2005－SLP－55　(6)
2005／2／4

実環境を考慮したマルチモーダル音声認識のための ストリーム重み最適化手法

田村 哲嗣　　岩野 公司　　古井 貞熙
東京工業大学　情報理工学研究科　計算工学専攻
〒 152-8552　東京都 目黒区 大岡山 2-12-1
E-mail: {tamura,iwano,furui}@furui.cs.titech.ac.jp

音声認識の頑健性向上の手法のひとつとして，口唇動画像の情報を利用するマルチモーダル音声認識の研究が進められている．実環境でのマルチモーダル音声認識の性能向上には，モデルとして用いるマルチストリーム HMM について，少量の適応データでも実行できるストリーム重み係数の自動最適化手法が必要不可欠である．本論文では，我々の従来手法（尤度比最大化法）を参考に，各 HMM の出力尤度平均を正規化するよう，尤度平均化基準による新たなストリーム重み最適化手法を提案する．車載カメラで収録した実環境データを用いた認識実験で，教師なし条件で提案法を評価したところ，音響特徴のみの結果と比べ，約 16%の正解精度が改善した．さらに MLLR 適応と提案手法を組み合わることで，約 23%の正解精度の改善に成功した．

A stream-weight optimization method for audio-visual speech recognition in real environments

Satoshi Tamura, Koji Iwano and Sadaoki Furui
Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
E-mail: {tamura,iwano,furui}@furui.cs.titech.ac.jp

Multimodal speech recognition which jointly uses acoustic and visual information has been actively investigated for increasing robustness of ASR. In order to improve performance of multimodal ASR in real environments, it is crucial to automatically optimize stream weights for multi-stream HMMs using a small size of data. This paper proposes a new stream-weight optimization method based on an output likelihood normalization (OLN) criterion; the stream weights are adjusted to equalize mean log likelihood values for all HMMs. Experiments were conducted using audio-visual data recorded in a driving car. A 16% improvement of recognition accuracy was achieved over an audio-only baseline by applying the unsupervised OLN stream-weight optimization. By additionally applying the MLLR adaptation, a 23% improvement was obtained.

1. はじめに

来るべきユビキタスコンピューティング時代に向け，特にカーナビや携帯電話などのユーザフレンドリーインターフェースとして，音声認識はいま最も注目されている技術のひとつである．加えて，情報化社会の発展にともない，大規模コンテンツの構築・活用といった分野においてインデキシングツール・書き起こし作成ツールとして，音声認識への期待が高まっている．しかし，現在の音声認識技術には，実環境など雑音が大きい状況の下では認識性能が著しく低下してしまうという問題があり，音声認識の実用化に向けての大きな課題となっている．

雑音下でも頑健に音声認識を行う手法のひとつとして，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が注目され，近年研究が進められている [1, 2, 3]．我々はこれまでに，口の動き情報を利用したマルチモーダル音声認識 [4] や，画像特徴量として口唇の幅・高さや歯の情報を用いたマルチモーダル音声認識システム [5] を構築している．さらに，モデルとして用いているマルチストリーム HMM における，ストリーム重み係数の自動最適化法として，尤度比最大基準による手法を提案しており，実環境データによる実験を行い，認識性能の改善を確認している [5]．しかし尤度比最大化法には，少量の最適化用データではストリーム重みを正しく推定できないという問題があった．

そこで，本論文ではこの改善手法として，尤度平均化基準によるストリーム重み最適化手法の提案を行い，実環境データを用いた認識実験により，提案手法と従来手法との性能比較と評価を行う．さらに，提案手法と MLLR による雑音適応を組み合わせた実験の結果について報告する．

2. ストリーム重み最適化手法

2.1. マルチストリーム HMM

本研究では音声認識のためのモデルとして，音響ストリームと画像ストリームより成るマルチストリーム HMM を用いている．このマルチストリーム HMM において，単語 w に対する音響–画像特徴量 $\mathbf{O}_t$ の対数尤度 $b_w(\mathbf{O}_t)$ は，式 (1) のように表される．

$$b_w(\mathbf{O}_t) = \lambda_{Aw} b_{Aw}(\mathbf{O}_{At}) + \lambda_{Vw} b_{Vw}(\mathbf{O}_{Vt}) \quad (1)$$

