# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題 | 雑音の基底信号を用いた耐雑音性の高い時間領域音声分離 |
| Title | Noise-robust time-domain speech separation with basis signals for noise |
| 著者 | 尾座本耕平, 岩野 公司, 宇都 有昭, 篠田 浩一 |
| Authors | Kohei OZAMOTO, Koji IWANO, Kuniaki UTO, Koichi SHINODA |
| 出典 | 電子情報通信学会技術研究報告, vol. 120, no. 399, pp. 63-67 |
| Citation | TECHNICAL REPORT OF IEICE, vol. 120, no. 399, pp. 63-67 |
| 発行日 / Pub. date | 2021, 3 |
| 権利情報 / Copyright | 本著作物の著作権は電子情報通信学会に帰属します。<br>Copyright(c) 2021 IEICE |

一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
EA2020-70,SIP2020-101,SP2020-35(2021-03)

# ［ポスター講演］雑音の基底信号を用いた耐雑音性の高い時間領域音声分離

尾座本耕平† 岩野 公司†† 宇都 有昭† 篠田 浩一†

† 東京工業大学 情報理工学院
†† 東京都市大学 メディア情報学部
E-mail: ozamoto@ks.c.titech.ac.jp, iwano@tcu.ac.jp, uto@ks.c.titech.ac.jp, shinoda@c.titech.ac.jp

**あらまし**　近年，深層学習を用いた音声分離が盛んに研究されている．波形を直接入力する時間領域の手法である TasNet は，音声を畳み込みによって特徴量に変換した上で分離を行い，分離した特徴量に対して畳み込みを行って波形を再構成する．畳み込みフィルタは基底信号と呼ばれ，話者間の分離精度を高めるように学習される．この手法は，音声に雑音が含まれている場合に分離性能が大きく低下する．そこで我々は，話者の基底信号に加え雑音の基底信号を追加することで分離性能を向上させる方法 TasNet with noise basis signals（TasNet-NB）を提案する．雑音の SN 比を徐々に小さくするカリキュラム学習と雑音の再構成損失の計算を用いる．WHAM!データセットを用いて評価した結果，SI-SDRi が 13.7 から 14.6 に向上した．

**キーワード**　音声分離，単チャネル，時間領域，深層学習

# [Poster Session] Noise-robust time-domain speech separation with basis signals for noise

Kohei OZAMOTO†, Koji IWANO††, Kuniaki UTO†, and Koichi SHINODA†

† Department of Computer Science, Tokyo Institute of Technology
†† Faculty of Environmental Studies, Tokyo City University
E-mail: ozamoto@ks.c.titech.ac.jp, iwano@tcu.ac.jp, uto@ks.c.titech.ac.jp, shinoda@c.titech.ac.jp

**Abstract**　Recently, speech separation using deep learning has been extensively studied. TasNet, a time-domain method that directly inputs waveforms, converts speech into features by convolution, performs separation, and reconstructs waveform by convolution on separated features. The convolutional filter is called basis signals and is trained to improve separation accuracy between speakers. The separation performance of this method is greatly degraded when the speech contains noise. Therefore, we propose TasNet with noise basis signals (TasNet-NB), a method to improve separation performance by adding noise basis signals to speaker's basis signals. We evaluate the method on WHAM! dataset and show that it improves SI-SDRi from 13.7 to 14.6.

**Key words**　Speech separation, single channel, time-domain, deep learning

## 1. まえがき

音声分離とは同時に複数人が話している状況において話者ごとの音声を分離する技術である．音声分離では一般に，音声をいったん時間周波数領域の特徴量に変換した上で分離する [1]．このシステムでは音声の周波数や振幅は用いるが位相は破棄される．その結果，波形を再構成するときに音声の位相も計算しなければならず，出力する音声の品質低下が避けられない．

これに対し，近年，深層学習を用いて時間領域の信号を直接入力とし end-to-end 学習を行う時間領域システム TasNet が登場した [2]．上述の問題を回避できるだけでなく，処理の遅延を大幅に減らすことができる．時間領域システムは一般に時間周波数領域システムに比べて高い分離性能をもつが，一方で分離対象の音声に雑音が含まれているときに性能が大きく低下する [3]．時間領域システムでは畳み込みによって波形を特徴量に変換する．畳み込みに用いるフィルタは基底信号と呼ばれ，話者間の分離精度を高めるように学習される．しかし，雑音を含む音声分離で学習した場合に話者と雑音が十分に分離されず，それが耐雑音性低下の一因になっていると考えられる．

