# 敵対的生成ネットワークによる振幅スペクトログラムの位相復元*

☆小山田圭佑 (筑波大), 亀岡弘和 (NTT), 金子卓弘 (NTT),
田中宏 (NTT), 北条伸克 (NTT), △安東弘泰 (筑波大)

## 1 はじめに

本稿では振幅スペクトログラムの位相再構成問題を扱う。実世界音響信号において，短時間 Fourier 変換 (Short-Time Fourier Transform; STFT) などにより得られる振幅スペクトログラムには特徴的な構造が現れやすく，振幅スペクトログラムを加工・合成する処理が有効となる場面が多い。特に最近は振幅スペクトログラムを生成する音声合成方式の有効性が示されている [1, 2]。振幅スペクトログラムは位相情報が欠損しているため，加工・合成した振幅スペクトログラムから音響信号を再構成するためには通常位相情報を再構成する必要がある。この位相再構成問題に対し，従来は Griffin-Lim らによる信号処理をベースとした方法 [3]（以後，Griffin-Lim 法）が広く用いられてきた。しかし，Griffin-Lim 法では高品質な音響信号を得るためには多数の反復計算を要することが多く，この点が実時間システムに応用する上での課題であった。また，入力の振幅スペクトログラムによっては，反復回数を増やしても高品質な音響信号を得られない場合があり，この点が品質面での課題として残されていた。

本稿では，以上の課題を解決するため，振幅スペクトログラムから音響信号を再構成するプロセスを深層ニューラルネットワーク (Deep Neural Network; DNN) によりモデル化し，敵対的生成ネットワーク (Generative Adversarial Networks; GAN)[5] の枠組を用いた学習ベースの位相再構成手法を提案する。

## 2 位相再構成問題

時間領域信号を $\mathbf{x} = [x(0), \ldots, x(T-1)]^{\mathsf{T}} \in \mathbb{R}^T$ とすると，その時間周波数表現 $c_{f,n}$（ただし $f$ は周波数，$n$ は時刻のインデックスを表す）は一般に，$\mathbf{x}$ と時刻 $t_n$ 周辺に局在する周波数 $\omega_f$ の複素正弦波 $\mathbf{w}_{f,n} = [w_{f,n}(0), \ldots, w_{f,n}(T-1)]^{\mathsf{T}} \in \mathbb{C}^T$ との内積 $c_{f,n} = \mathbf{w}_{f,n}^{\mathsf{H}}\mathbf{x}$ で与えられる。STFT の場合は $t_n$ がフレーム $n$ の中心時刻に相当し，$\mathbf{w}_{f,n}$ は窓関数を乗じた複素正弦波に当該フレーム以外の区間に 0 を詰めた信号となる。すべての時間周波数成分 $c_{f,n}$ を縦に並べたベクトルを $\mathbf{c} \in \mathbb{C}^{FN}$ とすると，$\mathbf{c}$ と $\mathbf{x}$ との間には

$$\mathbf{c} = \mathbf{W}\mathbf{x} \tag{1}$$

という関係が成り立つ。ただし，$\mathbf{W}$ は各行を $\mathbf{w}_{f,n}^{\mathsf{H}}$ とした $FN \times T$ 行列である。以下，この $\mathbf{c}$ を複素スペクトログラムと呼ぶ。通常，時間周波数点の総数 $FN$ は時間領域信号のサンプル点数 $T$ より大きくとるため，$\mathbf{c}$ は $\mathbf{x}$ の冗長表現となる。すなわち，$\mathbf{c}$ は $\mathbf{W}$ の各列ベクトルによって張られる $T$ 次元の線形部分空間 $\mathcal{C}$ に属する。この冗長性による $\mathbf{c}$ に関する制約は，STFT の場合，各フレームの複素スペクトルの逆 Fourier 変換（信号波形）が隣接するフレームの重複区間において無矛盾でなければならないという制約に相当する。ここで，$\mathbf{c}$ の各要素を絶対値化したベクトルを $\mathbf{a}$（振幅スペクトログラムと呼ぶ）とすると，位相再構成問題はこの制約を手がかりとして $\mathbf{a}$ のみから $\mathbf{x}$ を推定する問題と捉えられる。

