The 34th Annual Conference of the Japanese Society for Artificial Intelligence, 2020

画像変換手法による音声強調のためのスペクトログラム変換

Spectrogram Transforms for Speech Enhancement by Image-to-image Translation

黒澤 義明* 目良 和也* 竹澤 寿幸*
KUROSAWA Yoshiaki MERA Kazuya TAKEZAWA Toshiyuki

*広島市立大学大学院 情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

We aimed to examine well-known image-to-image translation technique, so-called pix2pix based on deep neural networks. Focusing on time-frequency analysis and implementing auxiliary classifier generative adversarial networks (ACGAN), we estimated the transform performance of spectrograms for sound enhancement. As a result using an image index, SSIM, we confirmed to slightly improve its performance compared to the original research.

1. はじめに

深層学習研究の進展に伴い，様々なアルゴリズムが提案されている．対象メディアは，画像，動画，音声等多岐にわたる．その中でも，画像に関する研究は非常に多く，その知見は日々更新されている．中でも，Generative Adversarial Networks（GAN）の研究は進展が著しい．

音声に対し GAN の適用を試みた研究も増えている．研究内容も様々で，ブラインド音源分離や声質変換に利用される．その基本技術としては，U-Net[Ronneberger 15]や pix2pix[Isola 17]等の畳み込みによる画像変換が挙げられる．

このように，画像変換技術による音声変換の様々な試みは成功していると言えよう（cf. [Michelsanti 17] [Hiroshiba 18]）．一方で，畳み込みカーネルが，音声の 2 次元特徴量表現の全領域に対して適用される点については問題がある．

視覚的画像は，全領域が走査の対象となる．画像中のどこにあっても，その形状に違いはないからである．一方，聴覚的 2 次元特徴量については，周波数方向で同じ特徴を持つとは限らない．例えば，聴覚的 2 次元特徴量（スペクトログラム，図 1（右））では，上方の高周波数帯と下方の低周波数帯とで，出現する波線の形状，濃度が異なる．すなわち，全領域でなく，特定の領域だけを走査すればよいことが示唆される．

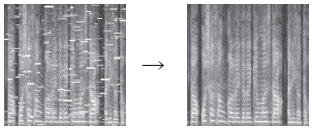


図 1　本研究の課題（スペクトログラム変換）

そこで本研究では，周波数に対応可能な枠組みを提案する．具体的には，先行研究の pix2pix に，[Odena 17]により提案された ACGAN（Auxiliary Classifier GAN）を組み込む．ACGAN に導入される分類課題は，スペクトログラムの周波数帯推定である．この推定の導入により，変換精度を高めることを目的とする．

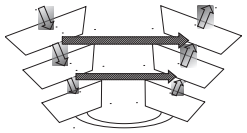


図 2　U-Net の模式図

連絡先：黒澤 義明，広島市立大学大学院 情報科学研究科，広島市安佐南区大塚東 3-4-1，082-830-1500(代)，kurosawa@hiroshima-cu.ac.jp

2. 関連研究

本研究は，2 次元で表現された音声特徴量（スペクトログラム，図 1（右））を用いる．具体的には，短時間フーリエ変換(STFT)を用いて振幅と位相に分離した後，振幅情報を時間と周波数の2次元で表現する．

2 次元特徴量を用いる利点は，①時間周波数情報の強調，②データ量の削減，③画像としての扱いやすさ，等が挙げられる．これにより，様々な画像に対するアルゴリズムの適用が可能となる．本研究は画像変換に有効な pix2pix を用いる[Isola 17]．

2.1 pix2pix

pix2pix は conditional Generative Adversarial Networks（GAN）の一種に基づく画像変換手法である[Isola 17]．線画への着色や画像セグメンテーション等，様々な変換が可能である．後述するように，音声の変換，音声に関する研究においても pix2pix が使われている．ここで，pix2pix の特徴を挙げる．

(1) Generative Adversarial Networks（GAN）

生成器 G(Generator)と判別器 D(Discriminator)という，2 つの畳み込みニューラルネットワークを対立的に使用する．

生成器 G の役割は，訓練画像に似た画像(fake)を生成することである．一方の判別器 D は，D への入力が訓練画像か，G が生成した fake かを判定する．そして，G は D を騙すように，D は G に騙されないように，お互いに競い合って訓練されるアルゴリズムである．

(2) U-Net

特徴の 2 つめは，畳み込みネットワークに U-Net を使用（生成器 G）している点である．医療分野での画像セグメンテーション（細胞の着色 etc）を目的として開発された技術である[Ronneberger 15]．ネットワークの構造が最も重要であり，名前の通り U という形状を有する（図 2）．

