

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

田村 哲嗣[†] 岩野 公司[†] 古井 貞熙[†]

[†] 東京工業大学 情報理工学研究所 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

あらまし 近年、音声認識の頑健性向上の手法のひとつとして、音声情報に加え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、多くの研究が進められている。マルチモーダル音声認識で広く用いられているマルチストリーム HMM では、ストリーム重み係数を自動的に調整することが認識性能向上に必要不可欠である。本研究では、正解（仮説）単語とその他の単語の尤度の差が最大となるよう、尤度比最大基準に基づく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により、教師なし条件で提案法の評価を行ったところ、MLLR 適応と提案手法をあわせて行うことで、音響のみの結果と比べ、約 29% の正解精度の改善、約 76% の誤り率の削減に成功した。

キーワード マルチモーダル音声認識, マルチストリーム HMM, ストリーム重み最適化, 尤度比最大基準, 実環境

Investigation of a stream-weight optimization method for multi-modal speech recognition

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

Abstract Researches on audio-visual multi-modal speech recognition have recently become very active for increasing the robustness of automatic speech recognition (ASR). For multi-stream HMMs that are widely used in multi-modal ASR, it is important to automatically and properly adjust stream weight factors. This paper proposes a stream-weight optimization technique based on a likelihood-ratio maximization criterion. Experiments were conducted using real-world data in an unsupervised manner. Combining the maximum likelihood linear regression (MLLR) adaptation and our optimization method, we achieved a 29% absolute accuracy improvement and a 76% relative error rate reduction compared with the audio-only scheme.

Key words Multi-modal speech recognition, Multi-stream HMM, Stream-weight optimization, Likelihood-ratio maximization criterion, Real-world environments

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、音声認識はユーザフレンドリーなインターフェースとして、いま最も注目されている技術のひとつである。しかし現在の音声認識技術は、雑音が大きい環境の下での認識性能が低く、これが音声認識の実用化における大きな問題となっている。そこで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システムが注目

され、近年研究が進められている [1]~[4]。

多くのマルチモーダル音声認識において、音響と画像の情報を効果的に融合する方法として、マルチストリーム HMM が用いられている。マルチストリーム HMM は、音響情報をモデル化した音響ストリームと画像情報による画像ストリームで構成されており、認識に用いる音響-画像尤度は、ストリーム重みと呼ばれるパラメータにより重み付けされた、音響尤度と画像尤度の和によって計算することができる。このストリーム重みには、HMM 学習時に最尤推定することができないという問題があり、さらに、より良い認識性能を得るためには、音響・

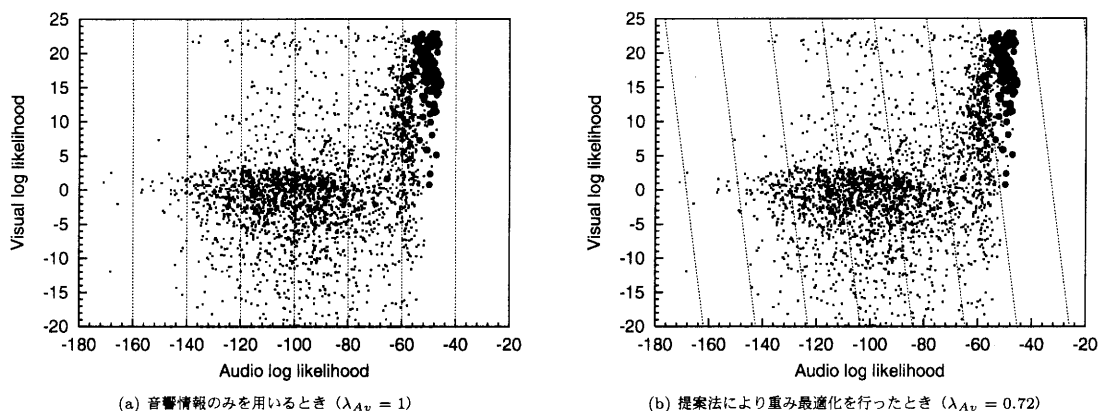


図1 単語 v に対する音響尤度・画像尤度の分布と等高線 (●印…第一仮説のとき、○印…それ以外のとき)

