深層学習を利用したユーザの重視している音響特徴の学習

芝浦工業大学
データ工学研究室

上野裕太郎

発表の流れ

- 背景
- 先行研究
- 目的
- 提案手法
- 実験，考察
- まとめ，今後の展望

背景

欧米を中心としてSpotifyやLast.fm，Pandora等の音楽ストリーミングサービスがトレンドとなっていて，日本でも2015年に入ってからAWA，LINE MUSIC，Apple Music，Google Play Musicが一斉にリリースされ大きな話題となった

ユーザが多くの楽曲を簡単に聴けるようになり
推薦システムの重要度が上がった

図1　Spotify

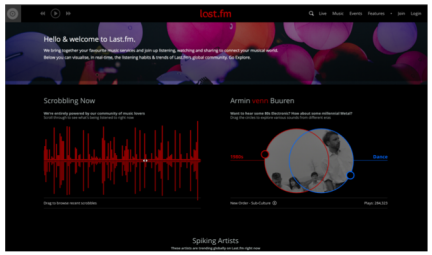
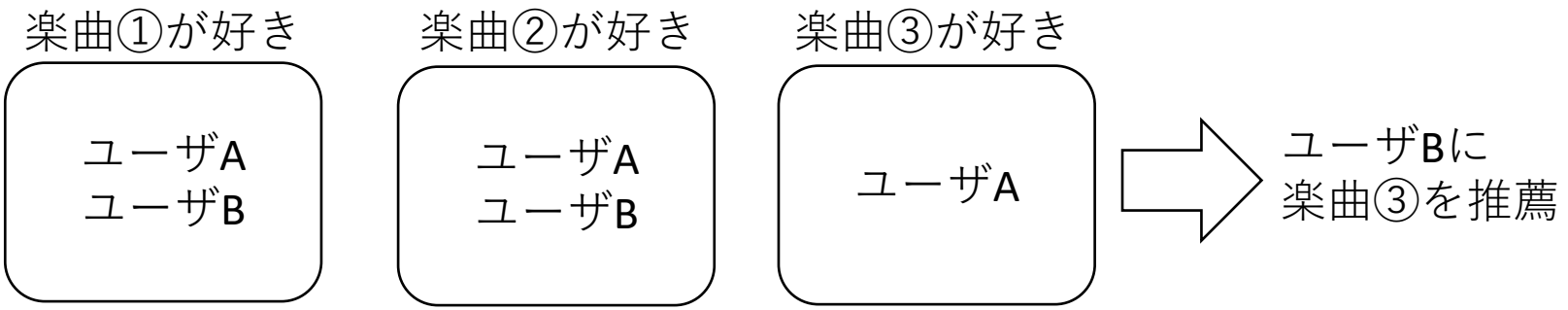


図2　Last.fm

先行研究1
協調フィルタリング

あるユーザに対し好みの似た他のユーザが好む楽曲を推薦する手法



- 楽曲に蓄積されているユーザ履歴を用いて推薦を行うためユーザ履歴の蓄積されていない新しい曲などの推薦に有効でない

[1] Badrul S. , et al. , “Item-based collaborative filtering recommendation algorithms”, Proceedings of the 10th international conference on World Wide Web. ACM, pp. 285-295, 2001.

先行研究2
Deep content-based music recommendation

Convolutional neural network(CNN)で音響特徴を学習させる手法

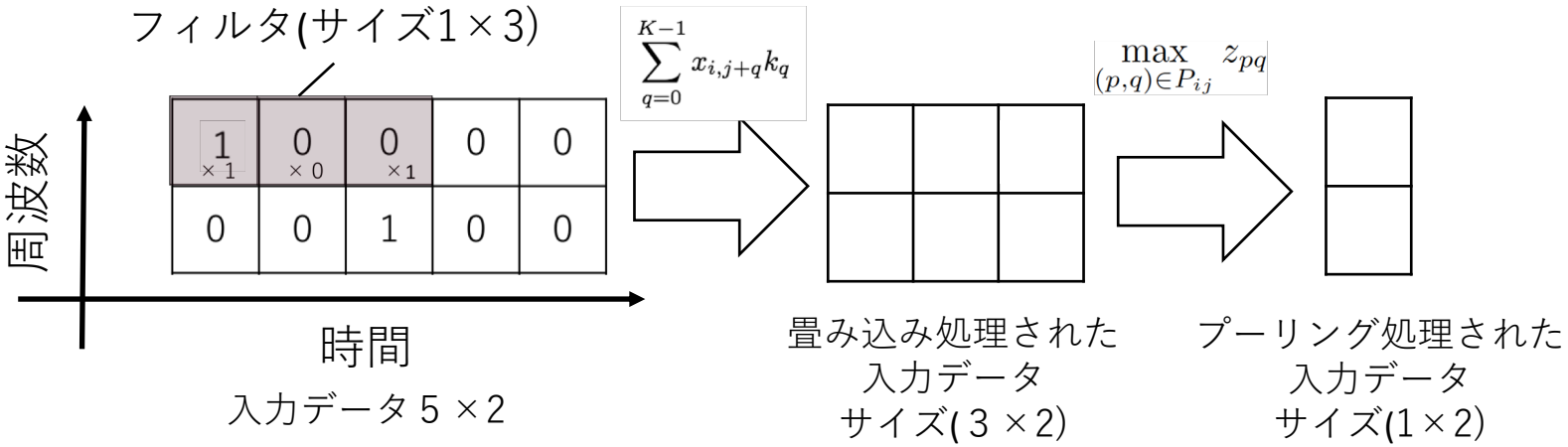
- 1つの楽曲全体をメル周波数ケプストラム係数に変換し，ユーザがその楽曲を聴いたかどうかを教師信号として音響特徴をCNNに学習させた
- 1つの楽曲全体に対してのみCNNで楽曲の特徴量を抽出することができている

[2] Aaron O. , et al. “Deep content-based music recommendation”, Advances in neural information processing systems,vol. pp. 2643-2651, 2013.

