論　文

オプティカルフローを用いた読唇

正　員　間瀬　健二[†]　　非会員　アレックス　ペントランド[††]

Automatic Lipreading by Optical-Flow Analysis

Kenji MASE[†], *Member and* Alex PENTLAND[††], *Nonmember*

あらまし　音声認識を行う場合に，音声情報と共に視覚情報を使うと認識率の向上に役立つと考えられる．特に音声信号レベルが雑音などによって低いときに有用である．ここでは，視覚情報だけに基づいて連続発声の単語を認識するシステムを報告する．口の周りのオプティカルフローを計算し，唇の動きを推定して認識を行う．この動きデータから筋肉の動作を推定できることになる．動きゼロの情報を単語境界の抽出に利用し，更に動きのパターンを単語の識別に用いた．数字の識別を限定数のデータで実験を行った結果，それぞれの単語についてパターンは安定しており認識に使えることがわかった．また話者を変えてもそのパターンは類似しており，不特定話者の認識手法としての可能性もある．3人の話者で実験し，連続発声した単語から，各単語の抽出と認識を行い約70％の識別率を得た．

1. まえがき

人間は口腔，鼻腔，口唇などで構成される声道を，筋肉を動かして変形させることによって調音を行っている．そのときに口唇と顎を動かす筋肉によって，顔の表情にも変化をきたす．読唇術はその表情の変化をもとに話している言葉を理解する方法である．実際には読唇情報は調音位置に関する情報を主に与える[(1)]とされ，文脈なしですべての言葉を読唇によって理解することは難しい．しかしながら，音声信号に基づく言葉の認識（以下，音声認識と呼ぶ）と補い合うところは多い．例えば鼻音の /n/，/m/，/ng/ は音声認識では識別しにくい音であるが，口唇の形は非常に異なる．また工場内や車内などの雑音環境において音声認識が困難であるときでも，視覚情報には影響がない*．このように，人間は知らないうちに聴覚障害のあるなしにかかわらず読唇をして会話を行っている．

計算機とのマン・マシンインタフェースにおいても，視覚情報と音声情報の両方をうまく結合すると，計算機の言語理解の能力を上げることができる．Petajan[(4)]はこれに着目して，音声認識ボードの認識候補出力に口形の画像処理結果を組み合わせて，認識率向上を図った．しかしながら，彼の実験では口腔の暗い部分と皮膚，歯，舌などを2値化処理により区別し口腔部の形状を解析するため，顔を特殊な照明・撮影条件におかなければならなかった．

そのほかにも松岡ら[(5)]，栗田ら[(6)]，Petajan ら[(7)]，Finn & Montgomery[(8)]，田村ら[(9)]により画像処理による読唇の報告がなされている．これらのほとんどは，口形あるいは口唇輪郭を解析し読唇を試みているが，口唇周りの画像は濃度変化が緩やかで，これらの形状を正確に求めるのは本質的に困難である．そのため，松岡ら[(5)]は唇に黒い口紅をつけて口唇輪郭の抽出を容易にしている．また，田村ら[(9)]はスプラインモデルのフィッティングにより，より正確な口唇輪郭を求めようとした．このように，読唇では映像信号からいかに必要な情報を取り出すかの工夫が重要な課題となっている．なお発声学の分野ではこれらより以前に，唇の形状が調音とどうかかわるかを調べるために，ストロボスコープなどを使って発声中の連続写真をとった[(1),(10)]り，LED発光素子を唇の上下左右に付けるなどして，手作業で口形を計測，解析する研究が多数行われている．

† NTTヒューマンインタフェース研究所，横須賀市
NTT Human Interface Laboratories, Yokosuka-shi, 238-03 Japan
†† MITメディア研究所，米国
MIT Media Laboratory, Cambridge, MA 02139 USA

* 最近，これらが相補的ではなくむしろ融合することがあるという報告もある[(2),(3)]．

対象が映像信号になるという違いはあるが，計算機による音声認識で解決すべき問題が，視覚情報を用いる読唇でも同様にあてはまる．すなわち，

（ｉ） 認識に必要な情報の確実な収集
（ii） 連続発声の単語認識
（iii） 不特定話者の認識
（iv） 大語いの認識

などである．（ii）～（iv）の問題は読唇情報を音声認識の補助として考えた場合は，重要とはならないが，騒音環境のように読唇のみでまたは読唇情報を主として言語認識を行うときには避けられない．ところがこれまでの報告では，（ｉ）の問題を解決するためにいかに唇の形状を正確に取り出すかに注目しており，（ii）～（iv）の問題を扱ったものはない．

