論文 / 著書情報
Article / Book Information

論題(和文)	口唇深度画像を利用したディープオートエンコーダに基づくマルチモーダル音声認識
Title(English)	
著者(和文)	安井 勇樹, 岩野 公司, 井上 中順, 篠田 浩一
Authors(English)	Yuki Yasui, Koji Iwano, Nakamasa Inoue, Koichi Shinoda
出典(和文)	日本音響学会2017年秋季研究発表会講演論文集, , , pp. 117-118
Citation(English)	, , , pp. 117-118
発行日 / Pub. date	2017, 9

口唇深度画像を利用したディープオートエンコーダに基づくマルチモーダル音声認識*

安井勇樹（東工大），岩野公司（都市大），△井上中順，○篠田浩一（東工大）

1　はじめに

音声認識の耐雑音性を向上させるための一つの手段として，音響情報と口唇の動画像情報を併用するマルチモーダル音声認識が検討されている．近年，Kinect などのデバイスの登場により，口唇やその周辺の 3 次元情報を手軽に取得できるようになり，その情報を利用したマルチモーダル音声認識が検討されるようになった[1, 2]．一方，音声と画像情報の融合に深層学習を利用する手法が提案され，その有効性が確認されている．Ngiam らは音響と口唇のカラー画像（色画像）の 2 つの特徴量を入出力とするディープオートエンコーダを構築し，その中間層から得られる融合特徴量を認識に利用する手法を提案している[3]．そこで本研究では，「音響情報」「口唇の RGB の色画像情報」「口唇の深度画像情報」の 3 者をディープオートエンコーダ（Trimodal Deep Autoencoder）を利用して融合する，マルチモーダル音声認識手法を提案する．

2　Trimodal Deep Autoencoder

提案するオートエンコーダ（Trimodal Deep Autoencoder）の様子を図 1 に示す．入力は各モード（音響，色画像，深度画像）の特徴ベクトルであり，中間層でそれらを圧縮，出力層で再び各モードの特徴ベクトルが復元されるようなネットワークとなっている．共有されている中間層によって，各モードを融合した情報が表現されるため，この層の出力を新たな一つの特徴量として認識に利用する．

本研究では，入力にノイズを付与した特徴ベクトル，出力にノイズ付与前の特徴ベクトルを利用した Denoising Autoencoder [4] としてネットワークを構成する．ノイズは，「確率 p で特徴量の各成分をランダムに 0 に置き換える操作」で与える．

事前学習は，下層から上層に向けて行うのではなく，外側（入出力側）の層から内側に向かって行う．まず，それぞれのモードごとに，最上層と最下層を繋げて構成したオートエンコーダの事前学習を行い，それを分割してディープオートエンコーダの最も外側の層の初期の重みパラメータとする．このプロセスを外側から中央に向けて繰り返し，中央に 3 モードを共有する中間層を挿入してディープオートエンコーダを構成する．中央の共有中間層の前後の重みパラメータは，3 モードの特徴ベクトルを連結した高次元ベクトルを利用して事前学習する．

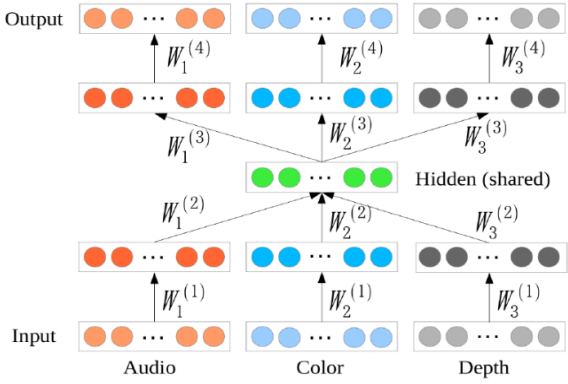

Fig. 1　Trimodal Deep Autoencoder

最後に 3 モードの連結特徴ベクトルを用いて全体のファインチューニングを行い，最終的な Trimodal Deep Autoencoder を得る．

3　評価実験

3.1 使用データ

20 名の話者による音響・画像データにより評価を行う．発声内容は ATR 音素バランス文（503 文）であり，各話者とも全 503 文を発声している．収録データ長は約 15 時間となった．音響データには 44.1kHz から 16kHz にダウンサンプリングしたものを，画像データには Microsoft Kinect 2.0 で撮影した，フレームレート 30Hz のものを用いる．各フレーム

* Deep autoencoder-based multimodal speech recognition using depth images of mouths, by Yuki Yasui (Tokyo Institute of Technology), Koji Iwano (Tokyo City University), Nakamasa Inoue, and Koichi Shinoda (Tokyo Institute of Technology).

