社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

AAMを用いた顔方位にロバストな唇領域特徴抽出と音声特徴による構音障害者の音声認識

宮本　千琴† 　駒井　祐人† 　滝口　哲也† 　有木　康雄† 　李　 義昭††

中林　稔堯†††

† 神戸大学工学研究科　〒 657–8501 兵庫県神戸市灘区六甲台 1–1
†† 追手門学院大学経済学部　〒 567-8502 大阪府茨木市西安威 2–1–15
††† 神戸大学発達科学部　〒 657–8501 兵庫県神戸市灘区鶴甲 3–11
E-mail: †{miyamoto,komai}@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki,nakaba}@kobe-u.ac.jp,
†††chao55@res.otemon.ac.jp

あらまし　本稿では，アテトーゼ型脳性麻痺による構音障害者の音声認識の検討を行う．アテトーゼ型の構音障害者の場合，筋肉の緊張のため発話が不安定になりやすく，発話時に頭が動いてしまう場合がある．これに対して，音声特徴としてデルタケプストラム係数のセグメント特徴量を用いる．また，発話時の頭部の動きに対しては，Active Appearance Model（AAM）を用いることで画像から顔方位にロバストな唇領域特徴を抽出し，音声特徴と共に用いることで，雑音の影響を受けず発話変動を考慮したマルチモーダル音声認識を検討する．
キーワード　構音障害，マルチモーダル音声認識，Active Appearance Model

Dysarthric Speech Recognition Using Pose-Robust Lip Area Feature Extraction Based on AAM and Acoustic Features

Chikoto MIYAMOTO†, Yuto KOMAI†, Tetsuya TAKIGUCHI†, Yasuo ARIKI†, Ichao LI††,

and Toshitaka NAKABAYASHI†††

† Graduate School of Engineering, Kobe University, 1–1 Rokkodaicho, Nada-ku, Kobe, Hyogo, 657–8501, Japan
†† Faculty of Economics, Otemon Gakuin University, 2–1–15 Nishiai, Ibaraki, Osaka, 567–8502, Japan
††† Faculty of Human Development, Kobe University, 3–11 Tsurukabuto, Nada-ku, Kobe, Hyogo, 657–8501, Japan
E-mail: †{miyamoto,komai}@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki,nakaba}@kobe-u.ac.jp,
†††chao55@res.otemon.ac.jp

Abstract　We investigated the speech recognition of a person with articulation disorders resulting from athetoid cerebral palsy. The articulation of speech tends to become unstable due to strain on speech-related muscles, and that causes degradation of speech recognition. Therefore, we use multiple acoustic frames as an acoustic feature to solve this problem. Further, in a real environment, the speech recognition systems do not have sufficient performance due to noise influence. In addition to acoustic features, visual features are used to increase noise robustness in a real environment. However, there is a recognition problem due to the tendency of his/her unsettling head movement. We investigate a pose-robust audio-visual speech recognition method using Active Appearance Model (AAM) to solve this problem.
Key words　articulation disorders, audio-visual speech recognition, Active Appearance Model

1. はじめに

情報技術が向上し，近年，福祉分野への情報技術の適用が行われている．例えば，画像認識技術を用いた手話認識 [1] や，文書内の文字の音声化などが行われている [2]．また，音声合成を用いて，発話障害者支援のための音声合成器の作成なども行われている [3]．

音声認識技術は近年，飛躍的に進歩し，様々な環境や場面での利用が期待されている．例えばカーナビゲーションの操作や会議音声の議事録化など様々な分野に応用されている．対象者が子供である場合などには精度が低下することがわかっている [4]．文献 [5] 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ものは非常に少ない．現在，日本だけでも構音障害者も含まれる言語障害者が 4 万 2000 人もいることから十分なニーズがあり，研究の必要性があるといえる [6]．

言語障害の原因の一つとして，脳性麻痺が考えられる．脳性麻痺の定義として，1968 年の厚生労働省脳性麻痺研究班は「受胎から生後４週以内の新生児までの間に生じた，脳の非進行性病変に基づく，永続的な，しかし変化しうる運動および姿勢の異常である．その症状は満２歳までに発現する.」としている．

