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
SP2017-20(2017-08)

深層生成モデルを事前分布に用いた教師なし音声強調

坂東　宜昭† 　三村　正人† 　糸山　克寿† 　吉井　和佳†,†† 河原　達也†

† 京都大学 大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
†† 理化学研究所 革新知能統合研究センター　〒 103-0027 東京都中央区日本橋 1-4-1
E-mail: {yoshiaki, mimura, itoyama, yoshii, kawahara}@sap.ist.i.kyoto-u.ac.jp

あらまし　本稿では，深層生成モデルを事前分布に用いた教師なし音声強調について述べる．近年，DNN を用いて，雑音を含む音声信号からクリーンな音声信号への写像を教師あり学習することで，高品質な音声強調が実現されつつある．しかし，このアプローチでは，大量の訓練データ（入出力のペア）を準備する必要があるうえ，未知の雑音環境下に対する汎化性能に問題があった．一方，音声スペクトルと雑音スペクトルの統計的な性質に着目することで，雑音環境に依存せずに，教師なし音声強調を行う方法も提案されている．しかし，このアプローチでは，仮定した音声スペクトルの統計モデルが貧弱で，強調された音声信号の品質に限界があった．これらの問題を解決するため，本研究では，DNN と従来の統計モデルを確率的に統合した教師なし音声強調法を提案する．本手法では，雑音スペクトルは非負値行列因子分解モデルから，音声スペクトルは深層生成モデルから確率的に生成され（事前分布），それらが重畳することで混合音スペクトルが生成される（尤度関数）と考える．このとき，大量のクリーンな音声信号を用いて，音声スペクトルの深層生成モデル（事前分布）をあらかじめ教師なし学習しておけば，混合音が与えられたときに，含まれている実際の音声スペクトル（事後分布）を MCMC を用いてベイズ推論することができる．シミュレーション混合音を用いた評価実験で，その有効性を確認した．

キーワード　ベイズ信号処理，深層生成モデル，変分オートエンコーダ，非負値行列因子分解

1. はじめに

雑音環境下でも頑健に動作する音声認識や遠隔対話システムを実現するために，音声強調が研究されている [1–8]．口元から離れたマイクロホンを用いて録音した音響信号には，目的音声だけでなく，周囲の雑音が混入し，音声認識や音声変換などの性能劣化を招く．音声強調は，入力音響信号に含まれる雑音を抑圧し音声を抽出する技術として広く研究されており，教師あり音声強調と教師なし音声強調に大別できる．

教師あり音声強調は，入力である混合音と教師信号である目的音声との間の写像を機械学習することで，高い品質で音声強調できる．教師あり音声強調では，深層ニューラルネットワーク (Deep Neural Network: DNN) に基づく音声強調 [1,2] が注目されており，例えば DAE (Denoising AutoEncoder) [2] が知られている．DNN は，高次元かつ非線形な写像を効率的に学習できる．そのため，DNN に基づく教師あり音声強調は，既知の雑音環境下で高い強調性能を発揮できる．一方で，未知の雑音環境下では必ずしも有効とは限らず，使用環境に応じて大量の訓練データを準備する必要があった．

教師なし音声強調は，音声信号と雑音信号の統計モデルを仮定し，それらの構造の違いから音声と雑音を推定する [3–5,9–12]．例えば，Wiener filter に基づく音声強調法 [9] では，雑音信号の定常性を仮定し，定常的な雑音信号を推定・除去する．非定常な信号のモデルとして，非負値行列因子分解 (Non-negative Matrix Factorization: NMF) が提案されている [4,12–14]．NMF は，雑音と音声のパワースペクトログラムが低ランクであると仮定する．つまり，各音源のスペクトルが少数の基底スペクトルの重み付き和で表現できると仮定する．事前にこの基底スペクトルを教師なし学習することで，抽出したい音源信号を観測信号から推定することができる [4,13]．NMF を用いた事前学習を行わない音声強調法としてロバスト NMF (Robust NMF: RNMF) が提案されている [15–17]．RNMF は音声スペクトログラムにスパース性を仮定し，NMF モデルを仮定した雑音との統計的構造の違いから，音声と雑音を事前学習せずに分離できる．しかし，統計モデルに基づく音声強調法には，観測とのモデル誤差によって性能が劣化する問題がある．たとえば，RNMF が仮定する音声のスパース性は，音声の調波構造や時間連続性を考慮できず，性能劣化の原因となっていた．

本稿では，NMF による雑音モデルと DNN を用いた音声モデルを統合した教師なし音声強調法について述べる．本手法は，深層生成モデルの一つである変分オートエンコーダ (Variational AutoEncoder: VAE) [18,19] を用いて音声をモデル化する．VAE は，訓練データが従う確率分布を DNN を用いて学習する手法である．提案法は，クリーン音声データセットを教師なし学習した VAE で音声の事前分布を構成するので, より自然な音声の推定を実現できる．また，雑音には NMF モデルを仮定することで，環境に依存しやすい雑音を事前学習せずに推定・抑圧する．VAE 音声モデルと NMF 雑音モデルは，それぞれ事前分布として単一の統計的生成モデル (以下，VAE-NMF) に統合される．VAE-NMF は，マルコフ連鎖モンテカルロ法

