ヒューマンインタフェース 74－8
（1997. 9. 12）

唇の動き情報による騒音環境下での音声認識性能の改善

奥村 晃弘　宮崎 敏彦
沖電気工業 (株)　研究開発本部　関西総合研究所

我々は、騒音環境でもロバストな発話理解の研究として、音声認識と唇情報の統合方式の研究を進めている。本稿では、唇の動きを使って音声認識に入力する音量を調節するパワーコントロールと、唇の動きの特徴による認識候補の絞り込みについて述べる。

Improvement of Speech Recognition under Noisy Environments Using Lip Movements

Akihiro OKUMURA　Toshihiko MIYAZAKI
Oki Electric Industry Co.,Ltd.
Crystal Tower 2-27 Shiromi 1-chome,Chuo-ku,Osaka 540,JAPAN

We have been studying a speech recognition system with robustness for background noises. Our major interests are the methods to fuse auditory information and lip movements information. In this paper, we propose two methods based on lip movements information. One is to control input voice level. The other is to reduce candidates of the words to be recognized.

1　はじめに

近年のコンピュータ性能の向上は目覚しいものがあり、特に最近のパーソナルコンピュータなどは、高い処理能力を有し、マイクやスピーカなどのマルチメディア関連機器をも標準で装備するようになってきている。このような環境になり、音声認識をソフトウェアだけで実現できる様になってきた。これにより音声認識の普及が期待されるが、実際に使用するとなると周囲の雑音の影響が問題になってくる。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システムであっても、周囲の物音や近くの人が発する声の影響により著しく認識率が低下してしまう。そこで、我々は騒音環境下でもロバストな音声認識を実現するために、発話中の話者の顔画像の情報を併用する研究を行なっている [1, 2]。

本論文では、唇の動き情報と音声認識の融合方法として、唇の動きと候補単語の文字列のマッチングによる認識候補の絞り込みと、唇の動き情報を使って音声認識が処理する音声の音量を変化させるパワーコントロール (詳細は後述) の二つの方法 (図1) を提案する。以下、第2章で唇の動きによる絞り込みについて、第3章でパワーコントロールについて、第4章で雑音量による候補単語絞り込みについて、第5章でこれらの性能を評価した結果について述べる。

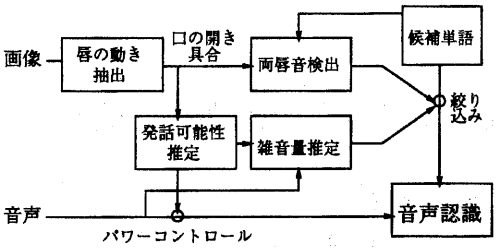

図 1: パワーコントロールと絞り込み

2 唇の動きによる絞り込み

我々はこれまでに行なってきた唇の動き情報と音声認識の融合の研究 [2] により以下の知見を得た。

- 唇情報と音声認識の融合には比較的個人性に依存しない両唇音[1]が有効である
- 日本語の発話速度は文節内で比較的一定であり、両唇音の発生すべき位置が発話区間と候補単語から推定できる
- 発話単語中の両唇音の数、両唇音の位置、両唇音の前後の状態により候補単語の絞り込みが可能である

しかしながら、「口の縦方向の開き具合 (height) がある閾値より小さいかどうか」 (図 2) を基準に両唇音を検出していたために、検出精度にやや不満があった。また、絞り込みの効果を大きくするために両唇音以外の情報を必要としていた。

一方、上記の融合方式の研究とは別に、唇の縦方向の開き具合 (*height*) を特徴量として (図 2) 以下の手法による単語認識実験 [1] を行なってきた。

1 認識に先だって、候補単語を発話した際の特徴量の時間的変化パターンを多数準備し、これらのパターンから各単語の標準パターンを作成しておく。

2 認識時には、発話によって得られたパターンと上記標準パターンとを DP マッチングすることによって単語を認識する。

この実験により、特定話者が同一単語を発話したとき口の縦方向の開き具合 (*height*) のグラフは、極大値や極小値 (山や谷のピークの値) のばらつきは大きいが、山や谷の大まかな形や位置の類似性が高いことがわかった (図 3)。

