Vol.2009-NL-191 No.24
Vol.2009-SLP-76 No.24
2009/5/22

マルチモーダル音声認識における ストリーム重みの教師なし推定法の検討

岩野公司[†1]　松尾俊秀[†2]　古井貞熙[†2]

本稿では，対話システムへの利用を想定したマルチモーダル音声認識のための，音響・画像ストリーム重みの教師なし推定手法の提案を行う．提案手法では，まず，クリーン環境における最適重みと，それぞれのストリームのエントロピーを事前に計算しておく．システム利用時の雑音環境において，それぞれのストリームのエントロピーを計算し，それがクリーン環境で観測されたものからどれだけ変化しているかに応じて，重みを調整する手法である．複数の男性話者の発声が収録されたマルチモーダルデータベースを利用して提案手法の評価を行ったところ，様々な雑音条件において，Misra らが提案している従来までのエントロピーに基づく教師なし推定手法よりも，提案手法が良好な性能を示すことが確認された．

A study on unsupervised stream-weight estimation for multimodal speech recognition

Koji Iwano ,[†1] Toshihide Matsuo [†2]
and Sadaoki Furui[†2]

This paper proposes an unsupervised stream-weight estimation method for an audio-visual speech recognizer constructed for spoken dialogue systems. In the proposed method, audio and visual stream weights are optimized and stream entropies of the audio and visual signals are calculated in advance in the clean condition. In the weight estimation process, the stream entropies under an actual noisy condition are calculated and compared with those of the clean condition. And then the stream weights are adaptively controlled according to the differences of entropies between the clean and noisy conditions. Evaluations were conducted by using an audio-visual speech database collected from multiple male speakers. Experimental results show that the proposed method yields better performance in various noise conditions than the entropy-based unsupervised weight estimation method proposed by Misra et al.

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．我々の研究室においても，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，ユーザがシステムを利用する際の音響雑音や画像への外乱の状況を素早く把握し，最適な音響・画像ストリーム重みを決定することが重要となる．

これまでに，音響・画像ストリーム重みの推定法としては，最小分類誤り（MCE）基準による推定手法[2)–4)] や，最大エントロピー基準による推定手法[3)]，ゆう度比最大化基準による推定手法[5)] などが提案されており，それぞれ有効性が確認されている．これらの手法は，モデルや状態ごとに最適重みを推定することが可能であり，全てのモデルで同じ重みを利用することを前提としたグローバルな重み推定に比べて，より精度良く最適重みを設定することが可能である．一方，これらの手法は正解の音素（状態）系列を必要とする教師あり推定であることから，対話システムへの利用を考えると，一度認識結果を得た上で，その結果を正解とみなして重み推定を行う必要があり，即時性に優れない．本研究では，即時性を重視し，グローバルな重みを教師なしで推定する手法について検討を行う．

従来までの重みの教師なし推定法としては，推定した音響信号の SN 比を利用する手法[6),7)] や，個々のストリームのエントロピー情報を利用した手法[8)] が挙げられる．このうち，Misra らの手法[8)] では，入力の個々のストリームに対して，それぞれのエントロピーを計算し，エントロピーが大きいストリームは信頼性が乏しいと判断してその重みを小さく設定するものである．具体的には，エントロピーの逆数の比でストリーム重みの配分を行っている．しかし，エントロピーの逆数と，最適となるストリーム重みの数値の間には直接の対応関係がないことから，推定された重みが最適値から大きくずれる可能性がある．そこで本研究では，音響雑音や画像への外乱がない，クリーン環境における音響・画像ストリームのエントロピーと，それぞれの最適重みを予め求めておき，それらの値を基準にして，システム利

†1 東京都市大学 環境情報学部 情報メディア学科
Faculty of Environmental and Information Studies, Tokyo City University
†2 東京工業大学大学院 情報理工学研究科 計算工学専攻
Department of Computer Science, Tokyo Institute of Technology

Vol.2009-NL-191 No.24
Vol.2009-SLP-76 No.24
2009/5/22

用時の各ストリームのエントロピーがクリーン環境からどれだけ変化したかに応じてストリーム重みを調整する手法を提案する．クリーン環境の最適重みとエントロピーの対応関係を基準とすることで，より安定した重み推定を実現することが可能となる．

本稿では，この新しいエントロピーに基づく重み推定手法の提案を行い，従来手法との性能比較を行う．次に，提案手法を実際の対話システムに利用することを考慮して，重み推定用データを発話単位とした場合の性能について論じる．さらに，音響雑音への対策としてスペクトルサブトラクションを施した場合にも，提案手法によって推定された重みによって画像情報が有効に作用し，認識性能が向上することを検証する．