ただし t は時刻，$b_{Aw}(\mathbf{O}_{At})$，$b_{Vw}(\mathbf{O}_{Vt})$ はそれぞれ音響特徴量 $\mathbf{O}_{At}$，画像特徴量 $\mathbf{O}_{Vt}$ に対する単語 w の対数尤度，λ_{Aw}，λ_{Vw} は単語 w の HMM における音響，画像ストリーム重みで，本研究では以下の制約を設けている．

$$\lambda_{Aw} + \lambda_{Vw} = 1 \ , \ 0 \le \lambda_{Aw}, \lambda_{Vw} \le 1 \quad (2)$$

認識性能の向上には，各ストリームの雑音状況や信頼度に応じてストリーム重みを適切に設定することが有効である．しかし，ストリーム重み係数は，HMM 学習時には最尤推定法により最適化できないという問題がある．そのため，新たなストリーム重みの自動決定手法が必要である．

2.2. 尤度比最大基準による最適化手法

いま，デコーダが単語系列 w_t（$1 \le t \le T$, $w_t \in W$, W は認識辞書）を出力したとする．w_t が正解単語と異なる認識誤りは，モデルと入力特徴量のミスマッチにより，本来は正解でない単語 w_t の尤度が一番大きくなることに起因する．そこで適応データとそのラベルが与えられたとき，ラベルがある程度以上正しければ，第一仮説単語の対数尤度とそれ以外の単語の対数尤度の差が最大となるようにストリーム重みを調整することで，認識誤りを抑制できると考えられる．すなわち，

$$L(\Lambda) = \sum_{t=1}^{T} \sum_{w \in W} \Big\{ b_{w_t}(\mathbf{O}_t) - b_w(\mathbf{O}_t) \Big\}^2 \quad (3)$$

として，$L(\Lambda)$ を最大にするストリーム重み $\Lambda = \{\lambda_{Aw}\}$ を推定する．式 (3) より，$r \in W$ に対するストリーム重み λ_{Ar} の変化分 $\Delta\lambda_{Ar}$ は，次のように求められる．

$$\begin{aligned} \Delta\lambda_{Ar} &= \frac{N}{D} \quad (4) \\ N &= \sum_{t=1}^{T} \Big[\delta_{w_t=r} \Big\{ N b_r(\mathbf{O}_t) - \sum_{w \in W} b_w(\mathbf{O}_t) \Big\} \\ &\qquad \delta_{w_t \neq r} \Big\{ b_r(\mathbf{O}_t) - b_{w_t}(\mathbf{O}_t) \Big\} \Big] \\ D &= \sum_{t=1}^{T} \Big[\delta_{w_t=r} \cdot N d_r(\mathbf{O}_t) + \delta_{w_t \neq r} \cdot d_r(\mathbf{O}_t) \Big] \\ d_w(\mathbf{O}_t) &= b_{Aw}(\mathbf{O}_{At}) - d_{Vw}(\mathbf{O}_{Vt}) \end{aligned}$$

ただし，δ_x は x が真のとき 1，偽のとき 0 を返す関数である．式 (4) により，全ての $\lambda_w \in \Lambda$ について $\Delta\lambda_{Aw}$ を計算し，その後 λ_{Aw} の値を更新する．この更新サイクルを繰り返すことにより，最適な Λ を推定することができる．この尤度比最大 (Likelihood Ratio Maximization) 基準による方法は，十分に最適化用データが得られる状況では，従来用いられてきた MCE-GPD による方法と比べて，高い性能を得ることができ，実用性・頑健性の点において有利である [6]．

2.3. 尤度平均化基準による最適化手法

前節で述べた尤度比最大化法では，重み推定に際しては多量の適応データが必要である．しかし，実際のアプリケーションでは，リアルタイム・オンラインでの重み最適化が必要不可欠であり，それゆえ