Table 1　ディープオートエンコーダのユニット数

種類	モード	各層のユニット数
DAE_A	音響	440-200-120-80-120-200-440
BDAE_A_C	音響	440-200-120-80-120-200-440
	色画像	352-200-120-80-120-200-352
BDAE_A_D	音響	440-200-120-80-120-200-440
	深度画像	352-200-120-80-120-200-352
TDAE_A_C_D	音響	440-200-120-120-120-200-440
	色画像	352-200-120-120-120-200-352
	深度画像	352-200-120-120-120-200-352

画像について口唇領域の検出を行い，その領域画像のサイズは 96×48 に正規化した．

3.2 実験条件

音響信号からは 10msec ごとに 40 次元のメルフィルバンク係数を抽出する．グレースケール化された口唇の色画像と深度画像は，主成分分析に基づく次元圧縮を行い，フレームごとに 32 次元のベクトルに変換する．音響と画像のフレームレートを一致させるため，画像の特徴ベクトルは時間方向に 3 次元スプライン補間を施して 100Hz になるように補正する．ディープオートエンコーダへの入力／出力は，連続する 11 フレームの特徴ベクトルを連結したものとする．

性能比較を行うため，以下の 4 つのディープオートエンコーダ（DAE）に基づいて特徴量化を行った場合の実験を行う．

- 音響のみについて，DAE に基づく特徴量抽出を行った場合（DAE_A）
- 音響と口唇の色画像について，Bimodal DAE で特徴量の融合を行った場合（BDAE_A_C）
- 音響と口唇の深度画像について，Bimodal DAE で特徴量の融合を行った場合（BDAE_A_D）
- 音響，口唇の色画像，口唇の深度画像について，Trimodal DAE で特徴量の融合を行った場合（提案手法：TDAE_A_C_D）

各 DAE の各層のユニット数を表 1 に示す．

音声認識の音響モデルには triphone を単位とした DNN-HMM [5]を用いる．DNN の隠れ層は 3 層で，それぞれの層のユニット数は 1,024，入力は前後 5 フレームずつを連結した特徴ベクトルとした．言語モデルは毎日新聞記事データから学習された語彙数 63,465 の前向き 3-gram を用いる．

Table 2　認識実験の結果（単語誤り率 %）

実験条件	SN 比		
	clean	20dB	10dB
DAE_A	21.8	24.6	35.0
BDAE_A_C	21.9	23.4	25.2
BDAE_A_D	21.1	22.3	23.9
TDAE_A_C_D	20.9	22.4	24.2

認識性能は 4 分割の交差検証で評価した．ノイズ付与の確率 p は 0.2 とした．また，雑音に対する頑健性を評価するため，評価データと学習データの音響信号に同じ SN 比で白色雑音を重畳した場合の実験も行った．

3.3 実験結果

表 2 に各手法における単語誤り率を示す．雑音環境下では，提案手法（TDAE_A_C_D）によって，音響のみを利用した場合（DAE_A）からの性能の改善が見られ，SN 比 10dB の条件で約 10%の性能向上を示している．また，提案手法と色画像情報を融合した場合（BDAE_A_C）とを比べると，深度情報の利用により約 1%の性能向上が得られていることがわかる．ただし，提案手法と深度情報のみを融合した場合（BDAE_A_D）の性能はほぼ変わらず，3 者の融合による効果がまだ十分に得られていないことがわかる．

4　おわりに

本研究では，音声情報と口唇の色画像情報，深度画像情報の 3 者をディープオートエンコーダによって融合するマルチモーダル音声認識手法の提案し，その有効性の検証を行った．今後は，画像に対する雑音成分を考慮した評価や，白色雑音以外の雑音による評価などを行う必要がある．

謝辞　この研究は科研費 16H02845 の支援を受けた．

参考文献

[1] G. Galatas, et al., Proc. EUSIPCO, pp.2714–2717, 2012.
[2] 押尾他，情処研報，vol.2014-SLP-102, no. 2, pp. 1-6, 2014.
[3] J. Ngiam, et al., Proc. ICML, pp. 689-696, 2011.
[4] P. Vincent, et al., Journal of Machine Learning Research, vol. 11, pp. 3371–3408, 2010.
[5] F. Seide, et al., Proc. INTERSPEECH, pp. 437–440, 2011.