以降では，まず，2 章で使用するマルチモーダル音声認識システムと音響・画像特徴量について説明する． 3 章で，Misra らが提案したエントロピーに基づく重み推定手法の説明と，エントロピーの変化量を利用した重み推定手法の提案を行う．提案手法の性能評価について 4 章で論じ， 5 章で本稿をまとめる．

2. マルチモーダル音声認識システム

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．図 1 に，システムの処理の流れを示す．音声・画像はそれぞれ標本化周波数 16kHz，60Hz でサンプリングされ，それぞれ 100Hz の音響特徴量と 60Hz の画像特徴量に変換される．音声，画像の特徴量をそれぞれ独立に抽出した後，両者のフレームレートを合わせるため，画像特徴量に 3 次スプライン補間を施して 100Hz の特徴量に変換し，2 つの特徴量をフレームごとに結合して，100Hz の音響・画像特徴量を作成する．この融合特徴量をマルチストリーム HMM に入力することで認識を行う．

2.1 音響・画像特徴量の抽出

音響特徴量には，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 次元ベクトルを用いる．分析窓にはハミング窓を用いており，その幅は 25ms である．なお，発話ごとにケプストラム平均正規化（Cepstral Mean Normalization: CMN）を行っている．

画像特徴量にはオプティカルフローに基づく特徴量[9)] を用いている．画像特徴量の計算は，動画データ中の各静止画像に対して行う．元の動画像は，解像度 720×480 の DV フォーマットで収録されているが，計算量削減のため，まず解像度を 640×480 に変換（削減）する．解像度変換後の画像に対し，口唇領域の検出を行い，切り出された口唇画像から画像特徴量を抽出する．口唇領域の検出には OpenCV[10)] で採用されている，Adaboost に基

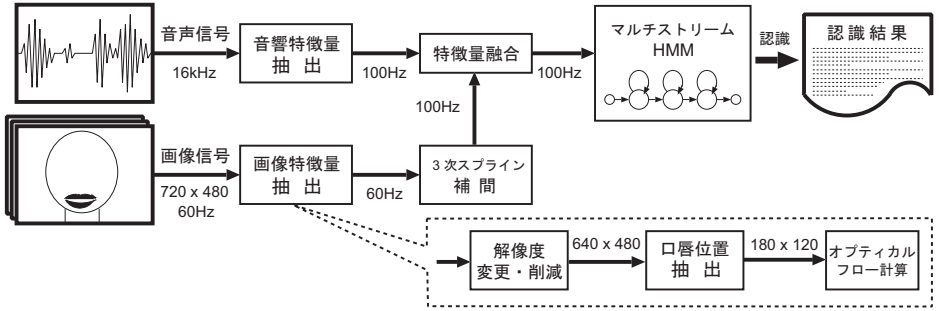

図 1 先行研究[1)] で構築したマルチモーダル音声認識システムの処理の流れ

づく複数識別器を利用した物体検出法[11)] を利用している．この手法の検出結果は，サイズ可変の矩形領域で得られるため，その領域の中心点を求め，その点を中心とした固定サイズ（180×120）の矩形領域で再切り出しを行い，口唇画像の大きさを揃えている．連続する 2 フレームの口唇画像から，Lucas-Kanade 法[12)] によってオプティカルフローを計算する．図 2 にフローベクトルの計算結果の例を示す．図 2 の (a) から (b) に口唇画像が変化したときに観測されたフローベクトルが (c) である．フローベクトルの抽出は，全ピクセル位置に対しては行わず，図に示すような 216 点 (= 18×12) を選択し，その点位置のみで行っている．得られたフローベクトル群の水平・垂直方向の分散値を求め，さらに，それらの Δ 成分をベクトル結合した計 4 次元のベクトルを最終的な画像特徴量とする．

