顔画像情報と音声情報の統合による発話認識

奥村晃弘† 濱口佳孝†
岡野健治† 宮崎敏彦†

我々は，騒音環境でもロバストな発話理解の研究として，音声認識と唇の動き情報の統合方式の研究を進めている．本論文では，顔および唇の位置検出方式，および，抽出した唇の動き情報を使った音声認識との統合方式を提案する．そして擬似的に雑音を加えた発話認識実験において，唇情報を併用することにより，雑音が6dBのときの認識率を5.24%から91.5%に改善することができ，本統合方式が有効に機能することを確認した．

Speech Recognition Based on Integration of Visual and Auditory Information

AKIHIRO OKUMURA,† YOSHITAKA HAMAGUCHI,† KENJI OKANO† and TOSHIHIKO MIYAZAKI†

We have been studying a speech recognition system with robustness for background noises. Our major interests are the methods to integrate auditory information and lip movements information. In this paper, we propose a new method for lip extraction and method to integrate auditory information and lip movements information. We realize the effectiveness of these methods. Because use our integrate metod, it can raise recognition rate to 91.5% from 5.24% in added 6 dB noise artificial environment.

1. はじめに

人間にとって最も自然な情報伝達手段である音声は，従来より計算機に対する情報入力手段の1つとして有望視されており，近年のマイクロプロセサ能力の向上とともに，音声認識機能を搭載しているパーソナルコンピュータも発売されるなど，普及の兆しが見えてきている．また単純な単語によるコマンド入力だけでなく，ユーザにとってより自然な対話文を使って，データベース検索など各種サービスが受けられるような自然言語対話システムの研究が多く報告されてきている[7)~9)]．

しかし現状の音声認識技術では，物音や他人の声などいわゆる背景雑音が認識率に及ぼす影響が非常に大きく，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，利用に際していくつかの制限が設けられている．

一方で，発話者の唇の動きは音響的な雑音に影響されにくいと考えられるため，唇を映像的にとらえ，その形状変化などの画像情報と音響情報とを統合することによって高雑音下での認識性能を高めようとする研究が従来より行われている．

萩原ら[5)]は音声と画像それぞれ独立にHMMを用いた認識を行い，得られた2つの尤度を適当な重み係数によって合成する（結果統合）という手法を報告している．しかしながら彼らの実験では，比較的単純な2値化処理により開口部分の縦横幅の比および口内面積を抽出しているため，より正確な特徴量を得るために2値化画像を人手によって修正している．また中村ら[3)]は，やはりHMMを用いて，音声と画像の情報を学習時点で統合する初期統合法と，萩原らと同様の結果統合法の両者を比較している．しかし，精度の良い映像を得るために発話者の頭を固定し口の周りの画像のみを撮影するという方法をとっている．

その他，松岡ら[6)]，Wuら[4)]，間瀬ら[2)]により画像情報のみによる発話認識（機械読唇）の報告がなされている．これらは口形や口唇輪郭あるいは唇周辺の動きを解析することによって読唇を試みているが，口唇周辺の特徴を抽出することが困難なため，口紅をつけて抽出を容易にしたり，Wuら[4)]のようにスプラインモデルのフィッティングや間瀬ら[2)]のようにオプティカ

† 沖電気工業株式会社研究開発本部関西総合研究所
Kansai Lab., OKI Electric Industry Co., Ltd.

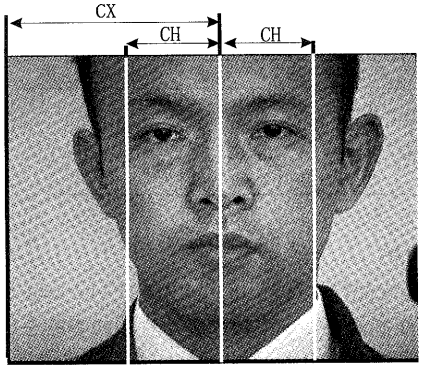


図1 左右対象領域の探索
Fig. 1 Search for symmetric area.

ルフローを用いるなど比較的計算量の多い手法を使ってより正確な情報抽出を試みている．

画像情報を統合した音声認識の実用化を考えた場合，画像処理側では取得する画像の性質や特徴量，あるいはアルゴリズムに関し次のような点を考慮する必要がある．

(1) できるだけ正確な唇の情報を得るためには唇を大きく写し出すことが望ましいが，このために発話者に対し発話位置や姿勢など多くの制約を課すのは望ましくない．

(2) 音響的に雑音の多い場所での利用を前提としており，画像情報に対する高い信頼性が非常に重要である．したがって，環境変動に対し頑健で信頼性が高く不特定話者の間で安定的な画像上の特徴量を利用する必要がある．

(3) 音声認識はすでに実時間処理のレベルにあり，画像情報の取得および統合にも実時間処理を念頭に置いた手法の開発が必要である．

本論文では以上の点をふまえ，**図1** に示すように，発話時に発生する動きなどによって顔（正確には唇）がフレームアウトしてしまわない程度の大きさで写された映像を前提とし，また特定の色を仮定しないような，顔および唇の位置検出方式を提案し，実験によってその有効性を確認する．さらに，抽出した唇情報（唇の動き情報）を使った音声認識との統合方式を提案し，擬似的な雑音を加えた実験によって騒音下での有効性を示す☆．