そこで筆者らはこれらの問題のうち特に（ｉ）～（iii）を解決するために，口唇の形状ではなく，口唇周りの動きに注目して読唇を試みた[11]．調音のもとになる筋肉の動きが表情を変化させ，その変化をたよりに読唇が行われているのであるから，筋肉の動きをとらえることができれば，その動きのパターンによって，発声している言葉を識別できるはずである．また，筋肉の動きは発声を行うための神経インパルスに直接関係しており，話者が変わっても一定であると推定される．そこで，口唇周りのオプティカルフローを求め，特徴パターンとなる速度を検討し，連続発声の単語（英数字）の識別実験を行った．オプティカルフローを使うことの利点は，

（１） 人間の視覚は変化する照明条件の中でも動きには敏感であり，読唇の場合も動きに注目している．オプティカルフローを直接求めることにより，口唇形状の抽出という問題から逃れられる．

（２） 単語の切れ目には一瞬の動作の停止がある．すなわち単語の切れ目は口唇の速度がゼロとなり，速度を特徴量とすることにより，連続発声した単語列を分解できる．形状解析では，分節は非常に難しい．

（３） 生理学的に見れば，同じ単語を発声するためには，話者が変わっても同じように筋肉を動かしていると考えられる．口唇形状の個人差やひげの有無の影響を受けない特徴量を得られる．

などが，考えられる．

以下，本論文ではオプティカルフローに基づく特徴を求め単語認識を行うシステムを示す．また，そのシステムで英語の数字列の発声に対して行った予備実験の結果の詳細を示し，上に示した仮説を検証する．

2. 筋肉モデル

2.1 発声にかかわる筋肉

顔には大きく分けて２種類の筋肉：表情筋とそしゃく（咀嚼）筋があって，発声の際に口形や口唇の動きを制御している．そしゃく筋は収縮によって下顎骨を頭蓋骨に近づけそしゃく運動に役立つ．すなわち，下顎を下げ，口唇を開閉させる．

表情筋は口裂の周りだけで約10種類あり，それらの収縮・弛緩が複雑に組み合わさって，表皮の変形を形成している．主なものは，円形の口輪筋，頬筋，頬骨筋，口角筋などである[12]．図１にこれらの筋肉の位置関係を示す．

これらの筋肉と表情とのかかわりを扱った研究としては，Waters[13]が，10種類の表情筋の動きをシミュレートして顔の表情をコンピュータグラフィックスのアニメーションで表現した例がある．Watersは Ekman-Friesen の FACS (Facial Action Coding System[14]) の考え方に基づき，表情の基本要素となるアクションユニットで表情の部分的変化を記述し，その組合せで「怒る」，「笑う」などの表情の表現に成功した．FACSでは筋肉の配置と機能に対応するようにアクションユニットを定めて，表情の記述を行ったため，このように表情のシミュレーションには非常に有効である．逆に表皮の変化から，表皮下で起こっている複数筋肉の動作を解析し個々の筋肉の動きを正確にとらえることは困難であると考えられる．しかしながら，筋肉のモデルとして代表的な（動きが顕著な）筋肉だけを使えば，ある程度の解析は可能と考えられる．

FACSのマニュアル[14]には，ある表情をアクションユニットで記述するための指針が書かれているが，その中で，口唇周りの形状を記述する言葉として，八つ

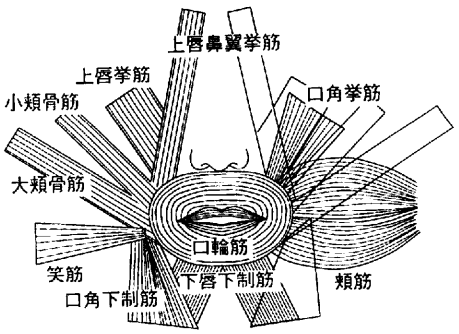


図１ 顔の筋肉
Fig. 1 Schematic of facial muscles.

