正負ラベルなし学習を用いた半教師付き深層学習に基づくモノラル音声強調 *

☆小川諒，△米倉悠記，伊藤信貴，高宗典玄，山岡洸瑛，齋藤佑樹，猿渡洋 (東大)

1 はじめに

モノラル音声強調は，単一のマイクロホンで観測された雑音が重畳した観測音声から雑音のない音声を抽出する技術である．従来は教師付き学習によるものが主流であり，観測音声と雑音のない音声の組からなる教師付きデータを大量に用いることで，高性能な音声強調が実現できる．しかし雑音のない音声の収録は，スタジオ等静かな環境が必要なため高コストである．

本稿では，正負ラベルなし (positive-negative-unlabeled: PNU) 学習 [1] を用いた半教師付き深層学習に基づくモノラル音声強調法を提案する．提案手法は，観測音声のみからなる教師なしデータを活用する．このようなデータは，スマートスピーカなどの定常的にモニタリングしているデバイスやウェブから容易に取得できる．提案手法により，教師付きデータが少量しか得られない状況でも，容易に大量に得られる教師なしデータを活用し，高性能な音声強調が可能になると期待される．

2 背景

2.1 二値分類

二値分類は，入力パターン $\boldsymbol{x} \in \mathbb{R}^d$ の属するクラス $y \in \mathcal{Y} := \{\pm 1\}$ を予測する問題であり，経験的リスクの最小化として定式化される．この予測は，分類器 $f_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}$ ($\boldsymbol{\theta}$ はパラメータ) に $\boldsymbol{x}$ と y の関係を学習させ，予測ラベルを $\widehat{y} := \operatorname{sgn} f_{\boldsymbol{\theta}}(\boldsymbol{x})$ で計算することにより行われる (sgn は符号関数)．パラメータ $\boldsymbol{\theta}$ は，理想的には次式のリスクの最小化により学習する [2]．

$$R(\boldsymbol{\theta}) := \mathbb{E}_{p(\boldsymbol{x},y)} \left[\ell(\boldsymbol{x}, y, \boldsymbol{\theta}) \right] \tag{1}$$

ここで，$\ell(\boldsymbol{x}, y, \boldsymbol{\theta})$ は，$f_{\boldsymbol{\theta}}$ のデータ点 $(\boldsymbol{x}, y)$ からの乖離度を表す非負の損失であり，$p(\boldsymbol{x}, y)$ はデータが従う同時確率密度である．$p(\boldsymbol{x}, y)$ は未知であるため，実際には (1) の期待値は計算できない．そこで，期待値を訓練データを用いた標本平均で近似した経験的リスクを用い，その最小化により $\boldsymbol{\theta}$ を学習する．

2.2 PNU 学習

PNU 学習 [1] は半教師付き二値分類法であり，訓練データとして正例，負例，ラベルなしデータを用いる．正例 $\mathcal{P} := \{\boldsymbol{x}_i^{\mathrm{P}}\}_{i=1}^{n_{\mathrm{P}}}$ は $y = +1$ の下での条件付き確率密度 $p(\boldsymbol{x} \mid y = +1)$ に，負例 $\mathcal{N} := \{\boldsymbol{x}_i^{\mathrm{N}}\}_{i=1}^{n_{\mathrm{N}}}$ は $y = -1$ の下でのそれ $p(\boldsymbol{x} \mid y = -1)$ に，それぞれ独立に従うと仮定する．また，ラベルなしデータ $\mathcal{U} := \{\boldsymbol{x}_i^{\mathrm{U}}\}_{i=1}^{n_{\mathrm{U}}}$ は周辺確率密度 $p(\boldsymbol{x})$ に独立に従うと仮定する．

PNU 学習では，$\mathcal{P}$,$\mathcal{N}$,$\mathcal{U}$ を全て活用するため，PN 学習 [2] と PU 学習 [3] (またはその変形である NU 学習) の経験的リスクの凸結合 (次式) を用いる．

$$\widehat{R}_{\mathrm{PNU}}(\boldsymbol{\theta}) \tag{2}$$
$$:= \begin{cases} \eta \widehat{R}_{\mathrm{PU}}(\boldsymbol{\theta}) + (1-\eta)\widehat{R}_{\mathrm{PN}}(\boldsymbol{\theta}), & 0 \le \eta \le 1 \\ -\eta \widehat{R}_{\mathrm{NU}}(\boldsymbol{\theta}) + (1+\eta)\widehat{R}_{\mathrm{PN}}(\boldsymbol{\theta}), & -1 \le \eta < 0 \end{cases}$$