U 字の左右をつなぐ結合を skip connection と呼ぶ（右向き車線矢印）．U 字の右側では，下層からの特徴マップと，対応する左側の層からの特徴マップが繋がる．これにより，結合のない単なる Encoder-Decoder モデルよりも，速く正確に学習することができる．

2.2 U-Net/pix2pix を用いた音声研究

次に，U-Net/pix2pix を用いた関連研究，特に音源分離に関連した研究について述べる．

(1) カラオケを用いたボーカル/曲分離 [Jansson 17]

ネットワークに U-Net を使用し，楽曲とカラオケを利用することで，楽曲からのボーカル分離（または，その逆）を試みた研究である．GAN は使われていないものの，U-Net を用いた変換技術という観点から紹介する．

人と楽器という周波数特徴が異なる対象を分離しているため，特徴が似た対象（例えば，人間同士）にも適用可能か，検討が必要と考えられる．

(2) 環境音からの音声強調 [Michelsanti 17]

pix2pix を利用し，ホワイトノイズや飛行機音等から，目的とする音声を強調することを試みている．

上の論文と同様，周波数帯が重なる対象への適用について検討する必要があろう．

(3) 発話音声の分離 [Takaichi 19]

人間の声同士を分離しようと試みている．「女性同士より，男性と女性の方が分離が容易」という結論を得ている一方で，「『高く発声した男性』と『低く発声した女性』を分離できない」という問題も生じている．

彼らの手続きは，2 人の声を同時に学習する方法であるため，個人に関する特徴を学習したというよりも，2 人の相対的な高低関係を学習しただけという可能性がある．この点，周波数を考慮した実験統制を要する．

2.3 カーネルの物体走査とスペクトログラム走査の差異

物体が写った画像と，音声を変換した 2 次元特徴量ははたして同一と考えられるのだろうか？　先行研究に欠けている点は，物体が写った画像と，音声を変換した 2 次元特徴量との差異に着目せず，画像に対する処理と同様の処理を行っているところである．ここでは両者の差異に着目する（表 1）．

表 1　物体と音声特徴量の違い

物体	輪郭がはっきり
	画像中のどこにあるかに非依存
	画像中の大きさ不定
音声	輪郭あいまい
	画像中の場所によって変化
	画像中の大きさ一定

「場所によって変化」は，特に周波数方向で観測される．例えば，図 1 の例では，低周波数（図の下方）と高周波数（図の上方）を比べると，波模様の大きさが異なることがわかる．つまり，全領域を走査する必要はなく，特定の周波数領域だけをサポートするようなカーネルを用いた方が，性能向上の可能性が高いと考えられる．

そこで本研究では，上に述べた周波数方向への対応を行うため，ACGAN（Auxiliary Classifier GAN）を用い，画像の変換に加え，周波数のクラス分類を同時に行う手法を提案する．

2.4 Auxiliary Classifier GAN（ACGAN）

通常の GAN の判別器 D は入力の「真/偽」を判定する．これに加え，ACGAN では入力の「クラス」を同時に推定する．

GAN の枠組みでは，適当に変形だけされた画像（例えば，「犬」）で，D を騙せたかもしれない．しかし，ACGAN では，その「クラス」の特徴，すなわち「犬の特徴」を維持する必要があり，特徴を失うような変形は許さず，質の高い変形を要求する．

本研究では先に述べた周波数に対し，この考えを適用する．どのように拡張するかについて，次節に述べる．

3. 提案手法

前節に述べたように，本研究の提案手法は pix2pix に周波数処理を導入することである．そして，そのために ACGAN を採用する．また，ノイズ除去についての検討を行う．

3.1 ACGAN による周波数拡張

今回はシンプルな手法～周波数ラベルを導入し，そのラベルを推定する～を試みる．具体的な手順は，概略図に示したように，入力されたスペクトログラムを分割し，その入力の「真/偽」と「周波数クラス」を推定する．例えば，256x256 の画像を 16x16 というサイズへの分割ならば，Frq0～Frq15 までの 16 分類を行う．

この分割数が適切で，かつ周波数の特徴をうまく表現できれば，生成スペクトログラムの改善が期待される．

3.2 ノイズ付与

前項に掲げたように，特定の 2 人を選んで学習させるだけでは，相対的な音高の高低だけを学習する可能性があり，汎用的な学習結果にはならない．また，周波数帯が明らかに異なる 2 種の音源を選んだ学習も同様である．

