横顔の動画像情報を用いたマルチモーダル音声認識

吉永智明　田村哲嗣　岩野公司　古井貞熙

東京工業大学大学院 情報理工学研究科 計算工学専攻
〒 152-8552 東京都目黒区大岡山 2-12-1
Email: {yossy, tamura, iwano, furui}@furui.cs.titech.ac.jp

本論文では，横顔から抽出された口唇動画像情報を利用した，雑音に頑健な音声認識手法を提案する．これまでのマルチモーダル音声認識手法では，主に顔の正面から撮影された口唇画像を用いているが，モバイル環境で利用することを考えると，ユーザは発話をしながらカメラ付き携帯電話を顔の正面で持たなければならず，音声入力が困難である．提案手法は，携帯電話のマイク部分に小型カメラを搭載し，その映像を用いることを想定しており，自然なスタイルで容易に音声と画像を取り込むことができる．画像特徴量はオプティカルフロー解析によって抽出される．フレームごとに画像特徴量を音響特徴量と結合し，マルチストリーム HMM を用いて認識を行う．白色雑音を重畳した連続数字音声による認識実験を行ったところ，画像情報を利用することによって，様々な SN 比条件で数字正解精度の改善が確認された．SN 比 5dB の時に画像情報の効果が最も高く，正解精度の改善は約 6% であった．

A Multi-Modal Speech Recognition Using Side-Face Images

Tomoaki Yoshinaga, Satoshi Tamura, Koji Iwano, and Sadaoki Furui

Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
Email: {yossy，tamura, iwano, furui}@furui.cs.titech.ac.jp

This paper proposes an multi-modal speech recognition method using lip movement extracted from side-face images for increasing noise-robustness in mobile environments. Although most previous multi-modal speech recognition methods use frontal face (lip) images, these methods are not easy for users since they need to hold a device with a camera in front of their face when talking. Our proposed method capturing lip movement using a small camera installed in a handset is more natural, easy and convenient. Visual features are extracted by optical-flow analysis and combined with audio features. HMMs are built by the multi-stream HMM technique. Experiments conducted using connected digit speech contaminated with white noise show improvement of digit accuracy by using the visual information in various SNR conditions. The best improvement is approximately 6% at 5dB SNR.

1　はじめに

モバイル環境における音声入力は非常に手軽かつ便利に行うことが出来るため，近年このようなデバイスを用いた音声認識技術の必要性が高まって来ている．しかし，モバイル環境下では周辺雑音の影響が大きく，音声認識を行う際に問題となる．そのため，雑音に頑健な音声認識手法が求められている．

そこで，音響雑音の影響を受けない発声時の口唇の動画像から得られる情報を，音声情報とともに利用するマルチモーダル音声認識システムが注目され，近年その研究が進められている [1-3]．これらの研究では，正面から撮影された口唇画像が用いられている．しかし，モバイル環境下でこのような方式を利用しようとすると様々な問題が生じる．カメラ付き携帯電話で音声と画像を入力することを考えると，ユーザは発話しながら携帯電話を顔の正面に持ってカメラ撮影を行うことになり，負担が大きく，自然な音声入力を行うことができない．また，撮影のために携帯電話と口との距離を離さなければならず，音声の SN 比が劣化するといった問題も生じる．

そこで本研究では，正面の顔画像ではなく，横顔

の動画像情報を用いたマルチモーダル音声認識手法を提案する．この方法では，携帯電話などのモバイル機器のマイクロフォン部分に微小カメラを搭載し，口唇動画像情報を取得することを想定している．横方向からの口唇の動き情報を音声認識に利用することによって，モバイル環境下において自然な形で音声入力が可能で，雑音に頑健な音声認識を行うことができる．

横顔の口唇画像からの特徴量の抽出にはオプティカルフローを用いる．オプティカルフローは口唇の動きの情報を反映しており，口唇の形状情報を利用するよりも，単語境界情報を含んだ有益な特徴量の抽出，不特定話者への適用といった点で有効であるという報告がある [4]．我々はこれまでに，正面から撮影された口唇画像から画像特徴量としてオプティカルフローを抽出し，音声認識を行う手法を提案し，その有効性を確認している [5]．そこで，この手法 [5] に基づいて横顔の動画像情報を用いたマルチモーダル音声認識システムを構築する．