Convolutional Neural Network (CNN)

画像の小領域の情報を畳み込んで作成する畳み込み層とプーリング層からなるNeural Network

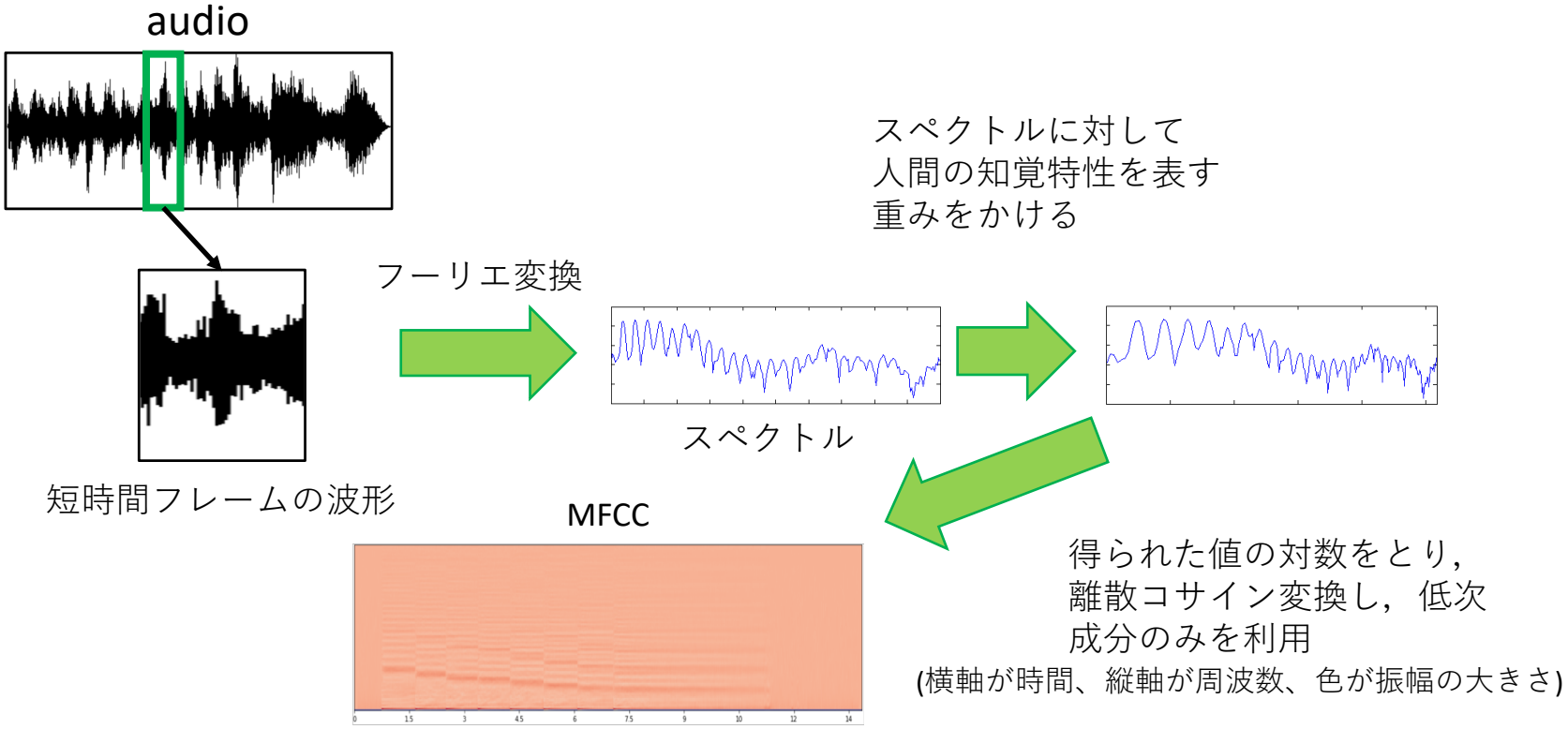
- フィルタと呼ばれる画像の小領域を1つの特徴量として自動的に特徴量を学習することができる
- 楽曲の音響波形を画像にすることができることから音響波形の特徴をCNNで学習し楽曲の特徴抽出に用いる



[3] K. Fukushima, ¥Pattern Recognition and the Neocognitron," The Journal of The Institute of Image Information and Television Engineers, vol.38, no.3, pp.212-218, 1984.

メル周波数ケプストラム係数 (MFCC)

人間の音高の知覚的尺度であるメル周波数から得られる量



[4] B.D. Steven, Comparison of parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," Proc. IEEE, no.4, pp.357{366, Aug. 1980.