2.2 マルチストリーム HMM

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，

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

の条件を満たすものとする．

融合モデルの構築方法は以下の通りである[1)]．

(1) 音響特徴量のみを用いて音響モデル（triphone HMM）を学習する．各モデルの状態数は 3 とする．この時点で，状態共有化も行う．

(2) 得られた音響モデルを用いて，学習データの強制切り出しを行い，状態ごとの時間情

Vol.2009-NL-191 No.24
Vol.2009-SLP-76 No.24
2009/5/22

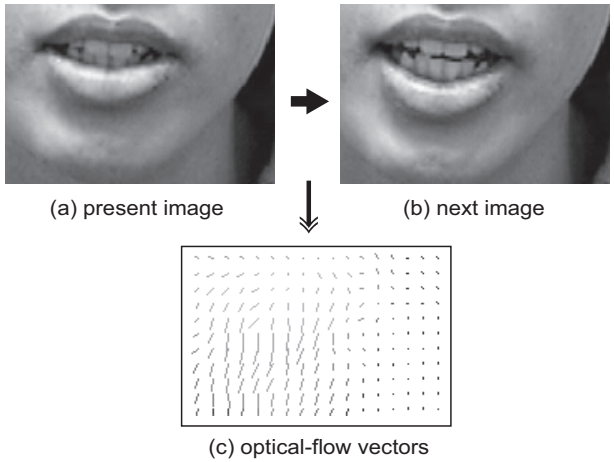

図 2 時間的に連続したフレームの口唇画像とそれらから抽出されたフローベクトル

報がついたラベルを作成する．

(3) このラベルを用いて，状態を単位として画像モデル（GMM）を学習する．

(4) 音響モデル，画像モデルのモデルパラメータを，一致する音素・状態ごとに音響，画像ストリームとして融合し，音響モデルの共有化構造と同じ構造を持つ音響・画像のマルチストリーム HMM を構築する．

3. エントロピーを用いた音響・画像重みの教師なし推定

3.1 Misra らの従来手法

エントロピーを用いたストリーム重み推定法である Misra らの手法[8] について説明する．この手法では，推定されるストリーム重み W_s $(s = A, V)$ は，以下のように定義される．

$$W_s = \frac{1/h_s}{\sum_s 1/h_s} \tag{3}$$

ここで，h_s はストリーム s に対するエントロピーであり，以下の式で与えられる．

$$h_s = -\sum_p P(\lambda_p|O_s) \cdot \log_2 P(\lambda_p|O_s) \tag{4}$$

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．本研究では，クラスとして表 1 に示す 42 の音素（モノフォン）を利用する．

表 1 エントロピーの計算に利用したクラス（モノフォン）

silB, silE, sp, a, aa, b, by, ch, d, ee, e, f, g, gy, h, hy, i, ii, j, k, ky, m, my, n, N, ny, o, oo, p, py, q, r, ry, s, sh, t, ts, u, uu, w, y, z

この手法では，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，エントロピーの逆数と，最適となるストリーム重みの数値の間に直接の対応関係がないことから，推定された重みが最適値から大きくずれる可能性がある．

なお，文献 8) では，この方法によって，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．本研究では，フレーム単位ではなく，発話全体を単位とした重みの推定を目的とするため，この手法を提案手法と比較する際には，式 (4) の事後確率 $P(\lambda_p|O_s)$ として，推定用データ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る．

3.2 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

ここでは，音響や画像に対する雑音のない（クリーンな）環境下で，事前に最適重みを求めておくことを前提として，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におけるエントロピーからどの程度変化したかに応じて，重みを調整する手法を提案する．

提案手法における推定重みは以下の式で定義される．

$$W_s = W_s^c \cdot \frac{h_{max} - h_s}{h_{max} - h_s^c} \tag{5}$$

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，

$$h_{max} = \log_2 n \tag{6}$$

で求めることができる．n はクラス（モノフォン）の総数であり，本研究では 42 である．

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て重みが 0 に近づく．この式は，エントロピーがクリーン環境から増加しているとき（$h_s \geq h_s^c$）にのみ適用することとし，減少（$h_s < h_s^c$）するときには推定値は不定として，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

Vol.2009-NL-191 No.24
Vol.2009-SLP-76 No.24
2009/5/22

は，クリーン環境の最適重みをそのまま用いる．

両ストリームともエントロピーが増加している（$h_s \geq h_s^c$）場合には，音響・画像それぞれのストリームの重みが式 (5) によって独立に算出されるため，式 (2) の制約を満たさなくなる可能性がある．その場合には，以下の式によって正規化した重み $\hat{W}_s$ を最終的な推定重みとする．

$$\hat{W}_s = \frac{W_s}{W_A + W_V} \tag{7}$$

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．その際には，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．

