論文 / 著書情報
Article / Book Information

論題(和文)	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
Title(English)	
著者(和文)	田村 哲嗣, 岩野 公司, 古井 貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会 2001年秋季講演論文集, Vol. , No. 1-1-14, pp. 27-28
Citation(English)	, Vol. , No. 1-1-14, pp. 27-28
発行日 / Pub. date	2001, 10

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 *

◎田村 哲嗣　岩野 公司　古井 貞熙 （東工大）

1. はじめに

雑音環境下で頑健に音声認識を行う手法の一つとして，唇動画像の情報を利用するマルチモーダル音声認識が注目され，近年研究が進められている [1, 2, 3]．我々はオプティカルフローによる読唇 [4] に注目し，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，マルチモーダル音声認識の手法の提案を行った [5]．本研究では，さらに高い認識性能が得られるよう，提案法の改善の検討を行った．

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]．すなわち，

$$I_x \cdot u + I_y \cdot v + I_t = 0 \tag{1}$$

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

ここで $I(x, y, t)$ は時刻 t における点 (x, y) の明度，$u(x, y)$，$v(x, y)$ はそれぞれ点 (x, y) のフローベクトルの水平成分，垂直成分である．式 (1)(2) から，各点でのフローベクトルを繰り返し演算により推定することで，オプティカルフローを計算することができる．

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．音声，画像はそれぞれ標本化周波数 16kHz，15Hz でサンプリングを行った．その後，表 1 の条件でそれぞれ 39 次元の音響特徴量と，2 種類の 2 次元画像特徴量 (a)，(b) に変換した．(a) は文献 [5] で用いたパラメータ，(b) は今回提案したパラメータである．(b) で利用しているオプティカルフローの積分結果の例を図 2 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．

$$u(x, y) = \frac{\partial}{\partial x} f(x, y)\ , \quad v(x, y) = \frac{\partial}{\partial y} f(x, y) \tag{3}$$

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 2 からわかるように，開閉に応じて最大値，最小値が変化する．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．

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)：フロー全体の水平成分分散値 　　：　〃　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　：2 次元

得られた音響特徴量と画像特徴量はパラメータレベルで融合し，41 次元の音声–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．HMM は学習時には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 において音声–画像特徴量 O_{AV} を観測する確率 $b_j(O_{AV})$ は式 (4) で表される．

$$b_j(O_{AV}) = {b_A}_j(O_A)^{\lambda_A} \cdot {b_V}_j(O_V)^{\lambda_V} \tag{4}$$

ここで ${b_A}_j(O_A)$，${b_V}_j(O_V)$ 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 は各々のストリーム重みである．

4. 実験

4.1. データベース

データベースは，筆者らが収録した 11 名の男声話者による連続数字読み上げ音声–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

実験では，10 名分のデータを用い連結学習によって HMM を学習した．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，それらの認識率の平均値をもっ

---

* Multimodal speech recognition using optical-flow analysis,
by Satoshi Tamura, Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology).

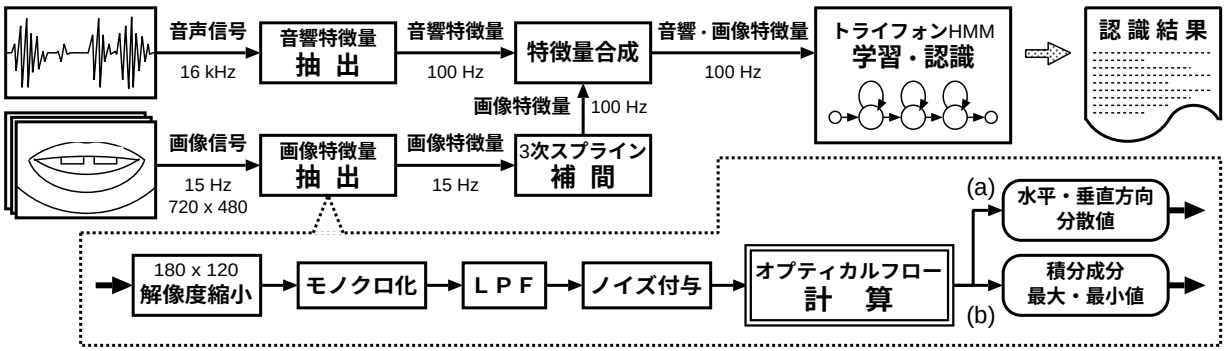

図 1: マルチモーダル音声認識システム

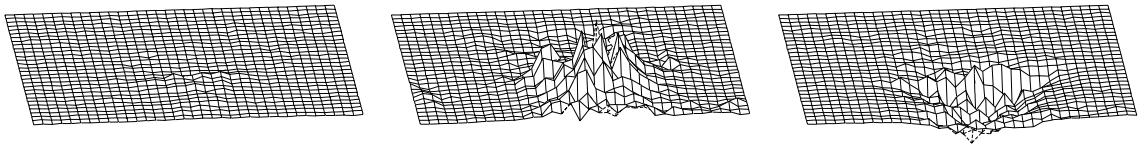

図 2: オプティカルフローの積分結果 （左: 無発声時，中央: 開口時，右: 閉口時）

てモデルの性能を評価した．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，silence のモデルのみ $\lambda_A + \lambda_V = 1$ の条件で変化させ，その他のモデルは $\lambda_A = 1$，$\lambda_V = 0$ とした．ま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．

4.3. 実験結果・考察

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，および画像特徴量 (a)，(b) を併せて用いたときの最適なストリーム重みでの認識率を図 3 に示す．これから，画像特徴量 (a) を加えた場合，SNR=10dB で最大約 12%，(b) では同じく最大約 15%の改善が見られ，オプティカルフローから得られた画像特徴量が有効に機能していることが確かめられた．また全ての環境下において，画像特徴量 (b) を加えたときの認識性能は，(a) を加えたものと比べて同等かそれ以上のものを示しており，今回提案したパラメータは口の動きの情報をより反映していることが判明した．

5. まとめ

本研究では，我々が提案した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について，抽出する画像特徴量を改善することで，さらに認識性能が向上できることを示した．また実験結果より，雑音環境下においては，発声の有無の検出が認識率を向上させる上で重要な役割を果たしていると推測される．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．

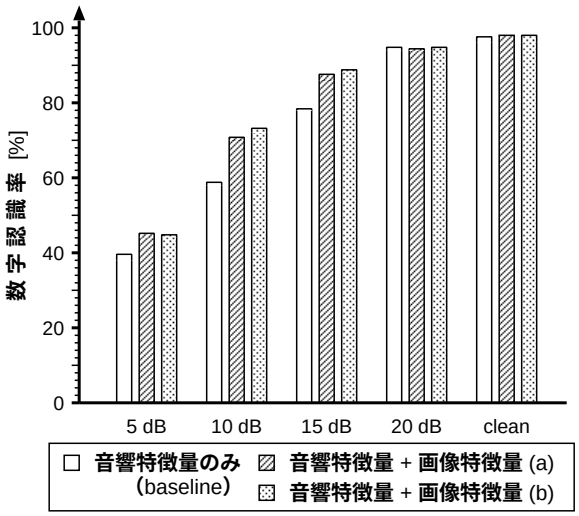

図 3: 認識結果

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

[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] 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] 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).

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