☆ 我々の最終目標は不特定話者を対象とした連続発話の認識であり，音声部分の認識には，不特定話者連続音声認識システムを用いた．なお，本論文で報告している実験評価で，顔の検出および唇の動き情報の抽出については複数話者を対象とした評価の結果である．音声との統合部分については，1人の発話者の発話データによる評価である．

2. 正面顔の検出

唇情報を取得する処理を開始するためには，まずシステムの前に話者が来たことを検出する必要がある．ここで，話者はシステムの方を向いて発話すると仮定すると，ほぼ正面を向いた顔が画面内に現れるのを検出すればよいことになる．カメラに正面顔が写っていることを検出するときに，照明条件などによる影響を少なくするために，

(1) 左右対称の領域を探索して顔の候補とすることにより，顔の色を仮定することなく顔の候補となる部分を検知し，

(2) 顔候補の中心線上の色情報から顔の色を逐次推定する．また，初期の唇の位置候補を得るために

(3) 推定された顔の色との違いから唇候補を抽出し，数フレームにわたって安定して唇候補を検出できた場合に正面顔とすることで，唇の抽出を開始するのに適したフレームを検出する．

以下，処理手順に従いアルゴリズムを説明する．

2.1 左右対象領域の検出

正面顔は左右対象となる部分が大きいと考えられるため，ある軸を対象に左右対象性の高い部分を顔の候補として検出する．

幅 W 高さ H の画像の中に，図1に示すようなある縦の直線 $x = CX$ を仮定し，その直線までの距離が CH 以内の部分画像について，$x = CX$ を軸に対象となる位置にある画素どうしの画素値の距離の総和を以下の式で求める．なお，座標 (x, y) での赤，青，緑の画素値を $R(x, y)$，$G(x, y)$，$B(x, y)$ とする．

$$\begin{aligned} ds(x,y,i) = &(R(x-i,y) - R(x+i,y))^2 \\ &+ (G(x-i,y) - G(x+i,y))^2 \\ &+ (B(x-i,y) - B(x+i,y))^2 \end{aligned}$$

$$cdist(x) = \sum_{y=0}^{H-1} \sum_{i=1}^{CH} \sqrt{ds(x,y,i)}$$

対象の軸が $CH <= x < W - CH$ の範囲において，$cdist(x)$ が最小となる位置を顔の中心の候補とする．

また，$cdist(x)$ の最小値がある閾値以下にならない場合は顔がフレーム内にないものと判定する．

2.2 色による顔領域の識別

同一フレーム内では顔の色相はほぼ一定であると考えられるので，顔の中心近傍の色相分布から顔の色の推定を行う．

顔の中心として検出された直線までの距離が NH 以内の部分画像について，各スキャンラインごとに色

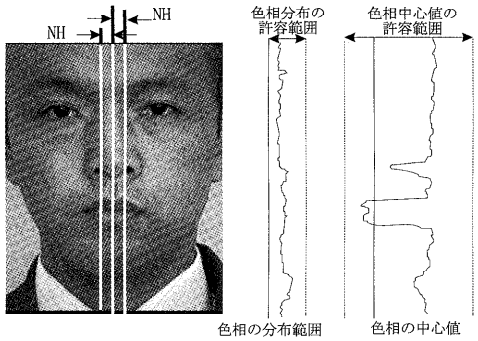


図2 顔の色の推定
Fig. 2 Face color detection.

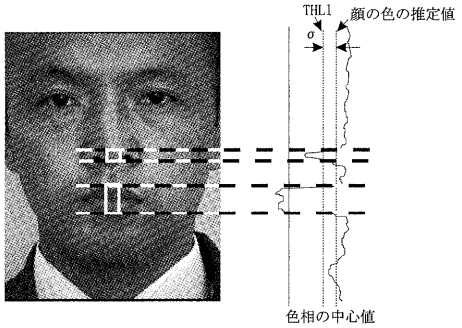


図3 色相による唇候補の検出
Fig. 3 Lip detection by color.

相の分布範囲 $SW(y)$ と色相の中心値 $SM(y)$ を検出する（図2）．$SW(y)$ と $SM(y)$ が一定範囲内となる y について色相の中心値の平均値 SF をとり，顔の色の推定値とする．

色相を利用したのは，顔の凹凸による照明の当たり具合の変化の影響が比較的少ないためである．

ここで，色相の分布範囲が一定値以下の部分が少なかった場合は顔がフレーム内にないものと判定する．また，顔の色の推定値が緑色等一定範囲外の場合も顔がフレーム内にないものと判定する．

2.3 唇位置候補の検出

顔の中心近傍において，唇は肌の色と異なる色相となると思われるので，色相が大きく異なる部分を唇の候補とする．

色相の平均をとって顔の色相の推定値 SF を推定するときの色相の標準偏差を σ とし，SF より σ だけ赤よりの色相 $THLl$ を設定する．色相の分布範囲 $SW(y)$ が一定値以下で，図3に示すように色相の中心値 $SM(y)$ が $THLl$ より赤くなる位置 y を唇位置の候補とする．

FN フレーム連続で同じ位置に唇の候補が得られた場合に，正面顔がフレーム内にあるものと判定する．

表1 評価結果
Table 1 Face detection result.