脳性麻痺とは，筋肉の動きをつかさどる脳の部分が受けた損傷が原因で筋肉の制御ができなくなり，けいれんや麻痺，そのほかの神経障害が起こる症状のことである．出生前，出生時，出生直後の脳への酸素供給，出生前の胎内感染，妊娠中毒症，分娩時の外傷，仮死状態，未熟出生，出生後の脳を覆う組織の炎症や外傷性損傷などが原因として考えられる．

脳性麻痺は，脳の損傷部分によって主に痙直型（大脳皮質），アテトーゼ型（中脳もしくは脳基底核），失調型（小脳），混合型（脳の広範囲）に分類される．痙直型は正常な筋の伸張反射が過度になる，アテトーゼ型はアテトーゼと呼ばれる筋肉の不随意運動を伴う，失調型は協調運動の障害，混合型はそれぞれの症状が混合して現れる，というような症状が見られる．

本稿では，アテトーゼ型の脳性麻痺による構音障害者を対象としている．アテトーゼ型は，脳性麻痺患者の約 20% に発生する．筋肉の随意運動や姿勢の調整を行っている大脳基底核（大脳皮質，視床や脳幹を結び付けている神経核の集まり）に損傷を受けたことにより，筋肉が不随に動き，正常に制御できないアテトーゼと呼ばれる症状が見られる．とくに緊張状態にあるときや，意図的動作を行うときに見られる．症状は軽度から重度まで様々であり，知能障害を合併していないケースや比較的知能障害の程度が軽いケースも多いのが特徴である [7] [8]．そこで本稿では，まず知能障害を合併していないアテトーゼ型に着目した．

アテトーゼ型の構音障害者の発話スタイルは，筋肉の緊張のため健常者と大きく異なり不安定になる場合がある．従来の音声認識では，対数スペクトルに対し離散コサイン変換を適用した MFCC (Mel Frequency Cepstral Coefficient) を特徴量として用いるが，我々は離散コサイン変換ではなく 2 回目以降のより安定したデータを利用した，PCA (Principal Component Analysis) による発話変動にロバストな手法を提案してきた [9]．また，我々は動的特徴量を用いた音声認識において，構音障害者の認識精度が健常者に比べて大きく低下することに着目した．構音障害者において，動的特徴量は時間特徴が十分に表現されていないと言える．これに対し，動的特徴量の代わりに音声特徴として，デルタケプストラム係数のセグメント特徴量を用い，認識精度の改善を行った [10]．本稿では音声特徴としてセグメント特徴量を用いる．

さらに，雑音の多い実環境下では音声特徴のみを用いて発話内容を認識することは難しい．そこで，音声特徴と同時に画像特徴を用いることで，発話の検出精度や，音声認識における耐雑音性を高める研究が盛んに行われている．近年では，Lucey らは，正面画像と横顔画像を用いた手法を提案している [11]．岩野らは，横顔から唇領域特徴を抽出して音声認識を行う手法を提案している [12]．構音障害者の場合は，発話時に頭が動いてしまう場合があり，認識精度に影響を及ぼす可能性がある．そこで本稿では，AAM [13] を用いることで顔方位にロバストなマルチモーダル音声認識を検討する．AAM は顔方位の変化のある画像からモデルを構築するため，顔方位が変化しても顔特徴点が追跡可能である．また，AAM を使用することで，顔の方位を正面に戻すことが可能である [14]．本稿では，構音障害者の画像及びクリーン音声，雑音重畳音声を用いて単語認識実験を行い，有効性を示す．

2. AAM

AAM は shape（特徴点の座標値）と texture（輝度値）をそれぞれ PCA によって次元削減することにより，少ないパラメータで顔の形状の変化とテクスチャの変化を表現できるようにしたモデルである．変形を伴う物体を高速かつ安定して追跡することが可能であり，顔特徴点抽出や発話認識において広く用いられている [15] [16]．また，AAM は顔画像の平均形状から学習により得られるパラメータにより，学習サンプルに十分近い画像を生成することもでき，入力された顔が横顔であっても正面の画像に補正することが可能である．