の特徴が挙げられている．すなわち，口形の延伸(elongate)，口形の収縮(de-elongate)，口唇の細め(narrow)，口唇の太め(widen)，薄い口唇(flatten)，口唇の突出し(protrude)，口唇の緊張(tighten)，口唇の引延(stretch)となっている．このうち発声時に大きく変化するのは口形の延伸と収縮(elongate/de-elongate)である．突出し(protrude)も発声時の特徴としては重要な情報を含んでいる[15]が，ここでは正面からの視覚情報だけを考慮するため直接的な特徴量としては使えない．しかし，突出しは「すぼめ」となって，収縮と似た変形を起こす．そのほかに，下顎の上下による口唇の開閉も発声時には重要な動作である．従って，口唇の上下方向の開閉，左右方向の伸縮が発声にかかわる主な二つの動きと考えられる．すなわちこのような動きにかかわる筋肉として，図1に示す頬筋，口輪筋と顎を動かすそしゃく筋に注目する．

2.2 口唇画像のオプティカルフロー

このような，筋肉と発声・表情のかかわりを考慮した上で，学習サンプルとしてとった発声中の口唇画像のオプティカルフローを解析した．オプティカルフローの抽出には，いろいろな方法があるが，隣接フレームだけから速度場が求まることと，パターンマッチングに必要な特徴点が特にいらないという理由から，我々は Horn-Schunk[16] のグラディエント法を使った．すなわち，時刻 t の画像を $I(x, y, t)$ で表し，微少時間では画像は平行移動していると仮定すると，その時間微分がゼロになることから，速度 (u, v) に関する1次方程式が求まる：

$$\frac{dI}{dt}=I_x u+I_y v+I_t=0 \qquad (1)$$

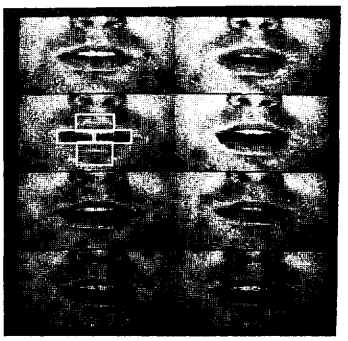
ここで，I_x，I_y，I_t は I の，x，y，t に関する偏微分である．これに，局所的にオプティカルフローのグラディエントが最小になるように制約条件を付けて解くことによって，各点の速度を求めることができる：

$$\left(\frac{du}{dx}\right)^2+\left(\frac{du}{dy}\right)^2\longrightarrow\min,$$

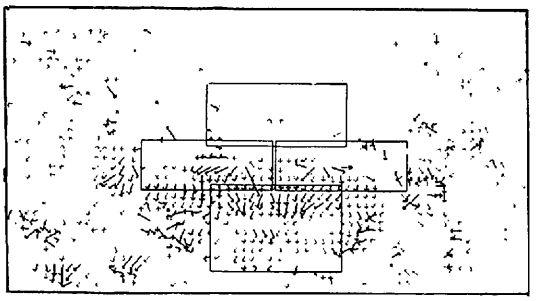
$$\left(\frac{dv}{dx}\right)^2+\left(\frac{dv}{dy}\right)^2\longrightarrow\min \qquad (2)$$

頬筋，口輪筋および顎の動きをとらえるために口唇の上下左右に窓を設定し，各窓内のオプティカルフローの平均値を計算し代表的な動きデータを測定した．図2(a)は口唇画像の1例で，窓が白枠で表示してある．図2(b)は画素ごとのオプティカルフローを矢印で示す．

図中のオプティカルフローは口を開く時点をとらえ



(a) Input Images and Window Positions



(b) Optical-flow around mouth

図2 入力画像とオプティカルフロー
Fig. 2 Input images and flow field.

ており，下唇，下顎，口唇両端に大きな動きが見られる．各窓の (x, y)-軸方向の速度成分 (u, v) を用いて，8次元の特徴量で各時刻の動きを記述できる．すなわち時刻 i の特徴ベクトル $\boldsymbol{x}^i$ は[上，下，左，右]の各窓の平均速度成分 $[(u_a(i), v_a(i)), (u_b(i), v_b(i)), (u_l(i), v_l(i)), (u_r(i), v_r(i))]$ を要素とする：

$$\boldsymbol{x}^i=(u_a(i)\ v_a(i)\ u_b(i)\ v_b(i)\ u_l(i)\ v_l(i)\ u_r(i)\ v_r(i)) \qquad (3)$$

