解説：特集
ダイナミクスと機械学習の融合に挑む

深層生成モデルを用いた音声音響信号処理

亀 岡 弘 和*

*日本電信電話株式会社　神奈川県厚木市森の里若宮 3–1
*Nippon Telegraph and Telephone Corporation, 3–1 Morinosato Wakamiya, Atsugi, Kanagawa, Japan
*E-mail: hirokazu.kameoka.uh@hco.ntt.co.jp

キーワード：深層生成モデル (deep generative model)，音声音響信号処理 (speech and audio signal processing).
JL 0003/19/5803–0195

1. はじめに

音声情報処理分野における永年の課題とされてきた音声認識や音声合成の性能は近年の深層学習アプローチの進展により飛躍的な向上を遂げた．周知のとおり深層学習の基礎となるニューラルネットワーク (Neural Network; NN) は，解が唯一であるような問題，学習データが豊富な場合の教師あり学習タスクにおいては高い解決能力をもつことがすでに示されている．しかし，複数の解が存在しうる問題，不完全データによる学習タスクなどのように，解くための条件や手がかりが不足している問題に対し単純適用した場合の解決能力はいまだ限定的である．このような，いわゆる不良設定問題に対するアプローチとして，NN を巧みに用いて複雑な確率分布を表現することを可能にする深層生成モデルの枠組が近年注目されている．本稿では，まず深層生成モデルを概説した上で，音声音響分野における不良設定問題のいくつかの例を題材に最近の深層生成モデルの応用事例を紹介する．

2. 深層生成モデル

深層生成モデルとは，深層ニューラルネットワーク (Deep Neural Network; DNN) を用いて表現される生成モデルである．データの生成過程が不明であったり複雑であったりする場合など，データの生成分布を手動で設計するのが容易でない場合において深層生成モデルはその真価を発揮する．

ある生成モデルが対象データの「良い」モデルとなっているかどうかは，学習したモデルからランダム生成したサンプルが実際の対象データらしいものになるかどうか，という観点で判断することができる．たとえば 16 kHz で標本化した音声信号の場合，わずか 1 秒間であっても実に 16,000 次元ものデータになるが，画像や音声などの実世界データをモデル化する上での難しさは，高次元データの要素の同時分布をいかに表現するかという点にある．深層生成モデルはこれを解決する有効なアプローチとして近年注目を集めている．

深層生成モデルの代表的な例として，自己回帰生成ネットワーク (Autoregressive Generative Network; AGN) [1),2)]，変分自己符号化器 (Variational Autoencoder; VAE) [3),4)]，敵対的生成ネットワーク (Generative Adversarial Network; GAN) [5)] などが現在広く知られている．以下，それぞれの動機と原理を解説する．

2.1 自己回帰生成ネットワーク (AGN)

時系列信号 $\mathbf{x} = [x_1, \ldots, x_N]^\mathsf{T}$ の同時分布 $p(\mathbf{x})$ を記述するモデルとしては自己回帰 (Autoregressive; AR) モデルが有名である．音声音響分野では特に音声信号のモデルとして音声の符号化，合成，強調など，さまざまな場面で重要な役割を果たしてきたことで広く知られている．AR モデルは，対象の信号が AR 過程

$$x_n = \sum_{q=1}^{Q} a_q x_{n-q} + \epsilon_n \quad (n = 1, \ldots, N) \qquad (1)$$

$$\epsilon_n \sim \mathcal{N}(\epsilon_n | 0, \sigma^2) \qquad (2)$$

に従うと仮定することにより，(1) 式から導かれる $\mathbf{x}$ と $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\mathsf{T}$ の関係 $\mathbf{A}\mathbf{x} = \boldsymbol{\epsilon}$ と，(2) 式から導かれる $\boldsymbol{\epsilon}$ の確率分布に関する仮定に基づいて立てられる同時分布 $p(\mathbf{x})$ のモデルである．$\mathbf{x}_{<n} = [x_1, \ldots, x_{n-1}]^\mathsf{T}$ と置くと，これは，$p(\mathbf{x})$ を条件付分布 $p(x_n|\mathbf{x}_{<n})$ の積

$$p(\mathbf{x}) = \prod_n p(x_n|\mathbf{x}_{<n}) \qquad (3)$$

に分解した上で，各条件付分布 $p(x_n|\mathbf{x}_{<n})$ を

$$p(x_n|\mathbf{x}_{<n}) = \mathcal{N}\Big(x_n \Big| \sum_q a_q x_{n-q}, \sigma^2\Big) \qquad (4)$$

と置いたものに相当する．このようにして立てられる $p(\mathbf{x})$ は未知パラメータ $\mathbf{a} = [a_1, \ldots, a_Q]^\mathsf{T}$ の尤度関数となっており，所与の信号 $\mathbf{x}$ のもとで $p(\mathbf{x})$ が最大となるように $\mathbf{a}$ を推定することで $\mathbf{x}$ に最も良く当てはまる $p(\mathbf{x})$ を得ることができる．

AR モデルの巧みな点の 1 つは，高次元データの同時分布 $p(\mathbf{x})$ をトップダウンにモデル化するのではなく，信号の局所的な相関関係に着目してモデル化の対象を 1 次元の分布 $p(x_n|\mathbf{x}_{<n})$ とすることによりボトムアップに同時分布 $p(\mathbf{x})$ を構築している点にある．一方で AR モデルの限界は，信号の局所的な関係を表わす式が (1) 式のような線形クラスに限定されている点，条件付分布 $p(x_n|\mathbf{x}_{<n})$ がガウス分布などの特定の分布クラスに限定されている点，表現可能な信号が定常過程に限られている点にある．

実際の音声は非定常であるため，音声分析への適用の際は，音声信号を近似的に定常とみなせる短時間フレームに分割した上で各フレームで個別にAR分析が行われる．信号を合成する場合は，各フレームで分析したARパラメータを用いて $\boldsymbol{\epsilon}$ をパルス列などに置き換えた上で信号 $\mathbf{x}$ を生成する方式がとられる．

AGN[1]は，(4)式を，$\mathbf{x}_{<n}$ を入力として $p(x_n|\mathbf{x}_{<n})$ の分布パラメータまたは分布そのものを出力するNNに置き換えることにより，低次元の条件付分布のモデルをボトムアップに積み上げて高次元の同時分布を構築するARモデルのスタイルを踏襲しつつ，ARモデルにおける前述の限界を取り払うことに成功した生成モデルである．近年，音声音響分野の域を超えて広く知られることとなったWaveNet[2]はAGNの一種であり，きわめて高品質な音声信号を生成できることが示されている．WaveNetでは，振幅値 x_n を離散値に量子化し，$p(x_n|\mathbf{x}_{<n})$ を各値の離散分布とすることで任意の分布形を記述できるようにしている点，$p(x_n|\mathbf{x}_{<n})$ を出力するNNとして拡張型畳み込みネットワーク (Dilated Convolutional Network) を用いることにより長期依存関係を捉えられるようになっている点などが特徴である．ARモデルのパラメータ推定と同様，$p(\mathbf{x})$ が未知のネットワークパラメータ θ の尤度関数となるため，所与の学習データ $\mathcal{X}=\{\mathbf{x}_i\}_{1\le i\le l}$ のもとで平均対数尤度 $\mathcal{L}(\theta)=\frac{1}{l}\sum_i \log p(\mathbf{x}_i)$ が大きくなるようにパラメータ学習を行うことで $\mathcal{X}$ に良く当てはまる $p(\mathbf{x})$ を得ることができる．