FN	平均フレーム数	最大フレーム数
2	2.001	4
3	3.002	6
4	4.006	8
5	5.034	10

2.4 評　　価

30 fps で記録した18種，計2182フレームの発話中の正面顔画像について顔の検出性能の実験を行った．唇候補の連続検出数 FN に対して，顔動画像の任意フレームから検出を開始し，正面顔があると判定されるまでのフレーム数の最大値と平均値は**表1**のようになった．

$FN=5$ で FN に対する平均フレーム数の比率が大きくなるが，これは発話中の画像で口の動きが大きい部分において，連続して同一位置に唇候補を検出できないことがあるためである．

以上の結果から，閾値を $FN=4$ に設定し，室内の背景動画像36種，計1047フレームについて再度検出実験を行ったところ，このうち2種の動画像で計6回，正面顔があると誤認識した．99.5%の背景画像について顔がないと判定しており，設置条件を選べば実用できる範囲内と考えられる．

3. 唇情報の抽出

基本的には従来我々が提案してきた手法[1]であるが，初期追跡点を正面顔の検出で得られた唇候補の位置に初期追跡点を設定することで初期の追跡点を減らし，処理の効率化を行っている．以下簡単にアルゴリズムを説明する．

3.1 追跡点の設定

前述の各唇位置候補に，図4に示すように一定間隔 DH で追跡点を設定する．これらの各初期追跡点ごとに，それを中心に矩形のテンプレートを作成する．

3.2 テンプレートマッチングによる追跡

テンプレートマッチングは画素値の差の自乗和が最小となる部分を探索して追跡点の新たな位置とする．ただし，発話しているときの唇形状は変化するために初期のテンプレートのみでマッチングを続けると形が許容範囲を超えてしまい追跡に失敗する場合が起こりうる．そこで，テンプレートマッチングの処理においてマッチングの許容量以内ではあるが，ある程度自乗和が大きくなった場合は，その画像で新たなテンプレートの追加を行う．このように，1つの追跡点に対し複数のテンプレートを用いることにより，変形する

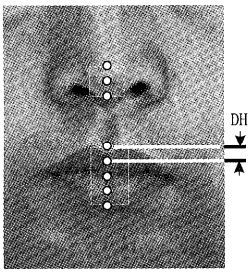


図4 初期追跡点の設定
Fig. 4 Set initial tracking point.

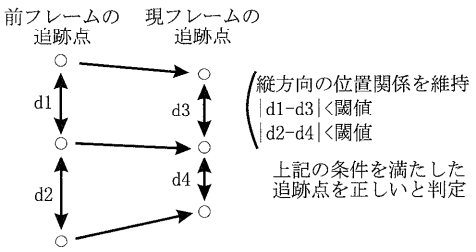


図5 追跡点の正誤判定
Fig. 5 Tracking error detection.

唇の動き追跡を実現している．

最初のテンプレートと最近追加されたテンプレート双方を用いてマッチングを行う．

3.3 追跡点の除去

顔の中心線上の各唇候補位置に設定された追跡点をテンプレートマッチングにより追跡すると，歯，舌など消滅するものや，肌の特徴のない部分などはテンプレートマッチングの処理が正常に行うことができないため，まったく異なる部位がマッチングすることが多い．これらは上下関係が頻繁に入れ替わる，あるいは大きく位置が変化することになる．

本処理では連続する3つの追跡点の位置関係の推移を用いてこのような追跡点の検出，除去を行い，唇の抽出性能の向上を図る．

追跡点の判定は

- その追跡点および上下の追跡点が直前のフレームで正しいと判定されている．
- 上下の追跡点と順序が入れ替わらない
- 上下の各追跡点との距離の変化が一定値以下である．

図5のように，以上の条件をすべて満たした場合に追跡点は正しいとする．この判定をすべての点について行い，一度も正しいと判定されなかった追跡点は誤りであったと判定し，以後テンプレートマッチングを行わない．

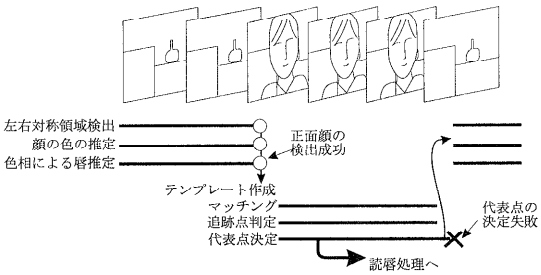


図6 全体の処理の流れ
Fig. 6 Flow of lip detection.

3.4 唇の代表点の決定

以上の処理で正しい追跡点と判定され続けている点のうち，上唇と下唇の追跡点は互いに最も距離の変化が激しいと考えられる．現フレームまで正しいと判定されている追跡点のうち，隣り合う追跡点どうしの距離の最大値 Dmax と最小値 Dmin を求め，Dmax-Dmin が最も大きくなる隣り合う2点をそれぞれ上唇，下唇の代表点とし，その間隔の変化を唇の動き情報として出力する．

3.5 唇の抽出のまとめ

処理全体の流れは**図6**のようになる．処理開始後，正面顔の判定を繰り返し行い，正面顔を検出した時点で唇の追跡を開始する．唇の追跡結果は後述の音声認識との統合に使用される．

