論文 / 著書情報
Article / Book Information

論題(和文)	マルチモーダル音声対話システムのための音響・画像重みの推定手法の検討
Title(English)	
著者(和文)	松尾 俊秀, 岩野 公司, 古井貞熙
Authors(English)	Toshihide Matsuo, Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会2009年春季講演論文集, , , pp. 115-116
Citation(English)	, , , pp. 115-116
発行日 / Pub. date	2009, 3

マルチモーダル音声対話システムのための音響・画像重みの推定手法の検討*

◎松尾俊秀 (東工大), 岩野公司 (武蔵工大), 古井貞熙 (東工大)

1 はじめに

近年，雑音環境における音声認識性能を向上させるため，音声と共に唇動画像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．我々の研究室でも，マルチストリーム HMM を利用したマルチモーダル音声認識の研究を行っており，それを利用した音声対話システムの構築について検討を進めている [1]．対話システム実現のためには，音声認識システムの即時性を確保する必要があり，特徴量抽出や探索の高速化を図る以外に，最適な音響・画像ストリーム重みを素早く決定することが重要となる．そこで本研究では，システム運用時に音響・画像ストリーム重みを素早く最適化する方法として，それぞれのストリームのエントロピーの変化情報を利用した教師なし重み推定手法の提案を行う．実験により，様々な条件における提案手法の有効性の評価を行う．

2 音響・画像重みの自動推定

本研究では，先行研究 [1] で構築されたマルチモーダル音声認識システムを使用している．このシステムではマルチストリーム HMM を利用しており，時刻 t の音響・画像特徴量 O_t の観測確率は，対数尤度 $b(O_t)$ を用いて以下の式で示される．

$$b(O_t) = W_A b_A(O_{At}) + W_V b_V(O_{Vt}) \quad (1)$$

ただし，$b_A(O_{At})$，$b_V(O_{Vt})$ はそれぞれ音響特徴量 O_{At}，画像特徴量 O_{Vt} に対する対数尤度である．W_A，W_V は音響，画像ストリーム重みであり，

$$W_A + W_V = 1\ ,\ 0 \leq W_A, W_V \leq 1 \quad (2)$$

の条件を満たすものとする．

本研究では，このストリーム重みを，それぞれのストリームのエントロピーを用いて推定する手法の提案を行う．

2.1 従来法

エントロピーを用いたストリーム重みの推定法として Misra らによる手法 [2] が従来法として挙げられる．この手法では，ストリーム $s(= A, V)$ に対するエントロピー h_s と，それを用いて推定される重み W_s は，以下のように定義される．

$$h_s = -\sum_p P(\lambda_p|O_s) \cdot \log_2 P(\lambda_p|O_s) \quad (3)$$

$$W_s = \frac{1/h_s}{\sum_s 1/h_s} \quad (4)$$

ここで，O_s はストリーム s の入力特徴量，λ_p はクラス p のモデルをあらわす．クラスとしては音素（モノフォン）を利用する．

Misra らの手法では，エントロピーの逆数の比によって各ストリーム重みを配分しており，エントロピーの大小に応じた信頼度を設定することができる．しかし，比によって得られた数値自体が最適重みとなる保証がなく，推定値が最適値から大きく外れる可能性がある．本論文では，事前に雑音のないクリーンな環境で重みの最適化を行うことを前提として，認識時の雑音環境下で各ストリームのエントロピーを計算し，その値がクリーン環境下でのエントロピーからどの程度変化したかに応じて，事前に最適化した重みを調整する手法を提案する．クリーン環境での最適重みを基準として用いることで，精度の良い重みの推定を行うことができる．

2.2 環境の違いによるエントロピーの変化を利用した重み推定手法

本論文における提案手法での推定重みは以下の式で定義される．

$$\hat{W}_s = W_s^c \cdot \frac{h_{max} - h_s}{h_{max} - h_s^c} \quad (5)$$

ここで W_s^c は，事前に求めたクリーン環境でのストリーム s の最適重みであり，h_s^c はクリーン環境でのストリーム s のエントロピーである．h_{max} はエントロピーの上限値であり，

$$h_{max} = \log_2 n \quad (6)$$

で求めることができる．n はクラス（モノフォン）の数であり，本研究では無音を表す silB, silE, sp を含め 42 とした．

この式により，雑音環境におけるストリーム s のエントロピー h_s がクリーン環境時のエントロピーと等しければ重みは変化せず，エントロピーが増加して上限値に近づくに従って，そのストリームの信頼度は低下していると見なされ，重みが 0 に近づく．したがって，この式は，$h_s \geq h_s^c$ となるときにのみ適用することとし，$h_s < h_s^c$ となるときには推定値は不定とし，もう一方のストリームの重みが決定された後にその値を 1 から減ずることで求める．(両ストリームの重みとも不定となった場合には，クリーン環境の最適重みをそのまま用いる.)

