

複数口唇領域を利用した多言語に有効な単語読唇

齊藤 剛史[†] 森下 和敏[†] 小西 亮介[†]

[†] 鳥取大学大学院工学研究科, 〒680-8552 鳥取市湖山町南 4-101

E-mail: †{saitoh,konishi}@ele.tottori-u.ac.jp

あらまし これまで読唇に関する研究では、特定の言語を対象とする報告がほとんどであり、言語と読唇手法に関する言及はない。さらに従来手法の多くは唇の外側輪郭領域あるいは口内領域を用いており、歯や舌の情報が特微量に反映されていない。本論文では、多言語に対して有効な読唇手法について言及する。まず、Active Appearance Model を適用し、唇の外側と内側の輪郭を同時に抽出する。次に歯領域と口内領域を抽出する。従来提案されている様々な特微量を計測して比較実験を行う。本論文では4カ国語を対象とし、各言語において20単語の発話シーンを撮影して認識実験を行った。その結果、内側唇輪郭の面積とアスペクト比と口内領域の面積の3形状に基づくトラジェクトリ特微量が、従来手法やその他の領域を用いた場合よりも高い認識率93.6%を得ることを確認した。

キーワード 読唇, 単語認識, 日本語, 中国語, モンゴル語, ネパール語

Efficient Lip Reading Method to Various Languages using Multiple Lip Regions

Takeshi SAITOH[†], Kazutoshi MORISHITA[†], and Ryosuke KONISHI[†]

[†] Tottori University, Tottori

E-mail: †{saitoh,konishi}@ele.tottori-u.ac.jp

Abstract The traditional researches targeted at only one language, and there is no research to refer the language and recognition method. Moreover, a lot of model-based methods use only an external lip or intraoral region, and tooth or tongue region is not reflected to the feature. This paper describes analysis of efficient lip reading method for various languages. First, we apply Active Appearance Model, and simultaneously extract the external and internal lip contour. Then, the tooth and intraoral regions are detected. Various features from five regions are fed to the recognition process. We set four languages to be the recognition target, and recorded twenty words per each language. As the result, proposed trajectory feature based on three shape features, the area and aspect ratio of internal lip region, and area of intraoral region, was obtained the highest recognition rates of 93.6%, compared with the traditional methods and other regions.

Key words lip reading, word recognition, Japanese, Chinese, Mongolian, Nepalese

1. はじめに

マン・マシンインタフェースにおける入力手段の一つとして音声がある。音声は、人間にとって最も自然な意思伝達手段である。音声認識に関する研究は、これまで盛んに取り組まれており、ゲーム機、カーナビゲーションシステム、携帯電話等で音声認識が実用されるようになってきている。これらの用途では、高い認識精度での

音声認識が可能になってきている。しかし、高騒音環境下や公共の場所で声を出せない場面、発声が困難な障害者など、明瞭な音声を得ることできない状態での音声認識は困難である。ここで音声の発声時には、唇の動きが生じる。唇の動きは、音響的な雑音に影響を受けないため、高騒音環境下でも情報が変わらない、声を出さなくてもよいなどの利点をもつ。読唇に関する研究では、正面の顔画像を用いる手法[1]~[12]、側面の顔画像を用い

る手法[13],[14], 音声情報と統合する手法[14]~[18] などが提案されている。

これまで読唇に関する多くの文献は、特定の言語に対する実験のみである。例えば日本語[2]~[8],[14]~[16], 英語[1],[13],[17],[18], フランス語[9],[10], 韓国語[11], ポルトガル語[12] などがある。しかし、言語と認識手法に関して言及した報告はない。

画像情報に基づく読唇では特徴量の計測が重要であり、これには画像ベース法[2]~[4],[9]~[11],[18]とモデルベース法[1],[5]~[8],[12]~[17]がある。前者は唇などの正確な輪郭情報を必要とせず口唇周辺領域の画像情報を利用するため、唇形状だけでなく歯や舌の情報を含めた特徴量を取得できる利点がある。しかし、画素情報に基づくため多量のデータが必要であり、領域サイズや撮影環境における明暗の影響を受けやすい問題がある。一方、後者は2値化処理や動的輪郭モデルなどにより口唇領域のモデルを計測し、口唇領域の幅や面積などの少ない情報でモデルを表現できる利点がある。唇輪郭を用いることより正確な認識が行える。しかしこれまでの報告の多くは、唇の外側輪郭あるいは口内領域などの形状しか用いられておらず、歯や舌の情報が特徴量に反映されていない。

