Vol.2014-SLP-102 No.2
2014/7/24

口唇の深度画像を用いたマルチモーダル音声認識

押尾 翔平[1,a)]　岩野 公司[2]　篠田 浩一[1]

概要：音声認識の雑音耐性の向上のための手段のひとつとして，唇動画像情報を音声情報とともに利用するマルチモーダル音声認識の研究が数多く行われている．本研究では，音声認識のための画像特徴量として，従来の正面画像に加え，Microsoft Kinect から得られる深度情報を用いる手法を提案する．HMM による口唇・口腔の輪郭抽出手法に深度情報を入力として加えるほか，唇の突出などで生じる凹凸を画像特徴量として抽出する手法を導入した．日本語文音声に対する連続音声認識実験の結果，複数話者のデータを用いた際に，単語正解精度が 66.0%から 67.0%に増加し，発声時に口を尖らせる音素や舌が口腔を塞ぐような動きをする音素に対して提案手法が特に有効であることが確認された．

1. はじめに

近年，音声認識技術は社会に広く普及しており，その技術を利用した様々な情報システムが実用化されている．しかし，現在の音声認識技術は，雑音環境下では認識率が著しく低下し，十分に機能しなくなってしまう．従って，雑音に対してより頑健な音声認識システムが必要となる．

耐雑音性の強化のための手段のひとつとして，発声時の唇動画像情報を音声情報とともに利用するマルチモーダル音声認識がある．動画像情報は音響雑音の影響を全く受けないため，マルチモーダル音声認識システムは実環境での利用に耐えうるシステムとして期待され，現在に至るまで数多くの研究が行われている．

これまでに行われているマルチモーダル音声認識に関する研究は，話者を正面から撮影した画像を用いているものがほとんどである [1], [2], [3], [4], [5]．ところが，日本語の音素，特に子音の中には，正面から見た場合の形状が似通っているものがあるため [6]，正面画像だけではこれらを適切に分類することができない．より高精度に音素の分類を行うための手段として，唇の奥行きや横顔の画像も含めた画像特徴量を用いるという手法があり，正面画像のみを用いる場合と比べ認識性能が向上するという報告がされている [7], [8], [9]．しかしこれらの報告では，特徴量を得るために話者に特殊な機械を取り付けたり，撮影に複数のカメラを用いたりするなど，実環境とかけ離れた撮影環境を構築しており．正面画像を用いるものに比べ実用性が低いという問題点があった．

一方近年になり，Microsoft Kinect に代表されるような，比較的安価であり，特殊な環境設定を必要としない RGB-D カメラが実用化され，カメラからの距離を表す深度情報の取得が容易になった．特定の点だけではなく，カラー画像全体の深度情報を画素単位で得ることにより，顔の表面の凹凸を観測できるため，周囲と比較し突出している唇の輪郭抽出の精度が向上し，取得できる特徴量の正確性が増すことが期待される [10], [11].

そこで本研究では，Microsoft Kinect から得られる色情報と深度情報の双方を用いた口唇・口腔輪郭の抽出手法，ならびに奥行きの情報を含んだ画像特徴量を用いたマルチモーダル音声認識手法を提案し，日本語音声認識の実験を通じてその有用性を評価する．

本稿では，まず 2 章で提案する音声認識システムの概要について説明する．続いて 3 章で口唇・口腔の輪郭抽出手法，4 章で認識に用いる画像特徴量について説明する．そして 5 章で評価実験の詳細を，6 章でその結果と考察を報告する．最後に，7 章で本研究のまとめと今後の課題について述べる．

2. 音声認識システム

図 1 に，本研究で構築した音声認識システムの概要を示す．

音声データに関しては，通常の音声認識で広く用いられる MFCC12 次元と対数パワー，およびこれらの Δ，ΔΔ 成分の計 39 次元からなる特徴量に変換する．

カラー画像と深度画像は，深度カメラの一つである Microsoft Kinect(以下 Kinect) を用いて撮影を行う．Kinect

[1] 東京工業大学
Tokyo Institute of Technology
[2] 東京都市大学
Tokyo City University
[a)] oshio.s.aa@m.titech.ac.jp