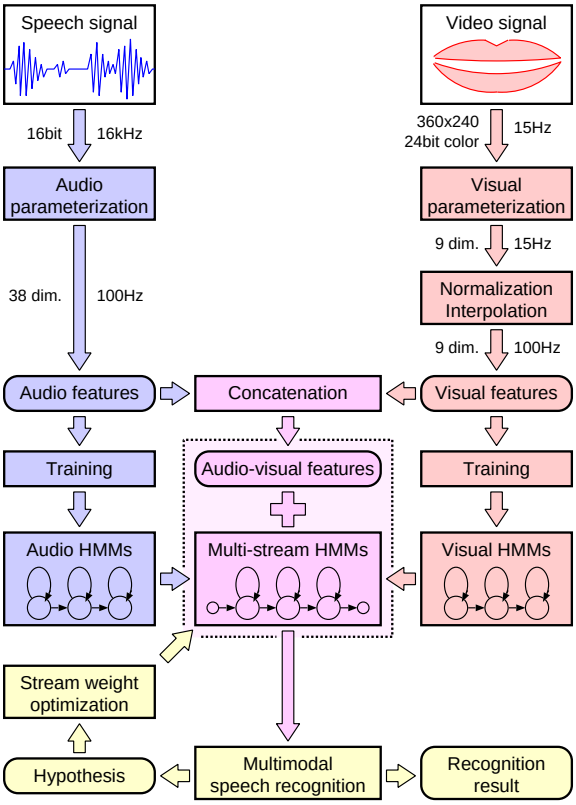

図 1: マルチモーダル音声認識システム

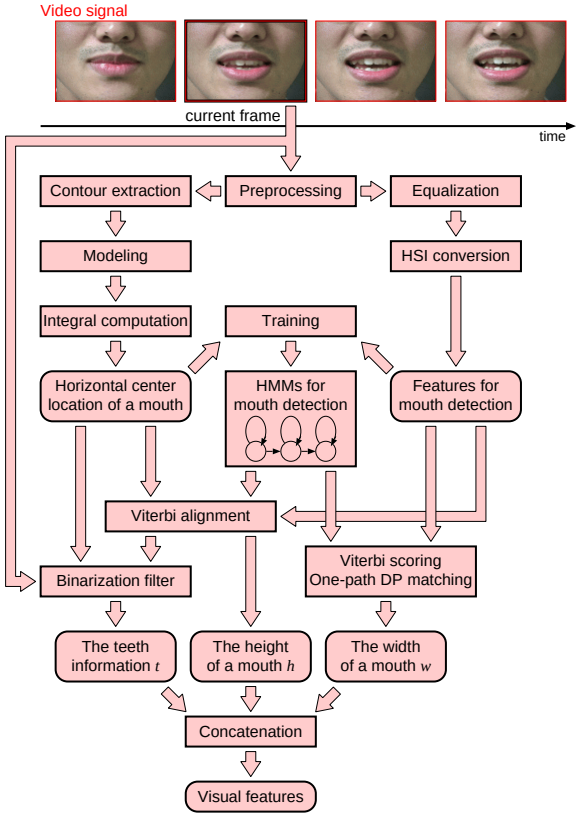

図 2: 画像特徴量抽出

少量の最適化用データでも適用可能なアルゴリズムが望まれる．そこで本論文では，尤度比最大化法をもとに，重み最適化手法の再検討を行った．

まず，尤度比最大化法によって得られたストリーム重みを用い，各モデルが出力する音響–画像対数尤度の評価と解析を行った．その結果，モデルごとの音響–画像対数尤度の平均がほぼ同じになることが判明した．このことから本論文では，尤度比最大化法に代わる簡便な方法として，新たに各モデルの出力尤度の平均が等しくなるように重み係数を推定する手法を提案する．具体的には，次式により単語 r に対する音響ストリーム重み λ_{Ar} を推定する．

$$\lambda_{Ar} = \frac{\dfrac{1}{NT}\displaystyle\sum_{t=1}^{T}\sum_{w\in W} b_{Aw}(\mathbf{O}_{At})}{\dfrac{1}{T}\displaystyle\sum_{t=1}^{T} b_{Ar}(\mathbf{O}_{At})} \qquad (5)$$

ただし $N = |W|$ である．式 (5) において，分母は観測系列 $\mathbf{O}_{At}$ を単語 r のモデルにあてはめたときの対数尤度の平均，分子は全ての単語のモデルから得られる対数尤度の平均である．得られた音響重みは，$0 \leq \lambda_{Ar} \leq 1$ となるよう $\lambda_{A_{MAX}} = \max_w \lambda_{Aw}$ により正規化し，次いで画像ストリーム重みを式 (2) により計算する．この尤度正規化 (Output Likelihood Normalization) 基準による手法は，従来の尤度比最大化法と比べて，繰り返し演算が不要で計算量・演算時間が削減できるという利点がある．

3. マルチモーダル音声認識システム

図 1 に，本研究で用いたマルチモーダル音声認識システムを示す [5].

