AAMを用いた音声・画像による連続発話認識への構想

楊 楠[1,a)]　滝口 哲也[2,b)]　有木 康雄[2,c)]

1. はじめに

近年，音声認識が広く実用化されている．例えば，自動音声応答システム，スマートフォンを用いた音声による文書作成，音声認識に対応したカーナビゲーションシステムなど，さまざまな音声認識技術が世の中に出ている。しかし，現在の音声認識技術には，雑音の大きい状況下では認識性能が著しく低下してしまうという問題点があり，音声認識の大きな課題となっている．

一方，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，唇の動きと音声が一致しない場合，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．

2. 現在に至る唇画像を用いる連続音声認識

音声情報に唇動画像情報を併用するマルチモーダル音声認識 [1] は，長く研究されてきている．画像特徴量の抽出手法及び音声・画像モデルの統合手法の研究 [2][3][4][5] が行われてきたが，その中の大半が単語や連続数字を対象としてきた [6][7][8]．また，認識の画像特徴量は，従来，主成分分析 PCA[11] や DCT[12] といった画素ベースの特徴量，唇の幅や高さ [13] といった輪郭ベースの特徴量などが用いられてきた．画素情報と輪郭情報の両方を用いたもの [14][17] などもあり，単語認識の領域で良好な精度が得られたと報告されているが，連続音声認識の領域では，まだほとんど研究が進んでいない．

3. 提案手法

本研究では，マルチモーダル連続音声認識において，[14][17] などで用いられている AAM のパラメータ [9] を用いたマルチモーダル連続音声認識システムを提案する．

AAM は，特徴点の形状と特徴点の輝度値であるテクスチャを主成分分析して部分空間を構成し，比較的低次元のパラメータにより顔モデルを表現する手法である．

AAM のパラメータ (c パラメータ) を画像特徴量として，発話者の位置に関係なく，唇領域を自動的に抽出することができるというメリットがある．AAM のパラメータを採用するもう一つの理由は，この特徴量が顔方位と相関関係があり，顔方位に依存しない音声認識システムことが実現できるという点が挙げられる [16].

図 1 に提案手法の流れを示す．発話動画が入力されると，画像情報に対しては，まず，Haar-like 特徴を用いた AdaBoost 法 [10] により，顔領域検出を行う．次に検出した顔領域に対して AAM を適用する．AAM によって顔特徴点の探索を行い，入力画像に最も近いモデルのパラメータを生成し，その中から唇の情報が集約されてる特徴量を抽出する．また，発話動画の音声情報から，音声特徴量を抽出する．その後，フレームレートを合わせて 2 つの特徴量を統合し，マルチストリーム HMM により認識を行う．

音声特徴量には音声認識において最も広く用いられている MFCC を用いる．

4. 認識手法

従来の認識手法においては，音声情報と唇動画像情報を統合する手法として，初期統合や結果統合，合成統合 [18] が提案されている．初期統合は音声と画像の相関を利用しているが，非同期性を表現することができない．結果統合は音声と画像の相関を考慮していないため，音声と画像が音素レベルで相補的であるという特徴が有効に利用されていない．合成統合は音声と画像の非同期性相関をうまく利用しながら音声と画像の非同期性も表現できるが，音素境界を越えたずれを表現することができない．マルチモーダル連続音声認識の研究では，[2] が初期統合，[3] が改良した結果統合，[19] が改良した合成統合を用いている．[3] や [19] が改良した統合モデルを用いているが、認識精度にわずかの改善しか報告されていない．また，[4][5][6][7][8] はマルチストリーム HMM を用いており，その有効性が確認されている．そこで，本研究のでは，マルチストリーム

[1] 神戸大学大学院システム情報学研究科　〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
[2] 神戸大学自然科学系先端融合研究環　〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
[a)] yang.nan@me.cs.scitec.kobe-u.ac.jp
[b)] takigu@kobe-u.ac.jp
[c)] ariki@kobe-u.ac.jp

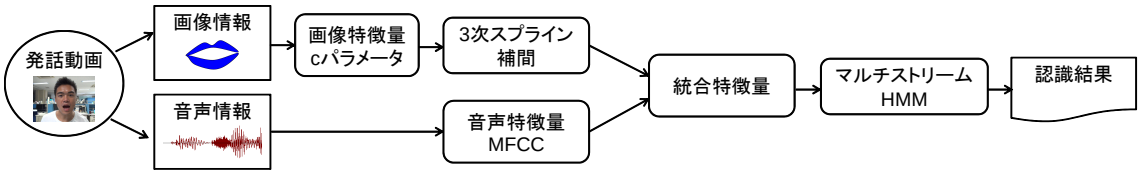

図 1 提案手法の流れ

HMM を認識手法として用いる．

マルチストリーム HMM では，音声・画像特徴量 O_t の観測確率は，対数尤度 $b(O_t)$ を用いて下記のように表記できる．

$$b(O_t) = \lambda_V b_V(O_{Vt}) + \lambda_A b_A(O_{At}) \quad (1)$$

t は時刻，$b_V(O_{Vt})$，$b_A(O_{At})$ はそれぞれ画像特徴量 O_{Vt}，音声特徴量 O_{At} に対する尤度，λ_V，λ_A は画像，音声ストリームの重みとして，認識時には以下の条件で変化させる．

$$\lambda_V + \lambda_A = 1\ , \quad 0 \leq \lambda_V, \lambda_A \leq 1 \quad (2)$$

