Convolutional Neural Network を用いた 重度難聴者のマルチモーダル音声認識*

☆柿原康博, 滝口哲也, 有木康雄 (神戸大), 三谷信之, 大森清博, 中園薫 (福祉のまちづくり研究所)

1 はじめに

現在，我が国の障害者手帳を持つ 18 歳以上の人口が 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，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].

2 提案手法の流れ

Fig. 1 に提案手法の流れを示す．音声信号に対しては，文献 [6] と同様に，Convolutional Neural Network(CNN)[7, 8] を適用するためメルマップ化を行う．また，画像に対しては，Constrained Local Model(CLM)[9, 10, 11] を用いて，画像上の 目, 口, 鼻, 眉，輪郭の位置決定を行い，唇領域の抽出 (唇の輝度画像の切り出し) を行う．

次に，抽出した唇画像の各画素の時系列に対して，音声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．

最後に，唇画像列とメルマップ列を，それぞれ事前に学習しておいたボトルネック構造のネットワークに入力し，画像と音声それぞれのネットワークからボトルネック特徴量を抽出する．その後，音声ボトルネック特徴量と画像ボトルネック特徴量を Hidden Markov Model(HMM) の入力とすることで，CNN を用いたマルチモーダル音声認識を実現する．

3 Constrained Local Model(CLM) による唇領域抽出

唇情報を用いるマルチモーダル音声認識において，画像上の 目, 口, 鼻, 眉，輪郭の位置決定 (フェイスアライメント) は重要な課題である．一般的にフェイスアライメントは, Point Distribution Model(PDM) で表現される顔モデルと，顔画像のアピアランス (濃淡パターン) によって，特徴検出器を作成し, 入力画像とのマッチングすることによって実現される．代表的な手法として，AAM(Active Appearance Model)[12], ASM(Active Shape Model)[13], CLM(Constrained Local Model)[9, 10, 11] 等が提案されており，顔がカメラに対しておおよそ正面を向いている場合に，顔の各パーツ及び輪郭を高い精度で計測することができる．

本稿の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 で表現し，CLM の枠組みで計算し実現する．CLM は顔モデルである PDM と濃淡パターンのアピアランスから作られた特徴点検出器から構成される．CLM の処理は，特徴点検出器により入力画像から顔の特徴点を検出する第 1 ステップと，顔モデルと特徴点との距離が最小となるようにモデルパラメータを最適化する第 2 ステップからなる．

3.1 Point Distribution Model(PDM)

PDM は複数人の表情から 2 次元座標データ (2D シェイプ) を取得しモデル化される．PDM の各点を示す 2 次元の位置ベクトルは,

$$\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, ..., \mathbf{X}_M^{\mathrm{T}})^{\mathrm{T}} \tag{1}$$

で表し，$\mathbf{X}_i = (x_i, y_i)^{\mathrm{T}}$ は PDM の i 番目の特徴点を示す．PDM は,

$$\mathbf{X} = \hat{\mathbf{X}} + \mathbf{\Phi_q} \tag{2}$$

と表し,$\mathbf{\Phi}$ は表情の動きと個人の違いを PCA でモデル化した行列，$\mathbf{q}$ はそのパラメータ，$\overline{\mathbf{X}}$ は 2D シェイプの平均を表す．モデルの i 番目の画像上の特徴点 $\mathbf{X}_i(\mathbf{p})$ は,

$$\mathbf{X}_i(\mathbf{p}) = s\mathbf{R}[\overline{\mathbf{X}}_i + \mathbf{\Phi q}] + \mathbf{t} \tag{3}$$

で表される．パラメータ $\mathbf{p} = \{s, \mathbf{R}, \mathbf{t}, \mathbf{q}\}$ の要素である s はスケール，$\mathbf{R}$ はピッチ α，ヨー β，ロール γ か

---

* Multimodal Speech Recognition using Convolutional Neural Networks for a Person with a Severe Hearing Loss. by Yasuhiro KAKIHARA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University), Nobuyuki MITANI, Kiyohiro OMORI, Kaoru NAKAZONO (Hyogo Institute of Assistive Technology)

