論文 / 著書情報
Article / Book Information

論題(和文)	横顔画像の口唇情報を利用し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 2003年秋季講演論文集, Vol. , No. 3-6-12, pp. 125-126
Citation(English)	, Vol. , No. 3-6-12, pp. 125-126
発行日 / Pub. date	2003, 9

横顔画像の口唇情報を利用し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

モバイル環境における音声入力は非常に手軽かつ便利に行うことが出来るため，近年このようなデバイスを用いた音声認識技術の必要性が高まって来ている．しかし，こうした環境下では周辺雑音の影響が大きく，音声認識を行う際に問題となる．

そこで，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，近年その研究が進められている [1-4]．これらの研究では，正面から撮影された口唇画像が用いられている．しかし，モバイル環境下でこのような方式を利用しようとすると様々な問題が生じる．カメラ付き携帯電話で音声と画像を入力することを考えると，ユーザは発話しながら携帯電話を顔の正面に持ってカメラ撮影を行うことになり，自然な音声入力を行うことができず，とても負担のかかるものとなってしまう．また，撮影のために携帯電話と口との距離を離さなければならず，音声の SN 比が劣化するといった問題も生じる．

そこで本研究では，正面の顔画像ではなく，横顔の動画像情報を用いたマルチモーダル音声認識手法を提案する．この方法では，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動き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．

2　特徴量抽出手法

口唇画像からの特徴量抽出には，オプティカルフローを用いる．オプティカルフローとは物体の動き情報を反映する特徴量であり，1) 時間的に連続した 2 枚の画像のみから計算が可能，2) 物体の形状といった事前知識が不要，3) パターンマッチングを用いないため特徴点抽出が不要，といった利点がある．我々はこれまでに正面から撮影された画像データから，オプティカルフローを特徴量として抽出して音声認識を行う手法を提案し，その有効性を確認している [4]．そこで，この手法に基づいて横顔の動画像情報を用いた音声認識システムを構築した．

2.1　音響特徴量

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．

2.2　画像特徴量

図 1 に使用した横顔画像の例を示す．横顔は毎秒 15 フレーム，解像度 720 × 480 の 24bit カラー画像としてキャプチャする．これを計算量削減のために 180 × 120bit に変換し，さらに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，ローパスフィルタリングと低レベルのランダム雑音付加を行う．こうして得られた画像に対して，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の計算を行う．

図 1. 横顔画像の例

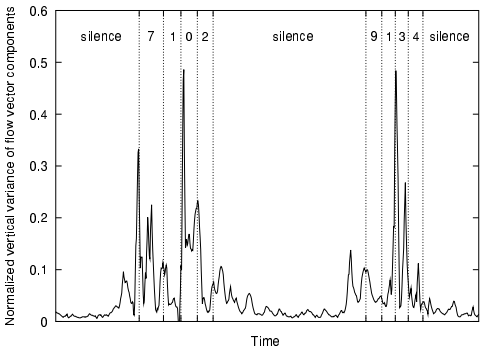

図 2. オプティカルフローの垂直成分の分散値

得られたフローベクトルの水平・垂直成分の分散値をフレーム画像ごとに計算し，入力発話毎に最大値によって正規化して 2 次元の画像特徴量を得る．この特徴量は，無発声時にはフローベクトルが観測されないので 0 に近い値となり，発声時には口唇の開閉方向にフローベクトルが表れるので値が大きくなる．したがって，この特徴量は発声時と無発声時の判別に有効である．図 2 に「7102，9134」と発声したときのオプティカルフローの垂直成分の分散値のグラフを示す．なお，水平成分の分散値もほぼ同じ傾向であった．

最後にフレーム周期を音響特徴量と同じ 10ms に合わせるため，3 次元スプライン関数によって補間を行う．

3　実験

3.1　データベース

クリーン環境下で 38 名の日本人男性話者に対し収録を行い，音響–画像データベースを構築した．各話者には日本語で 4 連続数字の読み上げ 10 回を 1 セットとし，これを 5 セットずつ発声してもらった．なお，連続数字間には 2 秒程度のポーズ区間が挿入されている．横顔の口唇画像は，話者の右頬から 10cm 程度の場所にデジタルカメラを配し，撮影を行った．

