G-011

オートエンコーダを用いた視覚と聴覚の統合学習によるマガーク効果の再現
Reproduction of Mcgurk Effect through Integrated Visual and Auditory Learning using Autoencoder

中村 健† 野口 渉†† 飯塚 博幸††,††† 山本 雅人††,†††
Ken Nakamura Wataru Noguchi Hiroyuki Iizuka Masahito Yamamoto

1. 序論

ヒトは物事を知覚する際に，複数の感覚を統合して知覚している．その一例として挙げられるのがマガーク効果である．これは言語音声の音韻知覚において視覚情報と聴覚情報の相互作用を示す現象の一つであり，Mcgurk らによって初めて示された[1].

また近年，深層学習は様々な分野において高い成果をあげており，人間と同じレベルの感覚入力の認識・生成が可能になりつつある．そこで，深層学習を用いて複数感覚の統合をモデル化することで人間の知能を理解しようとする研究が行われている．特に Ngiam らはマルチモーダルオートエンコーダを用いて感覚情報を統合して学習し，マガーク効果を再現可能なモデルを提案している[2]．錯覚を再現するためには結果だけでなくその処理過程まで同じである必要があるため，このモデルは錯覚を再現可能であるという点で人の知覚モデルに近いと考えられる．

しかし，先行研究においては人間の受容する高次元の情報から次元を削減することで過度な単純化が行われており，ヒトが受容する情報とはかけ離れている．本研究ではヒトが受け取るのと質的に同等な視覚情報と聴覚情報を統合して学習し，その相互作用が起こる例としてマガーク効果を再現できるモデルを構築する．

2. 関連研究

2.1 マガーク効果

マガーク効果[1]とは，人間が言語音声を知覚する際に聴覚情報と視覚情報を統合して認知していることを示す現象である．Mcgurk らは、ある音韻の発話の映像と別の音韻の音声を組み合わせて視聴すると発話の映像とも音韻の音声とも異なる第 3 の音韻が知覚されると報告した．例えば，発話の映像が「が」で音声が「ば」であったとき，「が」でも「ば」でもなく「だ」と聞こえる．これは人間が音韻を認識する際に音声情報だけでなく視覚情報など他の感覚情報からも影響を受けることを示しており，聴覚視覚情報統合の代表例となっている．

2.2 マルチモーダルオートエンコーダ

オートエンコーダは入力を再構築するように学習されるニューラルネットワークである．学習目標は入力と出力の誤差の最小化であるため，入力以外の教師信号が必要ないという点で教師なし学習に分類される．オートエンコーダにおいて隠れ層では入力を再構成するために入力情報を保持していなければならないため，抽象化した情報を保持していると考えることができる．

ここで，オートエンコーダで得られる抽象化能力を複数のモダリティ（画像や音声，文章など）を統合するために用いるとき，マルチモーダルオートエンコーダと呼ぶ[2]．マルチモーダルオートエンコーダは複数のモダリティを入力として受け取り，それぞれから特徴量を抽出した後にそれらの特徴量を統合する．そして再度モダリティの数だけ枝分かれしてそれぞれのモダリティを再構成する．隠れ層は次元が小さいため，二つのモダリティの情報を別々のニューロン群で排他的に処理すると再構成に十分な情報がエンコードできない．したがって再構成するために，同じニューロン群を用いて二つのモダリティの情報が統合されてエンコードされる． 特に各モダリティそれぞれを深いニューラルネットワークによって抽象化したのちに統合することで，画像や音声といった大きく異なる入力特徴をもつモダリティの統合学習が可能になる．マルチモーダルオートエンコーダは隠れ層において複数のモダリティを統合して学習しているため，人間が音声を聞かずに口の形から音声を推測することができるのと同じように，入力から欠けたモダリティが存在した場合でも欠けたモダリティを補って出力することができる．

Ngiam らはこのマルチモーダルオートエンコーダを用いてマガーク効果が再現できることを示した[2]．この研究では音声と動画から統合された特徴量を学習し，その統合された特徴量を SVM で分類する学習を行うことで，音声が「ば」，視覚情報が「が」であるときにモデルの予測が「だ」に分類されるというものである．その結果を表 1 に示す．

また Noguchi らはこのマルチモーダルオートエンコーダに LSTM を用いることで，時系列データを扱うことを可能にし，運動する際の関節情報と視覚情報の統合が可能であることを示した[3].