4. 性能評価実験

4.1 実 験 条 件

モデル学習用データ，評価用データには先行研究[1] で収録したマルチモーダルデータベースを用いている．モデル学習用データは，男性話者 15 名による ATR 音素バランス文の発声であり，各話者約 100 文，合計 1,509 発声で，全体の総時間長は約 2 時間半である．構築した triphone 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 となった．なお，重み推定用データに対するエントロピーを計算するためには，表 1 に示す音素クラスごとに確率 $P(O_s|\lambda_p)$ を求める必要がある．そこで本研究では，1) 認識用に構築したマルチストリーム triphone HMM の中から中心音素が p となるモデルを全て取り出し，2) それらを並列に結合してクラス p のモデル λ_p を作成し，3) このモデルに重み推定用のデータを入力したときの，ストリーム s から得られる観測確率を $P(O_s|\lambda_p)$ とした．

評価用データには，男性話者 10 名による模擬対話文を使用した．各話者 40 文，合計 400 発声で，総時間長は約 30 分である．模擬対話文は，飲食店舗検索のための音声対話システムへの入力を想定し，「柏で定食屋さんありますか？」などの文章となっている．なお，学習用データと評価用データの発声者に重なりはない．評価用データには，SNR = 10，15，20dB で，白色雑音と電子協騒音データベース[13] の駅雑音を重畳した．

言語モデルには 2-gram，逆向き 3-gram を用いた．学習にはシステムとの対話を想定して作成した 1,206 文の模擬対話文を用い，語彙数は 6,839 となった．

認識デコーダには，先行研究[1] でマルチストリーム HMM が扱えるよう改良を施した Julius を用いた．

表 2 各重み推定手法の認識性能（単語正解精度）

雑音条件		音響のみ	従来手法[8]	提案手法	手動最適重み
clean		74.3 %	74.6% (0.93)	74.9% (0.50)	74.9% (0.50)
白色雑音	20dB	47.2%	47.9% (0.92)	52.5% (0.43)	53.1% (0.35)
	15dB	23.1%	24.7% (0.84)	31.7% (0.40)	34.8% (0.25)
	10dB	7.1%	7.7% (0.76)	10.5% (0.40)	14.1% (0.20)
駅雑音	20dB	54.6%	56.2% (0.83)	58.2% (0.39)	59.0% (0.45)
	15dB	28.2%	29.4% (0.79)	32.1% (0.31)	33.3% (0.30)
	10dB	5.6%	6.4% (0.75)	7.9% (0.33)	8.1% (0.35)

4.2 提案手法と従来手法の比較

まず，提案手法と従来手法[8] の性能比較を行う．ここでは，評価用データ全体を利用して雑音環境ごとに重みを推定し，得られた重みによって各発話の認識を行った．表 2 に，従来手法，提案手法それぞれの単語正解精度，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，どの雑音条件についても，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いることがわかる．また，提案手法の方が従来手法よりも良好な結果が得られていることもわかる．

4.3 推定用データの条件の違いによる性能の変化

次に，提案手法において，実際の対話システムへの利用を想定し，重み推定用データの条件を様々に変化させた場合の性能について検討した．条件としては，

(1) 各雑音環境における評価用データ中の 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）
(2) 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）
(3) 複数の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）
(4) 評価用データ全体を利用して重みを推定した場合（all）

の 4 通りについて検討する．実験結果を 表 3 に示す．表中 “all” は，表 2 の提案手法の結果と同じである．また，重み推定に 1 発声のみのデータを利用した場合（one）と評価用データ全体を利用した場合（all）については，全ての発話の認識に対して単一の重みが使用されるため，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）でも全評価用データを用いた場合（all）の結果とほぼ変わらず，十分な性能が得られることが確認された．雑音が随時変化してしまう状況を想定す

Vol.2009-NL-191 No.24
Vol.2009-SLP-76 No.24
2009/5/22

表 3 推定用データの条件を様々に変化させたときの提案手法の認識性能（単語正解精度）

雑音条件		one	utr	pre	all
clean		74.9% (0.50)	74.9%	74.9%	74.9% (0.50)
白色雑音	20dB	51.6% (0.44)	50.6%	51.0%	52.5% (0.43)
	15dB	32.5% (0.42)	32.8%	31.9%	31.7% (0.40)
	10dB	10.5% (0.40)	10.5%	9.1%	10.5% (0.40)
駅雑音	20dB	57.2% (0.30)	58.7%	58.1%	58.2% (0.39)
	15dB	33.3% (0.30)	32.2%	31.4%	32.1% (0.31)
	10dB	8.0% (0.30)	8.1%	7.9%	7.9% (0.33)

