音声言語情報処理 12－18
（1996. 7. 27）

騒音環境下での音声理解のための唇認識と音声認識

宮崎敏彦　奥村晃弘　藤井明宏　岡野健治
沖電気工業(株) 研究開発本部 関西総合研究所

情報の入力手段の一つとして音声が有望視され、音声認識技術も比較的静かな環境では実用に耐えうる認識率が得られるようになってきている。しかし、実際の応用を考えた場合、その利用環境では様々なノイズが無視できない場合が多い。そこで我々は、画像認識を使った唇情報を併用することによって、音声認識の認識性能を向上させる方式について研究を進めている。我々の提案する唇情報と音声認識の融合方式は、音声認識から出力される複数の候補から、唇情報を用いて取捨選択するという方式である。本発表では、方式の概略と性能評価結果について報告する。

Lip-reading for Speech Recognition in Noisy Environment

Toshihiko Miyazaki　Akihiro Okumura　Akihiro Fujii　Kenji Okano
Kansai Laboratory, Research & Development Group, Oki Electric Industry Co., Ltd.

Most approaches of speech recognition using auditory information are very sensitive to background noise which include more than one speaker talk simultaneously. Humans solve these problems by using additional informations such as context information and visual information (such as lip movements). In this paper, we propose a method to fuse auditory information and lip movements in order to realize more robust speech recognition system in noisy environment. In our method, lip movements information are used to filter candidates generated by speech recognition system. As the experimental result, the recognition accuracy of isolated word utterance is improved.

1 はじめに

従来より、情報入力手段の一つとして音声が有望視されており、近年では音声認識を用いた自然言語対話システムが試作されている[5]。しかし現状の認識技術では、物音や近くの人が発する声などの背景ノイズが、認識率に及ぼす影響が大きいく、特定の静かな環境での利用に限定されていたり、マイクを口の近くに持ってくるなど、幾つかの制限が設けられている。

一方、電話や運転中の車内など、音声の利用が必然的であるような状況を除いて考えた場合、音声入力に期待される性質の一つとして、「手軽さ」をあげることができよう。例えば切符の自動販売機を考えてみると、目的地を言うだけで切符が買えればどんなにか楽だろうと思ったのは我々だけではないはずである。キーボード入力では、入力が確実である反面、キー配列を覚えたり、キーボードを打てる距離に近寄ったりする必要がある。地図上にタッチパネルを配置するという案も悪くはないが、表示にかなりの面積が必要だったり、画面上から目的地を探す手間が馬鹿にならない。その点音声は楽だと思われるが、残念ながら自動販売機が置かれているような環境では、認識誤りに対処する手間の方が大きくなってしまい、手軽さを活かせるまでは至っていない。

また、手軽さを活かすためには、ユーザがわざわざマイクを手にしたり、マイクに顔を近付けたりしなくて済むように、発話者からある程度離れた位置にあるマイクを使っても、音声認識が可能であることも望まれる。

我々はこのような考察のもとに、騒音環境下でもロバストな音声認識を実現するため、図1に示すような音声認識システムの研究を進めている。図において、音響処理部は、[4]のように、複数の人が同時に発話している混合音の中から特定の人の発声音を取り出す研究である。最終的には、図に示すような構成とは異り、音声認識と音響処理が相互に情報をフィードバックする必要があると思われるが、現状では完全な前処理として基本方式の検討を進めている。

画像解析部は、唇の動き情報の抽出と、顔画像による個人識別機能を試作し、これらの情報と音声認識や音響処理との融合について検討している。唇の動き情報を音声認識に利用するためには、特に以下の課題を解決する必要がある。

1. 画像の撮影条件に対しロバストな唇特徴の抽出
2. 個人性に左右されない特徴の利用
3. 音声認識との融合方法

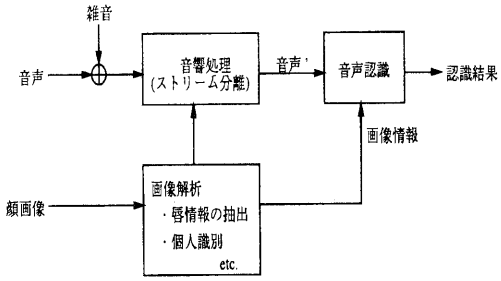

図 1: 騒音に対し頑健な音声認識に向けて

本論文では、図1の全体システムの内、特に唇情報の抽出と音声認識への適用について述べる。以下、2章で動画像から唇の動き情報を得る方式について、3章で、我々が提案する唇情報と音声認識の融合方法について、4章で、試作したシステムの唇情報を音声認識に適用した場合の認識性能について述べる。