上記の実験では発話区間全体のグラフ形状の類似性を単語認識に利用したが、両唇音の発声位置のような部分的なグラフ形状にも類似性を見出すことができる。そこで、部分的なグラフ形状の一つとしてグラフ中の谷部分の形状に注目し、以下の方針で絞り込みを行なうことにした。

両唇音の検出精度の向上 両唇音発声の際には一瞬口を閉じるので、口の開き具合のグラフでは必ず谷として現れる。この谷の形状から両唇音の判定を行なうことによって、両唇音の検出精度を高める。

両唇音以外の特徴量の利用 両唇音以外にも特徴的な谷の形状が存在する。谷の形状から両唇音以外の特徴を得ることによって絞り込みの効果を向上させる。

以下、2.1章でデータの収集方法について、2.2章で谷の形状による分類について、2.3章で谷種別の判定方法について、2.4章でマッチング方法について述べ、2.5章で絞り込みの性能を評価する。

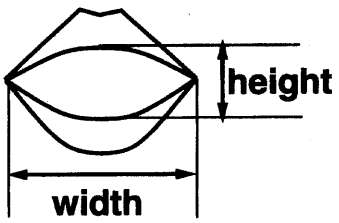

図 2: 利用する特徴量

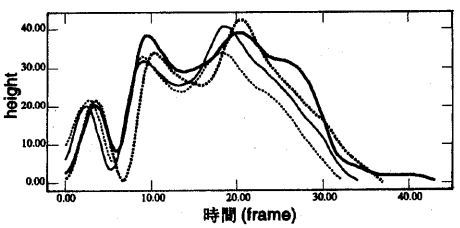

図 3: *height* の変化 (にまいがい)

[1] 両唇音とは 'b','m','p' で始まる音であり、発声するためには上唇と下唇が触れ合うことが不可欠な音である。

図 4: 撮影画像

2.1 データ収集

両唇音を含むものを中心とする３０種類の単語を選び、話者がこれらの単語を発話するところを正面から撮影し、総計１２２１の顔画像を得た(図 4)。

これらの顔画像からテンプレートマッチングで唇の動きを抽出し、この唇の動きから口の開き具合を求める。（紙面の都合により詳しいアルゴリズムは [1, 3] を参照されたい。）求めた口の開き具合は目の間の距離を使って正規化し撮影の倍率の違いを吸収する。

また、発話内容と動き情報の対応関係を解析するため、この動き情報と同期した音声データを使って、各音節区間のラベリングを人手にて行なった。

2.2 谷の形状の分類

音節ラベルと口の開き具合の時間変化のグラフを比較検討することにより、グラフ中に表れる谷の形状と該当区間の文字列との間に関係があることがわかった。そこで、谷の形状を該当部分の文字列の特徴から以下の 4 種類に分類する。以後、この谷の分類のことを谷種別とよぶ。

C タイプ 両唇音（'b','m','p' で始まる音）の発声によって形成される谷

XC タイプ 両唇音の直前に隣接した撥音または促音のために、撥音または促音の発声による谷と両唇音による谷が合成されてできた谷

p タイプ 音節の切れ目の部分や子音部分に形成される谷

a タイプ 撥音あるいは促音または母音の発声部分に形成される谷

前後の母音よりも該当部分の母音の口の開き具合が小さい場合に谷が形成される

2.3 谷種別の判定方法

発話中に現れた谷の形状から谷種別が判定できれば、その谷種別と候補単語との間で対応を取ることにより、絞り込みが可能となる。

そこで、収集した発話データを「学習用発話データ」と「評価用発話データ」に分け、学習用発話データ (610 データ) の中から 1163 個の谷をピックアップし、「学習用谷形状データ」とした。学習の際のパラメータは以下の 3 つの値を用い、これらを統計処理することによって得ることにする。

Vw 谷の幅 変曲点から変曲点までを谷の範囲として、その範囲に対応する時間（図 5a）。

Vh 谷底の高さ 谷底（谷の範囲内での最小値）にあたる部分の口の開き具合（図 5a）。

Sr 加速度波形の外接矩形の縦横比 口の開き具合の変化量から口の動く速度を算出し、速度の変化量から加速度を算出する。その加速度波形の最大値 Ah と谷の幅 Vw から $Sr = Ah/Vw \times 100$ として求める（図 5b）。