画像それぞれの雑音状況に応じて、認識時にストリーム重みを適切に設定する必要がある。このため、条件・制約を設けた最尤推定や最小分類誤り基準 (MCE) などにより、ストリーム重みを自動的に推定する方法がいくつか提案されている [4]~[6]。しかしこれらの手法では、用いている仮定条件の妥当性や、最適な制御パラメータの設定が困難であるといった問題がある。また実用性を考えた場合、教師なしの条件で重み係数を推定できることが重要となるが、これまでの手法は正解ラベルを用いる教師ありの条件を前提としており、実環境など実用条件での教師なしによる性能評価は行われていない。

そこで本研究では、尤度比最大基準に基づき適応的に重み係数を推定する新たなストリーム重み最適化手法を提案し、走行中の乗用車内で収録した実環境データを用いて教師なし条件で認識実験を行うことにより、手法の性能評価を行う。

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、音声認識時におけるモデルとして、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 を使用している。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 $\mathbf{O}_t$ の観測確率は、対数尤度 $b_w(\mathbf{O}_t)$ を用いて式 (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、以下の制約がある。

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

2.2 尤度比最大基準によるストリーム重み最適化

マルチストリーム HMM におけるストリーム重みに関して、HMM 学習時にはガウス分布の平均・分散や混合重みのような他のモデルパラメータとは異なり、最尤推定することができな

いという問題がある。一方で認識時には、音響・画像の各々のチャンネルの雑音状況や信頼度に応じてストリーム重みを変化させることが、認識性能の向上に有効である。そこで本研究では、尤度比最大基準により、適応的にストリーム重みを自動決定する手法の提案を行う。

いま、デコーダが単語 w_i ($1 \leq i \leq M$) を時刻 $T_{i-1} \leq t < T_i$ (ただし $T_0 = 0$, $T_M = T$, T はデータの総時間長) において出力したとする。すると、任意の $w \in W$ (辞書中の単語の集合、 $|W| = N$) に対して以下の関係が成り立つ。

$$\bar{b}_{w_i}(\mathbf{O}^i) \geq \bar{b}_w(\mathbf{O}^i) \quad (3)$$

上式で、 $\mathbf{O}^i$ は音響-画像観測系列で $\mathbf{O}^i = \{\mathbf{O}_{T_{i-1}}, \dots, \mathbf{O}_{T_i-1}\}$ である。また $\bar{b}_w(\mathbf{O}^i)$ は音響-画像平均対数尤度で、音響観測系列 $\mathbf{O}_A^i$ 、画像観測系列 $\mathbf{O}_V^i$ と、単語 w のモデルによる音響平均対数尤度 $\bar{b}_{Aw}(\mathbf{O}_A^i)$ 、画像平均対数尤度 $\bar{b}_{Vw}(\mathbf{O}_V^i)$ により、式 (1) と同様に次式 (4) で表される。

$$\bar{b}_w(\mathbf{O}^i) = \lambda_{Aw} \bar{b}_{Aw}(\mathbf{O}_A^i) + \lambda_{Vw} \bar{b}_{Vw}(\mathbf{O}_V^i) \quad (4)$$

w_i 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_i の尤度が一番大きくなってしまいうことに起因する。そこで適応データが与えられたとき、正解 (仮説) 単語の対数尤度とその他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により、同じ環境の認識データにおける認識誤りを抑制できると考えられる。教師なしの場合においても、適応に用いるラベルがある程度正しければ、統計的にみて、誤りを抑制するストリーム重みの推定は可能である。以上に基づき、本研究では、ストリーム重み $\Lambda = \{\lambda_{Aw}\}$ を適応的に求める手法を提案する。すなわち、

$$L(\Lambda) = \sum_{i=1}^M \sum_{w \in W} \left\{ \bar{b}_{w_i}(\mathbf{O}^i) - \bar{b}_w(\mathbf{O}^i) \right\}^2 \quad (5)$$