2.1 AAM の構築

まず，AAM の構築法について述べる．図 1 のように 63 点の特徴点の座標が与えられた複数の画像を学習データとする．顔画像の i 番目の特徴点座標を (x_i, y_i) とすると全特徴点を並べたベクトル $\boldsymbol{s}$ は，

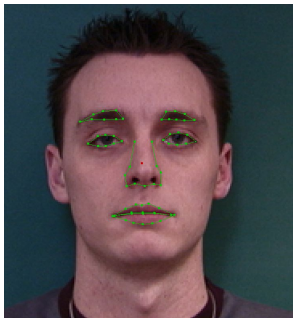

図 1　AAM の構築に用いた特徴点（緑色の点が特徴点を表す）

$$\boldsymbol{s} = (x_1, y_1, \cdots, x_n, y_n) \tag{1}$$

と表すことができる．この $\boldsymbol{s}$ を shape ベクトルと呼ぶ．これに対して PCA を行うと，shape ベクトル $\boldsymbol{s}$ は，

$$\boldsymbol{s} = \bar{\boldsymbol{s}} + \boldsymbol{P_s b_s} \tag{2}$$

と表現される．ここで，$\bar{\boldsymbol{s}}$ は全学習データに対する平均 shape ベクトルであり，$\boldsymbol{P_s}$ は shape 空間の基底ベクトル，$\boldsymbol{b_s}$ はそれに対応したパラメータベクトルである．

次に，全ての学習サンプルを共通の形状に変形し，形状について正規化されたテクスチャの輝度値を並べたベクトルを $\boldsymbol{g}$ とし，texture ベクトルと呼ぶ．texture ベクトル $\boldsymbol{g}$ も $\boldsymbol{s}$ と同様に PCA を行い，

$$\boldsymbol{g} = \bar{\boldsymbol{g}} + \boldsymbol{P_g b_g} \tag{3}$$

と表現される．ここで，$\bar{\boldsymbol{g}}$ は全学習データに対する平均 texture ベクトルであり，$\boldsymbol{P_g}$ は texture 空間の基底ベクトル，$\boldsymbol{b_g}$ はそれに対応したパラメータベクトルである．また，$\boldsymbol{b_s}$，$\boldsymbol{b_g}$ は平均からの変化を表すパラメータであり，これらを変動させることで shape と texture を変化させることができる．ここで，$\boldsymbol{b_s}$，$\boldsymbol{b_g}$ を結合し，式 (4) のように表現する．

$$\boldsymbol{b} = \begin{pmatrix} \boldsymbol{W_s b_s} \\ \boldsymbol{b_g} \end{pmatrix} = \begin{pmatrix} \boldsymbol{W_s {P_s}^T (s - \bar{s})} \\ \boldsymbol{{P_g}^T (g - \bar{g})} \end{pmatrix} \tag{4}$$

ここで，$\boldsymbol{W_s}$ は shape ベクトルと texture ベクトルの単位を正規化する行列である．さらに，このベクトルに対して PCA を行うと，ベクトル $\boldsymbol{b}$ は，

$$\boldsymbol{b} = \boldsymbol{Qc} = \begin{pmatrix} \boldsymbol{Q_s} \\ \boldsymbol{Q_g} \end{pmatrix} \boldsymbol{c} \tag{5}$$

と表すことができる．ここで，$\boldsymbol{Q}$ は基底ベクトル，$\boldsymbol{c}$ はそれに対応したパラメータベクトルで，combined パラメータベクトルと呼ぶ．$\boldsymbol{s}$，$\boldsymbol{g}$ を $\boldsymbol{c}$ を用いて表現すると式 (6)(7) のようになる．

$$\boldsymbol{s(c)} = \bar{\boldsymbol{s}} + \boldsymbol{P_s {W_s}^+ Q_s c} \tag{6}$$

$$\boldsymbol{g(c)} = \bar{\boldsymbol{g}} + \boldsymbol{P_g Q_g c} \tag{7}$$

このようにして，combined パラメータベクトル $\boldsymbol{c}$ のみで，shape の変化と texture の変化を制御することが可能となる．

