DNN マスク推定に基づく畳み込みビームフォーマによる音源分離・残響除去・雑音除去の同時実現*

☆髙橋理希 (筑波大)，中谷智広，落合翼，木下慶介，池下林太郎，
Marc Delcroix，荒木章子 (NTT)，牧野昭二 (筑波大)

1 はじめに

音声信号が遠方のマイクで観測されるとき，しばしば，残響，拡散性雑音，非目的話者の音声などの妨害音が含まれる．これらの成分は，観測音声を不明瞭にするとともに，ハンズフリー電話会議や自動音声認識（ASR）などの多くのアプリケーションで深刻な劣化を引き起こす．

マスクベースのビームフォーミングは，前述の妨害音成分を抑制するために非常に効果的であることが示されている．このアプローチでは，ビームフォーマ [1] と呼ばれる線形フィルタが，目的の音声を強調するために観測された信号に適用される．ここで，マスクは各時間周波数点においてどの音源が支配的であるかを示すために使用され，ビームフォーマの係数の推定に用いられる．

マスクベースのビームフォーミングにおいて重要な要素はマスクの推定である．マスク推定は大別してニューラルネットワーク (NN) に基づく手法と空間クラスタリングに基づく手法の 2 種類が開発されている．NN によるアプローチ [2,3] では，事前に学習された NN を使用して，観測信号の時間周波数点を目的の信号や雑音などの複数のクラスに分類する．雑音に非目的話者の音声が含まれる場合には，深層クラスタリング [4]，パーミュテーション不変学習（uPIT）[5] など，音源分離用に開発された高度な NN 技術を使用して，複数の話者に分類できる．NN によるアプローチの利点は，すべての周波数にわたる信号のスペクトルパターンの特徴をとらえてマスクを推定できる点である．一方で，空間クラスタリングによるアプローチ [6, 7] では，マルチチャネル信号から抽出された信号の到来方向などの空間特性に基づいて，時間周波数点を複数の音源にクラスタリングする．ただし，クラスタリングは各周波数で個別に実行されるため，個々の周波数で推定された時間周波数マスクをそれぞれの音源にさらにグループ化する必要がある．これはパーミュテーション問題と呼ばれ，マスク推定の全体的な精度にとって重要である．空間クラスタリングによるアプローチの利点は，推定が教師なし学習で行われるため，テスト環境に柔軟に適応できる点である．

しかし，雑音や残響に加えて複数の音源が存在するような環境（雑音残響下複数音源環境）でのマスク推定は依然として困難な課題である．予備実験では，NN を用いたマスク推定により雑音除去と残響除去を同時に実現しようとすると、性能が低下することが確認されている．また，空間クラスタリングでは，雑音と残響の存在下ではパーミュテーション問題の解決が困難になる．これらの問題により，マスクベースのビームフォーミングは，雑音除去，残響除去，および音源分離を同時に行う必要がある一般的な環境への適用が難しい．

上記の問題を解決するために，本稿では，マスクベースのビームフォーミングを単一の最適化フレームワークに統合する方法を提案する．マスク推定のために，提案手法は信号のスペクトルおよび空間特性に関する情報を確率的に統合する結合尤度関数 [8] を使用して，大きな受容野を持つ周波数領域畳み込み NN を用いた uPIT（CNN-uPIT）[9]，及び雑音重畳型複素ガウス混合モデルに基づく空間クラスタリング（noisyCGMM）[10] を統合する．この統合により，テスト環境に柔軟に適応し，雑音に強く，パーミュテーション問題を高精度に解決できるマスク推定が可能となる．さらに，ビームフォーミングには，重み付き最小パワー無歪み応答畳み込みビームフォーマ (WPD) [11] を採用し，上記のマスク推定手法と統合することで，残響に対してより頑健にする．