$$\hat{\Lambda} = \arg \max_{\Lambda} L(\Lambda) \quad (6)$$

このとき、任意の単語 $v \in W$ に対し、

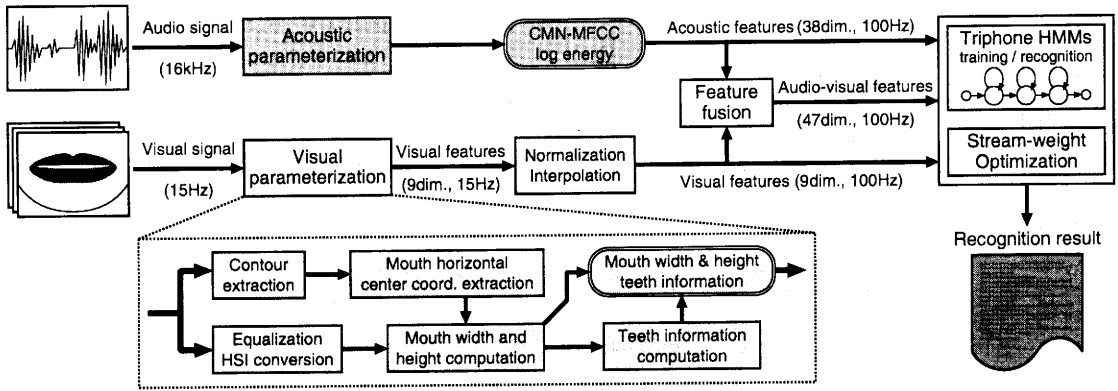


図2 マルチモーダル音声認識システム

$$\frac{\partial L(\Lambda)}{\partial \lambda_{Av}} = 0 \quad (7)$$

が成立するので、この式(7)を解くことにより λ_{Av} の変化分 $\Delta \lambda_{Av}$ を次のように求めることができる。

$$\Delta \lambda_{Av} = \frac{P}{Q}$$

$$P = \sum_{i=1}^M \left[\delta_{w_i=v} \cdot \left\{ N \bar{b}_v(\mathbf{O}^i) - \sum_{w \in W} \bar{b}_w(\mathbf{O}^i) \right\} + \delta_{w_i \neq v} \cdot \left\{ \bar{b}_v(\mathbf{O}^i) - \bar{b}_{w_i}(\mathbf{O}^i) \right\} \right]$$

$$Q = \sum_{i=1}^M \left[\delta_{w_i=v} \cdot N \bar{d}_v(\mathbf{O}^i) + \delta_{w_i \neq v} \cdot \bar{d}_v(\mathbf{O}^i) \right]$$

$$\bar{d}_w(\mathbf{O}^i) = \bar{b}_{Aw}(\mathbf{O}_A^i) - \bar{b}_{Vw}(\mathbf{O}_V^i) \quad (8)$$

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。同様に上式により、全ての $\lambda_{Aw} \in \Lambda$ について $\Delta \lambda_{Aw}$ を求め、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、 $\hat{\Lambda}$ を推定することができる。以上で得られるストリーム重みが、実際に音響-画像尤度を求める際にどのように影響しているのかについて、図1を用いて説明する。これは単語 v に関して、デコーダが第一仮説として v を出力したとき($w_i = v$, ●印)と、それ以外の単語を出力したとき($w_i \neq v$, ○印)の、単語 v のモデルに対する音響対数尤度 $\bar{b}_{Av}(\mathbf{O}_A^i)$ と画像対数尤度 $\bar{b}_{Vv}(\mathbf{O}_V^i)$ の分布をプロットしたものである。図において、横軸は音響対数尤度、縦軸は画像対数尤度で、図中の点線は同じ音響-画像平均対数尤度 $\bar{b}_v(\mathbf{O}^i)$ の点を結んだもの(等高線)である。(a)は音響情報のみを用いる場合($\lambda_{Av} = 1$, $\lambda_{Vv} = 0$)で、 x 軸に射影して音響-画像尤度が計算され、等高線は y 軸に平行になる。一方、(b)はストリーム重み最適化手法により得られた重みを用いた場合($\lambda_{Av} = 0.72$, $\lambda_{Vv} = 0.28$)のものである。提案手法は音響-画像尤度について、●印の尤度がなるべく大きく、○印の尤度がなるべく小さくなるようにストリーム重みを調整する。これは図1中の等高線の傾きを、 λ_{Av} , λ_{Vv} により制御することに相当する。

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-MFCC 12次元, : これらの 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	フレーム周期 : 10ms 抽出特徴量 : 口腔の高さ h と幅 w , : 見かけの歯のpixel数 t , : これらの Δ , $\Delta\Delta$ 成分 特徴量次元数 : 9次元

3. マルチモーダル音声認識システム

図2に、本研究で構築したマルチモーダル音声認識システムを示す。

3.1 音響特徴量

表1に、使用した音響および画像特徴量について示す。音声データは16kHz, 16bitでサンプリングし、毎秒100フレームで12次元のCMN-MFCCとこれらの Δ , $\Delta\Delta$ 成分、および対数パワーの Δ , $\Delta\Delta$ 成分の計38次元のパラメータに変換し音響特徴量とする。

