論文 / 著書情報
Article / Book Information

論題(和文)	マルチモーダル音声認識における音響・画像特徴の融合法に関する検討
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 2003年秋季講演論文集, Vol. , No. 3-6-11, pp. 123-124
Citation(English)	, Vol. , No. 3-6-11, pp. 123-124
発行日 / Pub. date	2003, 9

マルチモーダル音声認識における音響・画像特徴の融合法に関する検討 *

◎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ている [1, 2]．我々はこれまでに，画像についても頑健性を考慮したマルチモーダル音声認識手法を開発しており，実環境データの認識実験で，その有効性を示している [3]．本論文では，画像特徴量やその抽出法の改良，マルチストリーム HMM における音響，画像ストリームの重み係数最適化など，融合に関する種々の検討を行い，認識性能の改善を試みた．

2. 画像特徴量

2.1. 口唇の縦方向座標の推定

幅 W × 高さ H（単位 pixel）の口唇を含む入力画像に対し，各画素の RGB 値を HSI（色相，彩度，明度）値に変換して，これより各列毎に口唇座標推定のための H 点のパラメータを W 系列生成する．また，輪郭情報をもとに横方向（x 軸方向）の口唇中心位置 C を推定する．列 C のパラメータ系列に対し，強制切り出しにより縦方向（y 軸方向）の口唇座標を推定する．使用する HMM は上肌，上唇，口腔，下唇および下肌の 5 種類である．求められた座標情報から，口の高さ h を算出する [3]．

2.2. 口唇の横方向座標の推定

本論文では，新たに口の横方向に関する情報を抽出する（図 1）．前節で用いた HMM により，各列毎に口唇なし（上肌→下肌），口唇あり（上肌→上唇→下唇→下肌），口腔あり（上肌→上唇→口腔→下唇→下肌）の 3 種類の尤度を求め，これをスコアとして使用する．ワンパス DP マッチングにより左列から右列へスコアが最大となるパスを計算し，バックトラックにより境界情報を求め，口唇座標を推定する．そして口腔ありの開始および終了座標から，口の幅 w を算出する．

2.3. 画像特徴量抽出

以上で得られた h，w に加え，列 C の口腔内で輝度二値化により検出された歯の pixel 数 t（$h = 0$ のときは $t = 0$）の 3 次元パラメータ (h, w, t) を抽出する．これを画像系列ごとに正規化し，その Δ，ΔΔ 成分を求めて計 9 次元の画像特徴量とする．

3. ストリーム重み最適化

本研究では，音声認識時においてマルチストリーム HMM を使用している．マルチストリーム HMM では，単語 w に対する音響・画像特徴量 $\mathbf{O}_t$ の観測確率は，対数尤度 $b_w(\mathbf{O}_t)$ を用いて以下のように表される．

$$b_w(\mathbf{O}_t) = \lambda_{Aw} b_{Aw}(\mathbf{O}_{At}) + \lambda_{Vw} b_{Vw}(\mathbf{O}_{Vt}) \qquad (1)$$