なお，(3)式のとおり，$p(\mathbf{x})$ は $p(x_n|\mathbf{x}_{<n})$ の積の形になっており，$p(x_n|\mathbf{x}_{<n})$ は $\mathbf{x}_{<n}$ が決まらない限り決まらない．このため，学習した $p(\mathbf{x})$ を用いて信号 $\mathbf{x}$ を生成する際は，$p(x_n|\mathbf{x}_{<n})$ をもとに x_n を1サンプルずつ逐次的に生成する必要がある．原理上この処理は並列化することができないため，効率性の面では問題になる場合がある．

2.2 変分自己符号化器 (VAE)

たとえ高次元のデータであっても，要素間に何らかの相関構造や制約が存在するとき，データサンプルは低次元の潜在空間に圧縮することができる．このような低次元潜在空間と，元の高次元データを再構成するプロセスを見つけることができれば，これもまた高次元データをモデル化する1つの方式になりうる．Tippingら[6]は，主成分分析 (Principal Component Analysis; PCA) の生成モデル的解釈を示し，パラメータの最尤推定がPCAと等価な処理に帰着する生成モデルを示しているが，まさにこれは前述のモデル化の一例とみなすことができる．

$\mathbf{x}\in\mathbb{R}^N$ を平均が $\mathbf{0}$ のデータとし，低次元の潜在変数 $\mathbf{z}\in\mathbb{R}^M (N>M)$ のアフィン変換により得られたもの

$$\mathbf{x}=\mathbf{W}\mathbf{z}+\boldsymbol{\epsilon} \tag{5}$$

と仮定し，$\boldsymbol{\epsilon}$ を分散が σ^2 の正規白色雑音

$$\boldsymbol{\epsilon}\sim\mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0},\sigma^2\mathbf{I}) \tag{6}$$

と仮定すると，$p(\mathbf{x}|\mathbf{z})$ は

$$p(\mathbf{x}|\mathbf{z})=\mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z},\sigma^2\mathbf{I}) \tag{7}$$

となる．さらに，$\mathbf{z}$ の事前分布 $p(\mathbf{z})$ を標準ガウス分布

$$p(\mathbf{z})=\mathcal{N}(\mathbf{z}|\mathbf{0},\mathbf{I}) \tag{8}$$

とすると，$p(\mathbf{x})=\int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z}$ は

$$p(\mathbf{x})=\mathcal{N}(\mathbf{x}|\mathbf{0},\mathbf{W}\mathbf{W}^{\mathsf{T}}+\sigma^2\mathbf{I}) \tag{9}$$

となる．$\mathbf{W}$ は低次元の $\mathbf{z}$ から高次元の $\mathbf{x}$ を再構成するための変換行列であり，(9)式は未知パラメータ $\mathbf{W}$, σ^2 の尤度関数となっている．よって，所与のデータ $\mathcal{X}=\{\mathbf{x}_i\}_i$ のもとで平均対数尤度 $\mathcal{L}(\mathbf{W},\sigma^2)=\frac{1}{l}\sum_i \log p(\mathbf{x}_i)$ が最大となるように $\mathbf{W}$, σ^2 を推定することで全データ $\mathcal{X}$ に最も良く当てはまる $p(\mathbf{x})$ を得ることができる．また，$p(\mathbf{z}|\mathbf{x})$ はベイズの定理より

$$\begin{aligned} p(\mathbf{z}|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})} \\ &= \mathcal{N}(\mathbf{z}|\mathbf{M}^{-1}\mathbf{W}^{\mathsf{T}}\mathbf{x},\sigma^2\mathbf{M}^{-1}) \end{aligned} \tag{10}$$

と求まる．ただし，$\mathbf{M}=\mathbf{W}^{\mathsf{T}}\mathbf{W}+\sigma^2\mathbf{I}$ である．以上の $p(\mathbf{z}|\mathbf{x})$ と $p(\mathbf{x}|\mathbf{z})$ は，データ $\mathbf{x}$ を低次元ベクトル $\mathbf{z}$ に圧縮する符号化器，$\mathbf{z}$ から元データ $\mathbf{x}$ を再構成する復号化器とぞれぞれみなせる．

(7), (10)式から明らかなようにPCA型の生成モデルでは，潜在変数 $\mathbf{z}$ とデータ $\mathbf{x}$ の関係は線形式で表わされる．しかし空間上に平面的ではなく曲面状に分布するデータに対しては，$\mathbf{z}$ と $\mathbf{x}$ の関係をより柔軟に記述できるモデルのほうが望ましい．VAE[3),4)]は，上述の生成モデルにおいて $\mathbf{z}$ と $\mathbf{x}$ の関係をNNによる非線形関数に置き換えたものとみなせる．たとえば(7)式の平均と対角分散行列を $\mathbf{W}\mathbf{z}$, $\sigma^2\mathbf{I}$ と置く代わりに

$$p(\mathbf{x}|\mathbf{z})=\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\phi(\mathbf{z}),\mathrm{diag}(\mathbf{v}_\phi(\mathbf{z}))) \tag{11}$$

のように，$\mathbf{z}$ を入力とするパラメータ ϕ のNNの出力 $\boldsymbol{\mu}_\phi(\mathbf{z})$, $\mathrm{diag}(\mathbf{v}_\phi(\mathbf{z}))$ で表わすことを考える．ただし，$\mathrm{diag}(\mathbf{y})$ はベクトル $\mathbf{y}$ の要素を対角成分にもつ対角行列を表わすものとする．さて，ここで先述のPCA型の生成モデルと大きく事情が異なるのは，$p(\mathbf{x})$ を得るための周辺化も，ベイズの定理による $p(\mathbf{z}|\mathbf{x})$ の計算も解析的に行うことが困難である点である．そこで，$p(\mathbf{z}|\mathbf{x})$ を近似することを目的とする補助分布 $q(\mathbf{z}|\mathbf{x})$ を新たに導入し，$q(\mathbf{z}|\mathbf{x})$ と $p(\mathbf{z}|\mathbf{x})=p(\mathbf{x}|\mathbf{z})p(\mathbf{z})/p(\mathbf{x})$ ができるだけ近くなるように $q(\mathbf{z}|\mathbf{x})$ と $p(\mathbf{x}|\mathbf{z})$ を求める学習問題を考える．ここでたとえば，$q(\mathbf{z}|\mathbf{x})$ の分布の平均と対角分散行列を，(11)式と同様

$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\theta(\mathbf{x}), \mathrm{diag}(\mathbf{v}_\theta(\mathbf{x}))) \tag{12}$$

