ヒューマンインタフェース 97−6
音声言語情報処理 40−6
（２００２．２．１）

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

田村 哲嗣　　岩野 公司　　古井 貞熙
東京工業大学　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
E-mail: {tamura,iwano,furui}@furui.cs.titech.ac.jp

音声情報に加え，唇動画像の情報を利用するマルチモーダル音声認識は，特に雑音環境下で頑健な認識性能を発揮できる手法として注目され，近年多くの研究が進められている．本研究では，パターンマッチングなどに比べ，より頑健な画像特徴量抽出方法として，口唇の抽出やモデリングを必要としないオプティカルフロー解析を利用する手法を提案する．本手法は，オプティカルフロー解析から得られる特徴量を，音響特徴量とパラメータのレベルで結合し，HMM をモデルとして音声認識を行うものである．本手法を用いて，雑音を重畳したデータを用いて認識実験を行った結果，音響特徴量のみの結果と比べ，SNR=15dB において約 46%の誤り率の削減に成功した．

A study on multi-modal speech recognition using optical-flow analysis

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
E-mail: {tamura,iwano,furui}@furui.cs.titech.ac.jp

The research on multi-modal speech recognition based on the combination of acoustic features and visual features such as lip information has recently become very attractive for the purpose of increasing the robustness of automatic speech recognition in noisy environments. This paper proposes a new method using the optical-flow analysis. Since optical-flow is computed without extracting the speakers' lip contours, the visual features can be robustly extracted. The visual feature set is combined with the acoustic feature set in the framework of HMM-based recognition. Our multi-modal ASR system has achieved a 46% relative reduction of digit recognition error rate compared with the audio-only recognition scheme in 15dB SNR condition.

1. はじめに

今日，音声認識を利用した多くの製品が実用化され普及するようになり，音声認識は最も注目されている技術のひとつとなってきている．しかし現在の音声認識技術は，雑音が大きい環境の下での認識性能が低く，これが音声認識の実用化における大きな問題となっている．そこで雑音下でも頑健に音声認識を行う手法のひとつとして，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，近年研究が進められている [1, 2, 3]．

マルチモーダル音声認識において，画像特徴量の抽出法は，口唇のモデルをベースとした方法と，ピクセル（画素）をベースとした方法の 2 つに大別される [4]．前者は，画像の中から口唇の輪郭を抽出して統計的にモデル化し，そこから得られるモデルパラメータを画像特徴量とする手法である．しかし，口唇を正確に抽出することは本質的に難しく，頑健に特徴量を抽出するためには，リップマーキングや，

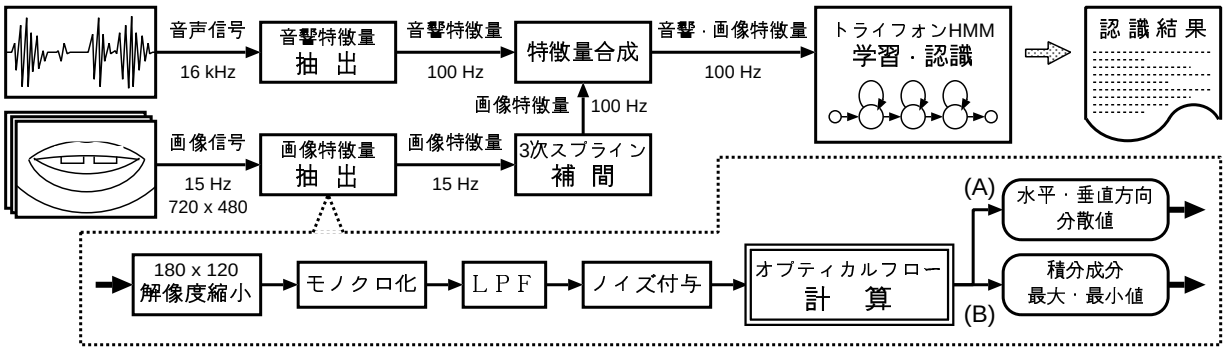


図 1: マルチモーダル音声認識システム

特殊な照明条件が必要となるなど，実用上の問題が多い．後者は，画像中のピクセル自身がもつ明度などの情報に，直接，フーリエ変換などの信号処理を施して画像特徴量を得る手法である．この方法は口唇モデルを仮定しないため，前者のような実用面での問題を回避することができる．