以上のことより、本論文では、特定の言語だけでなく多言語において単語認識に有効な特徴量を検討する。輪郭形状を利用するモデルベース法を用いて、従来手法のように唇の外側輪郭あるいは口内領域だけでなく、歯領域などの複数領域を抽出する。複数領域より計測できる様々な特徴量を用いて認識を試みて、多言語における読唇に有効な領域と特徴量を検討し、言語と認識手法について言及する。

2. 複数口唇領域の自動領域抽出

2.1 動的輪郭モデル

画像から口唇領域を抽出する技術は、読唇に限らず、顔や表情の認識などのヒューマンインタフェースにおける重要な要素技術の一つである。口唇領域の抽出に用いられる主な手法としてしきい値処理を用いる手法[5],[12],[13],[15]と動的輪郭モデルを適用する手法[6],[8],[17]がある。動的輪郭モデルには Kass らが提案した Snakes[19], Cootes らによって提案された Active Shape Model (ASM) および Active Appearance Model (AAM) [20] など様々な手法が提案されている。

しきい値処理は、シンプルな手法であるが、光源環境の変化に弱い問題がある。

Snakes は画像エネルギーを最小化しようその収縮動作を決定するため、収縮を繰り返すたびに各制御点につ

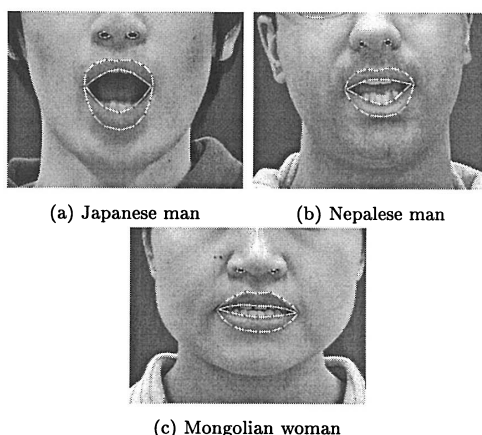


図1 Extracted contour results.

いてエネルギー最小化問題を解く必要がある。Snakes を用いた輪郭抽出法の多くは、輪郭上のエッジ検出に基づいてエネルギー関数を定義しているが、それらのエッジが不明瞭の場合に安定に抽出することが困難な場合がある。この問題を解決するために、色分布間の分離度を用いる手法[21]などが提案されている。

一方、ASM は事前に複数個の目的物体形状を学習させ、そのデータを制御点の動作に活用することで、効率的に輪郭抽出を行う動的輪郭モデルである。ASM は唇の外側と内側の輪郭など、複数の輪郭を同時に抽出することが可能である。また、学習データに依存する輪郭抽出が可能であり、その点で抽出後の形状が保障されない他の動的輪郭モデルと大きく異なるため、抽出精度の向上も期待される。ASM は画像上で顔や臓器などの平均的な形状から様々に変形する領域に関して、その形状を低次元で表現するが、AAM はさらに領域内部の明度分布を利用して、形状と同時に低次元で表現することのできる統計モデルである。これによりモデルを変形させて新しい形状を作り出し領域の抽出を行うことができる。AAM は、ASM では困難なエッジの弱い画像に対しても、明度分布を用いることにより高い抽出精度を実現できる。

以上のことより、本論文では AAM を用いて唇の外側と内側の輪郭抽出を実現する。

2.2 AAM による領域抽出

本論文で構築する AAM モデルの制御点は、図1に示す通り、唇の外側輪郭上に16点、内側輪郭上に12点、左右の鼻孔輪郭にそれぞれ6点、合計40点を与える。ここで意図的に発話者が動かず自然な状態では、鼻孔は発話中に大きな変化がほとんどない。一方、唇は発話により急激な変化が生じる。唇輪郭のみに AAM モデルを与

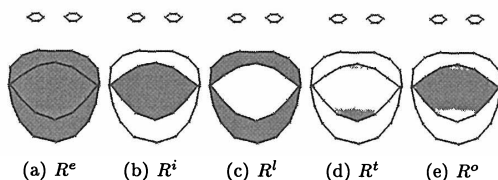


図2 Detected five regions of Fig.1(a).

