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

Speech extraction from conversation based on image-to-image translation using deep neural networks

髙市 晃佑[*1] 片上 敬雄[*2] 黒澤 義明[*1] 目良 和也[*1] 竹澤 寿幸[*1]

Kosuke Takaichi Yoshio Katagami Yoshiaki Kurosawa Kazuya Mera Toshiyuki Takezawa

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

[*2] 広島市立大学 情報科学部
School of Information Sciences
Hiroshima City University

We aim to separate sound sources by deep neural networks which has been active in recent years. We attempt to extract a certain human voice from usual conversation using the networks. We focus on image-to-image translation: pix2pix. The algorithm of pix2pix bases on purely procedure of the image processing. Therefore, we need an additional procedure, that is, we convert voice to spectrogram once. After that we perform to learn the networks to separate human voice, we especially pay attention to segmentation between the same sex and opposite sex. Form this point of view, we conducted two experiments using the sounds overlapped both sexes in this paper. Structure-Similarity (SSIM) index and color map representation were used as evaluation criteria. As a result, we confirmed the good extraction of the female voice from the one synthesized both sexes. However, we did not extract the female voice from same sex. Although we reached the conclusion that the separation did not work well, the generated voice seemed to be played naturally. This is not objective judgment. For this reason, it is our future work.

1. はじめに

これまでにも音源分離の技術が開発されてきた. 例えば, 複数の音源を複数のマイクで録音したデータから, それぞれの音源を分離するときに使われるブラインド信号分離等である. しかし, 音源が互いに独立であると仮定する等があり, これらが成り立たない場合, 適切な結果が出ない. また, 人間の声と音楽等の分解を行っていることが多く, 人間の声同士の分解は例が少ない.

そこで, 音源の重なりを近年盛んである深層学習を利用し分離することを目的とし, 合成音から一人の音声の抽出を行う. また, 音声はすべてスペクトログラムと呼ばれる画像に一度変換する. 一度画像に変換する理由は, 音声に関する研究よりも画像に関する研究の方が多く, 画像に関する知見が使用できると考えたからである. そのため, 本稿では, 様々な画像に対応でき, 画像から学習した通りに加工し画像を出力することのできる pix2pix を用いて, 音源の抽出を行う.

2. 提案手法

先行研究に楽曲からボーカルと音楽を分離している研究がある[Jansson 17]. 音楽と人間の声が分離できるのであれば, 人間同士の発話の重なりも分離できるのではないかと考えられる.

また, RNN（Recurrent Neural Networks）という手法で分離する研究も存在している[Huang 15]. 音声を一度スペクトログラムに変換する必要はない. しかし, 今回は画像に関する知見が使用できることが有用であると考え, 画像を中心に学習を行う.

本稿では, 音声をスペクトログラムへ変換し, pix2pix を用いて深層学習を行い, 音声抽出を試みる. pix2pix は GAN（Generative Adversarial Network）を利用した画像変換アルゴリズムの一種であり, 例えば, 白黒写真からカラー写真を生成したり, 地図から航空写真を生成したりできる. Generator と Discriminator の 2 つのモデルが敵対することで変換精度を高めていくモデルである. 図 1 に pix2pix の変換例を示す. 左が入力であり, 右が出力(predict)である.

連絡先：髙市晃佑, 広島市立大学情報科学部, 広島市安佐南区大塚東, takaichi@ls.info.hiroshima-cu.ac.jp

本稿は, まず女性と男性の声に現れる特徴の違いに着目する. 人間の男性と女性の声には明確に高低差が存在する. 具体的には, 男性と女性の声では平均基本周波数が異なる. 基本周波数とは,「周期的に生じる生体振動の時間間隔のうち, 最短の間隔として与えられる基本周期」[森勢 18]であり,「人間が知覚する声の高さにおおむね対応する」[森勢 18]. 本稿では, その差に注目する. すなわち, 男性の声の平均基本周波数は約 120Hz であり, 女性では 240Hz であると報告されている[寺澤 84]. 同様に, 男女の声の基本周波数と年齢の関係を報告している研究もある[粕谷 68]. その研究によると 14 歳以降では男女の声の平均基本周波数は大きく差がでている.