これらの成分は互いに独立ではないので，学習サンプルに対して主成分分析を行い，真の特徴ベクトルを求めた．

すなわち，上記の $\boldsymbol{x}^i$ の分散行列 S を求め，その固有ベクトルを調べた．

$$S=\sum_{i=1}^{N}(\boldsymbol{x}^i-\bar{\boldsymbol{x}})^T(\boldsymbol{x}^i-\bar{\boldsymbol{x}}) \qquad (4)$$

ここで，N は学習サンプル全体のフレーム数 $(N\approx 700)$，$\bar{\boldsymbol{x}}$ は $\boldsymbol{x}$ の平均値，$(\cdot)^T$ は転置行列である．

その結果，S の固有ベクトルは，

$$\begin{bmatrix} \underline{e_1} \\ \underline{e_2} \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{bmatrix} = \begin{bmatrix} -0.51 & 0.00 & 0.74 & \underline{0.94} \\ 0.20 & \underline{1.00} & 0.08 & 0.13 \\ 1.00 & 0.02 & 0.60 & -0.48 \\ -0.58 & -0.23 & 1.00 & -0.14 \\ -0.06 & 1.00 & 0.19 & -0.31 \\ -0.61 & 0.44 & 0.05 & -0.04 \\ 0.05 & 0.16 & 0.13 & 1.00 \\ -0.02 & 0.01 & -0.04 & -0.12 \end{bmatrix}$$

$$\begin{bmatrix} 0.47 & \underline{1.00} & 0.02 & \underline{1.00} \\ \underline{-0.83} & 0.08 & \underline{1.00} & 0.01 \\ 0.41 & -0.35 & 0.17 & -0.35 \\ -0.33 & -0.10 & -0.02 & -0.06 \\ 0.01 & 0.06 & -0.96 & -0.12 \\ 1.00 & -0.02 & 0.51 & -0.15 \\ 0.03 & -0.66 & -0.23 & -0.42 \\ 0.05 & -0.87 & 0.12 & 1.00 \end{bmatrix} \quad (5)$$

となり，またその固有値は，

$(\lambda_1\ \lambda_2\ \lambda_3\ \lambda_4\ \lambda_5\ \lambda_6\ \lambda_7\ \lambda_8)^T$

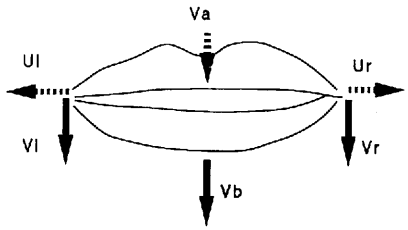
$=(3.57\ 2.41\ 1.23\ 0.45\ 0.19\ 0.14\ 0.03\ 0.01)^T$

となった．第1の固有値が全体の約45％，第2が約30％を占め，これら二つを主成分ベクトルとしてみなすことができる．そこで，第1，2固有ベクトル e_1，e_2 の各成分のうち，絶対値が1.0に近いものだけ（式(5)の下線部）を残すように量子化すると，二つの特徴量を得られる：

$$O(t)=v_b+v_l+v_r, \quad (6)$$

$$E(t)=\kappa v_a-u_l+u_r \quad (7)$$

ここで，κ は速度成分のアスペクト比である．これを図示すると，図3のようになり，$O(t)$ は顎の上下による口の開閉，$E(t)$ は口唇の伸縮に対応することがわかる．これは，オプティカルフローによる特徴抽出が筋肉の動作による変形をとらえるのに適していることを示す．



$$O(t) = v_b(t) + v_l(t) + v_r(t)$$

$$E(t) = \kappa v_a(t) - u_l(t) + u_r(t)$$

図3　特徴量 $O(t)$ と $E(t)$

Fig. 3　Principle motions; $O(t)$ and $E(t)$.