そこで，今回は人間の声の周波数帯に重なるよう，人工的なノイズ～ブロック様ノイズ（図 1（左））～を生成する．一見，スペクトログラムの波形状と見間違う部分もあり，このような同一周波数帯に混在するノイズが分離可能かについて議論する．

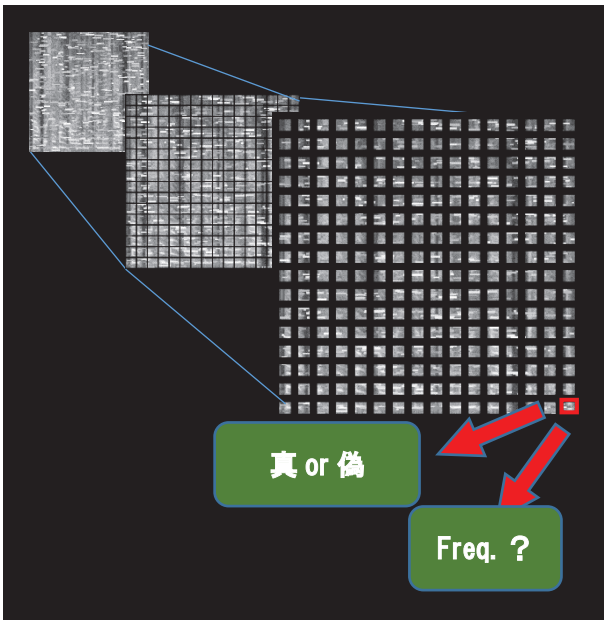


図 4　提案手法　（with Auxiliary Classifier）

4. 実験と考察

まず 3.2 で述べた pix2pix を用いたノイズ除去実験を行う．

4.1 本研究で使用するデータ

女性声優のラジオ CD（Web ラジオの CD 収録版）を用いた．深層学習に利用するため，長時間の音源を選んだ．

(1) 正解データ(ground-truth)作成

放送中の発話以外の区間（CM，音楽，プロデューサ等の該当声優以外の笑い声等）を手動で除去した上で，閾値（無音区間の音圧・間隔）により非発話区間を削除した．また，Sampling Rate を 16,000Hz にした．STFT 関係のパラメータについては，

Window Size: 512（ハミング窓），Shift Size を 128 とした．振幅と位相を分離し，振幅の対数パワーを求めた．

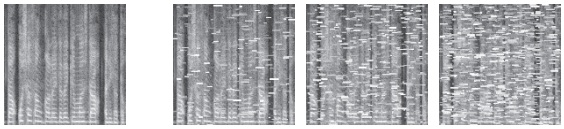
その後，256 x 256 のサイズを抽出し，各 bin を 8bit 整数で表現し，画像化した．これを，本実験の正解画像とする．

(2) ノイズ画像生成

矩形を画像中のランダムな位置に付加することで，前項に挙げたノイズを実現する．矩形の大きさ，輝度について記す．

- 高さ 1~3 のうちの整数（pixel）
- 幅 2~9 のうちの整数（pixel）
- 輝度は 128~255 のうちの整数

全ての画像に対し，上記の手続きを行い，矩形の pixel が総画素数のおよそ 10%，25%，50%の 3 種類になるように，ノイズ比データを 3 系列用意した．画像の例を図に示す（図 5）．



(a) オリジナル　(b) 付加後～10%，25%，50%

図 5　付加前後のスペクトログラム

データ数はそれぞれ，Train: 10,000，Validation: 1,000，Test: 1,000 件とした．重複はない．過学習を避けるため，モデルの選別と評価は別のデータを用いる．

4.2 実験 1 ~ノイズ除去に対する基本的検討~

pix2pix の実装として，原著者らが公開している U-Net256 を用いた実験を行い，前節に挙げたノイズの除去が可能かについて検討する．

なお本研究は，上記の実装にいくつかの変更を加え，実験を行った．変更点は以下の通り．

- spectral normalization の導入[Miyato 18]
- loss の変更（SSIM）

(1) SSIM の採用

本研究では，変換がどの程度向上したかについての指標として，以下の Structural Similarity（SSIM）を用い，G についての loss として採用する．MSE を採用した場合に起こりがちな，画像の局所的な破綻に対し敏感な指標だからである．また，GAN を採用した他の研究でも，この指標が使用されているからである[Yoo 16]．

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

SSIM は，画素中の小領域の平均 μ・分散 ρ・共分散 ρ_{xy} を用いた指標である．なお，SSIM 自体の最良値は 1 であり，1 から減じることで，loss として使用する．

