# 認識に使用する顔領域の違いによる読唇性能の比較

## Comparison of Lipreading Performance Using Different Facial Regions

池田　大輔*　桂田　浩一　入部　百合絵　新田　恒雄

Daisuke Ikeda, Kouichi Katsurada, Yurie Iribe, and Tsuneo Nitta

豊橋技術科学大学
Toyohashi University of Technology

**Abstract:** Lipreading is the technique to recognize speaker's utterances from the motion with changing shape of mouth. Although most of the previous approaches to lipreading focus on the limited region of mouth, utterances of some phonemes widely move together with its surrounding areas. We have compared three types of regions, entire face region, mouth and adjacent region, and mouth region, based on these facts. Experimental results of word recognition and vowel/consonant recognition show that vowel recognition using the entire face region results in the highest performance, while the mouth region outputs the best performance for recognizing consonants 's' and 'r'.

## 1　はじめに

顔画像からの読唇研究は，従来，主に唇の動きや形状を解析し発話内容を認識してきた[1][2]．しかし実際の発話を考えると，口の動作が大きく周辺の皺や顎の形状の変化が大きく現れる単音がある一方で，口の動作自体は小さい単音もある．このように発話する音声によって，変化する顔の部位が異なるにもかかわらず，顔領域の違いが認識に与える影響については余り報告されていない．そこで本論文では口唇領域，唇と顎などを含めた口周辺領域，顔全体領域の 3 つの領域で特徴量を抽出して単語認識を行い，各領域で母音・子音認識率を比較する．

## 2　読唇手法の概要

読唇の処理は大きく，特徴抽出部と学習・認識部からなる．まず，特徴抽出部では顔領域から読唇に有効な特徴量を抽出する．本論文では先行研究[2]で高い認識率を示した Active Appearance Models[3]（以下 AAM）のパラメータを特徴量として用いた．また，学習・認識部では特徴抽出部で得た特徴量を用いて学習と認識を行う．認識の単位としては口形素[4]を採用し，それらの分類には HMM を使用した．以下これらの概要を説明する．

### 2.1　AAM

AAM は特徴点情報と輝度情報を含む合成モデルである．AAM のモデル構築・パラメータ取得の概念図を図 1 に示す．AAM は図に示すように顔画像の形状と輝度をそれぞれ主成分分析して作成した形状モデルと輝度モデルを更に統合して作成したモデルである．構築した AAM のパラメータを変動させることにより様々な顔画像が合成できる．目的画像と AAM で合成された画像の輝度の差分が一定以下になったときの AAM パラメータを特徴量として取得する．

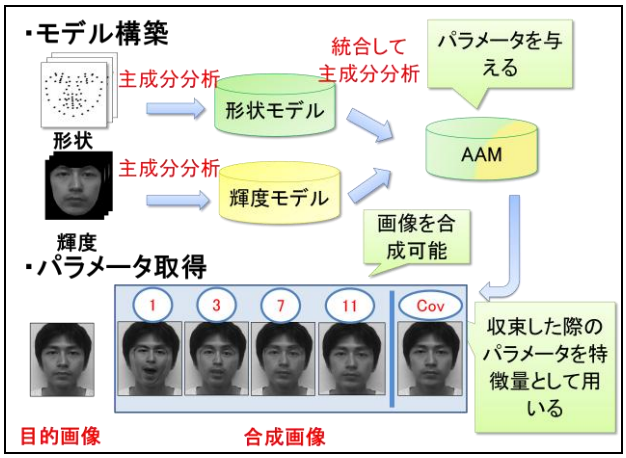



図 1：AAM の構築・パラメータ取得の概念図

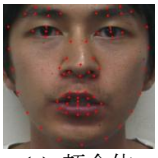
### 2.2　口形素

口形素とは発話の際に生じる口の形の最小単位である．本論文で用いた口形素は山口らの文献[4]を参考にした．表 1 に音素と口形素の対応表を示す．

表 1：音素と口形素の対応表

| 音素 | 口形素 | 音素 | 口形素 | 音素 | 口形素 | 音素 | 口形素 |
|---|---|---|---|---|---|---|---|
| a | a | p | p | ty | sy | ts | s |
| a: | | b | | hy | | z | |
| i | i | m | | ry | | s | |
| i: | | w | w | py | | k | vf |
| u | u | f | | ch | | g | |
| u: | | j | sy | dy | | h | |
| e | e | my | | sh | | q | 無し |
| e: | | ky | | r | r | N | N |
| o | o | by | | t | t | | |
| o: | | gy | | d | | | |
| y | y | ny | | n | | | |

* 連絡先：豊橋技術科学大学　情報・知能工学専攻
〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1
E-mail:ikeda@vox.cs.tut.ac.jp


(a) 顔全体


(b) 口周辺


(c) 口 唇

図 2：各領域の範囲.

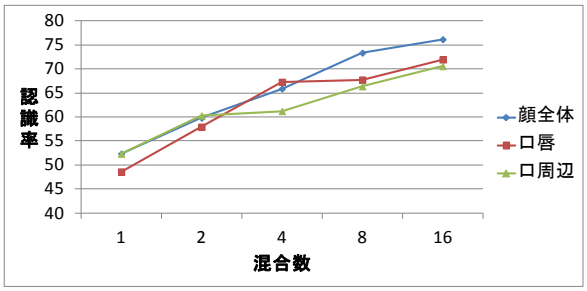



図 3：単語認識率.

