4R-5

方向と距離の変化に対してロバストな唇認識

山口 健 † 駒谷 和範 † 尾形 哲也 † 奥乃 博 †

† 京都大学大学院 情報学研究科 知能情報学専攻

1. はじめに

マンマシンインターフェースにおける入力手段として音声がある。しかし音声認識は雑音に対して非常に脆弱であり、この脆弱性の克服が移動ロボットのような動的環境において必要となってくる。

この脆弱性に対する克服方法として、一つのセンサのみではなく、複数のセンサを用いる方がよいと言われている [1]。その中でもマイクからの音声情報にカメラからの画像情報を用いて補正する方法については、CHMM [2] などの統合方法など、様々な 研究がなされている。

しかし、これらの研究において用いられている映像データは人を至近距離・単一方向から撮影したものである。ロボットのように人とシステムが共に動くシステムの場合、常に人がシステムに対して至近距離から一定の方向を向いて喋っている状況は不自然である。そのため、システムへの実装を考えたとき、カメラから人への方向と距離の変化に対しての頑健性が必要となってくる。

そこで、本研究では方向の問題に対しては簡潔な唇モデルを用いて 3 次元復元を行うことで、計算量を大きく増加させることなく対処する。また、距離の問題に対しては画像の低周波数成分を用いることで対処する。

2. 提案手法

この章では本稿において提案する、方向の問題への対処のための唇モデルについてと、距離の問題に対処するための低周波数成分を用いた特徴量抽出について述べる。

2.1 唇モデル

方向の問題に関しては唇モデルを用いることで対処する。唇モデルを用いて 3 次元復元を行い、特徴量を抽出することで方向によらない安定した認識を行うことができる。ここでは計算量を減らすために唇モデルは円柱の側面上に存在する簡潔な形とした。図 1 は唇モデルを頭頂部側からみた図である。頭頂部から見た場合、唇はあ

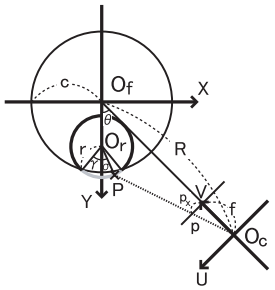

図 1: 唇モデル

る円の弧であると考えられる。ここでこの円のことを唇円とし、唇円の中心を O_r、唇円の半径を r とする。唇円上の唇がある弧の中心角を 2λ とする。唇円は顔の断面の円、顔円とは異なり顔円の中心を O_f、顔円の半径を c とする。カメラはピンホールモデルとし、顔の正面から左 θ 度の位置に顔の中心から R 離れているとする。カメラの投影中心は O_c、焦点距離は f とする。また、ワールド座標系を O_f を中心とした X,Y 座標系とし、カメラ座標系を O_c を中心とした U,V 座標系とする。

唇上の点 P を考える。P は正面から左 σ 度の位置にあり、ワールド座標系において $\mathbf{P_w} = (x,y)^T = (rcos\sigma, c - r + rsin\sigma)^T$ と表現される。一方カメラ座標系においての P は $\mathbf{P_c} = (u,v)^T$ と表現されるとすると、$\mathbf{P_c} = \mathbf{R}\mathbf{P_w} + \mathbf{T}$ と表すことができる。ここで、$\mathbf{R}$ は回転行列、$\mathbf{T}$ は並進ベクトルであり、

$$\mathbf{R} = \begin{pmatrix} -\cos\theta & \sin\theta \\ -\sin\theta & -\cos\theta \end{pmatrix}, \quad \mathbf{T} = \begin{pmatrix} 0 \\ R \end{pmatrix} \tag{1}$$

となる。この変換式によりワールド座標からカメラ座標に変換することができる。

この点 P をカメラに投影させた投影平面上の点を p とする。点 p はカメラ座標系において $\mathbf{p} = (p_x, f)^T$ と表せるので、$\mathbf{p} = \frac{f}{v}\mathbf{P_c}$ となる。