(MCMC: Markov Chain Monte Carlo) [20] による事後分布推論により，観測信号から音声と雑音を推定・分離する．

2. 深層生成モデル

本節では，VAE-NMF を設計するために深層生成モデルについて概観する．深層生成モデルは，訓練データの各サンプルが従う確率分布を DNN を用いて学習するために研究されている．表現力の高い DNN によって訓練データの分布を学習するため，画像や音響信号といった多変量変数の従う分布を効率的に学習できる．以降では，F 次元で T 個のデータからなる訓練データ $\mathbf{X}$ の各サンプルを $\boldsymbol{x}_t \in \mathbb{R}^F$ $(t = 1, \ldots, T)$ と表す．$\boldsymbol{x}_t$ は非負実数や複素数，離散値に拡張できるが，簡単のため実数の場合のみを扱う．

2.1 敵対的生成ネットワーク

敵対的生成ネットワーク (Generative Adversarial Network: GAN) が近年，深層生成モデルの一つとして大きく注目されている [21, 22]．GAN はまず，多変量標準ガウス分布に従う潜在変数 $\boldsymbol{z}_t \in \mathbb{R}^D$ を仮定し，訓練データの各サンプル $\boldsymbol{x}_t$ が，$\boldsymbol{z}_t$ を非線形関数 $f : \mathbb{R}^D \to \mathbb{R}^F$ で変換して得られると考える．

$$\boldsymbol{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{1}) \tag{1}$$

$$\boldsymbol{x}_t = f(\boldsymbol{z}_t) \tag{2}$$

ここで $\mathcal{N}(\mu, \sigma)$ は，平均 μ で分散 σ のガウス分布を表す．この関数 f は DNN で定義された Generator と呼ばれ，訓練データから学習される．潜在変数 $\boldsymbol{z}_t$ の各次元の役割は f の学習時に自動的に決定される．ある $\boldsymbol{x}_t$ の現れやすさは，対応する $\boldsymbol{z}_t$ の生起確率で表現される．

Generator ネットワークは，Discriminator と呼ばれる DNN と同時に訓練データを用いて学習する．Discriminator ネットワークは，あるサンプルが訓練データ内のサンプルか，Generator が生成したサンプルかを識別するネットワークである．GAN の学習では，この Discriminator が誤判定するように Generator ネットワークを学習する．GAN は，Generator のサンプルの品質を Discriminator を用いて与えるため，主観的に高品質なサンプルを生成でき，音声変換などに応用されている [23]．一方で，GAN により学習された分布の確率密度を計算するには非線形関数 f の逆関数を求める必要があるので，統計的生成モデルの事前分布への応用が困難である．

2.2 変分オートエンコーダ

訓練データの確率分布を学習する別の方法として，VAE が研究されている [18, 19]．VAE も標準ガウス分布に従う潜在変数 $\boldsymbol{z}_t \in \mathbb{R}^D$ を仮定する．VAE は，GAN における決定的な非線形変換関数 f の代わりに，訓練データの各サンプル $\boldsymbol{x}_t$ が条件付き分布 $p(\boldsymbol{x}_t | \boldsymbol{z}_t)$ から生成されると考える．

$$\boldsymbol{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{1}) \tag{3}$$

$$\boldsymbol{x}_t \sim p(\boldsymbol{x}_t | \boldsymbol{z}_t) \tag{4}$$

この条件付き分布 (尤度関数) は，計算が容易な確率密度関数として定式化され，その密度関数のパラメータを DNN による非線形関数で与える．例えば，Kingma ら [18] は，平均パラメータが非線形関数 $\mu_f^{\boldsymbol{x}}(\boldsymbol{z}_t) : \mathbb{R}^D \to \mathbb{R}$ であるガウス尤度を持つ VAE モデルを報告している．

$$x_{ft} \sim \mathcal{N}\left(\mu_f^{\boldsymbol{x}}(\boldsymbol{z}_t), 1\right) \tag{5}$$

VAE は $\boldsymbol{x}_t$ の条件付き確率を定義するので，他のベイズモデルと容易に統合できる．

VAE の学習の目的は，周辺尤度を最大にする尤度関数 $p(\boldsymbol{x}_t | \boldsymbol{z}_t)$ を求めることである．

$$\underset{p(\boldsymbol{x}_t|\boldsymbol{z}_t)}{\operatorname{argmax}} p(\mathbf{X}) = \underset{p(\boldsymbol{x}_t|\boldsymbol{z}_t)}{\operatorname{argmax}} \prod_{d,t} \int p(\boldsymbol{x}_t | \boldsymbol{z}_t) p(\boldsymbol{z}_t) d\boldsymbol{z}_t \tag{6}$$

この周辺尤度の計算は解析的に困難なので，VAE では変分ベイズ法 [20] を用いて周辺尤度を近似する．変分ベイズ法ではまず，$\boldsymbol{z}_t$ の事後分布を以下の変分事後分布 $q(\boldsymbol{z}_t)$ の積で近似する:

$$p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_t | \mathbf{X}) \approx \prod_t q(\boldsymbol{z}_t) = \prod_{d,t} q(z_{dt}) \tag{7}$$

$$= \prod_{d,t} \mathcal{N}(\mu_d^{\boldsymbol{z}}(\boldsymbol{x}_t), \sigma_d^{\boldsymbol{z}}(\boldsymbol{x}_t)) \tag{8}$$

