TR-I-0101

話者適応化における写像方法の比較

A Comparative Study of Spectral Mapping Method on Speaker Adaptation

中村　哲，鹿野清宏
Satoshi NAKAMURA, Kiyohiro SHIKANO

1989. 8

概要

近年、学習アルゴリズムの確立によりニューラルネットワークの研究が進み音声の分野にも適用され始めた。本稿では、教師つき話者適応化の枠組みの中で、ベクトル量子化話者適応アルゴリズムにおけるファジィ写像、線形写像、ニューラルネットを用いた非線形写像の写像精度の比較評価を行う。

ATR 自動翻訳電話研究所
ATR Interpreting Telephony Research Laboratories

目次

1. . まえがき

これまでに、話者適応をフレーム毎のスペクトルの写像としてとらえるベクトル量子化話者適応アルゴリズムにより話者適応化が良好にできることを示した[1,2,3,4]。一方、近年、学習アルゴリズムの確立によりニューラルネットワークの研究が進み音声の分野にも適用され始めた。話者適応化においても文献(5,6)にニューラルネット適用の試みがみられる。本稿では、教師つき話者適応化の枠組みの中で、ベクトル量子化話者適応アルゴリズムにおけるファジィ写像、線形写像、ニューラルネットを用いた非線形写像の写像精度の比較評価を行う。

2. ベクトル量子化話者適応アルゴリズム[1,2,3,4]

ベクトル量子化話者適応アルゴリズムは、ベクトル量子化により各話者の音声の特徴空間を離散的に表現し、話者間で離散表現された量子点の対応付けを学習によって求め写像を行うことにより話者適応化を行うものである。さらにファジィベクトル量子を導入することにより、ベクトル量子化のコードブックサイズを増加させずに入力音声を精度よく離散表現できること、ファジィ級関数を確率と見なすことにより量子点の対応付けを表すヒストグラムの推定精度の改善ができること、離散点の対応に基づき入力音声を連続的に標準話者の空間に写像できることなどの改善を行った。このアルゴリズムを次に示す。

2-1) 未知話者の発声した学習用音声を用いて未知話者のコードブックを作る。

2-2) 未知話者の発声した学習用音声を2-1)で作成したコードブックを用いてファジィベクトル量子化する。

2-3) 予めファジィベクトル量子化されている学習用単語と同一の未知話者の単語とのDTWを行う。

2-4) DTWの最適パスに沿って、コードベクトルの対応付けをファジィ級関数を確率と見なしてヒストグラムを積算する。

2-5) 全学習単語についてヒストグラムを積算したのち、変換コードブックを作成する。

2-6) 全学習単語に対する歪を計算し、歪が十分小さくなったら終了する。歪が大きいときは、変換コードブックを未知話者のコードブックとして2－2)から繰り返す。

未知話者の音声が入力されると次のアルゴリズムに従って、話者適応化が行われる。

2-7) 未知話者の音声を未知話者のコードブックを用いてファジィベクトル量子化する。

2-8) ファジィベクトル量子化の復号化の過程に於けるコードベクトルを未知話者のものから変換コードブックのコードベクトルへ入れ換え話者適応化を行う。この写像は、ファジィ級関数を保存したままのファジィ写像になっている。

3. 線形写像アルゴリズム

未知話者の入力音声に対し、未知話者のコードブックのk-近傍コードベクトルを選び、これらが区分線形空間を張ると仮定する。次に、この部分空間が対応する変換コードブックのコードベクトルが張る標準話者の部分空間に基底変換により線形に写像されるものとする。線形写像アルゴリズムとして次の二種類を検討する。

(3-a) ベクトル量子化話者適応アルゴリズムの2－7)及び2－8)のステップを線形写像に基づいて行う。

(3-b) 未知話者のコードブックから変換コードブックのコードベクトルへ全空間の線形写像を求めて写像する。

(3-a)は、ベクトル量子化話者適応アルゴリズムにおけるファジィ写像を線形写像に入れ換えたものである。アルゴリズムを次に示す。

3-1) 未知話者の音声からk-近傍となる未知話者のコードベクトルを選択する。

3-2) k-近傍の未知話者のコードベクトルと対応する変換コードブックのコードベクトルを各部分空間の基底とみなして擬似逆行列をにより線形写像を求める。

3-3) 未知話者の音声の入力ベクトルに対し線形写像を施し写像ベクトルを求める。

一方、(3-b)のアルゴリズムは(3-a)のアルゴリズムのおける近傍数をコードブックのベクトル全部を用いた場合に相当している。

4. ニューラルネットによる非線形写像アルゴリズム

本稿で用いたニューラルネットは層状のフィードフォワード型ネットワークである。ニューラルネットワークは、隠れ層を持ち各ユニットの出力に非線形のシグモイド関数がかかるので非線形の写像が行える。ここでは、次の2種類のアルゴリズムを考える。

(4-a) ベクトル量子化話者適応化アルゴリズムにおける未知話者のコードブックと変換コードブックを入力信号と教師信号としてネットワークを学習する。これは、線形写像の場合の(3-b)のアルゴリズムを非線形化したものに相当する。

(4-b) 話者適応化の全体の枠組みをニューラルネットワークを用いるように変更する。つまり、ベクトル量子化話者適応化におけるDTWの最適パス上で対応する未知話者と標準話者の特徴ベクトルを入力信号と教師信号としてニューラルネットワークの学習を行う。