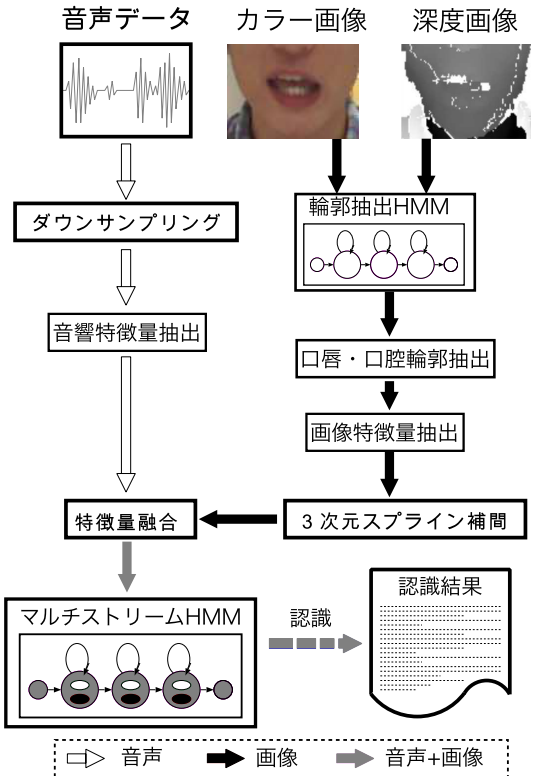

図 1　マルチモーダル音声認識システム

は，2010 年にマイクロソフトから発売されたセンサーである．元々はゲーム用の周辺機器であったが，2012 年に Windows コンピュータ向けの Kinect for Windows，および開発キットである Kinect for Windows SDK[12] が公開されたことにより，深度情報や骨格情報などをプログラム経由で容易に取得できるようになった．本研究で利用するカラー画像，深度画像の解像度は 640 × 480 であり，フレームレートはともに 30fps である．また，1 画素あたりのビット数はカラー画像で 24，深度画像で 13 である，取得できる深度値の範囲は通常は 0.8m から 4m であるが，「Near モード」を有効にすることで範囲が 0.4m から 3m に変更され，被写体がカメラにより近い状態での撮影が可能になる．深度は mm 単位で取得できるが，距離が離れるほどその分解能は低くなる．

各フレームのカラー画像・深度画像に対し，まず事前処理を施して画素列ごとに分解し，各々の列に対し口唇・口腔位置の推定を行う．推定は学習データから構築される HMM を用いて行い，列ごとの推定結果を統合することで，フレームごとの口唇・口腔の輪郭を抽出する．得られた輪郭情報と深度画像から，口腔の幅と高さ，唇の突出という口の形状を表す特徴量を抽出する．音響特徴量と画像特徴量は，3 次スプライン補間によりフレームレートをあわせた後，フレーム単位で融合される．その後，マルチストリーム HMM に基づくマルチモーダル音声認識により，融合特徴量を認識結果に変換する．

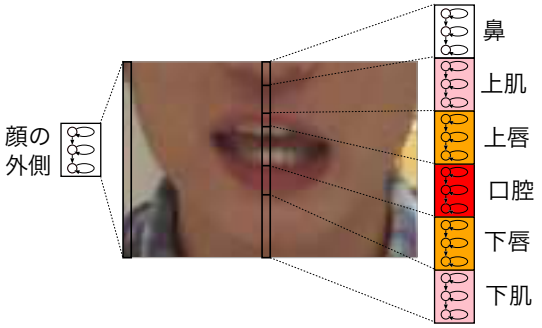

図 2　口唇・口腔位置推定用 HMM

3. 口唇・口腔の輪郭抽出手法

本研究では，田村ら [3], [13] の HMM による口唇・口腔輪郭の抽出手法を，カラー画像とともに深度画像を用いるものに拡張した．本章では，この抽出手法の詳細について説明する．