ここで，$\mu_d^{\boldsymbol{z}} : \mathbb{R}^F \to \mathbb{R}$ と $\sigma_d^{\boldsymbol{z}} : \mathbb{R}^F \to \mathbb{R}_+$ は，DNN を用いた非線形関数で，変分事後分布を表すガウス分布の平均と分散パラメータである．この変分事後分布を用いて，対数周辺尤度 $\log p(\mathbf{X})$ の下限 (変分下限) を以下のようにとり近似する．

$$\log p(\mathbf{X}) = \sum_k \log \int p(\boldsymbol{x}_t | \boldsymbol{z}_t) p(\boldsymbol{z}_t) d\boldsymbol{z}_t \tag{9}$$

$$\geqq \sum_k \int q(\boldsymbol{z}_t) \log \frac{p(\boldsymbol{x}_t | \boldsymbol{z}_t) p(\boldsymbol{z}_t)}{q(\boldsymbol{z}_t)} d\boldsymbol{z}_t \tag{10}$$

$$= \sum_k \mathbb{KL}\left[q(\boldsymbol{z}_t) | p(\boldsymbol{z}_t)\right] + \sum_k \mathbb{E}_q\left[\log p(\boldsymbol{x}_t | \boldsymbol{z}_t)\right] \tag{11}$$

ただし，$\mathbb{KL}[\cdot|\cdot]$ は Kullback-Leibler 義距離を表す．VAE の学習では，この変分下限が最大になるように $q(\boldsymbol{z}_t)$ と $p(\boldsymbol{x}_t | \boldsymbol{z}_t)$ を表す DNN を最適化する．式 (11) の第一項は解析的に計算可能で，第二項はモンテカルロ法で近似できるので，確率的勾配降下法 (Stochastic Gradient Descent: SGD) などを用いて最適化できる．

3. VAE と NMF に基づく混合音生成モデル

本節では，VAE に基づく音声モデルと NMF に基づく雑音モデルを統合した VAE-NMF を説明する．

3.1 問題設定

本稿で扱う音声強調の問題設定を以下に示す．

入力: 雑音と音声の混合音複素スペクトログラム $\mathbf{X} \in \mathbb{C}^{F \times T}$
出力: 音声強調された音声複素スペクトログラム $\mathbf{S} \in \mathbb{C}^{F \times T}$

ここで，F および T は，それぞれ周波数ビン数と，時間フレーム数を表す．複素スペクトログラムは，時間領域信号を短時間フーリエ変換 (Short Time Fourier Transform: STFT) することで得られる．

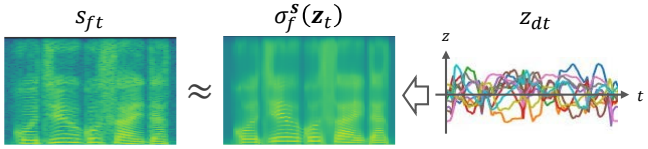


図 1: 音声スペクトログラムの VAE 事前分布による表現の概要

3.2 VAE に基づく音声事前分布

VAE に基づく音声事前分布では，各時間フレームごとの音声の特徴を表す D 次元潜在変数 $\mathbf{Z} \in \mathbb{R}^{D\times T}$ を仮定する．各時刻の潜在変数 $\boldsymbol{z}_t$ は，その時刻での F0 やスペクトル包絡，音素といった音声を表現する特徴量を想定するが，$\boldsymbol{z}_t$ が具体的にどういった特徴を表すかは，クリーン音声信号の訓練データから VAE を用いて機械学習する．従来の VAE と同じように，潜在変数 $\mathbf{Z}$ には，以下のように標準ガウス分布を仮定する．

$$z_{dt} \sim \mathcal{N}(0,1) \tag{12}$$

ある $\mathbf{Z}$ の音声らしさは，このガウス分布の生起確率を計算することで計測できる．

音声信号は，主にそのパワースペクトル密度 (Power Spectral Density: PSD) によって特徴付けることができる．よって，音声の複素スペクトログラム $\mathbf{S}$ は，分散が $\mathbf{Z}$ で定義される平均 0 の複素ガウス分布に従っていると仮定する (図 1).

$$s_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sigma_f^{\boldsymbol{s}}(\boldsymbol{z}_t)\right) \tag{13}$$

ここで，$\mathcal{N}_{\mathbb{C}}(\mu,\sigma)$ は，平均 μ かつ分散 σ の複素ガウス分布を表す．また，$\sigma_f^{\boldsymbol{s}}(\boldsymbol{z}_t): \mathbb{R}^D \to \mathbb{R}_+$ は，$\mathbf{Z}$ と音声信号 $\mathbf{S}$ の関係を表す DNN を用いた非線形関数で，VAE を学習して得る．

3.3 VAE 事前分布を用いた混合音の生成モデル

VAE-NMF では，入力スペクトログラム $\mathbf{X}$ が，音声スペクトログラム $\mathbf{S}$ と雑音スペクトログラム $\mathbf{N} \in \mathbb{C}^{F\times T}$ の和で表現できると考える．

$$x_{ft} = s_{ft} + n_{ft} \tag{14}$$

音声信号 $\mathbf{S}$ に対しては前節で述べた VAE に基づく階層事前分布 (式 (12) および (13)) を仮定する．一方で，雑音スペクトログラムはその PSD が低ランクであることを仮定し，NMF 事前分布を置く．以下のように，雑音事前分布の分散パラメータを，K 個の基底スペクトル $\mathbf{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K] \in \mathbb{R}_+^{F\times K}$ とそれらの重み行列 $\mathbf{H} \in \mathbb{R}_+^{K\times T}$ で表現する．

$$n_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_k w_{fk}h_{kt}\right) \tag{15}$$