先行研究2
Deep content-based music recommendation

Convolutional neural network(CNN)で音響特徴を学習させる手法

- 1つの楽曲全体をメル周波数ケプストラム係数に変換し，ユーザがその楽曲を聴いたかどうかを教師信号として音響特徴をCNNに学習させた
- 1つの楽曲全体に対してのみCNNで楽曲の特徴量を抽出することができている

[2] Aaron O. , et al. “Deep content-based music recommendation”, Advances in neural information processing systems,vol. pp. 2643-2651, 2013.

先行研究2での問題点

- 好みが似たユーザかどうかをもとに楽曲を推薦するため自分の好みと必ずしも一致するとは限らない
- CNNのみで特徴抽出を行っているため楽曲内の音響特徴の時間変化などに関しては抽出することができていない
- 1つの楽曲全体に対してCNNを学習させているため楽曲内の一部分だけ好きな場合などが考慮されていない
- フィルタが音響特徴を学習しているかをモデルの出力結果から考察しているためフィルタが音響特徴を学習しているかどうかを確認していない

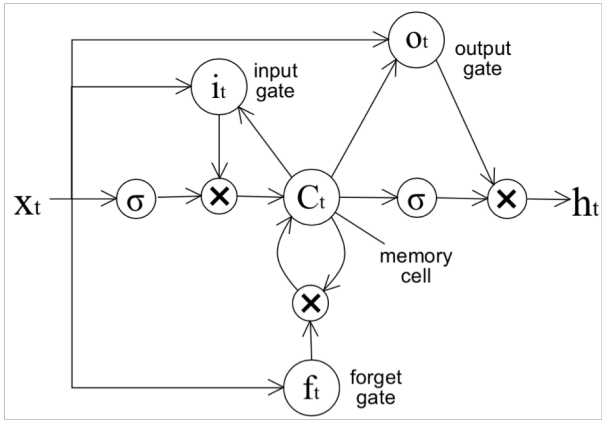
目的

ユーザ個人が重視している楽曲の一部分の特徴,
楽曲内の時間変化する特徴を考慮した
音楽推薦を行う

Long short-term memory (LSTM)

時系列データを扱うために隠れ層の値を再び隠れ層に入力でき，長期的な依存を学習するために過去の特徴量を記憶できる記憶セルを持つNeural network

- 本研究では人の知覚する音響特徴は時系列データであると考えLSTMによって楽曲に含まれる時間変化する音響特徴を学習させる
- CNNにより獲得した音響特徴に加え時間変化する音響特徴をLSTMによって学習させる



LSTMの概略図

[5] Hochreiter, S., Jurgen S. , “Long short-term memory“, Neural computation 9.8, vol.9, no.8, pp.1735-1780, 1997.