本稿では，この問題を解決するため，時間領域システムに対

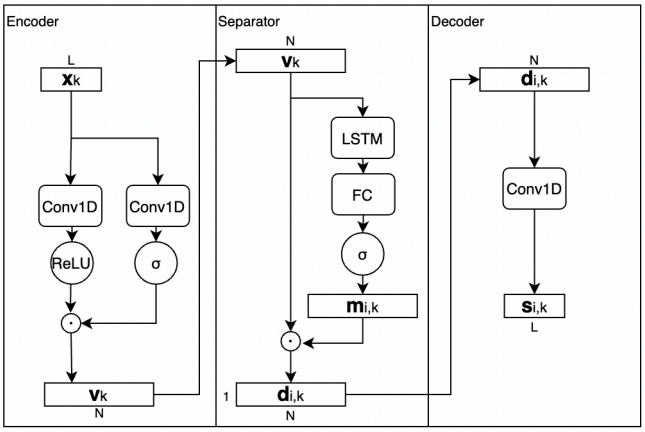


図 1　TasNet のネットワーク構造

して話者とともに雑音の基底信号の作成を行う手法 TasNet with noise basis signals（TasNet-NB) を提案する．学習時に雑音の再構成損失を計算する．また，データセットにおける雑音の相対量を段階的に増加させるカリキュラム学習を適用する．

## 2. 従来研究

単チャネル音声分離では，入力する混合音声の離散信号 $x$ を $C$ 個の話者 $s_1, ..., s_c$ に分離する．時間領域システムでは推定された信号$\hat{s}_1, ..., \hat{s}_c$と目標の信号 $s_1, ..., s_c$と から損失関数を Scale-Invariant Signal-to-Distortion Ratio（SI-SDR）[4] を目的関数として学習を行う．

### 2.1 TasNet

本稿では多くの時間領域システムの元になっている TasNet を使用する．ネットワークの構造を図 1 に記す．実装は Asteroid Toolkit [5] を用いた．

#### 2.1.1 ネットワーク構造

まず分離対象の混合音声の離散信号をフレーム長 $L$ ごとに $L/2$ ずつずらしながら区切って $\mathbf{x}_k$ を得る．$k$ 個目のセグメント $\mathbf{x}_k$ に対して，Encoder によって特徴量 $\mathbf{v}_k$ への変換を行う．

$$\mathbf{v}_k = \mathrm{ReLU}(\mathrm{Conv1D}(\mathbf{x}_k)) \odot \sigma(\mathrm{Conv1D}(\mathbf{x}_k)) \qquad (1)$$

ここで，2 つの Conv1D は 1 次元畳み込み層であり，それぞれのフィルタを Encoder における基底信号 $\mathbf{B}^e$ と呼ぶ．ReLU は活性化関数 Rectified Linear Unit を表し，$\sigma$はシグモイド関数を表す. $\odot$ はアダマール積を表す．以上の構造はゲート付き CNN と呼ばれる [6].

時系列データである $K$ 個のセグメント間の依存性のモデル化には再帰型ニューラルネットワーク（RNN）のひとつである双方向長・短期記憶（Bidirectional long short-term memory: BLSTM）ネットワークを用いる．システムをリアルタイムに用いる場合は双方向性でない LSTM を用いる．LSTM では従来の RNN における勾配消失問題を改善しており，長期依存性をモデル化できる．

続く Separator では LSTM とその後に続く全結合層とシグモイド関数によって，入力する $\mathbf{v}_k$に対して話者数 $C$ 個のマスク $\mathbf{M}_k = [\mathbf{m}_{1,k}, ..., \mathbf{m}_{C,k}]$ を出力する．ここで，マスク $\mathbf{m}_{i,k}$ は $\mathbf{v}_k$ における話者 $i$ の音声の割合を表す値である．出力したマスクを用いて $\mathbf{v}_k$ に対する分離を行う．この際，$\mathbf{v}_k$ に対し [7] で用いられた Global layer normalization により正規化を行う．