画像平面に投影された画像は横軸が p_x に比例する形で表されており、この変換式を用いることで横軸が λ に比例した画像に変換することができる。この変換式を用いて画像の見えている側半分のみを変換することで、どの方向からでも同じような画像を得ることができる。

2.2 低周波数成分を用いた特徴量抽出

距離の問題に対しては低周波数成分を用いた特徴量抽出を行うことで対処する。距離の変化によって生じる問題の中で最も認識に影響する点は、画像の粗さが増す点である。この画像の粗さによる影響を減らすために画像の低周波数成分を用いる。画像の周波数成分を用いることは JPEG 圧縮などに用いられており、さらにその中でも低周波数成分は画像の概略を表すことができる。

特徴量抽出の手順としては図 2 のような流れで行う。まず、抽出してきた唇画像に大きさに関しての正規化を

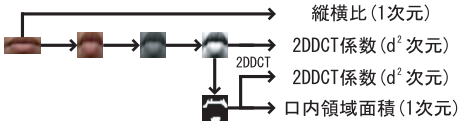

図 2: 特徴量抽出の流れ

行う。これは唇領域を 32×32 ピクセルの一定の大きさに拡大・縮小することで行う。カメラからの距離や唇の形の個人差による影響を減らすための処理である。次に輝度に関しての正規化を行う。これは輝度値頻度分布の平坦化によって行う。撮影状況における照明条件の変化による影響を減らすための処理である。

この 2 つの正規化を行った後、唇画像を輝度のある閾値によって 2 値化し、口内領域を強調する画像である口内領域画像を生成する。特徴量は、正規化を行う前の唇画像の縦横比、口内領域画像における口内領域の面積の 2 次元に加え、唇画像と口内領域画像をそれぞれ 2 次元離散コサイン変換にかけ得られる DCT 係数のうち低周波数成分を抽出し、特徴量として用いている。この特徴量を主成分分析によって次元数の圧縮を行ったものをモデルを構築するのに用いる特徴量としている。

Robustness Lip Reading against Changes of Face Direction and Distance
Takeshi Yamaguchi (Kyoto Univ.), Kazunori Komatani (Kyoto Univ.), Tetsuya Ogata (Kyoto Univ.), and Hiroshi G. Okuno (Kyoto Univ.)

3. 実験

この章では2章で提案した手法についての検証を行うために構築したシステム、そしてそのシステムを用いた実験について述べる。

3.1 実験システム

認識は図3のような流れで行う。まず、データとなる映

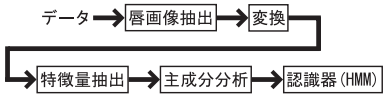

図 3: システムの流れ

像から、色情報、時系列情報を用いて唇画像を自動で切り出してくる。その唇画像に対して唇モデルを用いた変換を行い、変換後の画像から特徴量を抽出する。その特徴量を主成分分析を用いて圧縮し、その特徴量を HMM にかけて学習、認識を行う。

3.2 実験データと認識器

話者の周り4方向にカメラ（SONY EVI-G20）を設置し、その4台からの入力を4画面分割スイッチャ（SONY YS-Q440）にかけた図 4 のような映像を用いている。フレームレートは 30fps、24bit カラーであり、サイズ

図 4: 実験に用いたデータ

は 640×480 ピクセル、つまり 1 つの方向からの映像は 320×240 ピクセルである。方向は正面、左 15 度、左 30 度、左 45 度から撮影を行っており、カメラから話者の中心までの距離は 50cm である。話者は 20 代の男性 1 人であり、ATR 音素バランス単語 216 語を 1 人 5 セットずつ、1 単語毎に唇を閉じて発話している。

認識器として HMM を用いた。また、認識には唇からの特徴量のみを用いて、母音 5 種類と唇を閉じた発話していない状態を認識の最小単位として用いた。HMM は HTK を用いて実装した。HMM は 3 状態 8 混合のモノフォン、もしくはトライフォンモデルである。

