最適輸送によるデータ拡張と変分自己符号化器を用いた音声データの生成

2018 年度修了
東京大学大学院　新領域創成科学研究科　複雑理工学専攻
指導教員：佐藤　一誠　講師　学籍番号：47176097　氏名：鵜飼　翔馬
キーワード：最適輸送，データ拡張，変分自己符号化器，音声データ生成，音声補間

1. 序論

1.1 背景

音声認識や音声合成，話者認識，音声符号化など，我々の生活の中には音声データの処理が必要な問題が数多く存在している．近年では，機械学習 や深層学習を用いた手法により，これらの分野は目覚ましい発展を遂げているが，音声データの複雑さにより，解決されていない問題も数多く存在する．例として，音声データの補間に関する問題がある．ある音声とある音声の中間的な音声を生成するなど，複雑な音声の補間処理を行う場合，単純なパラメータの調整によって音声を生成し，補間を行うことは非常に難しい．そこで本論文では，音声データを滑らかに補間するように，音声データを生成することを目的とし，このような問題を解決する手法を提案する．

1.2 計算機における音声データの扱い

計算機上で音声データを扱う場合，音声データをスペクトログラムというデータ形式に変換してから処理を行うことが一般的である．スペクトログラムとは，音声などの信号を時間，周波数，各周波数成分の強さの 3 次元グラフで表したものである．スペクトログラムでは．スペクトログラムは，短時間フーリエ変換 (short-time Fourier transform, STFT) を計算することで作成される．STFT とは，音声データなどの，通常のフーリエ変換を施すことができない非周期関数に対して，窓関数をずらしながら掛けて，それにフーリエ変換を施すことでスペクトルを計算する方法である．

1.3 問題設定

2 つの音声データを用意し，ある音声データからある音声データをまでの間に存在するであろう音声データを擬似的に生成することを考える．具体的には、音源 A と音源 B を用意し，それらの間を補間するような音声データを連続的に生成することを目指す．更に，図 1 の赤い矢印のように、音源 A から音源 B へ向かうベクトルとは逆方向に向かう方向に移動しても，音源が生成されるようなモデルを設計することを目指す．

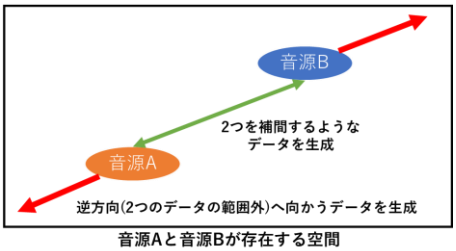

図 1. 本研究の問題設定

2. 変分自己符号化器

変分自己符号化器[1]とは，自己符号化器を拡張した生成モデルである．変分自己符号化器では，自己符号化器の中間層にある潜在変数 z に確率分布を仮定する．通常は z～N(0,1)と仮定する．更に，正規分布に従う乱数を学習時に取り入れることで，似たデータのものを近くに集めることができる．

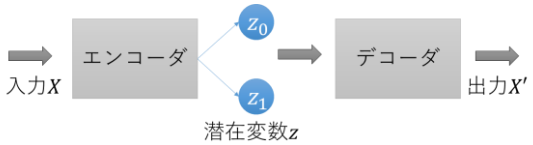

図 2. 変分自己符号化器の概略図

また，エンコーダでは中間層にある潜在変数 z に入力 X の特徴を押し込めていると解釈できる．ここで，z は比較的小さな次元に設計する．そして，デコーダでその潜在変数 z から元の入力を再現させるように X'を出力する．こうすることで，エンコーダでは入力 X が潜在変数 z として潜在空間に押し込められ、デコーダでは潜在空間内の任意の潜在変数からデータを生成することができる．図 2 に変分自己符号化器の概略図を示す．この図において，入力 X と出力 X'を比較し，誤差が最小になるようにエンコーダとデコーダそれぞれの重みを学習する．

3. 提案手法

3.1 最適輸送理論

最適輸送理論[2]とは，物質をある場所からほかの場所へ最小コストで移す理論である．画像などの幾何学的な領域においては，滑らかに補間データの生成を行えることが知られている．図 3 に，最適輸送を用いて 3 つの 3D データを滑らかに補間したときの図を示す．