唇の追跡処理において，唇と判定できる追跡点がなくなるか，一定時間唇の動きがない場合は追跡処理を終了し，再度正面顔の検出処理に戻る☆．

4. 唇情報と音声認識の統合方法

図1程度の大きさの顔画像から得られる唇の上下動だけから発話単語を正確に認識することは非常に困難である．しかし，音声認識と組み合わせれば，音響情報とはまた別の観点から得られた情報として有効に利用できる．周囲の雑音によって音声認識の性能が低下している場合は，特に効果が大きいであろう．

そこで，以下の2種類の組合せ方法を利用することにした．

(A) 認識候補の絞り込み
まず唇情報だけを使って認識を行い，その結果上位の成績をとった候補だけに絞り込んで音声認識の認識候補とする．

(B) 入力音量調節（従唇音量法）
唇の動きから発話の可能性を求め，そ

☆ 動き追跡に関するアルゴリズムの詳細と評価結果は参考文献1)を参照いただきたい．実験で用いたテンプレートサイズは32×16画素である．

れに基づいて音声認識への入力音量を調節する．

(A)の認識候補の絞り込みについては4.1節で，(B)の従唇音量法については4.2節で述べる．

なお，パラメータ作成および性能評価のために，以下の3種類の唇の動き情報を収集し利用した．

(a) 学習用発話データ（610データ）
(b) 評価用発話データ（611データ）
(c) 駅名発話データ（152データ）

発話単語数はいずれも30種類であり，(a)，(b)は両唇音を含むものを中心とする30種類の単語を，(c)は大阪市営地下鉄の30の主要な駅名をそれぞれ1人の話者が発話したものである．

4.1 認識候補の絞り込み

音節を区切って発話しない，いわゆる連続発話の場合には，唇の運動軌道は前後の音節の影響を受けて多様に変化する．そのため，読唇を行うには，唇を各音節に対応した静的なパターンとして考えるのではなく動きとしてとらえることが重要である．

そこで我々は，読唇の手法として以下に示す手順を採用することにした．

(1) 上下それぞれの唇の動きを抽出する．
(2) 上下の唇の距離から口の開き具合（以後，開口度）を求める．
(3) 開口度の時間変化から動きの特徴を抽出する．
(4) 各特徴の種類を判別し，その系列と文字列を比較する．

以下ではこの手順にそって詳しく述べる．

4.1.1 開口度の算出

3章で述べたように，我々の唇情報の抽出手法は，唇のある特定の位置（たとえば輪郭線，内側の境界など）を見つける必要がなく，同様の上下動をしている部分ならどこでもよい．そのため動き情報の抽出が撮影環境に対しロバストに行える反面，唇のどの点が代表点として選ばれるかが分からないという問題が発生する．すなわち，口を閉じたときの2点間の距離がこの時点では未知である．口を閉じたときが0となる開口度は，発話の有無や音節種類の推定に重要なので，以下の方法で開口度を求める．

まず，発話前後の区間から2点間の距離が比較的安定した状態を抽出し，そのときを口を閉じた状態として2点間の距離 H_c を求める．次に，この距離 H_c をすべての2点間の距離から引き，開口度とする．

このようにして得られた開口度はサンプリング周波数30 (Hz)，データの粒度20☆とともに低く，また追跡時の誤差も含まれるためこのままでは扱いにくい．

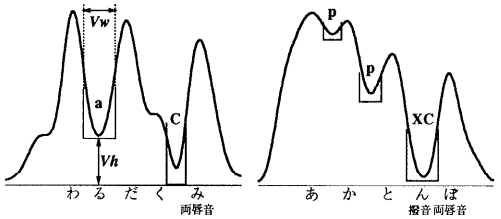


図7 谷種別の具体例
Fig. 7 Ravine shape examples.

表2 谷種別による分類
Table 2 Classification by ravine shape.

谷種別	文字列の特徴	谷形状の特徴 Vh	Vw	Sr	発生関係
C	両唇音	小	小	大	○
XC	(撥音 or 促音)+ 両唇音	小	大	小	○
p	音節の切れ目	大	小	小	△
a	・撥音 or 促音 ・前後の母音よりも開口度が小さい母音	大	大	小	△

そこで，口の動きは滑らかに変化すると仮定し，サンプル間に線形補間により4点を補間した後に，ローパスフィルタを使ってスムージングを行う．さらに，左右の目の間隔を使って正規化し，カメラの撮影倍率の違いを吸収する．このようにして，得られた開口度のデータ例を図7に示す．

4.1.2 唇の動きの特徴

“ば”，“ぱ”，“ま”のような両唇音は発音の際に上唇と下唇が触れ合うので一瞬口を閉じた状態になり，個人差が非常に少ない．また，発話区間内で口を閉じる動作のほとんどは両唇音によるものなので，開口度から両唇音の数および正確な位置の検出が期待できる．しかし，両唇音を1つも含まない単語も多く存在し，これだけで十分ではない．そこで，特徴のとらえ方を両唇音を中心として以下のように考える．

両唇音発音の際は一瞬口を閉じるため，開口度が急激に小さくなった後にまた元に戻るので，開口度の時間変化グラフ上に必ず谷として現れる．グラフ上に現れる谷に関して調査したところ，両唇音2種を含む4種類に大別でき，谷形状と該当部分の発話文字列の間には表2の関係があることが分かった．

表中の谷形状の特徴 Vh および Vw は，図7にも示すように，Vh は谷底部分での開口度☆☆を，Vw は変曲点から変曲点までを谷範囲としたときの谷幅を表

☆ 図1のような画像を320 × 240画素で扱うため，口を最も開いたときの上唇と下唇の距離は20ドット程度である．
☆☆ スムージングによりピークが欠けるので，スムージング前の開口度を利用する．

表3 評価用発話データ(b)での両唇音検出結果
Table 3 Feature detection result.

