重度難聴者音声認識のための Deep Canonical Correlation Analysis を用いた音響特徴量抽出の検討*

☆高島悠樹 (神戸大), 滝口哲也 (神戸大/JST さきがけ), 有木康雄 (神戸大)

1 はじめに

現在，我が国の障害者手帳を持つ 18 歳以上の人口は 350 万人を超えており，聴覚・言語障害者の数は 36 万人とされている [1]．文献 [2] では，構音障害者音声を対象とした音響モデル適応の検証を行っているが，言語障害者などの障害者を対象としている研究は非常に少ない．本研究は，コミュニケーション手段として口話を用いる重度難聴者を対象として，音声と唇形状によるマルチモーダル音声認識を実現し，ユビキタス社会における彼らの生活の支援をすることを目的としている．

人間は発話内容を理解する際，種々の情報を統合的に利用している．音声が聞き取り難い場合，発話者の顔，特に唇の動きに注目して発話内容を理解しようとし，逆に，唇の動きと音声が不一致の場合，唇の動きに影響されて発話内容を誤って理解してしまうこともある．これは，McGurk effect(マガーク効果) と呼ばれ，音韻知覚が音声の聴覚情報のみで決まるのではなく，唇の動きといった視覚情報からも影響を受けることが報告されている [3]．このように人間による発話内容の理解には，唇の画像と音声の情報の統合的利用が極めて重要である．

唇の動きからの発話内容の読み取りは，リップリーディング (読唇) と呼ばれ，聴覚障害者にとって重要なコミュニケーション手段の一つである．リップリーディングは，背景雑音に影響されることがないため，計算機上での実現が期待されている．例えば，監視カメラに収録された会話映像のように音声が聞き取りにくい場合であっても，リップリーディングであれば発話内容の分析が可能であり，犯罪の防止や抑止に繋がると考えられる．そのため，音声の雑音に対して頑健な発話認識を行う手法の一つとして，音声情報に唇動画像情報を併用して認識を行うマルチモーダル音声認識が注目され，研究が進められている [4, 5]．

重度難聴者は耳で音を聞くことができないため，正確な発音をすることが難しく，発話スタイルが健常者と異なる．彼らのコミュニケーション手段の一つとして口話があり，訓練により意図した発話の唇の形状を作ることが可能である．そこで，彼らの音声を認識するために，唇画像を併用した音声認識システムの構築が望まれる．重度難聴者を対象としたマルチモーダル音声認識として，CNN (convolutional neural network) を用いた手法 [6] が提案されている．一般に，音声をニューラルネットワークを用いてモデリングする際，教師信号として，入力特徴量に対応する音素ラベルが使用される．しかし，重度難聴者音声の場合，発話スタイルが健常者と異なるため，強制アライメントにより得られた音素ラベルは誤りを含む．そこで，文献 [6] では，ネットワークの中間層にボトルネック層を設け，この層のユニットを認識のための特徴量とすることで，教師信号の誤りに頑健な特徴抽出を行なっている．本稿では，このアライメントの誤りに対する異なるアプローチとして，教師信号を使わない，教師なし学習による検討を行なう．

マルチモーダル学習として，canonical correlation analysis (CCA) を非線形拡張した deep canonical correlation analysis (DCCA) [7] が提案されている．CCA は 2 変数間の相関を最大にするような線形射影行列を学習する手法であり，DCCA は，2 つのニューラルネットワークにより非線形マッピングされた変数間の相関を最大にするように拡張した枠組みである．CCA と異なり，DCCA はパラメトリックな手法であり，より複雑な変換を学習することができる．DCCA は様々な分類タスク [8, 9] に応用されており，精度の向上が報告されている．DCCA の目的関数は学習データに対する教師なし学習として設計されており，上述の誤りを含む音素ラベルを用いる必要がない．本研究では，DCCA を用いて音声と唇画像からマルチモーダル音声認識のための特徴量抽出法を提案する．マルチモーダルタスクにおいて，モダリティ間の関係性を考慮することは非常に重要なことだと考えられる．音声認識タスクにおいて，一般に唇画像は音声と比べて情報量が少ないと考えられる．しかし，認識に有効な相補的な特徴を抽出することができれば，音声特徴量の品質が劣化した際に，唇画像を用いて認識率を補償することができると考えられる．DCCA により互いに高相関な特徴量を抽出することができ，このような効果を持つ特徴量の抽出が期待できる．以下，第 2 章で CCA と DCCA について述べ，第 3 章で提案手法の流れを説明する．第 4 章で従来の特徴量と比較し，第 5 章で本稿をまとめる．