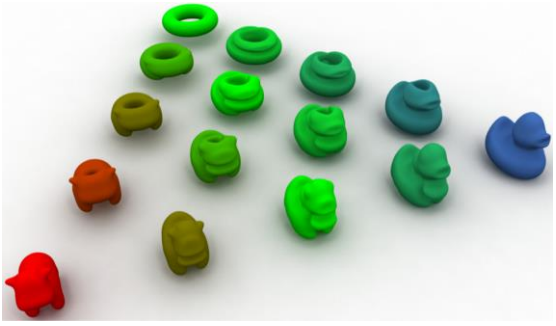

図 3. 最適輸送により補間された 3D データ
出典 ：[3] p. 7

3.2 従来手法の問題点

従来手法である変分自己符号化器のみを用いた音声データの補間の問題点として，潜在空間上に 2 つの音声データを滑らかに接続するような構造が無いことが挙げられる．変分自己符号化器のみでは，似ているデータが同じクラスとして潜在空間上で近い位置にプロットされるだけで，2 つのクラス間を滑らかに接続することは出来ない．また，変分自己符号化器を学習するためには大量の学習データが必要になるため，学習データを集めるために多くのコストが掛かるという問題点がある．

3.3 提案手法

まず，2 つ以上の音声データを用意し，それぞれの音声データのスペクトログラムを計算する．2 つのスペクトログラムを畳み込みワッサースタイン距離を用いた最適輸送を行い，滑らかに補間データを生成する．

次に，最適輸送によって生成されたスペクトログラムデータを学習用データとして，変分自己符号化器を学習させる．このように，最適輸送によって 2 つのデー

タの中間のデータを生成した後，変分自己符号化器を学習することで，図 4 のように 2 つのデータの範囲外のデータでも生成を行うことができる．

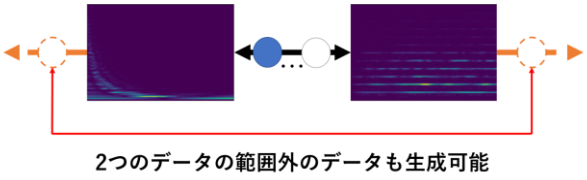

図 4．提案手法の概略図

4. 実験

4.1 データセットとパラメータの説明

実験には，サイン波を合成して人工的に生成した 3 クラスの音声データを使用する．3 クラスにはそれぞれ 4 種類の音声データを作成し，計 12 種類の音声データを用意した．そして，今回実験に使用した音声データの形式は，標本化周波数 16[kHz]，量子化ビット数 16[bit]，長さが 0.5[秒] (= 8,000 サンプル) の無圧縮の wav ファイルである．

次に，この 12 種類の音声データを全てスペクトログラムに変換する．そして，用意した 3 クラスのスペクトログラムの中から 2 クラスを選び，その中からクラスの異なる 2 つのデータを使用して最適輸送を行い，データを生成する．今回生成するスペクトログラムのデータは，輸送を行わないオリジナルのデータを含めて 32 個に設定した．これにより，1 つの組み合わせに対して，512 個のデータが生成される．そして，生成したスペクトログラムのデータ計 1,536 個を使用して，変分自己符号化器を学習させる．今回学習させた変分自己符号化器のパラメータは，入出力の次元は 15,876，中間層の数はエンコーダは 3 層で，次元はそれぞれ 9,500, 6,000, 2,500，デコーダは 1 層で次元は 9,500 に設定した．また，潜在変数の次元は 2，バッチサイズを 32，エポック数を 15 に設定した．活性化関数には，デコーダの出力層を除いて全て正規化線形関数を使用した．なお，デコーダの出力層にはシグモイド関数を使用している．

4.2 実験結果

はじめに,今回の提案手法との比較手法として，最適輸送を用いない単純な変分自己符号化器を学習させた．学習には提案手法と同様のデータを用意し，ノイズを加えてデータを増加させ，1 クラスあたり 3,780 個のデータを使用した．図 5 に，最適輸送を用いない単純な変分自己符号化器を学習させたときの潜在空間の図を示す．図 5 を見て分かる通り,単純な変分自己符号化器の場合，3 つのクラスが滑らかにつながっておらず,補間が全く出来ていないことが分かる．