	数	率
両唇音(C, XC)の数が一致	606	99.2%
両唇音の種類を誤判定(C → XC)	3	0.5%
両唇音の検出もれ(C → p)	2	0.3%

す．Sr は動きの激しさを表すパラメータとして，谷範囲内での最大加速度と Vw の比によって求める．また，発生関係の項目には文字列の特徴があれば必ず谷が現れる場合は○で，そうとは限らない場合は△で表してある．

このことから，グラフ上の谷（開口度がいったん減少し，また元に戻る動き）を両唇音をも包含する特徴として扱えることが分かる．

4.1.3 特徴の種類判別

谷の形状と谷種別の間に大まかな関係があることはすでに述べた．そこで，谷の形状の情報を以下の手順で統計処理し両唇音の判別実験を行った．

まず，学習用発話データ(a)の中から谷種別を特定できる1163個の谷を人手によりピックアップし，それぞれの谷に関する Vh, Vw, Sr 3つの値を求め，「学習用谷形状データ」とし，このデータから各谷種別の分散共分散行列を作成した．Vh, Vw は前章で説明した谷底での開口度と谷幅を表す．また，Sr は動きの激しさによって谷種別Cとpをより良く分離するための値である．

次に，評価用発話データ(b)から発話区間内の谷を自動抽出し，その谷の Vh, Vw, Sr と，先に求めた谷種別の分散共分散行列でマハラノビスの距離を求めることによって発話中に両唇音（谷種別がCまたはXC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，発話内容の文字列から決定される両唇音の正解数との比較を行う．ただし，発話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．表から分かるように，99%以上の精度で発話内の両唇音の数と種類を正しく検出できている．

4.1.4 特徴と文字列とのマッチング

日本語の発話速度は，文節内では比較的一定であるといわれている．今，発話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，観測されるべき谷種別とその位置を認識候補の文字列から予測することができる．

そこで，単語文字列から得られた特徴の系列と実際に観測された谷の特徴の系列とを比較してマッチングすることにする．複数の特徴どうしを比較する必要があるので，まず，1つの特徴どうしを比較したときの得点($Score$)を式(1)のように定める．

$$L_s = L_w/N_w, \quad D = |V_e - V_o|$$
$$pos = \begin{cases} 1 & \text{if } D \le L_s \\ (2L_s - D)/L_s & \text{if } L_s < D < 2L_s \\ 0 & \text{otherwise} \end{cases}$$
$$Score = (10 - dist) \cdot pos \qquad (1)$$

式(1)において，L_w は発話区間の長さを，N_w は候補単語の音節数を，L_s は1音節の長さの推定値を，V_e および V_o は谷の予測位置☆と観測位置を，$dist$ は観測した谷と予測谷種別とのマハラノビスの距離をそれぞれ表す．

複数の特徴どうしの比較には，特徴どうしの対応づけを行い，それらの $Score$ の合計点を算出する．これら2種の特徴は数が同じとは限らないので，対応する特徴がない場合も考慮する必要がある．基本的にこのような場合はペナルティとして合計点から5点減点する．しかしながら，表2の発生関係の項目にも示したように，文字列の特徴があっても必ず谷が発生するとは限らない谷種別もあるので，これらの場合は減点を免除する．

すべての対応づけについて合計点を算出しその中の最大値を該当単語の評価点とする．合計点の計算例を**図8**に示す．

4.1.5 各音節の開口度によるマッチング

母音発音時の開口度には，おおむね/a/,/e/ > /i/,/o/ > /u/の関係がある．そこで，特徴のマッチング結果から音節の切れ目を推定し，そこから，各音節の開口度を求めてさらに詳細なマッチングを行う．

まず，谷種別pはもともと音節の切れ目を表すことから，また，谷種別Cは両唇音の子音部分なので付近に音節の切れ目があると考えられることから，谷底部分に音節の切れ目を設定する．谷種別XC，およびaは谷部分がちょうど1音節に当たるので，谷の始まりと終わりの部分に音節の切れ目を設定する．この時点で現在比較している単語の文字数に必要な音節数に満たない場合は，以下の方法で音節の切れ目を推定し補充する．

母音発声時は比較的口形が安定しているので開口度の変化が少ないのに対して，子音発声の直前から子音

☆ 谷種別Cは両唇音にあたる音節とその直前の音節の境界に，XCは撥音または促音にあたる音節の中央に，aは該当する音節の中央に谷があると予測する．

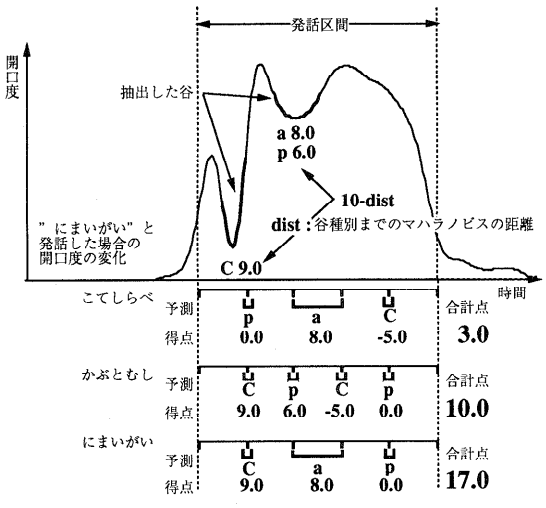


図8 候補文字列とのマッチング
Fig. 8 Matching with candidate letters.

