7R－2

口唇画像情報に基づく音声対話制御の一手法 ‡

黄英傑† 陳履恒† 土肥浩† 石塚満†

東京大学工学部電子情報工学科††

1 はじめに

コンピュータのヒューマンインタフェースの一つの有効な手段として、音声方式[1]が挙げられる。コンピュータに音声を認識させるとき、周辺の雑音が話者の言葉と誤認識される場合がある。話者の話し以外の音声が誤って認識されてしまうのを防ぐために、話しの始めを例えばフットスイッチとかトークボタンで音声認識装置に提示する必要がある。コンピュータとのより自然な対話を実現するためには、フットスイッチのような接触型のスイッチを踏みながら話すのではなく、話者の話し始めと終りを自動的に検出して、音声を認識する方法が必要になる。

本研究では、マイク付きの小型CCDカメラで話者の唇の動きを追跡して話しの始めと終りを自動的に検出し、音声認識装置と連動させる制御手法を提案する。

2 システムの構成

システム構成の概要を図１に示す。音声と画像を同時に入力するために、マイク付きのカメラを使用する。話者が話しをする時、顔が自然にマイクに近付くので、入力画像は顔の下の部分（鼻、唇、顎を含める）になる。

マイク付きカメラの映像信号はワークステーションに、音声信号は音声認識装置に入力する。画像認識の結果に基づき、話しの始めと終りを音声認識装置に提示する。音声認識の結果はテキストに変換され、ワークステーションに表示される。音声認識装置は市販のオージス総研製DS200を使用している。

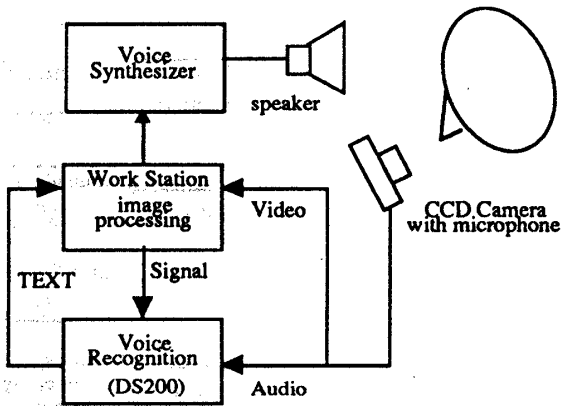

図１：システム概要図

3 口の開閉の判断

入力画像から特定の部分を取り出す場合に、対象部分の色、形状、輝度などの情報を利用することが考えられる。ここではシステムのリアルタイム性の要求を満たすために、輝度情報を利用する。入力画像は顔の下の部分で、１６０×１２０画素、２５６階調の白黒画像である。唇の部分はほぼ入力画像の１／３の面積を占めているが、輝度値は唇の周りの肌色よりやや低く、周辺環境の照明の影響で唇を安定的に直接抽出することは困難である。しかし、唇の間、即ち口の内部の輝度値は入力画像の中で最も低い[3]ので、まず口の内部の形状を抽出する。

口の内部の形状は口の開閉に従って変わり、この変化を口の開閉の特徴として捉える。一方、口の開閉を判断するとき、歯の影響も考慮に入れなければならない。ここでは、抽出した口の内部の上に判別ボックスも設けることによって口の開閉を判断する。

最後に、この検出した結果に基づいて適当な制限条件をかけて、音声認識装置に信号を出す。以下は、この手順に従って実験の結果を説明する。

3.1 口の内部の抽出

顔の下の部分の画像では、唇を含めて肌の輝度値を持つ画素が一番多い。唇の間の画素は低い輝度値に集中しているので、式(1)から求めた閾値で口の内

† Ying-jieh Huang, Lieu-Heng Chen, Hiroshi Dohi, and Mitsuru Ishizuka (huang@miv.t.u_tokyo.ac.jp)

‡ A method of dialogue management based on lip image recognition.