両ストリームとも $h_s \geq h_s^c$ となった場合には，音響と画像それぞれのストリームの重みは式 (5) によって独立に算出されるため，式 (2) の条件を満たさなくなる可能性がある．その場合には，以下の式によって最終的な調整を行う．

$$W_s = \frac{\hat{W}_s}{\hat{W}_A + \hat{W}_V} \quad (7)$$

---

* An audio-visual weight estimation method for multimodal dialogue systems By Toshihide Matsuo (Tokyo Institute of Technology), Koji Iwano (Musashi Institute of Technology) and Sadaoki Furui (Tokyo Institute of Technology)

Table 1　各手法の認識結果 (単語正解精度)

SNR	音響のみ	従来手法	提案手法	手動最適重み
10dB	7.1%	7.7% (0.76)	10.5% (0.40)	14.1% (0.20)
15dB	23.1%	24.7% (0.84)	31.7% (0.40)	34.8% (0.25)
20dB	47.2%	47.9% (0.92)	52.5% (0.43)	53.1% (0.35)
clean	74.3%	74.6% (0.93)	74.9% (0.50)	74.9% (0.50)

なお，本研究では，事前のクリーン環境下での重みの最適化は手動で行った．その際，重みを 0.05 から 1.00 まで 0.05 ずつ変化させながら最適化を行った．

3　実験

3.1　実験条件

認識デコーダには，先行研究 [1] においてマルチストリーム HMM が扱えるように改良の施された Julius を用いた．学習データ，評価データには先行研究 [1] で収録したデータベースを用いている．言語モデルは 2-gram と逆向き 3-gram であり，システムとの対話を想定して作成した語彙数 6,839 単語，1,206 文の模擬対話文から作成している．また，音響モデルは，男性話者 15 名による合計 1,509 発声（総時間長は約 2 時間半）のデータから構築された，left-to-right 型 triphone HMM を利用している．各モデルの状態数は 3 である．音響・画像の融合 HMM の構築方法は先行研究 [1] に準じている．HMM の混合数はクリーン環境における予備実験により最適化されており，音響ストリームで 8，画像ストリームで 2 とした．

音響特徴量としては，ケプストラム平均正規化法を行った MFCC 12 次元とその Δ，$\Delta\Delta$ 成分，および対数パワーの Δ，$\Delta\Delta$ 成分の計 38 次元を，画像特徴量としてオプティカルフローのフローベクトルの水平・垂直方向の分散値 2 次元にその差分成分をベクトル結合した計 4 次元を用いた．

評価データには，男性話者 10 名による模擬対話文発声（合計 400 発声）を用いており，SNR = 10，15，20dB で白色雑音，電子協騒音データベースの駅雑音とエレベータ雑音を重畳して実験を行った．すべての種類の雑音条件においてほぼ同じ傾向の結果が得られたため，本論文では白色雑音を重畳したときの結果のみを示す．

3.2　従来手法との比較

まず，評価データ全体を利用して重みを推定し，得られた重みによって認識を行った場合の，従来手法 [2] と提案手法の性能を比較した．その際，従来手法ではフレームごとにエントロピーの計算と重み推定を行っているのに対し，提案手法では発話全体に対するグローバルな重みを推定しているため，式 (3) の事後確率 $P(\lambda_p|O_s)$ として，推定用データ全体から得られる事後確率のフレーム平均値を利用し，エントロピーと重みを求めた．Table 1 にそれぞれの手法による単語正解精度，ベースラインとなる音響情報のみ (学習にも認識にも画像情報を用いていない) での単語正解精度，手動で重み最適化を行った場合の単語正解精度を示す．また，推定された音響重みを括弧内に示す．この結果より，従来手法，提案手法ともに音響のみの結果からの改善が得られており，さらに提案手法の方が従来手法よりも良好な結果となっていることが確認された．

Table 2　様々な条件下での認識結果 (単語正解精度)

SNR	one	utr	pre	all
10dB	10.5% (0.40)	10.5%	9.1%	10.5% (0.40)
15dB	32.5% (0.42)	32.8%	31.9%	31.7% (0.40)
20dB	51.6% (0.44)	50.6%	51.0%	52.5% (0.43)
clean	74.9% (0.50)	74.9%	74.9%	74.9% (0.50)

3.3　推定用データの条件の違いによる性能の変化

提案手法において，重み推定用データの条件を変化させた場合にどのような性能となるかを検討した．各雑音環境における評価データ中の 1 発声のみを利用して推定した重みを，その雑音環境での全評価データの認識に使用した場合（one），発話ごとに重みを推定し，その発話の認識を行った場合（utr），複数ある発話が続けてシステムに入力されたと想定し，1 つ前の発話から推定した重みを次の発話の認識に利用した場合（pre），評価データ全体を利用して重みを推定した場合（all），について認識実験を行った．結果を Table 2 に示す．また，重み推定に 1 発声のみのデータを利用した場合 (one) と評価データ全体を利用した場合 (all) は，全評価データの認識に利用する重みが等しいため，その音響重みを括弧内に示す．

この結果から，今回の実験では，推定用データを 1 発話に減らした場合（one）でも十分な性能が得られることが確認された．雑音環境が随時変化してしまう状況を想定すると，発話ごとの重み推定は有用であると考えられる．その際，当該発話から重みを推定した場合（utr）と，一つ前の発話から重みを推定した場合（pre）が考えられるが，どちらも認識性能の向上が得られている．対話システムへの応用を考えると後者の方式が実時間性があり，望ましいが，今回の実験では，前者に比べ若干性能が劣っていることが分かる．

4　まとめ

本論文では，マルチモーダル音声対話システムのためのエントロピーに基づく音響・画像重みの教師なし推定手法の提案を行った．提案手法は，クリーン環境における最適重みをエントロピーの変化に応じて調整することで重み推定を行う．認識実験の結果，従来手法よりも良好な結果が得られ，また発話を単位とした重み推定も有効であることが確認された．音声対話システムへの応用を考えると，より短い時間で正確な重みを推定することが求められる．今後の課題として，発話冒頭の数フレームを利用した重み推定の検討などが挙げられる．

謝辞　本研究は，株式会社東芝との共同研究として行われました．ここに深く感謝いたします．

参考文献

[1] 高山他，春季音講論，2-9-13, pp.61-62, 2007.
[2] H. Misra, et al., ICASSP, vol.2, pp.741-744, 2003.