2 唇の動き情報抽出

2.1 利用する特徴量

唇領域(あるいはその特徴点)を抽出する手法には、色情報を用いる方法[3]や、輪郭を抽出する方法[1]等が考案されている。しかしながら、いずれの手法も撮影条件の変化の影響を受けやすく、実際に使用する場合には、専用の照明を設置する必要があったり、処理パラメータの変更が必要であったりする。一方、我々の目的としている環境変化にロバストな音声認識への応用を考えた場合、唇の動き情報の抽出も環境変化にロバストであることが要求される。

そこで我々は、照明条件の変化の影響を受けにくくするために、色情報を用いた唇領域の抽出や、唇の輪郭抽出は行なわず、顔の中心線上の点を正確に追跡することにより、顔の動画像から唇の上下方向の動きを抽出する[1]。(紙面の都合により詳しいアルゴリズムは[6]を参照されたい。)

唇の動きから発話内容を捉えるためには、上下の動きだけでなく、左右の開き具合や前への突き出し具合も重要である[2]が、

1. 左右の動きや前後の動きは変化量が少なく、比較的離れた位置から、発話者にあまり制限を設けず

[1]音声で入力しようとしている機器には、ユーザに有用な何らかの情報が表示されており、(入力としての)発話時にはユーザがその方向を向いていると仮定している。

に撮影することを考えた場合、これらの情報を正確に抽出するのは非常に困難であると思われる。

2. 発話時に、顔が上下左右に動くことは十分考えられ[2]、その上で、変化量の少ない左右や前後の動きを捉えるのは非常に困難であろう。

といった理由により、当面我々は、安定的に抽出できると考えられる上下の動き情報のみを用いることにした。

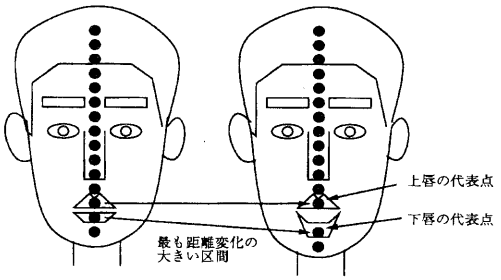

図 2: 代表点決定アルゴリズムの概要

2.2 抽出アルゴリズムの概要

我々の提案する唇の動き情報抽出アルゴリズムは、概略以下のステップよりなる。

目の位置の推定: 顔の対称性と眉や目の配置知識を基に、目の位置を推定する。目の位置情報は、唇情報を利用する際に正規化のための基準データとして用いる。

追跡点の設定: 最初のフレームに対し、顔の左右対称性を利用して、顔の中心線を求める。次に、図2に示すように、求めた中心線上に等間隔で追跡する点を設定し、それぞれの点に対して適当な大きさのテンプレートを作成する。

上唇と下唇の代表点の決定: 作成したテンプレートを用いて、各追跡点がどの位置に移動したかを、テンプレートマッチングによって探索する。この際、追跡精度を上げるため、探索の結果として追跡点間の位置関係が入れ替わったり、2点間の距離が閾値以上に変化した場合は、探索が失敗したものとみなす(図3)。また、唇の形状の変化や微妙な照明の変化などに追随するため、テンプレートとのマッチング度合がある閾値を越えた場合には、現在のフレームを用いてテンプレートの追加を行なう。

追跡点の探索を、ある程度の時間(約2秒)継続した後、その時点までで生き残っている点(探索が失敗しなかった点)の、隣合う点間の距離の時間的変化量が最も大きい点の組合せを求め、その2点を上唇と下唇の代表点とする(図2)。これは、顔の中心線上で相対的に最も上下に動くものが唇であると考えられるからである。

代表点の追跡: 以上のようにして求めた2点を、先と同様のテンプレートマッチングを用いて継続的に追跡する。この追跡によって得られた点の移動の時系列を唇の上下動情報として用いる。

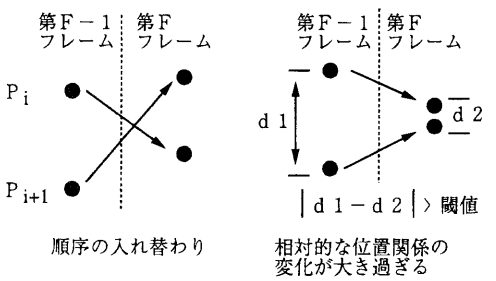

図 3: マッチングエラーの排除

2.3 唇の動き情報の抽出性能

唇の動き抽出性能の評価のために、予め決められた数単語を4人の話者に発話してもらい、これをビデオカメラで撮影した。撮影した動画データからの発話区間の切り出しは、音声の開始点と終了点をもとに人手で行ない、30frames/sec, 24bits/pixel, 320×240画素の画像としてA/D変換した。