そこで本研究では, 男女の声の平均基本周波数の違いから, 男女の声には大きな特徴の差異があると仮定し, 音声をスペクトログラムへ変換を行い, その違いを学習する. そして, 同性間においてもある程度違いが表れると考え, 同様に違いを学習する. 基本周波数が異なる男女の声の分離は容易であると思われる. しかし, 差が小さい女性同士の分離は困難であると思われる.

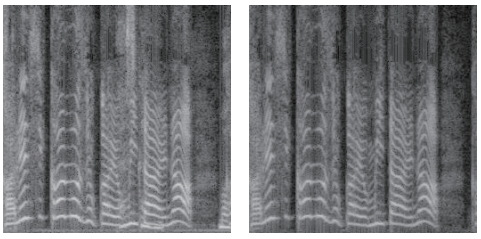


図 1 pix2pix の入出力例

3. 実験と結果

本稿では, 女性だけの音声と男女の合成音を変換したスペクトログラムをもとに, pix2pix による深層学習を行う実験（3.1 節）と女性だけの音声と女性同士の合成音を同様に変換し, 深層学習を行う実験（3.2 節）を行う.

また, 生成された画像は元の音声のスペクトログラムと比較して評価を行う. 比較方法として SSIM(Structural Similarity)(式 1)とカラーマップ（図 2）を求め精度を測る. SSIM は人間が感じる違いに近い結果を返し, 1 に近いほど同じと言える.

カラーマップは生成画像と元のスペクトログラム，2 つの画像の差で表し，青(左)に近いほど未変換，赤(右)に近いほど過変換，中央値の緑に近いほど変換ができていると言える．

図 2　カラーマップの見方

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \qquad (1)$$

3.1 男女の分離

実験の際に，Discriminator のネットワークパラメータを変更して 3 種の条件で実験を行い，各条件に対して validation，test を実行する．なお，本研究で使用したネットワークパラメータは，以下の通りである．

1. NLayerDiscriminator のカーネル数が 64
2. NLayerDiscriminator のカーネル数が 128
3. PixelDiscriminator のカーネル数が 64

1 と 2 の NLayerDiscriminator のネットワークと 3 の PixelDiscriminator のネットワークを図 3 と図 4 に記載する．なお、ここにある ngf はカーネル数のことである．

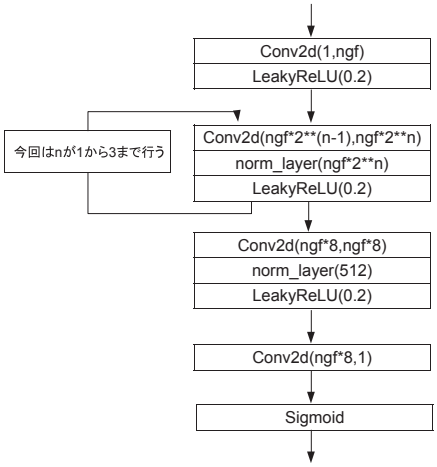


図 3　NLayerDiscriminator のネットワーク

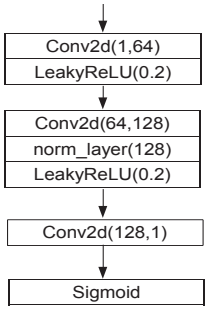


図 4　PixelDiscriminator のネットワーク

NLayerDiscriminator と PixelDiscriminator の違いは畳み込みを行う際に，PixelDiscriminator は 1pixel 毎に読み込みと書き出しを行う点にある．