実験により，これらの技術はすべて，雑音除去，残響除去，および音源分離を同時に達成するために重要な役割を果たすことを確認する．

2 信号モデルと WPD

2.1 信号モデル

雑音残響環境で，K 人の話者が話す音声信号が M 個のマイクで収録されると仮定する．収録された信号は、短時間フーリエ変換（STFT）領域で次のようにモデル化される．

$$\mathbf{x}_t = \sum_{k=0}^{K} \mathbf{x}_t^{(k)}. \tag{1}$$

ここで t は時間フレーム，$k > 0$ における $\mathbf{x}_t^{(k)} \in \mathbb{C}^M$ は k 番目の話者の観測信号，$k = 0$ における $\mathbf{x}_t^{(k)}$ は加法性雑音である．また本稿では簡単のため，同じ処理が各周波数で独立に適用されるという前提のもと，すべての記号の周波数インデックスを省略する．$k > 0$ における $\mathbf{x}_t^{(k)}$ は，さらに以下の式でモデル化できる．

$$\mathbf{x}_t^{(k)} = \mathbf{d}_t^{(k)} + \mathbf{r}_t^{(k)}, \tag{2}$$

$$\mathbf{d}_t^{(k)} = \mathbf{v}^{(k)} s_t^{(k)}, \tag{3}$$

$$\mathbf{r}_t^{(k)} = \sum_{\tau=\mathrm{D}}^{L_a+D-1} \mathbf{a}_\tau^{(k)} s_{t-\tau}^{(k)}. \tag{4}$$

ここで，$\mathbf{d}_t^{(k)}$ は k 番目の音源から生成された直接音と初期残響の合計であり，$\mathbf{r}_t^{(k)}$ は後部残響である．時間領域

*Simultaneous realization of denoising, dereverberation, and source separation, using DNN-supported mask-based convolutional beamforming. by Riki TAKAHASHI (University of Tsukuba), Tomohiro NAKATANI, Tsubasa OCHIAI, Keisuke KINOSHITA, Rintaro IKESHITA, Marc Delcroix, Shoko ARAKI (NTT), Shoji MAKINO (University of Tsukuba)

での初期残響の長さが分析窓よりも短いと仮定すると，式 (3) の $\mathbf{d}_t^{(k)}$ は，k 番目のクリーン音声 $s_t^{(k)} \in \mathbb{C}$ と，ステアリングベクトル $\mathbf{v}^{(k)} = [v_1^{(k)}, v_2^{(k)}, \ldots, v_M^{(k)}]^\top \in \mathbb{C}^M$ の積によってモデル化できる．式 (4) の $\mathbf{r}_t^{(k)}$ は，$s_t^{(k)}$ と畳み込み伝達関数 $\mathbf{a}_\tau^{(k)} \in \mathbb{C}^M (\tau = D, D+1, \ldots, L_a + D - 1)$ との間の周波数領域畳み込みによってモデル化される．ここで，D は後部残響が直接音に遅れて開始する時間フレームの番号，L_a は畳み込み伝達関数の長さである．

本稿では、各 $k(>0)$ における $\mathbf{d}_t^{(k)}$ を k 番目の音源の目的信号と呼び，推定される信号として扱う．また，提案手法の目的は， $\mathbf{d}_t^{(k)}$ を維持しながら，目的音源の後部残響 $\mathbf{r}_t^{(k)}$ と，非目的音源 $\mathbf{x}_t^{(k')} (k' \neq k)$ を $\mathbf{x}_t$ から低減することである．

2.2 WPD

WPD [11] は，残響除去フィルタ (WPE) [12] と雑音除去用の (非畳み込み) ビームフォーマを統合した畳み込みビームフォーマである．目的信号 $\mathbf{d}_t^{(k)}$ は以下の式による WPD の出力 $y_t^{(k)}$ として推定される．

$$y_t^{(k)} = (\mathbf{w}_0^{(k)})^{\mathsf{H}} \left(\mathbf{x}_t - \sum_{\tau=D}^{L+D-2} \mathbf{C}_\tau^{\mathsf{H}} \mathbf{x}_{t-\tau} \right), \quad (5)$$

$$= (\mathbf{w}_0^{(k)})^{\mathsf{H}} \mathbf{x}_t + \sum_{\tau=D}^{L+D-2} (\mathbf{w}_\tau^{(k)})^{\mathsf{H}} \mathbf{x}_{t-\tau}. \quad (6)$$

ここで，$\mathbf{w}_0^{(k)}$ は無歪み条件を満たすビームフォーマ，$\mathbf{C}_\tau$ は WPE の予測行列，L は畳み込みビームフォーマの長さ，H は共役転置である．式 (5) は，$\mathbf{w}_t^{(k)} = -\mathbf{C}_t \mathbf{w}_0^{(k)}$ とすることで，式 (6) に一致する．WPD は WPE とビームフォーマを統合することで，後部残響を削減し，雑音や非目的音声を抑圧しながら目的信号を強調することができる．