評価は4人の話者で合計821発話を対象とし、代表点決定性能と、動きの抽出性能の2項目について実施した。代表点決定については、正誤判定処理を入れた場合で93.8%の抽出率が得られている。動き抽出については、代表点の抽出に成功した動画データの中から28個のデータを選び、評価した。正解データは、自動抽出したデータを人手で修正することによって作成した。正解データとのずれは、平均的に1画素前後のずれとなることが確認された。[3]

[2]同意の「はい」や否定の「いいえ」を考えれば容易に想像できる。

[3]画像条件や各種パラメタおよび評価に用いた画像の種類等については紙面の都合上割愛する。詳しくは[6]

3 唇情報と音声認識の融合方法

唇の動き情報を、不特定話者音声認識と組み合わせて使う場合、唇情報は次の様な性質を持つ必要がある。

1. 個人性に左右されない特徴量であること
2. 認識候補の文字列から容易に推定可能な特徴量であること

音節を区切って発話しない、いわゆる連続発話の画像を観察すると、唇の動きは個人によって大きく異なっている。従って、不特定話者に対応するためには、従来の唇情報を用いた方式のように、縦と横の開き具合の時系列データを直接モデル化して扱うのでは不適当である。

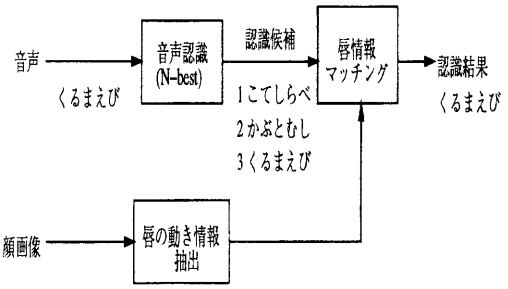

図 4: 唇情報と音声認識の融合

そこで我々は、比較的個人性に依存しない音の種類として両唇音に着目した。両唇音は、その名が示すように、唇を閉じた状態から始まる音であり[4]、音を正しく発話しようとする限り、誰が発話しても必然的に唇を閉じる動作が行なわれると考えられる。また、発話中に両唇音があるか否かは、認識候補の文字列から容易に得ることができる。

以上のことから、我々は両唇音を中心に、上記1と2の性質を満たす唇情報の利用を考えることとした。今回我々が試作した音声認識との融合方式の概略ステップを以下に示す。

1. 唇の上下動情報から、両唇音の位置を推定する。
2. 音声認識が出力した認識候補それぞれに対し、
   (a) 推定した両唇音の数が一致するかチェックする。
   (b) 推定した両唇音の位置と、認識候補の文字列から予想される両唇音の位置が一致するかチェックする。
   (c) 推定した両唇音の前後の唇の開き具合が、認識候補の文字列から予想される前後の状態に一致するかチェックする。

3.1 両唇音の位置の推定

先に説明した唇の動き情報抽出は、発話時の画像状態の下で唇付近の追跡しやすい適当な点を代表点として決め、その点をその後の入力画像に対して追跡するというものであった。このような方式を取ることによって、動き情報抽出が照明条件に対しロバストに行なえる反面、唇のどの点が代表点として選ばれるか分からないという問題が発生する。すなわち唇を閉じた状態で、2点間の距離が必ずしもゼロになるわけではなく、また同じ人間に対しても、代表点を選び直す度に異なった位置が代表点になる可能性がある。

このため、2点間の距離がどの程度のとき唇が閉じているか、を決定しなければならない。今回の試作では、発話開始前の、2点間の距離が比較的安定した状態を調べ、その区間の距離を「唇を閉じた状態」の距離(この距離を D_c とする)とした。

両唇音の発声位置は、2点間の距離の時系列の中から $D_c + \alpha$ 以下の距離に狭まった区間と仮定する。ここで α は、予め設定しておくマージンで、幾つかの発話データから統計的に決定した値である。

3.2 認識候補の絞り込み

本稿で述べる唇情報と音声認識の融合方式では、画像から得られる両唇音の数, 位置, 前後の状態を用いて、音声認識が出力する N 個の認識候補の中から絞り込みを行ない、残った認識候補を最終的な認識結果とする。(図5参照)