---

* Audio-Visual Speech Recognition for a Person with Severe Hearing Loss Using Deep Canonical Correlation Analysis, by Yuki Takashima (Kobe University), Tetsuya Takiguchi (Kobe University/JST PRESTO), Yasuo Ariki (Kobe University)

2 Canonical Correlation Analysis と非線形拡張

2.1 Canonical Correlation Analysis

$\mathrm{X}_{audio} \in \mathbb{R}^{d_1 \times N}$, $\mathrm{X}_{visual} \in \mathbb{R}^{d_2 \times N}$ をそれぞれ，d_1, d_2 次元の音声特徴量及び画像特徴量を N サンプル並べた行列とする．ここで，これらの行列は平均 0 に正規化されているものとする．CCA において，これらの変数間の相関係数は以下の式で計算される．

$$\rho(\boldsymbol{a}, \boldsymbol{b}) = \mathrm{corr}(\boldsymbol{a}^\top \mathrm{X}_{audio}, \boldsymbol{b}^\top \mathrm{X}_{visual}) \tag{1}$$

$$= \frac{\boldsymbol{a}^\top \Sigma_{av} \boldsymbol{b}}{\sqrt{\boldsymbol{a}^\top \Sigma_{aa} \boldsymbol{a}} \sqrt{\boldsymbol{b}^\top \Sigma_{vv} \boldsymbol{b}}} \tag{2}$$

ここで，$\boldsymbol{a} \in \mathbb{R}^{d_1}$, $\boldsymbol{b} \in \mathbb{R}^{d_2}$ は射影ベクトルであり，CCA において推定されるパラメータである．$\Sigma_{av} \in \mathbb{R}^{d_1 \times d_2}$, $\Sigma_{aa} \in \mathbb{R}^{d_1 \times d_1}$, $\Sigma_{vv} \in \mathbb{R}^{d_2 \times d_2}$ はそれぞれ，X_{audio} と X_{visual} の相互共分散行列，X_{audio} 及び X_{visual} の自己共分散行列を表す．$\rho(\boldsymbol{a}, \boldsymbol{b})$ は，$\boldsymbol{a}, \boldsymbol{b}$ のスケールに対して不変であるため，各標準偏差を 1 と仮定すると，CCA で解くべき問題は以下の最大化問題となる．

$$\max_{\boldsymbol{a}, \boldsymbol{b}} \boldsymbol{a}^\top \Sigma_{av} \boldsymbol{b} \ \text{ subject to } \ \boldsymbol{a}^\top \Sigma_{aa} \boldsymbol{a} = \boldsymbol{b}^\top \Sigma_{vv} \boldsymbol{b} = 1 \tag{3}$$

また，$L \leq min(d_1, d_2)$ 個の線形射影ベクトルを使用するとき，音声及び画像特徴量の射影行列はそれぞれ，$\mathrm{A} \in \mathbb{R}^{d_1 \times L}$ 及び $\mathrm{B} \in \mathbb{R}^{d_2 \times L}$ と書け，以下のように定式化される．

$$\begin{aligned} &\text{maximize} \ \ \mathrm{tr}(\mathrm{A}^\top \Sigma_{av} \mathrm{B}) \\ &\text{subject to} \ \ \mathrm{A}^\top \Sigma_{aa} \mathrm{A} = \mathrm{B}^\top \Sigma_{vv} \mathrm{B} = \mathrm{I} \end{aligned} \tag{4}$$