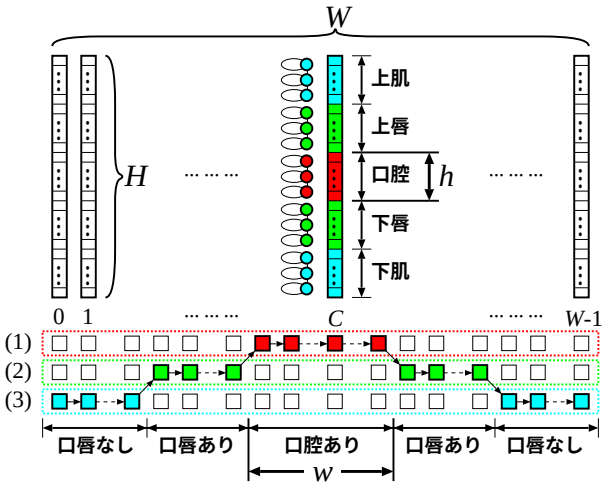

図 1: 口唇座標推定（ (1)：口腔あり尤度，(2)：口唇あり尤度，(3)：口唇なし尤度 ）

ただし t は時刻，$b_{Aw}(\mathbf{O}_{At})$，$b_{Vw}(\mathbf{O}_{Vt})$ はそれぞれ音響特徴量 $\mathbf{O}_{At}$，画像特徴量 $\mathbf{O}_{Vt}$ に対する単語 w の対数尤度，λ_{Aw}，λ_{Vw} は単語 w を構成する HMM における音響，画像ストリーム重みで，以下の制約がある．

$$\lambda_{Aw} + \lambda_{Vw} = 1 \quad , \quad 0 \le \lambda_{Aw}, \lambda_{Vw} \le 1 \qquad (2)$$

このストリーム重みは，HMM 学習時には他のモデルパラメータと異なり最尤推定によって決定することができない．一方，認識時には，音響・画像の各々の雑音状況に応じて重みを変化させることが性能向上に有効であると考えられる．このようなことから，ストリーム重み最適化については種々の研究がなされている [4, 5]．本論文では，認識時にストリーム重みを適応的に決定する手法について検討を行った．

時刻 t におけるデコーダ出力単語 w_t に対し，任意の $w \in W$（辞書中の単語の集合，$|W| = N$）について，

$$b_{w_t}(\mathbf{O}_t) \ge b_w(\mathbf{O}_t) \qquad (3)$$

が成り立つ．w_t が本来の正解単語でないという認識誤りは，ミスマッチにより正解でない単語の尤度が大きくなってしまうことによる．そこで適応データが与えられたとき，正解単語の尤度とその他の単語の尤度の差が最大となるようにモデルパラメータを調整すれば，同じ環境の認識データに対し，認識誤りを抑制できると考えられる．本論文では，以上の考えに基づき，ストリーム重み $\Lambda = \{\lambda_{Aw}\}$ を適応的に求める手法を提案する．すなわち，

$$L(\Lambda) = \sum_{t=1}^{T} \sum_{w \in W} \left\{ b_{w_t}(\mathbf{O}_t) - b_w(\mathbf{O}_t) \right\}^2 \qquad (4)$$

---

* Investigation of a fusion method for multi-modal speech recognition,
by Satoshi Tamura, Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology).

として，式 (4) が最大となる $\Lambda = \hat{\Lambda}$ を求めてゆく．このとき，任意の $\lambda_{Aw_r} \in \Lambda$ に対し，

$$\frac{\partial L(\Lambda)}{\partial \lambda_{Aw_r}} = 0 \tag{5}$$

が成り立つので，式 (5) を解くことにより λ_{Aw_r} の変化分 $\Delta\lambda_{Aw_r}$ は次のように求めることができる．

$$\begin{aligned} d_w(\mathbf{O}_t) &= b_{Aw}(\mathbf{O}_{At}) - b_{Vw}(\mathbf{O}_{Vt}) \\ A &= \sum_{\substack{t=1 \\ w_t=w_r}}^{T} \Big\{ N b_{w_r}(\mathbf{O}_t) - \sum_{w \in W} b_w(\mathbf{O}_t) \Big\} \\ &\quad + \sum_{\substack{t=1 \\ w_t \neq w_r}}^{T} \Big\{ b_{w_r}(\mathbf{O}_t) - b_{w_t}(\mathbf{O}_t) \Big\} \\ B &= \sum_{\substack{t=1 \\ w_t=w_r}}^{T} N d_{w_r}(\mathbf{O}_t) + \sum_{\substack{t=1 \\ w_t \neq w_r}}^{T} d_{w_r}(\mathbf{O}_t) \\ \Delta\lambda_{Aw_r} &= \frac{A}{B} \end{aligned} \tag{6}$$