$$\mathbf{d}_{i,k} = \mathbf{v}_k \odot \mathbf{m}_{i,k} \qquad (2)$$

最後に Decoder では Separator で得た 特徴量 $\mathbf{d}_{i,k} \in \mathbb{R}^{1\times N}$ を波形の形に変換する．変換には 1 次元畳み込みを用いる．

$$\mathbf{s}_{i,k} = \mathrm{Conv1D}(\mathbf{d}_{i,k}) \qquad (3)$$

ここで用いる畳み込み層のフィルタを Decoder の基底信号 $\mathbf{B}^d$ と呼ぶ．以上のようにして，$\mathbf{s}_i \in \mathbb{R}^{1\times L}$ として各話者の波形を得る．最終的に $K$ 個の波形を結合し，各話者の分離後の音声を構成する．

#### 2.1.2 目的関数

単チャネル音声分離の研究において広く使用されている SI-SDR を目的関数として用いる．

$$\mathbf{s}_{\mathrm{target}} = \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{\| \mathbf{s} \|^2} \qquad (4)$$

$$\mathbf{e}_{\mathrm{noise}} = \hat{\mathbf{s}} - \mathbf{s}_{\mathrm{target}} \qquad (5)$$

$$\mathbf{SI\text{–}SDR} := 10\log_{10} \frac{\| \mathbf{s}_{\mathrm{target}} \|^2}{\| \mathbf{e}_{\mathrm{noise}} \|^2} \qquad (6)$$

ここで，$\hat{\mathbf{s}}$ と $\mathbf{s}$ はそれぞれ推定信号と目標信号とする．$\langle x, y \rangle$ は $x, y$ の内積を表し，$\| x \|$ はL2 ノルムを表している．

#### 2.1.3 順列問題

学習を行う中では，ネットワークから出力された $\mathbf{s}_{i,k}$ が実際にはどの話者のものであるかがわからないという順列問題がある．これに対して Permutation invariant training（PIT）[8] を採用して目的関数として計算する SI-SDR が最小となるように順列を並び替える操作を行う．

### 2.2 基底信号を置換する手法

TasNet では Encoder と Decoder の基底信号を end-to-end 学習の中で最適化している．一方，学習された基底信号の代わりに多相ガンマトーンフィルタバンクを用いることで，SI-SDRi を改善できるだけでなく，基底信号のサイズを減らしても性能が変わらないことが報告された [9]．また Encoder と Decoder 双方ではなく片方のみ基底信号を置換した場合に最も分離性能が高くなるという結果も得られている．

### 2.3 雑音の再構成損失の計算

話者と同様に雑音に対してもマスク推定を行って分離し，再構成損失を計算する手法が雑音を含む音声分離と音声強調において用いられ [10] [11]，雑音の推定が頑健なモデルを学習するための正則化として機能することが述べられている [11]．しか

し，時間領域システムによる雑音を含む音声分離において，雑音を話者と同様に分離して再構成損失を計算することによる効果は確認されてこなかった．

## 3. 提 案 手 法

TasNet with noise basis signals（TasNet-NB）を提案する．TasNet において話者の基底信号に加え雑音の基底信号を用いることで話者間だけでなく雑音に対する分離性能も向上させる．これに加え学習精度を高めるため雑音の再構成損失の計算を行い，またカリキュラム学習も導入する．

### 3.1 雑音の基底信号の学習

TasNet で end-to-end 学習を行う中で Encoder・Decoder の基底信号は他のパラメータと同様に学習される．基底信号 $\mathbf{B}^x$（$x = e, d$）はフレーム長 $L$ の信号 $N$ 個によって構成されていると捉えることができる．以後，$N$ を基底信号の数と表記する．

$$\mathbf{B}^x \in \mathbb{R}^{N \times L} \tag{7}$$

まず，雑音を含まない話者のみのデータセット（“Clean”）を用いて基底信号 $\mathbf{B}^x$ を学習する．このとき $\mathbf{B}^x$ は “Clean” の環境下の話者間分離に特化していると考えられ，

$$\mathbf{B}^x = \mathbf{B}^x_{\text{clean}} \tag{8}$$

と表す．

次に，話者間の分離に加えて雑音への分離性能を得るために，雑音の基底信号を構成する．まず追加の基底信号 $\mathbf{B}^x_{\text{extend}} \in \mathbb{R}^{N' \times L}$ を作り，元の基底信号に結合する．

$$\mathbf{B}^x_{\text{conc}} = \text{Concatenate}(\mathbf{B}^x_{\text{clean}}, \mathbf{B}^x_{\text{extend}}) \in \mathbb{R}^{(N+N') \times L} \tag{9}$$