3.2 画像特徴量

口唇付近を撮影した入力動画は毎秒15フレーム、解像度 360×240 の24bitカラーでキャプチャされる。

横方向の口唇中心位置の推定 入力画像に対し輪郭抽出フィルタをかけ、おおその輪郭の抽出を行う。次に入力画像中の各列 i ($0 \leq i < 360$)に対して、 $v(i, y) = v_i(y)$ を最小自乗法により式(9)で近似し、パラメータ A_i , B_i を求める。

$$v_i(y) \simeq \left| A_i(y - y_0)e^{-B_i(y - y_0)^2} \right| \quad (9)$$

ただし $A_i > 0$, $B_i > 0$, y_0 は $v_i(y)$ の重心である。列中に口唇が含まれる場合は上唇と下唇の2箇所輪郭が抽出されるの

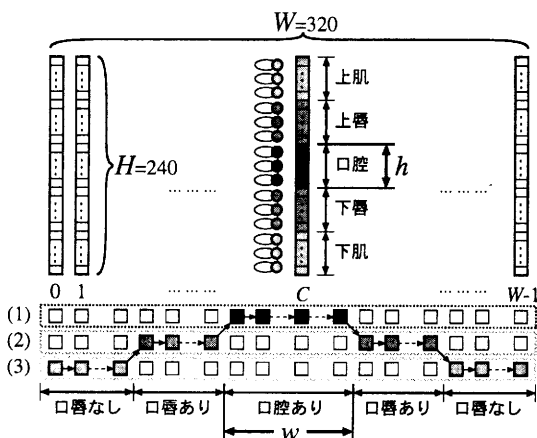


図3 口唇座標推定 ((1): 口腔あり尤度, (2): 口唇あり尤度, (3): 口唇なし尤度)

で、式(9)を用いることで効率的に輪郭情報を近似することができる。また口唇を含むときは、式(9)の積分値は大きくなるという特性がある。そこでこれを口唇を含んでいるもっともらしさの指標 $l(i)$ として、式(10)のように定義する。

$$l(i) = \int_{-\infty}^{\infty} v_i(y) dy = \frac{A_i}{B_i} \quad (10)$$

以上から式(11)により、 $l(i)$ の重心を計算することで、横方向の口唇中心位置 C を推定できる。

$$C = \sum_{i=0}^{W-1} i \times l(i) \quad (11)$$

縦方向の口唇座標(高さ)の推定 入力画像にヒストグラム分布を均一にする平準化フィルタをかけ、次いで各画素のRGB値を人間の知覚に近いHSI(色相, 彩度, 明度)値に変換する。これを各列ごとに上から下にスキャンし、240点の3次元パラメータ系列を生成する。このうち列 C のパラメータ系列を、口唇座標計算用のHMMを用いて強制切り出しし、縦方向の口唇の座標情報を推定する(図3)。使用するHMMは状態数3、混合数4のleft-to-right型で、上側肌、上唇、口腔(歯)、下唇および下側肌の5種類である。口が閉じている場合を考え、口腔HMMを飛ばして上唇HMMから下唇HMMへの直接遷移を可能としている。手作業で付与した座標ラベルありデータによって初期モデルを生成し、続いて座標情報なしデータを用いて連結学習を行い、不特定話者モデルを構築する。座標計算時には、Viterbiアルゴリズムによる強制切り出しで各HMMのアライメントを求める。そして口腔HMMの開始および終了座標から、口の高さ h を算出する。また、求められた口腔部分について、輝度二値化により歯の検出を行い、そのpixel数 t ($h=0$ のときは $t=0$) を算定する。

横方向の口唇座標(幅)の推定 さきに求めたパラメータ系列とHMMにより、各列毎に口唇なし(上肌→下肌)、口唇あり(上肌→上唇→下唇→下肌)、口腔あり(上肌→上唇→口腔→下唇→下

表2 実験条件(使用特徴量・適応の有無)

	使用特徴量	MLLR	重み最適化
(1)	音響のみ	×	×
(2)	音響-画像	×	×
(3)	音響-画像	×	○
(4)	音響-画像	○	×
(5)	音響-画像	○	○

肌)の3種類の尤度を計算する。制約つき(口唇なし→口唇あり→口腔あり→口唇あり→口唇なし、口腔ありはスキップ可)ワンパスDPマッチングにより左列から右列へスコアが最大となるパスを計算し、バックトラックにより境界情報を求め、横方向の口唇座標を推定する。そして口腔ありの開始、終了座標を求め、口の幅 w を算出する。