なお，本研究では音声に関する評価手法を使わず，この指標のみを評価に用いる．

(2) 訓練の実行

訓練のパラメータとして，G/D のカーネル数(ndf, ngf)を 64 とし，バッチサイズ 50，エポック数 200 の実験を行った．64，200 は，同実装のデフォルトである．また，5 エポックごとにモデルを生成する設定にした．

これ以外のパラメータについては，デフォルトのパラメータを使用した．GPU は Titan RTX，Geforce GTX 1080Ti を使用し，訓練を行った．学習時の loss の遷移を示す（図 6）．

なお実験条件により，使用 GPU が異なるため，今回，実行時間の計測は行っていない．

(3) モデルの選別

次に，得られたモデル 40 個に対し validation set を入力し，適切なモデルの選別を行う．SSIM の遷移を示す（図 7）．

各々の学習器が最大の値を示す epoch とその時の SSIM 値の平均値を表 2 に示す．

表 2　SSIM 平均の最大値

ノイズ比	epoch	SSIM
10%	120	0.9704
25%	130	0.9370
50%	165	0.8874

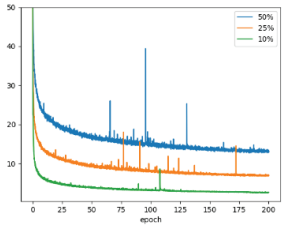


図 6　loss の推移（学習時）

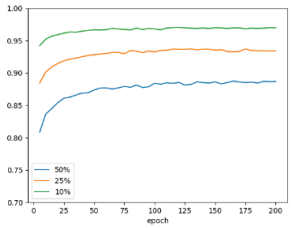


図 7　SSIM の推移（validation set）

次に，上で選別した学習器に対し，test set での評価を行う．なお今回は，120 epoch 時のノイズ比 10%のデータ，165 epoch 時の 50%のデータについて報告する．

10%データに対する最小値は 0.9287，最大値は 0.9788 となった．SSIM が最大，最小となった入力(input)，予測画像(predict)，正解画像(ground-truth)を図に示す（図 8）．

最小/最大の大きくないことがわかる．10%のノイズ付加であれば，pix2pix による変換が可能であることがわかった．実際の音声の分離が対象となる場合でも，この程度の重なりならば，分離ができると考えられる．

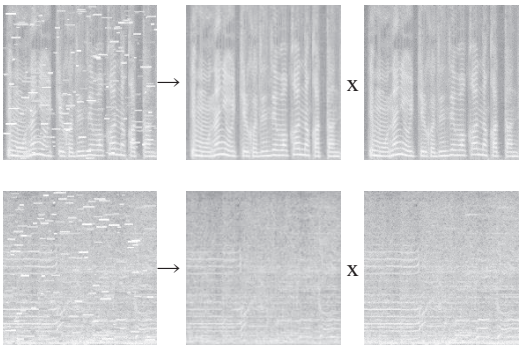


図 8　画像比較 1：input → predict x ground-truth
10% @ 120epoch

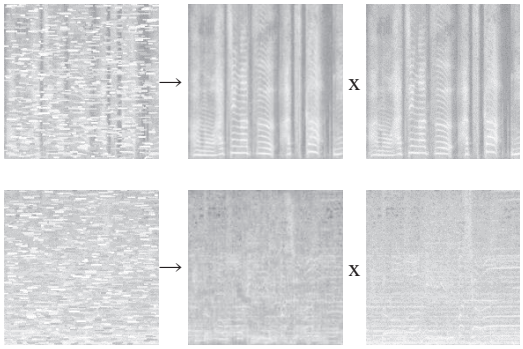


図 9　画像比較 2：input → predict x ground-truth 50% @ 165 epoch

一方，50%条件に対する最小値は 0.7598，最大値は 0.9148 となり，10%条件に比べ，両者の差が大きくなった．SSIM が最大，最小となった画像を示す（図 9）

元の模様が明瞭ならば（図 9 上），見かけ上，かなりのノイズがあっても復元できる．一方で，線が不明瞭な場合（図 9 下）には，過剰に適用し，復元に失敗することがわかる．50%条件は難しい課題と言えよう．

4.3 実験 2 ~周波数特性に着目したノイズ除去~

実験 1 の結果，矩形ノイズを除去できることがわかったため，周波数対応拡張実験を行う．実験1の実装に変更を加え，実験を行った．変更点は以下の通りであり，考察を加える．

・画像サイズ 16x16 への対応（U-Net16）
・ACGAN の導入[Odena 17]