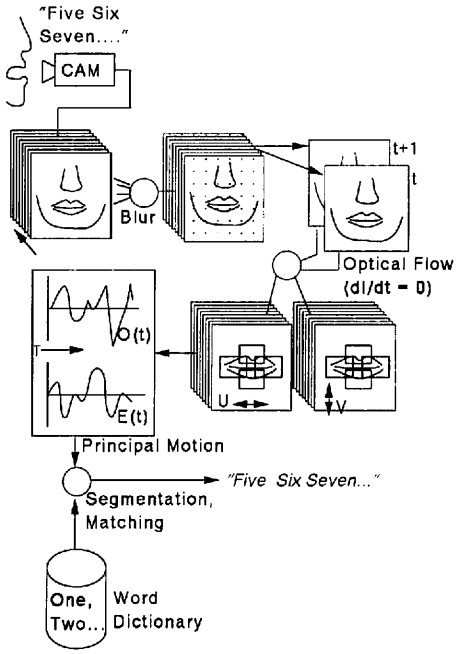


図4　オプティカルフローを用いた読唇システム

Fig. 4　Block diagram of optical-flow based lipreading system.

3. 英数字認識システム

3.1 画像入力と前処理

英数字読唇システムの構成図を図4に示す．動画像データはCCDカメラを使って画像処理装置（Detacube）でディジタイズする．本装置は最大256枚の128×64画素（8 bit/画素）の大きさの画像を1度（60枚/秒）に入力することができる．入力画像に対し，ノイズ除去のため時空間ガウスフィルタを掛ける．60 Hzで入力することによって，フレーム間の口唇の動きは十分小さくなり，簡単なオプティカルフロー抽出アルゴリズムが適用できる．前述のとおり，Horn-Schunk[16]の方法で各点の動きベクトルを推定する．窓を設定して，マクロな動きをとるので，推定のための繰返し計算は3回で打ち切ることにした．結論から言って，繰返しの回数はこの程度で十分である．文献[16]には，時刻 t の結果を使って時刻 $t+1$ のオプティカルフローを計算する手法も示されているが，舌や歯の出現などによる動きの不連続な場合も想定して，オプティカルフローの計算は隣接する2フレームのみ使い，時間に対して独立に行った．特徴量 $O(t)$，$E(t)$ を2.2に基づいて計算し単語のマッチングを行う．図5は，学習データの /five/ から /eight/ の発声の際の特徴量をプロットしたもので

ある．

3.2 単語検出とマッチング

図5に示す特徴量と原画像を見比べると，確かに単語の境界で $O(t)$ がゼロになっている．$O(t)=0$ となるのは，図5に示すように単語の境界以外にもあり，これは口唇の開閉が逆方向に動いたり，止まるときが単語境界以外にもあることを示す．これを使うと，音声信号で言う音素 (phoneme) に対応する視覚情報のプリミティブ (1種の Visime であるが，文献[17]の定義とは異なる) を隣接する $O(t)=0$ で決まる区間で定義でき，一つの単語は複数のプリミティブから構成されるセグメントであると考えられる．従って，学習データから単語セグメントの辞書を作成しておいて，対象とする実験データから一つ以上のプリミティブ列でセグメントを作って辞書とのマッチングを順に行うと，連続した単語識別が可能となる．ここでは，$O(t)$ と $E(t)$ の波形を単語別に16点で標本化して，辞書を作成した．

マッチングは，実験データから得られる特徴量から，単語セグメントとなる区間を仮定し，区間内をそれぞれ16点で標本化して時間軸の伸縮を行い，マッチング尺度 M を計算する．時刻 t_i から t_j の区間を単語 w と対応づけたときのマッチング尺度 $M(t_i, t_j, w)$ は辞書 (単語 w) の特徴パターン $\hat{O}_w(\tau)$，$\hat{E}_w(\tau)$ と実験データとの重み付き2乗誤差の和として計算される．すなわち，

$$M(t_i, t_j, w) = \sum_{\tau=0}^{\tau=15} (O(\tau')-\hat{O}_w(\tau))^2+\lambda(E(\tau')-\hat{E}_w(\tau))^2 \quad (8)$$

ここで λ は固有値 λ_1，λ_2 の比，$\tau'=t_i+(\tau/16)(t_j-t_i)$，$(t_i<t_j)$，$O(\tau=0)=0$ である．

図6に単語のマッチング手順を示す．マッチングは発話の先頭から順に単語の分離と識別を同時に行う．単語辞書とのマッチングで十分小さい $M(t_i, t_{i+1}, w)$ となるセグメント (区間 $[t_i, t_{i+1}]$) と，単語 w を探し，そのセグメントの終点 t_{i+1} を次の単語セグメント候補の始点として，繰り返し行う．最終的に一連の発話に対して，マッチング尺度の総和が最小となる単語列を認識結果とする：

$$\min_{(t_0,t_1,\cdots,t_n),(w_1,w_2,\cdots,w_n)} \sum M(t_i, t_{i+1}, w_n) \quad (9)$$