## 3 Griffin-Lim 法

以下，[4] の導出に従って Griffin-Lim 法の反復アルゴリズムを導く。

所与の $\mathbf{c}$ が時間領域信号に対応する複素スペクトログラムとしての制約を満たすかどうかは，部分空間 $\mathcal{C}$ への $\mathbf{c}$ の直交射影 $\mathbf{W}\mathbf{W}^{+}\mathbf{c}$ が $\mathbf{c}$ と一致するかどうかにより評価することができる。ただし，$\mathbf{W}^{+}$ は

$$\begin{aligned}\mathbf{W}^{+}\mathbf{c} &= \operatorname*{argmin}_{\mathbf{x}} \|\mathbf{c} - \mathbf{W}\mathbf{x}\|_2^2 \\ &= (\mathbf{W}^{\mathsf{H}}\mathbf{W})^{-1}\mathbf{W}^{\mathsf{H}}\mathbf{c}\end{aligned} \tag{2}$$

を満たす $\mathbf{W}$ の擬似逆行列であり，$\mathbf{W}$ が STFT のときは逆 STFT に相当する。よって，位相 $\phi_{f,n} \equiv e^{\mathrm{j}\theta_{f,n}}$ を要素にもつベクトルを $\boldsymbol{\phi}$ とすると，所与の $\mathbf{a}$ に対する位相再構成問題は，

$$\mathcal{J}(\boldsymbol{\phi}) = \|\mathbf{a} \odot \boldsymbol{\phi} - \mathbf{W}\mathbf{W}^{+}(\mathbf{a} \odot \boldsymbol{\phi})\|_2^2 \tag{3}$$

が最小となる $\boldsymbol{\phi}$ を推定する最適化問題として定式化される。ただし，$\odot$ はベクトルの要素ごとの積を表す。ここで，式 (2) より $\mathbf{W}\mathbf{W}^{+}(\mathbf{a} \odot \boldsymbol{\phi})$ は部分空間 $\mathcal{C}$ の中で $\mathbf{a} \odot \boldsymbol{\phi}$ に最も近い点を表すので，

$$\mathcal{J}(\boldsymbol{\phi}) = \min_{\tilde{\mathbf{c}} \in \mathcal{C}} \|\mathbf{a} \odot \boldsymbol{\phi} - \tilde{\mathbf{c}}\|_2^2 \tag{4}$$

が成り立つ。補助関数法の原理より，$\mathcal{J}^{+}(\boldsymbol{\phi}, \tilde{\mathbf{c}}) \equiv \|\mathbf{a} \odot \boldsymbol{\phi} - \tilde{\mathbf{c}}\|_2^2$ は $\tilde{\mathbf{c}} \in \mathcal{C}$ を補助変数とした $\mathcal{J}(\boldsymbol{\phi})$ の補助関数となり，

$$\tilde{\mathbf{c}} \leftarrow \operatorname*{argmin}_{\tilde{\mathbf{c}} \in \mathcal{C}} \|\mathbf{a} \odot \boldsymbol{\phi} - \tilde{\mathbf{c}}\|_2^2 = \mathbf{W}\mathbf{W}^{+}(\mathbf{a} \odot \boldsymbol{\phi}) \tag{5}$$

$$\boldsymbol{\phi} \leftarrow \operatorname*{argmin}_{\boldsymbol{\phi}} \|\mathbf{a} \odot \boldsymbol{\phi} - \tilde{\mathbf{c}}\|_2^2 = \angle\tilde{\mathbf{c}} \tag{6}$$

*Generative adversarial network-based approach to phase reconstruction from magnitude spectrogram. by OYAMADA, Keisuke (University of Tsukuba), KAMEOKA, Hirokazu(NTT), KANEKO, Takuhiro(NTT), TANAKA, Kou(NTT), HOJO, Nobukatsu(NTT), ANDO, Hiroyasu(University of Tsukuba).

のようなステップを反復的に行うことで $\mathcal{J}(\boldsymbol{\phi})$ の停留点を得ることができる。ただし，$\angle\cdot$ はベクトルの各要素をその絶対値で割る演算を表すものとする。式 (5) は，$\mathbf{a} \odot \boldsymbol{\phi}$ に対し逆 STFT を行った後 STFT を行う操作に相当し，式 (6) は，式 (5) で得られた $\tilde{\mathbf{c}}$ の各要素の偏角を $\boldsymbol{\phi}$ に移植する操作に相当する。これらの処理ステップは Griffin-Lim 法 [3] と手続き的に等価である。

Griffin-Lim 法では，高品質な音響信号を得るには多数の反復回数を要する場合が多い。また，入力振幅スペクトログラムによっては，反復回数を増やしても低品質な音響信号しか得られない場合がある。これらの課題を以下で提案する方法により解決する。