のように $\mathbf{x}$ を入力とするパラメータ θ の NN の出力 $\boldsymbol{\mu}_\theta(\mathbf{x})$, $\mathrm{diag}(\mathbf{v}_\theta(\mathbf{x}))$ で表わし，$q(\mathbf{z}|\mathbf{x})$ と $p(\mathbf{z}|\mathbf{x})$ のカルバック・ライブラ (KL) ダイバージェンス

$$\begin{aligned}\mathrm{KL}[q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}|\mathbf{x})] &= \log p(\mathbf{x}) \\ &\quad -\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] + \mathrm{KL}[q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})]\end{aligned} \tag{13}$$

が小さくなるように未知パラメータ ϕ, θ を求めることができれば，互いに矛盾のない符号化器と復号化器を得ることができる．(13) 式の第二項は符号化器と復号化器による入力 $\mathbf{x}$ の再構成誤差を表わす指標，第三項は符号化器の出力 $\mathbf{z}$ の分布と標準ガウス分布 $p(\mathbf{z})$ との乖離度を表わす指標となっている．したがって，(13) 式を小さくすることは，潜在変数 $\mathbf{z}$ の各要素ができるだけ無相関になるような自己符号化器を得ることを意味する．ところで，$\mathrm{KL}[q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}|\mathbf{x})] \geq 0$ であるので，(13) 式より VAE の学習規準 $\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - \mathrm{KL}[q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})]$ は，対数尤度 $\log p(\mathbf{x})$ の下界となっている．

ここで 1 点注意が必要なのは，(13) 式の第二項の $\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})$ に関する期待値計算である．$\log p(\mathbf{x}|\mathbf{z})$ は，$\boldsymbol{\mu}_\phi(\mathbf{z})$ および $\mathbf{v}_\phi(\mathbf{z})$ の具体形によって決まる $\mathbf{z}$ の非線形関数となっており，一般にその期待値を解析的に得るのは難しい．したがって，$\mathbf{z}$ のサンプリングによるモンテカルロ近似を用いて計算する方法が考えられるが，その場合，パラメータ ϕ がサンプリング元の分布の中に含まれるため，誤差逆伝播法において ϕ の勾配を評価することができなくなる．しかしここで，正規乱数 $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \mathbf{I})$ を用いて表わされる $\mathbf{z} = \boldsymbol{\mu}_\phi(\mathbf{x}) + \mathbf{v}_\phi(\mathbf{x}) \odot \boldsymbol{\epsilon}$ が $\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})$ の等価表現となっていることを利用すると，実は前述の期待値計算を $\boldsymbol{\epsilon}$ のサンプリングによるモンテカルロ近似で代替することができるのである．ただし，$\odot$ はベクトルの要素積を表わす．これにより，パラメータ ϕ を $\log p(\mathbf{x}|\mathbf{z})$ の中に移し替えることができ，ϕ の勾配を評価することが可能になる．この技法を変数変換トリックという．

以上では PCA の発展版という視点からの説明だったため，符号化器も復号化器もガウス分布で与えられる場合を想定したが，復号化器 $p(\mathbf{x}|\mathbf{z})$ に関してはガウス分布以外の分布を仮定してもさしつかえない．ただし，符号化器に関しては変数変換トリックを適用可能な分布とする必要がある (注1)．なお，Maddison ら[7] は離散分布の符号化器を扱えるようにするための変数変換トリックの方法を提案している．

VAE では，以上のように符号化器と復号化器にガウス分布などの特定クラスの分布を仮定するが，VAE の生成モデルとしての性質は仮定した分布に強く依存する．たとえばガウス分布を仮定した場合の VAE では，復号化器による生成データは実際のデータよりも平滑化されたものになる傾向がある．

(注1) 変数変換トリックさえ適用可能であれば，(13) 式第三項の KL ダイバージェンス項が解析的に記述できなくとも第二項と同様に第三項に関してもモンテカルロ近似を用いて ϕ の勾配を評価することが可能である．

2.3 敵対的生成ネットワーク (GAN)

AGN と VAE はともに NN により確率分布を表現した生成モデルであり，尤度関数またはその下界を学習規準とする点で共通しているのに対し，GAN[5] はこれらとは一線を画すタイプの生成モデルである．GAN の目的は，学習データ $\mathcal{X} = \{\mathbf{x}_i\}_{1\leq i\leq I} \in \mathbb{R}^N$ が従う未知の分布 $p_{\mathrm{data}}(\mathbf{x})$ と同一の分布からのサンプリングを可能にする生成器 G を得ることである．この目的を実現するため，識別器 D を用いて生成器 G を学習するアプローチをとる点が GAN の特徴である．識別器 D は，入力データが生成器 G が生成したサンプルなのか実データサンプルなのかを正しく識別するよう学習される一方で，生成器 G は，自身が生成したサンプルを識別器 D にできるだけ実サンプルと誤認識されるように学習される．このように D と G を「敵対的」に学習することで G を実データ分布 $p_{\mathrm{data}}(\mathbf{x})$ に従う乱数生成器とすることができる．これはつぎのように示される．

まず，生成器 G として，適当な確率分布（一様分布など）$p(\mathbf{z})$ に従う乱数 $\mathbf{z} \sim p(\mathbf{z})$ を入力とするパラメータ ϕ の NN $\tilde{\mathbf{x}} = G(\mathbf{z})$ を考え，識別器として，入力 $\mathbf{x}$ が実データサンプルかどうかを表わす確率を出力するパラメータ θ の NN $p = D(\mathbf{x}) \in [0, 1]$ を考える．識別器 D の目標は，実データサンプル $\mathbf{x} \sim p_{\mathrm{data}}(\mathbf{x})$ に対し高い確率，生成器 G から生成されたいわば「偽」のデータサンプル $G(\mathbf{z})$ に対して低い確率を返すことであるので，損失関数としてたとえばクロスエントロピー規準を用いた場合，

$$\begin{aligned}\mathcal{V}(G, D) = &-\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}(\mathbf{x})}\big[\log D(\mathbf{x})\big] \\ &-\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\big[\log(1 - D(G(\mathbf{z})))\big]\end{aligned} \tag{14}$$

を小さくすることが学習目標となる．一方で G の目標は D に誤認識させることであるので，$\mathcal{V}(G, D)$ を大きくすることが目標となる．仮に D に何の制約もなく無限の表現能力が備わっているとすると，$\mathcal{V}$ を最小にする $D(\mathbf{x})$ は，$\mathcal{V}$ の $D(\mathbf{x})$ に関する変分を 0 と置くことにより

$$\hat{D}(\mathbf{x}) = \frac{p_{\mathrm{data}}(\mathbf{x})}{p_{\mathrm{data}}(\mathbf{x}) + p_G(\mathbf{x})} \tag{15}$$