提案手法：学習方法

① audio形式の楽曲をMFCCに変換し入力データを作成

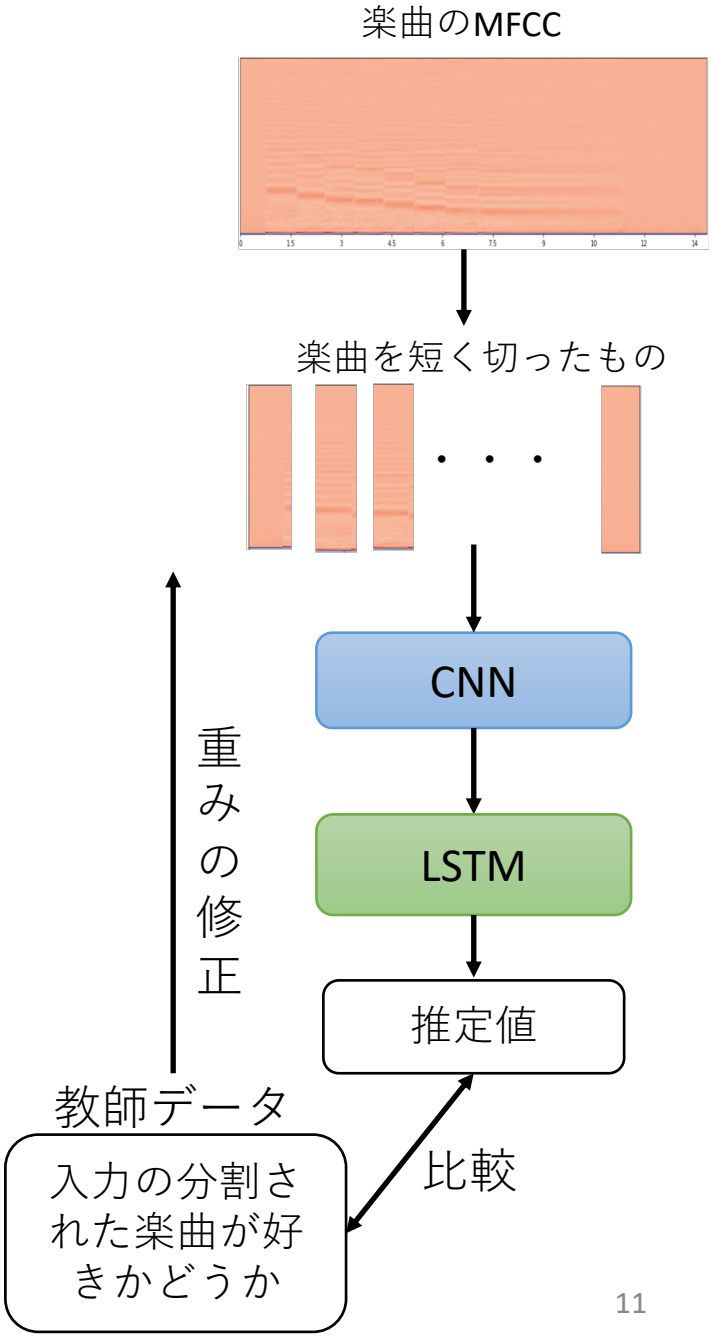
② 変換した楽曲を8秒ごとに短く切る

③ CNNで時間領域のみを畳み込み楽曲の特徴量を抽出

④ CNNによって抽出した特徴量をLSTMに入力し、LSTMによって好みの程度を出力

⑤ 出力値と教師データを比較してCNN・LSTMの重みを学習

上記を繰り返す



提案手法：推薦までの流れ

① audio形式の楽曲をMFCCに変換し入力データを作成

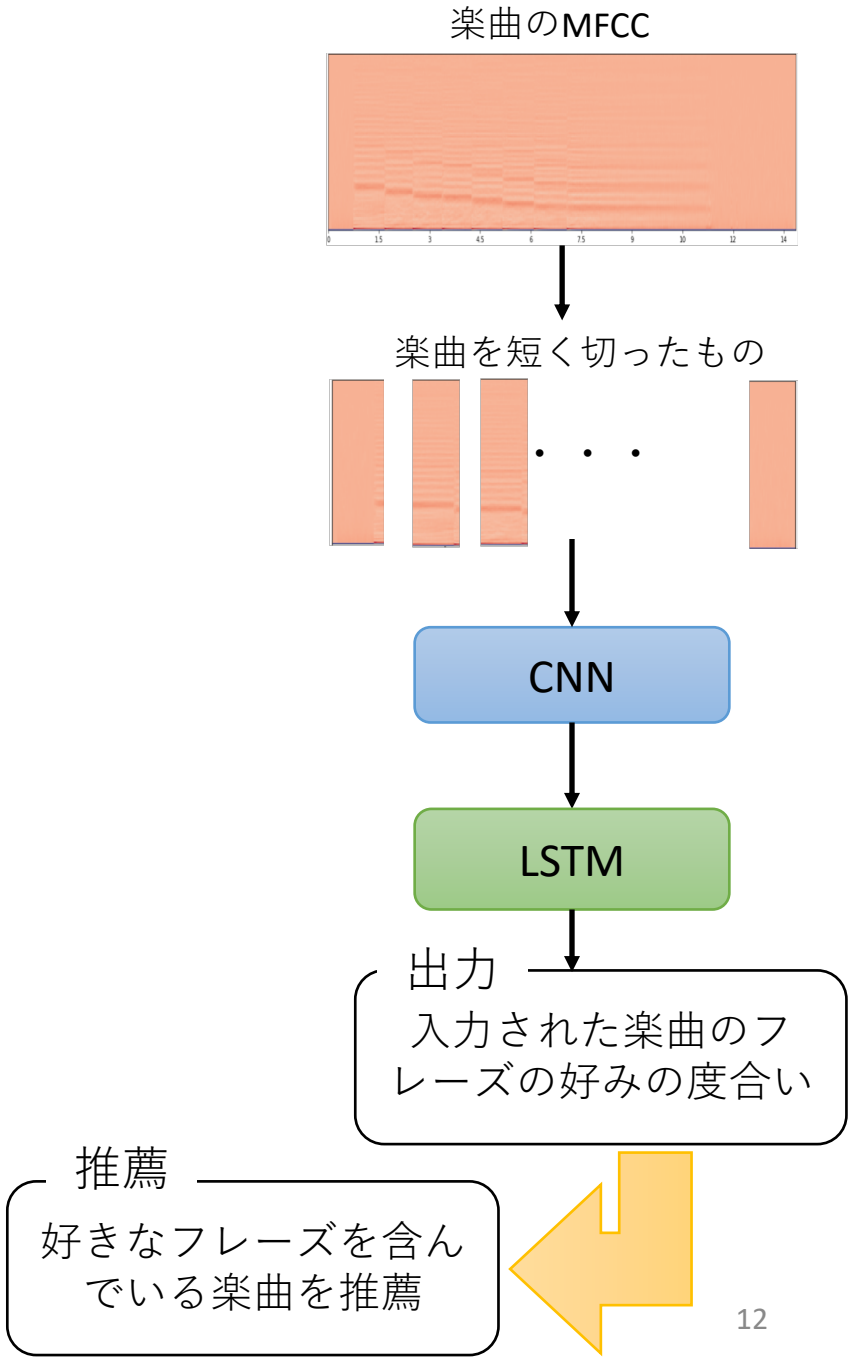
② 8秒の間隔の窓を1秒ずつずらし変換した楽曲を短く切る

③ CNNに②を入力

④ CNNの出力値をLSTMに入力

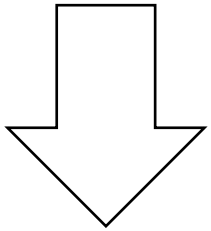
⑤ 入力された8秒ごとの楽曲にどれくらい好きなのかを出力

⑥ 8秒ごとの出力を足し合わせ，楽曲の時間で割った値を比較し一番値の高い楽曲を推薦



予備実験

実際に楽曲をMFCCに変換したデータでモデルを学習し，CNNで音響特徴が学習できるのかを確認したところ，CNNでどのような音響特徴が抽出できているの出力が複雑であり確認が困難であった.