## 4 提案法

### 4.1 位相再構成プロセスのモデル化

$\boldsymbol{\phi}$ の初期値を $\boldsymbol{\phi}^{(0)}$，$h(\mathbf{a}, \boldsymbol{\phi}) \equiv \mathbf{W}\mathbf{W}^{+}\mathbf{a} \odot \boldsymbol{\phi}$, $g(\mathbf{c}) \equiv \angle\mathbf{c}$ と置くと，Griffin-Lim の反復アルゴリズムは

$$\hat{\mathbf{c}} = h(\mathbf{a}, g(\cdots g(h(\mathbf{a}, g(h(\mathbf{a}, \boldsymbol{\phi}^{(0)})))) \cdots)) \qquad (7)$$

のような多層の合成関数に展開することができる。$h$ も $g$ も，入力に対し線形変換後に活性化関数を適用する演算となっているため，式 (7) は，固定の重みパラメータと活性化関数からなる多層ニューラルネットワーク (Deep Neural Network; DNN) と見なすことができる。この観点に立てば，より良い解へより早く収束するアルゴリズムを見つけることは，適切な重みパラメータ（及び活性化関数）を決定する DNN の学習問題と捉えることができる。幸い $\mathbf{c}$ と $\mathbf{a}, \boldsymbol{\phi}$ のペアデータはありとあらゆる時間領域信号を用いて複素スペクトログラムと振幅スペクトログラムを算出することで容易かつ無数に用意することができるので，DNN の学習問題としては大変有利である。

以下，$\mathbf{a}, \boldsymbol{\phi}$ を入力として $\mathbf{c}$（または $\mathbf{x}$）を出力とした DNN を生成器と呼び，$\hat{\mathbf{c}} = G(\mathbf{a}, \boldsymbol{\phi})$ と表す。

### 4.2 学習規準

DNN の学習では，NN 出力と教師データの誤差（$\ell_1$ ノルムなど）を学習規準とすることが多いが，これはデータが何らかの分布（$\ell_1$ ノルムを誤差規準とした場合は Laplace 分布）に従うことを仮定していることに相当する。このようにデータ空間における誤差規準を用いて NN を学習する場合，出力が教師データに対して平均的にフィットするような NN が最適と見なされることとなる。従ってこのように生成器 $G$ を学習すると，過剰に平滑化された信号を生成するようになる可能性が考えられる。実世界信号の多くは少なからずのランダム成分を含むものであり，ランダム成分が信号 (または位相) 再構成の過程で除去されることは再構成信号の聴感上の品質に限界を与える恐れがある。そこで生成器 $G$ とは別に，生成器 $G$ が生成した複素スペクトログラム $\hat{\mathbf{c}}$ なのか実データの複素スペクトログラム $\mathbf{c}$ なのかを識別する NN（以後，識別器 $D$）を導入し，識別器 $D$ の中間層の出力値間で測る誤差を生成器 $G$ の学習規準とすることを考える。学習がある程度進んだ識別器 $D$ の中間層では，$\hat{\mathbf{c}}$ と $\mathbf{c}$ を識別しやすいような（引き離すような）特徴量空間になっているはずであるため，そのような空間で $\hat{\mathbf{c}}$ と $\mathbf{c}$ をできるだけ近づけることで，$\mathbf{c}$ との違いをより見分けられないような $\hat{\mathbf{c}}$ を得られるようになることが期待される。そこで，識別器 $D$ の識別関数を $D(\cdot, \mathbf{a}) \in \mathbb{R}$ とし，

$$\begin{aligned} V(D) =& \frac{1}{2}\mathbb{E}_{(\mathbf{c},\mathbf{a})\sim p_{\mathbf{c},\mathbf{a}}(\mathbf{c},\mathbf{a})}\left[(D(\mathbf{c},\mathbf{a})-1)^2\right] \\ &+\frac{1}{2}\mathbb{E}_{\mathbf{a}\sim p_{\mathbf{a}}(\mathbf{a}),\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}\left[D(G(\mathbf{a},\mathbf{z}),\mathbf{a})^2\right] \end{aligned} \qquad (8)$$