同様にして全ての $\lambda_{Aw} \in \Lambda$ について $\Delta\lambda_{Aw}$ を求め，その後 λ_{Aw} の値を更新する．この更新サイクルを繰り返すことにより，$\hat{\Lambda}$ を推定することができる．

4. 実験条件

4.1. データベース

学習データにはクリーン環境で収録した男性話者 11 名の，テストデータには高速道路走行中の車内で収録した，6 名の数字連続読み上げデータを用いた [6]．各話者は 2〜6 桁の数字を，学習データでは 250 個，テストデータでは 115 個発声している．

4.2. 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 次元である．認識時には音響特徴量と画像特徴量を融合し，47 次元の音響・画像特徴量として使用した．

4.3. 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．連結学習によって音響 HMM を作成し，Viterbi アルゴリズムにより時間情報つきラベルを生成し，これにより画像 HMM の学習を行った．得られた音響 HMM と画像 HMM を融合し，音響–画像マルチストリーム HMM を生成した．ストリーム重みについては，(1) 全てのモデルに対し同じ重みを使用，(2) その認識結果を用い，前章で説明した手法により最適化（教師なし適応）したものを使用，の 2 種類の方法で推定を行った．

5. 実験結果・考察

表 1 に，(a) 今回提案する特徴量，および (b) 以前用いていた口唇の縦方向長さと歯の pixel 数 [3] の 2 種類のパラメータに対する，(1) および (2) の重み決定法を適用したときの数字正解精度を示す．これより音響のみの結果（ベースライン）と比較して，提案したパラメータにおいて (2) のとき最大で 11%の正解精度向上，29%の誤り率削減に成功した．(a) と (b) の結果を比較すると，いずれにおいても改良した特徴量の方が高い性能を示している．これは口の横方法の情報を新たに加えたことで識別性能が向上したためと考えられる．一方 (1) と (2) については，画像特徴量を加え，全モデルに同じ重みを設定する (1) の手法によりある程度認識率が改善し，さらに (2) のストリーム重み最適化手法によって認識性能が大幅に向上していることがわかる．また (1) で得られる認識結果には認識誤りが含まれているが，その影響を受けても (2) の再推定によって認識率が改善していることから，提案したストリーム重み最適化手法は，雑音などさまざまな環境の変化に対して高い頑健性を有すると考えられる．

表 1: ストリーム重み最適化による認識結果（数字正解精度）

	(1) 全モデルに同じ重みを使用	(2) その結果を用いて最適化
音響のみ	61.96%	
音響 · 画像 (a)	64.35%	73.04%
音響 · 画像 (b)	63.19%	70.94%

音響 · 画像 (a)· · · 口唇の縦，横方向長さ h，w，歯の pixel 数 t とこれらの Δ，ΔΔ 成分
音響 · 画像 (b)· · · 口唇の縦方向長さ h，歯の pixel 数 t とこれらの Δ，ΔΔ 成分

6. まとめ

本論文では，従来の口唇抽出アルゴリズムを改良し新しい画像特徴量を提案するとともに，尤度差最大基準によるストリーム重み係数の自動適応化手法を提案した．実環境データにより認識実験を行った結果，最大で約 11%正解精度が向上し，提案手法の有効性と頑健性が確かめられた．今後の課題としては，decision fusion 法による認識手法の検討，別のタスクにおける本手法の評価などが挙げられる．

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, pp.309-312 (2002-5).

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

[3] 田村 哲嗣, 岩野 公司, 古井 貞熙, “マルチモーダル音声認識のための画像特徴量の改善,” 2003 年春季音講論, 3-Q-22, pp.195-196 (2003-3).

[4] Javier Hernando, “Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition,” Proc. ICASSP’97, pp.1267-1270 (1997-4).

[5] 宮島 千代美, 徳田 恵一, 北村 正, “多次元確率分布 GMM に基づく話者識別モデルにおけるストリーム重みの推定,” 2001 年春季音講論, 1-3-3, pp.5-6 (2001-3).

[6] 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).