2.2　AAM による探索

入力画像が与えられた時，構築した AAM を用いて特徴点の探索を行う．AAM に対して平行移動や回転を行い，$\boldsymbol{c}$ を用いて変形させ合成されたモデル画像の texture と入力画像の texture を比較し，その誤差を最小にする $\boldsymbol{c}$ を再急降下法によって求める．このようにして入力画像に最も近い AAM を探索することができ，特徴点を抽出できる．

2.3　AAM による正面画像の生成

顔方位の成分は combined パラメータベクトル $\boldsymbol{c}$ の低次元に現れる．よって，$\boldsymbol{c}$ を低次元のベクトルとすると，

$$\boldsymbol{c} = \boldsymbol{c_0} + \boldsymbol{c_1} * \theta \tag{8}$$

とおける．ここで，$\boldsymbol{c_0}$，$\boldsymbol{c_1}$ を定数，θ を顔方位角度とする．これに基づいて学習データから最小二乗法によって $\boldsymbol{c_0}$，$\boldsymbol{c_1}$ を求める．次に，入力顔画像から，

$$\theta' = (\boldsymbol{c'} - \boldsymbol{c_0}) / \boldsymbol{c_1} \ (c_1 \neq 0) \tag{9}$$

のようにして取得した $\boldsymbol{c'}$ から，顔方位角度 θ' を求める．また，顔方位を正面に戻すには，式 (8) に $\theta = 0$ を代入し，$\boldsymbol{c_{res}}$ を残差ベクトルした

$$\boldsymbol{c_{front}} = \boldsymbol{c_0} + \boldsymbol{c_{res}} \tag{10}$$

によって算出される $\boldsymbol{c_{front}}$ を用いる．

3.　マルチモーダル音声認識

本章では，音声特徴と画像特徴を用いた音声認識システムについて述べる．

3.1　音声特徴抽出

音声認識システムにおいて従来は，音声特徴量として MFCC や特徴量の線形回帰係数である ΔMFCC や ΔΔMFCC が広く用いられている．ΔMFCC は以下のように求められる．

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2} \tag{11}$$

ここで，c_t は時間 t におけるケプストラム係数，Θ は窓幅を表す．同じように，ΔΔMFCC は ΔMFCC に適用することで求められる．しかし，構音障害者の発話スタイルは不安定であるため，MFCC を用いた特定話者モデルでの音声認識には限界がある．特に動的特徴量である ΔMFCC を用いた音声認識において，構音障害者の認識精度は大きく低下する (図 7)．そこで，ΔMFCC の代わりに，デルタケプストラム係数のセグメント特徴量を音声特徴量として用いる．図 2 に