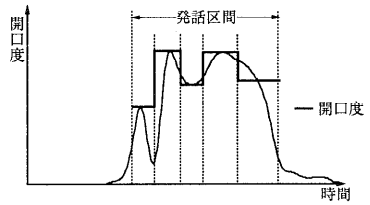


図9 各音節の開口度
Fig. 9 Mouth opening of each syllable.

発声に至る区間では，新しい口形を形作る必要性から開口度の変化が激しくなる．このことから開口度の加速度が極大値を示す部分を音節の切れ目と推定できる．

各音節の区間が決まれば，次に，各音節区間の中で速度が0になる部分を探し，そのときの開口度をその音節での開口度とする．ただし，速度が0になる部分が複数あるときは最も区間中央に近いものを，速度が0になる部分がないときは区間中央を利用する（**図9**）．

最後に，各音節の開口度を認識候補の文字列の各母音にあてはめ，“ae”，“io”，“u” の3つの母音グループに分類し，“ae”>“io”>“u” の関係に違反する程度 P を式(2)より求める．

$$P=\sum_{x\in V}\sum_{i}^{3}\begin{cases}\alpha_{ig(x)}(E(x,\mu_i)+E(x,n_i)\beta) & \text{if } O_{g(x)}<O_i\\ \alpha_{ig(x)}(E(\mu_i,x)+E(m_i,x)\beta) & \text{if } O_{g(x)}>O_i\\ 0 & \text{otherwise}\end{cases}\quad(2)$$

式(2)において，V は観測した全母音の開口度を，$g(x)$ は開口度 x が属する母音グループを，μ_i，m_i，n_i は母音グループ i に含まれる開口度の平均値，最大値，最小値をそれぞれ表す．また，α_{ij} は母音グループ i と母音グループ j の距離を表す係数，β は平均値を使う場合と最大値最小値を使う場合を統合する係数である．さらに，$E(i,j)$ は $i>j$ のとき $i-j$ をその他のときは0の値をとり，母音グループ i より母音グループ j の方が開口度が小さいという関係がある場合を $O_i>O_j$ とする．

この P の値と特徴のマッチングによる評価点から総合得点を候補単語ごとに算出し順位づけする．この順位づけに従って上位候補だけに絞り込んで音声認識の認識候補とする．

4.1.6 絞り込みの性能評価

評価用発話データ(b)，および，駅名発話データ(c)に対する候補単語の絞り込み性能の評価結果を**表4**に示す．発話区間既知は正しい発話区間を与えた場合を，発話区間未知は後述する発話可能性から発話区間を自動推定☆した場合を表す．絞り込み基準値は絞り込みの程度を決定する数値であり，全候補中の最高得点との得点差が絞り込み基準値以下の候補だけが絞り込み後の候補単語となる．また，正解含有率は絞り込みを行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

表4 絞り込みの性能
Table 4 Reduce candidates result.

評価用発話データ(b)

	発話区間既知			発話区間未知		
絞り込み基準値	正解含有率	絞り込み後の候補数 平均	最大	正解含有率	絞り込み後の候補数 平均	最大
0.0	93.4%	1.6	5	72.6%	2.7	7
0.5	98.7%	2.6	7	74.4%	3.1	8
1.0	99.0%	3.3	11	76.7%	3.4	12
2.0	99.3%	4.3	18	94.3%	4.4	15
3.0	99.3%	5.0	18	98.7%	5.1	20
5.0	99.8%	5.9	21	99.0%	6.2	24

駅名発話データ(c)

	発話区間既知			発話区間未知		
絞り込み基準値	正解含有率	絞り込み後の候補数 平均	最大	正解含有率	絞り込み後の候補数 平均	最大
0.0	69.1%	1.9	5	75.7%	4.6	9
0.5	84.9%	3.5	7	82.2%	5.5	12
1.0	88.8%	4.6	9	84.2%	5.9	17
2.0	92.8%	5.7	13	92.1%	6.5	19
3.0	96.7%	6.9	18	95.4%	7.2	19
5.0	99.3%	8.6	23	98.0%	8.8	24

☆ 手法は欄外に後述する．

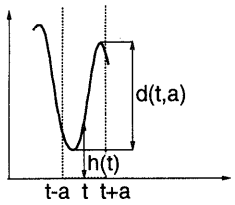


図10 発話可能性を求めるパラメータ
Fig. 10 Parameters for speech possibility.