3.1 画像データの事前処理

撮影する画像データは顔全体を含むものであるため，まず Kinect for Windows SDK の機能である Face Tracking SDK[14] を用いてフレームごとに顔の下半分の領域を抽出し，バイキュービック法により大きさを 120×80 ピクセルに統一する．なお，顔の検出に失敗したフレームについては，前フレームにおける抽出位置と同じ領域を抽出する．その後，カラー画像と深度画像の座標補正を行うことで，各画素に対応する距離情報を正しく取得できるようにする．

3.2 輪郭抽出用パラメータ抽出

まず，カラー画像と深度画像それぞれから輪郭抽出のためのパラメータを抽出する．

カラー画像は，RGB から HSV に変換し，彩度と輝度，連続性を持たせるために sin および cos 関数で変換した色相値をパラメータとして抽出し用いる．

深度画像は，撮影日の違いなどで生じる被写体とカメラとの距離の変動を排除するため，まず最大値と最小値を上下限とした正規化を行う．その後，凹凸による影響で赤外線が届かず深度情報を取得できなかった箇所を縦方向の 3 次スプライン補間により補う．そして，“1 ピクセル上の画素の深度値 − 現在の画素の深度値” の値，すなわち縦方向の凹凸変化量をパラメータとして抽出する．

以上の 5 要素を統合し，さらにこれらの 1 次微分成分を加えて列ごとの輪郭抽出用パラメータを生成する．

3.3 HMM による口唇・口腔位置推定

生成されたパラメータ列を用いて，口唇・口腔位置推定用の HMM を構築する．構築する HMM は，鼻，上肌，上唇，口腔，下唇，下肌，顔の外の 7 種類であり，全て状態

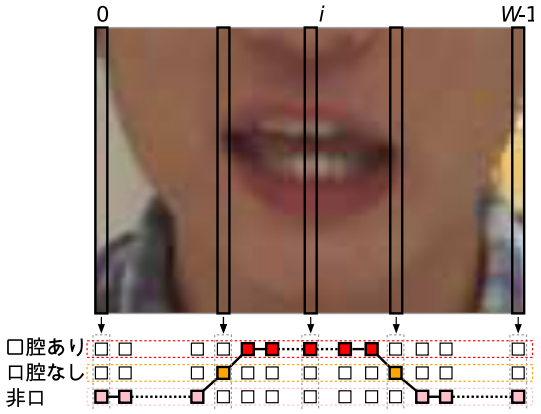

図 3　DP による横方向の状態推定

数 3，混合数 8 とする (図 2).

HMM の学習には，実験に用いるデータ全体の 10%のフレームを用いる．Face Tracking SDK により唇の中心列の位置が大まかに得られるため，その列データを用いて HMM を学習する．ただし，「顔の外」HMM の学習については，各フレームの両端の列データを利用した．初期学習に必要な座標ラベルは一部のデータのみに対して手作業で付与し，残りのデータは口が開いているか閉じているかの情報を与え，これらを用いて連結学習を行った．

次に，学習した HMM を用いて，全データに対してアライメントを行い口唇及び口腔の境界を列ごとに求める．各画像列は，画像の上から順に

- 上肌→下肌 (→顔の外)(非口領域)
- (鼻→) 上肌→上唇→下唇→下肌 (口腔なし口領域)
- (鼻→) 上肌→上唇→口腔→下唇→下肌 (口腔あり口領域)

のいずれかの並びをとると仮定できるため，各列に対し上記の 3 種類すべてのアライメント結果を求めておき，同時に，それぞれの並びをとった際の尤度値を求めておく．

その後，実際に各列がどの並びを取るかを推定するために，尤度値を用いたワンパス DP を適用し，「非口領域→口腔なし口領域→非口領域」または「非口領域→口腔なし口領域→口腔あり口領域→口腔なし口領域→非口領域」のいずれかの状態遷移を満たし，かつ全体の尤度を最大にするような状態分類を取得する (図 3).