†† University of Tokyo, Faculty of Engineering, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113, JAPAN

部の形状を抽出する。

$$Th = M - Sd \quad (1)$$

Thは二値化の閾値で、MとSdは入力画像の平均値と分散である。図２には、(a)と(b)が原画像で、(c)と(d)が２値化した後抽出した口の内部である。

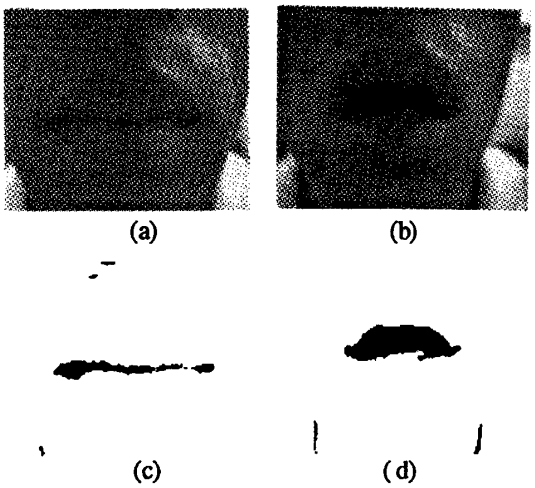

図２： 原画像(a),(b)と２値化した画像(c),(d)

3.2 判別ボックス

口の開閉を判断するために、抽出した口の内部の周りに図３に示したような判別ボックスを設置する。２値化した画像のY軸への加算投影量を計算して、投影量の最大値の位置と口内部の長さの1/2の所を判定ボックスの中心位置にする。そして、口内部の長さによって領域Ra,Rb、Rcを決める。口の開閉の判断は次の二つの条件によって決められる。

１）RaとRbの領域にある口内部の画素の数。

２）Rc領域に白い画素（歯の部分）の数。

また演算量を削減するために、現フレームの判別ボックスの範囲を次のフレームの加算投影量計算の範囲とする。

3.3 雑音の抑制

画像認識の結果によって音声認識装置にON（話者の口が開く）、OFF（話者の口が閉じる）の信号を送る。音声認識装置が音声を認識している最中、誤ってOFFの信号が入ると、認識結果の信頼性が低下してしまう。これを防ぐために、ある一定時間以上話者の口が閉じているのを認識した場合にOFF信号を出す。

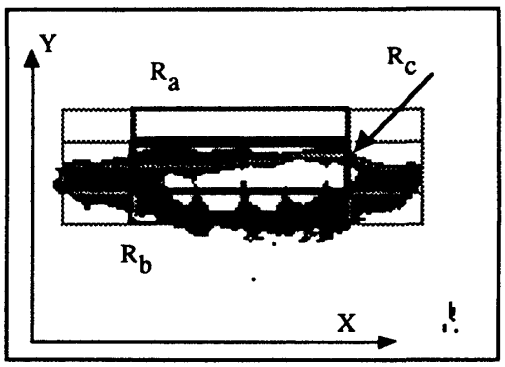

図３：判別ボックス

4 実験結果

実験は普通の室内の照明の条件下で行った。口が映る限り、話者の頭はある程度左右に揺れたり前後に移動してもかまわない。認識率について、口が閉じている時の正解率は約９７％、口が開いている時の正解率は約９３％である。SGI I Indyワークステーションで一枚の入力画像を処理するのに平均68msをかかる。

5 おわりに

本研究では、マイク付きカメラを使って話者の顔画像から口の動きを自動検出し、音声認識装置に話者の話し始めと終りを提示する手法を提案した。一枚の入力画像を平均68msの時間で処理することができる。約13frame/secの処理能力を持つ。コンピュータが発話中にユーザの発話があった場合、コンピュータの発話を中断してユーザ発話の認識を行うことにも有効であると考えている。

参考文献

[1] Y. Hiramoto, H. Dohi, M. Ishizaka: "A Speech Diaglogue Management System for Human Interface employing Visual Anthropomorphous Agent," Proc. 3rd IEEE Int' l Workshop on Robot and Human Communication, Nogoya, July 1994

[2] 田村ら：エネルギー関数とオプティカルフローを用いた口形輪郭の抽出・補完と追跡、信学研資、PRU89-20

[3] 黒田ら：顔画像からの口部領域の自動抽出法、信学研資 IE91-3