なお，分割数を 16 に決定し，画像を 16x16 に分割し，それぞれを学習に用いた．このため，実験に使用した画像数は，全て 256 倍となった．分割数以外は，実験 1 で用いた画像と同一である．

画像以外の変更点としては，G/D のカーネル数(ndf, ngf)を 64 から 256 に増やした点が挙げられる．これは単純に入力画像サイズが小さくなったため，GPU のメモリに余裕ができたからである．

ノイズ比 10%条件の結果を，図 10 に示す．図中「10%_unet16_ACGAN」が分割した提案手法（ACGAN 有）であり，「10%_unet16_each」は分割だけの手法（ACGAN 無）である．そして，「10%_unet256」は比較のための pix2pix の結果である．

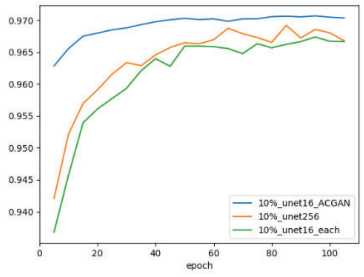


図 10　SSIM 値の比較（validation set）

他の手法と比べ，わずかながら提案手法（ACGAN 有）の方が SSIM の値が向上している．また，20epoch 程度の学習の早い段階から，学習が進んでいる．ACGAN 無の条件と比べても速い．このことは，単なる分割だけの効果ではなく，ACGAN 導入の有効性を示していると言える．

ただし，その差はわずかであり，分割数/分割法の検討とともに議論が必要である．

5. おわりに

本研究は，ACGAN を用いた，周波数特性に特化した学習法を提案した．実験の結果，pix2pix 手法に比べ，わずかながら性能が向上することが明らかになった．また，単に分割するだけと比べても，より早い学習段階で精度が上がることがわかった．

今回は音声に復元することをせず画像のみの評価を行ったため，今後は，音声に復元した上での評価を行う必要がある．また，他の手法（ex. DeepClustering [Isik 17]）との比較を行う等，手法の有効性を吟味する必要がある．

謝辞

本研究の一部は，国立研究開発法人科学技術振興機構（JST）の研究成果展開事業「センター・オブ・イノベーション（COI）プログラム」グラント番号 JPMJCE1311 の支援によって行われている．また，広島市立大学特定研究費（先端学術研究費 H27～29，30 年度科研費獲得支援費）の支援を得ている．

参考文献

[Donahue 18] C. Donahue, B. Li, R. Prabhavalkar: Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition, arXiv:1711.05747v2, 2018.

[Hiroshiba 18] 廣芝, 能勢, 宮本 , 伊藤 , 小田桐: 畳込みニューラルネットワークを用いた音響特徴量変換とスペクトログラム高精細化による声質変換, 研究報告音声言語情報処理（SLP）, 2018-SLP-122(27), 2018.

[Isik 16] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, J. R. Hershey: Single-Channel Multi-Speaker Separation using Deep Clustering, in Proceedings of Interspeech2016, 2016.

[Isola 17] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros : Image-to-Image Translation with Conditional Adversarial Networks, In IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[Jansson 17] A. Jansson, E. J. Humphrey, N. Montecchio, R. Bittner, A. Kumar, T. Weyde: Singing Voice Separation with Deep U-Net Convolutional Networks, in Proceedings of the 18th International Society for Music Information Retrieval Conference, 2017.

[Michelsanti 17] D. Michelsanti, Z.-H. Tan: Conditional Generative Adversarial Networks for Speech Enhancement and Noise-Robust Speaker Verification, arXiv:1709.01703, 2017.

[Miyato 18] T. Miyato, T. Kataoka, M. Koyama, Y. Yoshida : Spectral Normalization for Generative Adversarial Networks, arXiv:1802.05957, 2018.

[Odena 17] A. Odena, C. Olah, J. Shlens: Conditional Image Synthesis with Auxiliary Classifier GANs, in Proceedings of the 34th International Conference on Machine Learning, PMLR 70:2642-2651, 2017.

[Ronneberger 15] O. Ronneberger, P. Fischer, T. Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation, arXiv:1505.04597, 2015.

[Takaichi 19] 高市, 片上, 黒澤, 目良, 竹澤: 深層学習を用いた画像変換に基づく会話からの音声抽出, 人工知能学会全国大会論文集, 2019.

[Yoo 16] D. Yoo, N. Kim, S. Park, A. S. Paek, I. S. Kweon: Pixel-Level Domain Transfer, arXiv:1603.07442, 2016.