3 条件中で最も SSIM 値が高かった条件(0.806)について述べる．元音声のスペクトログラムと合成音のスペクトログラム(図 5)，Discriminator が生成したスペクトログラムと SSIM を基にしたカラーマップ(図 6)を示す．次に精度比較用に，最も精度が低かった条件(0.652)による，同一音声区間の生成画像とカラーマップ(図 7)を示す．

図 5，図 6 の比較から，元音声の特徴は正確に残しつつ，図 5 右側の左下付近，特に音が重なっている部分は変換ができた．このことは，図 6 のカラーマップからもわかる．しかし，図 7 の左では左下部分は合成音が残っており，図 7 のカラーマップからも確認できる．つまり，ネットワークパラメータによって変換精度に大きく差があると言える．また，全体的に高周波域や無音区間の変換は過変換気味であることが，カラーマップから読み取れる．これにより SSIM 値が低下したと考えられる．

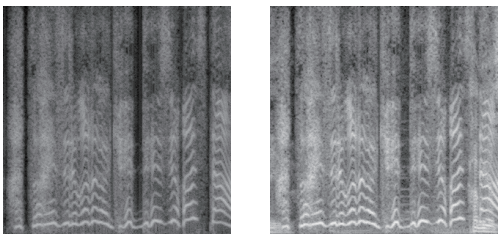


図 5　元音声（左）と合成音（右）のスペクトログラム

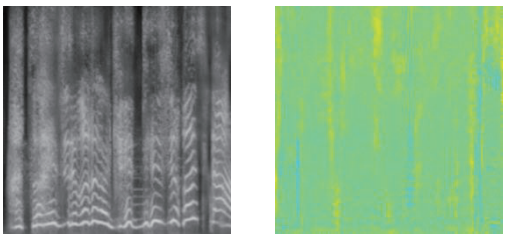


図 6　生成されたスペクトログラム（左）とカラーマップ

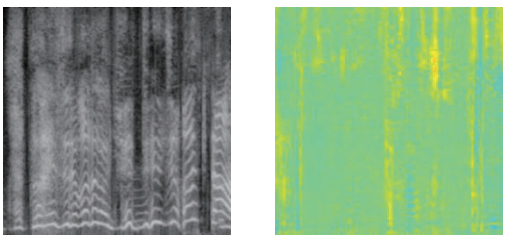


図 7　別のネットワークの生成画像（左）とカラーマップ

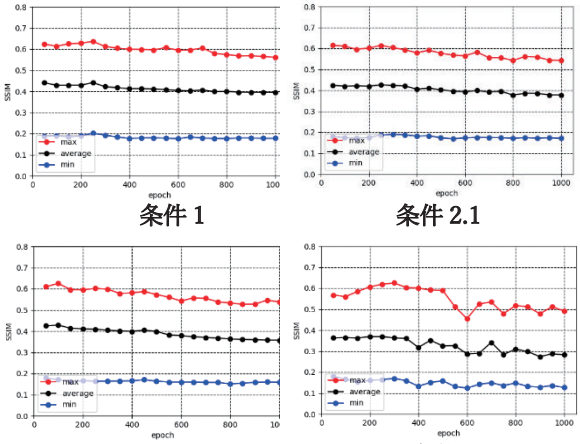
3.2 女性同士の分離

女性同士の場合では，学習データを以下のように変え，学習を行った．

1. 単独の発話のみを学習
2. 1. 対話からの発話を全体の 10%になるよう追加し学習
2. 2. 対話からの発話を全体の 20%になるよう追加し学習
3. 無音区間を発話の前後に追加し，全体の 10%になるよう追加し学習

条件 1 では，学習データにラジオ CD から収集した単独の発話をスペクトログラムで 9,235 件収集し，これを学習した．他の条件ではこの 9,235 件に追加していく．各条件 1000epoch まで学習し，SSIM 値の推移とテスト結果にはそれぞれ最も良い世代の変換を記載する．図 8，図 9 では，条件 1(左上)，条件 2.1(右上)，条件 2.2(左下)，条件 3(右下)とする．