数の平均値と最大値を示した．

4.2 唇の動きによる音量制御

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．たとえば，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うのは非常に困難であるといえる．このような場合に唇の動きから発話の有無が分かれば有効であるといえる．

口を開いているときや，唇が動いているときでも発話していない場合は多いので，唇の動きから発話の有無を完全に知ることは困難である．しかし，口を閉じた状態がある程度持続するときは発話していないといってよいであろう．

そこで，時間 t における発話可能性 $p(t)$ を時間 t での開口度 $h(t)$ と，その前後 a フレーム内での開口度の最大値 $\max(t,a)$ と最小値 $\min(t,a)$ を使って式(3)のように定義する（図**10**）．ただし，L は係数である．

$$
\begin{aligned}
&d(t,a) = \max(t,a) - \min(t,a)\\
&p_0(t) = h(t) + d(t,a)\\
&p(t) = \begin{cases} 0 & \text{if } p_0(t) \le 0 \\ p_0(t)/L & \text{if } 0 < p_0(t) < L \\ 1 & \text{otherwise} \end{cases}
\end{aligned} \tag{3}
$$

これにより，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．

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．我々はこの手法を従唇音量法と呼んでいる．

実際には変化特性を表す係数 γ_p を用いて式(4)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．

$$
P_{out}(t) = pow\left(p(t), \frac{1}{\gamma_p}\right) \cdot P_{in}(t) \tag{4}
$$

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．

4.3 雑音量の測定とそれに基づく絞り込み

音声認識は周囲雑音がほとんどないときは実用上問題のない性能を発揮できるが，周囲雑音に非常に弱く雑音が増えると性能が大幅に低下する．それに対して，機械読唇は性能は高くないが，周囲雑音の影響をほとんど受けない．このことから，周囲の雑音の大きさに応じて読唇情報による絞り込みの程度を制御することにする．

周囲雑音は発話がないときの音量であるから，発話をしていない可能性 $1-p(t)$ から重みを算出し，入力音声の振幅の大きさ $|P_{in}(t)|$ の加重平均として雑音量 n を求める（式(5)）．

$$
n = \frac{\displaystyle\sum_{t=0}^{T}\left\{pow\left(1-p(t), \frac{1}{\gamma_n}\right)\cdot|P_{in}(t)|\right\}}{\displaystyle\sum_{t=0}^{T} pow\left(1-p(t), \frac{1}{\gamma_n}\right)} \tag{5}
$$

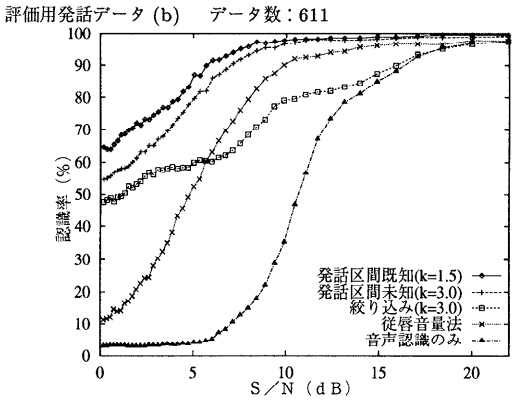
ただし，T は入力データの終る時間を表す．また，γ_n は変化特性を表す係数で，$\gamma_n = 0.5$ とした．

この雑音量 n を使って k/n（k は係数）を絞り込み基準値とすれば，雑音が少ないときは絞り込みが緩く，雑音が大きいときほど絞り込みがきつくなるので，雑音の増加による認識性能の低下をおさえることができる．

5. 統合による認識性能の評価

撮影時の音声データ（雑音なし）に対して0～22 dBの雑音を付加した合成音を認識するシミュレーションにより，雑音量に対する認識性能を調べた（図**11**）．実験に用いた音声認識システムは不特定話者連続音声認識システムである．また，雑音には「電子協騒音データベース」に収録されている展示会場（ブース内）を利用した．なお，図中の「絞り込み」「従唇音量法」はそれぞれの手法を単独で音声認識と組み合わせた場合を，「発話区間既知」「発話区間未知」は上記手法を双方ともに組み合わせたうえで，発話区間を人手で指定した場合と，発話可能性から自動推定☆した場合とを

☆ 観測開始後，最初に $p = 1$ になる時点を発話開始とし，その後最初に $p < 0.3$ となる時点を越えずにその時点に最も近い $p = 1$ となる時点を発話終了とした．



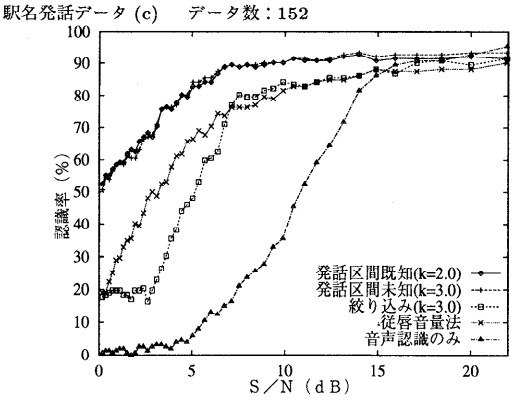


図11 雑音量に対する認識性能

Fig. 11 Speech recognition result in artificial noisy environment.