ビームフォーマの係数の最尤解は，目的信号のステアリングベクトル $\mathbf{v}^{(k)}$ が与えられた場合に推定でき [12]，$\mathbf{v}^{(k)}$ は目的信号のマスクに基づいて推定することができる．マスクに基づいて $\mathbf{v}^{(k)}$ を推定するための手法は多く提案されているが，本稿では，雑音共分散白色化による固有値分解に基づく手法 [13,14] を用いる．この手法では、$\mathbf{v}^{(k)}$ は次のように推定される．

$$\mathbf{v}^{(k)} = \Phi^{(\backslash k)} \mathrm{MaxEig}((\Phi^{(\backslash k)})^{-1} \Phi^{(k)}). \quad (7)$$

ここで，MaxEig(·) は最大固有値に対応する固有ベクトルを計算する関数である．また，$\Phi^{(k)}$ および $\Phi^{(\backslash k)}$ は，目的信号とそれ以外の信号の空間共分散行列であり，これらはそれぞれ次のように推定される．

$$\Phi^{(k)} = \frac{\sum_t \lambda_t^{(k)} \mathbf{z}_t \mathbf{z}_t^{\mathsf{H}}}{\sum_t \lambda_t^{(k)}} \quad \text{and} \quad \Phi^{(\backslash k)} = \frac{\sum_t (1 - \lambda_t^{(k)}) \mathbf{z}_t \mathbf{z}_t^{\mathsf{H}}}{\sum_t (1 - \lambda_t^{(k)})}. \quad (8)$$

ここで，$\lambda_t^{(k)}$ は目的信号のマスクであり，$\mathbf{z}_t$ は WPE の機能によって残響除去のなされた混合音声である．

3 マスク推定

WPD においては，ビームフォーマの係数を推定するために目的信号のステアリングベクトルが必要となる．従来の WPD [15] では，noisyCGMM [10] によって目的信号のマスクを推定することで，ステアリングベクトルの推定を行っていた．しかし，noisyCGMM などの空間クラスタリングによるマスク推定では，雑音残響環境下でパーミュテーション問題の解決が困難になるという問題があった．一方，NN ベースのマスク推定を用いる場合でも，残響除去と雑音除去を同時に精度よく行うことは困難であった．これらの問題を解決するため，本稿では NN ベースのマスク推定手法である CNN-uPIT [9] と，空間クラスタリングベースのマスク推定手法である noisyCGMM を統合した手法によって，目的信号のマスクを推定する．

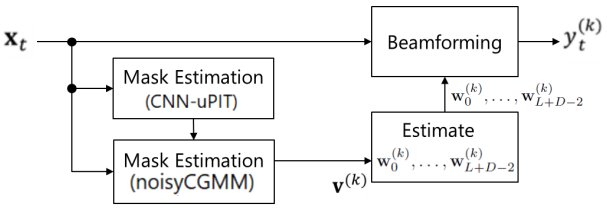

Fig. 1 Overall processing flow of the proposed method

図 1 は，提案手法全体の処理の流れを示している．マスクは初めに CNN-uPIT によって推定され，noisyCGMM によって改善される．次に，推定されたマスクに基づいて WPD が設計され，観測信号に適用されて，雑音除去，残響除去，音源分離がなされた目的信号が生成される．

以下では，CNN-uPIT，noisyCGMM，およびそれらを統合したマスク推定手法について説明する．

3.1 CNN-uPIT

uPIT [5] は，混合音声を入力とし，個々の音源の推定マスクを出力する NN である．本稿では CNN-uPIT [9] を使用する．CNN-uPIT は uPIT の一種であり，そのネットワーク構造として Conv-TasNet [16] で使用されるものと同様の大きな受容野を持つ．また，STFT 係数の実数部と虚数部の連結をネットワークの入力とし，推定マスクを観測信号にかけて得られる信号の信号対歪み比（SDR）を損失関数として設定する．

提案手法では，目的信号 $\mathbf{d}_t^{(k)} (k = 1, \ldots, K)$ に関するマスクを，CNN-uPIT によって推定する．推定されたマスクは，$\lambda_{\mathrm{uPIT},t}^{(k)}$ で表される．この推定のために，次のような 3 つの CNN-uPIT の構成を検討し，実験によって有効性を比較する．

CNN-uPIT(derev) この構成では，CNN-uPIT は，(A) 雑音＋残響＋複数音源を含んだ入力信号を受け取り，(B) 各目的信号に対応するマスクを出力するように学習される．