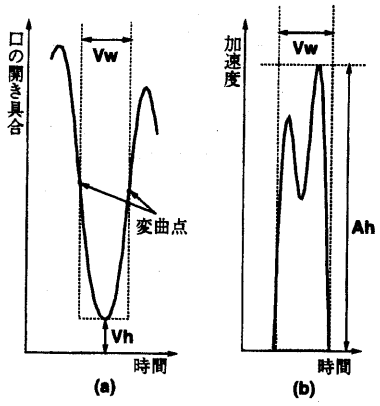

図 5: 谷形状のパラメータ

判別は谷種別を小分類 (表 1) でクラスタリングし、各小分類までのマハラノビスの距離によって行う。

表2に学習に使ったデータに対する判別結果を示す。aタイプとpタイプに多少の混同が見られるが、両唇音に関してはうまく判別できていることがわかる。

次に評価用発話データを用いた両唇音の判別実験を以下のような手順で実施し性能評価を行なった。

まず、それぞれの「評価用発話データ(611個)」から発話区間内の谷を抽出する。そして、抽出したそれぞれの谷の谷種別の判定を行ない、各単語から両唇音(谷種別がCまたはXCタイプのもの)がいくつ検出されるかを調査する。最後に、検出した両唇音の数と、発話単語の文字列から決定される両唇音の正解数との比較を行なう。但し、両唇音はCタイプ,XCタイプに分けてカウントし、単語の先頭の両唇音は検出されないので、対象から除外する。

この方法によって評価した結果を表3に示す。表からわかるように、我々のアルゴリズムを使うことによって、99%以上の精度で単語内の両唇音の数を正しく検出できている。

表 1: 学習用谷形状データの内訳

大分類	小分類	個数	大分類	小分類	個数
C	C	588	a	i	94
XC	NC	60		o	9
	QC	20		u	88
p	p	218		N	20
	ps	21		Q	22
	px	23	合計 1163 データ		

表 2: 学習用谷形状データの谷種別判別結果

谷種別	誤り	正答数	母数	正答率
C	なし	588	588	100.0%
XC	なし	80	80	100.0%
a	a→p 12	250	262	95.4%
p	p→a 2	231	233	99.1%

表 3: 評価用発話データでの両唇音検出結果

	数	率
両唇音(C,XC)の数が一致	606	99.2%
両唇音タイプの誤り(C→XC)	3	0.5%
Cタイプ検出もれ(C→p)	2	0.3%

2.4 単語文字列とのマッチング

以下の方法で観測データと候補単語とのマッチングを行なう。

まず、口の開き具合のグラフから発話区間内にある谷部分を抽出し、谷形状データを取得する。また、候補単語からpタイプ以外の文字列上の特徴を取り出す。

次に、谷形状データと文字列上の特徴とを対応させて下記の2項目から評価点を算出し、最も評価点が高くなる対応づけとそのときの評価点を求める。

- 谷形状データと文字列上の特徴とのマハラノビスの距離
- 発話区間における谷の位置と文字列上の特徴の音節位置との一致度

発話区間から抽出した谷形状データの数と文字列上の特徴の数が違うときには、以下の2つの場合がある。

- 谷形状データに対応する文字列上の特徴がない
- 文字列上の特徴に対応する谷形状データがない

前者の場合は、谷形状データとpタイプとのマハラノビスの距離から評価点を算出する。また、後者の場合はペナルティとして該当する特徴に応じた得点を評価点から減じる。

2.5 絞り込みの性能評価

2.4章の手法により計算した得点を用いて、候補単語の絞り込み性能を評価した。絞り込みの基準としては、最高得点との得点差を用いた。

「評価用発話データ(611個)」に対して絞り込みを行なった結果を表4に示す。正解含有率は絞り込みを行なった後の候補単語の中に正解単語が含まれる率を表す。さらに、絞り込みによって候補単語をいくつに減らすことができたかを知るために、絞り込み後の候補数の平均(μ)を示した。また、参考のために平均(μ)と分散(σ)から$\mu + 3\sigma$の値を示した。