3.1. 特徴量抽出

音響特徴量には，CMN-MFCC 12 次元とこれらの Δ，ΔΔ 成分，および正規化対数パワーの Δ，ΔΔ 係数の計 38 次元を用いる．次に画像特徴量抽出の流れを図 2 に示す．各フレーム画像から計算した口の幅 w，高さ h，および歯の情報 t の 3 次元と，これらの Δ，ΔΔ 成分の計 9 次元を抽出し，3 次元スプライン関数で時間方向に補間して，画像特徴量とする．そして認識に用いる音響–画像特徴量を，音響特徴量と画像特徴量をフレーム毎に連結することにより生成する．

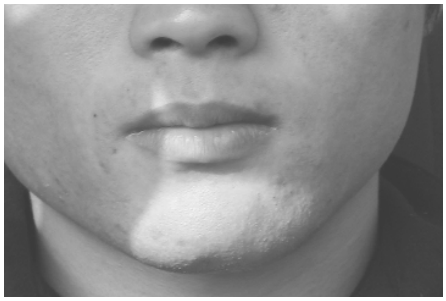

図 3: テストデータ（顔の一部に日光が射している例）

3.2. モデリング

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．HMM は音響と画像それぞれ別に学習する [7]．初期モデル生成・連結学習によって音響 HMM を作成した後，Viterbi アルゴリズムで時間情報つきラベルを生成し，これにより画像 HMM のラベルつき学習を行う．得られた音響 HMM と画像 HMM を融合し，音響–画像マルチストリーム HMM を生成する．

3.3. ストリーム重み最適化・認識

音響–画像特徴量とマルチストリーム HMM により，デコーディングを行い認識仮説を生成する．これを用いて，ストリーム重み係数の最適化を行う．得られた重みをマルチストリーム HMM に反映し，再度デコーディングを行うことで，最終的な認識結果を得ることができる．

4. 認識実験

従来の尤度比最大基準によるストリーム重み最適化と，今回新たに提案した尤度平均化基準による最適化手法の比較を行うため，実環境データによる認識実験を行った．

4.1. データベース

学習データ，テストデータはともに，連続数字読み上げタスクである [8]．学習には音響・画像ともにクリーン環境で収録した男性話者 11 名によるデータを，テストには高速道路走行中の車内で収録した，学習セットには含まれない男性話者 6 名によるデータを使用した．各話者は 2〜6 桁の数字を，学習データでは 250 個，テストデータでは 115 個発声している．テストデータ中の音響雑音としてはエンジン音，風切り音やウィンカー音などが観測され，SNR 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．このテストデータの画像の例を，図 3 に示す．

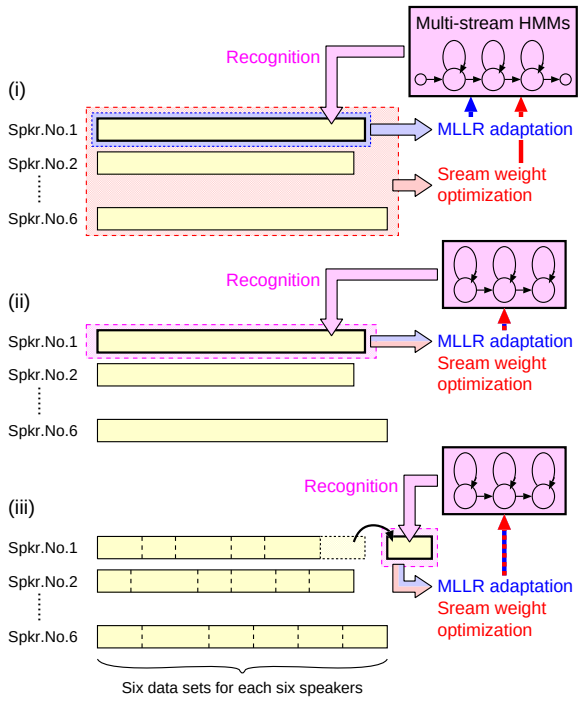

図 4: 実験条件

4.2. 実験条件

尤度比最大基準，尤度平均化基準について，教師なし重み最適化による認識実験を行った．尤度比最大化法における繰り返し演算回数は 50 回とした．また，雑音適応（MLLR[9]）と組み合わせる実験についてもあわせて行った．MLLR を行う場合には，ストリーム重み最適化よりも先に適用し，音響ストリーム中の正規分布の平均と共分散行列を適応化した．テストセットは，各話者ごとにデータを 6 つに分け，合計 36 個のデータセットに分割した．そして，図 4 で示すような 3 種類の条件で，ストリーム重み最適化および MLLR 適応を行った．