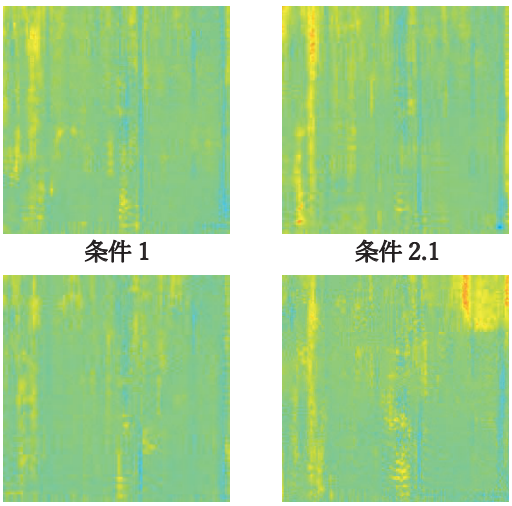
SSIM 値については，図 8 より 100～300epoch においてピークを迎え，以降は緩やかに減少している．図 9 より，中央下の重なりが分離できていない等，変換できていない部分は同じであることがわかる．また，図 9 右下の条件 3 の結果は右上に過変換部分が見受けられる．これは右上が無音区間であると判断したことが原因だと考えられる．また，SSIM 値については，図 8 より，いずれも 0.6 ほどであり，右上以外の部分は変換できていると考えられる．しかし，男女間の実験と同様に高周波域の変換は過変換気味である．



条件 1　条件 2.1

条件 2.2　条件 3

図 8　各条件の SSIM 値の推移



条件 1　条件 2.1

条件 2.2　条件 3

図 9　各条件の SSIM 値が最も大きいテスト結果

4. まとめ

本研究では，重なった音源からの分離を目的とした．そのために本論文では pix2pix を用いてスペクトログラムから一人の発話を抽出する手法を提案した．男女間では，SSIM 値とカラーマップからも抽出できていると判断できる．しかし，同性間では，黄色の箇所が残っており，抽出できているとは言いづらい．学習データが少なかったことが原因と考えられる．

今後の課題として，学習データに音声の前後ではなく途中に無音区間を考慮したデータを含めて再学習をし，抽出の精度を高めていく必要がある．また，学習に使用するネットワークに関しても，適宜パラメータを調整し，精度の向上を模索する必要がある．また，音声を聞いた限り大きな差はないように感じられた点は補足しておく．主観評価についても今後の課題である．

謝辞

この研究は，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」の補助を得ている．

参考文献

[Huang 15]Po-Sen Huang　Minje Kim Mark Hasegawa-Johnson Senior and Paris Smaragdis，Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation，IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 23, NO. 12, DECEMBER 2015

[Isola 17]Phillip Isola,Jun-Yan Zhu,Tinghui Zhou,Alexei A. Efros, Image-to-Image Translation with Conditional Adversarial Networks，In CVPR 2017.

[Isola 17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros，Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks，In ICCV 2017.

[Jansson 17]Andreas Jansson, Eric J. Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, Tillman Weyde, Singing Voice Separation with Deep U-Net Convolutional Networks, Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR), 2017.

[粕谷　68] 粕谷 英樹, 鈴木 久喜, 城戸 健一," 年令, 性別による日本語 5 母音のピッチ周波数とホルマント周波数の変化," 日本音響学会誌，24(6), pp.355-364, 1968.

[森勢　18]森勢 将雅，“音声分析合成，一般社団法人　日本音響学会，pp47，コロナ社，東京，2018 年

[寺澤　84] 寺澤 るり子 垣田 有紀　平野 実，“平均呼気流率,声の基本周波数および声の強さの同時測定―正常成人男女各 30 名の成績―,” 音声言語医学, 25(3), pp189-207, 1984.