表 1　マガーク効果の再現

Visual/ Audio Setting	Model Prediction		
	/ga/	/ba/	/da/
/ga/, /ga/	82.6%	2.2%	15.2%
/ba/, /ba/	4.4%	89.1%	6.5%
/ga/, /ba/	28.3%	13.0%	58.7%

3. 視覚・聴覚を統合する認知モデル

3.1 モデルの満たすべき要件

本研究では，時間的に変化する信号である動画と音声を統合して学習し，聴覚情報と視覚情報が相互に影響を与えるプロセスをシミュレーションするモデルを構築する．そ

†北海道大学 情報科学院 Graduate School of Information Science and Technology, Hokkaido University
††北海道大学 人間知・脳・AI 研究センター Center for Human Nature, Artificial Intelligence, and Neuroscience, Hokkaido University
†††北海道大学 情報科学研究院 Faculty of Information Science and Technology, Hokkaido University

こで前節で述べたマルチモーダルオートエンコーダを用いる．さらに，人間が受け取るのと質的に同等と考えられる情報を扱うためには以下の 2 つの条件を満たす必要がある．

・時間情報を扱うことが可能
・高次元の感覚情報を認識可能

3.1.1 時間情報を扱うことが可能なモデル

聴覚情報も視覚情報も時間的に変化するため，時間変化から文脈を見出すことのできるモデルである必要があり，そのためには時系列から文脈のある特徴量を抽出する必要がある．

3.1.2 高次元の感覚情報を認識可能なモデル

人間は高次元な視覚入力から物体や物体の動きを知覚することが可能である．Ngiam らは前処理として次元削減された視覚情報を用いていたが，本研究では，高次元な視覚情報をそのまま入力として受け取る，という点で，人間が受け取るのと質的に同等な感覚情報の認識が可能なモデルを構築する．

3.2 3D 畳み込みマルチモーダルオートエンコーダ

前節で述べた条件をすべて満たすモデルとして 3D 畳み込みマルチモーダルオートエンコーダを提案する．高次元の感覚情報を認識可能なモデルとして 3D Convolutional Neural Network (3DCNN) [4]を用いる．3DCNN は時間方向と空間方向 両方の情報を畳み込むことによって時系列データから特徴量を抽出することができ，高次元な情報から特徴量を抽出することが可能である．そのモデルの概要を図 1 に示す． 提案モデルが入力を受取り，再構築する過程は大きく以下の 3 段階で表すことができる．

1. 各感覚についてエンコーダが時系列入力を受取り，時系列方向のサイズを保持しながら特徴量を抽出する．
2. 音声のエンコーダから出力される特徴量を成形して三次元に変換し，動画の特徴量とチャネル方向に結合した後に 3D 畳み込みで各感覚の特徴量の統合を行う．
3. 統合された特徴量から各感覚についてのデコーダが感覚情報を再構築する．

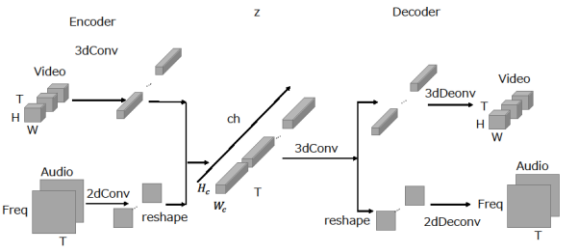

図 1　モデルの概要

4. 実験

4.1 データセット

本研究ではモデルに学習させる視覚・聴覚データセットとして，ヒトの発話を撮影したものを用いる．撮影は定点カメラによって行われ，撮影対象の人物は 1 人である．動画の内容は五十音の「や行」と「わ行」を除く 40 音を 10 回発音した 400 データに加え，マガーク効果が再現できているかを確認するための「が，だ，ば」をそれぞれ 8 音ずつ録音した計 424 データである．動画のフレームレートは 30fps であり，各フレームは，サイズが 1080 x 1920 の RGB 画像である．また，背景が学習に影響を与えないようにするために撮影はすべて白い壁の前で行った．音声のサンプリングレートは 44,100Hz である．

4.2 学習データ

前節で述べたデータセットの中から，五十音の「や行」と「わ行」を除く 40 音を 8 回と「が，だ，ば」それぞれ 8 音ずつ，計 344 データを学習データとして用いる．さらに，感覚の統合を促進するために，音声のみから画像，画像のみから音声を復元するように学習を行う．音声の入力が欠けた場合には発声している音声の平均を入力し，動画の入力が欠けた場合には発声している動画の平均を入力する．つまり，収集したデータ 1 つにつき，以下の 3 種類の（動画，音声）の組を入力とする学習データを作成し，学習を行う．