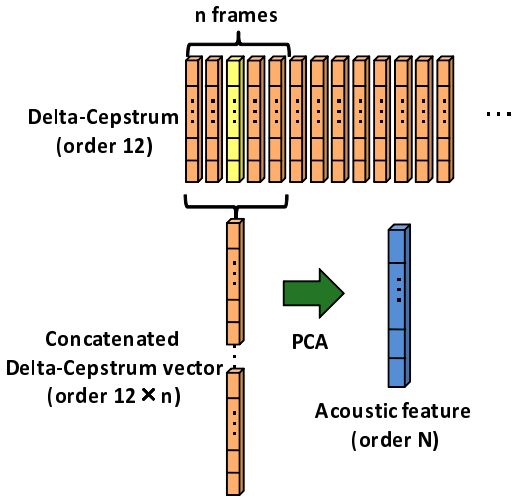

図 2　セグメント特徴抽出

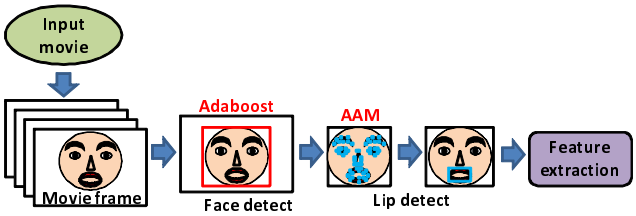

図 3　画像特徴抽出

セグメント特徴量抽出の流れを示す．当該フレームとその前後数フレームの計 n フレームを連結させ，PCA により N 次元に圧縮を行ったものを音声特徴量とする．実際には MFCC と組み合わせたものを音声特徴量として用いる．

3.2　画像特徴抽出

図 3 に画像特徴抽出の流れを示す．AAM は入力画像中における初期探索位置と実際の顔特徴点の位置が大きく離れていると正確に特徴点を抽出することができない．本稿では安定して特徴点を抽出するためにまず，Haar-like 特徴を用いた AdaBoost 法による顔領域検出 [17] を行い，抽出された顔領域を初期探索位置とする．次に，AAM による特徴点の探索を行い，顔の方位を正面に戻す．AAM を用いることで頭が動いている画像も正面に戻すことができ，顔方位にロバストな特徴抽出が可能となる．これにより唇領域が得られ，その領域を 32×32 ピクセルにリサイズすることで，画面内の唇サイズに影響を受けないようにする．このリサイズされた唇領域をブロック分割し，それぞれのブロックにおける平均輝度値を求めることでモザイク化する．モザイク化した画像に対して二次元離散コサイン変換を行い，得られた特徴ベクトルを PCA により次元削減したものを画像特徴量して用いる．

3.3　音声情報と画像情報の統合

音声特徴量，画像特徴量をそれぞれ用いて，音声 HMM，画像 HMM を構築する．音声 HMM を画像 HMM と統合することで，音響的な雑音にロバストな認識が可能であるだけでなく，雑音がない環境下においても，構音障害者の音声認識率が低下するという問題に対する精度の改善が期待できる．認識時において，両 HMM の尤度に対して，式 (12) を用いて統合を行う．

$$L_{Aco+Vis}^{\acute{w}_{N-best}} = (1-\alpha)\cdot L_{Aco}^{\acute{w}_{N-best}} + \alpha \cdot L_{Vis}^{\acute{w}_{N-best}} \qquad (12)$$

ここで，L_{Aco}, L_{Vis} はそれぞれ，音声 HMM，画像 HMM の尤度を表す．統合の精度を高めるために，単語認識結果の N-best 単語に対してのみ統合を行う（図 4）．

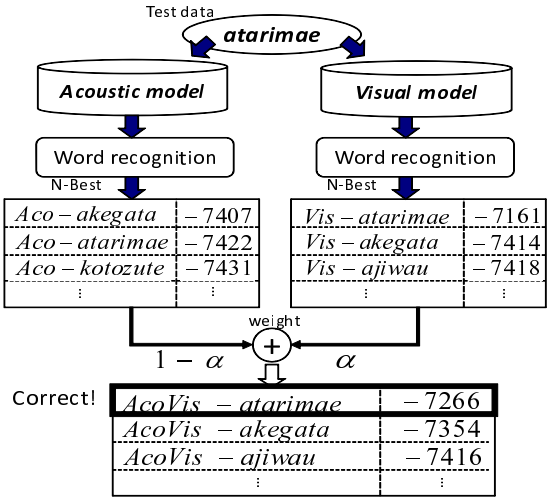

図 4　Example of integrated recognition

4.　認 識 実 験

4.1　実 験 条 件

実験用データとして構音障害者 1 名のデータを収録した．発話内容として ATR 音素バランス単語 1065 発話（216 単語 ×5 回）と ATR 音声データベース 5240 発話（2620 単語 ×2 回）を使用し，各発話を手動で切り出した．音声データのサンプリング周波数は 16 kHz，フレーム窓長は 25 mec，フレーム周期は 10 msec である．画像データの解像度は 720×480，フレームレートは 30 fps である．図 5 に構音障害者，図 6 に健常者のスペクトログラム例を示す．構音障害者の場合，子音など高域のパワーが弱く，明瞭度が劣化している．音響モデルの学習には 5240 発話のクリーン音声データを使用し，1065 発話のクリーン音声データ及び雑音を重畳したデータを評価データとして使用した．初期モデルの作成，学習，認識には HTK [18] を用いた．