2　オプティカルフロー

オプティカルフローは「画像中の明度パターンの見かけ上の速度分布」と定義される．本研究では，最も一般的な Horn-Schunck のアルゴリズムによりオプティカルフローを計算した [6, 7]．

今，ある時刻 t における物体上の点 (x, y) の明度が，微小時間内においては不変であると仮定する．

$$\frac{dI}{dt} \simeq \frac{\partial I}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial I}{\partial y} \cdot \frac{dy}{dt} + \frac{\partial I}{\partial t} = 0 \qquad (1)$$

ここで，$I(x, y, t)$ は時刻 t における点 (x, y) の明度である．さらに，$u = dx/dt$，$v = dy/dt$ とおくと，式 (1) は，

$$I_x \cdot u + I_y \cdot v + I_t = 0 \qquad (2)$$

となる．これは「オプティカルフローの拘束式」と呼ばれ，$u(x, y)$，$v(x, y)$ はそれぞれ点 (x, y) のオプティカルフローのベクトルの水平成分，垂直成分となる．式 (2) だけではフローベクトルを決定できないので，フローベクトル全体の自乗和が最小となるよう，次式で表される新たな拘束式を導入する．

$$\iint \left\{ ({u_x}^2 + {u_y}^2) + ({v_x}^2 + {v_y}^2) \right\} dxdy \rightarrow \min \quad (3)$$

表 1: 音響特徴量，画像特徴量

音響	フレーム長	: 25ms
	フレーム周期	: 10ms
	抽出特徴量	: MFCC 12 次元，
		: その $\Delta, \Delta\Delta$ 成分
		: $\Delta, \Delta\Delta$ 対数パワー
	特徴量次元数	: 38 次元
画像	フロー演算繰り返し回数	: 5 回
	抽出特徴量	: フロー垂直成分の分散値
		: フロー水平成分の分散値
	特徴量次元数	: 2 次元

式 (2)(3) から，変分法に基づいて，以下の繰り返し演算により，フローベクトルを推定することができる．

$$u_{p,q}^{k+1} = \bar{u}_{p,q}^{k} - \mu \frac{I_x \bar{u}_{p,q}^{k} + I_y \bar{v}_{p,q}^{k} + I_t}{1 + \mu(I_x^2 + I_y^2)} I_x \qquad (4)$$

$$v_{p,q}^{k+1} = \bar{v}_{p,q}^{k} - \mu \frac{I_x \bar{u}_{p,q}^{k} + I_y \bar{v}_{p,q}^{k} + I_t}{1 + \mu(I_x^2 + I_y^2)} I_y \qquad (5)$$

ここで $\bar{u}_{p,q}$，$\bar{v}_{p,q}$ はそれぞれ点 (p, q) における u，v の近傍平均値，k は繰り返し数であり，μ は明度の精度によって決まる定数で，本研究では経験的に 0.01 とした．この手法には，時間的に連続した 2 枚の画像のみから計算できること，物体の形状といった事前知識が不要であること，およびパターンマッチングを用いないので特徴点抽出が不要であること，といった利点がある．

3　マルチモーダル音声認識システム

3.1　特徴量抽出・融合

本研究で構築したマルチモーダル音声認識システムの流れを図 1 に，音響，画像特徴量それぞれの概要を表 1 に示す．16kHz でサンプリングされた音声データを，100Hz で 12 次元の MFCC と，その 1 次，2 次微分，そして対数パワーの 1 次，2 次微分の計 38 次元のパラメータに変換したものを音響特徴量として用いる．

動画像データは，市販のビデオカメラを用い，携帯電話の受話口付近にカメラがあるように見立てて，口唇の右下部分から，やや見上げるような角度で撮影した (図 2(a)，(b))．画像は毎秒 15 フレーム