1. （発声している動画，発声している音声）
2. （発声している動画，発声している音声の平均）
3. （発声している動画の平均，発声している音声）

合計で学習するデータ数は 344 x 3=1032 データとなる．

4.3 実験設定

マルチモーダルオートエンコーダの動画の入力として，8 枚の連続した画像の各フレームから口の周りのみトリミングし 64 x 64 に縮小したものを用いる．また，音声の入力として音声をフーリエ変換し，最大値を 1，最小値を 0 とした正規化を行ったものを用いる．マルチモーダルオートエンコーダの構造を表 2，表 3 に示す．また，最適化アルゴリズムは Adam[5]，学習率は10^{-3}，バッチサイズは 4 とする．また活性化関数は f(x) = max (x, 0.2x)，で定義される Leaky ReLU[6]であり，LReLU と表す．BN は Batch Normalization[7]，Conv , Deconv はそれぞれ畳み込み層と逆畳み込み層を示す．

表 2　マルチモーダル AE の動画部分の構造

	HxWxTimexCh	Filter size	Activation Func
Input	64x64x8x3		
Conv3d x5	2x2x8x128	3x3x3	BN, LReLU
shared layer	2x2x8x256		
Conv3d	2x2x8x128	3x3x3	BN, LReLU
Deconv3dx4	32x32x8x128	3x3x3	BN, LReLU
Deconv3d	64x64x8x3	3x3x3	Sigmoid

表 3　マルチモーダル AE の音声部分の構造

	BinxTimexCh	Filter size	Activation Func
Input	128x8x2		
Conv2d x 5	4x8x128	3x3	BN, LReLU
shared layer	(2x2)x8x256		
Conv3d	(2x2)x8x128	3x3x3	BN, LReLU
Deconv2dx5	64x8x128	3x3	BN, LReLU
Deconv2d	128x8x2	3x3	Sigmoid

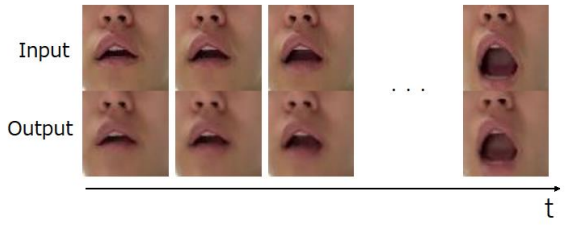

図２　動画の入力と出力

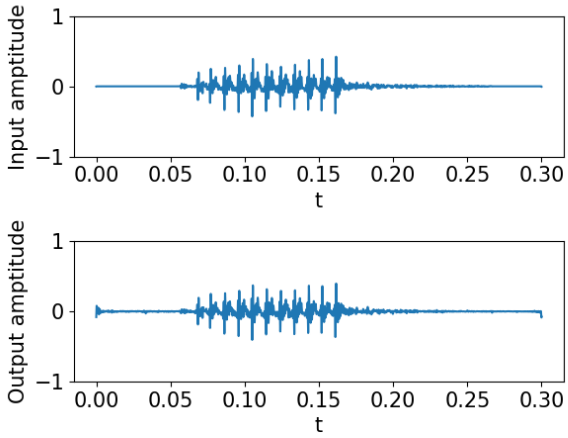

図３　音声の入力と出力

4.4 評価方法

本研究では聴覚情報と視覚情報を統合した特徴量を学習することを目的としている．そこで，ただ単に再構築された時系列と入力の類似度を比較するだけでは感覚情報の統合度合いを評価することはできない．そこで（動画，音声）の組が（が，ば）という音声と動画で整合性のない入力が与えられた場合のモデルの出力について，「が」「ば」「だ」それぞれの音声との類似度を評価する．つまり，入力動画の「が」と入力音声の「ば」の双方と一致しない「だ」に近い音声を出力することが確認できればマガーク効果を再現しているといえる．類似度の評価は音声を短時間フーリエ変換して最大値を 1，最小値を 0 とした正規化を行ったものの平均二乗誤差を用いる．

5. 結果

5.1 3D 畳み込みマルチモーダルオートエンコーダの再構成能力

3D 畳み込みマルチモーダルオートエンコーダが動画，音声をともに再構成できていることを確認する．「さ」と発音している音声と動画を入力としたときに提案モデルによって生成された画像の系列の一例を図２に，音声波形を図３に示す．この画像系列と音声は訓練データ内に含まれない．図２を見てみると，画像系列は口の形がわかる程度に再構成されていることがわかる．さらに，音声も図３に示すように似たような振幅を再現できており，実際に聞いてみることで母音だけでなく子音まではっきりと再構成されていることが確認できた．