のような規準を考える。ここで，実データに対応するラベルを 1，生成器 $G$ から生成された合成データに対応するラベルを 0 とすると，この規準は，識別器 $D$ が入力の複素スペクトログラムが実データなのか生成器 $G$ から生成されたものなのかを正しく識別できている場合に小さい値をとる識別スコアを表す。よって，$D$ の目標はこの規準を小さくすることである。一方生成器 $G$ の第一の目標は，再構成した複素スペクトログラムが，識別器 $D$ に（誤って）実データと識別されるようにすること，すなわち

$$U(G) = \frac{1}{2}\mathbb{E}_{\mathbf{a}\sim p_{\mathbf{a}}(\mathbf{a}),\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}\left[(D(G(\mathbf{a},\mathbf{z}),\mathbf{a})-1)^2\right] \qquad (9)$$

を小さくすることである。また第二の目標は，出力 $\hat{\mathbf{c}}$ と教師データ $\mathbf{c}$ との誤差を小さくすることである。そこで，$U(G)$ に加えて $\hat{\mathbf{c}}$ と $\mathbf{c}$ の識別器 $D$ の中間層におけるそれぞれの出力値の誤差

$$I(G) = \sum_{l=0}^{L} w_l \|D_l(\mathbf{c}) - D_l(G(\mathbf{a},\mathbf{z}))\|_2^2 \qquad (10)$$

を含めたものを学習規準とする。ただし，$w_l$ は非負の重み定数で，$D_l(\cdot)$ は識別器 $D$ の第 $l$ 層の出力値を表す。よって，$D_0(\mathbf{c}) = \mathbf{c}$ である。

以上より $D$ と $G$ の学習規範は以下となる。

$$D : V(D) \rightarrow \text{minimize} \qquad (11)$$

$$G : U(G) + \lambda I(G) \rightarrow \text{minimize} \qquad (12)$$

ただし，$\lambda$ は非負の重み定数である。

このように識別器と生成器を競争させることにより生成器を学習する方法論は，敵対的生成ネットワーク (Generative Adversarial Network; GAN)[5] と呼ばれる。提案法はこの方法論を，以上のように式 (10) を考慮することで信号（または位相）再構成問題に合った形に導入した点が新しい。式 (8), (9) に示した規準は，Mao らによって導入されたもので，この基準を用いた GAN の枠組みは LSGAN (Least Squares GAN; GAN) と呼ばれる [6]。この他にも [7] などさ

まざまな学習規準が提案されている。本学習方法において設計すべき最適化関数として，式 (8), (9) に限る必要性はなく，[5] や [7] などで提案されているものを用いても良い。

## 5 実験的評価

### 5.1 実験設定

本実験では，音声強調など向けに公開されているデータセット [8] を用いた。[8] にはノイズを含む音声データとノイズを含まないものがあるが，ノイズを含まない音声のみを用いた。また，訓練データとして 28 人，テストデータとして 2 人の音声データが用意されているので，訓練データで学習を行い，テストデータを用いて評価実験を行なった。学習時は，音声データを 0.5 秒の重複を持たせ 1 秒ごとに切り分けた。サンプリング周波数は 16 kHz とした。振幅スペクトログラムは，窓幅 1024 点，シフト幅 512 点，窓関数をブラックマン窓とした STFT により得た。Fig.1 に本研究で構築した DNN の構造を示す。点線より左半分が生成器 G，右半分が識別器 D の構造である。ここで，緑色の層は畳み込み層を示し，その上に記述されている記号はそのハイパーパラメータを示す。例えば”2D k11×11 s1 c64”について考えると，”2D”は二次元の畳み込み演算，”k11×11”はカーネルのサイズが横 11 縦 11，”s1”はストライド幅が 1，”c64”はチャネル数が 64 であることを示している。次に黄色の層は，活性化関数を示す。生成器 G では PReLU[9] を用い，識別器 D では Leaky ReLU[10] を用いた。また，水色の層は”要素和”もしくは，チャネル方向への結合を示す。”concat”と記述されている層は，チャネル方向への結合を行なっている。紫色の層は，全結合層を示し，上についている数字は出力ユニットの数を示す。特に記号のない層は，前の層と同じ設定を用いている。モデルの構造については，Ledig, *et al.*(2016)[11] を参考とした。

生成器 G には振幅スペクトログラムと位相の初期値を直接与えるのではなく，反復回数を 5 回とした Griffin-Lim 法により再構成した複素スペクトログラムを与えた。また，識別器 D に入力された複素スペクトログラムには逆 STFT を適用し，時間領域信号を DNN の入力とした。ここで生成器 G の入出力である複素スペクトログラムは実部と虚部でチャネルを分けた。生成器 G に入力する複素スペクトログラムについて，周波数方向の各次元が平均 0，分散 1 になるような正規化を行う。一方，生成器 G が出力した複素スペクトログラムに対しては，スケールを元に戻す処理を行った。また，式 (10) について，重み定数 $w_l$ は $l = 0$ のとき 0，それ以外の層については 1 とした。式 (12) の $\lambda$ は 1 とした。最適化アルゴリズムは RMSprop[12] を用い，学習率は $5 \times 10^{-5}$，$\alpha = 0.5$ とした。バッチサイズは 10，epoch 数は 73 回で学習を止めた。