ここで，$\mathrm{tr}(\cdot)$ と I はそれぞれ，対角成分の和と単位行列を表す．

CCA は，$\mathrm{T} = \Sigma_{aa}^{-1/2} \Sigma_{av} \Sigma_{vv}^{-1/2}$ の特異値分解により計算される．k 個の射影ベクトルを扱うとき，射影行列は $(\mathrm{A}, \mathrm{B}) = (\Sigma_{aa}^{-1/2} \mathrm{U}_k, \Sigma_{vv}^{-1/2} \mathrm{V}_k)$ で与えられる．ここで，$\mathrm{U}_k \in \mathbb{R}^{d_1 \times k}$, $\mathrm{V}_k \in \mathbb{R}^{d_2 \times k}$ は，T の最初から k 個の左及び右特異ベクトルを並べたものである．実際には，共分散行列 Σ_{aa}，Σ_{vv} は，正則行列となるよう正則化を加えて計算される．

2.2 Deep Canonical Correlation Analysis

DCCA は，CCA にニューラルネットワークを組み込んだものであり，複数層による非線形変換を行なう．音声及び画像特徴量 (X_{audio},X_{visual}) が与えられた時，音声及び画像のニューラルネットワークの出力を $f(\mathrm{X}_{audio}; \theta_1) \in \mathbb{R}^{o \times N}$ と $f(\mathrm{X}_{visual}; \theta_2) \in \mathbb{R}^{o \times N}$ とする．ここで，θ_1, θ_2 はそれぞれ，音声及び画像のネットワークのパラメータを表す．DCCA における相関係数は以下の式で計算される．

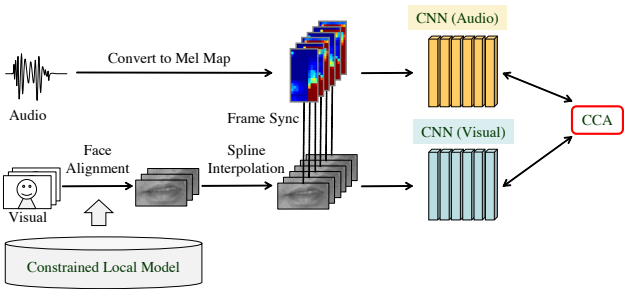

Fig. 1　Deep CCA using CNNs

$$corr(\boldsymbol{a}^\top f(\mathrm{X}_{audio}; \theta_1), \boldsymbol{b}^\top f(\mathrm{X}_{visual}; \theta_2)) = \mathrm{tr}(\mathrm{T}^\top \mathrm{T})^{\frac{1}{2}} \tag{5}$$

ここで，$\mathrm{T} = \hat{\Sigma}_{aa}^{-1/2} \hat{\Sigma}_{av} \hat{\Sigma}_{vv}^{-1/2}$ は 2.1 節と同様である．$\hat{\Sigma}_{av} = \frac{1}{N-1} \mathrm{X}_{audio} \mathrm{X}_{visual}^\top$ and $\hat{\Sigma}_{aa} = \frac{1}{N-1} \mathrm{X}_{audio} \mathrm{X}_{audio}^\top + r_1 \mathrm{I}$ は，正則化制約 $r_1 > 0$ を用いて計算された共分散行列であり，$\hat{\Sigma}_{vv}$ に対しても同様に計算される．DCCA は 2 つのデータ群に対する相関を最大にするように，パラメータ $\{\theta_1, \theta_2, \boldsymbol{u}, \boldsymbol{v}\}$ を同時に学習を行なう．パラメータ $\{\theta_1, \theta_2\}$ は式 (5) より，誤差逆伝搬法により学習され，その勾配は以下の式で計算される．

$$\frac{\partial corr(\boldsymbol{a}^\top f(\mathrm{X}_{audio}; \theta_1), \boldsymbol{b}^\top f(\mathrm{X}_{visual}; \theta_2))}{\partial f(\mathrm{X}_{audio}; \theta_1)} = \frac{1}{N-1} (2 \nabla_{aa} \mathrm{X}_{audio} + \nabla_{av} \mathrm{X}_{visual}) \tag{6}$$