える場合、発話時の変化に対応できず、抽出が失敗することがある。発話中にも安定な抽出が可能な鼻孔をAAMモデルに含むことにより、抽出精度の安定化を図る。

AAMを適用することにより、外側と内側の唇輪郭を同時に抽出できる。本論文では、これらの輪郭情報より、外側唇内領域 R^e 、内側唇輪郭内領域 R^i 、 R^e から R^i を除いた唇領域 R^l 、 R^i 内においてしきい値処理により抽出する歯領域 R^t と口内領域 R^o の5領域を定義する。歯領域 R^t の抽出に関しては次節で述べる。これらの領域には $R^e = R^i \cup R^l$ 、 $R^i = R^t \cup R^o$ の関係がある。

図1(a)の R^e 、 R^i 、 R^l 、 R^o 、 R^t の5領域を図2に示す。図中のグレー部分がそれぞれの領域を表現している。

2.3 歯領域の抽出

AAMモデルにおいて歯領域を設けられることが望ましい。しかし、閉口時だけでなく、開口時においても歯領域が見えないことがある。そこでAAMモデルに歯領域を抽出するためのモデルを組み込まず、内側唇輪郭領域 R^i 内に対して、次に述べるしきい値処理を適用することにより歯領域 R^t を抽出する。ここで、本論文では、口唇領域が及ぼす認識精度の影響を検討することを主眼としており、光源環境の変化に対する認識精度の影響を議論しない。そのため、実験で用いる画像は同一環境下で撮影する。これより本処理では、単純なしきい値処理を適用するだけでも歯領域が正しく抽出されていることを目視により確認している。

R^t を抽出するために、まず R^i 内の全画素をRGB色空間からHLS色空間に変換する。座標 (x, y) における L 値を $L(x, y)$ と表記するとき、 $L(x, y) < L_t$ の場合は R^o 、 $L(x, y) \geq L_t$ の場合は R^t に属させる。 L_t は経験的に0.2を与えた。

また舌の情報については、目視で観測しても、明確な領域の抽出は困難であった。このため、舌領域は R^o に含ませることとする。

2.4 領域サイズと傾きに対する正規化処理

本研究では、発話シーンを撮影するために、被験者に単語を発話させる際、椅子に着席させて、かつカメラを三脚で固定して撮影する。被験者には自然な姿勢で発話させる。自然な姿勢で発話しているにもかかわらず発話

時の被験者の顔のわずかな動きにより、画像中の口唇領域の位置とサイズ、傾きが異なる場合がある。位置に関しては、次節で述べる特徴量は不変であるが、領域のサイズと傾きの違いは特徴量に影響を与える。そこで、サイズと傾きに対する正規化処理を適用する。

本研究では、被験者が発話する前後に、唇を閉じるように指示している。また概して唇は横長であり、かつ発話時の唇の変化は縦方向の方が横方向より大きい。またAAMにより輪郭を抽出しているため、唇の左右端点の位置は既知である。そこで初期数フレームにおける左右端点の距離（唇の横幅）より平均横幅 W を求め、更に標準横幅 W^* を与え、両者のスケール比 $s = W/W^*$ を求める。また左右端点の傾き θ を算出する。サイズと傾きに対する正規化は s, θ を用いる。ただし、ここでは $W^* = 200$ pixel とする。

2.5 発話区間検出

前述の通り、本研究では、発話開始前後の口は閉じられている。そのため口の開閉情報を利用することにより発話区間の抽出が可能である。発話区間の抽出に関しては先行文献[8]で提案した唇の高さの変化を利用する。

発話区間の抽出は、外側唇輪郭内領域 R^e において、初期数フレームの高さの平均値 $\bar{h}$ を求める。そして、高さが $\bar{h}$ より大きくなった時を発話開始点とする。発話終了点は、 $\bar{h}$ との差が数フレーム連続してしきい値以下になるフレームとする。

3. 特徴量と認識手法

荻原らはしきい値処理により口内領域を抽出し、口内面積と開口部分のアスペクト比を特徴量として用いている[15]。両特徴量を入力として隠れマルコフモデル(HMM)による認識を報告している。李は口唇内側のアスペクト比のみを用いて、部分空間法による認識を行っている[5]。