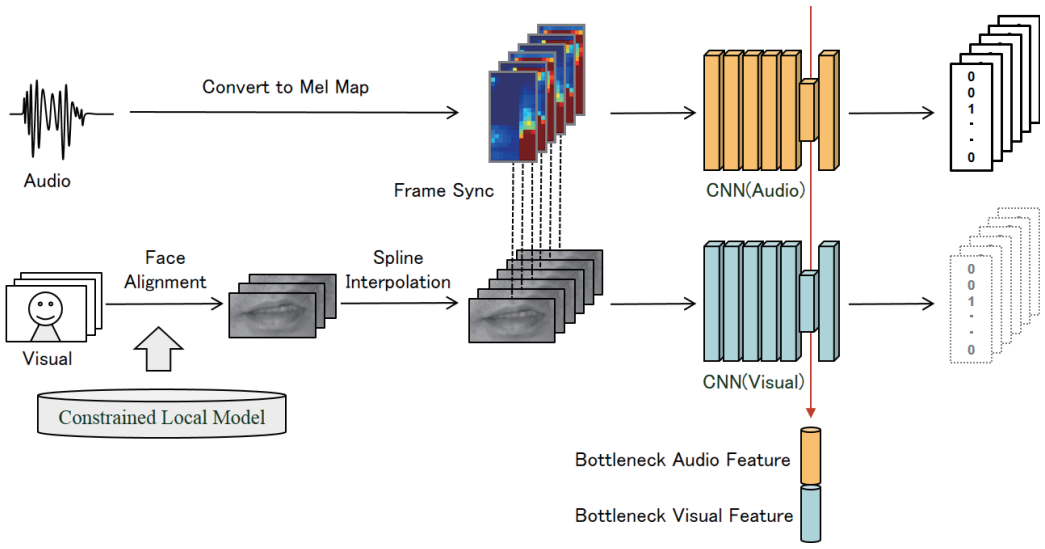

Fig. 1　Flow of the feature extraction.

らなる回転行列，$\mathbf{t}$ は平行移動，$\mathbf{q}$ は変形パラメータである．なお，$\mathbf{\Phi}_i$ は $\mathbf{\Phi}$ の i 番目の要素である．フェイスアライメントとはこの $\mathbf{p}$ を求めることに相当する．

3.2　アピアランスと特徴検出器

顔モデルを構成するフェイスアライメントの各点は，それぞれ特徴検出器を持っており，入力画像上の対応する特徴点を検出する．本稿では，検出器として Support Vector Machine(SVM) を用いて，複数人の顔特徴点のアピアランスを学習して作成する．

3.3　モデルパラメータの最適化

$\mathbf{X}_i(\mathbf{p})$ と入力画像から検出された特徴点 $\hat{\mathbf{X}}_i$ を用いて

$$Q(\mathbf{p}) = \sum_{i=1}^{M} \|\hat{\mathbf{X}}_i - \mathbf{X}_i(\mathbf{p})\|^2 + \|\mathbf{q}\|^2_{\mathbf{\Lambda}^{-1}} \qquad (4)$$

を最小化することで $\mathbf{p}$ を求める．$\mathbf{q}$ が大きくなった時に明らかに人間の顔の形から外れたフェイスアライメントの結果となる場合があるので，パラメータ $\mathbf{q}$ は平均 0，分散 $\mathbf{\Lambda}$ の正規分布に従うと仮定し，右項の第二項において $\mathbf{q}$ が極端な値を取らない様に制約を加えている．

4　CNN のボトルネック特徴量

4.1　Convolutive Bottleneck Network

提案手法では，Fig. 2 に示すようにボトルネックの構造を持つ CNN(以下 CBN) を考える．入力層からの数層は，フィルタの畳み込みとプーリングをこの順で何度か繰り返す構造をとる．つまりフィルタ出力層，プーリング層の 2 層を，プーリング層を次の層の入力層とする形で積み重ねる．出力層は識別対象のクラス数と同じサイズを持つ一次元ベクトルであり，そこに至る何層かは畳み込み・プーリングを挟まない全結合の NN(MLP) とする．提案手法では，MLP を 3 層に設計し，中間層のユニット数を少なく抑える (ボトルネック) 構成をとっている．ボトルネック特徴量はボトルネック層のニューロンの線形和で構成される空間であり，少ないユニットで多くの情報を表現しているため，入力層と出力層を結び付けるための重要な情報が集約されていると考えられる．そのため，LDA や PCA と同じような次元圧縮処理の意味合いも合わせ持つ．提案手法においては，音声・画像それぞれの CBN を学習し，CBN から得られる音声及び画像のボトルネック特徴量をマルチモーダル音声認識に用いる．