ここで，$\widehat{R}_{\mathrm{PN}}(\boldsymbol{\theta}) := \pi_{\mathrm{P}} \widehat{R}_{\mathrm{P}}^{+}(\boldsymbol{\theta}) + \pi_{\mathrm{N}} \widehat{R}_{\mathrm{N}}^{-}(\boldsymbol{\theta})$ は，$\mathcal{P}$ と $\mathcal{N}$ からの教師付き学習法である PN 学習 [2] の経験的リスクである．$\widehat{R}_{\mathrm{PU}}(\boldsymbol{\theta}) := \pi_{\mathrm{P}} \widehat{R}_{\mathrm{P}}^{+}(\boldsymbol{\theta}) + \widehat{R}_{\mathrm{U}}^{-}(\boldsymbol{\theta}) - \pi_{\mathrm{P}} \widehat{R}_{\mathrm{P}}^{-}(\boldsymbol{\theta})$ は，$\mathcal{P}$ と $\mathcal{U}$ からの弱教師付き学習法である PU 学習 [3] の経験的リスクである．$\widehat{R}_{\mathrm{NU}}(\boldsymbol{\theta}) := \pi_{\mathrm{N}} \widehat{R}_{\mathrm{N}}^{-}(\boldsymbol{\theta}) + \widehat{R}_{\mathrm{U}}^{+}(\boldsymbol{\theta}) - \pi_{\mathrm{N}} \widehat{R}_{\mathrm{N}}^{+}(\boldsymbol{\theta})$ は，$\mathcal{N}$ と $\mathcal{U}$ からの弱教師付き学習法である NU 学習の経験的リスクである．ただし，$\pi_{\mathrm{P}} := p(y = +1)$，$\pi_{\mathrm{N}} := p(y = -1) = 1 - \pi_{\mathrm{P}}$ はクラス事前確率 (既知と仮定) であり，$\widehat{R}_z^{\pm}(\boldsymbol{\theta}) := (1/n_z) \sum_{i=1}^{n_z} \ell(\boldsymbol{x}_i^z, \pm 1, \boldsymbol{\theta})$ ($z = \mathrm{P}, \mathrm{N}, \mathrm{U}$, 複号同順) と定義する．(2) は (1) の不偏推定量であり，不偏 PNU リスクと呼ばれる．

上記の方法では，過学習の問題があることが知られており，解決策として経験的リスクの非負補正 [2] が提案されている．過学習の原因として，深層ニューラルネットワーク (deep neural network: DNN) などの自由度の高いモデルを用いる場合，(1) が非負であるにも拘らず，その推定値 (2) が負になりうることが指摘されている [4]．非負補正を行う場合，(2) が常に非負になるよう，(2) の $\widehat{R}_{\mathrm{PU}}(\boldsymbol{\theta})$，$\widehat{R}_{\mathrm{NU}}(\boldsymbol{\theta})$ を非負の項

$$\widehat{R}_{\mathrm{nnPU}}(\boldsymbol{\theta}) := \pi_{\mathrm{P}} \widehat{R}_{\mathrm{P}}^{+}(\boldsymbol{\theta}) + \left(\widehat{R}_{\mathrm{U}}^{-}(\boldsymbol{\theta}) - \pi_{\mathrm{P}} \widehat{R}_{\mathrm{P}}^{-}(\boldsymbol{\theta}) \right)_+$$
$$\widehat{R}_{\mathrm{nnNU}}(\boldsymbol{\theta}) := \pi_{\mathrm{N}} \widehat{R}_{\mathrm{N}}^{-}(\boldsymbol{\theta}) + \left(\widehat{R}_{\mathrm{U}}^{+}(\boldsymbol{\theta}) - \pi_{\mathrm{N}} \widehat{R}_{\mathrm{N}}^{+}(\boldsymbol{\theta}) \right)_+$$

に置き換えた非負 PNU リスク $\widehat{R}_{\mathrm{nnPNU}}(\boldsymbol{\theta})$ の最小化により $\boldsymbol{\theta}$ を学習する (ただし，$(x)_+ := \max\{x, 0\}$)．

3 提案手法

提案手法は，マスクを用いて音声強調を行う．提案手法の処理フロー (推論時) を Fig. 1 に示す．まず雑音が重畳した観測音声の短時間 Fourier 変換 (short-time Fourier transform: STFT) を計算する．次に観測音声の STFT の絶対値を取って振幅スペクトログラムを計算する．この振幅スペクトログラムをマスク予測部 (DNN を用いる) に入力し，時間周波数点ごとの重みであるマスクを得る．マスクは，signal-to-noise ratio (SNR) が所定の閾値 T を超えた時間周波数点で 1，それ以外で 0 を取る．最後に，マスクを観測音声の STFT と掛けて強調音声を得，逆 STFT により時間領域に変換する．上述のマスク予測部で用いる DNN を適切に学習させることが鍵となる．

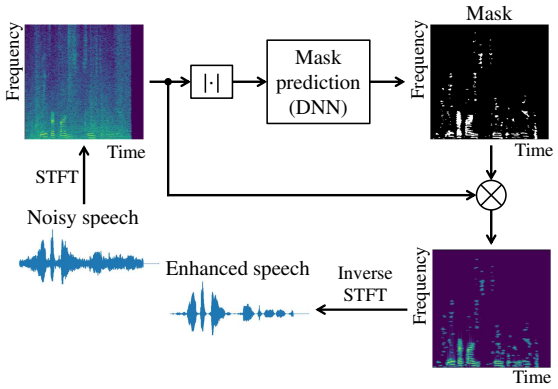

Fig. 1 Processing flow during inference in the proposed method.