Silveira らはしきい値処理によって得られた唇領域をもとに、四つの長さ特徴量 (W, H_1, H_2, H_3) を提案している[12]。 W は唇領域の横幅、 H_1, H_2, H_3 は W を4等分した位置における垂直方向の長さである。本論文では、この特徴量を WH と記す。三つの縦幅を用いるため正確な輪郭情報が必要である。一方、Kumar らはしきい値処理により抽出した唇領域をもとに、三つの長さ特徴量 (W, H_u, H_l) を提案している[13]。 W はSilveira らと同じ唇領域の横幅、 H_u と H_l はそれぞれ W を基準とした唇上側と唇下側の縦幅である。3特徴量を用いてHMMを適用した認識手法を提案している。

菅原らは唇輪郭形状に対して、輪郭点の平均座標を中心点 C として、 C から 30° ごとに直線を引き、これと

輪郭線との交点までの距離 $r_i, i = 0, 1, \dots, 11$ を特徴量（半径特徴量と呼ぶ）として用いている [6]。本論文では特徴量 r_i を半径特徴量 r と記す。認識には HMM を用いている。12 個の値を用いることにより、唇輪郭の動きを表現するが、特徴量数が多い。

3.1 形状特徴量

モデルベース法において、多くの従来手法は、唇輪郭あるいは口内領域のいずれか 1 領域をもとに特徴量を抽出し認識に用いていた。一方、本論文では 5 領域を抽出している。これら 5 領域のうち、外側唇輪郭内領域 R^e は従来手法で用いられてきた唇領域であり、口内領域 R^o は従来手法で用いられてきた口内領域に対応する。本論文では、これまで考慮されていなかった、その他の 3 領域を含めて下記に示す 17 個の形状特徴量を定義する。

- 外側唇輪郭内領域 R^e : 横幅 W^e , 縦幅 H^e , 面積 S^e , アスペクト比 A^e , 長さ特徴量 WH^e , 半径特徴量 r^e
- 内側唇輪郭内領域 R^i : 横幅 W^i , 縦幅 H^i , 面積 S^i , アスペクト比 A^i , 長さ特徴量 WH^i , 半径特徴量 r^i
- 唇領域 R^l : 面積 S^l
- 口内領域 R^o : 面積 S^o , アスペクト比 A^o
- 歯領域 R^d : 面積 S^d , アスペクト比 A^d

3.2 トラジェクトリ特徴量

著者らは読唇に関する先行研究で、日本語単音および日本語 10 単語に有効な特徴量として、唇領域あるいは口内領域に基づくトラジェクトリ特徴量 (TF) を提案した [8], [22]。TF は、フレーム毎に計測される特徴量の時間的変化を特徴量空間上にプロットし、時間的変化を軌道として表現した特徴量を提案する。ただしフレーム数だけでプロットすることは、情報量の不足のために認識率の低下を誘発するため、プロット点間を B-Spline で補間している。本論文でも TF を用いる。ただし、文献 [8], [22] では処理対象領域として唇輪郭と口内領域の 2 領域を用いている。一方、本論文では前述のとおり 5 領域を用いる。そこで 2 次元だけでなく、もう 1 次元を加えた 3 次元 TF を提案する。4 次元以上の TF を定義することも可能であるが、この場合、TF の特長である発話により生じる特徴量の時間的変化を口形の軌道として表現することが困難であるため、ここでは次元数を 3 までにする。

図 3 に日本語 “こんにちは” の発話シーンにより得られた 3 次元 TF を示す。図 3 の 3 軸、すなわち 3 特徴量は、4. で後述する認識実験において、最も高い認識率を得られた組み合わせ S^i, A^i, S^o である。図中、○印は各フレームにおける特徴量の値、太い実線は B-Spline 曲線で補間した値である。時間の流れを矢印①～⑦で示し

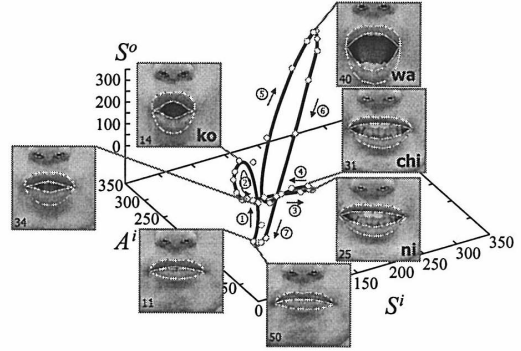


図 3 Trajectory feature (word: “ko n ni chi wa”).