5.2 マガーク効果の再現

5.2.1 各音声の距離による比較

次に聴覚情報と視覚情報の統合学習が行われていることを確認するために，マガーク効果が起きるかどうかを検証する．表４に（動画，音声）の組が（が，が），（ば，ば），（が，ば）として入力した場合に出力された音声と，「が，ば，だ」それぞれの音声の誤差を示す．このとき，それぞれの音声をフーリエ変換し，最大値を 1，最小値を 0 として正規化したものの周波数領域での二乗誤差を算出している．表４から，（動画，音声）の組が（が，が）や（ば，ば）であった時には「が」や「ば」という元の音声との誤差が小さい．しかし，（が，ば）を入力した際には「が」「ば」「だ」どの音声との誤差も小さくならなかった．このことより，口の形と音声の整合性がある場合には元の音声を復元できているが，整合性がない場合には「が」でも「ば」でもない音声を出力していることがわかる．

表４マガーク効果の再現の検証

Visual/ Audio Setting	Mean Squared Error		
	/ga/	/ba/	/da/
/ga/, /ga/	6.37×10^{-4}	4.33×10^{-3}	7.39×10^{-3}
/ba/, /ba/	3.57×10^{-3}	8.59×10^{-4}	6.52×10^{-3}
/ga/, /ba/	3.05×10^{-3}	1.82×10^{-3}	5.55×10^{-3}

5.2.2 出力音の分布

出力音がそれぞれの音声ごとにどのような特徴を持っているかを確認する．出力音は 128 x 8 x 2 の高次元ベクトルであるため，PCA によって二次元平面上に射影することで可視化を行う．その結果を図４に示す．図中の output は（動画，音声）の組が（が，ば）であるときにモデルの出力する音声である．この図より，（動画，音声）の組が（が，ば）であるときに，分布としては「が」「ば」の分布から「だ」の分布に近づいた音声を出力していることが確認できる．

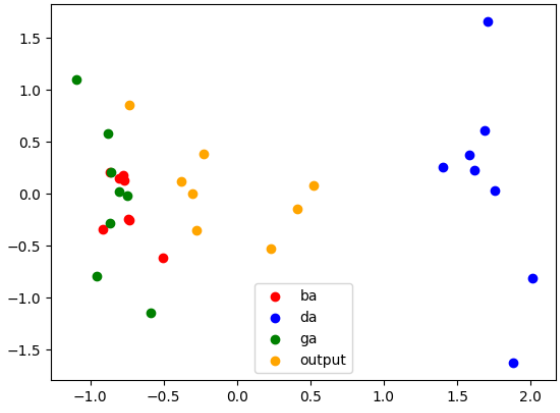

図４　音声を PCA によって二次元平面上に射影したときの分布

5.2.3 出力音のコサイン類似度

次に，モデルがどの程度人間と同じ錯覚を再現出来ているかを評価するためにモデルの出力音と，入力音である

「ば」という音声をコサイン類似度を用いて評価する．このとき，入力音「ば」から「だ」へ向かうベクトルと，入力音「ば」からモデルの出力音へ向かうベクトルのコサイン類似度は 0.997 となった．また，入力音「ば」から「が」へ向かうベクトルと，入力音「ば」とモデルの出力音へ向かうベクトルのコサイン類似度は 0.239 となった．以上のことから，入力動画に対応する音声である「が」よりも「だ」に近づく音声を出力しており，人間に起きるのと同じようにマガーク効果が再現できたことが確認できる．

6. おわりに

本研究ではマルチモーダルオートエンコーダに 3D 畳み込みを用いた 3D 畳み込みマルチモーダルオートエンコーを提案した．さらに，3D 畳み込みマルチモーダルオートエンコーダを用いて視覚情報と聴覚情報を統合した学習を行うことで人間の感覚統合によって起きるマガーク効果を再現できることを確認した．さらに今後の展望としてモデルの出力する音声がヒトにとって何の音声に聞こえるかを実験して確認することや，ほかの音声と動画の組み合わせで起きるマガーク効果について検証することが挙げられる．

謝辞

本研究の一部は北海道大学情報基盤センター人工知能対応先進的計算機システム共同研究の助成を受けたものである．

参考文献

[1] Harry Mcgurk and John Macdonald. Hearing lips and seeing voices. Nature, 264, pages746–748, 1976.
[2] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In ICML, pages 689–696, 2011.
[3] Wataru Noguchi, Hiroyuki Iizuka, and Masahito Yamamoto. Proposing multimodal integration model using lstm and autoencoder. volume 3, 12 2016.
[4] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition.
[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.
[6] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In in ICML Workshop on Deep Learning for Audio, Speech and Language Processing, 2013.
[7] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. CoRR, abs/1502.03167, 2015.