従来のピクセルベースの手法としては，FFT を用いたもの [1]，線形判別分析を利用したもの [3] などが提案されている．これらはいずれも 1 枚の静止画像ごとに，口唇の形状を反映した特徴量を抽出し用いている．しかし一方で，口唇の形状よりも，口唇およびその周辺の動きに注目した方が，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もなされている [5]．そこで我々は，口唇の動き情報を利用したオプティカルフローによる読唇 [5] に注目した．オプティカルフローとは明度の見かけの速度分布のことであり，連続した複数枚の画像から計算することによって，ベクトルという形で動き情報を抽出することができる．

以上から，我々は頑健な画像特徴量の抽出ができ不特定話者に適用可能な，オプティカルフローを用いたマルチモーダル音声認識の手法の提案を行う．

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．

$$\iint \Big\{ ({u_x}^2 + {u_y}^2) + ({v_x}^2 + {v_y}^2) \Big\} dxdy \rightarrow \min \qquad (3)$$

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．

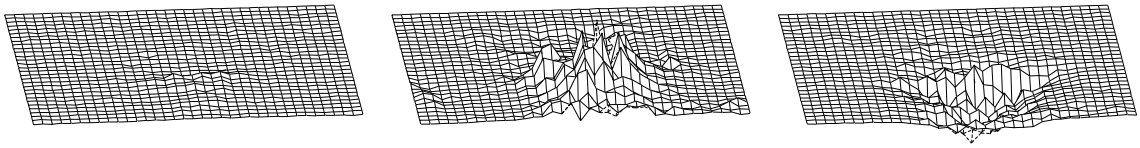
3. マルチモーダル音声認識システム

3.1. 特徴量抽出・融合

図 1 に，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示す．また表 1 に，使用した音響・画像特徴量について示す．音声データは 16kHz でサンプリングし，毎秒 100 フレームで 12 次元の MFCC と対数パワー，それらの一次・二次微分の計 39 次元のパラメータに変換し用いた．動画像は毎秒

(a) 時間的に連続したフレームの画像 (b) (c) フローベクトル

図 2: オプティカルフローの計算結果



(a) 無発声時 (b) 開口時 (c) 閉口時

図 3: オプティカルフローの積分結果

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，
		: 対数パワー，
		: これらの Δ，$\Delta\Delta$ 成分
	特徴量次元数	: 39 次元
画像	フロー演算繰り返し回数	: 5 回
	抽出特徴量 (A)	: フローの水平成分分散値
		: 〃 垂直成分分散値
	抽出特徴量 (B)	: フロー積分成分の最大値
		: 〃 最小値
	特徴量次元数	: いずれも 2 次元

15 フレームでキャプチャし，計算量削減のため，720 × 480 の 24bit カラー画像から，180 × 120 の 8bit グレースケール画像に変換した．さらに得られた画像に対して，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与を行った．その後，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．このオプティカルフローの計算結果の例を図 2 に示す．(a) はある時刻におけるフレームの画像，(b) はそれよりも 1 フレーム後の画像である．これらからオプティカルフローを計算し，これを図示したものが (c) である．

そして得られたフローベクトルから，2 種類の 2 次元特徴量 (A)，(B) を抽出し，比較・検討を行った．(A) ではフローベクトルの水平・垂直方向の分散値の計 2 次元を用いた．このパラメータは無発声時にはフローベクトルが観測されないので零となり，発声時には口唇の周りにのみフローベクトルが表れるので値が大きくなることを利用している．(B) ではフローベクトルを積分し，その最大値と最小値の 2 次元をもってパラメータとした．このオプティカルフローの積分結果の例を図 3 に示す．この曲面を $f(x,y)$ とすると，フローベクトルとは次の関係にある．