表 4 SS 法による雑音処理を行った場合の提案手法の効果（単語正解精度）

雑音条件		音響のみ	one	utr	pre	all	手動最適重み
clean		75.6%	76.0% (0.50)	76.0%	76.0%	76.0% (0.50)	76.0% (0.50)
白色雑音	20dB	56.6%	60.8% (0.50)	60.9%	60.8%	60.6% (0.52)	60.8% (0.50)
	15dB	35.1%	40.0% (0.51)	40.7%	41.0%	40.2% (0.53)	42.9% (0.30)
	10dB	14.6%	17.9% (0.50)	18.4%	18.1%	17.7% (0.52)	21.9% (0.30)
駅雑音	20dB	59.2%	61.4% (0.41)	60.1%	60.4%	60.7% (0.45)	61.6% (0.40)
	15dB	39.6%	45.3% (0.34)	44.3%	44.8%	44.3% (0.40)	45.5% (0.35)
	10dB	18.5%	22.2% (0.34)	21.1%	21.5%	21.2% (0.47)	22.0% (0.40)

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即時性が高く，より望ましいが，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わかる．

4.4 音響雑音に対する雑音処理を施した場合の有効性の検証

最後に，音響雑音対策としてスペクトルサブトラクション法（SS 法）[14] を適用し，音響のみでの認識性能を向上させた場合における，提案手法の有効性について検証する．

表 4 に，SS 法を適用した場合における，提案手法による認識性能を示す．推定用データの 4 つの条件（one, utr, pre, all）は，前節 4.3 と同じである．表 4 と表 2 の「音響のみ」の結果を比較すると，SS 法の効果により，全ての雑音条件で音響のみの認識性能が向上していることが分かる．また，SS 法を適用し，音響のみの認識性能が向上した条件であっても，提案する重み推定手法によって全ての雑音条件で性能の改善が見られ，手動で重み最適化を行った場合と近い性能を示すことがわかる．

5. ま と め

本稿では，音声対話システムへの利用を想定したマルチモーダル音声認識のための，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，クリーン環境で最適化された重みをエントロピーの変化量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ることがわかった．重み推定用データの条件を変化させた実験を行い，提案手法の重み推定が，発話を単位とした場合でも有効であることが確認された．また，スペクトルサブトラクションによる音響雑音処理を施した場合についても，提案手法が有効に機能することが確認された．今後の課題としては，実際の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ため，発話の冒頭部分のみを利用した場合の提案手法の有効性の検討や，モデルや状態を単位とした，より詳細な重み推定への改良などが挙げられる．

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) 高山俊輔，松尾俊秀，岩野公司，古井貞熙，“対話システムへの利用を想定したマルチモーダル音声認識の検討，” 電子情報通信学会技術研究報告，SP2007-4, vol.107, no.77, pp.19–24, 2007.
2) C. Miyajima, K. Tokuda, and T. Kitamura, “Audio-visual speech recognition using MCE-based HMMs and model-dependent stream weights,” *Proc. ICSLP2000*, vol.2, pp.1023–1026, Beijing, China, 2000.
3) G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, “Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR,” *Proc. ICASSP2002*, vol.1, pp.853–856, Orlando, FL, 2002.
4) K. Kumatani and S. Nakamura, “Audio-visual speech recognition based on optimized product HMMs and GMM based-MCE-GPD stream weight estimation,” *IEICE Trans. on Information and Systems*, vol.E86-D, no.3, pp.454–463, 2003.
5) 田村哲嗣，岩野公司，古井貞熙，“マルチモーダル音声認識におけるストリーム重み係数最適化の検討，” 電子情報通信学会技術研究報告，SP2003-153, vol.103, no.519, pp.241–246, 2003.
6) S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” *IEEE Trans. Multimedia*, vol.2, no.3, pp.141–151, 2000.

Vol.2009-NL-191 No.24
Vol.2009-SLP-76 No.24
2009/5/22

7) H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition," *Proc. ICASSP2001*, vol.1, pp.173–176, Salt Lake City, 2001.
8) H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," *Proc. ICASSP2003*, vol.2, pp.741-744, Hong Kong, 2003.
9) K. Iwano, S. Tamura, and S. Furui, "Bimodal speech recognition using lip movement measured by optical-flow analysis," *Proc. HSC2001*, pp.187–190, Kyoto, Japan, 2001.
10) `http://opencvlibrary.sourceforge.net`
11) P. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," *Proc. CVPR2001*, vol.1, pp.511–518, 2001.
12) B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proc. DARPA Image Understanding Workshop*, pp.121–130, 1981.
13) S. Itahashi, "Recent speech database projects in Japan," *Proc. ICSLP1990*, vol.2, pp.1081–1084, Kobe, Japan, 1990.
14) S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech, Signal Processing*, vol.27, no.2, pp.113–120, 1979.