## 2.3 認識手法

学習・認識には音声認識で広く用いられる HMM を使用した．HMM への入力は AAM から得られる特徴量，出力は口形素とした．また，単語認識では単語辞書により出力の口形素列を制限する．

# 3 実 験

## 3.1 実験条件

発話の際に生じる皺や顎の形状の変化が認識に与える影響を調べるため 3 種類顔領域を対象に認識実験を行った．本報告で使用した顔領域（顔全体領域，口周辺領域，口唇領域）を図 2 に示す．実験に用いた AAM は，顔画像 80 枚で構築し，形状モデル，輝度モデル，AAM のそれぞれの主成分分析の累積寄与率を 97%とした．各領域の AAM のパラメータの次元数は顔全体，口周辺，口唇領域がそれぞれ 21 次元，13 次元，11 次元となった．HMM に与える顔全体，口周辺，口唇領域の特徴量はこのパラメータに前後のフレームのパラメータから計算した線形回帰係数Δ，ΔΔを加えた 63 次元，39 次元，33 次元の特徴量とした．

発話動画像は VCV バランス単語 258 単語[5]を 8 セット，ATR 音素バランス単語 215 単語を発話した動画像を用いた．動画像のフレームレートは 30fps，話者は男性 1 名ではっきりとした口調で発話してもらった．実験では 8 セットの VCV バランス単語で HMM を学習し，ATR 音素バランス単語 215 単語を認識に用いた．HMM は 5 状態 3 ループの口形素の monophone モデルのサブワード型 HMM を用い，混合数を 1，2，4，8，16 とした．

## 3.2 実験結果

単語認識率の結果を図 3 に，母音認識率，子音認識率を各々表 2 と表 3 に示す．単語認識率では，顔全体領域が最も高い値を示した．これは母音認識率の高さが影響したものと考えられる．

以上の点は，表 2 に示す母音の平均認識率からも知られる（顔全体，口周辺，口唇領域の順に認識率が下がる）．顔全体を使用した際に母音認識率が向上する理由は，発話の際の口周辺の皺や顎の形状を捉えることができるためである．母音/i/の発話例を図 4 に示す．図から知られるように，母音では発話の際，口周辺に皺が明瞭に表れる．こうしたことから口周辺，顔全体領域は，口唇領域よりも認識率が向上したと推測できる．

一方，子音の認識率は領域により異なる．/r/や/s/など口の動きが小さい発音は，口唇領域が最も高い認識率を示している．図 4 に前後を母音/a/で挟んだ/r/の発話画像を示す．顔全体領域では/a/と認識されるが口唇領域では/r/として認識されており口唇領域は他の領域よりも唇の細かな動きを捉えていることができている．

# 4 まとめ

顔領域の違いによる単語認識率，母音・子音認識率の比較を行った．母音など口を大きく動かす場合は顔全体領域が，また子音の/s/や/r/など口の動きが小さい場合は口唇領域の認識率が良いことが分かった．今後は話者を増やしての実験や，領域を組み合わせての実験を行いたい．

表 2. 母音認識率

| | a | i | u | e | o | ave |
|---|---|---|---|---|---|---|
| 顔全体領域 | 76.9% | 70.2% | 81.7% | 76.9% | 93.8% | 80.7% |
| 口周辺領域 | 71.0% | 67.9% | 78.9% | 82.4% | 90.7% | 78.2% |
| 口唇領域 | 70.4% | 58.8% | 78.0% | 80.2% | 87.6% | 75.4% |

表 3. 子音認識率

| | p | r | sy | y | w | t |
|---|---|---|---|---|---|---|
| 顔全体領域 | 94.8% | 32.2% | 64.2% | 29.4% | 77.8% | 59.8% |
| 口周辺領域 | 90.9% | 27.1% | 60.9% | 35.3% | 77.8% | 59.8% |
| 口唇領域 | 88.3% | 44.1% | 59.6% | 23.5% | 83.3% | 63.4% |

| | s | vf | N | ave |
|---|---|---|---|---|
| 顔全体領域 | 75.9% | 42.1% | 67.3% | 60.9% |
| 口周辺領域 | 74.1% | 48.8% | 59.6% | 59.8% |
| 口唇領域 | 82.8% | 40.5% | 57.7% | 60.2% |


(a) /i/の発話画像


(b) /r/の発話画像

図 4：発話画像の例

# 参考文献


[1] I.Matthews, T.F. Cootes, J.A. Bangham, S. Cox, and R. Harvey: Extraction of Visual Features for Lipreading, IEEE Transactions of Pattern Analysis and Machine Intelligence, Vol.24, No.2, (2002)

[2] 駒井祐人，宮本千琴，滝口哲也，有木康雄: 唇領域の AAM を用いた発話認識における画像特徴量の音素解析, IS3-31, pp.1771-1778, (2010)

[3] T.F. Cootes, G.J. Edwards, and C.J. Taylor: Active Appearance Models, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.23, No.6, pp.681-685,(2001)

[4] 山口健，山本俊一，駒谷和範，尾形哲也，奥乃博: 多方向の唇画像を利用した音声認識，人工知能学会全国大会, 1E2-02, pp.1-4, (2004)

[5] 松浦博，新田恒雄: SMQ/HMM 方式に基づく不特定話者大語い単語認識，電子情報通信学会論文誌 D-Ⅱ, Vol. J76-D-Ⅱ, No. 12, pp.2486-2494,(1993)