### 5.2 学習方法補足

時間領域信号に対して Fourier 変換を行なって得られる複素スペクトログラムの位相成分に対して任意の値を足し，逆 Fourier 変換により時間領域信号に戻すことで，その波形は視覚的に異なるが，人間の聴覚的には元の時間領域信号と同一に知覚されることが知られている。この性質を利用し，各振幅スペクトログラムに対応する音響信号として波形の異なるものを多数用意して，識別器 D の学習に用いた。これにより生成器 G は，位相成分が乱数で与えられた時，聴覚的に同一と知覚される音響信号の中でも復元が容易な波形をもつものを復元するように学習されることが期待できる。本研究では，$[-\pi, \pi]$ の範囲で一様乱数を発生させ，位相スペクトログラムを変化させた。

また，生成器 G の学習時に，位相スペクトログラムをランダム生成するが，あるフレームの位相スペクトログラムは固定する。音響信号の位相成分について考えると，ある瞬間にどのような位相をもっているかべきかは絶対的に決まるものではなく，前後の信号の位相との相対的な関係性からどのような位相をもつべきかは決まる。この性質を利用し，位相をランダムサンプリングする時に，例えば 1 フレーム目の位相スペクトログラムだけ固定しておくと，生成器 G は 2 フレーム目以降の位相スペクトログラムを再構成するにあたり 1 フレーム目との相対的な関係性を学習するだけで良いので効率的に学習が進む可能性がある。本研究では，1 フレーム目の位相スペクトログラムを固定した。その際に，実際の音響信号から得られる位相スペクトログラムを用いた。一方，テスト時には全てのフレームについて位相成分をランダムサンプリングした。

また，STFT により得た複素スペクトログラムの実部は偶関数，虚部が奇関数となる。例えば，STFT のフレーム長を 1024 点にした場合，得られる複素スペクトログラムの周波数ヒ ン数は負の周波数も含めれは 1024 となるか ，対称性を利用すれは ，0 からナイキスト周波数まて の周波数に対応する 513 点のみの情報さえあれは 時間領域信号を構成するのに十分て ある。よって，本研究では生成器 G の入出力とした複素スペクトログラムの周波数ビン数は，0 からナイキスト周波数までとした。

### 5.3 主観的評価実験

本研究では，AB テストを用いて提案手法と既存手法で復元された音響信号の品質について評価した。ここで，音響信号の品質は，不快感がなく自然に聞こえるかどうかという基準で評価させた。既存手法は，アルゴリズムの反復回数を 400 回とした Griffin-Lim 法とした。主観的評価実験の被験者数は 5 人とし，1 人の被験者につき，発話内容が同一である復元音声のペ