4.2　構音障害者モデルでの認識実験

健常者音声で学習した音響モデルでの認識は，健常者と発話スタイルが異なるため困難であることから，まず構音障害

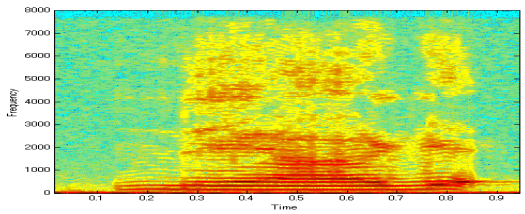

図 5　構音障害者のスペクトログラム例//n e a g e

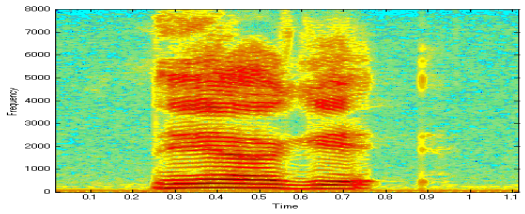

図 6　健常者のスペクトログラム例//n e a g e

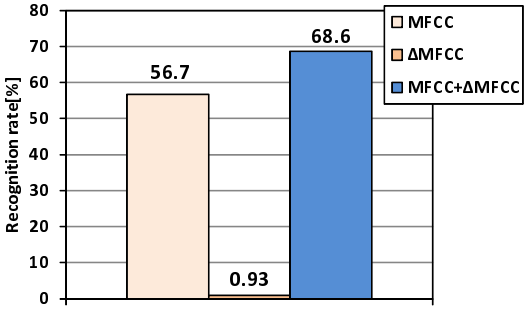

図 7　特定話者モデルでの認識結果

者の音響モデルを作成し認識実験を行った．特徴量として 12 次 MFCC，ΔMFCC，MFCC+ΔMFCC を用い，音響モデルは monophone（54 音素，6 混合）を用いた．認識結果を図 7 に示す．

ΔMFCC における認識率が 0.93% とほかの特徴量に比べて著しく低下していることがわかる．これは発話という意図的動作時で筋肉の緊張によってアテトーゼが生じて調音が困難になり，明瞭度が劣化したため，ΔMFCC では時間特徴が十分に表現されていないと考えられる．

4.3　セグメント特徴量による認識実験

ΔMFCC の代わりに，デルタケプストラム係数 12 次元に対しセグメント特徴量を求め，これを音響特徴量として用いた結果を示す．今回は予備実験より，フレーム数 n を 3 として実験を行った．ΔMFCC の結果と比較するために，PCA により ΔMFCC と同次元（N=12）に次元圧縮したセグメント特徴量を用いた場合のの認識結果を図 8 に示す．

ΔMFCC の代わりにセグメント特徴量を用いることにより，セグメント特徴量単体では 58.9% まで認識率が改善さ

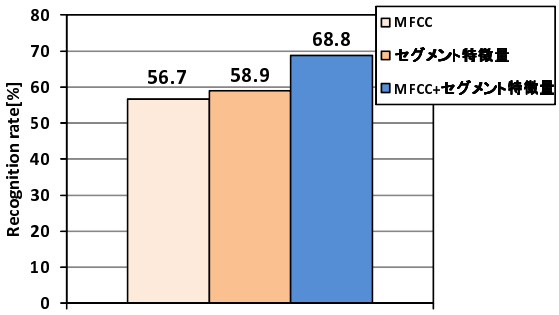

図 8　セグメント特徴量を用いた場合の認識結果（3 フレーム，12 次元）

表 1　音声のみ，画像のみ，音声＋画像の認識率の比較 (%)