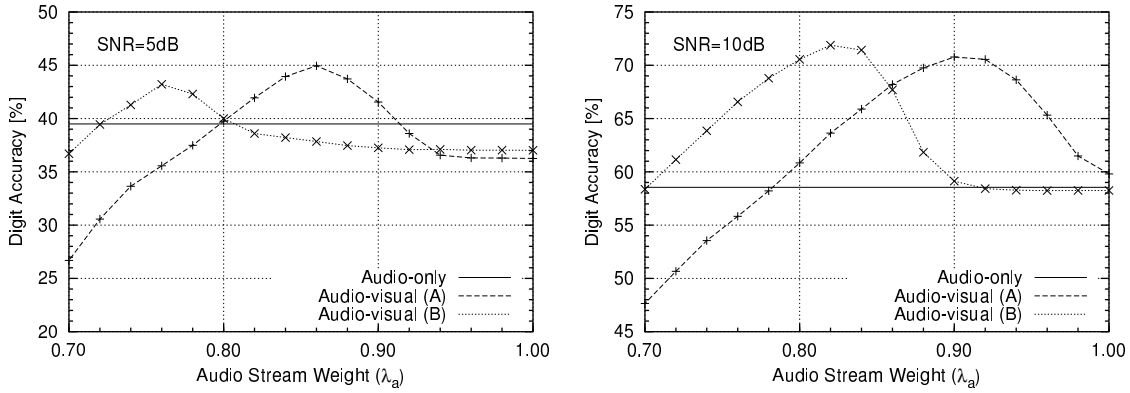
$$u(x,y) = \frac{\partial}{\partial x} f(x,y)\,, \quad v(x,y) = \frac{\partial}{\partial y} f(x,y) \qquad (6)$$

発声の際には口の動きに合わせ，開口時は拡散，閉口時は収束する方向にフローベクトルが観測される．(A) ではいずれの場合も同様に値が変化するが，(B) では図 3 からわかるように，無発声時の (a) に対して，開口時には (b) のように山ができ最大値に，閉口時は (c) のように谷ができ最小値に反映される．これより (A) は口の動きの有無，(B) ではさらに開閉の情報が加えられており，(A) よりも (B) の方が口の動きの情報をより反映していると考えられる．

以上で得られた 39 次元の音響特徴量と 2 次元の画像特徴量（(A)，(B) のいずれか一方）をパラメータレベルで融合し，41 次元の音響–画像特徴量を得た．ただし音響特徴量のフレーム周期に合わせるため，画像特徴量は 3 次スプライン関数により補間を行った．

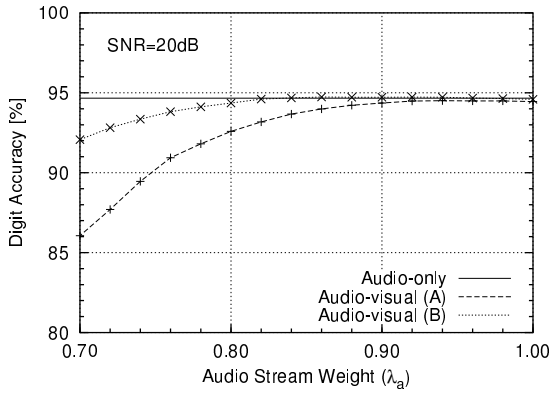
3.2. 学習・認識

モデルには，状態数 3，混合数 2 の left-to-right 型トライフォン HMM を用いた．この HMM は，学習時は，通常音声認識で用いられているシングルストリーム HMM であるが，認識時には音響ストリーム

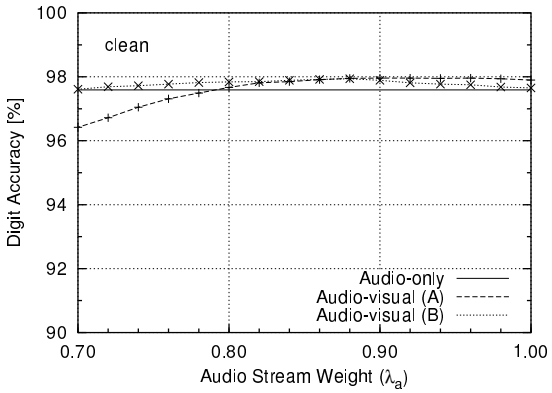


(a) SNR=5dB

(b) SNR=10dB



(c) SNR=20dB



(d) clean

図 4: 認識結果

と画像ストリームから成るマルチストリーム HMM に変換した．このとき，HMM の状態 j において音響–画像特徴量 O_{AV} を観測する確率 $b_j(O_{AV})$ は式 (7) で表される．

$$b_j(O_{AV}) = b_{Aj}(O_A)^{\lambda_A} \cdot b_{Vj}(O_V)^{\lambda_V} \qquad (7)$$