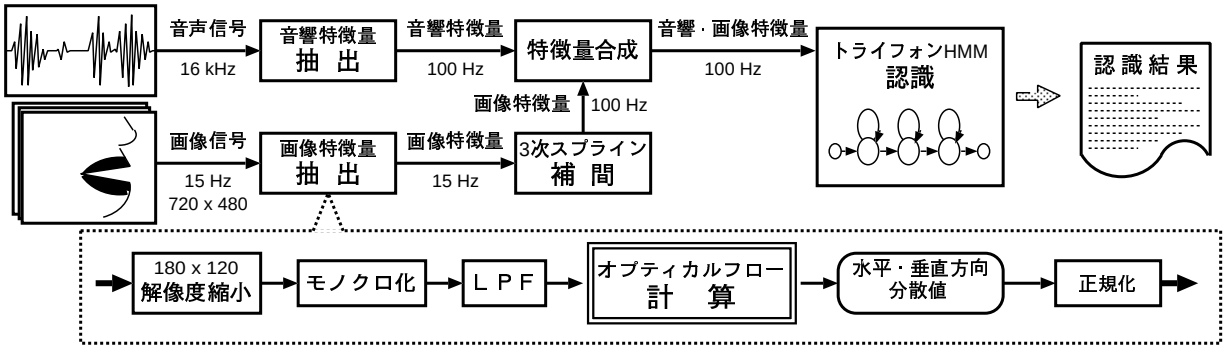

図 1: 横顔マルチモーダル音声認識システム

(a) 横顔画像例　(b) (a) の 1 フレーム後の画像　(c) オプティカルフロー

図 2: 横顔画像の例

でキャプチャされ，720 × 480 の 24bit カラー画像となる．これを計算量削減のために 180 × 120bit に変換し，さらにエッジや明度の平坦な部分におけるフローベクトルの抽出精度を向上させるために，ローパスフィルタリングと低レベルのランダム雑音付加を行う．こうして得られた画像に対して．時間的に隣接する 2 フレームの画像を用いてオプティカルフローを計算した．図 2(c) に，オプティカルフローの計算結果を示す．これは図 2(a)，(b) の連続する 2 フレームの画像からオプティカルフローの値を計算し，図示したものである．

得られたフローベクトルの水平・垂直成分の分散値を，フレーム画像ごとに計算し，入力発話毎に最大値によって正規化して 2 次元の画像特徴量を得る．この特徴量は，無発声時にはフローベクトルが観測されないので，0 に近い値となり，発声時には口唇の開閉方向にフローベクトルが表れるので値が大きくなる．したがって，この特徴量は発声時と無発声時の判別に有効である．図 3 に「7102，9134」と発声したときのオプティカルフローの垂直成分の分散値のグラフを示す．これからも，発声している時には特徴量は大きな値を取り，無発声時はほぼ 0 に近い値を取っていることが見て取れる．なお，水平成分の分散値も，ほぼ同じ傾向であった．

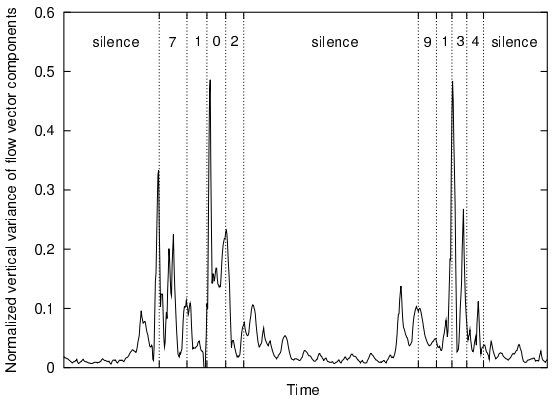

図 3: オプティカルフローの垂直成分の分散値

以上のようにして得られた，音響特徴量 38 次元と画像特徴量 2 次元をフレームごとに結合し，40 次元の音響–画像特徴量を作成し，この特徴量を用いて HMM による認識を行う．なお，フレーム周期は音響特徴量と同じ 10ms とし，それに合わせるため，画像特徴量は 3 次元スプライン関数によって補間を行った．