となる．ただし，$p_G(\tilde{\mathbf{x}})$ は $\mathbf{z} \sim p(\mathbf{z})$ のとき $\tilde{\mathbf{x}} = G(\mathbf{z})$ が従う分布である．この $\hat{D}(\mathbf{x})$ のもとでの $\mathcal{V}(G, D)$ は，

$$\begin{aligned}&\mathcal{V}(G, \hat{D}) \\ &= -\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}(\mathbf{x})}\left[\log \frac{p_{\mathrm{data}}(\mathbf{x})}{p_{\mathrm{data}}(\mathbf{x}) + p_G(\mathbf{x})}\right] \\ &\quad -\mathbb{E}_{\mathbf{x}\sim p_G(\mathbf{x})}\left[\log \frac{p_G(\mathbf{x})}{p_{\mathrm{data}}(\mathbf{x}) + p_G(\mathbf{x})}\right]\end{aligned}$$

$$= -\mathrm{KL}\left(p_{\mathrm{data}}(\mathbf{x}) \middle\| \frac{p_{\mathrm{data}}(\mathbf{x}) + p_G(\mathbf{x})}{2}\right) - \mathrm{KL}\left(p_G(\mathbf{x}) \middle\| \frac{p_{\mathrm{data}}(\mathbf{x}) + p_G(\mathbf{x})}{2}\right) + \log(4)$$
$$= -\mathrm{JS}\big(p_{\mathrm{data}}(\mathbf{x}) \| p_G(\mathbf{x})\big) + \log(4) \qquad (16)$$

となり，$\mathcal{V}(G, \hat{D})$ を G に関して大きくすることは，$p_{\mathrm{data}}(\mathbf{x})$ と $p_G(\mathbf{x})$ との間のイェンセン・シャノン (JS) ダイバージェンス $\mathrm{JS}\big(p_{\mathrm{data}}(\mathbf{x}) \| p_G(\mathbf{x})\big)$ を小さくすることを意味することがわかる．このことから，$\mathcal{V}(G, D)$ を D に関して小さくし，G に関して大きくする GAN のミニマックス学習法は，実データの分布 $p_{\mathrm{data}}(\mathbf{x})$ に従う乱数生成器 G を獲得する方法になっていることが示される．

GAN は，上記のようなミニマックス最適化により，実データ分布 $p_{\mathrm{data}}(\mathbf{x})$ を明示的にモデル化することなく $p_{\mathrm{data}}(\mathbf{x})$ を見つけ出すことができる利点がある反面，学習が不安定になることで知られる．しばしばその原因として指摘される問題の 1 つに，同じようなデータサンプルを生成するよう G が学習されるモード崩壊 (mode collapse) と呼ぶ現象がある．この問題を解決する目的で，学習規準，正則化法，ネットワーク構造などを工夫する試みが盛んに行われている[8)–12)]．

3. 音声音響信号処理問題への応用

本章では各種音声音響信号処理問題に対する最近の深層生成モデルの応用事例を紹介する．

3.1 音声分析合成

音声分析合成（ボコーダ）技術は，音声生成過程を模擬したモデルに基づき，音声信号の各短時間フレームにおける声帯の音源特性（基本周波数や非周期性指標など）と声道の共振特性などの情報を表現した音響特徴量を抽出する音声分析部と，抽出した音響特徴量から音声信号を再構成する音声合成部からなる．ボコーダ技術は，音声符号化，音声合成，音声変換などにおいてきわめて重要な役割を果たしてきた．たとえばテキスト音声合成 (Text-to-Speech; TTS) や声質変換 (Voice Conversion; VC) では，テキストや変換元音声から目標音声を予測する回帰問題となるが，音響特徴量は音声信号をコンパクトに圧縮した表現であるため，回帰問題の解きやすさの面では予測対象を直接音声信号波形とするよりも音響特徴量とするほうが有利となる．このため，限られた学習データのもとでの TTS や VC では，音響特徴量予測と音声合成のパイプライン方式がとられることが多い．しかし，従来のボコーダによる合成音は品質に限界があり，合成音とただちに知覚できるものとなっている．一般に音声信号から音響特徴量への変換は不可逆であり，従来のボコーダでは音声信号の再構成は手動設計されたモデルとヒューリスティックな仮定に基づいて行われるが，これが合成音声と実際の音声との間にギャップを生む要因となっていた．

音声信号と音響特徴量のペアデータは，ありとあらゆる音声データに対して音声分析を行うことで容易かつ無数に用意することが可能なので，ボコーダにおける音声合成部を機械学習方式に置き換えることに考えが至るのは自然なことである．しかし，通常，音響特徴量と音声信号はいわゆる一対多の関係にあるため，音響特徴量から音声信号に変換する問題を一対一対応を想定した単純な回帰問題として扱うことは難しい．そこで登場したのが AGN に基づく前述の WaveNet[2)] である．音声信号を $\mathbf{x} = [x_1, \ldots, x_N]^{\mathsf{T}}$，対応する音響特徴量系列を $\mathbf{h} = [\mathbf{h}_1; \ldots; \mathbf{h}_M]$（ただし，; は行列中の改行を表わす.）とすると，この方式では $\mathbf{x}$ と $\mathbf{h}$ のペアデータ $\{\mathbf{x}_i, \mathbf{h}_i\}_{1 \le i \le I}$ を用い，$\mathbf{h}$ で条件付けられた $\mathbf{x}$ の分布

$$p(\mathbf{x}|\mathbf{h}) = \prod_n p(x_n|\mathbf{x}_{<n}, \mathbf{h}) \qquad (17)$$

が大きくなるように $p(x_n|\mathbf{x}_{<n}, \mathbf{h})$ を出力する NN を学習することで，時間方向の依存関係を考慮しつつ所与の音響特徴量系列 $\tilde{\mathbf{h}}$ に適合した音声信号 $\tilde{\mathbf{x}}$ を生成する音声合成器を得ることができる．この方式におけるブレイクスルーのポイントは，信号サンプルの長期依存関係を考慮することにより従来のフレーム処理が不要になったこと，特定の分布形を仮定しないことにより音声波形のランダム性をきわめて良く再現できるようになったこと，などが挙げられる．近年，WaveNet の成功を皮切りに，深層生成モデルに基づくボコーダ方式がつぎつぎと提案されており[13)–16)]，今後も動向が注目される．

3.2 ブラインド音源分離 (BSS)

音声分析，音源定位，音源分離などのように音声音響信号処理問題の多くは，観測信号からその原因を推定する逆問題とみなせる．逆問題を解決する有効なアプローチの 1 つは，観測信号の原因となる情報をパラメータとした生成モデルを立て，逆問題をその生成モデルのパラメータ推論（機械学習）の問題に落とし込むアプローチである．すなわち，原因 θ によって観測信号 x が生成される過程を条件付分布 $p(x|\theta)$ の形で記述し，θ に関する先験的知識 $p(\theta)$ を手がかりに θ の事後分布 $p(\theta|x) \propto p(x|\theta)p(\theta)$ から θ を推論するアプローチである．特に，同じ観測信号 x を与える原因 θ が複数個存在しうる問題では，$p(x|\theta)$ を規準にするだけでは θ の解を一意に決めることはできず，適切な $p(\theta)$ をいかにして立てるかが解決の鍵となる．

