論文 / 著書情報
Article / Book Information

論題(和文)	マルチモーダル音声認識のための画像特徴量の改善
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 2003年春季講演論文集, Vol. , No. 3-Q-22, pp. 195-196
Citation(English)	, Vol. , No. 3-Q-22, pp. 195-196
発行日 / Pub. date	2003, 3

マルチモーダル音声認識のための画像特徴量の改善 *

◎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]．我々はこれまでに，オプティカルフローを用いた手法を提案し，実環境データの認識実験により，その頑健性と有効性を示した [3, 4, 5]．しかし，現在用いている情報は口の開閉と動きの有無のみであり，認識性能の改善には限界がある．認識性能のさらなる向上のためには，口の縦横の長さや口腔・歯などの情報が必要と考えられる．本研究では，これらの情報を反映した画像特徴量に改良することによる，提案手法の認識性能の改善について検討を行った．

2. 画像特徴量

本研究では，画像中の横方向の口唇中心位置を推定し HMM により縦方向の座標を求めるアルゴリズムを検討する．さらにこれを用いて，口唇と口腔の情報を求め画像特徴量とする新たな手法を提案する．

2.1. 横方向の口唇中心位置の推定

幅 W × 高さ H（単位 pixel）の口唇付近を撮影した入力画像に対し，輪郭情報の抽出を行う．ここではオプティカルフローを利用して輪郭を抽出する．フローベクトルは明度勾配のあるところのみ観測されるので，一種の輪郭抽出フィルタとして利用できる．入力された画像から作為的に数 pixel ずらした画像を生成し，これらからフローベクトルを計算して，その垂直成分の絶対値 $v(x, y)$ を輪郭情報として用いる．次に入力画像中の各列 i $(0 \leq i \leq W-1)$ に対して，$v(i, y) = v_i(y)$ を最小自乗法により式 (1) で近似し，A_i，B_i を求める．

$$v_i(y) \simeq \left| A_i (y - y_0) e^{-B_i (y-y_0)^2} \right| \tag{1}$$

ただし $A_i > 0$，$B_i > 0$，y_0 は $v_i(y)$ の重心である．列中に口唇が含まれる場合は上唇と下唇の 2 箇所で輪郭が抽出されるので，式 (1) を用いることで効率的に輪郭情報を近似することができる．また口唇を含むときは，式 (1) の積分値は大きくなるという特性がある．そこでこれを口唇を含んでいるもっともらしさの指標 $l(i)$ として，式 (2) のように定義する．

$$l(i) = \int_{-\infty}^{\infty} v_i(y) dy = \frac{A_i}{B_i} \tag{2}$$

以上から式 (3) により，$l(i)$ の重心を計算することで，横方向の口唇中心位置 C を推定することができる．

$$C = \sum_{i=0}^{W-1} i \times l(i) \tag{3}$$

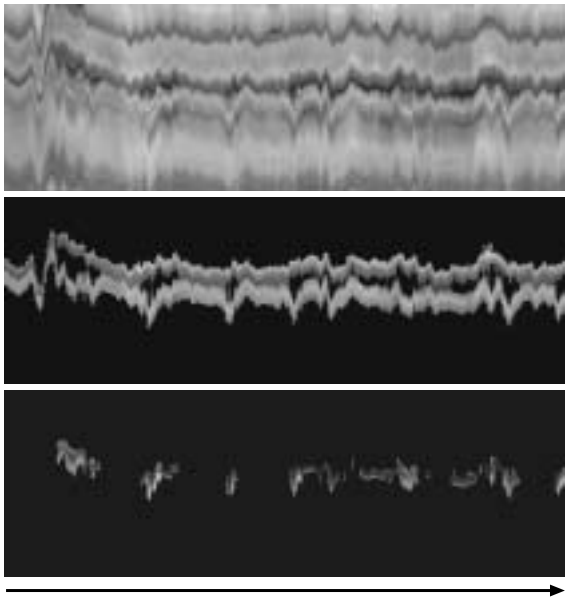

図 1: 口唇抽出結果（ 上：原画像，中：口唇抽出結果，下：口腔抽出結果 ）

2.2. 口唇位置推定

原画像より列 C を抽出し，各画素の RGB 値を人間の知覚に近い HSI（色相，彩度，輝度）値に変換する．照明条件に影響する輝度の値を除去し，一方で輪郭情報 $v_C(y)$ を加え，それぞれの画素に対して 3 次元のベクトルを計算する．これを列 C の上から下にスキャンし，H 点の 3 次元パラメータ系列を生成する．

このパラメータ系列を HMM を用いて強制切り出しし，縦方向の口唇や歯などの位置情報を推定する．使用する HMM は状態数 3，混合数 4 の left-to-right 型で，上側肌，上唇，口腔（歯），下唇および下側肌の 5 種類である．ただし口が閉じている場合を考え，口腔 HMM を飛ばして上唇 HMM から下唇 HMM への直接遷移を可能とする．手作業で付与した座標ラベルありデータによって初期モデルを生成し，続いて座標情報なしデータを用いて連結学習を行い，不特定話者モデルを構築する．認識の際は精度を高めるために，予めテストデータで MLLR によりモデルの適応を行い，Viterbi アルゴリズムによって各 HMM の座標情報（アライメント）を求める．

図 1 に，後述する学習データで HMM を構築し，実環境テストデータの画像系列の座標情報を求め，これを図示したものを示す．横軸は時刻，縦軸は各時刻の C における画素列である．

2.3. 画像特徴量抽出

前節で得られた座標情報から，上唇 HMM 終了座標と下唇 HMM 開始座標を求め，口唇の縦方向の長さ h

---

* Improvement of visual features for multi-modal speech recognition,
by Satoshi Tamura, Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology).