認識に用いる triphone HMM は，音響特徴量と画像特徴量を別のストリームとしたマルチストリーム HMM でモデル化する．この HMM において，状態 j における音響–画像特徴量 O_{AV} を観測する確率 $b_j(O_{AV})$ は式 (6) で表される．

$$b_j(O_{AV}) = b_{Aj}(O_A)^{\lambda_a} \cdot b_{Vj}(O_V)^{\lambda_v} \qquad (6)$$

ここで $b_{Aj}(O_A)$，$b_{Vj}(O_V)$ はそれぞれ状態 j で音響特徴量 O_A，画像特徴量 O_V を観測する確率，λ_a，λ_v はストリーム重みである．λ_a，λ_v は，例えば雑音環境下では音響特徴量の信頼度が下がるので，相対的に λ_v を大きくするといったように，各々のストリームの信頼度に応じて変化させるパラメータとなっていて，λ_a+λ_v=1 という制約を設けている．

3.2 マルチストリーム HMM の構築

マルチストリーム HMM は，まず音響 HMM と画像 HMM をそれぞれ別々に学習し，それらを融合することで構築される．以下にその流れを示す．

1. 38 次元の音響特徴量を用いて学習を行い音響 HMM を作成する．なお sp (short pause) モデルのみ状態数を 1 とし，その他の HMM は全て状態数 3 である．

2. 音響 HMM を用いて強制切り出しを行い，各音素の時間ラベルを得る．

3. 画像 HMM を 2 次元画像特徴量と，2. で得られた音素の時間ラベルを用いて学習し，構築する．なお，画像 HMM では sp と sil (silence) のモデルを状態数 1 とし，その他のモデルは 3 である．

4. これら 2 つの HMM からマルチストリーム HMM を構築する．マルチストリーム HMM は，音響 HMM と同じ状態数である．マルチストリーム HMM における音響ストリーム中の各状態の混合分布は，同じ音素の音響 HMM の，対応する状態の混合分布をそのまま利用する．画像ストリームの混合分布についても同様に画像 HMM 中の混合分布を利用するが，sil モデルについては，マルチストリーム HMM では 3 状態であるのに対し，画像 HMM では 1 状態であるので，3 状態全ての画像ストリームの混合分布について，同一の混合分布を用いる．

なお，予備実験の結果から，全ての HMM で混合数を 2 とした．

4 実験

4.1 データベース

クリーン環境下で 38 名の日本人男性話者に対し収録を行い，音響–画像データベースを構築した．各話者には日本語で 4 連続数字の読み上げ 10 回を 1 セットとし，これを 5 セットずつ各話者に発声してもらった．なお，連続数字間には 2 秒程度のポーズ区間が挿入されている．横顔の口唇画像は，受話口付近に微小カメラが取り付けられたモバイル機器を使用していると仮定し，話者の右頬から 10cm 程度の場所にデジタルカメラを配し，撮影を行った．データベースの総時間長は約 2 時間である．

4.2 学習・認識

実験は，leave-one-out 法を用いて行った．すなわち全 38 名分のデータのうち 37 名分のデータを使い HMM の学習を行った．その後，この HMM に，残り 1 名のデータをテストセットとして用いて認識実験を行うこととした．これらをテストセットを変え，19 名分行い，それらの数字正解精度を求め，その平均値をモデルの性能評価に用いた．認識に用いるテストセットには，実際に clean 環境下で収録されたデータの他に，このデータに SN 比 5，10，15，20dB の白色雑音を付加したものを用いた．

4.3 実験結果

表 2 に各 SN 比における，音響特徴量のみを用いて認識を行った場合と，音響–画像特徴量を用いて認識を行った場合の数字正解精度をまとめた．音響と画像のストリーム重みは各 SN 比条件ごとに最適化を行っており，最適な音響ストリーム重みの値 (λ_a) を表の Audio-visual の括弧内に表記した．表 2 より，画像特徴量を用いたことで全ての SN 比