この DP の結果と HMM によるアライメントの結果を統合し，最終的な口唇・口腔の輪郭を決定する．

4. 画像特徴量の抽出

抽出した口唇・口腔の輪郭情報から，音声認識に用いる画像特徴量を計算する．本研究では，従来用いられてきた口腔の幅，高さの値 [13](図 4) に加え，深度画像から得られる唇の凹凸を表す特徴量を導入した．

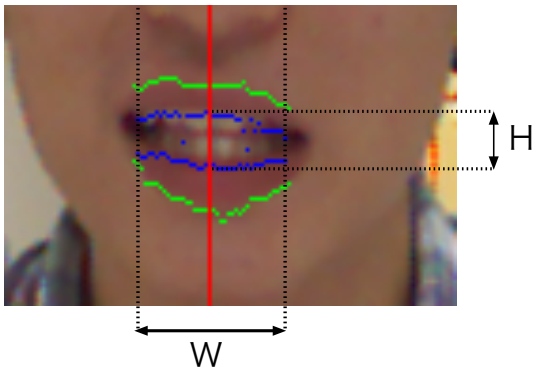

図 4　口腔の幅，高さの抽出

図 5　深度値ベクトルの平均

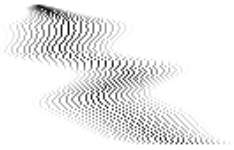

図 6　第 1 固有ベクトル

図 7　第 2 固有ベクトル

4.1 高さ

DP マッチングの結果「口腔なし口領域」または「口腔あり口領域」に分類された列の中で中央にあたる列を唇の中心とみなし，その列が「口腔あり口領域」に分類されていた場合は口腔のピクセル数を高さ H とし，「口腔なし口領域」に分類されていた場合は口が閉じていると判断し，$H = 0$ として扱う．

4.2 幅

DP マッチングの結果，「口腔あり」に分類された画素列の数を口腔の幅 W として抽出する．ただし，「口腔あり」に該当する列が存在しない場合は口が完全に閉じていると判断し，$W = 0$ として扱う．また，$H = 0$ となる場合は，口腔領域が検出されてもそれが誤検出である可能性が高いため $W = 0$ として扱うことにしている．

4.3 凹凸

より詳細な唇形状を取得するための手段として，カラー画像からは取得できない，唇の突出を表す特徴量を深度画像から抽出する．

まず，唇中心列の深度値のみに対して，最小値が 0, 最

大値が 100 となるような標準化を行う．この列において，深度値が取得できていない場所は口腔内であることがほとんどであるため，その範囲については輪郭抽出時のような補間処理を行わず，その列内で最も奥にある，すなわち値 100 を取ると仮定する．得られた深度値の列ベクトルの平均の様子を図 5 に示す．

続いて，得られた深度値列を入力として PCA を適用し，スコアを計算する．ここでの固有ベクトルは学習データ全体の内 10 得られた第 1, 第 2 固有ベクトルの様子を図 6, 7 にそれぞれ示す．画像特徴量には，第 2 固有ベクトルに対応するスコアを利用する．第 1 固有ベクトルに対応するスコアを用いないのは，このスコアによる変化が顔の形状そのものの違いを表すものであると考えられ，口唇近くの領域のみを考える本研究においては認識率を下げる要因なると判断したためである．

5. 評価実験

本研究で示される画像特徴量の有効性を測るため，日本語音声認識による評価実験を行った．本章では，実験設定の詳細について説明する．

5.1 認識タスク

本研究では，大語彙連続音声認識デコーダ Julius[15] を用いた日本語の連続音声認識タスクにより提案する画像特徴量の評価を行う．言語モデルは，毎日新聞記事から抽出された単語約 6 万語を用いて構築された，前向き 2-gram, 逆向き 3-gram のものを用いる．また，正解となる文章の単語分割は，形態素解析ツール ChaSen（茶筅）を用いて行った．

