論文 / 著書情報
Article / Book Information

論題(和文)	横顔画像から抽出した口唇角度情報を用いたマルチモーダル音声認識
Title(English)	
著者(和文)	吉永 智明, 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 2004年秋季講演論文集, Vol. , No. 3-1-19, pp. 147-148
Citation(English)	, Vol. , No. 3-1-19, pp. 147-148
発行日 / Pub. date	2004, 9

横顔画像から抽出した口唇角度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，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が注目され，近年研究が進められている ([1,2] 他)．これらの研究は主に顔の正面から撮影された画像を用いているが，モバイル環境下での適用を考えた場合には，発話を行いながら正面から顔画像を撮影することは困難，かつ，不自然であるといった問題が生じる．そこで，我々は横向きの顔画像から得られる口唇情報を用いた音声認識手法を提案している [3]．この手法は，マイクロフォン部分に微小カメラを搭載した携帯電話によって，通常の使用姿勢で口唇動画像情報を取得することを想定している．このような横方向からの口唇の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．

本論文では，この音声認識手法の更なる頑健性の向上を目指し，新たな画像特徴量を提案する．

2　特徴量抽出手法

従来研究 [3] では画像特徴量としてオプティカルフローの水平・垂直方向の分散値を用いていた．この特徴量は口唇の動き情報を反映していることから，特に発声区間の検出に有効であり，雑音環境下における認識性能の向上に寄与することが確認されている [3]．更なる認識精度向上には発話内容の推定精度（音素の識別精度）を高める必要があり，そのためには口唇の動き情報だけでなく，形状の情報も不可欠である．そこで，本研究では上下唇の成す角度を口唇形状を表す特徴量として抽出し，利用する．具体的には，上唇と下唇を模する 2 つの口唇ラインを抽出し，その成す角度を求める．本研究では，右方向から撮影された横顔を利用することを想定しており，口唇ラインとは「横顔画像中における口唇の最左点を基準点として，この点から上唇，下唇それぞれの口唇領域を最も含むように引かれた 2 つの直線」のことである．

2.1　口唇領域抽出

口唇ライン抽出のための最初の処理として，画像中から口唇の存在する矩形領域を抽出する (図 1(a))．

そこで，まず原画像に対し Sobel フィルタをかけ，エッジ画像を抽出する．エッジは主に，横顔と背景の境界部分や口唇の輪郭部分に検出される．このエッジ出現位置をもとに，水平方向に口唇領域の推定を行う．具体的には，1) 各列についてエッジ部に相当する画素の数を求めて，その数が最大となる列を見つけ，2) その列を開始列として左右に閾値以下となる列を探索し，3) 最初に閾値以下となった列を，口唇領域の左右端とする．なお，本実験では，最大値に 0.3 を掛けた値を閾値に用いる．

次に，垂直方向の口唇領域を推定する．そこで，得られた左右端で挟まれる領域内の原画像を色相値によって二値化する (以降，この画像を色相画像と呼ぶ)．ここでは，口唇の色に近い 1.5〜2π の色相値を有する画素を 1 として二値化を行う．この色相画像

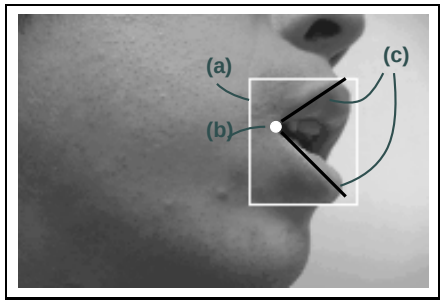

図 1. 口唇ライン抽出例

にラベリング処理を施して最大の面積を持つ領域を抽出し，その領域について，左右端決定時と同様の処理を行ごとに行うことで口唇の上下端を決定する．

2.2　基準点抽出

2.1 節で得られた口唇領域の画像から口唇ラインの基準点を決定する (図 1(b))．これにはまず，原画像から口唇内部の領域を抽出する．口唇内部は画像中において最も暗くなることから，輝度値による二値化を行うことで抽出する．この際の輝度の閾値は経験的に 15 とし，閾値以下となる画素を口唇内部領域とする．この口唇内部領域における最左点を基準点とする．

2.3　上下唇ラインの決定

以上によって得られた口唇領域と基準点から上下唇ラインを抽出する (図 1 (c))．

処理の流れを説明する．1) エッジ画像と色相画像の AND を取った二値画像を生成する．2) この画像において，基準点を始点として画像の右半分に放射状に半直線を引き，各半直線上の 1 となる画素の総数を求める．3) 最大値を持った線を探索開始線とする．4) 色相画像上で 2) と同様に放射状に直線を引き，探索開始線より下側にある直線の中で画素数が最大である直線を下唇ライン，上側で最大となる直線を上唇ラインとして抽出する．

3　実験

3.1　データベース

クリーン環境下で収録した男性話者 38 名による連続数字読み上げデータを使用した [3]．各話者は 4 桁数字 10 回を 1 セットとして 5 セット発声している．

3.2　音響・画像特徴量

音響特徴量には 12 次元の MFCC と，その 1 次，2 次微分，および対数パワーの 1 次，2 次微分の計 38 次元のパラメータを用いる．なお，フレーム長は 25ms，フレーム周期は 10ms であり，入力音声ごとに CMS を行っている．