ここで，Concatenate はベクトルの結合を表す．雑音を含む音声分離（“Noisy”）において $\mathbf{B}^x_{\text{conc}}$ を用いて学習を行うが，$\mathbf{B}^x_{\text{clean}}$ 部分に関してはパラメータを固定して値を更新せずに $\mathbf{B}^x_{\text{extend}}$ のみ学習を行う．このように学習を行うことで，話者間の分離性能を保ちつつ新たに設けた $\mathbf{B}^x_{\text{extend}}$ に雑音に対する分離性能を獲得させる（図 2）．

$$\mathbf{B}^x_{\text{extend}} = \mathbf{B}^x_{\text{noise}} \tag{10}$$

このとき $\mathbf{B}^x_{\text{noise}}$ を雑音の分離に特化した雑音の基底信号とする．

### 3.2 雑音の再構成損失の計算

Separator では Encoder から出力された特徴量 $\mathbf{w} \in \mathbb{R}^{1 \times N}$ から話者数 $C$ 個のマスクを求める．雑音も対象としてマスクを $C+1$ 個推定し話者だけでなく雑音の再構成に関しても損失を計算する．雑音の基底信号と組み合わせて利用する際に $\mathbf{B}^x_{\text{extend}}$ を雑音の再構成，$\mathbf{B}^x_{\text{clean}}$ を話者の再構成にのみ用いるように制限する．実装上はそれぞれ異なる畳み込み層を用いることで実現した．

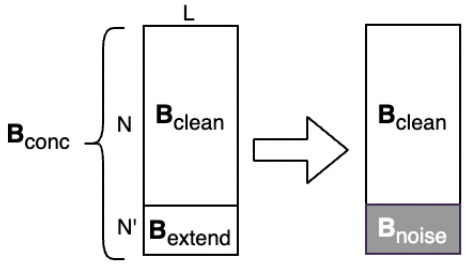


図 2　雑音の基底信号の学習

### 3.3 雑音を段階的に増やすカリキュラム学習

$\mathbf{B}^x_{\text{extend}}$ を設けて雑音の基底信号 $\mathbf{B}^x_{\text{noise}}$ を学習する際，雑音が話者に対して相対的に大きすぎると学習できない可能性がある．そこで，雑音の相対量を段階的に増やすカリキュラム学習 [12] を用いる．カリキュラム学習とは，人間や動物が簡単な概念から徐々に難しい概念を理解していく学習戦略を機械学習において形式化したものであり，簡単なタスクから徐々に難しいタスクを学習させるというカリキュラムによって学習効率及び学習性能の向上を狙う．ここでは SN 比を段階的に小さくすることでタスクを徐々に難化させる 3 段階のカリキュラムを構成した．いずれの段階においても前段階のモデルを初期値として用いる．なお，雑音を含む音声分離で得られたモデルを 1 段階目において初期値として使用した．

（1） 最終的に学習に用いる雑音を含むデータセットより，SN 比を 20dB 向上させたデータセット

（2） 最終的に学習に用いる雑音を含むデータセットより，SN 比を 10dB 向上させたデータセット

（3） 最終的に学習に用いる雑音を含むデータセット

## 4. 実　　験

### 4.1 データセット

通常の音声分離（“Clean”）に用いる 2 話者の混合音声のデータセット wsj0-2mix [13] と，雑音を含む音声分離（“Noisy”）に用いるデータセット WHAM! [3] を用意する．

#### 4.1.1 wsj0-2mix

2 話者の混合音声として，シングルチャネル音声分離において広く使用される wsj0-2mix を使用する．WSJ0 コーパス [14] から 2 発話をランダムに選び，SN 比を-5〜+5dB の範囲でランダムに選択して混合した音声により構成する．学習セットは 20,000 個の計 30 時間ほどの混合音声，バリデーションセットは 5,000 個の計 10 時間ほどの混合音声から成り，WSJ0 コーパスのうち si_tr_s から選んでおり一部の話者は共通している．一方で評価セットは 3,000 個の 5 時間ほどの混合音声から成り，WSJ0 コーパスのうち si_dt_05 と si_et_05 にのみ含まれる話者を用いており他のセットと共通の話者を持たない．バージョンとして “min” と “max” が存在し “max” では発話を切り取らずに混合しているのに対し，“min” では選んだ 2 発話のうち短い方の長さに切り取った上で混合する．本稿では他の多くの研究