ここで，$\nabla_{ab} = \hat{\Sigma}_{aa}^{-1/2} \mathrm{U} \mathrm{V}^\top \hat{\Sigma}_{vv}^{-1/2}$, $\nabla_{aa} = -\frac{1}{2} \hat{\Sigma}_{aa}^{-1/2} \mathrm{U} \mathrm{D} \mathrm{V}^\top \hat{\Sigma}_{aa}^{-1/2}$ であり，$f(\mathrm{X}_{visual}; \theta_2)$ に対する勾配も同様に計算される．

3 DCCA を用いたマルチモーダル特徴量抽出

3.1 提案手法の流れ

Fig. 1 に提案する特徴量抽出法の流れを示す．従来研究 [6] より，特徴量抽出において CNN が有効であることが示されているため，本研究では，DCCA における非線形変換に CNN を利用する．

重度難聴者の発話した音声と唇画像からネットワークへの入力特徴量を用意する．音声特徴量として，音声信号から計算されたメル周波数スペクトルを数フレーム束ねた 2 次元のメルマップを使用する．唇領域抽出のためのフェイスアライメントは，顔モデルを PDM (point distribution model) で表現し，CLM (constrained local model) の枠組みで計算し実現する．抽出された唇画像は，音声特徴量のサンプリ

ング周波数に合わせるため，3 次スプライン補間を適用する．

音声と唇画像の CNN は，DCCA の目的関数を偏微分し誤差逆伝搬法により学習される．各 CNN の学習後，入力音声特徴量と唇画像特徴量を CNN に対して順伝搬させ，得られた出力ユニットを用いて以下の式より線形射影を行なう．

$$\boldsymbol{\alpha}_t = \hat{\Sigma}_{aa}^{-1/2} \mathrm{U}_k f(\mathrm{X}_t; \theta_1) \quad (7)$$

$$\boldsymbol{\beta}_t = \hat{\Sigma}_{vv}^{-1/2} \mathrm{V}_k f(\mathrm{Y}_t; \theta_2) \quad (8)$$

ここで，$(\mathrm{X}_t, \mathrm{Y}_t)$ はそれぞれ，時刻 t における音声と唇画像の 2 次元の入力特徴量を表し，($\boldsymbol{\alpha}_t \in \mathbb{R}^k$, $\boldsymbol{\beta}_t \in \mathbb{R}^k$) は対応する射影後の特徴量を表す．これらの特徴量を連結し，$[\boldsymbol{\alpha}_t^\top \ \boldsymbol{\beta}_t^\top]^\top \in \mathbb{R}^{2k}$ を HMM への入力特徴量として扱い，音声認識を行なう．

3.2 重度難聴者音声への応用

重度難聴者の発話スタイルは健常者と異なるため，強制アライメントにより得られた音素ラベルは誤りを含む．Deep neural network (DNN) を用いた手法は教師信号を必要とし，一般に音声を扱う場合，教師信号として入力に対応する音素ラベルが用いられる．しかし，音素ラベルに誤りを含む場合，ネットワークの学習が十分に行なえず，性能向上の妨げになると考えられる．DCCA は教師なし学習の枠部みの 1 つであり，2 変数 (モダリティ) 間の相関を最大にするように学習するため，音素ラベルを使用しない．DCCA を用いることにより，音声及び唇画像はそれぞれ，互いに高相関となるように変換される．そのため，雑音環境下において，音声信号が劣化しても，唇画像は背景雑音に対して不変であるため，抽出された特徴量はノイズに対する頑健性を持つと期待される．