5.2 データ収録

従来までの音声認識の研究において，カラー画像と深度画像をともに含む日本語音声データベースが存在しないため，本研究では Microsoft Kinect を用いて独自にデータの収録を行った．発話内容は ATR 音素バランス文であり，2 名分の発話 (話者 A は 503 文，話者 B は 200 文) を収録した．

収録は研究室内にて行った．そのため，幾つかの文に生活雑音が入っている．音声収録は iPhone4 を用いて，iOS7.0.4 に標準搭載されているボイスメモアプリによる録音を行った．これにより得られる音声は周波数 44.1kHz, 量子化ビット 16bit であり，特徴量抽出の際は周波数を 16kHz にダウンサンプリングしたものを用いる．

カラー画像及び対応する深度画像の撮影は，Kinect Studio [*1] を用いて行い，その後非圧縮 avi ファイルに変換した．音声と画像の時間同期は，Kinect Studio の撮影を終了する瞬間のクリック音を使って行い，カラー画像と深度画像の時間同期は，Kinect for Windows SDK が取得するフレームごとのタイムスタンプの値を用いて行った．Kinect と話者の顔の表面との間は 65cm から 90cm の間に収まるようにし，背景の深度値の影響を無くすため，背景となる壁と Kinect の間は Near モードの取得範囲外である 3m 以上の距離を取った．また，太陽光による照明や赤外線センサーへの影響を取り除くため，撮影は全て日没後に行った．

Kinect は文章を表示するディスプレイの上に設置し，その真横にアプリを起動させた iPhone を設置した．

5.3 学習・評価データ

本研究では，不特定話者のデータに対する頑健性を測るため，話者 A の発話 503 文を用いるタスクと，話者 A と話者 B の発話 200 文ずつを用いるタスクの 2 種類を行った．

実験の際は，データ全体を話者 1 人の場合は 5 分割，話者 2 名の場合は 4 分割し，クロスバリデーションにより評価を行う．そして，全文の正解数および誤り数を集計し，全体の単語正解精度を算出した．

5.4 画像特徴量

実際に使用する画像特徴量としては，深度画像の有用性を検証するため，

(1) 画像特徴量なし
(2) 深度画像を用いずに抽出した幅および高さ
(3) 深度画像を組み合わせて抽出した幅および高さ
(4) (3)+凹凸特徴量

の 4 種類を用意し，それぞれの性能を比較した．なお (1) 以外は，Δ，ΔΔ 成分も特徴量として用いている．

5.5 音響・画像モデル

本研究では，マルチストリーム HMM を用いて音響情報と画像情報の融合を行う．HMM は monophone 単位で構築し，音響ストリームは状態数 3 混合数 8，画像ストリームは状態数 3 混合数 3 でモデル化する．画像情報単体では音素境界の推定精度が低いため，画像ストリームの学習の際は，事前に学習した音響ストリームの HMM を用いて学習データに対する強制切り出しを行い，時間情報つき音素ラベルを作成して，モデル学習に利用する．得られたそれぞれの HMM を，対応する音素の対応する状態ごとに融合することでマルチストリーム HMM を構築する．認識の際は，実験条件ごとに音響ストリームと画像ストリームの重みを変化させ，正解精度が最大となるように最適化する．その際，両者の重みの合計が 1.0 になるような制約を設けた．なお，今回の実験では，ストリーム重みを音素ごとに最適化せず，全ての音素で同じ値を共有して最適化を行っている．

[*1] Kinect ソフトウェアのデバッグを主目的とした RGB 画像と深度画像の録画ソフト [16].