雑音事前分布のパラメータ $\mathbf{W}$ および $\mathbf{H}$ は，ベイズ推定を行うために，複素ガウス分布の共役事前分布であるガンマ分布を以下のように仮定する．

$$w_{fk} \sim \mathcal{G}(a_0, b_0) \tag{16}$$

$$h_{kt} \sim \mathcal{G}(a_0, b_0) \tag{17}$$

ここで，$\mathcal{G}(a,b)$ は，シェイプパラメータ a とレートパラメータ b を持つガンマ分布を表し，a_0 および b_0 は，それぞれ $\mathbf{W}$ と $\mathbf{H}$ のハイパーパラメータである．

本モデルは，音声スペクトログラム $\mathbf{S}$ および雑音スペクトログラム $\mathbf{N}$ を積分消去することで，以下の尤度関数が得られる．

$$x_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sigma_f^{\boldsymbol{s}}(\boldsymbol{z}_t) + \sum_k w_{fk}h_{kt}\right) \tag{18}$$

また，この尤度関数は入力スペクトログラム $\mathbf{X}$ の位相成分に依存しないので，さらに位相を積分消去すると，以下の指数分布に基づく尤度関数が得られる．

$$\|x_{ft}\|^2 \sim \mathrm{Exp}\left(\sigma_f^{\boldsymbol{s}}(\boldsymbol{z}_t) + \sum_k w_{fk}h_{kt}\right) \tag{19}$$

ここで，$\|x_{ft}\|^2$ は，x_{ft} のパワーを表し，$\mathrm{Exp}(\lambda)$ は平均 λ の指数分布を表す．パワースペクトログラムに対する指数分布に基づく尤度関数の最大化は，音源分離で広く用いられている板倉斎藤伊距離の最小化に対応している．

3.4 VAE 事前分布の学習

VAE 事前分布の学習の目的は，クリーン音声の訓練データ (本節では $\mathbf{S} \in \mathbb{C}^{F\times T}$ と表記する) から以下に示す周辺尤度 $p(\mathbf{S})$ を最大にする $p(\mathbf{S}|\mathbf{Z})$ を見つけることである．

$$p(\mathbf{S}) = \int p(\mathbf{S}|\mathbf{Z})\, p(\mathbf{Z})\, d\mathbf{Z} \tag{20}$$

式 (13) に示す $p(\mathbf{S}|\mathbf{Z})$ は DNN による非線形変換を含むので，この周辺尤度を計算することができない．そこで，従来の VAE と同じく，$\mathbf{Z}$ の事後分布を近似した変分事後分布 $q(\mathbf{Z})$ を仮定し，周辺尤度の変分近似を行う．本モデルの $p(\mathbf{S}|\mathbf{Z})$ は，音声スペクトログラム $\mathbf{S}$ の位相成分に依存しないので，本稿では，$q(\mathbf{Z})$ も位相を無視して以下のように設定する．

$$q(\mathbf{Z}) = \prod_{d,t} q(z_{dt}) = \prod_{d,t} \mathcal{N}\left(\mu_d^{\boldsymbol{z}}\left(\|\boldsymbol{s}_t\|^2\right), \sigma_d^{\boldsymbol{z}}\left(\|\boldsymbol{s}_t\|^2\right)\right) \tag{21}$$

ここで，$\mu_d^{\boldsymbol{z}}: \mathbb{R}_+^F \to \mathbb{R}$ および $\sigma_d^{\boldsymbol{z}}: \mathbb{R}_+^F \to \mathbb{R}_+$ は，それぞれ DNN を用いた非線形関数で，変分事後分布を表すガウス分布の平均パラメータと分散パラメータである．対数周辺尤度は，変分近似により以下のように近似計算できる．

$$\log p(\mathbf{S}) \geqq \mathbb{KL}\left[q(\mathbf{Z})\,|p(\mathbf{Z})\right] + \mathbb{E}_q\left[\log p(\mathbf{S}|\mathbf{Z})\right] \tag{22}$$

$$= \sum_{d,t} \frac{1}{2}\left\{\left(\mu_d^{\boldsymbol{z}}(\|\boldsymbol{s}_t\|^2)\right)^2 + \sigma_d^{\boldsymbol{z}}(\|\boldsymbol{s}_t\|^2) - \log \sigma_d^{\boldsymbol{z}}(\|\boldsymbol{s}_t\|^2)\right\}$$

$$+ \sum_{f,t} \mathbb{E}_q\left[-\log \sigma_f^{\boldsymbol{s}}(\boldsymbol{z}_t) - \frac{\|s_{ft}\|^2}{\sigma_f^{\boldsymbol{s}}(\boldsymbol{z}_t)}\right] + \mathrm{const.} \tag{23}$$

この変分下限が最大となるように，$\sigma_f^{\boldsymbol{s}}$ および $\mu_n^{\boldsymbol{z}}$，$\sigma_n^{\boldsymbol{z}}$ を SGD を用いて最適化する．