今回、候補単語数は30であるから、最高得点との得点差を5.0として絞り込みを行なえば、危険度3%以下で平均で1／6に悪くても1／2程度に候補数を絞り込めることがわかる。

表 4: 唇の動きによる候補単語絞り込み性能

最高得点との差	正解含有率	絞り込み後の候補数	
		平均 (μ)	$\mu+3\sigma$
0.0	71.5%	2.2	6.7
1.0	76.9%	2.8	8.8
2.0	84.8%	3.3	10.3
3.0	95.1%	3.9	11.1
4.0	97.2%	4.6	13.3
5.0	97.7%	5.0	14.8
10.0	99.3%	6.4	21.1

3 パワーコントロール

唇の動き情報と不特定話者音声認識とを組み合わせて使う方法として、唇の動きから発話の有無を推測することが考えられる。例えば、対象となっている話者が発話していないときに他の人の話し声が入った場合、この音声は雑音として取り扱うべきであり認識の対象としてはならない。しかし、これを音声情報だけで行なうのは非常に困難であると言える。このような場合に唇の動きから発話の有無がわかれば有効であると言える。

口を開いているときや、唇が動いているときでも発話を行なっていない場合は多く存在する。そのため、唇の動きから発話しているかどうかは完全にはわからない。しかし、口をずっと閉じたままの状態のときはまず発話していないと言えるであろう。

そこで、時間 t における発話可能性 $p(t)$ を時間 t での口の開き具合 $h(t)$ と、その前後 a フレーム内での最大高低差 $delta(t,a)$ を使って式1の様に定義する。

$$\left.\begin{array}{l} delta(t,a) = max(t,a) - min(t,a) \\ p_0(t) = h(t) + delta(t,a) \\ p(t) = \left\{ \begin{array}{ll} 0 & if \;\; p_0(t) \leq 0 \\ p_0(t)/L & if \;\; 0 < p_0(t) < L \\ 1 & otherwise \end{array} \right. \end{array}\right\} \quad (1)$$

但し、$max(i,j), min(i,j)$ は t が $i-j$ から $i+j$ に変化したときの $h(t)$ の最大値および最小値をそれぞれ表す。また、L および a は係数である。

これにより、p が1に近い値にもかかわらず実際には発話が行なわれていない場合はあるにしても、p が0に近い値の場合に発話が行なわれていることは極めて少ないようになる。言い換えると、p が0である範囲では対象となっている話者は発話していないと考えられるので、人の声が入っていてもそれを認識すべきではない。このような範囲では音声認識への入力音声の音量を0にして、音声認識が反応しないようにすれば効果的である。

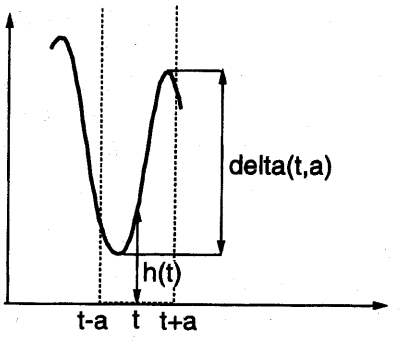

図 6: 発話可能性を求めるパラメータ

そこで、p の値に基づいて音声認識への入力音声の音量を調節することを試みる。我々はこの手法をパワーコントロールと呼んでいる。

実際には変化特性を表す係数 γ_p を用いて式2のようにして、入力音声の音量を変化させる。なお、γ_p の値を変えて実験した結果 γ_p の値には0.1を採用した。

$$P_{out}(t) = pow\left(p(t), \frac{1}{\gamma_p}\right) \cdot P_{in}(t) \quad (2)$$

ここで、$P_{in}(t)$ および $P_{out}(t)$ は、時間 t における入力音声および出力音声の振幅を表す。また、$pow(x,y)$ は x^y を表す。