CNN-uPIT + WPE この構成では，(C) 入力信号 (A) に対し WPE 前処理によって残響除去された後の信号を受け取り，(D) 入力信号 (C) に含まれる各音源信号に対応するマスクを出力するように学習される．すなわち，残響除去は WPE のみで行われ，CNN-uPIT は残響除去は行わない．

CNN-uPIT(derev) + WPE この構成では，上記

(C) を受け取り，(B) を出力するように学習される．すなわち，CNN-uPIT は，WPE 前処理で除去されずに残った残響も低減するように学習される．

3.2 noisyCGMM

複素ガウス混合モデルに基づく手法（CGMM）[6] などの従来の空間クラスタリングアプローチでは，W-disjoint Orthogonality(W-DO) [17] がすべての音源で満たされると仮定されているが，拡散雑音に関してはこの仮定を満たさない場合が多い．これを解決するため，noisyCGMM [10] では，拡散雑音はすべての時間周波数点に存在し、W-DO は他のすべての音源についてのみ満たされると仮定する．各時間周波数点はひとつの支配的な音源と拡散雑音によってモデル化され，その確率分布は以下のようになる．

$$p(\mathbf{z}_t \mid I_t = k; \theta_{\mathrm{SC}}) = \mathcal{N}(\mathbf{z}_t; 0, \psi_t^{(k)} \Psi^{(k)} + \gamma_t \Gamma). \quad (9)$$

ここで、I_t は時間周波数点の支配的な音源のインデックスであり，$p(\mathbf{z}_t \mid I_t = k)$ は支配的な音源が $I_t = k$ である場合の $\mathbf{z}_t$ の条件付き確率分布である．また，$\mathcal{N}(\mathbf{z}; 0, \Psi)$ は多変量複素ガウスの確率密度関数，$\theta_{\mathrm{SC}} = \{\psi_t^{(k)}, \gamma_t, \Psi^{(k)}, \Gamma\}_t^{(k)}$ はモデルパラメータの集合である．$\Psi^{(k)}$ と Γ は k 番目の音源と拡散雑音の時不変の空間共分散行列であり，$\psi_t^{(k)}$ と γ_t はそれぞれの信号の時変パラメータである．

テストデータが与えられた場合のモデルパラメータ θ_{SC} の推定は，期待値最大化（EM）アルゴリズムによって実現できる [10]．モデルパラメータが推定されると，k 番目の音源のマスクは，支配的な音源の事後確率として定められ，次のように推定される．

$$\lambda_t^{(k)} = \frac{p(I_t = k) p(\mathbf{z}_t \mid I_t = k; \theta_{\mathrm{SC}})}{\sum_{k'=0}^{K} p(I_t = k') p(\mathbf{z}_t \mid I_t = k.\theta_{\mathrm{SC}})}. \quad (10)$$

ここで，$p(I_t = k)$ は時間周波数点が k 番目の音源によって支配される事前確率である．上記の noisyCGMM のモデリングにより，従来の CGMM よりも雑音に対してより頑健に空間クラスタリングを行える．

3.3 noisyCGMM と CNN-uPIT の統合

noisyCGMM では，空間クラスタリングは各周波数で独立して実行されるため，パーミュテーション問題を解決する必要がある．つまり，異なる周波数で取得したクラスタを個々の音源にグループ化する必要がある．この目的のために多くの技術が提案されている [7] が，雑音の多い残響環境では性能が著しく低下する．一方，CNN-uPIT は noisyCGMM よりも頑健に，異なる周波数で各音源の信号成分を関連付けることができる．

よって本稿では，パーミュテーション問題の解決のために，CNN-uPIT を noisyCGMM に統合する [8]．具体的には，式 (10) の事前確率 $p(I_t = k)$ を，CNN-uPIT によって推定されるマスク $\lambda_{\mathrm{uPIT},t}^{(k)}$ に置き換える．

$$p(I_t = k) = \lambda_{\mathrm{uPIT},t}^{(k)}. \quad (11)$$

この統合により，すべての周波数のスペクトルパターンを考慮して空間クラスタリングを実行できるため，noisyCGMM による空間クラスタリングと同時にパーミュテーション問題を解決できるようになる．