楽曲の代わりに単純なデータで特徴が抽出できるかどうかを段階的に確認する必要がある

提案手法：音響特徴学習の確認

- どのような音響特徴を学習しているのか確認するために単純なデータから音響特徴を学習できるのかを確認する必要がある

- MFCCには変換せずaudio形式のデータに対して短時間フーリエ変換を行いベクトルを求め，その各成分の絶対値の2乗を入力データとし，以下の順にデータを複雑化して実験を行った

- 1~4のデータに対してデータに対して実験1を行い，5~7に対して実験2を行った．データ3に対しての実験1，データ7に対する実験2を示す

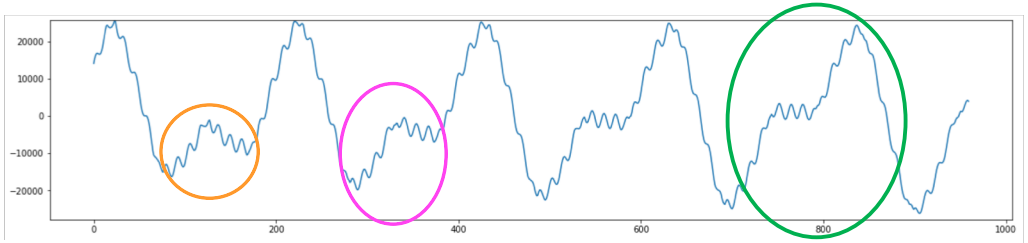
1. 単音のとき(周波数が1つのみ)
2. 和音のとき(周波数が1つ以上)
3. 1つの楽器の場合
4. 複数の楽器の場合
5. 単音で時間で音階が変化する場合
6. 和音で時間で音階が変化する場合
7. 1つの楽器で時間変化する場合
8. 楽曲の場合

実験1

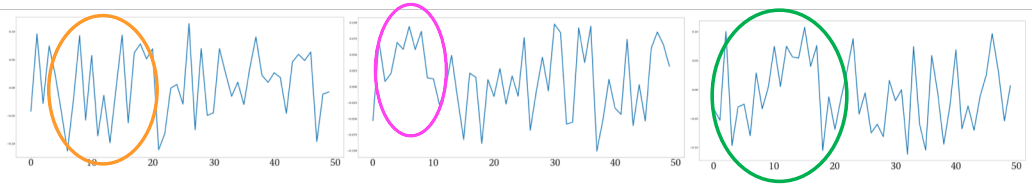
音階の変化のないデータに対してCNNが音響特徴を学習できるのか確認するため入力に対する畳み込み層の出力を可視化した

- 訓練データ
  音の高さの違うギターとフルートの2種類の楽器をそれぞれ10個用意し，フルートに対して1，ギターに対して0を教師信号として与えた

実験1　結果・考察



教師信号 1 を与えた学習データの audio 波形．横軸が時間，縦軸が振幅．

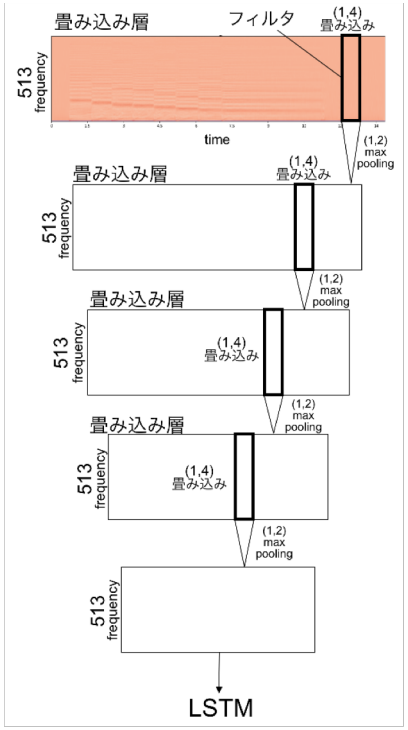


(a) 入力層から数えて 1 つ目の畳み込み層のうちの 1 つのフィルタの値の折れ線グラフ．

(b) 入力層から数えて 2 つ目の畳み込み層のうちの 1 つのフィルタの値の折れ線グラフ．

(c) 入力層から数えて 3 つ目の畳み込み層のうちの 1 つのフィルタの値の折れ線グラフ．

学習後のモデルのフィルタの値．横軸がフィルタの時間，縦軸がその時点における フィルタの値．



- フルートの音の波形と似た形を部分的にもつフィルタが確認できた
- 畳み込み層が深くなるに連れてaudio波形の概形を捉えているように見える

実験2

LSTMが音階の変化などの特徴を学習できているのかを確認するためCNNのみのモデルと提案手法での予測値の違いを比較した

- 訓練データ
  楽器Aが流れたあとに楽器Bが流れる時間変化のある楽曲に教師データとして1を与え，それ以外の楽曲に対して教師信号として0を与えた

- テストデータ
  楽器Bが流れたあとに楽器Aが流れる楽曲，楽器Aが流れたあとの時間間隔を狭め楽器Bが流れる楽曲を利用し，狭める時間感覚は2frame，6frame，10frame，50frameの順に狭めた