4 雑音量による候補単語の絞り込み

絞り込みの手法については2.5章で述べた。この章では実際に音声認識と組み合わせる方法について述べる。

我々が以前に行なった絞り込みの実験 [2] においては、先に音声認識を行ない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って、残った認識候補を最終的な認識結果とする方法を用いた。絞り込みの方法としては上記の他に、音声認識を行なう前に音声認識に与える候補単語に対して絞り込みを行なう方法がある。前者の方法の場合は複数の認識結果を比較するため、音声認識から上位N個の認識結果を得る必要がある。後者の場合、N-best アルゴリズムを組み込んだ音声認識装置でなくても

構わないというメリットがある一方、絞り込みに失敗すると正解が得られないのというデメリットがある。そこで、絞り込みの失敗による性能低下を低減させるために、雑音量に応じて絞り込みの度合を変化させることを試みた。雑音量 n は発話可能性 $p(t)$ を使って式3から求める。

$$n = \frac{\sum_{t=0}^{T} \left\{ pow\left(1 - p(t), \frac{1}{\gamma_n}\right) \cdot P(t) \right\}}{\sum_{t=0}^{T} pow\left(1 - p(t), \frac{1}{\gamma_n}\right)} \qquad (3)$$

但し、$P(t)$ は時間 t のときの入力音声の振幅を、T は入力データの終る時間を表す。また、γ_n は変化特性を表す係数で、$\gamma_n = 0.5$ とした。

このようにして求めた雑音量 n から k/n(k は係数) を算出し、絞り込みに使う得点差として利用する。これにより、雑音が少ないときは絞り込みを緩く、雑音が大きいときほど絞り込みをきつくすることができる。

5 性能評価

撮影時の音声データに、0 ～ 22dB の雑音を付加して認識させるシミュレーションを行い性能を調べた。雑音には、「電子協騒音データベース」(NOS-9601) に収録されている展示会場 (ブース内) を利用し、認識の手法として「パワーコントロール」と「絞り込み」それぞれの手法を使う場合と使わない場合の組み合わせにより以下の 4 つの方法について試みた。

- PC+LR+VR
  「パワーコントロール」と「絞り込み」を音声認識と併用する
- LR+VR　「絞り込み」と音声認識を併用する
- PC+VR
  「パワーコントロール」と音声認識を併用する
- VR　音声認識のみで認識を行なう

「評価用発話データ (611 個)」に対して上記シミュレーションを行なった結果を図 7に示す。比較的雑音が少ない範囲 (S/N が 10dB 以上) では「パワーコントロール」が有効に働き、雑音が多い範囲 (S/N が 10dB 以下) では「絞り込み」が威力を発揮していることがわかる。

S/N=10dB の場合、音声認識のみでの認識率は 40% 程度であるが、「パワーコントロール」と「絞り込み」を音声認識と併用することにより、これを 90% 以上に改善することができた。

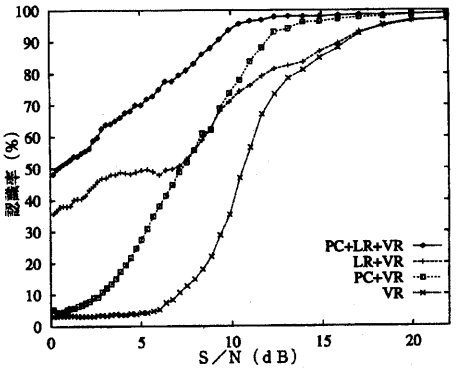

図 7: 雑音量に対する認識性能

6 おわりに

雑音が多い場合の音声認識の誤認識を減らす方法として、唇の動き情報が有効であることを示した。現在、唇の動きによる認識を行なうには発話区間の入力を必要としている。今後、この点を改良する必要がある。また、パワーコントロールは発話区間の入力を必要としないが、データ撮影時に発話の前後はなるべく口を閉じるように指導した。このように、現在は発話者の協力を必要としている。これも、今後対処していくべき課題である。

参考文献

[1] 奥村晃弘, 岡野健治, 宮崎敏彦, 藤井明宏:" 唇の動き情報を利用した単語認識" , 情処研報,96-HI-68,pp.1-8(1996)

[2] 宮崎敏彦, 奥村晃弘, 藤井明宏, 岡野健治,: "騒音環境下での音声理解のための唇認識と音声認識", 情処研報,96-SLP-12,pp.97-102(1996)

[3] 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,96-CV-98, pp.13-18(1996)