マスクの予測は，時間周波数成分の二値分類に帰着する．正クラスを SNR が T を超える時間周波数成分，負クラスをそれ以外と定義する．各時間周波数点における特徴ベクトル $\boldsymbol{x}$ として，その点を中心とする矩形の局所領域 (パッチ) を振幅スペクトログラムか

---

* Monaural speech enhancement based on semi-supervised deep learning using positive-negative-unlabeled machine learning by OGAWA, Ryo, YONEKURA, Yuki, ITO, Nobutaka, TAKAMUNE, Norihiro, YAMAOKA, Kouei, SAITO, Yuki, and SARUWATARI, Hiroshi (the University of Tokyo).

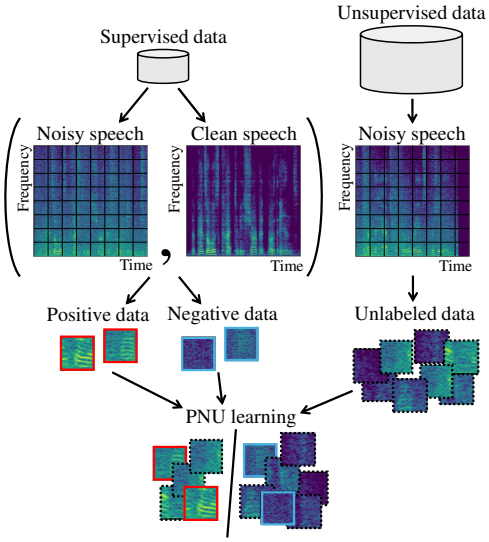

Fig. 2　Training DNN with PNU learning using supervised and unsupervised data.

ら切り出す．$\boldsymbol{x}$ を正または負に分類する $f_{\boldsymbol{\theta}}$ を用いて，マスク $m = \mathbb{I}(f_{\boldsymbol{\theta}}(\boldsymbol{x}) > 0)$ を得る ($\mathbb{I}$ は指示関数)．

分類器 $f_{\boldsymbol{\theta}}$ は，教師付き及び教師なしデータを用いた PNU 学習により訓練できる (Fig. 2 参照)．教師付きデータについては，観測音声と雑音のない音声の組から，時間周波数点ごとの SNR が分かる．よって，この教師付きデータから $\mathcal{P}$,$\mathcal{N}$ が得られる．一方，教師なしデータについては，雑音のない音声が得られないため，時間周波数点ごとの SNR が分からない．よって，教師なしデータからは $\mathcal{U}$ のみ得られる．このように得られた $\mathcal{P}$,$\mathcal{N}$,$\mathcal{U}$ を用いて PNU 学習を行い，時間周波数成分を分類するように $f_{\boldsymbol{\theta}}$ を学習させ，これを用いてマスクが得られる．

4　実験

提案手法により教師付き及び教師なしデータを用いたモノラル音声強調が実現可能であることを確認するため，小規模データを用いた予備実験を行った．

4.1　実験条件

本実験では，提案手法 (PNU 学習) と従来手法 (PN 学習) の 2 手法を比較した．PNU 学習では 2 節の $\widehat{R}_{\mathrm{nnPNU}}(\boldsymbol{\theta})$ において $\eta = 0.2$ としたものを用い，PN 学習では $\widehat{R}_{\mathrm{PN}}(\boldsymbol{\theta})$ を用いた．いずれもクラス事前確率 $\pi_{\mathrm{P}} = 0.2$ とした．また，損失 $\ell(\boldsymbol{x}, y, \boldsymbol{\theta})$ は振幅重み付きシグモイド損失 [5] を用いた．