画像特徴量の抽出 以上で得られた h, w, t に対し、それぞれ $\Delta, \Delta\Delta$ 成分を求め9次元のパラメータを求める。さらに画像系列(動画ファイル)ごとに正規化を行い、音響特徴量とフレームレートを合わせるため3次元スプライン関数により補間し、最終的な画像特徴量とする。

3.3 音響-画像特徴量

音声認識時に用いる音響-画像特徴量は、以上で得られた音響特徴量と画像特徴量をフレーム毎に連結・融合することで生成される。

3.4 音声認識用HMM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。HMMは音響と画像それぞれ別に学習する。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のラベルあり学習を行う。認識時には得られた音響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。

4. 実験条件

4.1 データベース

学習データ、テストデータともに連続数字読み上げデータを用いて実験を行った。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。テストデータの話者は学習データには含まれておらず、また音響チャンネルのSNRはおおよそ10~15dBであった。

4.2 MLLR適応・ストリーム重み最適化

認識実験は、表2に示す5つの条件で行った。(1)は音響特徴量のみで学習・認識を行うもので、ベースラインである。音響-画像特徴量を用いる(2)の場合のストリーム重みは、全てのHMMに同じ音響ストリーム、画像ストリームを設定して認識を行った。(3)~(5)では、MLLR適応[7]と提案したストリーム重み最適化法の、いずれか一方あるいは両方を適用した。

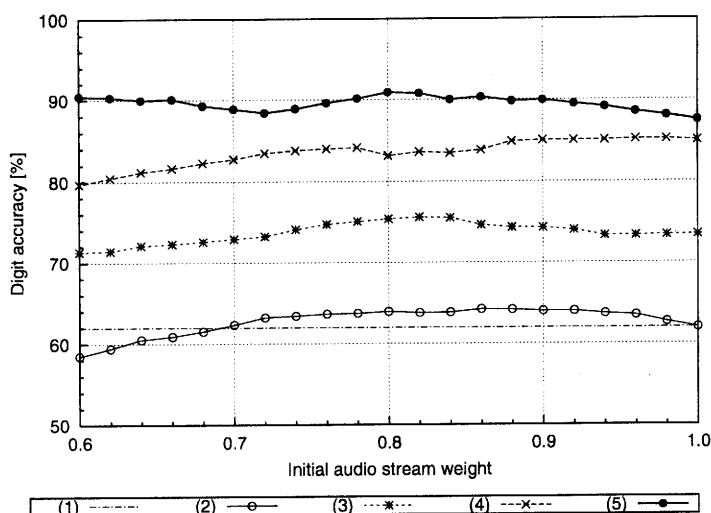


図 4 各種実験条件における認識結果

適応・最適化は教師なしで行い、使用するラベルは(2)と同様に全てのHMMに同じストリーム重みを設定して生成した(以降、このときの音響ストリームを初期音響ストリームとする)。MLLRは音響ストリームのガウス分布の平均・分散のみに適用し、各話者ごとにパッチ適応を行った。ストリーム重み最適化では、テストデータ6名全てのデータを用いて重みの推定を行い、重み最適化計算の繰り返し回数は50回とした。なお(5)のケースでは、MLLR適応を行ってからストリーム重み最適化を行った。

5. 実験結果・考察

図4に、(1)～(5)の実験条件それぞれについて、初期音響ストリーム重みに対する数字正解精度を示す。グラフにおいて、縦軸は数字正解精度、横軸は初期音響ストリーム重みである。また、最も高い認識性能を示したときの数字正解精度を表3に示す。図4および表3から、(1)と(2)の比較により、音響-画像特徴量を用いることで、正解精度において約2%認識性能が改善した。さらにストリーム重みの最適化を行った(3)とベースラインを比べると、約14%正解精度が改善し、誤り率では約36%の削減に成功した。これより、ストリーム重みを最適化することで認識率が大幅に向上することが確かめられた。また(4)と(5)を比べると、約6%正解精度が向上し、約39%誤り率が削減された。MLLR適応の有無に関わらずほぼ同じ誤り率の削減を達成できており、提案手法はさまざまな条件においても有効に機能することが示された。最終的に(1)と(5)の比較から、MLLRとストリーム重み最適化手法を組み合わせることで、約29%の数字正解精度の改善、約76%の誤り率の削減を達成した。