表 1　単語正解精度
Table 1　Word Accuracy

画像特徴量	(1)	(2)	(3)	(4)
話者 1 名	75.3	78.3	78.3	78.5
話者 2 名	61.7	65.9	66.0	67.0

表 2　音素認識率
Table 2　Phoneme Accuracy

画像特徴量	(1)	(3)	(4)
話者 1 名	83.8	85.9	86.0
話者 2 名	77.5	79.5	79.6

6. 実験結果と考察

6.1 結果

表 1 に実験結果を示す．表中の番号は，5.4 節の各画像特徴量の種類を表している．

文献 [17] で示された検定法を用いて危険率 5%での有意差検定を行った．その結果，特徴量抽出段階で深度情報を用いることによる認識性能の向上は有意であるとは言えないが，一方で，特徴量として深度情報を用いた場合は，複数話者のデータを用いたタスクにおいて有意な性能向上が確認された．このことから，本研究で提案した深度画像からの特徴量は，唇形状の個人差が存在する場合においても有効に機能することが示された．

6.2 考察

音素ごとの認識率を確認するため，特徴量 (1)，(3)，(4) を利用した実験に対して，音素間の境界情報を用いてデータを分割し，それらに対して単音素認識を行った．音素単位の境界情報は，手動での付与が困難であるため，あらかじめ評価用のデータも含めた全ての音響データで HMM を学習し，それを用いた強制アライメントを行うことで得ている．音素数は，発話前後や発話中に生じる無音部分および発話中一度も現れなかった音素/dy/を除いた 39 種類とし，実験は先述の連続音声認識と同様のクロスバリデーションで行った．実験の結果を表 2 に示す．

検定の結果，画像情報を用いない (1) に比べ，(3) や (4) といった画像特徴量を用いたときには，認識性能が有意に向上していたが，(3) と (4) の間には全体としての有意差は確認されなかった．音素単位での認識率を比較したところ，口を尖らせる動きを伴う音素が特に向上しており，特に/o:/は 1.5%，/u:/は 0.8%というように，口を尖らせる長母音は認識率が上昇しており，特に短母音に誤認識される割合が減少していた．

また，話者 2 名のタスクにおいては/n/(子音)，/N/(「ん」)，がそれぞれ 0.8%，/t/が 0.7%，話者 1 名の場合にはそれらに加え/d/が 1.0%認識率が向上していた．これらの音素は，瞬間的に舌を上下させる動きが伴うという特徴があり [6]，発音中唇はほとんど動かないため，口腔の幅や高さといった情報だけで判別することが難しい．一方，深度画像を用いた場合は，発音の過程で一瞬口腔が舌で塞がれるため，深度値が比較的小さい値になる瞬間を観測できる．そのため，他の音素との区別が容易になったと考えられる．

本研究の特徴量はこれらの形状に対しては有効であったが，音素/a/に関しては両タスクに対して認識率が減少していた．/a/のように発音の際に口腔が大きく開く音素の場合に，Kinect で口腔内の深度値が取得できないケースが多く見られたことから，それによる影響が性能劣化の原因になったものと考えられる．

7. おわりに

本研究では，Microsoft Kinect を用いたマルチモーダル音声認識システムとして，唇の凹凸を考慮した輪郭抽出手法ならびに，口唇の突出を表現する特徴量を音声認識に用いる手法を提案した．日本語文音声に対する認識実験を行った結果，複数の話者データを用いるタスクで凹凸の特徴量を導入した際に有意な性能の向上がみられ，その有効性が示された．

また，音素単位での認識率を確認したところ，提案した特徴量は，口を尖らせる形をとる音素や，口腔を舌で塞ぐ動きをする音素に対して特に有効であることが確認された．その一方で，日本語の音声認識において重要となる母音に対して認識率の低下が見られるという問題点も存在することがわかった．

今回は HMM による輪郭抽出を画像全体に対して行っているため，画像特徴量抽出に膨大な時間がかかるという点も問題点として挙げられる．現状では実時間での認識が達成されていないため，実用化のためにはより高速な特徴量抽出手法を確立することが今後の課題となる．

本研究で検証に用いたデータは話者 2 名分からなるものであり，はっきりとした口調で，正面を向いて話したものである．話者数を増やした場合や自然な口調で発話した場合，顔の動きの変化が大きい場合など，様々な状況への対処法の導入，および性能の検証が必要となる．

