論文 / 著書情報
Article / Book Information

論題(和文)	実環境におけるマルチモーダル音声認識の評価
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 2002年春季講演論文集, Vol. , No. 3-5-5, pp. 151-152
Citation(English)	, Vol. , No. 3-5-5, pp. 151-152
発行日 / Pub. date	2002, 3

実環境におけるマルチモーダル音声認識の評価 *

◎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]．我々もこれまでに，オプティカルフローを用いた手法を提案してきた [3, 4]．しかし従来のマルチモーダル音声認識の研究では，音声のみに雑音成分を含めた実験は多く行われているが，画像情報の雑音や外乱に対する頑健性についてはほとんど議論されていない．そこで本研究では，実環境で収録したデータを用いて，我々が提案したマルチモーダル音声認識手法の，音響・画像的な頑健性の評価を行った．

2. マルチモーダル音声認識手法

本研究で用いたマルチモーダル音声認識手法 [3, 4] を概説する．まず，入力された画像列に対して「画像中の明度パターンの見かけ上の速度分布」であるオプティカルフローを計算し，その結果から画像特徴量を生成する．オプティカルフローには，パターンマッチングや物体の形状といった事前知識を用いずに計算できるという利点があり，このため頑健な画像特徴量を得ることができる．そして画像特徴量と音声信号から抽出した音響特徴量を融合して音響–画像特徴量を生成し，これを用いて HMM の学習および認識を行う手法となっている．

3. データベース

3.1. 学習データ

学習データとしては，音響・画像ともにクリーンな環境で収録した，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 個発声しており，全体の総時間長は約 2 時間半である．

3.2. テストデータ

評価用データとしては，高速道路を走行中の乗用車内で収録した，約 1 時間分の数字連続読み上げデータを用いた．学習データに含まれない 6 名の男性話者に，助手席で 2〜6 桁の数字を 115 個発声してもらい，音声は襟元に着用したピンマイクで，画像は顔下半分を正面に設置したビデオカメラで収録した．この画像データの例を図 1 に示す．音響雑音には，ほぼ定常なエンジン音や風切り音のほか，ウィンカー音や橋板接合部の通過音などの非定常な雑音があり，S/N 比はおよそ 10〜15dB であった．画像外乱としては，陸橋や標識の影による瞬間的な明度の変化，走行振動による顔のブレ，カーブ通過時には日射角度の移動にともなう陰影の変化などが観測された．

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　：CMN を行った MFCC 12 次元 およびその Δ，$\Delta\Delta$ 成分， 対数パワーの Δ，$\Delta\Delta$ 成分 特徴量次元数：38 次元
画像	フロー演算繰り返し回数：5 回 抽出特徴量　：フロー積分成分の最大・最小値 特徴量次元数：2 次元

4. 実験条件

4.1. 音響・画像特徴量

今回の実験に使用した音響特徴量および画像特徴量を表 1 に示す．音響特徴量については，雑音やマイクロホンによる歪みを除去するために CMN を行った．画像特徴量は文献 [4] で提案した 2 種類のパラメータのうち，口の動き情報をより多く含んでいるオプティカルフローの積分結果の最大・最小値を用いた．

4.2. 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．11 名分の学習データを用い EM アルゴリズムによって HMM の学習を行い，6 名分のテストデータの認識実験を行って，その数字正解精度で性能を評価した．

実験に使用した 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 は各々のストリーム重みである．本研究では，使用した画像特徴量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，それ以外のモデルは $\lambda_A = 1$，$\lambda_V = 0$ とした．

5. 実験結果・考察

図 2 (a) に，音響–画像特徴量で学習・認識したときの認識率，およびベースラインである音響特徴量のみでの認識結果を示す．これより最適なストリーム重み

---

* Evaluation of multi-modal speech recognition in real environments,
by Satoshi Tamura, Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology).

