論文 / 著書情報
Article / Book Information

論題(和文)	マルチストリームHMMにおける重み係数決定法に関する検討
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 2004年秋季講演論文集, Vol. , No. 3-1-18, pp. 145-146
Citation(English)	, Vol. , No. 3-1-18, pp. 145-146
発行日 / Pub. date	2004, 9

マルチストリーム HMM における重み係数決定法に関する検討 *

◎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，発声時の口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ている．我々はこれまでに，マルチモーダル音声認識における，マルチストリーム HMM のストリーム重み係数の最適化手法を提案しており，実環境データを用いた認識実験で，認識率の改善を確認している [1]．本論文では，これまでの手法をもとに，少量のデータでも適用可能な重み最適化手法の検討を行い，認識実験によりその有効性について検証した．

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，音声認識時においてマルチストリーム HMM を使用している．いま，デコーダが出力した単語列を $w_1, w_2, \cdots, w_M$，単語 w_i $(1 \leq i \leq M)$ は時刻 $T_{i-1} \leq t < T_i$ で生起されたものとし，この時間の音響–画像観測系列を $\mathbf{O}^i$ とおく．このとき，単語 w のモデルに対する音響–画像平均対数尤度 $\bar{b}_w(\mathbf{O}^i)$ は，

$$\bar{b}_w(\mathbf{O}^i) = \lambda_{Aw}\bar{b}_{Aw}({\mathbf{O}_A}^i) + \lambda_{Vw}\bar{b}_{Vw}({\mathbf{O}_V}^i) \qquad (1)$$

と表される．ただし $\bar{b}_{Aw}({\mathbf{O}_A}^i)$，$\bar{b}_{Vw}({\mathbf{O}_V}^i)$ はそれぞれ音響観測系列 ${\mathbf{O}_A}^i$，画像観測系列 ${\mathbf{O}_V}^i$ に対する単語 w のモデルにおける音響，画像平均対数尤度，λ_{Aw}，λ_{Vw} は単語 w を構成する 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．

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

2.2. 尤度比最大基準による重み最適化

ストリーム重みの最適化について，我々は尤度比最大基準による手法を提案している [2]．この手法では，最適化用データについて，正解（仮説）単語とその他の単語の対数尤度の差 $L_P(\Lambda)$ が最大となるよう，次式によりストリーム重み $\Lambda = \{\lambda_{Aw}\}$ を推定する．

$$L_P(\Lambda) = \sum_{i=1}^{M} \sum_{w \in W} \left\{ \bar{b}_{w_i}(\mathbf{O}^i) - \bar{b}_w(\mathbf{O}^i) \right\}^2 \rightarrow \max \quad (3)$$

ここで W は認識に用いる辞書 $(|W| = N)$ で，$w \in W$ である．式 (3) から $\Delta\lambda_{Aw}$ を求める式が得られるので，繰り返し演算を行うことで，ストリーム重みを決定することができる．この尤度比最大による方法は，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，また制御パラメータをもたないことから，実用性・頑健性の点において有利である．

2.3. 尤度平均化基準による重み最適化

前節で述べた尤度比最大基準による手法では，重み推定に多くの最適化用データが必要なため，オンラインでの適用に不向きという問題がある．本論文では，この点を改善するため，ストリーム重み最適化手法の再検討を行った．

まず，ストリーム重み最適化後の各単語のモデルが出力する音響–画像対数尤度について解析を行った．その結果，各モデルの出力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 v に対する音響ストリーム重み λ_{Av} を推定する．

$$\lambda_{Av} = \frac{\dfrac{1}{MN}\displaystyle\sum_{i=1}^{M}\sum_{w \in W} \bar{b}_{Aw}({\mathbf{O}_A}^i)}{\dfrac{1}{M}\displaystyle\sum_{i=1}^{M} \bar{b}_{Av}({\mathbf{O}_A}^i)} \qquad (4)$$

式 (4) において，分母は観測系列 ${\mathbf{O}_A}^i$ $(1 \leq i \leq M)$ を単語 v 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，$0 \leq \lambda_{Av} \leq 1$ となるように絶対値最大の係数で正規化を行う．その後，画像ストリーム重みを $\lambda_{Vv} = 1 - \lambda_{Av}$ により計算する．この手法は前節の方法と比べ，繰り返し演算が不要で計算量・計算時間が削減できるという利点がある．

3. 重み最適化手法の比較・認識実験

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最適化手法との比較を行うため，実環境データを用い，マルチモーダル音声認識システム [2] により認識実験を行った．

3.1. データベース

学習データにはクリーン環境で収録した男性話者 11 名，テストデータには高速道路走行中の車内で収録した 6 名の，数字連続読み上げデータを用いた [3]．各話者は 2〜6 桁の数字列を，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．

3.2. 実験条件

(A) 尤度比最大基準，(B) 尤度平均化基準のそれぞれについて，教師なしの条件で認識実験を行った．重み最適化と認識は，(i) テストデータ全てを使って重み推定しその後認識する，(ii) テストデータを 36 個のセットに分割し，各セットごとに重み推定・認識する，

---

* Investigation of stream-weight optimization for multi-stream HMMs,
by Satoshi Tamura, Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology).