実験2　結果・考察

入力データ	CNN	提案手法
音A→音B	0.999	0.999
学習時教師信号0を与えたデータ	0.000	0.000
音A→音B 2frame	0.910	0.999
音A→音B 6frame	0.323	0.998
音A→音B 10frame	0.119	0.993
音A→音B 50frame	0.000	0.014
音B→音A	0.000	0.002

- 50frameが1秒
- それぞれの値は予測に対して答えがどのくらいあってたかを表す正解率の値

CNNでは数frameずれただけで値が大きく下がったが，提案手法では学習させた時間間隔に近くなればなるほど値が大きくなっている
LSTMを加えることで10frameまでのずれは許容できている

実験　考察

楽器の演奏間隔を変えても学習した音楽の共通した特徴をひろえている結果が得られた．このことからCNNとLSTMを組み合わせた方法により時間間隔が変化しても楽器固有の波形を学習することができると考えられる

まとめ・今後の展望

- 音響特徴が学習できているのかを確認するための方法として単純な楽曲データから複雑化して音響特徴が学習できているのかの実験を行った
- CNNで楽器固有の波形を学習することができた
- LSTMを加えることで演奏の時間間隔が異なっても楽器固有の波形を学習することができると考えられる
- 実際に複数の楽器を含む楽曲データや音楽などのより複雑なデータに対して構築したモデルが音響特徴を抽出できるかどうかの実験，実際に推薦システムの構築が必要である

参考文献

[1] Badrul S. , et al. , “Item-based collaborative filtering recommendation algorithms”, Proceedings of the 10th international conference on World Wide Web. ACM, pp. 285-295, 2001.

[2] Aaron O. , et al. “Deep content-based music recommendation”, Advances in neural information processing systems,vol. pp. 2643-2651, 2013.

[3] K. Fukushima, ¥Pattern Recognition and the Neocognitron," The Journal of The Institute of Image Information and Television Engineers, vol.38, no.3, pp.212-218, 1984.

[4] B.D. Steven, Comparison of parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," Proc. IEEE, no.4, pp.357{366, Aug. 1980.

[5] Hochreiter, S., Jurgen S. , “Long short-term memory“, Neural computation 9.8, vol.9, no.8, pp.1735-1780, 1997.

[6] G. Cybenkot. , “Approximation by Superpositions of a Sigmoidal Function”, Mathematics of Control, Signals, and Systems (MCSS) 2.4, vol.2, no. 4, pp. 303-314, 1989.

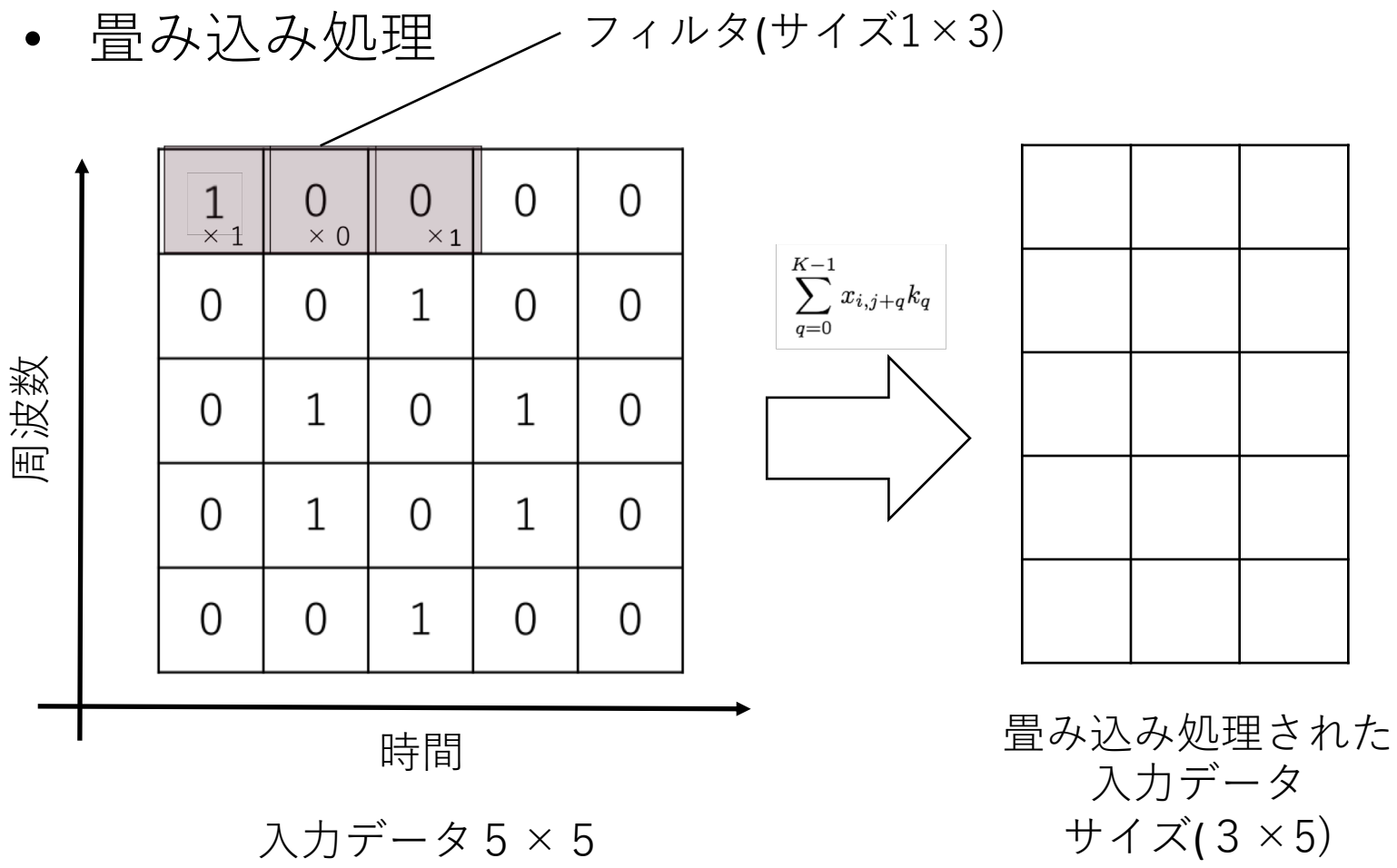
[7] Logan B. , “Music Recommendation from Song Sets”, The International Society for Music Information Retrieval (ISMIR), pp. 425-428, 2004.