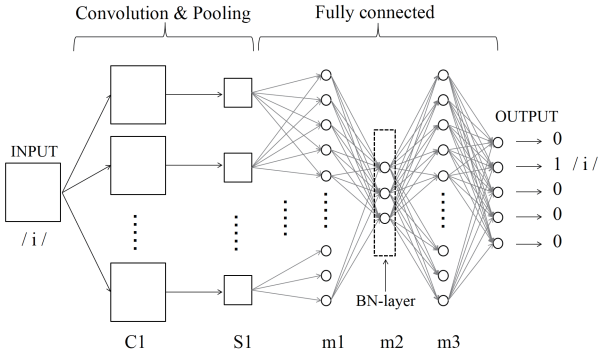

Fig. 2　Convolutive Bottleneck Network.

4.2　ボトルネック特徴量の抽出

はじめに，重度難聴者の発話音声データと発話時の唇領域画像列データを用いて，音声 CNN 及び画像 CNN の学習を行う．音声 CNN の入力層 (in) には，学習データ (音声) のメル周波数スペクトルを，オーバーラップを許しながら数フレームごとに分割して得られた 2 次元画像 (以下メルマップ) を用いる．出力層 (out) の各ユニットには，入力層のメルマップに対する音素ラベル (例えば，音素/i/のメルマップであれば，/i/に対応するユニットだけが値 1，他のユニットが値 0 になる) を割り当てる．音素ラベルを用意す

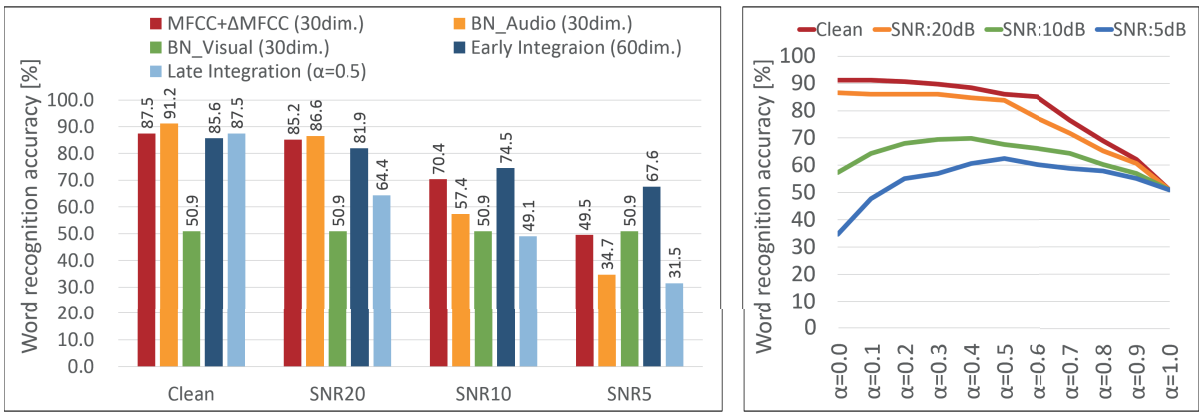

Fig. 3　Word recognition accuracy using HMMs.

るために必要な学習データの音素境界ラベルは，学習データを用いて構築された音響モデルと，その読み上げテキスト (意図された音素列) を用いた強制切り出し (forced alignment) によって求める．画像 CNN の入力層 (in) には，スプライン補間によって音声メルマップと同期のとれた唇領域画像列を用いる．出力層 (out) の各ユニットには，音声 CNN の出力層で用いた音素ラベルと同じものを割り当てる．ここで，音声 CNN 及び画像 CNN は，ランダムな初期値から学習を開始し，確率的勾配降下法 (Stochastic Gradient Descent, SGD) を用いた誤差逆伝搬により，結合パラメータを修正する．

次に，学習した音声及び画像のネットワークを用いて特徴量抽出を行う．学習データと同様に，テストデータのメルマップ及び唇領域画像を生成し，学習した音声 CNN と画像 CNN への入力とする．その後，畳み込みフィルタとプーリングによって入力データの局所的特徴を捉えて，後部の MLP 層によって音素ラベルへと非線形に変換する．入力データの情報はボトルネック層上に集約されているため，提案手法では，このボトルネック特徴量を用いてマルチモーダル音声認識を行う．