データセットは，Matlab toolbox for DNN-based speech separation [6] で用いられるデータセットと DEMAND [7] から作成した．前者は教師付き訓練及び検証データとして，後者は教師なし訓練データ及びテストデータの雑音として用いた．これは，多くの場合，教師付き訓練及び検証データの雑音環境はテストデータと適合しないが，教師なしデータはテストデータと適合した雑音環境で収録できるためである．教師付き訓練データには，文献 [6] の訓練データの 12 番目のサンプルを用いた．提案手法で用いる教師なし訓練データは，DEMAND の DKITCHEN, DLIVING, DWASHING, NFIELD, NRIVER の環境からランダムに切り出した雑音に，文献 [6] の訓練データの 81〜480 番目のサンプルの雑音のない音声を加算して作成した．提案手法では，作成した 400 クリップの教師なし訓練データのうち，1, 200, 400 クリップを用いる 3 つの場合について実験を行った．検証データには，文献 [6] の訓練データの 481〜600 番目のサンプルを用いた．テストデータは，DEMAND の STRAFFIC, TCAR の環境からランダムに切り出した雑音に，文献 [6] のテストデータの雑音のない音声 120 クリップを加算して作成した．教師付き訓練データの SNR は 5 dB とし，その他のデータの SNR は $[-5, 10]$ dB の範囲から一様にサンプリングした．すべてのデータのサンプリング周波数は 16 kHz，信号長は 3.125 s であった．

Table 1　SI-SNRi on the test data. Q is the number of noisy speech clips in the unlabeled training data

Method	Q	SI-SNRi (dB)
Conventional (PN)	0	3.52 (±3.04)
Proposed (PNU)	1	1.32 (±0.28)
Proposed (PNU)	200	5.38 (±2.13)
Proposed (PNU)	400	6.55 (±1.35)

計算効率化のため，パッチごとに DNN を適用する代わりに，振幅スペクトログラム全体を入力とし，全ての時間周波数点での $f_{\boldsymbol{\theta}}$ の値を出力する convolutional neural network (CNN) を用いた [5]．振幅スペクトログラムの計算において，STFT の窓長は 64 ms，シフト長は 16 ms とし，窓関数はハミング窓とした．CNN は次の 7 層からなる：Conv2d(1,8,3)-Conv2d(8,8,3)-Conv2d(8,16,3)-Conv2d(16,16,3)-Conv2d(16,32,1)-Conv2d(32,32,1)-Conv2d(32,1,1)．ここで，Conv2d($C_{\mathrm{in}}, C_{\mathrm{out}}, K$) は入力チャネル数 C_{in}，出力チャネル数 C_{out}，カーネルサイズ $K \times K$，ストライド $(1, 1)$，same パディングの 2 次元畳み込み層である．最終層以外で畳み込みの後に正規化線形ユニットを適用し，0.05 の確率でドロップアウトを行った．

実験は PyTorch を使い，4 台の NVIDIA V100 による分散学習により CNN を訓練した．最適化手法は Adam，エポック数は 100，学習率は 5×10^{-5}，バッチサイズは 8 とした．提案手法においては，各バッチは教師付き及び教師なしデータを半数ずつ含むように作成した．1250 ステップを 1 エポックと定義した．各エポック終了時にモデルを保存し，検証データにおける音声強調性能 scale-invariant signal-to-noise ratio improvement (SI-SNRi) が最大となるモデルを選択した．時間周波数成分のクラスの閾値 SNR は，$T = 0$ dB とした．

4.2　実験結果

Table 1 にテストデータにおける SI-SNRi を示す．表の数値は，5 回の試行に対する SI-SNRi の平均と標準偏差である．提案手法の SI-SNRi の平均は，教師なし訓練データのクリップ数 $Q = 1, 200, 400$ のとき，それぞれ 1.32, 5.38, 6.55 dB であり，Q の増加に伴い音声強調性能が向上した．$Q = 200, 400$ のとき，提案手法の SI-SNRi は従来手法の 3.52 dB をそれぞれ 1.86, 3.03 dB 上回った．このように，今回の予備実験では，提案手法は教師なしデータを活用し，従来の教師付き学習に基づく手法より高い音声強調性能を実現できた．大規模データを用いた評価は今後の課題とする．

謝辞　本研究は，公益財団法人立石科学技術振興財団 2023 年度研究助成 (S) 及び東京大学 Beyond AI 研究推進機構の支援により実施された．実験に産総研の AI 橋渡しクラウド (ABCI) を利用した．ABCI について中村友彦氏，中田亘氏に助言を受けた．

参考文献

[1] Sakai *et al.*, Proc. ICML, 2998–3006, 2017.
[2] Sugiyama *et al.*, “Machine Learning from Weak Supervision: An Empirical Risk Minimization Approach,” MIT Press, 2022.
[3] du Plessis *et al.*, Proc. ICML, 1386–1394, 2015.
[4] Kiryo *et al.*, Adv. Neural Inf. Process. Syst., 30, 1674–1684, 2017.
[5] Ito and Sugiyama, Proc. ICASSP, 2023. doi:10.1109/ICASSP49357.2023.10095988
[6] https://u.osu.edu/pnlab/software/
[7] Thiemann *et al.*, Proc. Mtgs. Acoust., 19 (1) , 2013.