ご清聴ありがとうございました

予備スライド

Convolutional Neural Network (CNN)

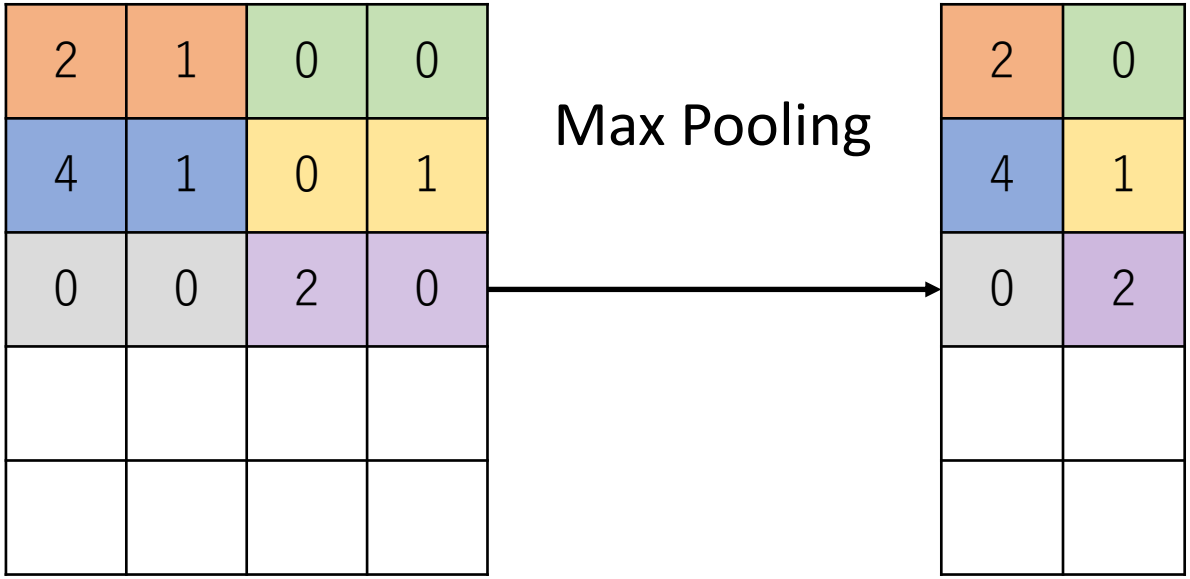
- 畳み込み処理



- 楽曲の音高情報を保つために楽曲の周波数領域の畳み込みはせず，時間領域のみ畳み込み処理を行う

Convolutional Neural Network (CNN)

