論文 / 著書情報
Article / Book Information

論題(和文)	ニューラルネットワークを用いたHMMの雑音適応の研究
Title(English)	
著者(和文)	伊藤大輔, 古井貞熙
Authors(English)	Daisuke Ito, SADAOKI FURUI
出典(和文)	日本音響学会 2000年春季講演論文集, Vol. , No. 1-8-7, pp. 13-14
Citation(English)	, Vol. , No. 1-8-7, pp. 13-14
発行日 / Pub. date	2000, 3

1—8—7 ニューラルネットワークを用いたHMMの雑音適応の研究 *

◎伊藤 大輔　古井 貞熙 (東工大)

1. はじめに

大語彙連続音声認識における問題の一つとして，背景に雑音や音楽を含む音声に対する認識性能の劣化が挙げられる．この雑音や音楽に音声認識系を適応させる方法の1つで有効な認識結果を示しているものにHMM合成法 [1, 2] があるが，これは領域変換を含む非線形演算を必要とし計算時間が増大するという問題があった．そこで本稿ではその非線形演算の関係をニューラルネットワーク [3] を用いて学習することで，実存する雑音に適応したHMMを得る手法を提案する．

2. 雑音適応の手法

2.1 学習時に必要となるHMM

HMM合成法では雑音のHMMと音声のHMMから雑音の重畳したHMMを合成する．本研究では合成時に各状態の出力確率分布に対して行う，領域変換及び線形演算を併せもつ関数をニューラルネットワークを用いて学習するため，学習時に雑音HMM，音声HMM，及び予め雑音の情報を含ませたHMM (以下「目標HMM」) が必要となる．

音声HMMは，クリーンな環境の下で学習された多数話者の音声から作成した不特定話者HMMを使用する．雑音HMMは，各 Noisy 音声データから，音声を含まない部分を雑音データとして抽出し，学習して作成する．目標HMMは，上と同一の Noisy 音声データに音声モデルを適応させ作成する．適応アルゴリズムには，少量データへの適応の際用いる変換行列によりパラメータ変換を行う手法を使用する．

2.2 ニューラルネットワークの学習

本研究では各出力確率分布の平均ベクトルに対してのみ雑音適応を行う．よって 音声及び雑音HMMの各平均ベクトル及びパワーに関する情報を入力，目標HMMの平均ベクトルを目標出力とする関数をニューラルネットワークに学習する．

学習手法として誤差逆伝搬法を使用する．また評価尺度としては最も一般的に用いられている平均2乗誤差を採用し，目標HMMの平均ベクトルとネットワークからの出力変数との平均2乗誤差を極小化するよう，1組の目標入出力対が与えられる度に逐次的に更新する．

2.3 雑音への適応

学習したネットワークを用いて，認識すべき音声データのうち音声を含まない部分より学習した雑音HMM，及び音声HMMの平均ベクトルとパワーに関する情報をネットワークに入力することで，対象となる音声の背景にある雑音に適応したHMM (以下「出力HMM」) を得ることができる．

3. 実験

3.1 学習用データ 及び 評価用データ

1996年7月に実際に放送されたニュース音声から，背景に雑音や音楽が乗っている発話や記者レポートなどの発話 (Noisy) のうち男性話者によるものを50文抽出し，うち10文を雑音適応の学習用データ，残り40文を評価用データとして使用した．

今回は2種類の学習セット (各 Train-1, Train-2) に対して雑音適応のための学習を行い，学習セット及び対応する評価セット (各 Test-1, Test-2) について認識実験を行った．学習セットと評価セットのSN比の変動を表1に示す．尚 Train-1 と Train-2 に同一の音声データは含まれていない．

表 1. 学習セットと評価セットの SN 比の変動 [dB]

	数	最小～最大	平均値	標準偏差
Train-1	10	13.78～19.81	17.06	1.97
Test-1	40	8.15～26.34	17.27	5.26
Train-2	10	9.31～23.06	14.11	3.93
Test-2	40	8.15～26.34	18.01	4.66

3.2 音声HMM

今回の実験では，音声HMMとして tree-based clustering により状態共有化を行った不特定話者文脈依存音素HMMを用いた．音響特徴量としては 1～16次の LPC ケプストラムと正規化対数パワー，及びそれらの一次微分の計34次元を使用した．学習用音声データは，ATR 音声データベース B セット，日本音響学会連続音声データベース，および同模擬対話データベースから，男性53名による13,270発話を用いて性別依存モデルを作成した．モデルの総状態数は2,106，各状態のガウス分布の混合数はすべて4である．

3.3 雑音HMM 及び 目標HMMの作成

雑音HMMは，各 Noisy 音声データのうち音声を含まない部分より Baum-Welch アルゴリズムを用いて作成した．また1状態1混合とし音声HMMと同様の音響特徴量を用いた．

目標HMMは MAP-MLLR[4] 及び VFS[5] の手法を併用し，音声HMMを各 Noisy 音声データに適応させることで作成した．尚 MLLR の変換行列数は7つ (無音・子音，各母音毎) とした．また目標HMMの作成方法として，Noisy 音声データに「教師あり適応1回」「同2回」「教師なし適応1回」の3種類の適応方法を用いた．

3.4 学習

以上より得られる各種HMMを用いて，音声HMM及び雑音HMMの1～16次の LPC ケプストラムと正規化対数パワーを入力，目標HMMの1～16次元の LPC ケプストラムを出力とする関数をニューラルネットワークに学習した．また目標HMMの作成方

*A Study on HMM-based Noise Adaptation using Neural Networks
By Daisuke Itoh and Sadaoki Furui (Tokyo Institute of Technology)