3.3 方向の変化に対する実験

唇モデルによる性能向上を見るために、唇モデルを適用した場合と適用しない場合を比較した。どちらも正面のデータを学習データとし、その他の方向のデータをテストデータとして認識を行い、評価は 5-fold cross-validation で行った。モデルはモノフォンモデルとトライフォンモデルを使用している結果は図 5 のようになった。結果からわかるように、モデルによる変換を行うことで方向が変化した場合の認識率の低下を抑えることができた。ただ 45 度においてはモデルによる効果はあらわれなかった。これは唇の自動切出しの精度が悪いためと、モデルを簡潔にしたためであると考えられる。しかし、人がシステムと会話している場面においては 45 度以上方向を変化させることはあまり考えられない。そのため、このモデルは十分実用に耐えると考えられる。

3.4 距離の変化に対する実験

低周波数成分を用いた特徴量抽出の性能を見るために距離を変化させたデータを用いて実験を行った。今

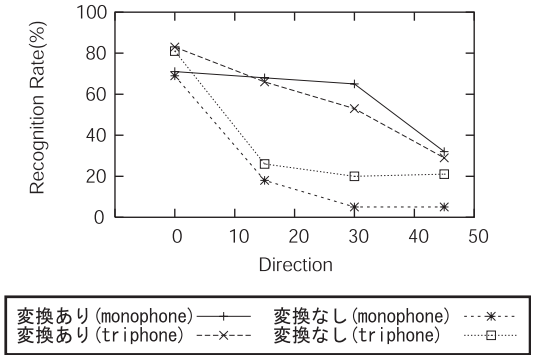

図 5: 方向に関する実験結果

回はデータの大きさを縮小することにより仮想的に距離を変化させたものとしている。データの縮小率は 7/8(0.6m)、3/4(0.7m)、5/8(0.8m)、1/2(1m)、3/8(1.3m)、1/4(2m)、1/8(4m) である。元の大きさのデータを学習データとし、縮小したデータをテストデータとして認識を行った。モデルはトライフォンモデルのみである。結果から

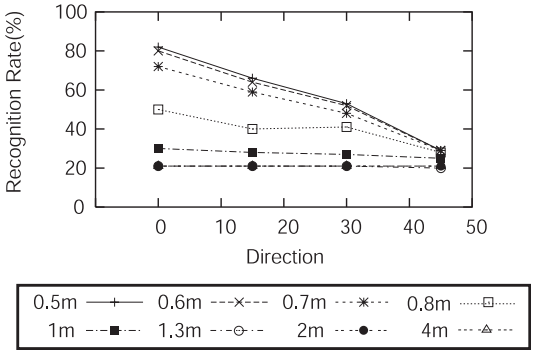

図 6: 距離に関する実験結果

0.7m あたりまでは認識率が大きく低下することがないことが確認できた。今回の実験では 320×240 ピクセルという低い解像度の画像を用いており、もっと高い解像度のカメラを用いることでさらに距離は伸びると考えられる。ロボットなどのシステムに良く使われる 640×480 ピクセルのカメラを用いれば 1.4m あたりまで使えることになり、実用に十分であると考えられる。

4. おわりに

本稿では唇情報を用いる際の人とカメラとの方向・距離の変化に対して頑健な認識を行うために、唇モデルと低周波数成分を用いた特徴量抽出を提案した。唇情報のみを用いた認識においてこれらの提案手法を用いることで、方向・距離の変化に対して頑健であることを実験で示した。今後は、人を増やして一般性を高める必要がある。また音声の特徴量も組み合わせることで、実環境において頑健に用いることのできる音声認識システムの構築を行う。おわりに、本研究の一部は、科研費、21 世紀 COE の支援を受けた。

参考文献

[1] Bernstein, L. E. , Benoit, C. , “For speech perception by humans or machines, three senses are better than one.”, Proc.ICSLP,pp. 1447-1480, 1996.

[2] Xiao Xing Liu, Yibao Zhao, Xiaobo Pi, Lu Hong Liang and Ara V Nefian, “Audio-visual continuous speech recognition using a coupled hidden Markov model”, IEEE ICSLP, pp. 213-216, September 2002.