また，音響情報と同様に，同じ音素であっても口唇や口腔の形状は前後の音素によって変化することがある [6]．音声認識用の HMM を音響，画像ストリームともに triphone で構築することにより，より高い認識性能を獲得できることが期待される．

参考文献

[1] Dupont, S. and Luettin, J.: Audio-visual speech modeling for continuous speech recognition, *Multimedia, IEEE Transactions on*, Vol. 2, No. 3, pp. 141–151 (2000).
[2] Potamianos, G., Neti, C., Luettin, J. and Matthews, I.: Audio-visual automatic speech recognition: An

overview, *Issues in visual and audio-visual speech processing*, Vol. 22, p. 23 (2004).
[3] 田村哲嗣，岩野公司，古井貞熙：マルチモーダル音声認識のための画像特徴量の改善，日本音響学会 2003 年春季講演論文集， Vol. 4, No. 1, pp. 195–196 (2003).
[4] 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，画像の認識・理解シンポジウム, MIRU2010, IS3-31，pp. 1771–1778 (2010).
[5] Tariquzzaman, M., Gyu, S. M., Young, K. J., You, N. S. and Rashid, M.: Performance Improvement of Audio-Visual Speech Recognition with Optimal Reliability Fusion, *Internet Computing & Information Services (ICICIS), 2011 International Conference on*, IEEE, pp. 203–206 (2011).
[6] Fukuda, Y. and Hiki, S.: Characteristics of the mouth shape in the production of Japanese-Stroboscopic observation., *Journal of the Acoustical Society of Japan (E)*, Vol. 3, No. 2, pp. 75–91 (1982).
[7] 大浦央子，川波弘道，李　晃伸，猿渡　洋，鹿野清宏：発話における口唇運動の三次元座標を用いた音声認識，日本音響学会 2003 年秋期講演論文集， Vol. 9, No. 3-Q-26, pp. 177–178 (2003).
[8] Kumar, K., Chen, T. and Stern, R. M.: Profile view lip reading, *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, Vol. 4, IEEE, pp. IV–429 (2007).
[9] 吉永智明，田村哲嗣，岩野公司，古井貞熙：横顔の動画像情報を用いたマルチモーダル音声認識，情報処理学会研究報告, 2003-SLP-46-11， Vol. 2003, No. 58, pp. 61–66 (2003).
[10] Galatas, G., Potamianos, G. and Makedon, F.: Audio-visual speech recognition incorporating facial depth information captured by the Kinect, *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, IEEE, pp. 2714–2717 (2012).
[11] Yargic, A. and Dogan, M.: A lip reading application on MS Kinect camera, *Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium on*, IEEE, pp. 1–5 (2013).
[12] Microsoft: Kinect for Windows SDK, Microsoft Developer Network (online), available from ⟨http://msdn.microsoft.com/en-us/library/hh855347.aspx⟩ (accessed 2014-03-10).
[13] 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， Vol. 9, No. 3-6-11, pp. 123–124 (2003).
[14] Microsoft: Face Tracking, Microsoft Developer Network (online), available from ⟨http://msdn.microsoft.com/en-us/library/jj130970.aspx⟩ (accessed 2014-03-10).
[15] 李　晃伸，河原達也，堂下修司：単語トレリスインデックスを用いた大語彙連続音声認識エンジン JULIUS，電子情報通信学会技術研究報告. SP, 音声， Vol. 98, No. 32, pp. 17–24 (1998).
[16] Microsoft: Kinect Studio, Microsoft Developer Network (online), available from ⟨http://msdn.microsoft.com/en-us/library/hh855389.aspx⟩ (accessed 2014-03-10).
[17] 中川聖一，高木英行：パターン認識における有意差検定と音声認識システムの評価法，日本音響学会誌，Vol. 50, No. 10, pp. 849–854 (1994).