and Language Processing*, 2017.
[64] Y. Koizumi, *et al.*, “End-to-end sound source enhancement using deep neural network in the modified discrete cosine transform domain,” in *Proc. ICASSP*, 2018.
[65] Y. Xu, *et al.*, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal Processing Letters*, pp.65–68, 2014.
[66] Y. Xu, *et al.*, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Trans. Audio, Speech and Language Processing*, pp.7–19, 2015.
[67] T. Kawase, *et al.*, “Application of neural network to source PSD estimation for Wiener filter based sound source separation,” in *Proc. IWAENC*, 2016.
[68] K. Niwa, *et al.*, “Supervised source enhancement composed of non-negative auto-encoders and complementarity subtraction” in *Proc. ICASSP*, 2017.
[69] D. E. Rumelhart, *et al.*, “Learning representations by back-propagating errors,” *Nature*, 323, pp.533–536, 1986.
[70] K. Kinoshita, *et al.*, “Deep mixture density network for statistical model-based feature enhancement,” in *Proc. ICASSP*, 2017.
[71] A. A. Nugraha, *et al.*, “Multichannel audio source separation with deep neural networks,” *IEEE/ACM Trans. Audio, Speech and Language Processing*, 2016.
[72] J. Hershy, *et al.*, “Deep clustering: Discriminative embeddings for segmentation and separation,” In *Proc. ICASSP*, 2016.
[73] S. Pascual, *et al.*, “SEGAN: Speech enhancement generative adversarial network,” In *Proc INTERSPEECH*, 2017.
[74] E. S. Sutton, *et al.*, “Reinforcement learning: An introduction,” *A Bradford Book,* 1998.
[75] D. Silver, *et al.*, “ Mastering the game of Go with deep neural networks and tree search,” *Nature*, pp.484—489, 2016.
[76] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Machine Learning*, Vol. 8, 1992.
[77] ITU-T Recommendation P.862, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” 2001.
[78] C. H. Taal, *et al.*, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Trans. Audio, Speech and Language Processing*, 2011.
[79] Y. Koizumi, *et al.*, “DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements,” in *Proc. ICASSP*, 2017.
[80] 小泉ほか, “聴感評点を向上させるための DNN 音源強調関数のブラックボックス最適化,” 音講論 (秋), 2017.
[81] Y. Koizumi, *et al.*, “DNN-based source enhancement to increase objective sound quality assessment score,” *IEEE/ACM Trans. Audio, Speech and Language Processing*, 2018 (*in print*).