(i) 各話者ごとのデータで MLLR を，全テストデータで重み最適化を行い，得られたモデルで各話者ごとにテストデータを認識する

(ii) MLLR，重み最適化ともに各話者ごとのデータを用いて行い，認識も各話者ごとに行う

(iii) MLLR，重み最適化ともに 36 個のデータセットごとに行い，認識も各セットごとに行う

表 1: 各種条件における数字正解精度（MLLR なし）

		尤度比最大化	尤度平均化
音響のみ		62.0%	
音響–画像	全モデル同じ重み	64.2%	
	(i) 最適化	75.6%	76.4%
	(iii) 最適化	59.4%	77.8%

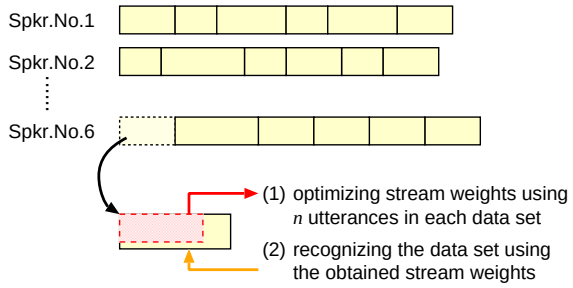

図 5: 重み最適化と用いる発話数による性能変化を調べる実験

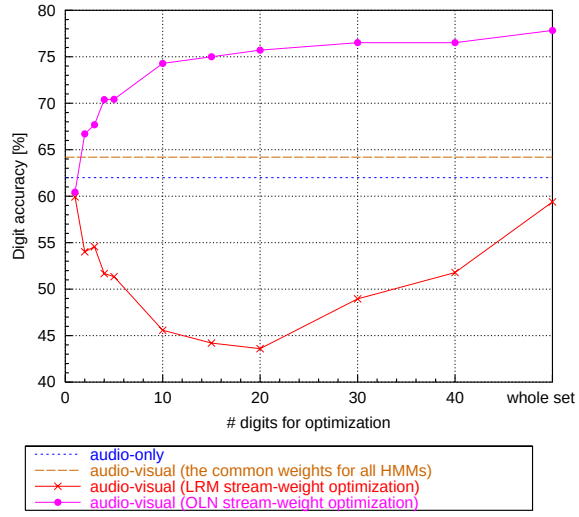

図 6: 重み最適化用データ数の違いによる認識率の変化

また，最適化で使用する認識仮説の生成や，尤度比最大化法における繰り返し演算の初期値に用いるマルチストリーム HMM の初期重みは，全モデル共通に $\lambda_{Aw} = 1$，$\lambda_{Vw} = 0$ とした．

4.3. 実験結果

はじめに，条件 (i) および (iii) における，ストリーム重み最適化のみを行ったときの数字正解精度を示す．表 1 は，音響特徴量のみを用いた場合の認識率（ベースライン），全ての単語モデルに同じ重みをマニュアルで設定した場合の最も高い認識率，および尤度比最大化法と尤度平均化法それぞれの認識率で

表 2: 各種条件における数字正解精度（MLLR あり）

		MLLR のみ	尤度比最大化	尤度平均化
音響–画像	(i)	85.1%	91.1%	90.2%
	(ii)		88.7%	90.4%
	(iii)	78.1%	76.2%	84.5%

ある．全モデル同一重みの場合と (i) を比較すると，尤度比最大化および尤度平均化による自動ストリーム重み最適化法により，それぞれ 11%，12%認識率が改善し，これらの手法の有効性が確かめられた．一方 (iii) の結果から，尤度平均化法は (i) のときよりも高い性能を示し約 16%の改善がみられたのに対し，尤度比最大化法はベースラインよりも性能が劣化していることが判明した．