ている。また発話において目視で判断した特徴的な 7 フレーム画像を図に示しており、各フレーム画像の左下の数値は、フレーム番号を意味する。フレーム番号 11 が発話開始フレーム、フレーム番号 50 が発話終了フレームである。フレーム番号 14, 25, 31, 40 でそれぞれ ‘こ’, ‘に’, ‘ち’, ‘わ’ を発声しているフレームである。この図より、発話フレームは曲線のコーナーに位置する傾向にあることが観測できる。

3.3 認識手法

読唇に関する従来手法の多くは HMM を用いた手法が提案されており、本論文でも形状特徴量を用いた認識手法として HMM を適用する。

TF を用いた認識手法として、文献 [8] と同様に DP マッチングを適用する。ただし、2 次元 TF であれば 2 次元 DP マッチング、3 次元 TF であれば 3 次元 DP マッチングを適用する。具体的には、処理対象の未知単語トラジェクトリ $\mathbf{X} = \{x_1, x_2, \dots, x_J\}$ とデータベースに登録されている単語のトラジェクトリ $\mathbf{R}_n = \{r_1, r_2, \dots, r_J\}$ を用いて、二つのトラジェクトリ $\mathbf{X}$ と $\mathbf{R}_n$ の距離 $D(\mathbf{X}, \mathbf{R}_n)$ を求める。 $\hat{n} = \arg \min_n D(\mathbf{X}, \mathbf{R}_n)$ を満たす単語 $\hat{n}$ を結果とする [23]。ここで x_j, r_i は、2 次元ベクトルまたは 3 次元ベクトルとする。DP マッチングは、次のように初期値を設定する。

$$\begin{cases} g(i, 0) = 0 & (i = 0, 1, \dots, I) \\ g(0, j) = \infty & (j = 1, 2, \dots, J) \end{cases}$$

各格子点 (i, j) における累積距離 $g(i, j)$ を次式により求める。

$$g(i, j) = \min \begin{cases} g(i-1, j) + d(i, j) \\ g(i-1, j-1) + 2d(i, j) \\ g(i, j-1) + d(i, j) \end{cases}$$

ただし、局所距離 $d(i, j)$ はユークリッド距離とする。これより、 $D(\mathbf{R}_n, \mathbf{X}) = g(I, J)/(I+J)$ で求まる。

Japanese	Nepalese	Chinese	Mongolian
こんにちは	नमस्ते	你好	ᠵᠢᠰᠤᠨᠠᠨᠤ
おはよう	शुभप्रभात	早上好	ᠵᠢᠰᠤᠨᠠᠨᠤ ᠵᠢᠰᠤᠨᠠᠨᠤ
おやすみなさい	शुभरात्री	晚安	ᠵᠢᠰᠤᠨᠠᠨᠤ ᠵᠢᠰᠤᠨᠠᠨᠤ
ありがとう	धन्यवाद	谢谢	ᠵᠢᠰᠤᠨᠠᠨᠤ ᠵᠢᠰᠤᠨᠠᠨᠤ
おめでとう	बधाई छ	祝贺	ᠵᠢᠰᠤᠨᠠᠨᠤ ᠵᠢᠰᠤᠨᠠᠨᠤ
行く	जान्छु	去	ᠵᠢᠰᠤᠨᠠᠨᠤ
来る	आउँछु	来	ᠵᠢᠰᠤᠨᠠᠨᠤ
教える	सिकाउँछु	教	ᠵᠢᠰᠤᠨᠠᠨᠤ
分かる	बुझ्छु	明白	ᠵᠢᠰᠤᠨᠠᠨᠤ
遊ぶ	खेल्छु	玩	ᠵᠢᠰᠤᠨᠠᠨᠤ
食べる	खान्छु	吃	ᠵᠢᠰᠤᠨᠠᠨᠤ
飲む	पिउँछु	喝	ᠵᠢᠰᠤᠨᠠᠨᠤ
どこ	कहाँ	哪里	ᠵᠢᠰᠤᠨᠠᠨᠤ
時間	समय	时间	ᠵᠢᠰᠤᠨᠠᠨᠤ
場所	ठाउँ	地点	ᠵᠢᠰᠤᠨᠠᠨᠤ
いますか	छ?	在吗	ᠵᠢᠰᠤᠨᠠᠨᠤ
いません	छैन	不在	ᠵᠢᠰᠤᠨᠠᠨᠤ
いいですか	हुन्छ?	好吗	ᠵᠢᠰᠤᠨᠠᠨᠤ
してください	गर्नुस्	请作	ᠵᠢᠰᠤᠨᠠᠨᠤ
また会いましょう	फेरि भेटौंला	再会	ᠵᠢᠰᠤᠨᠠᠨᠤ ᠵᠢᠰᠤᠨᠠᠨᠤ