4. 実験結果

3人の被験者を対象に初期実験を行った．1人 (被験者1) からは学習データと実験データの両方をとり，学習データから辞書を作成した．この辞書をすべての被験者の実験データとのマッチングに用いた．また別の1人 (被験者3) には口ひげがあった．

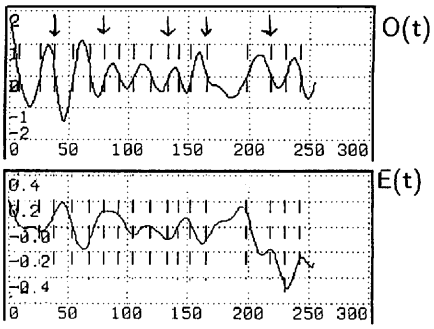


図5 /five/ から /eight/ までの数字の連続発声のときの $O(t)$ と $E(t)$．$O(t)$ のゼロクロッシングが破線で示してある．正確な単語の境界は矢印で示す．これから，テンプレートを作る

Fig. 5 An example of template pattern for /five/ through /eight/. Broken lines show zero crossings of $O(t)$. Arrows show correct word boundaries.

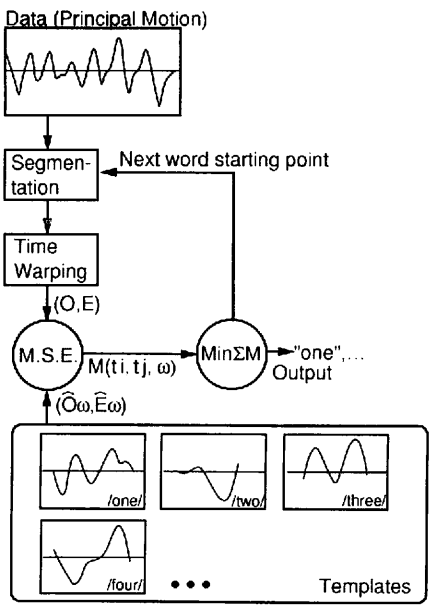


図6 単語のマッチング手順

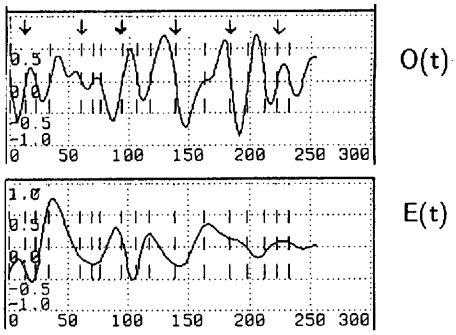
Fig. 6 Block diagram of word matching.

3から5個の数字 (英語) を含む発話を複数回収集し認識実験を行った．被験者は通常の照明で約2メートルの距離から撮影し，被験者には通常の速度で発声するよう指導した．全体では6回の発話 (延べ単語数=21) による実験となった．また，頭部を特に固定する装置は使用しなかったが，全体の揺れを軽減するために，

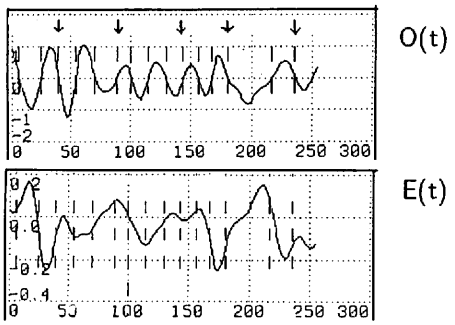
表1 単語認識結果-被験者1の辞書とのマッチング
(被験者1-[A, B, C, D], 2, 3)

実験データ	区間 開始点 (t_s)	区間 終了点 (t_e)	認識結果 (w)	マッチング尺度 (M)	正解	備考
1-A	11	84	one	5.43	one	
	84	166	two	5.65	two	
	166	232	three	4.91	three	
1-B (図7(a))	33	94	six	7.61	one/two	error
	94	138	three	4.43	three	
	138	184	four	3.53	four	
	184	222	five	2.97	five	
1-C (図7(b))	22	88	six	8.19	five	error
	88	143	six	5.93	six	
	143	180	seven	5.24	seven	
	180	236	eight	4.30	eight	
1-D	17	91	—	—	eight	n/f
	91	145	nine	3.72	nine	
	145	215	zero	6.18	zero	
2	9	68	one	5.64	one	
	68	134	two	5.89	two	
3	—	—	—	—	one	n/f
	-	-	--	—	two	n/f
	158	197	three	8.11	three	
	197	251	four	6.19	four	

n/f：not found，被験者3：ひげあり



(a)An utterance of /one/ though /five/ (sample 1-B)