図 1: 評価用データ（ 顔の一部に日光が射している例）

表 2: 実験結果のまとめ

	Audio–only	Audio–visual
(a)	58.30%	63.80% (λ_A=0.70)
(b)	82.46%	85.47% (λ_A=0.78)
(c)	76.30%	78.95% (λ_A=0.76)

においては，正解精度でみると絶対値で約 6%の改善，誤り率でみると相対値で約 13%の削減に成功した．次に，音響特徴量のみ，音響–画像特徴量のそれぞれについて，HMM の各状態の平均および分散を MLLR 法 [5] により適応化した．このときのクラスタ数は 3 を用いた．この HMM を用いて認識を行ったときの結果を図 2 (b)(c) に示す．(b) は各話者のテストデータ全てを一度に用いて適応化するバッチ方式，(c) は個々の話者のデータを 6 つに分け逐次的に適応化を行うインクリメンタル方式で，それぞれ教師なし適応を行ったときの認識率である．また，図 2 (a)〜(c) の実験結果をまとめたものを表 2 に示す．これらより音響特徴量のみの結果と比べ，バッチ方式で約 17%，インクリメンタル方式で約 11%の誤り率の削減が確認された．このように認識性能が向上した理由としては，音響–画像特徴量を用いることで無音区間の推定精度が向上し，silence の部分における数字の挿入誤りなどが抑制されたことが考えられる．以上のことから，音声のみならず，画像に雑音が入った実環境においても，本手法は有効であることが確認できた．

6. まとめ

本研究では，以前我々が提案したマルチモーダル音声認識手法について，音響・画像両面の頑健性を，実環境データによる認識実験によって評価した．その結果，MLLR 法による雑音適応手法と合わせることで，音響特徴量のみバッチ式 MLLR 適応の場合と比べ，約 17%の誤り率の削減に成功した．このことから，我々のマルチモーダル音声認識法は，実環境でも有効に機能することが確かめられた．今後の課題としては，より有効な画像特徴量の抽出法や，雑音に応じたストリーム重みの決定法の検討が挙げられる．

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．

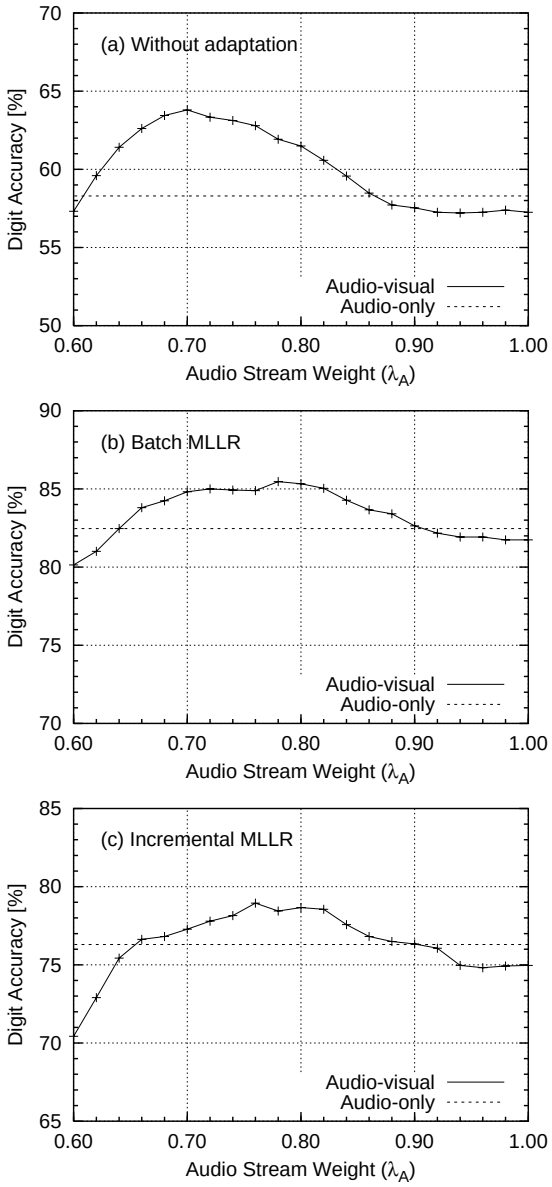

図 2: 実験結果（ (a) 適応なし，(b) バッチ適応方式，(c) インクリメンタル適応方式 ）

参考文献

[1] 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).

[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] 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).

[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] C.J. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, pp.171-185 (1995-4).