4 評価実験

4.1 実験条件

データセットとして，重度難聴者の男性 1 名の音声及び唇動画を収録し使用した．発話内容は ATR 音素バランス単語 A セット [10] から選択し，2,620 単語を学習，216 単語を評価に使用した．音声の標本化周波数は 16kHz，語長 16bit であり，音響分析には Hamming 窓を用いた．STFT におけるフレーム幅，シフト幅はそれぞれ 25ms，5ms である．本稿で用いる音響モデルは，54 音素の monophone-HMM で，各 HMM の状態数は 5，状態あたりの混合分布数は 6 である．

ケプストラム特徴量である MFCC+Δ+ΔΔ (36 次元) をベースラインとし，提案手法との比較を行なう．さらに，画像特徴量である discrete cosine transform (DCT) を加えた MFCC+Δ+ΔΔ+DCT (66 次元) をマルチモーダル特徴量として比較する．また，雑音環境下での認識性能を比較するため，音声データに白色雑音 (SNR:20dB, 10dB, 5dB) を加えて評価を行なった．なお，ネットワークの学習にはクリーン音声のみを用いた．

4.2 ネットワーク構成

音声 CNN の入力層には，39 次元のメル周波数スペクトルをフレーム幅 13，シフト幅 1 で分割したメルマップを用いる．画像 CNN の入力層には，発話時に顔正面から 60fps で撮影された動画を，(1) 画像列に変換し，(2)CLM により唇領域の輝度画像を抽出，(3)12 × 24pixel にリサイズを行った上で，(4) スプライン補間によってアップサンプリング (メルマップとの同期) を行った唇画像を用いる．

Table 1 に実験に用いたネットワーク構成を示す．ボトルネック特徴量の有効性を確認するため，ボトルネック層を設定している．学習率 0.0001，モーメンタム 0.99 として，確率的勾配法を用いてモデルを学習した．

Table 1 Filter size, number of feature maps and number of MLPs units for each architecture. The value for C indicates the filter size of the convolution layer that has #1 maps. The convolution layer is associated with the pooling layer. The value of S means the pooling factor. The value for M indicates the number of units for each layer in the MLP part.

	Input	C	S	#1	M
Audio CNN	39×13	4×2	3×3	13	108, 30, 108
Visual CNN	12×24	5×5	2×2	13	108, 30, 108

4.3 実験結果と考察

DCCA では，各変数に対する共分散行列を計算する必要がある．より正確な共分散行列の計算のため，ミニバッチサイズは大きく設定する必要がある [8]．そこで，クリーン環境下において，ミニバッチサイズによる影響を調査した．Fig. 2 に，ミニバッチサイズを変化させた時の実験結果を示す．ミニバッチサイズを大きくすると認識率が向上することが分かる．以降の実験では，ミニバッチサイズを 2,100 に設定する．

Fig. 2 に，従来手法との比較結果を示す．DCCA と DCCA bottleneck はそれぞれ，最終射影層とボトルネック層から抽出された特徴量を表す．DCCA は DCCA bottleneck と比べて精度が劣化する傾向が見られた．この理由として，音声特徴量が唇画像空

Table 2　Word recognition accuracy for each mini-batch size

# of mini-batches	1,200	1,500	1,800	2,100	2,400
Recognition accuracy [%]	63.89	65.28	66.20	71.76	71.76

間の近くへ射影されたことにより，音声特徴量が持つ情報が失われたためと考えられる．また，DCCA bottleneck は MFCC+DCT と比べて，SNR10dB において，良い精度が得られた．これは，DCCA を用いることで従来の特徴量よりもノイズにロバストな特徴量が得られたためと考えられる．