(b)An utterance of /five/ though /eight/ (sample 1-C)

図7 被験者1の実験データの特徴量の例
Fig. 7 Examples of experimental data of subject #1.

後頭部を壁にもたれかけるように指導した．

実験データの詳細と認識結果を表1に，被験者1の実験データから得られた特徴量をプロットした図を図7に示す．これから認識率を集計すると，[被験者1]—73％(11/15)，[被験者2]—100％(2/2)，[被験者3]—50％(2/4)となった．

認識結果を解析すると，ほとんどの失敗は各発声の第1単語で起こっている．表からは読み取れないが，特に第1単語の開始点の識別が困難となっている．すべての区間でマッチング尺度を試算してみると正しい単語の開始点からは，実際に発話された単語の辞書パタンに対してマッチング尺度が最小となることから，音声情報などにより，単語の開始点に関する情報が提供されれば，単語の認識率を上げることができる．試算によると，被験者1では93％(14/15)，被験者3では75％(3/4)になる．

5. むすび

読唇の仕組みを発声学に基づいて検討し，コンピュータによる読唇の一手法を提案した．口唇の周りのオプティカルフローを解析してシステムを構築し，不特定話者，連続発声の単語認識が通常の撮影条件で可能であることを示した．文脈のない言葉に対する人間の読唇の能力は50から70％前後であるとも言われ，それと比較して，かなり高い認識率を得られることがわかった．

本手法の特に注目すべき点は，オプティカルフローを原情報とすることにより連続発声した単語の境界の候補となる時点を簡単に得られることである．また，オプティカルフローを計算することにより，これまで煩わしかった口唇の形状を正確に抽出する必要がなくなった．特定の単語を発声させるための筋肉の動作の

話者独立性については，発声生理学での検討が必要となるが，今後実験データを増やすことにより実験的に検証できると思われる．被験者2，3についてはごくわずかの実験データによる結果のみを示したが，被験者1の辞書パターンに対して約70％の認識率を得たことは十分可能性があると言える．なお田村ら[9]らも，読唇のためにオプティカルフローを利用することを本研究とは独立に行っているが，その目的はスプラインモデルのフィッティングのための補助的情報として使っており，本論文で示したようなオプティカルフローの利点が生かされていない．

一方，本論文で示した実験結果は少量のデータに対して成功しており，手法の一般性の評価については今後の実験データの追加によるところが大きい．オプティカルフローの平均化のために設定した窓は現在，会話的に各発話に対して1回，設定しているが，鼻孔の検出ができれば自動位置合せも可能である．また頭部全体の揺れがフローに及ぼす影響については検討が必要である．例えば，頭部全体の動きを検出して，その動きで補償する手法が考えられる．

本手法で使ったマッチング処理は単純なもので，処理は高速であるが現在最適とは言えない．調音結合や発音の伸縮等についての考慮を含めることにより認識率を向上できると思われる．また辞書として用いるパターンは現在，認識データと同様に /one/ から /zero/ までを連続発声して得られたパターンを使っている．このため調音結合のある単語の認識能力については不明である．また，語いを増やすためには，窓を増やすなりして特徴空間の次元を上げる必要があると思われる．

本論文では示さなかったが，英語を母国語としない被験者のデータも採取し，被験者1の辞書に対して同様の実験を行ったところ被験者3と同程度の結果を得た．これは，以下のことを示唆する．(1)地域や原言語発声体系の違いにより生ずるアクセントや発声方法の違いに対しては，話者依存性がある．(2)一方で，ある程度の類似性もある．(3)標準的な発声方法との違いを生理学的，視覚的に示すことができ，発声法の治療，他言語の発声法の学習等に利用できる．この実験結果の詳細については別の機会に報告する．