4 評価実験

本章では，提案手法の性能を実験により評価し，個々の処理ブロックの役割を分析する．

4.1 実験条件

データセットとして，REVERB Challenge データセット（REVERB）[18] を用いて，雑音残響下混合音声（REVERB-MIX）を準備した．REVERB の各発話には，少しの定常拡散雑音を含む単一話者の音声と残響が含まれている．テストデータを生成するために，REVERB から抽出された 2 つの発話（1 つは開発セットから、もう 1 つは評価セットから）を混合した．混合発話の各ペアは，同じ部屋で，同じマイクアレイで，同じ条件（近距離または遠距離，SimData または RealData）で録音されているものを選んだ．テストデータとして，REVERB 評価セットと同じ数の混合音声を作成した．REVERB 評価セットの各発話は，テストデータのいずれかの混合音声に含まれている．テストデータの各混合音声の長さは，REVERB 評価セットの対応する発話の長さと同じにした．また，CNN-uPIT の学習データと評価データを生成するために、WSJ-CAM0 コーパス [19] からランダムに抽出された 2 つの発話と，REVERB 学習セットから抽出された 2 つの部屋のインパルス応答と背景雑音を使用して，混合音声を作成した．

実験では，各混合音声から 2 つの音声信号を推定し，REVERB 評価セット用に用意されたベースライン評価ツールを使用して，REVERB 評価セットに対応する音声信号のみを評価した．この際，推定されたそれぞれの信号と REVERB 評価セットの元の信号との相関に基づいて，評価する信号を選択した．また，音声強調性能の客観的な尺度として，ケプストラム距離（CD），周波数重み付きセグメンタル SNR（FWSSNR），および PESQ を使用した．ASR の単語誤り率 (WER) を評価するために，Kaldi [20] を使用して開発された REVERB のベースライン ASR システムを使用した．

分析窓には Hann 窓を使用し，フレーム長とシフトをそれぞれ 32 ミリ秒と 8 ミリ秒に設定した． サンプリング周波数は 16 kHz で，$M = 8$ 個のマイクを使用した．WPD のフィルタ長は，0〜0.8 kHz，0.8〜1.5 kHz，および 1.5〜8 kHz の周波数範囲に対して，それぞれ $L = 17, 13,$ および 5 に設定した．予測遅延は $D = 4$ に設定した．

4.2 実験結果

図 2 は，CNN-uPIT(derev) によって推定されたマスクのみで音声強調を行った場合の結果 (Masking) と、マスクベースビームフォーマを用いた場合の結果を示している．ビームフォーマとしては，MPDR，WPE と MPDR のカスケード接続（WPE+MPDR），および WPD の 3 つを用いた．またベースラインとして，音声強調処理を行わなかった場合 (No Enhancement) の結果も示している．実験結果は，マスクベースのビームフォーミングの有効性を明確に示しており，中でも WPD はすべてのビームフォーマの中で最高の性能を示した．

図 3 は，WPD に対してそれぞれ異なるマスク推定手法を使用して得られた評価結果を示している．このグラフでは，マスク推定において，noisyCGMM と CNN-

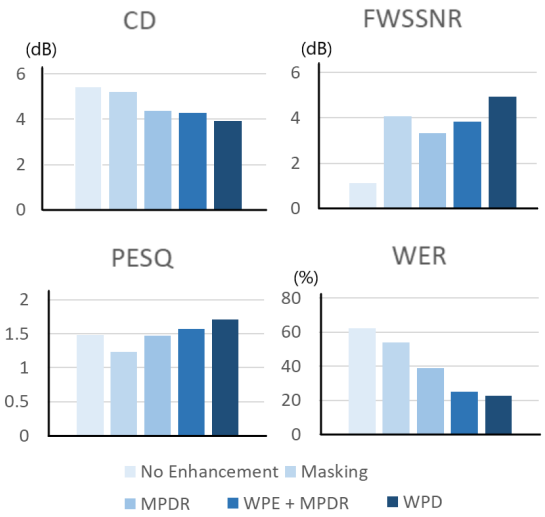

Fig. 2　CD (dB), FWSSNR (dB), and PESQ for SimData, and WER (%) for RealData in REVERB-MIX obtained with different enhancement schemes using masks estimated by CNN-uPIT(derev).

uPIT の 3 つの構成のそれぞれを，単独で用いた場合と統合して用いた場合の結果を示している．実験結果から，noisyCGMM と CNN-uPIT の統合により，一貫して性能が改善されたことを確認できる．中でも，マスク推定のために WPE と CNN-uPIT の両方で残響除去が実行される CNN-uPIT(derev) + WPE は，すべての指標で最高の性能を示した．