図 4 Target words.

4. 単語認識実験

本実験では、日本人成人男性、ネパール人成人男性とモンゴル人成人女性の 3 人に協力を得て、認識対象言語として日本語、ネパール語、中国語とモンゴル語の 4 カ国語の発話シーンを撮影した。モンゴル人女性はモンゴル語と中国語の 2 カ国語の発話可能であったため、両言語の発話シーンを撮影した。認識対象の単語は、図 4 に示す電話の会話で利用される 20 単語とした。各単語 10 サンプル、1 言語につき 200 サンプルの発話シーンをデジタルビデオカメラで撮影した。画像サイズは 640 × 480 画素、撮影時のフレームレートは 30 frame/sec. である。

発話シーンの撮影は、2.4 で述べたとおり、被験者を椅子に着席させて、三脚でビデオカメラを固定して撮影した。全ての言語は同一環境下で撮影した。また発話開始前後を閉口させた。

本論文では、2.2 で述べたとおり、40 点の制御点から構成される AAM モデルを構築し、AAM を適用することにより唇領域を抽出した。このときモデルを表現する次元数は、固有値の累積寄与率 98% 以上となる次元とした。AAM モデルを学習する際に、登録した口唇形状数、すなわち画像数は、4 カ国語でそれぞれ 12~14 枚であった。その結果、ほぼ正しく輪郭が抽出されていることを目視により確認した。

発話区間検出について、検出結果の各言語のシーン毎のフレーム数の分布を図 5 に示す。日本語、ネパール語、中国語、モンゴル語の各フレーム数はそれぞれ 36.3, 43.0, 77.3, 82.1, つまり発話時間はそれぞれ 1.21, 1.43, 2.58, 2.74 秒であった。日本人とネパール人の 2 名は自

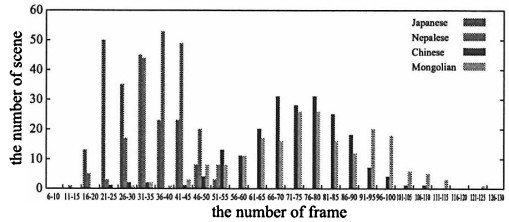


図 5 Distribution of number of utterance frames of each language.

然な速度で発話したが、モンゴル人はゆっくりと発話していたことがわかる。

4.1 認識結果

本実験では、抽出された 5 領域をもとに、様々な特徴量の組み合わせに対する認識実験を行った。形状特徴量と HMM による認識では 35 通りの形状特徴量の組み合わせ、TF と DP マッチングによる認識では 15 通りの組み合わせについて実験した。それらの中で、それぞれ 14 通りと 8 通りの代表的な組み合わせの認識結果を表 1 に示す。表中、 S^i, A^i, S^t +HMM は、三つの形状特徴量 S^i, A^i, S^t を用いて HMM により認識したことを意味し、 $TF(S^i, A^i, S^t)+DP$ は、三つの形状特徴量 S^i, A^i, S^t より求まる 3 次元 TF を用いて DP マッチングにより認識したことを意味している。HMM の実装にはケンブリッジの HMM ツールである HTK (Hidden Markov Model Tool Kit) [24] を利用した。HMM の状態数は経験的に 10 とした。また認識実験では正識別率を少数サンプルから推定するために leave-one-out 法を用いた。すなわち、学習データ 9 サンプルと認識データ 1 サンプルとして 10 通りの認識実験を行った。その結果、 $TF(S^i, A^i, S^t)+DP$ を用いた場合に最も高い 4 カ国語の平均認識率 93.6% を得た。