と同様に 8kHz の “min” バージョンを用いる．

#### 4.1.2 WHAM!

wsj0-2mix データセットよりも実環境に近い状況を作るために，wsj0-2mix データセットに環境音を雑音として混合したデータセットを作成した．雑音はサンフランシスコ湾エリアのコーヒーショップ，レストラン，バー，オフィスビル，公園などで採取されたものを用いており，音声分離の妨げになると考えられる-6dB 以上の話し声を含むものは使用しない．まず，雑音の SN 比を-6〜+3dB の範囲でランダムに選択し，混合対象の 2 話者のうち音量の大きい一方に対して選択した SN 比になるようにゲインを定めて混合する．次に，もう一方の話者に対しても同じゲインを用いて混合を行う．学習セット・バリデーションセット・評価セットの数は wsj0-2mix データセットと同じである．

### 4.2 ネットワーク

ベースラインとして用いる Asteroid Toolkit における TasNet について説明する．音声を分割するフレーム長 $L$ は 40 とし，Encoder および Decoder の基底信号の数 $N$ は 512 とするが，雑音の基底信号数 $N'$ は 128 とする．Separator には各方向に 600 個のユニットを持つ 4 層の LSTM を使用し，最後の層を除く LSTM の出力には 0.3 のドロップアウトを適用する．基本的に双方向 LSTM を使用して分離性能を評価するが，リアルタイムで処理を行うときは単方向 LSTM を用いる．実装及びパラメータ操作は全てフレームワーク PyTorch [15] 上で行う．学習はバッチサイズを 32 として 200 エポック学習し，学習率は初期値を $1.0e^{-3}$ とした上で 3 エポック連続でバリデーション損失が改善しなかった場合に半減させる．また，10 エポックの間バリデーション損失が改善しなかった場合に学習を停止する early stopping を採用した．正規化として $1.0e^{-5}$ の重み減衰を使用した．最適化アルゴリズムとして Adam [16] を使用する．

ここでは話者数が既知で特に 2 人である状況を想定するが，話者数が未知の場合に時間領域システムで 1 人ずつ音声を抽出するという操作を再帰的に行うことで分離が可能である [17].

### 4.3 実　　験

TasNet-NB によって雑音を含む音声分離を行った結果を表 1 に記す．雑音の基底信号（Noise Basis）や雑音の再構成損失の計算（Noise Loss）およびカリキュラム学習（Curriculum）の有無による分離性能を比較する．評価にはいずれも SI-SDR の分離前後での向上値（SI-SDR improvement: SI-SDRi）を用いた．単位は dB（デシベル）である．

雑音の基底信号を用いてカリキュラム学習したときに SI-SDRi14.6 と最も分離性能が高かったが，雑音の基底信号のみの使用では SI-SDRi14.0 と性能向上はわずかだった．雑音の再構成損失の計算をした場合 SI-SDRi が 0.2 向上した．

基底信号による分離性能の違いを確かめるため，Encoder と Decoder における基底信号の組み合わせを変え，雑音を含む音声分離において比較した．ネットワーク構造が大きく異なるため，雑音の再構成損失の計算を行わない場合と行う場合に分けて検証した．基底信号を表 2 に分類する．学習に用いたデータセットと対応するように命名した．

表 1　各構造による SI-SDR 向上値

| Method | SI-SDRi |
|---|---|
| TasNet (Asteroid Toolkit) | 13.7 |
| +Noise Basis | 14.0 |
| +Noise Loss | 14.2 |
| +Noise Basis +Curriculum | 14.6 |
| +Noise Basis +Noise Loss +Curriculum | 14.6 |

表 2　基底信号の分類

| $\mathbf{B}^x_{\text{clean}}$ | $\mathbf{B}^x_{\text{extend}}$ | 基底信号の名称 |
|---|---|---|
| × | × | Noisy Speech |
| ○ | × | Clean Speech |
| ○ | ○ | Clean Speech&Noise |