以下、各絞り込みステップについて簡単に説明する。

3.2.1 両唇音の数による絞り込み

各認識候補の文字列をもとに、その候補が正しい場合の両唇音の数を計算し、画像の動き情報から推定した両唇音の数と比較する。図5の例では、「こてしらべ」は両唇音が一つであるが、画像から推定した両唇音の数は二つであり、認識候補から除外される。他の2

を参照。

[4]厳密には閉じる前から始まっている場合もある。実際、音声認識との融合では両唇音の位置の推定を行なう際、両唇音の発話区間は閉じた状態を中心に適当な幅を持ったものとして考える。

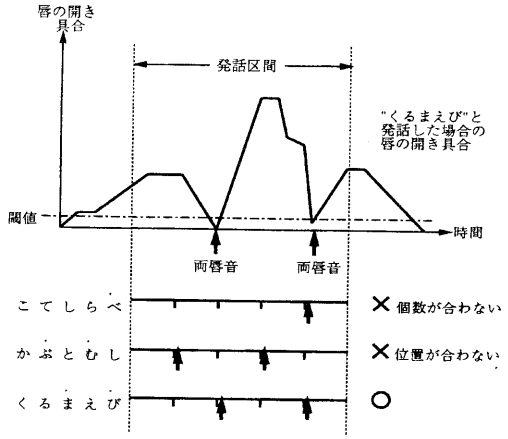

図 5: 唇情報マッチング

単語は推定結果と同じ二つの両唇音を持つため、認識候補として残される。

ただし、例えば「マンドリン」のように、発話の先頭に両唇音があるような場合には、その母音(この場合「あ」)を唇を閉じた状態から始めるのと本質的に区別がつかない。従って、先頭の両唇音は対象から除外する。

3.2.2 両唇音の位置による絞り込み

日本語の発話速度は、文節内で比較的一定であると言われている。今、発話全体の時間区間が分かっているとし、この発話速度一定という仮定を使うと、唇の動き情報の時系列の中でどの位置に両唇音が観測されるべきかが、認識候補の文字列から予測でき、これを使って認識候補を絞り込むことができる。

図5の例で考えると、音声認識から得られる認識候補の一つ「かぶとむし」は5音節であるので、発話区間を5等分し、両唇音である「ぶ」と「む」の位置(第2音節と第4音節)に両唇音が観測されると予測する[5]。この予測と画像の動き情報から得られた両唇音の推定位置とを比較し、一致していない場合は認識候補から除外する。

ただし、画像から推定しているのは、厳密には両唇音の位置ではなく、唇を閉じた位置である。両唇音を発声する場合、唇を閉じる動作は両唇音の音節の先頭で行なわれる。また実際の発話では、必ずしも等速度で発話されるとは限らないため、位置の比較にはある程度幅を持たせる必要がある。

このことから今回の我々の試作では、両唇音の推定位置の一致条件を以下のように定義している。

$$t_k - \frac{L}{2 \times N_w} \leq \text{両唇音の推定位置} \leq t_k + \frac{L}{2 \times N_w}$$

ここで、

t_k : k番目の音節の開始時刻

$$t_k = (T_s + \frac{L}{N_w} \times (k-1))$$

N_w : 認識候補wの音節数

T_s : 発話開始時刻

L : 発話区間の長さ

k : k番目の音節

3.2.3 両唇音の前後の状態による絞り込み

日本語の場合、両唇音の唇を閉じた状態の前後の唇の開き具合は、発話の先頭や文節の区切りを除くと、基本的には母音の発話形状であると言える。母音を発話する場合、唇の開き具合は、統計的には概ね

$$/a/, /e/ \geq /o/, /u/, /i/$$

という関係にある。勿論この関係は必ず満たされるというものではないが、連続的に発話されている文節内の両唇音の前後の相対的な関係としては、使える情報である可能性がある。そこで今回の試作では、上記大小関係を認識候補と唇の動き情報との一致条件として用いた。

4 融合による認識性能の評価

唇の動き情報から得た両唇音の数, 位置, 前後の開き具合を使って、音声認識性能がどの程度向上するかを評価するために、音声認識に入力する音声データに、人工的にホワイトノイズを加え、認識実験を行なった。認識タスクは30単語の単語認識である。評価データは一人の話者による299個の発話データである。なお、実験に用いた音声システムは、ノイズを付加しない299発話データに対しては100%の認識率が得られる。

実験結果を図6に示す。図において、横軸は上位何位までの候補を選ぶかを表し、縦軸は、その順位までに正解データが入っていた割合を表している。「音声のみ」とは、雑音を付加した音声データを音声認識に入力し、出力された認識候補の上位N位中に正解が含まれる率の変化である。他のグラフは、両唇音の個数, 位置, 前後の開き具合それぞれの情報を順に加えていった場合の正解含有率を表している。