一例として，ここではブラインド音源分離 (Blind Source Separation; BSS) の問題を考える．BSS は，音源信号と音源からマイクロホンまでの伝達特性がともに未知のもとで，各マイクロホンの観測信号から音源信号を復元する問題である．この場合，音源信号が原因 θ に相当し，観測信号 x から音源信号 θ を復元することが目的となるが，観測信号の分解の仕方は一般に無数に存

在する．その無数の候補の中から実際の音源信号に対応する解を見つけ出すには，音源信号に関する先験的知識 $p(\theta)$ が手がかりとなる．この際，手がかりの 1 つとして音源信号間の統計的独立性が挙げられるが，これだけでは十分ではない場合がある．

J 個の音源信号の混合信号を I 個のマイクロホンで観測するとき，時不変な混合過程を仮定できる場合，各マイクロホンで観測される信号は時間周波数領域で

$$x_i(f,n) = \sum_j a_{i,j}(f) s_j(f,n) + u_i(f,n) \tag{18}$$

と表わされる．ただし，$a_{i,j}(f) \in \mathbb{C}$（$f$ は周波数インデックス）は音源 j からマイクロホン i までの音響経路の周波数応答，$x_i(f,n)$，$s_j(f,n)$, $u_i(f,n) \in \mathbb{C}$ はそれぞれマイクロホン i の観測信号，音源 j の音源信号，マイクロホン i に混入する雑音信号の短時間フレーム n における短時間フーリエ変換 (Short-Time Fourier Transform; STFT) を表わす．$\mathbf{a}_j(f) = [a_{1,j}(f), \ldots, a_{I,j}(f)]^{\mathsf{T}}$，$\mathbf{A}(f) = [\mathbf{a}_1(f), \ldots, \mathbf{a}_J(f)]$, $\mathbf{s}(f,n) = [s_1(f,n), \ldots, s_J(f,n)]^{\mathsf{T}}$, $\mathbf{u}(f,n) = [u_1(f,n), \ldots, u_I(f,n)]^{\mathsf{T}}$ とすると (18) 式は

$$\mathbf{x}(f,n) = \sum_i \mathbf{a}_i(f) s_i(f,n) + \mathbf{u}(f,n) \tag{19}$$

$$= \mathbf{A}(f)\mathbf{s}(f,n) + \mathbf{u}(f,n) \tag{20}$$

と書ける．また，$\mathbf{u}(f,n) = \mathbf{0}$ で，音源数 J とマイク数 I が等しく $\mathbf{A}(f)$ が逆行列 $\mathbf{A}(f)^{-1} = \mathbf{W}^{\mathsf{H}}(f)$ をもつ場合，(20) 式は以下のように書ける．

$$\mathbf{W}^{\mathsf{H}}(f)\mathbf{x}(f,n) = \mathbf{s}(f,n) \tag{21}$$

(20) 式や (21) 式の混合過程を基にした BSS の枠組を周波数領域 BSS と呼ぶ．今，$s_j(f,n)$ および $u_i(f,n)$ が独立に平均が 0 の複素ガウス分布に従う場合を考える．

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n)) \tag{22}$$

$$\mathbf{u}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{u}(f,n)|\mathbf{0}, \mathbf{\Sigma}_{\mathbf{u}}(f,n)) \tag{23}$$

ただし，$v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$ は音源 j のパワー，$\mathbf{\Sigma}_{\mathbf{u}}(f,n) = \mathbb{E}[\mathbf{u}(f,n)\mathbf{u}(f,n)^{\mathsf{H}}]$ は雑音の分散共分散行列を表わす．(22) 式のように周波数 f と時刻 n に依存する分散 $v_j(f,n)$ をもつ $s_j(f,n)$ を局所ガウス音源モデル (Local Gaussian Model; LGM) と呼ぶ．このとき，(19) 式と (21) 式 のいずれの混合過程を仮定した場合も観測信号 $\mathbf{x}(f,n)$ は下記に従う．

$$\mathbf{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}(f,n)\,|\mathbf{0}, \mathbf{\Sigma}_{\mathbf{x}}(f,n)\right) \tag{24}$$

ただし，$\mathbf{\Sigma}_{\mathbf{x}}(f,n)$ は，(19) 式, (21) 式の場合，それぞれ

$$\mathbf{\Sigma}_{\mathbf{x}}(f,n) = \begin{cases} \sum_j v_j(f,n)\mathbf{R}_j(f) + \mathbf{\Sigma}_{\mathbf{u}}(f,n) \\ \mathbf{W}^{\mathsf{H}}(f)^{-1}\mathbf{\Sigma}_{\mathbf{s}}(f,n)\mathbf{W}(f)^{-1} \end{cases} \tag{25}$$

で与えられ，$\mathbf{R}_j(f) = \mathbf{a}_j(f)\mathbf{a}_j^{\mathsf{H}}(f)$，$\mathbf{\Sigma}_{\mathbf{s}}(f,n) = \mathrm{diag}([v_1(f,n), \ldots, v_J(f,n)]^{\mathsf{T}})$ である．(24) 式より観測信号 $\mathcal{X} = [x_i(f,n)]_{i,f,n}$ の確率分布は

$$p(\mathcal{X}|\theta) = \prod_{f,n} \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}(f,n)\,|\mathbf{0}, \mathbf{\Sigma}_{\mathbf{x}}(f,n)\right) \tag{26}$$

となり，周波数領域 BSS はこれを未知パラメータ $\mathcal{V} = [v_j(f,n)]_{j,f,n}$, $\mathcal{A} = [\mathbf{A}(f)]_f$ の尤度関数とした最尤推定問題として定式化される．ただし，(26) 式の対数は周波数 f ごとの項に分解されるため，以上の定式化では，各 f において分離信号のインデックス j にはパーミュテーションの任意性が残る．また，音源数がマイクロホン数より多い劣決定条件においては，$\mathbf{A}(f)$ が決まったとしても $\mathbf{s}(f,n)$ を一意に決めることができない．よって，解を絞り込むためのさらなる手がかりが必要となる．

たとえば音声は有声音の場合倍音構造やフォルマント構造を有するように，自然界の音源はパワースペクトログラム（いわゆる声紋）に特徴的な構造をもつものが多い．そこで，上述の問題の手がかりとして音源パワースペクトログラム $v_j(f,n)$ にモデルを導入する方法が考えられる．たとえば，BSS の代表的アプローチの 1 つとして知られる多チャンネル非負値行列分解 (Multichannel Non-negative Matrix Factorization; MNMF) 法[17)~20)]では，$\mathsf{V}_j = [v_j(f,n)]_{f,n}$ が 2 つの非負値行列 B_j, H_j の積

$$\mathsf{V}_j = \mathsf{B}_j\mathsf{H}_j \tag{27}$$

でモデル化される．これは，各時間フレームの音源パワースペクトルを，スペクトルテンプレートの重みつき和で表わすことに相当する．このモデルのもとで (26) 式を最大にするパラメータを推定することにより，音源のスペクトル構造を手がかりにしながら周波数ごとの音源分離とパーミュテーション整合を同時解決することが可能となる．