3.5 MCMC に基づくベイズ推論

雑音と音声の混合音から音声成分を推定するために，事後分布 $p(\mathbf{W},\mathbf{H},\mathbf{Z}|\mathbf{X})$ の計算する．本事後分布は解析的に計算が困難なので，MCMC [20] を用いて事後分布を近似する．MCMC

Algorithm 1 VAE-NMF モデルの事後分布サンプリング

for $i = 1, 2, 3, \ldots$ **do**
  for $k = 1, 2, 3, \ldots, K$ **do**
    式 (26) と (27) から補助変数を更新
    式 (24) を用いて $\boldsymbol{w}_k = [w_{1k}, \ldots, w_{Fk}]^T$ をサンプル
    式 (26) と (27) から補助変数を更新
    式 (25) を用いて $\boldsymbol{h}_k = [h_{k1}, \ldots, h_{kT}]$ をサンプル
  end for
  for $t = 1, 2, 3, \ldots, T$ **do**
    式 (28) を用いて $\boldsymbol{z}_t$ をサンプル
  end for
end for

は，事後分布を有限個のサンプル点で近似する手法で，各潜在変数 ($\mathbf{W}$ および $\mathbf{H}$，$\mathbf{Z}$) を他の変数を固定した条件付き事後分布から交互にサンプルする (Algorithm 1)．

雑音の潜在変数 $\mathbf{W}$ と $\mathbf{H}$ は以下の条件付き事後分布からサンプルできる．

$$w_{fk}|\mathbf{H}, \mathbf{Z} \sim \mathcal{GIG}\left(a_0, b_0 + \sum_t \frac{h_{kt}}{\lambda_{ft}}, \sum_t |x_{ft}|^2 \frac{\phi_{ftk}^2}{h_{kt}}\right) \tag{24}$$

$$h_{kt}|\mathbf{W}, \mathbf{Z} \sim \mathcal{GIG}\left(a_0, b_0 + \sum_f \frac{w_{fk}}{\lambda_{ft}}, \sum_f |x_{ft}|^2 \frac{\phi_{ftk}^2}{w_{fk}}\right) \tag{25}$$

ここで，$\mathcal{GIG}(\gamma, \rho, \tau) \propto x^{\gamma-1}\exp(-\rho x - \tau/x)$ はパラメータ γ と ρ, τ を持つ一般化逆ガウス分布を表す．また，λ_{ft} と ϕ_{ftk} は補助変数を表し，一つ前のサンプルを用いて以下で与えられる．

$$\phi_{ftk} = \frac{w_{fk}h_{kt}}{\sum_k w_{fk}h_{kt} + \sigma_f^{\boldsymbol{s}}(\boldsymbol{z}_t)} \tag{26}$$

$$\lambda_{ft} = \sum_k w_{fk}h_{kt} + \sigma_f^{\boldsymbol{s}}(\boldsymbol{z}_t) \tag{27}$$

一方で，音声の潜在変数 $\mathbf{Z}$ は条件付き事後分布を計算できないので，以下の提案分布を用いたメトロポリス・ヘイスティング法 (Metropolis-Hasting: MH) を用いてサンプルする．

$$z_{dt}^* \sim q(z_{dt}^*|z_{dt}) = \mathcal{N}(z_{dt}, \sigma^*) \tag{28}$$

ここで，σ^* は提案分布の分散パラメータを表す．

3.6 複素スペクトログラムの復元

本稿では，事後確率 $p(\mathbf{S}|\mathbf{X}, \mathbf{W}, \mathbf{H}, \mathbf{Z})$ が最大となる $\mathbf{S}$ を音声強調結果として出力する．事後確率を最大にする $\mathbf{S}$ を $\hat{\mathbf{S}} \in \mathbb{C}^{F\times T}$ とすると，$\hat{\mathbf{S}}$ は以下で得られる．

$$\hat{s}_{ft} = \frac{\sigma_f(\boldsymbol{z}_t)}{\sum_k w_{fk}h_{kt} + \sigma_f(\boldsymbol{z}_t)} x_{ft}. \tag{29}$$

4. 評価実験

騒音環境下音声認識の国際技術評議会 CHiME-3 Challenge [24] で使用されたデータセットを用いて性能評価を行った．

4.1 実験設定

CHiME-3 では，タブレット端末に装着したマイクロホンアレイに対して読み上げた音声の認識が行われた．バス (BUS)，カフェテリア (CAF)，歩行者エリア (PED)，車道 (STR) の 4 種類の雑音環境での実録音発話が提供されている．また，これらの環境での雑音のみの録音信号も公開されている．付属のツールキットを用いることで，新聞読み上げ音声コーパス WSJ0 を，任意の信号対雑音比 (Signal-to-Noise Ratio: SNR) で混合したシミュレーション混合音を生成できる．

本実験では，CHiME-3 で提供されたツールキットを用いたシミュレーション混合音を用いて音声強調性能を評価した．目的音声は，WSJ0 に含まれる男女 2 名ずつでそれぞれ 2 発話，計 8 発話である．これを，上記の 4 種類の雑音信号に SNR が 0 dB となるように混合した 32 個の混合音で評価した．CHiME-3 では，6 チャネルのマイクロホンアレイで音響信号が収録されているが，このうち 5 チャネル目を本実験での入力音響信号とした．混合音のサンプリング周波数は 16 kHz である．評価尺度には，強調音の信号対歪比 (Signal-to-Distortion Ratio: SDR) [25] を用いて計測した．SDR は総合的な音声の強調精度を表し，計算には MIR-EVAL [26] を用いた．