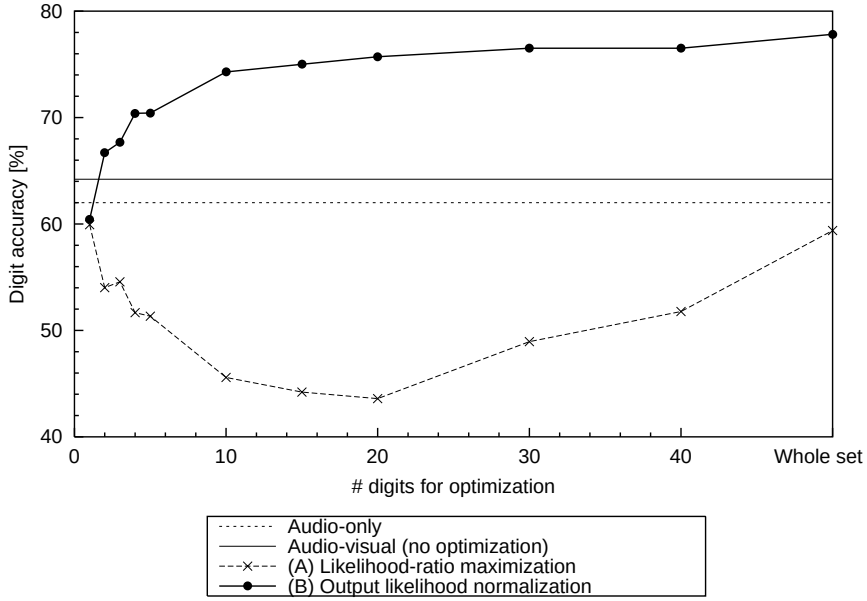

図 1: 重み最適化に用いるデータ数の違いによる認識率の変化 （Whole set: 各セット内の全発声を使用）

表 1: 重み最適化の条件の違いによる認識性能の比較

(A) 尤度比最大法		(B) 提案手法	
(i) 全データ	(ii) セット毎	(i) 全データ	(ii) セット毎
71.6%	59.4%	76.4%	77.8%

の 2 種類の方法で行った．(ii) において最適化に用いる数字の発声数は，1〜5，10，15，20，30，40 個，およびセット内の全発声（セットにより 47〜126 個）の計 11 種類とした．また，最適化で使用する参照ラベル（認識結果）の生成や，尤度比最大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，$\lambda_{Aw} = 1$，$\lambda_{Vw} = 0$ とした．

3.3. 実験結果・考察

表 1 に，(i) および (ii) のセット内全発声を最適化に用いた場合の実験結果を示す．全テストデータを用いる (i) では，(A)，(B) ともに重み最適化を行わない場合（64.2%）よりも，正解精度でそれぞれ約 7%，12%と大幅に性能が改善した．対照的に 36 セットに分割する (ii) では，(A) の認識率が音響のみの結果（62.0%）をも下回ってしまうのに対し，(B) は (i) の結果を上回る性能を示した．

さらに，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てみた．図 1 に，教師なし条件における (A)，(B) の，各セットにおいて重み最適化に用いた数字発声の個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，縦軸は数字正解精度である．グラフから，(A) 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，(B) では少量のデータでも認識率が改善することが確認できた．

以上から，(A) の尤度比最大法は (i) では十分なデータが得られ重み推定できたが，(ii) では重み最適化に必要なデータが不足し，推定が正しく行われずに性能が劣化することが判明した．一方，(B) の提案手法は，(ii) のような少量データの場合でも最適化を行えることが示された．さらに，表 1 において (i) より (ii) の方が認識率が高いことから，(B) の尤度平均化法では，各セットごとに雑音状況に応じて適切にストリーム重みを推定できたものと考えられる．また (B) では最適化用データ 10 個程度で十分に性能改善しており，これは約 10 秒の発声に相当することから，提案手法はオンラインでの最適化が可能であるといえる．

4. まとめ

本論文では，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った．認識実験により，従来法と比べ認識性能が向上し，少量のデータでも最適化可能なことを実証した．

今後の課題としては，発話情報をより多く含んだ画像特徴量の開発，大語彙連続音声認識へのマルチモーダル音声認識の適用などが挙げられる．

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．

参考文献

[1] 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).

[2] 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).

[3] 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).