横顔は毎秒 30 フレーム，解像度 720 × 480 の 24bit カラー画像としてキャプチャし，計算量削減のために 180 × 120 に変換した後，口唇ライン抽出を行う．得られた上下の口唇ラインの成す角度を「口唇角度」とし，その値と微分値を入力発話毎に最大値を用いて正規化することで 2 次元の画像特徴量を得る．最

∗ Multi-modal speech recognition using lip-angle information extracted from side-face images

By Tomoaki Yoshinaga†, Satoshi Tamura, Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology)
† 現在は日立製作所中央研究所所属

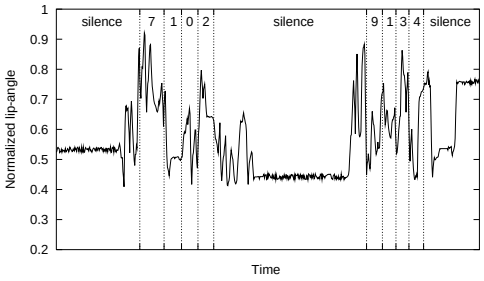

図 2. 正規化口唇角度値の例

後にフレーム周期を音響特徴量と合わせるため，3 次スプライン関数によって補間を行う．図 2 に「7102, 9134」と発声したときの口唇角度の値のグラフを示す．無発声時には口が動かないため固定値となり，発声時には発声内容によって異なる値をとっている．

実験には，比較のため，オプティカルフローの垂直・水平方向の分散値で構成された 2 次元の画像特徴量 [3] も用意する．さらに，提案した口唇角度情報による画像特徴量とオプティカルフローを用いた画像特徴量を結合して得られる 4 次元の融合画像特徴量も用意する．

最終的には，これら 3 種類の画像特徴量をそれぞれ音響特徴量とフレーム単位で結合することで，40 (42) 次元の音響–画像特徴量を作成し，認識に利用する．

3.3 音響・画像モデル

モデルとしては，状態数 3，混合数 2 の left-to-right 型 triphone HMM を用いる．まず音響特徴量のみを用いて音響 HMM を学習し，得られた HMM を用いて強制切り出しを行い各音素の時間ラベルを作成する．このラベルと画像特徴量のデータを用いて画像 HMM の学習を行い，得られた 2 つの HMM を融合し，音響–画像マルチストリーム HMM を構築する．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 (1) で表される．

$$b_j(O_{AV}) = b_{Aj}(O_A)^{\lambda_a} \cdot b_{Vj}(O_V)^{\lambda_v} \qquad (1)$$

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，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おり，$\lambda_a+\lambda_v=1$ という制約を設けている．

3.4 実験手法

leave-one-out 法を用いて実験を行い，その数字正解精度の平均を評価に用いた．テストセットには，SN 比 5, 10, 15, 20dB の白色雑音を付加したものを用い，音響特徴量のみを利用した場合と，音響–画像特徴量を利用した場合の認識結果を比較した．

また，MLLR を用いてマルチストリーム HMM の適応化を行った上で，同様の評価を行った．その際，適応化の対象は，音響ストリームの平均と分散のみとした．

3.5 実験結果

表 1 に SN 比条件ごとの，各画像特徴量を用いた場合の数字正解精度を示す．音響と画像のストリーム重みは各実験ごとに事後的に最適化を行っている．MLLR あり，なしに関わらず全ての雑音環境下において，口唇角度情報を用いたことで正解精度が向上し，本特徴量の有効性を確認することができた．また，融合画像特徴量を用いることで，MLLR あり SN 比 20dB 以外の全ての条件下で更なる正解精度の向上が達成された．SN 比 5 dB，MLLR ありの場合，融合画像特徴量を用いることで，音響特徴量のみでの結果に比べて，絶対値で 18.9%の正解精度の向上が確認された．

表 1. 各特徴量を用いた場合の数字正解精度の比較

使用特徴量	MLLR なし	MLLR あり
(SN 比 5 dB)		
Audio-only	28.4%	39.5%
Optical-flow	34.7%	52.6%
Lip-angle	36.4%	53.1%
Combined	39.3%	58.4%
(SN 比 10 dB)		
Audio-only	51.9%	69.4%
Optical-flow	56.7%	76.9%
Lip-angle	57.5%	77.2%
Combined	59.4%	79.5%
(SN 比 15 dB)		
Audio-only	75.6%	91.5%
Optical-flow	78.7%	93.3%
Lip-angle	79.1%	93.3%
Combined	79.9%	93.4%
(SN 比 20 dB)		
Audio-only	91.5%	97.0%
Optical-flow	92.2%	97.2%
Lip-angle	92.3%	97.4%
Combined	92.6%	97.2%

また，どの画像特徴量でも全ての条件下において，λ_a の変化における数字正解精度の変化は小さくゆるやかで，広い範囲にわたる本画像特徴量の効果が確認された．中でも融合画像特徴量での正解精度の変化は小さく，その推移の様子は他の 2 種類の画像特徴量の変化の様子を足し合わせた形状となった．

4 まとめ

本研究では，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における新たな特徴量を提案し，その有効性を示した．

今後の課題としては，1) 実環境での利用を考慮し，画像の外乱に対する頑健性向上のための手法の提案と，その効果の検証，2) 音響・画像ストリーム重みの自動的な最適化，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] 吉永 智明, 田村 哲嗣, 岩野 公司, 古井 貞熙, “横顔の口唇情報を用いたマルチモーダル音声認識,” 2003 年秋季音講論, 3-6-12, pp.125-126 (2003-9).