Fig. 3 に，教師なし学習である従来研究 [6] と比較した認識結果を示す．DCCA は教師なし学習であるため，抽出された特徴量が認識に有効な特徴を示すとは限らない．従来法は誤りを含む音素ラベルを学習に用いているが，明示的に音韻情報をモデルの学習に組み込んでいるため, 提案手法と比べて高い認識率が得られたと考えられる．提案手法は教師あり学習である従来手法と比べて，平均 14%の精度の劣化が見られた．

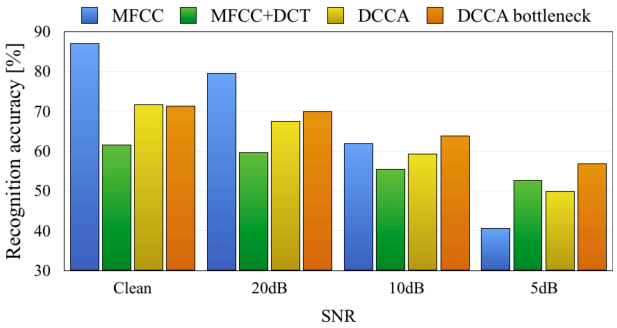

Fig. 2　Word recognition accuracy using HMMs

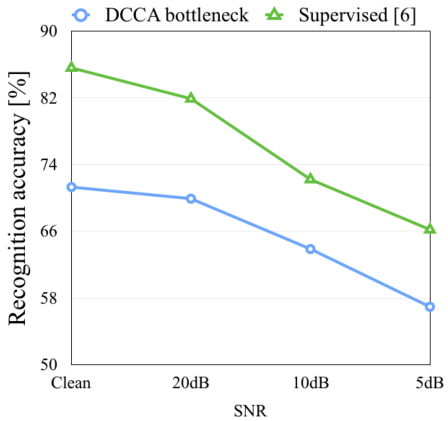

Fig. 3　Word recognition accuracy of unsupervised and supervised training procedure

5　おわりに

本稿では，DCCA を用いたマルチモーダル音声認識のための特徴量抽出法を提案した．重度難聴者の音声と唇画像を用いた雑音環境下単語認識実験により，提案手法の評価を行なった．評価実験において，提案手法は従来のケプストラム特徴量と比べて，高い認識率を示した．しかし，教師あり学習による従来研究と比べて，精度が改善されなかった．この原因として，DCCA では音韻情報を使用しないため，認識に有効な特徴量が抽出されなかったためと考えられる．今後は，音韻情報に対しても高い相関を持つような拡張を検討を行なう．

謝辞　本研究の一部は，JST　さきがけ JPMJPR15D2，JSPS 科研費 JP17J04380 の支援を受けたものである．

参考文献

[1] 内閣省，“平成 25 年版障害者白書，”．
[2] 中村圭吾 *et al.*，“発話障害者音声を対象にした健常者音響モデルの適応と検証，”日本音響学会講演論文集, pp. 109–110, 2015.
[3] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” Nature, vol. 264, pp. 746–748, 1976.
[4] G. Potamianos *et al.*, “Audio-visual automatic speech recognition: An overview,” Issues in visual and audio-visual speech processing, vol. 22, pp. 23, 2004.
[5] Y. Mroueh *et al.*, “Deep multimodal learning for audio-visual speech recognition,” in *ICASSP*, 2015, pp. 2130–2134.
[6] Y. Takashima *et al.*, “Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss,” in *INTERSPEECH*, 2016, pp. 277–281.
[7] G. Andrew *et al.*, “Deep canonical correlation analysis,” in *ICML*, 2013, pp. 1247–1255.
[8] W. Wang *et al.*, “Unsupervised learning of acoustic features via deep canonical correlation analysis,” in *ICASSP*, 2015, pp. 4590–4594.
[9] N. E.-D. El-Madany *et al.*, “Multiview learning via deep discriminative canonical correlation analysis,” in *ICASSP*, 2016, pp. 2409–2413.
[10] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol. 9, no. 4, pp. 357–363, 1990.