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る．

4. 実験

4.1. データベース

データベースは，本研究にあたり収録した，クリーン環境下における 11 名の男声話者による連続数字読み上げ音響–画像データを使用した．各話者は 2〜6 桁の数字を 250 個発声しており，全体の総時間長は約 2 時間半である．

4.2. 学習・認識

実験は，leave-one-out 法により行った．10 名分のデータを用い連結学習によって HMM の学習を行い，その後 HMM をマルチストリーム変換し，残る 1 名分のデータをテストセットとして認識実験を行った．この実験をデータの組み合わせを変えて 11 通り行い，それらの数字正解精度の平均値をモデルの性能の評価に利用した．ストリーム重みについては，今回用いたパラメータが口の動きの検出に有効と考えられるので，無発声をあらわす silence のモデルのみ $\lambda_A + \lambda_V = 1$ の条件で変化させ，その他の音素モデルは $\lambda_A = 1$，$\lambda_V = 0$ と固定し，音響特徴量のみを用いた．テストセットには，音声がクリーンのもののほか，5，10，15，20dB の白色雑音を加えたものを用意した．

5. 実験結果・考察

5.1. 実験結果

各種雑音およびクリーン音声に対して，音響特徴量のみでの認識率（数字正解精度），および画像特

表 2: 認識結果の比較

SNR	音響のみ	音響–画像 (A)	音響–画像 (B)
5dB	39.50%	44.95% (0.86)	43.22% (0.78)
10dB	58.55%	70.78% (0.90)	71.89% (0.82)
15dB	78.25%	87.65% (0.96)	88.21% (0.90)
20dB	94.66%	94.51% (0.94)	94.74% (0.86)
clean	97.59%	97.96% (0.96)	97.94% (0.88)

（括弧内は λ_a の値）

表 3: 認識誤りの分析結果（100 発話あたり）

	silence	脱落	置換	挿入
音響のみ	あり	14.36	25.01	4.12
	なし	9.64	18.90	12.91
音響–画像 (A)	あり	15.41	9.61	1.81
	なし	14.56	12.92	1.74
音響–画像 (B)	あり	14.47	9.52	2.35
	なし	14.01	11.83	2.27

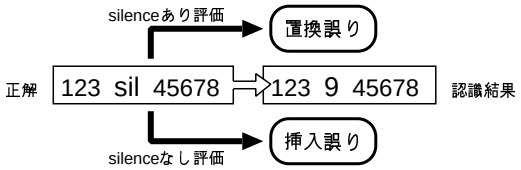


図 5: silence の評価の有無による誤り分類の違い

徴量 (A) または (B) を併せて用いたときの認識率を図 4 に示す．これらのグラフにおいて，縦軸は数字正解精度，横軸は silence モデルの音響ストリーム重み（λ_a）を示している．また，実線は音響特徴量のみでの認識率でベースラインに相当し，点線は音響–画像特徴量による認識率である．音響–画像特徴量を用いた場合には，各 SNR ごとに数字認識率が最大となるような最適な λ_a が存在する．このときの認識率と λ_a を表 2 にまとめる．

図 4 および表 2 から，ベースラインに比べ，SNR=5dB のとき約 9%，10dB のとき約 32%，15dB では約 46%，音響–画像特徴量を用いる提案手法によって誤り率が削減された．また SNR=20dB およびクリーンな条件での実験では，ベースラインとなる音声のみの手法での認識率が十分に高かったため，提案手法は音声のみの手法とほぼ同等の認識性能となった．