辞書とのマッチング方法，文脈に対する知識の応用など，自然言語理解へ向けた検討は，従来音声認識で行われてきた手法を取り入れることが必要となろう．実用的には，本システムを音声認識システムに組み込んで，視覚情報と聴覚情報を組み合わせることによって，より強固な言語認識システムを構築することが可能となろう．本論文では，筋肉の動きを推定して言語認識を行うシステムを構成したが，図2(b)の例にもあるように，頬のあたりの特徴点の無い場所でも動きベクトルが観測できる．これらを使って表情認識への発展も可能である．

謝辞 研究の協力を頂いた米国MITメディア研究所Vision Science Groupの諸氏に感謝する．MIT電子研究所鈴木規子博士には発声学に関して貴重な助言を頂いた．日ごろ御指導を頂く，視覚情報研究部小林幸雄部長，石井健一郎主幹研究員，末永康仁主幹研究員に感謝します．

文　献

(1) Y. Fukuda and S. Hiki: "Characteristics of the mouth shape in the production of Japanese — stroboscopic observation", J. of Acoust. Soc. Jpn., (E)3. 2. pp. 75-91 (1982).
(2) H. McGurk and J. McDonald: "Hearing lips and seeing voices", Nature, 264. pp. 746-748 (1976).
(3) 積山　薫，東倉洋一："読唇情報が音声知覚に果たす役割"，テレビ学技報，13. 44. pp. 31-36 (1989-09).
(4) E. D. Petajan: "Automatic lipreading to enhance speech recognition", PhD thesis, U. of Illinois (1984).
(5) 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，22. 2. pp. 191-198 (昭61-02).
(6) 栗田知好，本多清志，垣田有紀："口唇画像情報を併用する音声の分析"，信学技報，SP88-94 (1988).
(7) B. Petajan, E. D. Bischoff and D. Bodoff: "An improved automatic lipreading system to enhance speech recognition", ACM SIGCHI'88, pp. 19-25 (1988).
(8) E. K. Finn and A. A. Montgomery: "Automatic optically-based recognition of speech", Pattern Recognition Letters, 8. 3. pp. 159-164 (1988).
(9) 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"，信学技報，PRU89-20 (1989-06).
(10) O. Fujimura: "Bilabial stop and nasal consonants: a motion picture study and its acoustical implications", J. of Speech and Hearing Research, 4. 3. pp. 233-247 (1961).
(11) K. Mase and A. Pentland: "Automatic lipreading by computer", 信学画像理解の高度化と高速化シンポジウム予稿，pp. 65-70 (平1-04).
(12) 上條雍彦："口腔解剖学　第2巻"，アナトーム社 (1974).
(13) K. Waters: "A muscle model for animating three-dimensional facial expression", Computer Graphics (SIGGRAPH '87), 21. 4. pp. 17-24 (1987).
(14) P. Ekman and W. V. Friesen: "The Facial Action

Coding System". Consulting Psychologists Press, Inc., San Francisco, CA (1978).
(15) J. S. Perkell : "Coarticulation strategies : preliminary implications of a detailed analysis of lower lip protrusion movements". Speech Communications, 5, pp. 47-68 (1986).
(16) B. K. P. Horn and B. G. Schunk : "Determining optical flow". Artifitial Intelligence, 17, pp. 185-203 (1981).
(17) D. W. Massaro : "Speech Perception by Ear and Eye : A Paradigm for Psychological Inquiry". Lawrence Erlbaum Accoc. Publ., New Jersey, USA (1987).

(平成元年9月28日受付)

間瀬　健二

昭54名大・工・電気卒，昭56同大大学院修士(情報)課程了．同年日本電信電話公社入社．以来，画像情報システム，コンピュータグラフィックス，画像処理の研究に従事．現在NTTヒューマンインタフェース研究所視覚情報研究部主任研究員．1988～89年米国MITメディア研究所客員研究員．IEEE，情報処理学会各会員．

アレックス　ペントランド

1982年米国MITにて博士号取得．同年SRI・AIセンタに入所．1983年からスタンフォード大講師を兼任，1986年優秀講師賞を受賞．1987年MITメディア研究所準教授．人工知能，画像理解，人間の視覚機能，コンピュータグラフィックスの研究に従事．1984年米国人工知能学会(AAAI)よりテクスチャと形状の認知に関する研究で最優秀論文賞．著書「From Pixels to Predicates」(編)，「The Vision Machine」．