比較手法として，RNMF [27] を評価した．この RNMF は以下のように，観測の振幅スペクトログラム $\mathbf{X} \in \mathbb{R}_+^{F\times T}$ を NMF モデルで表す雑音成分と，スパース音声スペクトログラム $\mathbf{S} \in \mathbb{R}_+^{F\times T}$ に分解する．

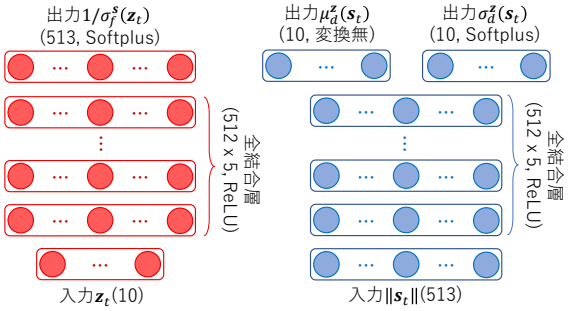
$$x_{ft} \approx \sum_k w_{fk}h_{kt} + s_{ft} \tag{30}$$

ここで，w_{fk} と h_{kt} はそれぞれ，雑音スペクトログラムの基底スペクトルとその重みを表す．VAE-NMF では音声の複素スペクトログラムに VAE 事前分布を仮定したが，RNMF では非ゼロの時間周波数ビンの個数が少なくなるように振幅スペクトログラムにスパース事前分布が仮定されている．

VAE-NMF の各パラメータは以下を使用した．STFT のシフト長と窓幅はそれぞれ，160 サンプルと 1024 サンプルとした．NMF 雑音モデルの基底数 K は 5 とし，$\mathbf{W}$ と $\mathbf{H}$ のハイパーパラメータ a_0 と b_0 はそれぞれ，1.0, $\sqrt{K/scale}$ とした．ここで，$scale$ は入力のパワースペクトログラムの平均値を表す．音声の潜在変数 $\mathbf{Z}$ の次元 D は 10 とした．$\mathbf{Z}$ をサンプルするための提案分布のパラメータ σ^* には，0.01 を用いた．これらの値は実験的に決定した．VAE-NMF のサンプリングは，$\mathbf{W}$，$\mathbf{H}$，$\mathbf{Z}$ を交互に 1000 回サンプルしたのち，これらを 50 回サンプルした結果の平均を出力とした．

4.2 VAE 事前分布の学習

図 2 に示す DNN を用いて，音声の事前分布 $p(\boldsymbol{s}_t|\boldsymbol{z}_t)$ と変分事後分布 $q(\boldsymbol{z}_t|\boldsymbol{s}_t)$ を構成した．それぞれ 5 層の中間層を持つ．本実験ではこれらを，英語新聞読み上げ音声コーパス WSJ0 と，日本語新聞読み上げ音声コーパス JNAS [28] で学習した．WSJ0 コーパスには，約 15 時間の読み上げ音声が含まれている．ただし，本学習で用いた WSJ0 コーパスは，評価に用いる混合音のクリーン音声と同じデータセットであり，WSJ0 コーパスで学習した VAE-NMF での評価はクローズド・テストである．オープン・テストを実施するため，日本語の読み上げコーパ



(a) $p(\boldsymbol{s}_t|\boldsymbol{z}_t)$ (b) $q(\boldsymbol{z}_t|\boldsymbol{s}_t)$

図 2: DNN による $p(\boldsymbol{s}_t|\boldsymbol{z}_t)$ と $q(\boldsymbol{z}_t|\boldsymbol{s}_t)$ の構成

表 1: 音声強調結果 (SDR)

手法	平均	BUS	CAF	PED	STR
VAE-NMF (WSJ0)	6.26	7.30	**5.30**	5.28	7.14
VAE-NMF (JNAS)	**6.80**	**8.62**	5.17	**5.55**	**7.86**
RNMF	5.00	6.72	3.94	3.79	5.54
入力	2.01	1.82	1.96	2.11	2.18

ス JNAS を使用した．JNAS コーパスのうち，約 23 時間の音素バランス文読み上げ音声を学習に使用した．学習には，SGD の一種であり，鞍点での学習効率が高い Adam [29] を用いた．

4.3 実験結果

表 1 に示すように，どちらのコーパスを用いた場合でも，RNMF より高い強調性能となった．RNMF と比較し，WSJ0 を用いた場合は，SDR が平均で 1.26 dB 向上した．JNAS を用いた場合は，SDR が平均で 1.80 dB 向上した．また，JNAS を用いた VAE-NMF の評価はオープン・テストになっているが，クローズド・テストになっている WSJ0 を用いた場合と比較して，SDR は同程度以上となった．JNAS は日本語コーパスで，入力信号と言語が違うが，VAE-NMF は時間フレームごとに独立して事前分布を仮定するので，言語の違いは SDR に大きく寄与しなかったと考えられる．