5　まとめ

本稿では，雑音除去，残響除去，及び音源分離を同時に実行できる，マスクベースの畳み込みビームフォーミングの統合手法を提案した．この手法は，CNN-uPIT，noisyCGMM，および WPD の 3 つの技術で構成されている．CNN-uPIT および noisyCGMM によって抽出された信号のスペクトルおよび空間特性を統合し，WPD の残響除去機能を利用することで，雑音残響下複数音源環境でも時間周波数マスクを確実に推定でき，最適な WPD によって音声強調を実行することができる．また，雑音残響下混合音声を使用した実験を行った結果，提案手法に使用されるすべての技術が，雑音残響下複数音源環境で効果的な音声強調を達成するために重要な役割を果たすことを示した．

参考文献

[1] B. D. Van Veen *et al.*, *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, April 1988.

[2] J. Heymann *et al.*, in *Proc. IEEE ICASSP*, 2017, pp. 5325–5329.

[3] H. Erdogan *et al.*, in *Proc. IEEE ICASSP*, 2015, pp. 708–712.

[4] Z.-Q. Wang *et al.*, in *Proc. IEEE ICASSP*, 2018, pp. 1–5.

[5] M. Kolbæk *et al.*, *IEEE/ACM TASLP*, vol. 25, no. 10, pp. 1901–1913, 2017.

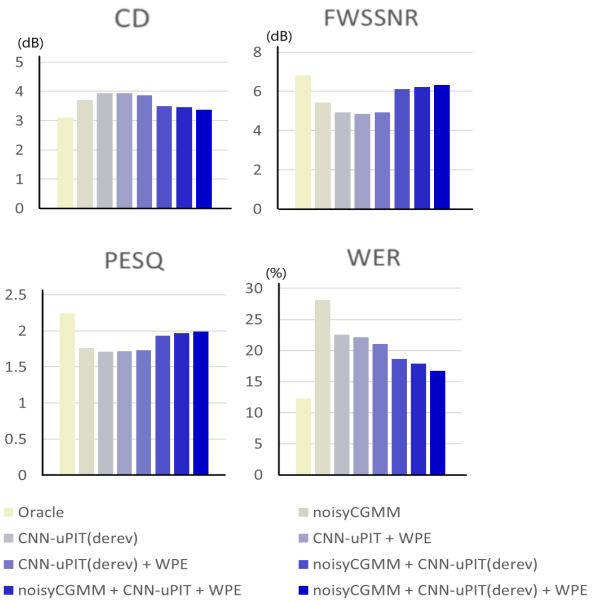

Fig. 3　CD (dB), FWSSNR (dB), and PESQ for SimData, and WER (%) for RealData in REVERB-MIX obtained with the WPD convolutional beamformer using different mask estimators.

[6] T. Yoshioka *et al.*, in *Proc. IEEE ASRU*, 2015, pp. 436–443.

[7] H. Sawada *et al.*, *IEEE/ACM TASLP*, vol. 19, no. 3, pp. 516–527, 2010.

[8] T. Nakatani *et al.*, in *Proc. IEEE ICASSP*, 2017, pp. 286–290.

[9] F. Bahmaninezhad *et al.*, in *Proc. Interspeech*, 2019, pp. 4574–4578.

[10] N. Ito *et al.*, in *Proc. EUSIPCO*, 2018, pp. 1662–1665.

[11] T. Nakatani *et al.*, *IEEE Signal Processing Letters*, vol. 26, pp. 903–907, April 2019.

[12] T. Yoshioka *et al.*, *IEEE/ACM TASLP*, vol. 20, no. 10, pp. 2707–2720, 2012.

[13] N. Ito *et al.*, in *Proc. IEEE ICASSP*, 2017, pp. 681–685.

[14] S. Markovich-Golan *et al.*, in *Proc. IEEE ICASSP*, 2015, pp. 544–548.

[15] T. Nakatani *et al.*, in *Proc. IEEE WASPAA*, 2019, pp. 224-228.

[16] Y. Luo *et al.*, *IEEE/ACM TASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[17] A. Jourjine *et al.*, in *Proc. IEEE ICASSP*, 2000, pp. 2985–2988.

[18] K. Kinoshita *et al.*, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE WASPAA*, 2013.

[19] T. Robinson *et al.*, in *Proc. IEEE ICASSP*, 1995, pp. 81–84.

[20] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, Dec. 2011.