3.2　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

∗ A multi-modal speech recognition using lip movement information extracted from side-face images

By Tomoaki Yoshinaga, Satoshi Tamura, Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology)

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 を構築した．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，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おり，λ_a+λ_v=1 という制約を設けている．

3.3 実験手法

評価には 38 名中 19 名分のデータを用いた．各話者について 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，クリーン環境下で収録されたデータに 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[5] を用いてマルチストリーム HMM の適応化を行った．適応化する対象は，音響ストリームの平均と分散のみとし，適応後のマルチストリーム HMM について同様の評価を行った．

3.4 実験結果

実験結果を表 1 に示す．音響と画像のストリーム重みは各 SN 比条件ごとに最適化を行っており，最適な音響ストリーム重みの値 (λ_a) を表中の Audio-visual の括弧内に表記した．MLLR あり，なしに関わらず全ての SN 比において画像特徴量を用いたことで正解精度が向上し，本手法の有効性を確認することができた．SN 比 5dB の時に最も数字正解精度の絶対値が向上し，MLLR なしの時で 5.9%，MLLR ありの時には 11.4%改善した．また，最適な λ_a の値は SN 比が劣化するほど小さくなる傾向がみられた．したがって，雑音が大きくなるほど画像情報に重点を置くことで，認識性能の最適化が行われていることが分かる．

図 3 に，MLLR なし，SN 比 5dB 条件での，音響特徴量のみでの正解精度，および画像特徴量を併せて用いた場合における λ_a を変化させたときの数字正解精度の推移の様子を示す．λ_a の変化に対し，緩やかに正解精度が変化していく様子が見て取れ，広い範囲のストリーム重みに対して画像特徴量の効果が表れていることがわかる．なお，MLLR ありの場合や，その他の SN 比においても同様のグラフが得られた．

なお，これらの正解精度の向上は，画像特徴量の利用による発声開始時刻の推定精度の向上に起因していることが確認されている [6]．

4 まとめ

本研究では，モバイル環境における雑音に頑健な音声認識手法として，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を提案した．提案手法を連続数字音声認識実験で評価したところ，様々な SN 比条件で画像特徴量を用いることによる耐雑音性の向上が確認された．また，MLLR による雑音適応化との併用により更なる耐雑音性の向上が実現された．

今後の課題としては，1) 実環境での利用を考慮し，画像情報の外乱成分に対する対策手法の提案と，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手法の提案，3) 口唇の開閉だけでなく，音素種の識別にも効果のある画像特徴量の提案，などが挙げられる．

表 1. 各条件下における数字正解精度の比較

SN 比 (dB)	MLLR なし		MLLR あり	
	Audio	Audio-visual	Audio	Audio-visual
20	89.6%	90.5% (0.60)	96.3%	96.7% (0.90)
15	71.8%	75.4% (0.55)	89.8%	92.3% (0.60)
10	48.4%	53.1% (0.60)	65.8%	73.3% (0.45)
5	26.4%	32.3% (0.45)	37.5%	48.9% (0.45)

（括弧内は最適な λ_a の値）

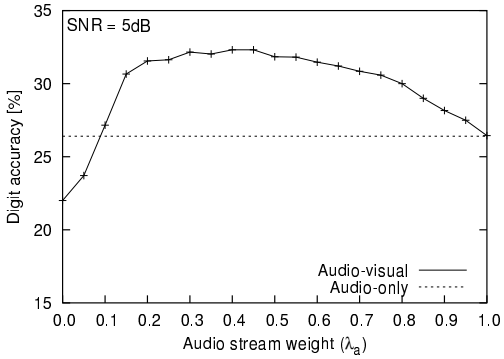

図 3. 音響ストリーム重みによる数字正解精度の推移

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] 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] 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法の提案と評価，” 情処研報, 2002-HI-97-6 / 2002-SLP-40-6, vol.2002, no.10, pp.33–38 (2002-2).

[5] C.J. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, vol.9, no.2, pp.171-185 (1995-4).

[6] 吉永智明，田村哲嗣，岩野公司，古井貞熙，“横顔の動画像情報を用いたマルチモーダル音声認識，” 情処研報, 2003-SLP-46-11, vol.2003, no.58, pp.61–66 (2003-5).