法の違いにより，学習セット毎にネットワークへの学習を3種類行った．さらに今回は各状態ごとに異なるネットワークを用いて学習を行った．従ってネットワークの数はモデルの総状態数と同一の2,106．各ネットワークに学習させる目標入出力対の数は40 (=学習データ数10 × 各状態の混合分布数4) である．

3.5 言語モデル

言語モデルの学習に用いたデータは放送ニュース原稿テキスト5年分(1992年7月～1996年5月)．約50万文のデータである．単語出現頻度上位2万語を認識語彙とし，間投詞と読みを考慮した言語モデル[6]を使用した．

3.6 学習セットに対する認識実験結果

まず，ネットワークが少なくとも学習セットに対し良く訓練されているかを調べるため，学習セットに対する認識実験を行った．実験結果を表2に示す．

表 2. 学習セットに対する評価 (単語正解精度 [%])

言語モデル	作成方法	Train-1 *NN* (Target)	Train-2 *NN* (Target)
bigram	(base)	55.1	67.5
	Sup	**64.6** (65.8)	**77.6** (81.2)
	Sup2	**64.4** (65.5)	**74.3** (83.8)
	UnS	**63.2** (61.2)	**76.3** (72.7)
trigram	(base)	56.8	72.7
	Sup	**66.1** (67.5)	**82.0** (84.5)
	Sup2	**67.3** (69.6)	**82.0** (86.5)
	UnS	**64.1** (63.5)	**80.4** (76.7)

ここで base(=baseline) 欄は音声HMMによる単語正解精度を，Sup(=教師あり適応1回)，Sup2(=同2回)，UnS(=教師なし適応1回) 欄は，出力HMM(左側)，及び認識の対象となる音声データを用いた目標HMM(右側括弧内) による単語正解精度をそれぞれ表している．

表より出力HMMでの単語正解精度が，目標HMMのそれにかなり近付いており，少なくとも学習セットに対しては良く訓練されていることが解る．

3.7 評価セットに対する認識実験結果

ネットワークが本研究で意図する関数を適切に学習しているか調べるため，評価データに対する認識実験を行った．結果を表3に示す．

表 3. 評価セットに対する評価 (単語正解精度 [%])

言語モデル	作成方法	Test-1 *NN* (Target)	Test-2 *NN* (Target)
bigram	(base)	65.4	62.2
	Sup	**70.0** (76.7)	**66.7** (72.8)
	Sup2	**70.7** (78.3)	**66.2** (73.7)
	UnS	**68.7** (69.1)	**66.5** (66.2)
trigram	(base)	70.1	65.9
	Sup	**73.8** (79.3)	**70.5** (75.1)
	Sup2	**75.8** (81.3)	**68.4** (77.0)
	UnS	**72.2** (73.4)	**69.1** (70.0)

認識すべきNoisy音声を用いて作成した目標HMMの性能には及ばないものの，音声及び雑音HMMの各平均ベクトルとパワーのみからネットワークを用いて推定したHMM(出力HMM) により，Noisy音声に対する認識性能を向上させることができることがわかる．本実験では，音声HMMと比較して2.5%～5.7% (平均では4.0%) 単語正解精度が向上している．

"教師あり適応で作成した頑健な目標HMM" を学習時に用いることで，評価用データに対して，" 音声データ全てを用いる教師なし適応" と比較すれば，雑音データのみしか必要としない本手法の方が計算量が少なく，しかも同程度或いはそれ以上の単語正解精度が得られている．評価用音声を用いた" 教師なし適応" での認識性能の向上度 (=適応後[%]−baseline[%]) は平均3.8%，これに対し "教師あり適応によるHMM" を目標として学習を行ったときの，出力HMMによる向上度は平均4.3%であった．

さらに目標HMMの作成法として教師なし適応を用いた，学習用音声に対する頑健さの限られたHMMを目標として学習を行った場合でも，認識性能の向上が測られることがわかる．本実験では平均3.2%単語正解精度が向上した．

4. まとめ

HMM合成法を使用する際の計算時間の上昇を回避するため，雑音HMMと音声HMMを合成する過程で，各状態の出力確率分布の平均ベクトルに対して行う非線形演算を，ニューラルネットワークを用いて学習する手法を提案した．これを，背景に雑音・音楽を含むニュース音声での認識に適用したところ，わずかな計算時間で認識性能を向上させることができた．

尚，本研究では音声モデルとして不特定話者HMMを使用し，目標HMM作成方法として変換行列を共有する手法を用いたが，実際には音声HMMから目標HMMを作成する過程で，学習用Noisy音声データの話者に関する情報を含有してしまう．本実験で学習セットとしてTrain-2を使用したときに，" 教師あり適応1回" により目標HMMを作成した場合と比べ，" 同2回" の方が，学習したネットワークより推定したHMM(出力HMM) の認識性能が良くならなかったのは，目標HMMがより話者に対し適応してしまい，ネットワークが雑音を重畳するための関数を適切に学習することができなかったためと思われる．純粋に雑音を重畳する関数を学習させ雑音に対する適応効果を調べるには，その個人差を取り除いて学習することが必要となり，今後検討していきたい．

謝辞

ニュース原稿及び音声データを提供して頂いたNHK放送技術研究所に感謝します．ご助言を頂いたNTTサイバースペース研究所の大附克年氏に感謝します．日頃討論頂く東工大の研究室の方々に感謝します．

参考文献

[1] Martin, 他, Proc.Eurospeech,pp.1031-1034,1993
[2] Gales, 他, Proc.ICASSP,pp.233-236,1992
[3] Hornik, 他, Neural Networks, :pp.359-366,1989
[4] 石井, 他, 秋季音学講論,pp.119-120,1996
[5] 大倉, 他, 信学論,Vol.J76-D-II pp.2468-2476,1993
[6] 桜井, 他, 春季音学講論,pp.57-58,1999