5S-3

Deep Neural Network を用いた マルチモーダル音声認識の為の特徴量学習

山口 雄紀 † 野田 邦昭 ‡ 中臺 一博 § 奥乃 博 † 尾形 哲也 ‡

† 京都大学 大学院情報学研究科 知能情報学専攻 ‡ 早稲田大学 基幹理工学部表現工学科

§HRI-JP (ホンダ・リサーチ・インスティチュート・ジャパン)

1. はじめに

近年，音声インターフェースが注目されているが，実用上の問題点として雑音の多い環境では認識率が著しく低下してしまう点がある．この問題に対するアプローチとして音声以外のモダリティ信号を利用するマルチモーダル音声認識が提案されており，特に唇画像と音声を組み合わせた視聴覚音声認識の研究が進められている．本研究では視聴覚音声認識の精度に重要な影響を及ぼす画像・音響特徴量の学習において特徴量学習の新たな手法として近年機械学習の分野で注目を集めている Deep Neural Network (DNN) を用いて唇画像および雑音音声から有用な特徴量を学習によって抽出することを目指す．

2. Deep Neural Network

DNN とは，深い階層を持つニューラルネットワークである．従来の誤差逆伝搬では，ネットワークの層が深くなるほど，内部パラメータの更新部分が拡散してしまう為に DNN の学習は困難であったが，近年 DNN を効果的に学習するためのアルゴリズムが提案されており，画像 [3] や音声 [5] などの高次元データの識別・特徴量学習において有効性が示されている．Hinton らは，DNN の各層を一層ずつ Restricted Boltzmann Machine を用いた教師無学習 (pre-training) によって準備し，それらの層を重ね合わせて誤差逆伝搬などで調整学習 (fine-tuning) を行う Greedy Layer-wise Training を提案した [1]．Martens は二次勾配に基づく学習方法として Hessian-free を提案している [2]．この手法では，pre-training を行わずに効率的に DNN を学習することができる．また，画像認識の分野で階層の深い畳み込みニューラルネット (Deep Convolutional Neural Network: DCNN) を利用することで高精度の認識を達成できることが報告されている [3]．本研究では画像特徴量の学習には DCNN を，音響特徴量の学習には雑音入り音声からクリーンな音声を出力する denoiseDNN を利用し，denoiseDNN の学習アルゴリズムとして Hessian-free を利用した．

3. 関連研究

DNN を用いて視聴覚音声認識を行っている研究として Ngiam らの研究がある [4]．Ngiam らは唇画像と音声スペクトログラムにそれぞれ主成分分析 (Principal Component Analysis: PCA) で圧縮したもの数フレーム分を入力として DNN を学習させることでマルチモーダルな特徴量の学習を行っている．アルファベット発声データを用いた実験で，マルチモーダル特徴量空間において線形識別器を用いて直接アルファベットの識別を行い，既存の手法と比べて同等以上の識別精度があることを示している．しかし，特徴量空間で直接単語識別しているために認識において時系列の繋がりが考慮されておらず，長い単語や連続音声への応用や単語数の増加についての検証は行われていない．本研究では，DNN により学習した特徴量を隠れマルコフモデル (Hidden Markov Model: HMM) への入力ベクトルとして認識を行うことでより大きな単語セットでの認識精度を検証する．

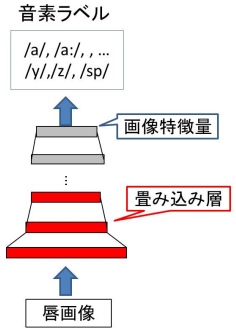

図 1: 画像特徴量の学習

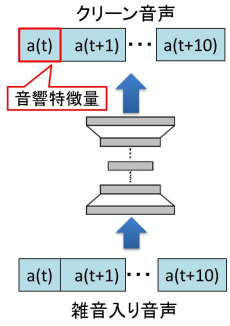

図 2: 音響特徴量の学習

4. DNN を用いた特徴量学習

4.1 畳み込みニューラルネットによる画像特徴量の学習

画像特徴量の学習には画像認識において高精度を達成している [3] で提案されている DCNN を利用した．図 1 のようにネットワークの入力として唇領域の画像を利用し，同フレームにおける音素ラベルを出力とするように学習を行った．音素ラベルは音声データで学習した HMM の強制アラインメントによって作成した 40 音素の時系列を用いている．予備実験においてすべての話者データを一括で学習すると，話者の違いに影響を受けるために認識率が低下していたため，本研究では話者ごとにネットワークを学習するものとした．

4.2 denoiseDNN による音響特徴量の学習

雑音に頑健な音響特徴量を学習するために図 2 に示すように DNN の入出力データとして Mel Frequency Cepstrum Coefficient (MFCC)12 次元とパワー，それらの一次微分および二次微分計 39 次元 11 フレーム分を 1 サンプルとし，雑音入り音声が入力された時にクリーンな音声を出力とするように学習を行った．認識の際には

Feature Learning for Audio-Visual Speech Recognition Using Deep Neural Networks: *Yuki Yamaguchi (Kyoto Univ.), Kuniaki Noda (Waseda Univ.), Kazuhiro Nakadai (HRI-JP), Hiroshi G. Okuno (Kyoto Univ.), and Tetsuya Ogata (Waseda Univ.)*