図 3 に入力信号と強調音声の抜粋を示す．入力信号と比較すると，VAE-NMF の強調音は，より調波構造が鮮明になっている．また，4 kHz 以上の周波数帯域に現れている調波構造を持たない無声音も強調されている．一方で RNMF は，特に BUS 以外の雑音条件において，無声音が抑圧されている．無声音はスパース性より低ランク性が強いので，低ランク成分に分離されたためと考えられる．また，RNMF の出力スペクトログラムは全体にごま塩ノイズ状のミュージカルノイズが生じている．VAE-NMF は，クリーン音声から事前学習した音声事前分布を用いるので，低ランク性がある非調波成分も強調でき，音声らしくないミュージカルノイズが抑圧されていると考えられる．

VAE-NMF は，カフェテリア (CAF) と歩行者エリア (PED) の条件で性能が劣化している．CAF と PED の条件では，周囲の会話が雑音として混入していた．音声のスペクトログラムは，一般に低ランク性が低いので，背景雑音に含まれる音声も目的音声として推定されやすい．VAE-NMF は，音声成分を各フレームごとに独立して推定するため，目的音声が存在しない時間フレームでは，背景雑音に含まれる音声成分を目的音声成分として推定していると考えられる．

5. 考察と今後の課題

シミュレーション混合音を用いた評価実験によって，VAE-NMF の有効性を確認した．VAE-NMF は，クリーン音声から教師なし事前学習した音声事前分布を用いているので，スパース性の高い調波成分だけでなく低ランク性がある非調波成分も強調できた．また，雑音の事前分布に低ランク性を仮定した NMF を用いているため，雑音を事前学習せずに音声強調できた．VAE-NMF は，時間依存性の導入と多チャネルモデルへの拡張によって，さらなる性能向上が期待できる．

5.1 時間依存性の導入

本稿で述べた VAE による音声事前分布は，音声スペクトログラムの各時間フレームごとに独立に定義されている．音声には時間依存性があるので，これを導入することで，より自然な音声の推定が期待できる．特に前節で述べた，背景雑音に含まれる音声が強調される問題は，音声の時間依存性を事前分布に導入することで低減が期待できる．VAE を時系列モデルに拡張した再帰型 VAE [30] が提案されており，時間依存性の導入に有用である．

5.2 多チャネルモデルへの拡張

本研究では，人手で設計することが難しい音声信号の事前分布を VAE を用いて機械学習し，ベイズ推論の枠組みに組み込む方法を実現した．本稿で述べた VAE 音源事前分布は，単チャネル音声強調モデルだけでなく，多チャネル音源分離モデルの事前分布にも適用できる．VAE-NMF では，観測信号が音声信号と雑音信号の和であるという単純な混合モデルを仮定した．多チャネル音源モデルでは，音源信号の空間伝搬モデルを扱えるので，音源位置の空間的な違いを分離の指標に導入でき，性能向上が期待できる．NMF 音源モデルを導入した多チャネル音源分離のための階層ベイズモデル [31] が提案されており，本モデルに VAE 音源モデルが導入できる．

6. おわりに

本稿では，NMF による雑音モデルと VAE を用いた音声モデルを統合した音声強調法 (VAE-NMF) について述べた．VAE-NMF は，クリーン音声のデータセットを事前学習した VAE で音声の事前分布を構成するので，自然な音声の推定を実現できる．また，雑音には NMF モデルを仮定することで，環境に依存しやすい雑音を事前学習せずに推定・抑圧できる．実環境で収録された雑音信号と音声を混合したシミュレーション混合音の音声強調性能を評価し，その有効性を確認した．

今後は，より高精度な音声強調を実現するために，VAE 事前分布への時間依存性の導入と多チャネル音源分離モデルへの拡張を行う．また本稿では，音声強調の性能を SDR でのみ評価したので，音声認識率の評価や主観評価を行う．