ここで、本研究で提案しているストリーム重み最適化手法の頑健性について調べるため、教師ありで重み最適化を行った場合との比較実験を行った。表4に、教師なしで重み最適化を

表 3 各種実験条件における最高数字正解精度

(1)	(2)	(3)	(4)	(5)
62.0%	64.2%	75.6%	85.2%	91.0%

表 4 教師なし／あり重み最適化による認識率の比較

教師なし (3)	教師あり (3')
75.6%	78.4%

表 5 認識誤りの比較

	削除誤り	置換誤り	挿入誤り
(1)	226	613	211
(3)	270	294	110

行ったとき(3)、および教師ありで最適化したとき(3')の、最も高い認識性能のときの数字正解精度を示す。教師なし(3)で最適化時に参照しているラベルは、全モデル同一重みで得られた認識結果(2)で、その正解精度は約64%であった。それにもかかわらず最終的に得られた認識率は、教師あり(3')のものと比べ、約3%程度しか劣化していなかった。このことから提案手法は、教師なしであってもある程度正しいラベルが得られれば、教師ありの結果に近いところまで認識性能を改善できることがわかり、提案手法の頑健性を確認することができた。

ところで我々は以前、無発声区間の推定に有効な画像特徴量を用い、無音モデルのみストリーム重みを変化させ他のモデルでは音響情報のみを使うという手法で、認識率の改善に成功している[1]。このときの画像特徴量は、口唇の動きの有無程度の情報しか含有していなかったため、音素や数字の同定に関しては効果は限定的であった。一方で、今回用いた画像特徴量は口唇の幅や高さといった情報を含んでおり、ストリーム重み最適化により音素・数字ごとに重み最適化が可能となったことから、マルチモーダル音声認識全体として音素・数字の識別性能が向上していると思われる。このことを確かめるため、認識結果の

分析を行った。表5に、(1)と(3)について、最も認識率が高かったときの脱落誤り、置換誤り、挿入誤りの個数を調べた結果を示す。これより、特に置換誤りについて削減の度合いが大きいことがわかる。置換誤りの減少は、音響特徴のみでは誤認識されていた数字が、マルチモーダル音声認識により正しく認識されるようになったことを示している。このことから、画像特徴量とストリーム重み最適化によって、音素・数字の同定性能が向上したことが確かめられた。

6. ま と め

本研究では、尤度比最大基準に基づき、正解（仮説）単語とそれ以外の単語の尤度比が最大となるよう、教師なし適応で重み係数を自動決定するストリーム重み最適化手法を提案した。口唇の幅・高さや歯の情報を含んだ画像特徴量を用い、実環境で収録したテストデータの認識実験を行ったところ、MLLRと提案手法を組み合わせることで、最大で約29%の正解精度の改善、誤り率では約76%の削減に成功した。また、提案手法の教師なしと教師あり条件との認識率の差は約3%であったことから、提案手法は実環境・教師なしの条件でも十分な頑健性を有することが確認できた。

今後の課題としては、MCE-GPDなど他のストリーム重み推定法との性能比較・評価、マルチモーダル音声認識手法の大語彙音声認識や対話システムといった他のタスクへの適用、音声と画像でそれぞれ独自に認識を行った結果を融合する decision fusion（結果統合）法の検討などが挙げられる。

謝 辞

本研究はNTTドコモ株式会社の援助を受けて行われました。ここに深く感謝いたします。

文 献

- [1] K. Iwano and S. Tamura and S. Furui, "Bimodal speech recognition using lip movement measured by optical-flow analysis," Proc. HSC2001, pp.187-190, 2001.
- [2] G. Potamianos and E. Cosatto and H.P. Graf and D.B. Roe, "Speaker independent audio-visual database for bimodal ASR," Proc. AVSP'97, pp.65-68, 1997.
- [3] C. Miyajima and 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, 2000.
- [4] S. Nakamura and H. Ito and K. Shikano, "Stream weight optimization of speech and lip image sequence for audio-visual speech recognition," Proc. ICSLP2000, vol.3, pp.20-24, 2000.
- [5] J. Hernando, "Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition," Proc. ICASSP'97, vol.2, pp.1267-1270, 1997.

- [6] 宮島, 徳田, 北村, "多次元確率分布 GMM に基づく話者識別モデルにおけるストリーム重みの推定," 春季音講論, 1-3-3, pp.5-6, 2001.
- [7] 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.