次に，最適化に用いるデータ数と認識性能との関係について調べた．図 5 に示すように，36 個のデータセットそれぞれについて，はじめの n 個の数字発声のみを用いてストリーム重み最適化を行い，得られた重みを用いて各データセットの音声認識を行った．図 6 に，尤度比最大化法 (LRM)，尤度平均化法 (OLN) それぞれについて，各セット中で重み最適化に用いた数字発声数に対する認識性能の変化を示す．グラフの横軸は数字発声数 n，縦軸は数字正解精度である．横軸の「whole set」は各セット中の全発声（セットにより 47〜126 個の数字発声）を用いた場合の性能を示しており，表 1 の (iii) の結果と同等である．グラフから，尤度比最大化法はデータ数が少ないと性能が著しく低下してしまうのに対し，尤度平均化法では少量のデータでも認識率が改善し，データ量が増えるほど認識性能も向上することが確かめられた．

最後に，MLLR 適応とストリーム重み最適化を併用した場合の，各条件での認識性能を表 2 に示す．表 1 および表 2 より，まず MLLR 適応によって認識性能が向上し，ストリーム重み最適化を行うことでさらに認識率が改善することが示された．条件 (iii) においては，MLLR 適応のみの結果からみて，尤度比最大化法では性能向上がみられなかったが，尤度平均化法ではさらに約 6%正解精度が改善した．

4.4. 考察

以上の結果から，従来の尤度比最大化法では，少量の最適化データでは，適切なストリーム重みを決定することができず認識率が低下してしまうのに対し，本論文で提案する尤度平均化法は，少量データ

でも頑健に重み係数を推定し性能が大きく改善することが確認された．このことから，尤度平均化によるストリーム重み最適化法は，逐次的に入力データセットの雑音状況に応じて重み係数を最適化することにより，認識性能を改善できると考えられる．また図 6 より，例えば尤度平均化法は 10 個の数字発声を用いただけでも，最適化を行わない結果と比較して約 10%認識率が向上した．10 個の数字発声は，約 10 秒の発声に相当し，このことからも本最適化手法はオンラインでのストリーム重み最適化が可能であるといえる．最後に表 2 より，尤度平均化によるストリーム重み最適化と MLLR 適応を用いることで，条件 (iii) でベースラインと比べて約 23%と大幅に数字正解精度が改善し，MLLR のみの結果からも約 6%向上した．以上から，MLLR によって音響モデルの適応を行った場合であっても，尤度平均化法によりストリーム重みを最適化することで，さらに認識精度を向上できることが確かめられた．

5. まとめ

本論文では，マルチストリーム HMM におけるストリーム重みの最適化手法として，新たに尤度平均化基準による手法の提案を行った．車載カメラで収録した実環境データによる認識実験を行ったところ，尤度平均化法は，従来の尤度比最大化法よりも高い性能を示し，特に最適化用データが少量のときに有効に機能することが確認された．さらに MLLR 雑音適応と組み合わせることで，音響のみのベースラインに比べ，約 23%正解精度の改善に成功した．

今後の課題としては，(1) 発話情報をより多く含んだ画像特徴量および特徴量抽出アルゴリズムの計算量削減，(2) 大語彙連続音声認識や情報検索システムなどへのマルチモーダル音声認識の適用，(3) よりよい音響と画像の同期手法と融合アルゴリズムの検討，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] 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] G. Potamianos, J. Luettin and C. Neti, “Hierarchical discriminant features for audio-visual LVCSR,” Proc. International conference on ICASSP 2001, pp.165-168 (2001-5).

[4] 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).

[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音響・画像特徴量の融合法に関する検討,” 2003 年秋季音講論, 3-6-11, pp.123-124 (2003-9).

[6] 田村 哲嗣, 岩野 公司, 古井 貞熙, “尤度比最大基準によるストリーム重み最適化を用いたマルチモーダル音声認識の性能評価,” 2004 年春季音講論, 3-8-1, pp.123-124 (2004-3).

[7] 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).

[8] 田村 哲嗣, 岩野 公司, 古井 貞熙, “実環境におけるマルチモーダル音声認識の評価,” 2002 年春季音講論, 3-5-5, pp.151-152 (2002-3).

[9] 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).