5　評価実験

5.1　実験条件

評価対象として，重度難聴者の男性 1 名が発話する ATR 音素バランス単語 B セット (216 単語) を用いた．CNN 及び HMM の学習データとして，同じ重度難聴者が発話する ATR 音素バランス単語 A セット (2620 単語) を用いた．重度難聴者の発話スタイルは，健聴者の発話スタイルと大きく異なるため，文献 [6] と同様に特定話者モデルにより認識を行う．音声の標本化周波数は 16kHz，語長 16bit であり，音響分析には Hamming 窓を用いている．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．また，本稿で用いる唇画像モデルは，音響モデル同様，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．ボトルネック層のユニット数が 30 の音声 CNN と画像 CNN を用意し，そこで得られたボトルネック特徴量を音声特徴量 (30 次元)・画像特徴量 (30 次元) として用いる．ケプストラム特徴量である MFCC+ΔMFCC(30 次元) をベースラインとし，提案手法との比較を行う．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) を加えて評価を行った．なお，音声 CNN・音声 HMM ともにクリーン音声を用いて学習を行っている．

5.2　ネットワークのサイズ

本稿では，Fig. 2 に示すように，畳み込み層とプーリング層からなる CNN と，ボトルネック層を含む 3 層の MLP とが階層的に接続されたネットワークを考える．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)$12 \times 24 pixel$ 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．

音声 CNN 及び画像 CNN の各層における特徴マップのサイズには Table 1 の値を用いた．畳み込みフィルタの数とサイズ，及びプーリングサイズは，これらの値から一意に決定される．なお，音声 CNN・画像 CNN ともに，MLP の各層 (ボトルネック層を除く) のユニット数は 108，ボトルネック層のユニット数は 30，出力層のユニット数は 54 としている．

Table 1　Size of each feature map. $(k, i \times j)$ indicates that the layer has k maps of size $i \times j$.

	Input	C1	S1
Audio CNN	1, 39×13	13, 36×12	13, 12×4
Visual CNN	1, 12×24	13, 8×20	13, 4×10

5.3　評価結果

評価を行った特徴量及び統合方法は，以下の通りである．(以降では，ボトルネック特徴量を BN 特徴量と表記する．)

- MFCC+ΔMFCC(30 次元)
- 音声 BN(30 次元)
- 画像 BN(30 次元)
- 音声 BN と画像 BN の初期統合 (60 次元)
- 音声 BN と画像 BN の結果統合 ($\alpha = 0.5$)

ただし，本稿における初期統合とは，音声特徴量と画像特徴量を繋いで 1 つの HMM に入力する統合方法を指す．また，本稿における結果統合とは，音声特徴量を音声認識のための HMM に入力し，画像特徴量をリップリーディング (読唇) のための HMM に入力し，音声 HMM と画像 HMM から出力される尤度を式 (5) で統合する方法を指す．

$$L_{A+V} = \alpha L_V + (1-\alpha)L_A, \quad 0 \leq \alpha \leq 1 \qquad (5)$$

ここで L_{A+V} は統合後の尤度，L_A，L_V，は音声と画像それぞれの尤度，α は重みである．例として，$\alpha = 0.5$ のとき，音声尤度と画像尤度の重みの比は，1 : 1 となる．

Fig. 3 の左図に，雑音環境下 (Clean, SNR:20dB, 10dB, 5dB) における単語認識結果を示す．まず，Clean 及び SNR20dB においては，ベースラインの MFCC と比べて，音声 BN 特徴量を用いた認識結果が最も良い．これは従来のケプストラム特徴量では考慮していない平行移動不変性によって，重度難聴者特有の発話変動によるスペクトルの微小な変化に対応することが可能になったためと考えられる．

また，画像 BN 特徴量のみを用いるリップリーディング (読唇) については，認識率は 50.9[%] であり，雑音に影響されない．SNR10dB においては，音声 BN と画像 BN の初期統合を行った場合，ベースラインの MFCC に対して 4.1% の認識率の改善がみられた．SNR5dB においては，音声 BN と画像 BN の初期統合を行った場合，ベースラインの MFCC に対して 18.1% の認識率の改善がみられた．従って，雑音が大きい程，音声特徴と画像特徴の統合による効果が大きいことが分かる．