謝辞 本研究は，科研費特別研究員奨励費 No. 15J08765，および ImPACT「タフ・ロボティクス・チャレンジ」の支援を受けた．

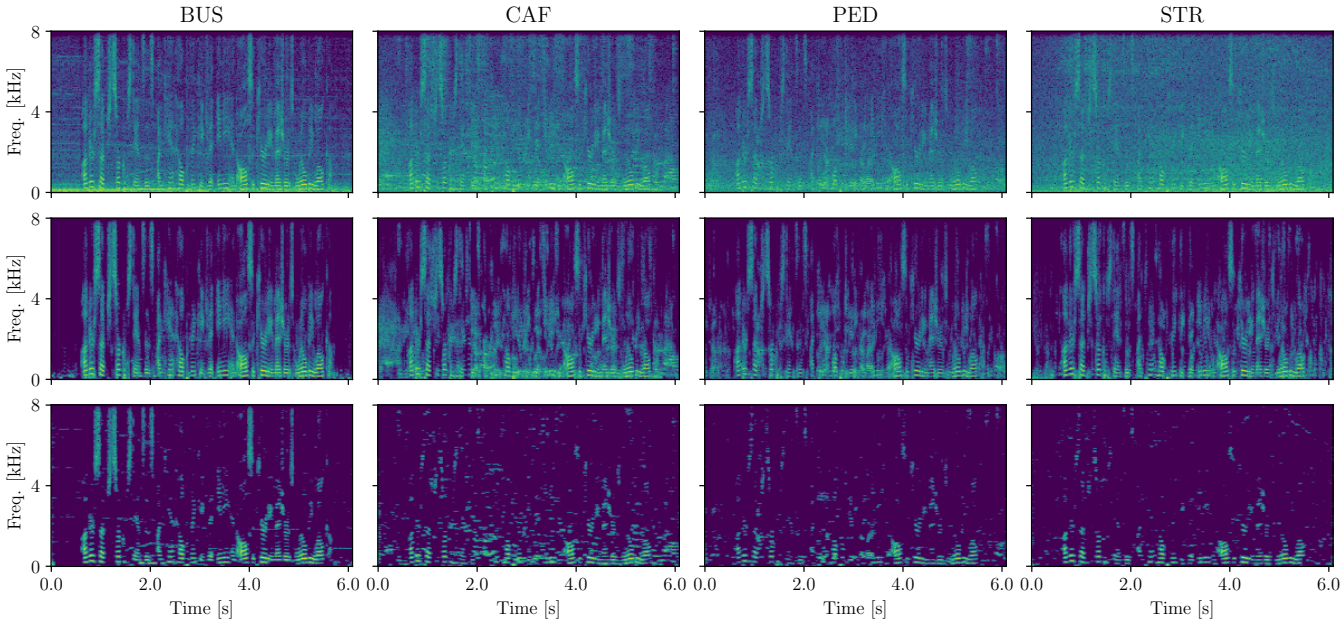


図 3: 音声強調結果の抜粋．上から順に，入力の混合音信号および VAE-NMF (WSJ0) の強調結果，RNMF の強調結果を示す．

文　　献

[1] J. Heymann et al. Neural network based spectral mask estimation for acoustic beamforming. In *IEEE ICASSP*, pages 196–200, 2016.

[2] X. Lu et al. Speech enhancement based on deep denoising autoencoder. In *Interspeech*, pages 436–440, 2013.

[3] Y. Ephraim et al. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE TASLP*, 32(6):1109–1121, 1984.

[4] N. Mohammadiha et al. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *IEEE TASLP*, 21(10):2140–2151, 2013.

[5] Y. Li et al. Speech enhancement based on robust NMF solved by alternating direction method of multipliers. In *IEEE MMSP*, pages 1–5, 2015.

[6] S. Araki et al. Spatial correlation model based observation vector clustering and MVDR beamforming for meeting recognition. In *IEEE ICASSP*, pages 385–389, 2016.

[7] N. Ono. Stable and fast update rules for independent vector analysis based on auxiliary function technique. In *IEEE WASPAA*, pages 189–192, 2011.

[8] Antoine Deleforge et al. Phase-optimized K-SVD for signal extraction from underdetermined multichannel sparse mixtures. In *IEEE ICASSP*, pages 355–359, 2015.

[9] P. C. Loizou. *Speech enhancement: theory and practice*. CRC press, 2013.

[10] C. Sun et al. Noise reduction based on robust principal component analysis. *JCIS*, 10(10):4403–4410, 2014.

[11] Z. Chen et al. Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition. In *IEEE WASPAA*, pages 1–4, 2013.

[12] M. D. Hoffman. Poisson-uniform nonnegative matrix factorization. In *IEEE ICASSP*, pages 5361–5364, 2012.

[13] B. Cauchi et al. Reduction of non-stationary noise for a robotic living assistant using sparse non-negative matrix factorization. In *SMIAE*, pages 28–33, 2012.

[14] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *CIN*, 2009(785152):1–17, 2009.

[15] C. Févotte et al. Nonlinear hyperspectral unmixing with robust nonnegative matrix factorization. *IEEE TSP*, 24(12):4810–4819, 2015.

[16] N. Dobigeon et al. Robust nonnegative matrix factorization for nonlinear unmixing of hyperspectral images. In *WHISPERS*, pages 1–4, 2013.

[17] M. Sun et al. Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback–Leibler divergence. *IEEE/ACM TASLP*, 23(7):1233–1242, 2015.

[18] D. P. Kingma et al. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013.

[19] C. Doersch. Tutorial on variational autoencoders. *arXiv:1606.05908*, 2016.

[20] C. M. Bishop. Pattern recognition. *Machine Learning*, 128, 2006.

[21] I. Goodfellow et al. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[22] A. Radford et al. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015.

[23] C. Hsu et al. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. *arXiv:1704.00849*, 2017.

[24] J. Barker et al. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In *IEEE ASRU*, pages 504–511, 2015.

[25] E. Vincent et al. Performance measurement in blind audio source separation. *IEEE TASLP*, 14(4):1462–1469, 2006.

[26] C. Raffel et al. mir eval: a transparent implementation of common MIR metrics. In *ISMIR*, pages 367–372, 2014.

[27] Y. Bando et al. Variational Bayesian multi-channel robust NMF for human-voice enhancement with a deformable and partially-occluded microphone array. In *EUSIPCO*, pages 1018–1022, 2016.

[28] K. Itou et al. The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus. In *ICSLP*, 1998.

[29] D. Kingma et al. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[30] O. Fabius et al. Variational recurrent auto-encoders. *arXiv:1412.6581*, 2014.

[31] K. Itakura et al. Bayesian multichannel nonnegative matrix factorization for audio source separation and localization. In *IEEE ICASSP*, pages 551–555, 2017.