4 カ国語の認識率の傾向を調べると、日本語が最も高く、ネパール語と中国語がほぼ同じ認識率、モンゴル語が最も低い認識率であった。詳細を解析するため、各言語における誤認識の傾向を調べた。 $TF(S^i, A^i, S^t)+DP$ における混合行列を求めたところ、ネパール語は 20 単語中 11 単語が 100% の認識率であり、5 単語が 90%, 2 単語が 80%, 2 単語が 70% であった。中国語も同じ傾向であり、11 単語が 100%, 5 単語が 90%, 3 単語が 80%, 1 単語が 70% であった。一方、モンゴル語は 12 単語が 100% であるにもかかわらず、2 単語が 90%, 2 単語が 80%, 2 単語が 70% であり、残り 2 単語 (ローマ字表記で irehu と idehu) が 50% であった。この 2 単語は発音が似ており相互に間違えていた。日本語は他の言語よりも高い認識率を得ている。この結果と図 5 より、認識率

表 1 Recognition results

method	Japanese	Nepalese	Chinese	Mongolian	average
WH^e+HMM	99.5	73.5	79.5	67.0	79.9
WH^i+HMM	91.5	67.0	68.0	58.5	71.3
r^e+HMM	98.5	76.0	78.5	65.5	79.6
r^i+HMM	92.0	82.0	79.5	83.0	84.1
W^e, H^e+HMM	99.5	74.5	79.5	63.5	79.3
W^i, H^i+HMM	99.0	82.0	77.0	71.5	82.4
S^e, A^e+HMM	99.5	90.0	86.0	82.5	89.5
S^i, A^i+HMM	97.0	72.5	75.0	78.5	80.8
S^e, A^e+HMM	93.5	62.0	54.0	46.0	63.9
S^e, A^e+HMM	97.0	81.5	83.0	80.5	85.5
S^i, A^i+HMM	92.0	77.0	76.0	64.5	77.4
S^e, A^e, S^i+HMM	99.5	90.0	86.0	82.5	89.5
S^i, A^i, S^e+HMM	98.5	88.5	88.0	86.0	90.3
S^i, A^i, S^e+HMM	98.5	81.5	87.0	82.5	87.4
$TF(S^e, A^e)+DP$	100.0	86.5	81.0	78.0	86.4
$TF(S^i, A^i)+DP$	99.5	84.5	84.0	81.5	87.4
$TF(S^e, A^e)+DP$	94.0	63.5	69.0	67.0	73.4
$TF(S^e, A^e)+DP$	99.5	84.0	90.0	82.5	89.0
$TF(S^i, A^i)+DP$	97.0	82.0	75.5	65.0	79.9
$TF(S^e, A^e, S^i)+DP$	100.0	92.5	86.5	90.0	92.3
$TF(S^i, A^i, S^e)+DP$	100.0	91.0	89.5	91.0	92.9
$TF(S^i, A^i, S^e)+DP$	100.0	92.5	93.0	89.0	93.6

は発話時間による影響を受けないことがわかる。また日本語は他の言語に比べ口の動きが大きいために認識率が高いと推測する。

言語による認識精度の違いが見られたものの概して、我々が提案したトラジェクトリ特徴量は従来の形状特徴量よりも高い認識率を得ている。また R^e と R^i の認識率は低く、 R^e , R^i , R^o は認識に有効であることがわかった。さらに R^e あるいは R^i だけでは区別できない発話時の歯の見え隠れを表現できる S^e あるいは S^o を組み合わせることにより、さらに高い認識率が得られることを確認した。

5. おわりに

本論文では、読唇においてこれまであまり検討されていなかった多言語において単語認識に有効な特徴量の検討を試みた。さらに複数の口唇領域を用いて有効な領域の検討を行った。日本語、ネパール語、中国語とモンゴル語の4カ国語に対する認識実験を行ったところ、唇内側輪郭内領域の面積とアスペクト比、口内領域の面積の3特徴量に基づくTFが、言語に依存せずに、従来手法や他の領域特徴量に比べて最も高い認識率93.6%を得ることを確認した。