を算出する．認識結果に口腔の情報がある場合は，口腔中の画素の輝度の二値化を行って歯を検出し，その pixel 数 t を計算する．口腔情報がないときは，本研究では $t = 0$ とした．以上の 2 次元のパラメータ (h, t) を画像特徴量として抽出する．

3. 実験条件

3.1. 学習データ

クリーン環境で収録した男性話者 11 名による連続数字読み上げデータを使用した．各話者は，2〜6 桁の数字を 250 個発声している．

3.2. テストデータ

高速道路走行中の乗用車内で収録した数字連続読み上げデータを用いた [5]．学習データに含まれない 6 名の男性話者が 2〜6 桁の数字を 115 個発声している．なおこのうち 3 名分のデータについて，前章で述べた口唇の位置推定が正しく行われていなかったため，本研究では使用していない．

3.3. 音響・画像特徴量

音響特徴量は CMN–MFCC 12 次元とその Δ，ΔΔ 成分，および対数パワーの Δ，ΔΔ 成分の計 38 次元を使用した．画像特徴量は前章で説明したパラメータ 2 次元を画像系列ごとに正規化し，さらにその Δ，ΔΔ 成分を求めて計 6 次元とした．

3.4. 学習・認識

音声認識のモデルには，状態数 3，混合数 2 の left-to-right 型 triphone HMM を用いた．まず学習用音響データを用い，連結学習によって音響 HMM の学習を行った．その後 Viterbi アルゴリズムで，学習データに時間情報のラベルを付与し，このラベルを用いて画像 HMM の学習を行った．そして得られた画像 HMM と音響 HMM を融合して，音響–画像マルチストリーム HMM を生成した．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 (4) で表される．

$$b_j(O_{AV}) = b_{Aj}(O_A)^{\lambda_A} \cdot b_{Vj}(O_V)^{\lambda_V} \qquad (4)$$

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 は各々のストリーム重みで，$\lambda_A + \lambda_V = 1$ である．画像 HMM ではパラメータ間の相関を考慮して，確率を表すガウス分布において全共分散行列を用いている．

4. 実験結果・考察

今回提案したパラメータによる認識実験の数字正解精度（音響 · 画像 (a)），および比較のため，ベースラインとして音響特徴量のみの結果（音響のみ）と，従来使用していたオプティカルフローの積分成分の最大・最小値 2 次元のパラメータ [4] を用いたときの結果（音響 · 画像 (b)）を表 1 に示す．ストリーム重みについては，silence（無音）モデルのみ変化させ他のモデルは $\lambda_A = 1$，$\lambda_V = 0$ で固定とした場合と，全ての HMM で同じ重みを用いた場合の 2 通りの条件で実験を行った．silence モデルのみ重み付けを行った場合，提案した特徴量は従来のものと同等の性能を示し，ベースラインと比べ相対的に約 18%誤り率が削減された．このように認識性能が改善した要因として，無音区間の推定精度の向上により，silence の部分の数字の挿入誤りなどが抑制されたことが考えられる．また，全モデルに同じ重み付けを行ったときは，従来法では認識率の改善がみられないのに対し，提案法はベースラインからみて約 10%誤り率が削減された．silence のみ重み付けを行ったときの結果にはまだ及ばないものの，従来のパラメータに比べると数字の同定精度が改善したことが確認された．

表 1: 実験結果（数字正解精度）

	sil モデル 重み付け	全モデル 重み付け
音響のみ	61.45%	
音響 · 画像 (a)	68.55% (λ_A=0.82)	65.29% (λ_A=0.58)
音響 · 画像 (b)	68.55% (λ_A=0.78)	61.74% (λ_A=0.96,0.98)

音響 · 画像 (a)： 口唇の縦方向長さ h，歯の pixel 数 t とこれらの Δ，ΔΔ 成分
音響 · 画像 (b)： オプティカルフローの積分成分の最大・最小値

5. まとめ

本研究では，HMM を用いた口唇抽出アルゴリズムを用いて，新たに口唇や口腔の情報を反映した画像特徴量を提案した．このパラメータにより認識実験を行った結果，従来のものと比べて数字の同定精度が改善し，認識性能を向上させることに成功した．今後の課題としては，口唇抽出アルゴリズムの性能向上，口唇の水平方向情報の利用方法の検討，および音素ごとのストリーム重みの最適化が挙げられる．

謝辞

本研究は NTT ドコモ株式会社の援助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] S. Nakamura, K. Kumatani and S. Tamura, “Robust bi-modal speech recognition based on state synchronous modeling and stream weight optimization,” Proc. ICASSP2002, Orlando, U.S.A., pp.309-312 (2002-5).

[2] M. T. Chen, “Stream-weighted HMM for audio-visual ASR,” Proc. MMSP2002, St. Thomas, U.S., (2002-12).

[3] K. Iwano, S. Tamura and S. Furui, “Bimodal speech recognition using lip movement measured by optical-flow analysis,” Proc. HSC2001, Kyoto, Japan, pp.187-190 (2001-4)

[4] 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).

[5] 田村 哲嗣, 岩野 公司, 古井 貞熙, “実環境におけるマルチモーダル音声認識の評価,” 2002 年春季音講論, 3-5-5, pp.151-152 (2002-3).