(4-b)のアルゴリズムを次に示す[5]。

4-1) 未知話者の発声した学習用音声と予め用意されている標準話者の同一単語を用いてDTWを行う。

4-2) DTWの最適パスに沿って、未知話者の音声の特徴ベクトルと標準話者の特徴ベクトルの対応付けを求める。

4-3) 対応付けされた未知話者の音声の特徴ベクトルと標準話者の特徴ベクトルを入力信号と教師信号としてニューラルネットワークの学習を行う。

4-4) 未知話者の入力音声に対しその特徴ベクトルをニューラルネットワークにより標準話者の空間に写像し話者適応化を行う。

5. 実験

各方法の写像精度の比較評価のため、未知話者の特徴ベクトルをフレーム毎に標準話者へ写像し、標準話者の実際の特徴ベクトルとDTWを行い時間整合を行ったケプストラムのユークリッド距離を話者間のスペクトル歪として評価尺度に用いる。分析条件は、12kHzサンプリング、分析窓長21.3msec、フレーム周期3msecで16次のLPCケプストラムを用いる。実験に用いたデータベースは、音韻バランス216単語で、この前半100単語を話者適応の学習用に用い、次の100単語を話者適応化の写像精度の評価用に用いる。話者は未知話者が男性1名、標準話者が男女各1名の計3名である。

ベクトル量子化話者適応アルゴリズムは、ファジネス1.6、k=6でベクトル量子化のコードブックサイズは、256とする。線形写像アルゴリズムでは、近傍数は、予備実験の結果最適であったk=3と全空間の写像の場合のk=256を用いる。また、ニューラルネットワークによる非線形写像アルゴリズムでは、16個のユニットからなる入出力層と、いずれも50個のユニットからなる隠れ層よりなる3層から5層の3種類の構成とした。実験の結果を表1に示す。表1より次の事が明かとなった。

- 線形写像アルゴリズム(3-a)は、ベクトル量子化話者適応アルゴリズムより1.6%歪が大きかった。これは、(3-a)のアルゴリズムでは、入力ベクトルから遠いコードベクトルに大きな係数が割り当てられることがありうるが、ベクトル量子化話者適応アルゴリズムでは、ファジィベクトル量子化が入力ベクトルからの距離を考慮して係数を決定するため写像における線形性が比較的うまく保存できることが原因と考えられる。
- 線形写像アルゴリズム(3−b)は著しく大きい話者間歪を示している。ランク落ちなどの問題があり表1の数値の信頼性は低いが、話者間の特徴ベクトルの写像を全特徴空間の線形写像で表わすことが困難であることを示していると考えられる。。
- ニューラルネットワークによる非線形写像(4-a)は、5層構成が最も良かったがベクトル量子化話者適応アルゴリズムより6%悪い結果であった。しかしながら、全空間の写像が可能になり非線形の導入の効果が大きいことがわかる。
- ニューラルネットワークによる非線形写像アルゴリズム(4−b)は、4層構成の場合に最も高い認識率を示したが、ベクトル量子化話者適応アルゴリズムより5.5%精度が低かった。

また、実際のスペクトログラムでは、線形写像による方法では時間方向の不連続が、ニューラルネットワークによる方法では細部の表現能力がベクトル量子化話者適応化より若干劣ることが観察された。

6. まとめ

本稿では、教師つき話者適応化の枠組みの中で、ベクトル量子化話者適応アルゴリズムにおけるファジィ写像、線形写像、ニューラルネットを用いた非線形写像の写像精度の比較を行った。この結果、線形写像では近傍数を3として区分線形部分空間の写像を行なうときベクトル量子化話者適応化アルゴリズムより1.6%低い写像精度となる、ニューラルネットよる話者適応化ではいずれの場合も5.5%から6%程度ベクトル量子化話者適応化アルゴリズムより低い写像精度となることが示された。

7. 謝辞

日頃ご指導頂く樽松社長、御討論頂いた音声情報処理研究室の皆様に感謝致します。

8, 文献

(1) K.Shikano,et al.,"Speaker Adaptation Through Vector Quantization" ICASSP86

(2) 中村、鹿野,"ベクトル量子化を用いたスペクトログラムの正規化"信学技報SP87-17(1987-06)

(3) 中村、鹿野,"ファジィベクトル量子化を用いたスペクトログラムの正規化"信学技報SP87-123(1988-02)

(4) 中村、鹿野,"ファジィベクトル量子化に基づく話者適応化のHMM音素認識による評価"音講論集2-P-20,1988

(5) 磯、麻生川、吉田、渡辺,"ニューラルネットワークによる話者適応化"音講論集1-6-16,1989

(6) C.Montacie,K.Choukri,G.Chollet,"Speech Recognition using Temporarl Decomposition and Multi-Layer Feed-Foward Automata" ICASSP 89 S8.6

Table 1. Intra-Speaker Distortion

	Fuzzy Mapping	Linear Mapping		Nonlinear Mapping with Neural Network					
				(4-a)			(4-b)		
		(3-a) k=3	(3-b) k=256	3-layer	4-layer	5-layer	3-layer	4-layer	5-layer
Male1→ Male2	0.288	0.294	(78.7)	0.349	0.335	0.310	0.311	0.301	0.309
Male1→ Female1	0.305	0.309	(65.3)	0.354	0.348	0.319	0.354	0.325	0.332
Average	0.296	0.301	(72.0)	0.351	0.342	0.314	0.332	0.313	0.320