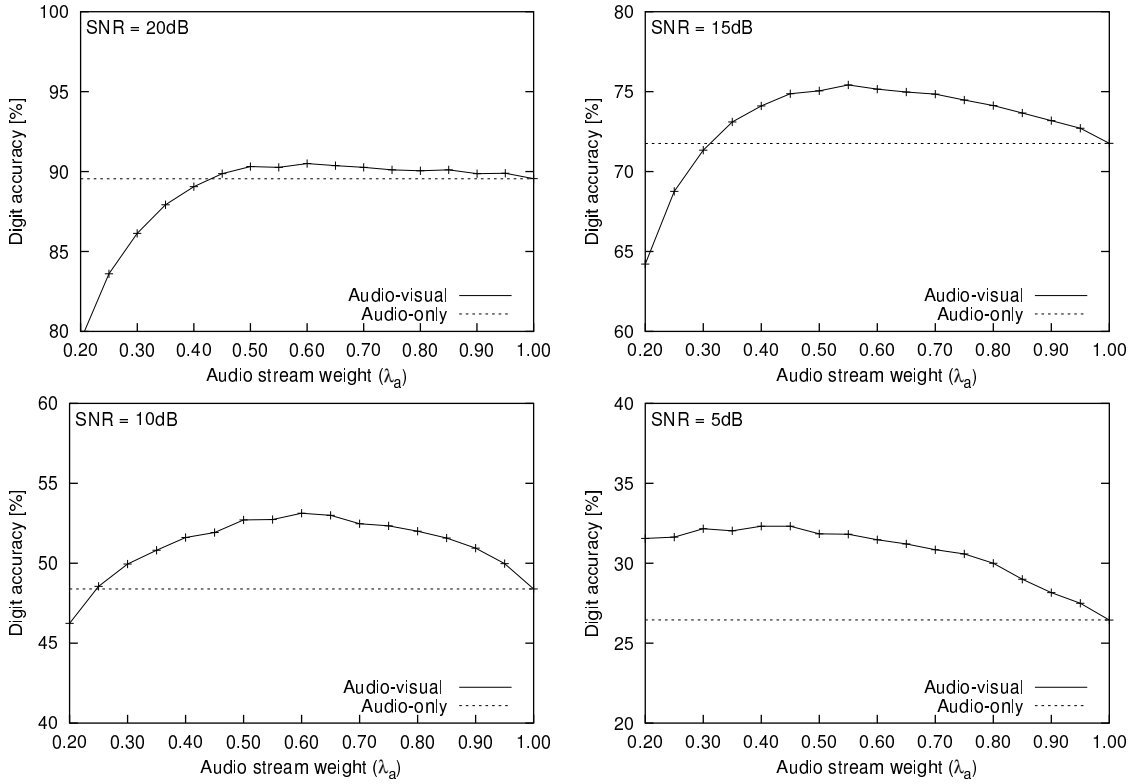

図 4: 音響ストリーム重みによる数字正解精度の推移

表 2: 各 SN 比における数字正解精度の比較

SN 比 (dB)	Audio-only	Audio-visual
∞ (clean)	99.3%	99.4% (0.80)
20	89.6%	90.5% (0.60)
15	71.8%	75.4% (0.55)
10	48.4%	53.1% (0.60)
5	26.4%	32.3% (0.45)

（括弧内は λ_a の値）

において正解精度が向上したことが分かる．SN 比 5dB の時に性能向上は最大となり，絶対値で 5.9% 数字正解精度が向上した．また，最適な λ_a の値は SN 比が小さいものほど小さくなる傾向がみられる．したがって，雑音が大きくなるほど画像情報に重点を置くことで，認識性能の最適化が行われていることが分かる．

図 4 に，各 SN 比における音声データに対して，音響特徴量のみでの正解精度，および画像特徴量を併せて用いた場合での λ_a を変化させていったときの数字正解精度の推移の様子を示す．全ての SN 比条件において，λ_a の変化に対し，緩やかに正解精度が変化していく様子が見て取れ，広い範囲のストリーム重みに対して画像特徴量の効果が表れていることがわかる．

これらの結果に対し，正面から撮影された口唇動画像から同様にオプティカルフローを用いて特徴量を抽出し音声認識を行った研究 [5] では，SN 比 10dB で 58.6 → 70.8%，SN 比 15dB で 78.3 → 87.7% という数字正解精度の向上が報告されている．本研究とはデータ量，収録環境などが異なるため単純に比較することはできないが，本研究ではこの半分程度の性能改善となっている．正面から撮影されたものに比べ，横顔では口唇の片面のみしか表れず，動き情報を十分に抽出できなかったことから，性能向上が小さかったものと考えられる．