SNR	Audio-only	Visual-only	Audio-visual (optimized α)
clean	68.8	35.9	74.1 (0.15)
20 dB	68.0	35.9	73.7 (0.15)
10 dB	57.7	35.9	64.3 (0.1)
5 dB	51.6	35.9	58.9 (0.3)
0 dB	4.9	35.9	35.9 (1.0)

れた．また，MFCC とセグメント特徴量を組み合わせたものと MFCC+ΔMFCC を比較した場合はわずかであるが認識率が改善された．これらの結果より，セグメント特徴量は ΔMFCC に比べ，認識率の改善に貢献していることがわかる．

4.4　音声情報と画像情報の統合

単語認識時の 5Best 単語に対して，音声 HMM と画像 HMM の統合を行った結果を示す．画像に対する重み α を 0.0〜1.0 として実験を行った．表 1 に音声データの SNR 比を変化させた音声のみの結果，画像のみの結果，音声情報と画像情報の統合結果, 最適な重み α を示す．

画像情報との統合によって，SNR 比が 5 dB のとき 58.9% まで認識率が改善された．また，全ての雑音条件に対して音声のみを用いた時よりも認識率が改善された．図 9，図 10，図 11 にそれぞれクリーン音声の場合，及び SNR 比が 10 dB，5 dB の場合における重み α を変化させたときの結果を示す．音声情報に画像情報を統合することで音声情報のみの認識率を下回る場合もあり，最適な重みを選択して統合することが必要であることがわかる．画像情報を用いることで，雑音の影響を受けず，発話変動を考慮することができ，認識率の改善が得られたと考えられる．