[5] この例では説明を簡単にするために、全て5音節からなる単語を使っているが、認識候補間で音節数が異っていても良い。

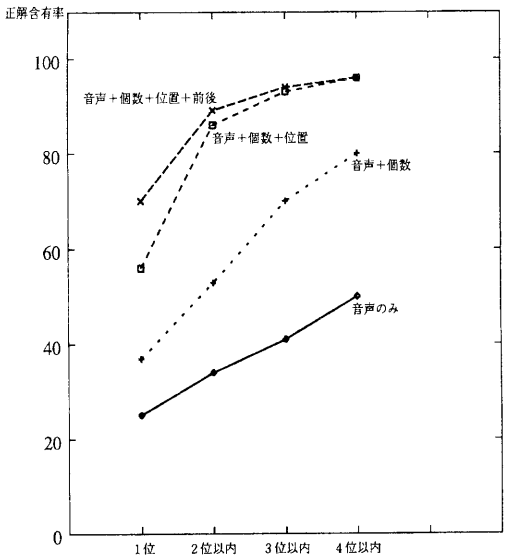

図 6: 唇情報との融合による認識性能の変化

5 今後の課題

今後の課題を以下に簡単に列挙する。

発話区間の推定: 現在の評価システムでは、両唇音の位置の一致検査の手掛かりとして、人手によって与えた発話区間の情報を用いているが、この発話区間の推定は今後の課題である。発話の開始と終了を調波構造の有無で判定する等の実験は行なっているが、システムとしては組み込まれていない。また、音声認識のマッチング結果として推定される発話区間を使うことも考えられる。予備的な実験として、雑音を付加した場合としなかった場合で、音声認識が推定する発話区間の差を比較したところ、正解データに対しては、概ね音声の分析窓に換算して 10 フレーム前後の誤差に収まっていた。我々の音声認識では、これは約 80msec であり、画像では 2 ～ 3 フレーム分に相当する。通常の発話速度では、1 音節が画像のレートで 4 ～ 6 フレーム程度であり、かなりの誤差であると思われる。

音声認識との結合: 上記認識実験で用いた音声認識システムの出力は、最大尤度の候補のみであった。このため、認識候補の上位 N 個を得るために、30 単語のそれぞれに対し、1 単語のみの認識を 30 単語分繰り返し、これによって得られた尤度を用いている。今後は、実際に N-best アルゴリズムを組み込んだ音声認識と結合し、評価する必要があろう。

複数話者での検証: 今回の実験では話者が一人であったが、今後は多くの話者で、より多くの発話データを用いて本方式の有効性を確認する必要がある。

6 おわりに

騒音環境下での音声認識の補助として唇情報を用いるために、画像中の唇の特徴点の決定方式と、唇情報の中から両唇音の情報を取り出し、音声認識と組み合わせる方式について報告した。試験的に、音声認識の出力結果をフィルタリングする形で両者を結合し評価したところ、認識率の改善に関しある程度の効果を得る見通しが得られた。今後は、先に課題として挙げた 3 点、特に発話区間の切り出しの自動化と、複数話者による評価を中心に検討を継続する予定である。

参考文献

[1] 田村進一, 梶見直樹, 岡崎耕三, 光本浩士, 河合秀夫, 副井裕: “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”, PRU89-20, pp.9-1(1989).

[2] 中野政身, 渡辺富夫: “ステレオ視による母音口形の識別”, 日本機械学会論文集 (C 編), 60 巻 570 号, pp.161-166 (1994).

[3] 黒田勉, 渡辺富夫: “色彩情報処理による顔画像の唇抽出法”, Human Interface, Vol.10, pp.13-18 (1995).

[4] 中谷智広, 奥野博, 川端豪: “音環境理解のためのマルチエージェントによる調波構造ストリームの分離”, 人工知能学会誌, Vol.10 No.2, pp.232-241 (1995).

[5] 宮崎敏彦, 須崎昌彦, 久野裕次, 田川忠道: “マルチモーダルインタラクションシステムの試作”, 情処研報, SLP95-7, pp.67-72 (1995).

[6] 岡野健治, 宮崎敏彦, 奥村晃弘, 藤井明宏: “動き情報を用いた唇の抽出法”, 情処研報, CV98-3, pp.13-18 (1996).

[7] 藤井明宏, 岡野健治, 宮崎敏彦, 奥村晃弘: “騒音環境下での音声理解のための唇認識”, 人工知能学会第 10 回全国大会予稿集 14-05 (1996).