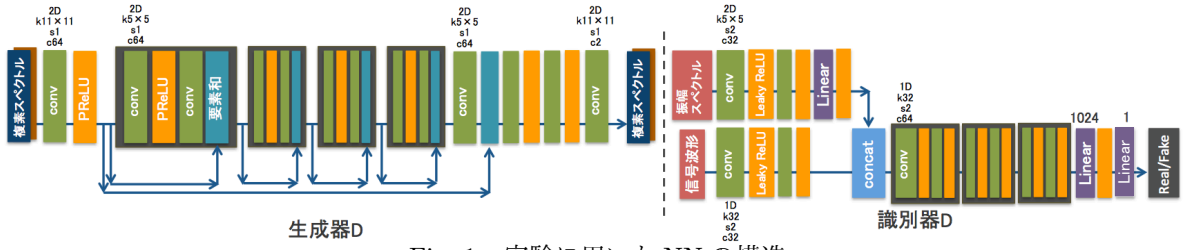


Fig. 1　実験に用いた NN の構造

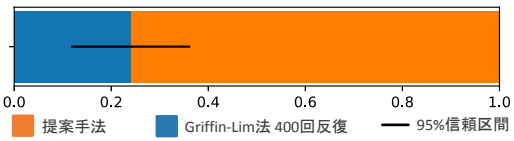


Fig. 2　AB テストの結果

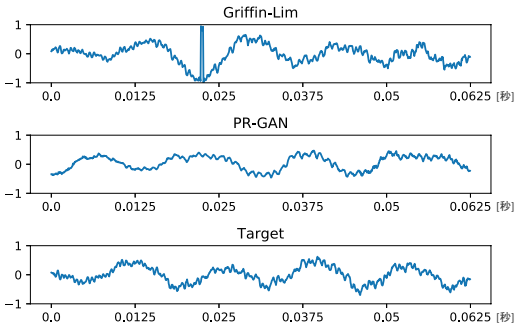


Fig. 3　音楽データの復元結果

アをランダムで 10 ペア与えた。このときテストデータのうち 2-5 秒の長さの音声のみを評価対象とした。

Fig.2 に AB テストの結果を示す。Fig.2 からわかるように，本実験では 50 ペア (10 ペア/人 × 5 人) のうち約 76%のペアについて，反復回数を 400 回とした Griffin-Lim 法よりも提案手法を用いて復元した音声の方が品質が高いという評価が得られた。よって，Griffin-Lim 法よりも提案手法により復元した音響信号の方が品質が高いことが示された。

### 5.4　汎化性能に関する考察

Fig.3 に音楽データ [13] より得た振幅スペクトログラムから音響信号を復元させた結果を示す。1 段目は Griffin-Lim 法，2 段目は提案手法により復元された音響信号を示し，3 段目は実世界音響信号を示す。なお，Griffin-Lim 法の反復回数は 400 回とし，提案手法は 5.1 に示した設定で学習したモデルを用いた。Griffin-Lim 法で復元した音楽データは，Fig.3 にあるように非連続な変化が度々見られた。一方で，提案手法で復元した音楽データは，Griffin-Lim 法よりも非連続な変化が少なかった。音声データのみで学習した提案手法が，音楽データに関しても音響信号を復元できていることが分かる。

## 6　まとめ

本研究では，振幅スペクトログラムの高速・高精度な位相再構成アルゴリズムを実現することを目的として、振幅スペクトログラムから時間領域信号を生成するプロセスを DNN によりモデル化し、GAN の枠組みにより学習するアプローチを検討した。主観的評価実験により反復回数を 400 回とした Griffin-Lim 法よりも提案手法の方が品質の高い時間領域信号を復元可能であることを示した。

**謝辞**　本研究は，筑波大学人工知能科学センターの研究の一環として行われたものである。また，本研究は，JSPS 科研費 17H01763 の助成を受けた。

## 参考文献

[1] S. Takaki, H. Kameoka, J. Yamagishi, “Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis,” in Proc. Interspeech, pp. 1128–1132, 2017.

[2] Y. Wang, et al., “Tacotron: A fully end-to-end text-to-speech synthesis model,” arXiv preprint arXiv:1703.10135, 2017.

[3] D. W. Griffin and J. S. Lim, “Signal estimation from modified short-time Fourier transform,” IEEE Trans. ASSP, vol. 32, no. 2, pp. 236–243, 1984.

[4] J. Le Roux, H. Kameoka, N. Ono, S. Sagayama, “Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency,” in Proc. DAFx, pp. 397–403, 2010.

[5] I. Goodfellow, et al. “Generative adversarial nets,” in Adv. NIPS, pp. 2672–2680, 2014.

[6] X. Mao, et al., “Least squares generative adversarial networks,” arXiv preprint ArXiv:1611.04076, 2016.

[7] M. Arjovsky, et al., “Wasserstein GAN,” arXiv preprint arXiv:1701.07875, 2017.

[8] C. Valentini-Botinhao, et al., “Superseded-Noisy speech database for training speech enhancement algorithms and TTS models,” University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR), 2016.

[9] K. He, et al. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.,” in Proc. ICCV, pp. 1026–1034, 2015.

[10] A. Maas, et al. “Rectifier nonlinearities improve neural network acoustic models. ,” in Proc. ICML vol. 30, no. 1, 2013.

[11] C. Ledig, et al., “Photo-realistic single image super-resolution using a generative adversarial network,” arXiv preprint arXiv:1609.04802, 2016.

[12] T. Tieleman and G. Hinton, “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. ,” COURSERA: Neural networks for machine learning, 4(2), 26-31, 2012.

[13] CAFÉ DEL CHILLIA, “In The Story That We Say,” https://www.jamendo.com/track/1455877/in-the-story-that-we-say , 2017.