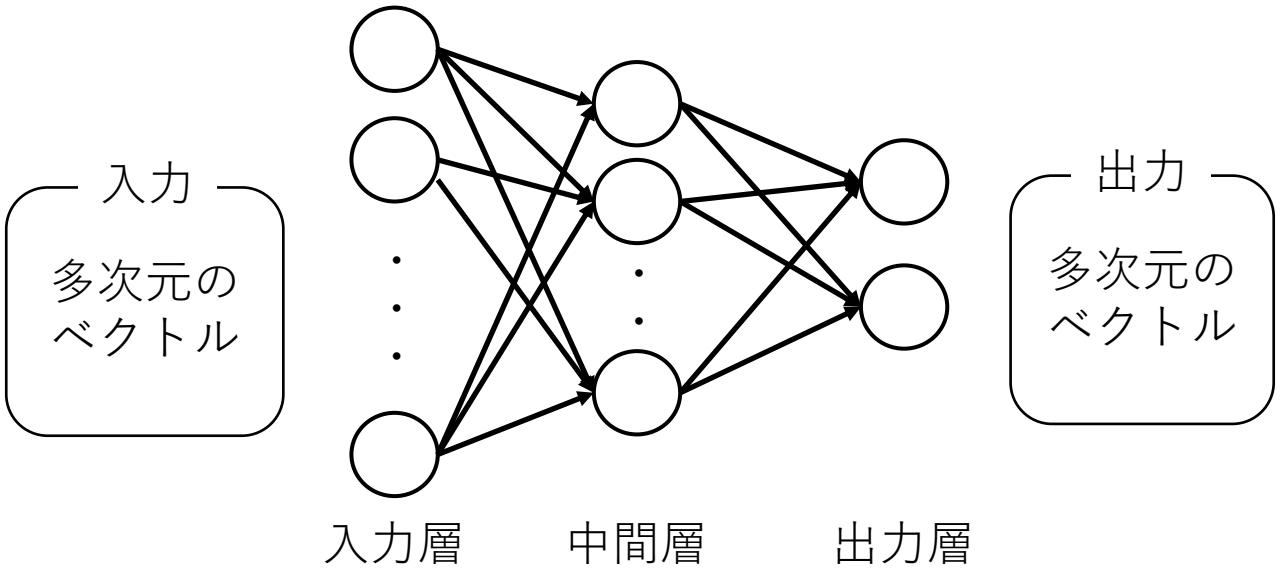
- プーリング処理



- 隣接したデータのうち値の大きい方を次の層へ渡す
- 畳み込み処理と同様，楽曲の音高情報を保つために楽曲の周波数領域のプーリングはせず，時間領域のみプーリング処理を行う

Feedforward Neural Network (FFNN)

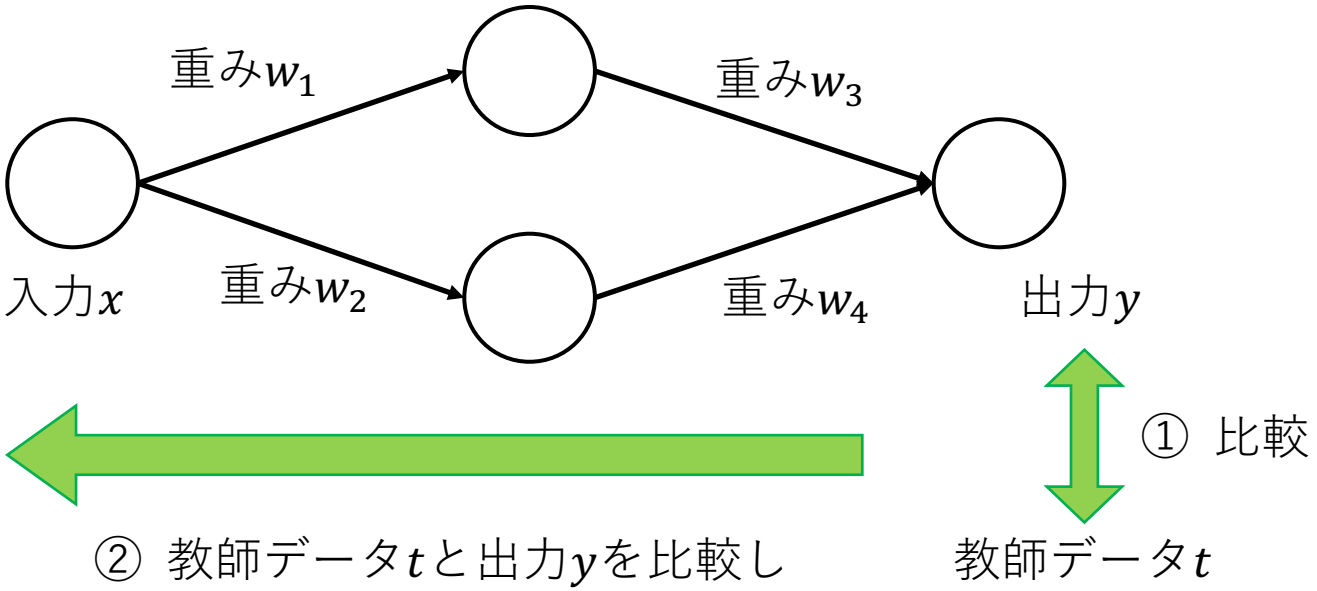
入力層、中間層、出力層からなる単一方向へのみ信号が伝播するニューラルネットワーク



[6] G. Cybenkot. , “Approximation by Superpositions of a Sigmoidal Function”, Mathematics of Control, Signals, and Systems (MCSS) 2.4, vol.2, no. 4, pp. 303-314, 1989.

逆誤差伝播法

教師データと出力を比較して重みなどを修正していく学習方法



内容に基づくフィルタリング

１つの楽曲に対し音響的な特徴量を計算し類似した楽曲を推薦する手法

問題点

- １つの楽曲に対して類似した音響情報から推薦するため複数の楽曲に対しての推薦を行うことができない
- 音の似た楽曲ばかりでジャンルの異なる曲など推薦の多様性に欠ける

[7] Logan B. , “Music Recommendation from Song Sets”, The International Society for Music Information Retrieval (ISMIR), pp. 425-428, 2004.