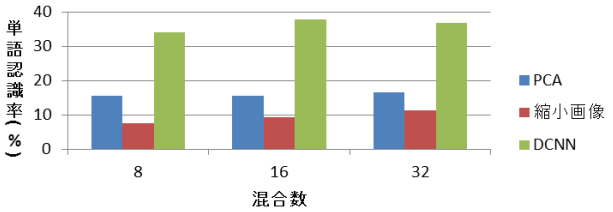

図 3: 各画像特徴量による単語認識率

出力された 11 フレーム分の MFCC のうち最初のフレームの値（39 次元）を音響特徴量として利用した．

5. 視聴覚統合音声認識

画像・音響特徴量を統合するためにマルチストリーム HMM を利用した．マルチストリーム HMM は，HMM への入力を複数のストリームに分け，各ストリームに重みを指定することで認識時に重みに応じて尤度を統合する手法である．本研究では，音声・画像間のストリーム重みを [0，0.2，0.4，0.6，0.8，1.0] と変化させて認識を行い，最も認識精度の高いものを認識結果として採用した．

6. 評価実験

DNN で学習した特徴量について検証するために獲得された特徴量を入力ベクトルとして，HMM を用いた孤立単語認識を行った．実験には男性 6 人，1 人当たり 300 単語 （ATR 音素バランス単語 216 単語と ATR 重要単語 84 単語）の発話を収録した視聴覚データセットを使用した．音声データはクリーンな環境で 16bit，16kHz で収録し，画像データはクリーンな環境で 8bit モノクロ，640 × 480 ピクセル，100Hz で収録した．画像データは，顔全体の画像となっているため手動で唇領域の抽出を行い，32 × 32 ピクセルに揃えて使用した．

6.1 画像特徴量の性能比較

画像特徴量の性能を検証するために孤立単語認識を行った．認識には混合数 8，16，32 のモノフォン HMM を利用し，学習セットとして ATR 音素バランス単語 216 単語分を利用し，残りの 84 単語を評価に用いた（話者クローズ単語オープン）．比較対象として，PCA に基づく 40 次元特徴量と単純に画像を縮小した 36 (6 × 6 ピクセル) 次元特徴量を用いた．図 3 にそれぞれの特徴量の単語正解率を示す．PCA や縮小画像に比べて DCNN で学習した特徴量が高い認識率を示していることが分かる．

6.2 視聴覚音声認識

denoiseDNN による音響特徴量および視聴覚統合の効果の検証を行った．学習セットは 216 単語分の音声・画像データであり，音声データについては denoiseDNN の学習にはクリーンデータと Signal-to-Noise Ratio (SNR) が 0，15，30 となるように白色雑音を付加した音声を利用した．HMM は混合数 16 のモノフォン HMM を使用し，HMM の学習にはいずれの特徴量についてもクリーンな音声のみを利用した．テストセットは 84 単語分の音声 (クリーンおよび SNR30〜SNR-20)・画像データであり，話者クローズ単語オープンの評価となっている．

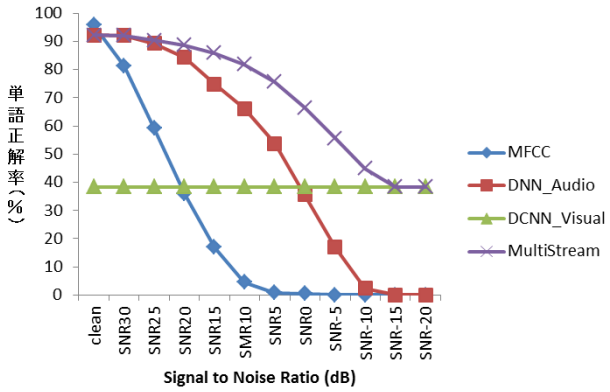

図 4: 各モダリティによる単語認識率

テストセットに対する単語認識率を図 4 に示す．グラフより，denoiseDNN によって学習した特徴量を利用することで MFCC に比べて雑音頑健性が向上しているのが分かる．また，マルチストリーム HMM を用いたモダリティ統合によって全ての雑音レンジにおいて単一モダリティの場合よりも高い認識精度が達成できている．

7. まとめと今後の課題

本研究では，視聴覚音声認識に利用するための DNN を用いた画像・音響特徴量の学習方法について検証した．画像特徴量については，PCA 等に比べ DCNN で学習した特徴量が高い認識精度を示した．音響特徴量については，denoiseDNN を利用することで MFCC に比べて雑音頑健性が向上することが確認できた．また，獲得された画像・音響特徴量を統合することで単一モダリティの場合と比べて認識精度が向上した．

今後検証すべき点としては，画像にノイズがある場合や解像度が低い場合に認識精度にどのような影響があるかを検証する必要がある．

謝辞　本研究は，さきがけ領域研究「情報環境と人」，科研費新学術領域研究「構成論的発達科学」 (24119003) の助成を受けた．

参考文献

[1] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks”, *Science (New York, N.Y.)*, 2006.

[2] J. Martens, “Deep Learning via Hessian-free Optimization”, *ICML*, 2010.

[3] A. Krizhevsky, I. Sutskever and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, *NIPS*, 2012.

[4] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee and A. Y. Ng., “Multimodal deep learning”, *ICML*, 2011.

[5] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-Dependent Pre-Trained Deep Neural Networks for Large Vocabulary Speech Recognition ”, *IEEE Trans.*, 2011