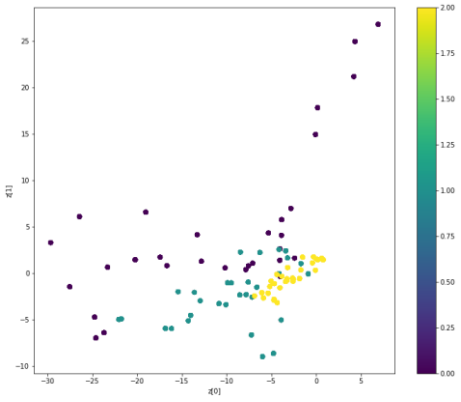

図 5．単純な変分自己符号化器の潜在空間

次に，提案手法の実験結果として，図 6 に最適輸送を行ってから学習した変分自己符号化器の潜在空間を示す．この図において，3 クラスの音源をそれぞれ A, B, C とすると，A と B で最適輸送を行ったデータ郡がラベル 0，A と C で最適輸送を行ったデータ郡がラベル 1，B と C で最適輸送を行ったデータ郡がラベル 2 を表している．図 6 より，それぞれの最適輸送により生成したそれぞれのクラスのデータが，変分自己符号化器のエンコーダの潜在空間上で滑らかに接続できていることが確認できる．

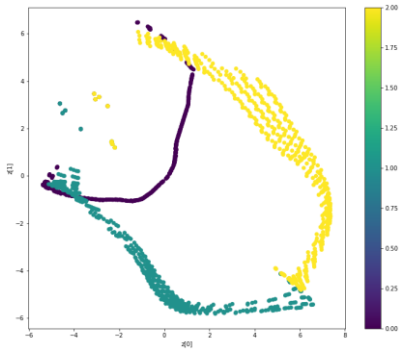

図 6．潜在変数から生成されたスペクトログラム

また，図 7 に潜在変数から生成されたスペクトログラムを並べた図を示す．図 7 を見ると，学習した変分自己符号化器の潜在変数から上手くスペクトログラムを生成できていることが分かる．また，図 6 で示したデータ間のつながりの通りスペクトログラムを生成できており，A と B, A と C, B と C の範囲外の点からも上手くスペクトログラムを生成できていることが確認できる．

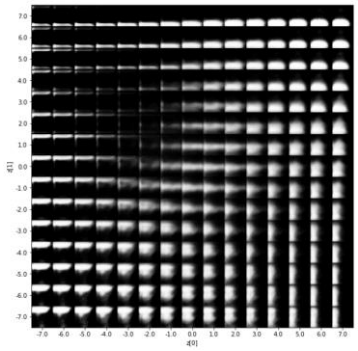

図 7．潜在変数から生成されたスペクトログラム

4.3 評価

実験結果を比較して分かるように，従来手法では潜在変数が滑らかにつながっていないが，提案手法では潜在変数が滑らかにつながっており，データの補間が行えていることが確認できる．学習データの数に関しても，最適輸送を用いない単純な変分自己符号化器の場合は,1 種類のデータに対して 3,780 個ものデータを使用して学習を行っているにもかかわらず，上手く学習が出来ていないが，提案手法の場合は 1 種類のデータに対して 4 個のデータのみを使用しても上手くスペクトログラム間の補間ができていることが分かる．以上の点から，今回提案した手法を用いた 2 つのスペクトログラムを補間するデータの生成は，従来の単純な生成モデルを使用する場合に比べて優れていると言える．

5. 結論

本論文では，まず音声データ補間において変分自己符号化器を単純に用いることの問題点を議論し、次に最適輸送を用いた新たな手法を提案した．最後に，提案手法を用いて音声のスペクトログラムデータの生成をする実験を行い，提案手法では潜在空間を上手く学習できることを示した．最適輸送を用いない単純な変分自己符号化器を用いた手法と比較して，より少ないデータ数で学習が上手くいくことや，スペクトログラムが滑らかに補間できることを示し，提案手法を用いた方が優れた結果であることを示した．

今後の課題として，実世界に存在するより複雑な音声データを用いて，音声データの補間を行うということが挙げられる．

参考文献

[1] Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, Max Welling, “Semi-supervised Learning with Deep Generative Models,” : Conference on Neural Information Processing Systems (NIPS), 2014.

[2] G. Monge, “Mémoire sur la Théorie des Déblais et des Remblais.,” Histoire de l’Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année, pp. 666-704, 1781.

[3] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, Leonidas Guibas , “Convolutional wasserstein distances: efficient optimal transportation on geometric domains,” ACM Transactions on Graphics (TOG), Volume 34, Issue 4, pp. pp.66:1-66:11, 2015.