5.2. 考察

雑音環境下において認識性能が改善した要因としては，音響–画像特徴量により無音区間の推定精度が向上したことで，数字の挿入誤りや無音の脱落誤りなどが抑制され，境界のずれとそれによる誤りの波及が抑えられた，といったことが考えられる．このことを確かめるため，通常は認識結果にカウントされない silence を，他の数字と同様単語とみなしたときの認識結果についての解析を行った．いま例として，図 5 のように，無音区間を別の数字に誤って認識された場合を考える．silence を他の数字と同様に扱ったとき（以下，silence あり評価とする）には，このエラーは置換誤りとなるが，silence を単語とみなさない（以下，silence なし評価とする）ときは，誤認識された数字の挿入誤りとして分類される．このように，無音区間の誤認識は silence あり評価，なし評価の誤りの分布の違いとして現れるので，この性質を利用することで，認識結果への影響を調べることができる．表 3 に，SNR=10dB における，音響特徴量のみ，および音響–画像特徴量 (A)，(B) の 3 種類の認識結果について，最も認識率が高かったときの脱落誤り，置換誤り，挿入誤りの個数を調べ，これを 100 発話あたりの頻度に換算した値を示す．音響特徴量のみの場合は，silence あり評価となし評価とで，置換誤りと挿入誤りの値の変化が大きく，エラーの分布に変化が生じている．これより，silence が正しく認識されず，無音区間が他の音素・数字に誤って認識されていると考えられる．また脱落誤りの差も大きく，無音区間が正確に推定できず silence 自身が脱落していることも推測できる．これに対し音響–画像特徴量を用いた場合には，誤りの分布の変化はわずかであることから，マルチモーダル音声認識では，無音部分の誤認識が抑制され，その結果認識精度が向上したものと考えられる．

ストリーム重みに関しては，図 4 や表 2 から，雑音が大きくなるほどより画像情報をより重視し利用するようになり，その結果 λ_a が小さく（λ_v が大きく）なることが確認できた．

また，画像特徴量 (A) と (B) について比較してみると，条件によって優劣が異なるものの，最高となる数字正解精度で比べるとほぼ同じ程度の認識性能を示している．(B) の方が (A) よりも口の動き情報をより多く持っているのにもかかわらず，両者の結果に差が出なかった原因としては，(A) の特徴量では分散値を用いているために正規化は不要であるが，(B) では同じ発声でも話者により，口の開き具合や画像中の口唇の大きさによって値が異なってくるため，正規化が必要であったためと考えられる．今回は silence にのみ重みづけを行うという制限下での実験であったため，動き情報の利用が不十分であったことも一因と思われる．ただし，図 4 を全体的に見てみると，(B) の方が広い範囲で (A) よりも高い性能を示しており，実用上は (B) のように，λ_a にあま

り依存せず性能を発揮するようなパラメータの方が有利であるといえる．

6. まとめ

本研究では，オプティカルフローを用いたマルチモーダル音声認識の手法の提案を行った．このマルチモーダル音声認識システムを用いて評価実験を行ったところ，音響特徴量のみの場合に比べ，SNR=10dBで約 32%，15dB で約 46%誤り率が削減し，雑音環境下での本手法の有効性が確認された．この要因として，雑音下における無音区間の推定精度の改善による数字認識率の向上といったことが挙げられる．今後の課題として，口の動きの方向性や程度をより反映した画像特徴量の検討，各モデルに対し最適なストリーム重みを決定する手法の検討などが挙げられ，これによりさらなる認識性能の向上が期待できる．また現在，本手法の実環境における評価を行うため，車載カメラによる高速道路走行時のデータを収録し，これをテストデータとして実験を行っている．この結果については，また別の機会に報告する予定である．

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] 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).

[2] 宮島 千代美, 徳田 恵一, 北村 正, “最小誤り学習に基づくバイモーダル音声認識,” 2000 年春季音講論, 1-Q-14, pp.159-160 (2000-3).

[3] G. Potamianos, J. Luettin and C. Neti, “Hierarchical discriminant features for audio-visual LVCSR,” Proc. International conference on ICASSP 2001, pp.165-168 (2001-5).

[4] G. Potamianos, A. Verma, C. Neti, G. Iyengar and S. Basu, “A cascade image transform for speaker independent automatic speechreading,” Proc. International conference on Multimedia and Expo, pp.1097-1100 (2000-8).

[5] 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).

[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).

[8] K. Iwano, S. Tamura and S. Furui, “Bimodal speech recognition using lip movement measured by optical-flow analysis,” Proc. International workshop on HSC 2001, pp.187-190 (2001-4).

[9] 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の検討,” 2001 年秋季音講論, 1-1-14, pp.27-28 (2001-10).