4.4 発声開始時刻の検出性能

本研究で用いた画像特徴量は発声・無発声の境界検出に有効であると考えられ，その検出精度の向

表 3: 発声開始時刻の平均誤差

SN 比 (dB)	Audio-only	Audio-visual
20	41.2	35.1
15	54.9	45.9
10	76.2	63.1
5	104.1	86.6

（単位：ms）

上が雑音環境下での正解精度の改善に寄与したものと思われる．そこで，発声開始時刻の検出精度が，雑音環境下でどの程度改善されているか調べるための追加実験を行った．

まず，音響 HMM，マルチストリーム HMM それぞれを用いて，clean データとそれに雑音が重畳したデータの強制切り出しを行う．sil から数字に変わる時刻を発声開始時刻と定義し，clean データの発声開始時刻を正解として，雑音重畳データで検出された発声開始時刻の誤差（ms）を求める．音響特徴量のみを用いた場合と，音響–画像特徴量を用いた場合の, 各 SN 比条件における平均誤差（ms）を表 3 に示す．

画像特徴量を用いることで，全ての SN 比のデータにおいて平均誤差が削減されていることが分かる．SN 比 5dB，10dB の時に効果が大きく，約 17% 誤差が削減されている．表 2 の数字正解精度の改善も同様の傾向となっていることから，発声開始時刻の検出性能の改善が，雑音環境下における認識性能の改善に寄与しているものと考えられる．

5 まとめ

本研究では，モバイル環境における雑音に頑健な音声認識手法として，横顔の口唇動画像情報を利用したマルチモーダル音声認識手法を提案した．横顔の動画像を利用することにより，ユーザに自然な形での音声入力を提供することができる．提案手法を連続数字音声認識実験で評価したところ，様々な SN 比条件で画像特徴量を用いることによる耐雑音性の向上が確認された．最も性能改善が得られたのは SN 比 5dB の白色雑音条件で，絶対値で約 6% 数字正解精度が改善した．また，提案手法を用いることによって，雑音環境下における発声開始時刻の検出性能の向上が得られることが示された．

今後の課題としては，1) 実環境での利用を考慮し，画像情報の外乱成分に対する対策手法の提案と，その効果の検証，2) 雑音適応手法との組み合わせによる頑健性の向上，3) 音響・画像ストリーム重みの自動的な最適化手法の提案，4) 口唇の開閉だけでなく，音素種の識別にも効果のある画像特徴量の提案，などが挙げられる．

謝辞

本研究は NTT ドコモ株式会社の研究委託を受けて行われました．ここに深く感謝いたします．

参考文献

[1] C. Bregler and Y. Konig, “ “Eigenlips” for robust speech recognition,” *Proc. ICASSP94*, vol.2, pp.669–672, Adelaide, Australia (1994-4).

[2] G. Potamianos, E. Cosatto, H.P. Graf, and D.B. Roe, “Speaker independent audio-visual database for bimodal ASR,” *Proc. AVSP97*, pp.65–68, Rhodes, Greece (1997-9).

[3] 熊谷建一，中村　哲，猿渡　洋，鹿野清宏，“HMM 合成を用いたバイモーダル音声認識，” 2000 年秋季音講論，2-Q-11，pp.111–112 (2000-9)．

[4] 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，” 信学論 D-II, Vol.J73-D-II, No.6, pp.796-803 (1990-6).

[5] 田村哲嗣，岩野公司，古井貞熙，“オプティカルフローを用いたマルチモーダル音声認識法の提案と評価，” 情処研報, 2002-HI-97-6 / 2002-SLP-40-6, vol.2002, no.10, pp.33–38 (2002-2)．

[6] B.K.P. Horn and B.G. Schunck, “Determining optical flow,” *Artificial Intelligence*, vol.17, nos.1–3, pp.185–203 (1981-8).

[7] 浅田　稔, ダイナミックシーンの理解，電子情報通信学会, pp.16–30 (1994-3).