5. 今後の予定

本論文で構想した AAM を用いた音声・画像による連続発話認識の有効性を確認するため，まずは ATR 音素バランス文 503 文に基づいて，人の発話動画を収録する．その後，連続音声認識実験を行い，この構想の有効性を証明する．さらに AAM を活用して，顔方位に依存しない連続発話認識への展開も行う予定である．

参考文献

[1] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.Senior, “Recent Advances in the Automatic Recognition of Audio-Visual Speech, ” In Proceedings of the IEEE, vol. 91, pp. 1306-1326, 2003.

[2] 石川剛，澤田裕子，全炳河，南角吉彦，宮島千代美，徳田恵一，北村正，“ 初期統合によるバイモーダル大語彙連続音声認識，” 情報科学技術フォーラム一般講演論文集，2002(2),no. F-5,pp. 203-204, Sep.2002.

[3] 石川剛，全炳河，南角吉彦，宮島千代美，徳田恵一，北村正，“ 音響尤度のリスコアリングによる結合結果を用いたバイモーダル連続音声認識，” 日本音響学会研究発表会講演論文集，2003(1), pp. 193-194, Mar.2003.

[4] 高山俊輔, 松尾俊秀, 岩野公司, 古井貞熙，“ 対話音声を対象とした不特定話者マルチモーダル音声認識の検討，” 日本音響学会 2007 年春季講演論文集，no. 2-9-13, pp. 61-62, Mar.2007.

[5] 松尾俊秀, 岩野公司, 古井貞熙，“ マルチモーダル音声対話システムのための音響・画像重みの推定手法の検討，” 日本音響学会 2009 年春季講演論文集，pp. 115-116, Mar.2009.

[6] 田村哲嗣, 岩野公司, 古井貞熙，“ 実環境におけるマルチモーダル音声認識の評価，” 日本音響学会 2002 年春季講演論文集, no. 3-5-5, pp. 151-152, Mar.2003.

[7] 田村哲嗣, 岩野公司, 古井貞熙，“ マルチモーダル音声認識における音響・画像特徴量の融合法に関する検討，” 日本音響学会 2003 年秋季講演論文集, no. 3-6-11, pp. 123-124, Sep.2003.

[8] 田村哲嗣, 岩野公司, 古井貞熙，“ 実環境を考慮したマルチモーダル音声認識のためのストリーム重み最適化手法，” 情報処理学会研究報告, 2005-SLP-55-6, vol. 2005, no. 12, pp. 29-34, Feb.2005.

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

[10] P. Viola, M. Jones, “Rapid Object Detection Using a Boosted Cascade of Simple Features, ” In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 511–518, Kauai, HI, USA, Dec.2001.

[11] M.J. Tomlinson, M.J. Russell, and N.M. Brooke, “Integrating audio and visual information to provide highly robust speech recognition, ” IEEE International Conference on Acoustics, Speech, and Signal Processing 1996(ICASSP 1996), vol. 2, pp. 821-824, Atlanta, GA, May.1996.

[12] He Jun and Zhang Hua, “Research on visual speech feature extraction, ” International Conference on Computer Engineering and Technology 2009(ICCET 2009), vol. 2, pp. 499-502, Singapore, Jan.2009.

[13] Takami Yoshida and Kazuhiro Nakadai, “Audio-visual speech recognition system for a robot, ” International Conference on Auditory-Visual Speech Processing 2010(AVSP 2010), pp. 8-13, Hakone, Kanagawa, Japan, Sep–Oct.2010.

[14] 齋藤剛史，久木貢，森下和敏，小西亮介，“複数の口唇領域を用いた単語認識，” 画像の認識・理解シンポジウム 2008(MIRU2008), no. IS-17, pp. 434-439, 軽井沢, Jul.2008.

[15] T.F. Cootes, KWalker, and C.J. Taylor, “View-based active appearance models, ” Image and Vision Computing, vol. 20, pp. 657-664, Aug.2002.

[16] 駒井祐人，楊楠，滝口哲也，有木康雄，“AAM を用いた顔方位に依存しない発話認識，” 画像の認識・理解シンポジウム 2012(MIRU2012), no. IS1-04, 福岡, Aug.2012.

[17] Yuxuan Lan, Barry-John Theobald, Richard Harvey, Eng-Jon Ong, and Richard Bowden, “Improving visual features for lipreading, ” International Conference on Auditory-Visual Speech Processing 2010(AVSP 2010), pp. 142-147, Hakone, Kanagawa, Japan, Sep–Oct.2010.

[18] J. Luettin, G. Potamianos, and C. Neti, “Ansynchronous stream modeling for large vocabulary audio-visual speech recognition,” Proc. ICASSP 2001, vol. 1, pp. 169–172, Salt Lake City, UT, May.2001. d

[19] 中村哲，熊谷健一，田村哲嗣，“バイモーダル音声認識における音素境界を越えた同期性のモデル，” 音響学会講演論文集, vol. 1, pp. 25–26, Oct.2001.