しかし，(27) 式では実際に表現可能なスペクトログラムの範囲は限られるため，音源信号のモデル化には改良の余地が大きく残されている．そこで考えられるのが，深層生成モデルの柔軟な表現能力を活かした $\mathsf{S}_j = [s_j(f,n)]_{f,n}$ の確率分布のモデル化である．多チャンネル VAE (MVAE) 法[21)~23)]は，VAE の復号化器を (22) 式の LGM と同形となるように S_j のモデルとして導入した BSS の枠組である．MVAE 法では，S_j の生成モデルを，

$$p(\mathsf{S}_j|\mathbf{z}_j, c_j) = \prod_{f,n} \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n)) \tag{28}$$

$$v_j(f,n) = g_j \cdot \sigma_\theta^2(f,n;\mathbf{z}_j,c_j) \tag{29}$$

のように音源クラスインデックス c_j[(注2)] を補助入力とし，分散 $\boldsymbol{\sigma}_\theta^2(\mathbf{z}_j,c_j) = [\sigma_\theta^2(f,n;\mathbf{z}_j,c_j)]_{f,n}$ を出力する NN を復号化器とした VAE によりモデル化する．ただ

(注2) ここで音源クラスインデックスは，たとえば音声の場合は話者 ID，楽音の場合は楽器の種類などを表わすインデックスを表わす．

し，g_j はパワースペクトログラムのスケールを表わす変数，$\sigma_\theta^2(f,n;\mathbf{z},c)$ は復号化器の出力 $\boldsymbol{\sigma}_\theta^2(\mathbf{z},c)$ の第 (f,n) 要素を表わす．クラスラベルつきの学習サンプルを用いて事前学習された上記音源モデルは，さまざまなクラスの音源のスペクトログラムを表現可能な生成モデルとなっており，c_j が音源クラスを表わすパラメータ，$\mathbf{z}_j$ がクラス内変動を表わすパラメータとみなせる．(28) 式は LGM となっているため，未知パラメータ $\mathcal{A}$, $\mathcal{Z}=\{\mathbf{z}_j\}_j$, $\mathcal{C}=\{c_j\}_j$, $\mathcal{G}=\{g_j\}_j$ の尤度関数は (24) 式と同形の尤度関数に帰着する．これにより MVAE 法の最適化アルゴリズムを導くことができる．MVAE 法は優決定条件と劣決定条件のいずれの場合も MNMF 法を上回る分離性能を達成しており，深層生成モデルによる音源モデル化の効果が示されている[21)~23)]．なお，モノラル音源分離タスクにおいても VAE や GAN を用いた手法が提案されている[24),25)]．

3.3 声質変換 (VC)

入力音声の言語情報（発話文）を保持したまま非言語・パラ言語情報（話者性や発話様式など）のみを変換する技術を VC といい，話者性変換，発声障がい者支援，非母語音声の発音自動修正などへの応用が拓かれる技術である．VC の問題は，変換元の音声の音響特徴量から変換目標の音声の音響特徴量への写像関数を推定する回帰分析の問題として定式化することができる．VC の従来法の中でも混合ガウス分布モデル (Gaussian Mixture Model; GMM) を用いた手法[26),27)] はその有効性と汎用性から広く用いられてきた．また，近年では，DNN を用いた手法や NMF などを用いた事例（Exemplar）ベースの手法の検討もなされている．これらの手法の多くは，時間整合された同一発話内容の音声ペアで構成されるパラレルデータを用いて変換音声の音響特徴量が目標音声の音響特徴量にできるだけ近くなるように変換関数の学習が行われる．しかし，用途によっては同一発話内容の変換元音声と目標音声のペアデータを用意することが難しい場面は多くある．また，仮にそのようなペアデータが用意できる場合でも，高い精度の時間整合が必要となり，これを自動処理で行う際は整合ミスを修正する人手の作業が必要となる．これを問題意識とし，パラレルデータを必要としない非パラレル VC の研究も取り組まれている．非パラレル VC は不完全データ問題の一例であり，これを解決する目的で深層生成モデルに着目した手法が近年提案されている．

1 つは，VAE に基づくアプローチ[28)~30)] である．この方式ではまず，音響特徴量 $\mathbf{x}$ とその属性のクラスインデックス（話者性変換タスクの場合は話者 ID）c を補助入力として潜在変数 $\mathbf{z}$ の条件付分布 $q_\phi(\mathbf{z}|\mathbf{x},c)$ を出力する符号化器と，潜在変数 $\mathbf{z}$ と属性クラス c を入力として音響特徴量 $\mathbf{x}$ の条件付分布 $p_\theta(\mathbf{x}|\mathbf{z},c)$ を出力する復号化器を学習サンプル $\{\mathbf{x}_m,c_m\}_{m=1}^M$ を用いて学習する．直感的には，符号化器は入力音声の言語情報を抽出するいわば音声認識器としての役割を担い，復号化器は符号化器により抽出された言語情報と補助入力情報（目標属性クラス）をもとに目標音声を合成する音声合成器としての役割を担っていると解釈できる．これにより，学習後，符号化器と復号化器に対し変換元音声の音響特徴量 $\mathbf{x}$ と目標属性クラス c' を入力することで，変換元音声の発話内容を保持したまま属性 c' をもった音響特徴量 $\hat{\mathbf{x}}$ を生成することが可能となる．この方式はパラレルデータを必要としない利点がある一方で，生成される音声の特徴量が過剰平滑化される傾向にある．これは復号化器の条件付分布にガウス分布などのパラメトリックな確率分布形を仮定することに起因し，仮定した分布形が実際の音響特徴量が従う真の確率分布と一致しないことによる．

2 つ目は，CycleGAN-VC と呼ぶ GAN に基づくアプローチ[31)] である．CycleGAN[32)~34)] は元々画像のスタイル変換を目的として考案された手法であり，CycleGAN-VC はこれを VC に応用したものである．この方法では，異なる属性の音声間の変換関数 G,F と，入力が実音声の特徴量か合成音声の特徴量かを識別する識別器 D をそれぞれ NN によりモデル化し，これらを敵対的学習規準，循環無矛盾性規準，恒等変換規準と呼ぶ三種の規準を用いて学習する．敵対的学習規準は識別器 D の損失関数をさし，D はこれが小さくなるように学習されるのに対し，G と F はこれが大きくなるように学習される．これはすなわち，G と F により変換された音響特徴量が D に実音声と誤識別されるように促すことを意味する．循環無矛盾性規準は，G または F により変換された音響特徴量をもう一方の変換関数により逆変換した際の復元誤差をさし，これを小さくするように G と F を学習することで，G および F が一対一変換となるよう学習を促進させることができる．また，恒等変換規準は，変換先の属性の音響特徴量を変換関数の入力とした際の変化の大きさを表わす変換誤差をさし，G と F はこれができるだけ小さくなるように学習される．これらの規準が，パラレルデータを用いずとも発話内容を保持したまま属性のみを変換する関数を得ることを可能にしている．この方法は，敵対的学習規準の導入により，VAE 方式のように音響特徴量の確率分布形を陽に仮定することなく実音声の音響特徴量が従う真の確率分布に近い音響特徴量を生成することができるという特長をもつ．一方で，この方法は二種の属性間の相互変換に限ったものであるため，多種の属性への変換を行うには，属性のペアの全組み合わせ分の変換関数を学習する必要があり，学習すべきパラメータの数が属性クラス数の二乗のオーダで増大するという問題がある．