表 3　基底信号ごとの性能の比較

| Encoder | Decoder | SI-SDRi | Curriculum |
|---|---|---|---|
| Noisy Speech | Noisy Speech | 13.7 | |
| Clean Speech | Clean Speech | 13.9 | |
| Clean Speech&Noise | Noisy Speech | 14.0 | 14.6 |
| Clean Speech&Noise | Clean Speech | 13.9 | |

まず，雑音の再構成損失の計算を行わない場合の結果を表 3 に記す．ともに “Noisy Speech” の場合，雑音を含む音声分離においてスクラッチで学習したことを示す．話者の基底信号のみを用いても分離性能が向上することを確認した．

次に，雑音の再構成損失の計算を行う場合の結果を表 4 に記す．スクラッチで学習した場合と話者の基底信号を用いた場合に最も分離性能が高くなることを確認した．また，カリキュラム学習を適用しないとき雑音の基底信号を用いただけでは分離性能が向上しなかった．

いずれの場合も雑音の基底信号を使用した中で分離性能が最も高い基底信号の組み合わせに対しカリキュラム学習を適用した．

## 5. 結　　論

本稿では WHAM!データセットを用いた雑音を含む音声分離を TasNet-NB によって行い，カリキュラム学習を行うことで SI-SDRi が 13.7 から 14.6 に向上した．また，学習基準の一つとして雑音の再構成損失を用いることの有効性を確認した．

今後は，残響を含むデータセット WHAMR! [18] を用いて加算性の雑音だけでなく乗算性の雑音に対する効果を検証したい．

### 文　　献

[1] Z. Wang, J. L. Roux, and J. R. Hershey. Alternative objective func-

表 4　基底信号ごとの性能の比較
（雑音の再構成損失を計算）

| Encoder | Decoder | SI-SDRi | Curriculum |
|---|---|---|---|
| Noisy Speech | Noisy Speech | 14.2 | |
| Clean Speech | Clean Speech | 14.2 | |
| Clean Speech&Noise | Noisy Speech | 14.1 | |
| Clean Speech&Noise | Clean Speech | 14.1 | 14.6 |


tions for deep clustering. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 686–690, 2018.

[2] Y. Luo and N. Mesgarani. Tasnet: Time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700, 2018.

[3] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. Wham!: Extending speech separation to noisy environments. In *Proc. Interspeech*, September 2019.

[4] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630. IEEE, 2019.

[5] Manuel Pariente, Samuele Cornell, Joris Cosentino, Sunit Sivasankaran, Efthymios Tzinis, Jens Heitkaemper, Michel Olvera, Fabian-Robert Stöter, Mathieu Hu, Juan M. Martín-Doñas, David Ditter, Ariel Frank, Antoine Deleforge, and Emmanuel Vincent. Asteroid: the PyTorch-based audio source separation toolkit for researchers. In *Proc. Interspeech*, 2020.

[6] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pp. 933–941. PMLR, 2017.

[7] Y. Luo and N. Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 27, No. 8, pp. 1256–1266, 2019.

[8] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 10, pp. 1901–1913, 2017.

[9] David Ditter and Timo Gerkmann. A multi-phase gammatone filterbank for speech separation via tasnet. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 36–40. IEEE, 2020.

[10] Hakan Erdogan and Takuya Yoshioka. Investigations on data augmentation and loss functions for deep learning based speech-background separation. *Proc. Interspeech 2018*, pp. 3499–3503, 2018.

[11] Keisuke Kinoshita, Tsubasa Ochiai, Marc Delcroix, and Tomohiro Nakatani. Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7009–7013. IEEE, 2020.

[12] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

[13] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35. IEEE, 2016.

[14] John S., Garofolo, David Graff, Doug Paul, and David Pallett. Csr-i (wsj0) sennheiser ldc93s6b. *Philadelphia: Linguistic Data Consortium*, 1993.

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Naoya Takahashi, Sudarsanam Parthasaarathy, Nabarun Goswami, and Yuki Mitsufuji. Recursive speech separation for unknown number of speakers. *arXiv preprint arXiv:1904.03065*, 2019.

[18] Matthew Maciejewski, Gordon Wichern, and Jonathan Le Roux. Whamr!: Noisy and reverberant single-channel speech separation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.