示す．雑音量から絞り込み基準値を算出するための係数 k については，予備実験によって試した何種類かの値の中から最も良好な結果が得られたものを用いた．

図11から2つの手法を同時に組み合わせると効果が高いことが分かる．特に，雑音が6 dBのときは認識率を5.24%から91.5%に大幅に改善できている．また，「発話区間未知」のときは絞り込みの性能が低い（表4）にもかかわらず，音声認識と統合したときには「発話区間既知」にかなり近い性能を得ている（図11）．そこで，このシミュレーションでどの程度の絞り込みが行われているかを調査した（**表5**）．表5を見ると絞り込みの失敗は非常に少なくなっている．特に，「発話区間未知」のときは「発話区間既知」と比較して絞り込みを緩めることによって正解含有率を98%以上にし，絞り込みの失敗を回避していることが分かる．これらのことから，絞り込みに失敗しない範囲で候補数を極力削減できるように，絞り込みの程度を制御することが，認識率の改善に最も重要であるといえる．

表5 雑音量に基づく絞り込みの結果

Table 5 Reduce candidates result in artificial noisy environment using noise based reduction.

評価用発話データ (b)

S/N (dB)	発話区間既知（$k=1.5$） 正解含有率	絞り込み後の候補数 平均	最大	発話区間未知（$k=3.0$） 正解含有率	絞り込み後の候補数 平均	最大
0.2	99.2%	3.6	15	98.2%	4.9	20
3.1	99.2%	4.0	18	98.7%	5.5	23
6.0	99.2%	4.8	18	98.9%	6.2	24
10.5	99.5%	5.8	21	99.0%	7.2	27
14.0	99.7%	6.7	24	99.0%	7.7	28
20.0	99.7%	7.7	28	99.0%	8.1	28

6. おわりに

正面顔画像から唇の動きを抽出する方法，および，唇の動き情報と音声認識の統合方式について述べた．唇の動き抽出では，口紅の塗布などの補助手段なしに唇の動きを実時間で抽出できることを，音声認識との統合では，上記の手法で取得した唇の動き情報を利用することにより，雑音の増加による音声認識の性能低下を抑制できることを示した．

今後は，複数人の発話データおよび複数種類の雑音による性能評価実験を行っていく．

参考文献

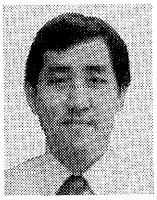
1) 岡野健治，宮崎敏彦，奥村晃弘，藤井明宏：動き情報を用いた唇の抽出法，情報処理学会研究報告，96-CV-98-3, pp.13–18 (1996).
2) 間瀬健二，アレックスペントランド：オプティカルフローを用いた読唇，電子情報通信学会論文誌，Vol.J73-D-II, No.6, pp.796–803 (1990).
3) 中村 哲，山本英里，永井 論，鹿野清宏：HMMを用いた音声と唇画像の統合による音声認識と唇画像生成，情報処理学会研究報告，97-SLP-15-17, pp.93–98 (1997).
4) Wu, J-T., 田村進一，光本浩士，河合秀夫，黒須顕二，岡崎耕二：音声・口形特徴量を併用するニューラルネットを用いた母音認識，電子情報通信学会論文誌，Vol.J73-D-II, No.8, pp.1309–1314 (1990).
5) 荻原昭夫，新谷 輝，土居直史，福永邦雄：視聴覚融合を用いたHMM音声認識，電気学会論文誌，Vol.115-C, No.11, pp.1317–1324 (1995).
6) 松岡清利，古谷忠義，黒須顕二：画像処理による読唇の試み，計測自動制御学会論文集，Vol.22, No.2, pp.67–74 (1986).
7) Nagano, K. and Takeuchi, A.: Speech Dialogue with Facial Displays: Multimodal Human-Computer Conversation, *Proc. 32nd*

Annual Meeting of the Association for Computational Linguistics (*ACL-94*), pp.102–109 (1994).
8) 宮崎敏彦，須崎昌彦，久野祐次，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，情報処理学会研究報告，96-SLP-7-11, pp.67–72 (1995).
9) 山本幹男，伊藤敏彦，肥田野勝，中川聖一：人間の理解手法を用いたロバストな音声対話システム，情報処理学会論文誌，Vol.37, No.4, pp.471–482 (1996).

(平成10年 3 月30日受付)
(平成10年10月 2 日採録)

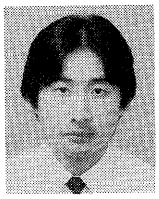
奥村 晃弘（正会員）

1990年大阪工業大学工学部電子工学科卒業．同年沖電気工業（株）入社．以来，同社関西総合研究所にて，グループウェア，マルチモーダルインタフェース等の研究に従事．1998年より（財）イメージ情報科学研究所出向中．



濱口 佳孝（正会員）

1991年名古屋大学大学院理学研究科物理専攻修士課程修了．同年沖電気工業（株）入社．同社関西総合研究所およびマルチメディア研究所にて文書認識，画像識別の研究に従事．日本物理学会会員．



岡野 健治

1991年九州大学工学部情報工学科卒業．同年沖電気工業（株）入社．以来，同社関西総合研究所にて，文字認識，顔識別，アイリス認識等の研究に従事．

宮崎 敏彦（正会員）

1981年大分大学工学部卒業．同年沖電気工業（株）入社．1984年より（財）新世代コンピュータ技術開発機構に出向し，並列論理型言語の研究に従事．1988年復職．現在，関西総合研究所勤務．コンピュータヒューマンインタラクション，画像認識，画像合成等に興味を持つ．日本ソフトウェア科学会会員