3 つ目は，StarGAN-VC と呼ぶ GAN に基づくアプローチ[35)] で，CycleGAN-VC の利点を継承しつつ，単一の変換器で多種の属性への変換を可能にする方式となっている．CycleGAN と同様 StarGAN[36)] も元々は画像の

スタイル変換の手法であり，StarGAN-VC はこれを VC に応用したものである．この方法では，音響特徴量 $\mathbf{x}$ と目標属性クラスのインデックス c を入力として属性クラス c の音響特徴量 $\mathbf{y}$ を出力する変換器 G を NN によりモデル化し，変換器の出力 $\hat{\mathbf{y}} = G(\mathbf{x}, c)$ が実音声らしく，かつ属性 c らしくなるように G を学習することが目標となる．そこで，入力音響特徴量が実音声か合成音声かを識別する識別器 D と，どの属性クラスに属しているらしいかを識別する識別器 C を導入し，各属性の学習データ $\{\mathbf{x}_m, c_m\}_{m=1}^{M}$ を用いて敵対的学習規準，循環無矛盾性規準，復元誤差規準，属性識別規準からなる学習規準により D, C, G を学習する方法となっている．CycleGAN-VC と同様，循環無矛盾性規準と恒等変換規準を学習規準が，入力音声の発話内容を保持しつつ属性のみを変換する関数 G を得ることを可能にしている．

3.4 テキスト音声合成 (TTS)

TTS では，言語特徴量から音響特徴量を予測する回帰問題において目標値との誤差とともに GAN の目的関数を学習規準に含めることで高品質な音響特徴量予測を行えるよう改良した枠組が提案されている[37]．また，近年の TTS 研究のトレンドの 1 つとなっているのが，テキスト列から音響特徴量系列への変換則を直接学習する End-to-End アプローチである．これは，テキスト列の各テキストがどの時刻の音響特徴量に対応しているかが明示されていない中で系列間の変換則を学習するタスクとなるため，不完全データ問題の一種とみなせる．特に最近は，テキスト列からテキスト列への変換タスクである機械翻訳問題を解決する目的で導入された系列変換 (Sequence-to-Sequence; Seq2Seq) モデル[38] を TTS に適用した手法が注目されている[39)~45]．Seq2Seq モデルは符号化器と復号化器からなり，Seq2Seq モデルを用いた TTS では符号化器はテキスト列 $\mathbf{t}$ から潜在情報 $\mathbf{h}$ を抽出する役割，復号化器は潜在情報 $\mathbf{h}$ をもとに音響特徴量系列 $\mathbf{x}$ を生成する役割を担う．さらに注意 (Attention) 機構[46] を導入することで系列間の各要素の対応関係を見つけ出しながら変換則を学習することが可能になる．ここで注意すべきは，テスト時においてはテキスト列 $\mathbf{t}$ しか与えられないという点である．このため復号化器は，時刻 n における音響特徴量 x_n が，自身がこれまでに生成した音響特徴量系列の履歴 $\mathbf{x}_{<n}$ と潜在情報 $\mathbf{h}$ をもとに逐次的に決定される仕組みをもっている必要があり，ちょうど (17) 式の AGN と同形の深層生成モデルとなる．なお，条件付分布 $p(x_n|\mathbf{x}_{<n}, \mathbf{h})$ を記述する NN としては再帰型 NN (Recurrent NN; RNN) や畳み込み NN (Convolutional NN; CNN) を用いることができる．以上のようなモデルのもとで，復号化器で生成された系列が目標系列とできるだけ一致するように符号化器と復号化器を学習することで，所与のテキスト列 $\tilde{\mathbf{t}}$ に対応する音響特徴量系列 $\tilde{\mathbf{x}}$ を生成するモデルを得ることができる．

3.5 音声強調

音声強調は，雑音や残響などで劣化した音声を回復することを目的とするタスクと音声合成の後処理として合成音声の品質を改善することを目的とするタスクに大きく分類され，それぞれのタスクを対象に GAN を適用した手法がいくつか提案されている．前者のものとしてはたとえば時間領域信号強調法[47] や時間周波数マスク推定法[48]，後者のものとしては特徴量強調法[49] や時間領域信号強調法[50] などが提案されている．

4. おわりに

近年の深層学習の研究の発展により，これまで難しいとされてきた多くの既存タスクにおいて特に性能面で数々のブレイクスルーがもたされたが，深層生成モデルの登場により解決可能なタスクや問題の範囲が大きく拡がってきている．本稿では，深層生成モデルの代表的な例として AGN，VAE，GAN の原理を概説し，音声音響信号処理問題のいくつかの例（音声分析合成，BSS，VC，TTS，音声強調）を題材に深層生成モデルの最近の応用事例を紹介した．この分野は今なお目まぐるしく発展しており，今後も動向が注目される．（2019 年 2 月 27 日受付）

参 考 文 献

1) A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu: Pixel Recurrent Neural Networks, in *Proc. ICML* (2016)
2) A. van den Oord, et al.: WaveNet: A Generative Model for Raw Audio, arXiv:1609.03499 [cs.SD] (2016)
3) D. P. Kingma and M. Welling: Auto-Encoding Variational Bayes, in *Proc. ICLR* (2014)
4) D. P. Kingma, D. J. Rezendey, S. Mohamedy, and M. Welling: Semi-Supervised Learning with Deep Generative Models, in *Adv. NIPS*, 3581/3589 (2014)
5) I. Goodfellow, et al.: Generative Adversarial Nets, in *Adv. NIPS*, 2672/2680 (2014)
6) M. E. Tipping and C. M. Bishop: Probabilistic Principal Component Analysis, *J. R. Statist. Soc. B*, **61**–3, 611/622 (1999)
7) C. J. Maddison, A. Mnih, and Y. W. Teh: The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables, in *Proc. ICLR* (2017)
8) S. Nowozin, B. Cseke, and R. Tomioka: f-GAN: Training Generative Neural Samplers Using Variational Divergence Minimization, in *Adv. NIPS*, 271/279 (2016)
9) X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley: Least Squares Generative Adversarial Networks, in *Proc. ICCV*, 2794/2802 (2017)
10) M. Arjovsky, S. Chintala, and L. Bottou: Wasserstein Generative Adversarial Networks, in *Proc. ICML*, 214/223 (2017)
11) I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville: Improved Training of Wasserstein GANs, in *Adv. NIPS*, 5769/5779 (2017)
12) A. Radford, L. Metz, and S. Chintala: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, in *Proc. ICLR* (2016)
13) A. van den Oord, et al.: Parallel WaveNet: Fast High-Fidelity Speech Synthesis, arXiv:1711.10433 [cs.LG] (2017)
14) N. Kalchbrenner, et al.: Efficient Neural Audio Synthesis, arXiv:1802.08435 [cs.SD] (2018)
15) Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu: FFTNet:

A Real-Time Speaker-Dependent Neural Vocoder, in *Proc. ICASSP*, 2251/2255 (2018)
16) R. Prenger, R. Valle, and B. Catanzaro: WaveGlow: A Flow-Based Generative Network for Speech Synthesis, arXiv:1811.00002 [cs.SD] (2018)
17) A. Ozerov and C. Févotte: Multichannel Nonnegative Matrix-factorization in Convolutive Mixtures for Audio Source Separation, *IEEE Trans. ASLP*, **18**–3, 550/563 (2010)
18) H. Kameoka, T. Yoshioka, M. Hamamura, J. Le Roux, and K. Kashino: Statistical Model of Speech Signals Based on Composite Autoregressive System with Application to Blind Source Separation, in *Proc. LVA/ICA*, 245/253 (2010)
19) H. Sawada, H. Kameoka, S. Araki, and N. Ueda: Multi-Channel Extensions of Non-Negative Matrix Factorization with Complex-Valued Data, *IEEE Trans. ASLP*, **21**–5, 971/982 (2013)
20) D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari: Determined Blind Source Separation Unifying Independent Vector Analysis and Nonnegative Matrix Factorization, *IEEE/ACM Trans. ASLP*, **24**–9, 1626/1641 (2016)
21) H. Kameoka, L. Li, S. Inoue, and S. Makino: Semi-Blind Source Separation with Multichannel Variational Autoencoder, arXiv:1808.00892 [stat.ML] (2018)
22) S. Seki, H. Kameoka, L. Li, T. Toda, and K. Takeda: Generalized Multichannel Variational Autoencoder for Underdetermined Source Separation, arXiv:1810.00223 [stat.ML] (2018)
23) L. Li, H. Kameoka, and S. Makino: Fast MVAE: Joint Separation and Classification of Mixed Sources Based on Multichannel Variational Autoencoder with Auxiliary Classifier, arXiv:1812.06391 [cs.LG] (2018)
24) Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara: Statistical Speech Enhancement Based on Probabilistic Integration of Variational Autoencoder and Non-Negative Matrix Factorization, in *Proc. ICASSP*, 716/720 (2018)
25) Y. Subakan and P. Smaragdis: Generative Adversarial Source Separation, in *Proc. ICASSP*, 26/30 (2018)
26) Y. Stylianou, O. Cappé, and E. Moulines: Continuous Probabilistic Transform for Voice Conversion, *IEEE Trans. SAP*, **6**–2, 131/142 (1998)
27) T. Toda, A. W. Black, and K. Tokuda: Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory, *IEEE Trans. ASLP*, **15**–8, 2222/2235 (2007)
28) C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang: Voice Conversion from Non-Parallel Corpora Using Variational Auto-Encoder, in *Proc. APSIPA-ASC* (2016)
29) C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang: Voice Conversion from Unaligned Corpora Using Variational Autoencoding Wasserstein Generative Adversarial Networks, in *Proc. Interspeech*, 3364/3368 (2017)
30) H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo: ACVAE-VC: Non-Parallel Many-to-Many Voice Conversion with Auxiliary Classifier Variational Autoencoder, arXiv:1808.05092 [stat.ML] (2018)
31) T. Kaneko and H. Kameoka: Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks, arXiv:1711.11293 [stat.ML] (2017)
32) J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros: Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, in *Proc. ICCV*, 2223/2232 (2017)
33) T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim: Learning to Discover Cross-Domain Relations with Generative Adversarial Networks, in *Proc. ICML*, 1857/1865 (2017)
34) Z. Yi, H. Zhang, P. Tan, and M. Gong: DualGAN: Unsupervised Dual Learning for Image-to-Image Translation, in *Proc. ICCV*, 2849/2857 (2017)
35) H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo: StarGAN-VC: Non-Parallel Many-to-Many Voice Conversion Using Star Generative Adversarial Networks, in *Proc. SLT*, 266/273 (2018)
36) Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo: StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation, arXiv:1711.09020 [cs.CV] (2017)
37) Y. Saito, S. Takamichi, and H. Saruwatari: Statistical Parametric Speech Synthesis Incorporating Generative Adversarial NetWorks, *IEEE/ACM Trans. ASLP*, **26**–1, 84/96 (2018)
38) I. Sutskever, O. Vinyals, and Q. V. Le: Sequence to Sequence Learning with Neural Networks, in *Adv. NIPS*, 3104/3112 (2014)
39) Y. Wang, et al.: Tacotron: Towards End-to-End Speech Synthesis, in *Proc. Interspeech*, 4006/4010 (2017)
40) S. Ö. Arık, et al.: Deep Voice: Real-Time Neural Text-to-Speech, in *Proc. ICML* (2017)
41) S. Ö. Arık, et al.: Deep Voice 2: Multi-Speaker Neural Text-to-Speech, in *Proc. NIPS* (2017)
42) J. Sotelo, et al.: Char2Wav: End-to-End Speech Synthesis, in *Proc. ICLR* (2017)
43) H. Tachibana, K. Uenoyama, and S. Aihara: Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention, in *Proc. ICASSP*, 4784/4788 (2018)
44) W. Ping, et al.: Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning, in *Proc. ICLR* (2018)
45) J. Shen, et al.: Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, in *Proc. ICASSP*, 4779/4783 (2018)
46) M.-T. Luong, H. Pham, and C. D. Manning: Effective Approaches to Attention-Based Neural Machine Translation, in *Proc. EMNLP* (2015)
47) S. Pascual, A. Bonafonte, and J. Serrá: SEGAN: Speech Enhancement Generative Adversarial Network, in *Proc. Interspeech*, 3642/3646 (2017)
48) T. Higuchi, K. Kinoshita, M. Delcroix, and T. Nakatani: Adversarial Training for Data-Driven Speech Enhancement without Parallel Corpus, in *Proc. ASRU*, 40/47 (2017)
49) T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino: Generative Adversarial Network-Based Postfilter for Statistical Parametric Speech Synthesis, in *Proc. ICASSP*, 4910/4914 (2017)
50) K. Tanaka, T. Kaneko, N. Hojo, and H. Kameoka: Synthetic-to-Natural Speech Waveform Conversion Using Cycle-Consistent Adversarial Networks, in *Proc. SLT*, 632/639 (2018)

［著 者 紹 介］

亀 岡 弘 和（かめおか ひろかず）君

1978 年生．2002 年東京大学工学部計数工学科卒業．2007 年同大大学院情報理工学系研究科システム情報学専攻博士課程修了．情報理工学博士．同年日本電信電話（株）コミュニケーション科学基礎研究所入社．2011 年東京大学大学院情報理工学系研究科システム情報学専攻客員准教授，2016 年国立情報学研究所客員准教授．音声音響信号処理の研究に従事．日本音響学会，電子情報通信学会，情報処理学会，IEEE の会員．