文 献

- [1] 間瀬, アレックス: “オブティカルフローを用いた読唇”, 信学論 (D-II), **J73-D-II**, 6, pp. 796–803 (1990).
- [2] 清田, 内村: “口唇周辺画像情報を用いた発話単語認識”, 信学論 (D-II), **J76-D-II**, 3, pp. 812–814 (1993).
- [3] 中田, 安藤: “色抽出法と固有空間法を用いた読唇処理”, 信学論 (D-II), **J85-D-II**, 12, pp. 1813–1822 (2002).
- [4] アラー, 鶴田, 谷口: “日本語リップリーディングシステム”, 信学技報, 第104巻, pp. 31–36 (2004).
- [5] 李, 山崎, 黒畑, 小川: “部分空間法による読唇”, 信学技

報, 第97巻, pp. 9–14 (1997).

- [6] 菅原, 新地, 岸野, 小西: “パーソナルコンピュータ上での読唇システムの実時間実現”, 計測自動制御学会論文集, **36**, 12, pp. 1145–1151 (2000).
- [7] 大槻, 大友: “オブティカルフローとHMMを用いた駅名発話画像認識の試み”, 信学技報, 第102巻, pp. 25–30 (2002).
- [8] 齊藤, 小西: “トラジェクトリ特徴量に基づく単語読唇”, 信学論 (D), **J90-D**, 4, pp. 1105–1114 (2007).
- [9] O. Vanegas, K. Tokuda and T. Kitamura: “Lip location normalized training for visual speech recognition”, IEICE Trans. Inf. & Syst., **E83-D**, 11, pp. 1969–1977 (2000).
- [10] 南角, 徳田, 北村, 小林: “隠れマルコフモデルを用いた視覚音声認識のための正規化学習”, 信学論 (D-II), **J86-D-II**, 2, pp. 163–172 (2003).
- [11] J. Kim, J. Lee and K. Shirai: “An efficient lip-reading method robust to illumination variations”, IEICE Trans. Fundamentals, **E85-A**, 9, pp. 2164–2168 (2002).
- [12] L. G. ves da Silveira, J. Facon and D. L. Borges: “Visual speech recognition: a solution from feature extraction to words classification”, XVI Brazilian Symposium on Computer Graphics and Image Processing, pp. 399–405 (2003).
- [13] K. Kumar, T. Chen and R. M. Stern: “Profile view lip reading”, IEEE International Conference on Acoustics, Speech and Signal Processing, No. 4, pp. 429–432 (2007).
- [14] 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, **2007**, (2007). doi:10.1155/2007/64506.
- [15] 荻原, 新谷, 土居, 福永: “視聴覚融合を用いたHMM音声認識”, 電学論 (C), **115**, 11, pp. 1317–1324 (1995).
- [16] 奥村, 濱口, 岡野, 宮崎: “顔画像情報と音声情報の統合による発話認識”, 情報学論, **39**, 12, pp. 3232–3241 (1998).
- [17] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox and R. Harvey: “Extraction of visual features for lipreading”, IEEE Trans. Pattern Anal. & Mach. Intell., **24**, 2, pp. 198–213 (2002).
- [18] A. V. Nefian, L. Liang, X. Pi, X. Liu and K. Murphy: “Dynamic bayesian networks for audio-visual speech recognition”, EURASIP Journal on Applied Signal Processing, **2002**, 11, pp. 1274–1288 (2002). doi:10.1155/S1110865702206083.
- [19] M. Kass, A. Witkin and D. Terzopoulos: “Snakes: active contour models”, International Journal of Computer Vision, **1**, pp. 321–331 (1988).
- [20] T. F. Cootes, G. J. Edwards and C. J. Taylor: “Active appearance models”, European Conference on Computer Vision, No. 2, pp. 484–498 (1998).
- [21] 若杉, 西浦, 山口, 福井: “色分布間の分離度を用いた唇輪郭抽出”, 信学論 (D), **J89-D-II**, 9, pp. 2025–2032 (2006).
- [22] T. Saitoh, M. Hisagi and R. Konishi: “Japanese phone recognition using lip image information”, IAPR Conference on Machine Vision Applications (MVA 2007), pp. 134–137 (2007).
- [23] 中川: “パターン情報処理”, 丸善株式会社 (1999).
- [24] “HTK hidden markov model toolkit”.
http://htk.eng.cam.ac.uk/.