Fig. 3 の右図は，結果統合に関するグラフである．横軸は式 (5) の重み α，縦軸は単語認識率を表す．ただし，$\alpha = 0.0$ は音声のみによる認識結果 (音声認識) であり，$\alpha = 1.0$ は画像のみによる認識結果 (読唇) である．このグラフから Clean 及び SNR:20dB, 10dB, 5dB のそれぞれの雑音環境下における最適な重みが読み取れる．SNR20dB 及び 5dB に関して，グラフは上に凸であり，音声尤度に画像尤度を加えていくことで認識率が改善されることが分かる．

6 おわりに

本稿では重度難聴者の特定話者モデルを用いて，CNN(CBN) によるマルチモーダル音声認識の検討を行った．ベースラインの MFCC と比べて，ボトルネックの構成を持つ CNN(CBN) を用いた特徴量抽出を行った場合，認識性能の改善が見られた．これは従来のケプストラム特徴量では考慮していない平行移動不変性によって，重度難聴者特有の発話変動によるスペクトルの微小な変化に対応することが可能になったためと考えられる．唇領域画像によるリップリーディングについては，CLM を用いることで正確にフェイスアライメントを行えることを確認した．また，唇領域画像を CNN(CBN) の入力とし，音素ラベルを教師信号として与え，ボトルネック構造とすることで，リップリーディングのための唇画像特徴量を抽出した．今後は，音声及び画像のネットワーク構造の変更とメルマップ・唇領域画像の処理の再改良を行う予定である．

参考文献

[1] 内閣府, “平成 25 年版障害者白書”.

[2] 中村圭吾, 田村直良, 鹿野清宏, “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, 3-7-4, pp.109-110, 2005.

[3] McGurk Harry, MacDonald John, “Hearing lips and seeing voices”, Nature 264(5588), pp.746-748, 1976.

[4] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, “Audiovisual automatic speech recognition ’an overview”, In Issues in Visual and Audio-Visual Speech Processing, MIT Press(In Press), 2004.

[5] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, Andrew W. Senior, “Recent Advances in the Automatic Recognition of Audio-Visual Speech”, In Proceedings of the IEEE, Vol.91, pp.1306-1326, 2003.

[6] 柿原康博, 滝口哲也, 有木康雄, 三谷信之, 大森清博, “発話に不自由のある聴覚障害者の発話音声認識の検討”, 日本音響学会 2014 年秋季研究発表会, 1-R-19, pp.109-110, 2014-09.

[7] Y. Lecun and Y. Bengio, “Convolutional networks for images, speech, and time-series”, in The Handbook of Brain Theory and Neural Networks, 3361, 1995.

[8] T. N. Sainath, A. R. Mohamed, B. Kingsbury, and B. Ramabhadran, “Deep convolutional neural networks for LVCSR”, in Advances in ICASSP, pp.8614-8618, 2013.

[9] Cristinacce, David, and Timothy F. Cootes, “Feature Detection and Tracking with Constrained Local Models”, British Machine Vision Conference, Vol.2. No.5. 2006.

[10] Saragih, Jason M., Simon Lucey, and Jeffrey F. Cohn, “Deformable model fitting by regularized landmark mean-shift”, International Journal of Computer Vision 91.2, pp.200-215, 2011.

[11] 高野博幸, 出口 光一郎, “輪郭によるフェイスアライメントにおける姿勢変化への対応のための顔輪郭の利用について (一般セッション, コンピュータビジョンとパターン認識のための機械学習及び企業ニーズセッション)”, 電子情報通信学会技術研究報告, PRMU, パターン認識・メディア理解 112.197 (2012): 65-72.

[12] T.F.Cootes, “Active Appearance Models”, Proc. European Conference on Computer Vision, Vol.2, pp.484-498, 1998.

[13] K.L. Sum, WH. Lau, S.H. Leung, Alan WC, Liew, and K. WTse, “A new optimization procedure for extracting the point-based lip contour using active shape model”, IEEE International Conference on Acoustics, Speech, and Signal Processing 2001(ICASSP 2001), pp.1485-1488, 2001.