5.　おわりに

動的特徴量である ΔMFCC を用いた音声認識において，構音障害者の認識精度が大きく低下することに着目し，ΔMFCC

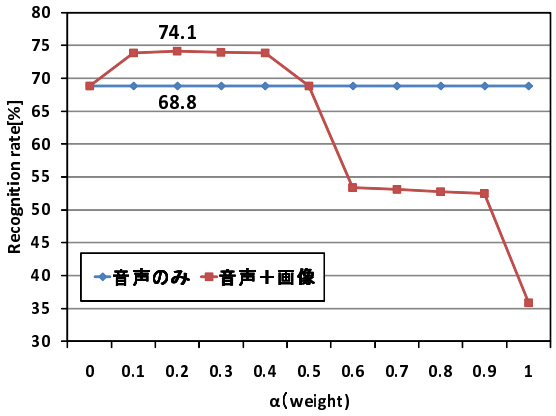

図 9　clean 音声の場合における音声のみ，音声＋画像の認識結果

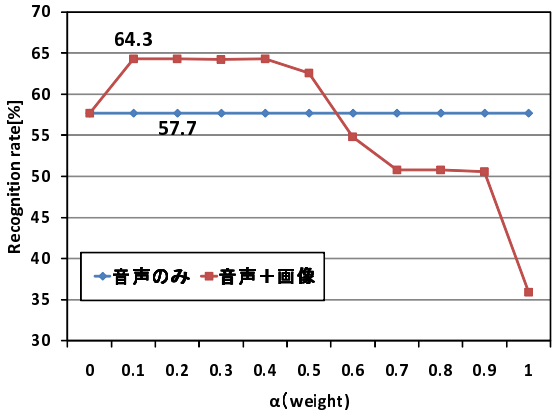

図 10　10 dB の場合における音声のみ，音声＋画像の認識結果

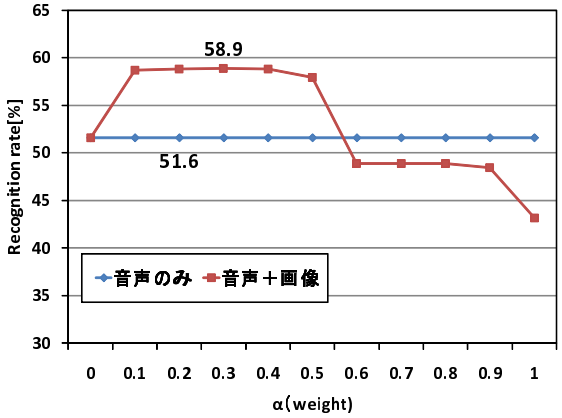

図 11　5 dB の場合における音声のみ，音声＋画像の認識結果

の代わりにセグメント特徴量を用いた．また，音声情報と画像情報の統合を検討した．その結果，7.3%（51.6% → 58.9%）の改善が得られた．今後は，構音障害者特有の特徴量の検討や，音質改善の試み認識率の改善に取り組んでいく．また更に対象者を増やし有効性を確認していく予定である．

文　献

[1] 佐川浩彦，酒匂裕，大平栄二，崎山朝子，阿部正博，“圧縮連続 DP 照合を用いた手話認識方式，” 電子情報通信学会論文誌，Vol.J77-D2，No.4，pp. 753–763，1994.

[2] 鈴木悠司，平岩裕康，竹内義則，松本哲也，工藤博章，大西昇，“視覚障害者のための環境内の文字情報抽出システム，” 電子情報通信学会技術研究報告，WIT2003-314，pp. 13–18，2003.

[3] 藪謙一郎，濱篤志，伊福部達，青村茂，“発話障害者支援のための音声合成器—その研究アプローチと設計概念—，” 電子情報通信学会技術研究報告，SP2006-164，pp. 25–30，2007.

[4] 鮫島充，李晃伸，猿渡洋，鹿野清宏，“子供音声認識のための音響モデルの構築および適応手法の評価，” 電子情報通信学会技術研究報告，SP2004-114，pp. 109–114，2004.

[5] 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.

[6] 内閣府，“平成 20 年版障害者白書，” http://www8.cao.go.jp/shougai/

[7] S.Terry Canale，落合直之，藤井克之，“キャンベル整形外科手術書 第 4 巻 小児の神経障害/小児の骨折・脱臼，” エルゼビア・ジャパン，2004.

[8] Mark H. Beers，福島雅典，“メルクマニュアル医学百科 最新家庭版，” 日経 BP 社，2004.

[9] H. Matsumasa, T. Takiguchi, Y. Ariki, I. LI and T. Nakabayashi, ”PCA-Based Feature Extraction for Fluctuation in Speaking Style of Articulation Disorders,” INTERSPEECH-2007, pp. 1150–1153, 2007.

[10] 宮本千琴，滝口哲也，有木康雄，李義昭，中林稔堯，“構音障害者の音声認識における動的特徴量の考察，” 電子情報通信学会技術研究報告，SP2009-55，pp. 37–42，2009.

[11] P. lucey, G. Potamianos and S. Sridharan, “A Unified Approach to Multi-Pose Audio-Visual ASR,” INTERSPEECH-2007, pp. 650–653, 2007.

[12] K. Iwano, T. Yoshinaga, S. Tamura and S. Furui, “Audio-Visual Speech Recognition Using Lip Information Extracted from Side-Face Images,” EURASIP Journal on Audio, Speech, and Music Processing, vol.2007, ID 64506, 2007.

[13] T. F. Cootes, G. J. Edwards and C. J. Taylor, “Active appearance models,” ECCV, volume 2, pp. 484–498, 1998.

[14] T. F. Cootes, K. Walker and C. J. Taylor, “View-based active appearance models,” Image and Vision Computing 20, pp. 227–232, 2002.

[15] Akihiro Kobayashi, Junji Satake, Takatsugu Hirayama, Hiroaki, Kawashima, Takashi Matsuyama, “Person-Independent Face Tracking Based on Dynamic AAM Selection,” CVIM，No.3，pp. 35–40，2008.

[16] 齊藤剛史，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，MIRU2008，IS1-17，pp. 434–439，2008.

[17] P. Viola and M. Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.1–9, 2001.

[18] “HTK (Hidden Markov Model Toolkit),” http://htk.eng.cam.ac.uk/.