論文 / 著書情報
Article / Book Information

論題(和文)	MLLR変換行列を特徴量として用いた年齢推定
Title(English)	
著者(和文)	和田俊也, 篠崎隆宏, 古井貞熙
Authors(English)	Toshiya Wada, Takahiro Shinozaki, SADAOKI FURUI
出典(和文)	日本音響学会2010年春季講演論文集, , No. 2-6-13, pp. 83-84
Citation(English)	, , No. 2-6-13, pp. 83-84
発行日 / Pub. date	2010, 3

MLLR変換行列を特徴量として用いた年齢推定*

☆和田俊也，篠崎隆宏，古井貞熙（東工大）

1　はじめに

一般に、タバコやアルコール等を購入するときに成人判定が必要である。またマーケティングリサーチやアンケートをとるときに相手の年齢層のデータが欲しい場合がある。このように様々なところで年齢を確認する場面があり、年齢を推定する技術は重要になっている。従来音声を用いた年齢推定においては、MFCC を特徴量として GMM 判別器を作り尤度比較を行う方法[1]が用いられてきた。

他方話者認識の分野では、GMM 判別器に代えて話者性をより安定に表すものとして MLLR 行列を特徴量として用いる手法が提案され、有効性が示されている[2]。本研究ではこの MLLR 変換行列を特徴量として用いる手法を年齢推定へ応用する手法を提案する。日本語話し言葉コーパス(CSJ)を用いた実験により、提案手法を用いることで従来の GMM 判別器を用いる手法よりも高い認識率が得られることを示す。

2　提案手法

提案手法では、まず MLLR 適応の際に必要となる初期モデルを用意する。初期モデルとしては、訓練話者のばらつきを取り除くため話者正規化学習(SAT)を行った SAT モデルを用いた。次いで初期モデルから各話者の特徴量に対して MLLR 適応を行う。適応の際に推定された MLLR 変換行列を特徴量として用い、SVM 判別器で学習し、推定に用いる。カーネル関数は線形カーネルを用いた。以下で話者正規化学習および MLLR 特徴量について詳しく述べる。

2.1 話者正規化学習

本研究では MLLR 適応の初期モデルとして GMM および HMM を用いた。訓練話者のスペクトラム分布にばらつきがある状態で不特定話者モデルを学習すると、識別能力が低い広がったモデルが作られてしまう。そのため訓練話者のばらつきを抑えたモデルを作成するために、SAT を行った[3]。

SAT は GMM を用いて MLLR 特徴量を抽出する場合は GMM を、HMM を用いて抽出する場合は HMM を用いて行った。SAT の初期モデルとして用いた GMM は、全ての訓練データを 1 つにまとめて学習したものである。HTK [4]を用いて、まず訓練データから初期モデルを作り、そのモデルの混合数を 2 倍にした後再学習するプロセスを繰り返してモデルを作成した。HMM は、日本語話し言葉コーパス(CSJ)の学会講演音声 254 時間より EM 学習した 3000 状態状態共有混合ガウス分布トライフォン音響モデルを利用した。GMM および HMM で用いた音声の特徴量は MFCC12 次元とパワー、およびそれらのΔとΔΔの計 39 次元である。

SAT のための変換としては CMLLR を用いた。まず GMM, HMM 各々の初期モデルを用いて、各訓練話者に対して CMLLR 変換行列を求める。そして各話者の特徴量にその逆変換を作用させると話者性が取り除かれた特徴量が得られる。その話者性を取り除いた訓練データを用いて、GMM の場合は始めからモデルを学習し、HMM の場合は元の不特定話者 HMM を再び EM 学習しガウス分布の平均値を更新する。この操作を数回繰り返し作成した SAT モデルを MLLR 特徴量抽出のための初期モデルとして用いた。

2.2 MLLR 特徴量

MLLR 適応では、式(1)によりアフィン変換 (A, b) を用いてガウス分布の平均値を不特定話者のもの (μ) から特定話者のもの (μ') へ尤度が最大になるように変換する。

$$\mu' = A\mu + b \qquad (1)$$

上記の MLLR 変換行列 (A, b) の要素を 1 つの特徴量ベクトルに並べて、SVM でモデル化し、推定を行う。元の音響特徴量が 39 次元ならば、A は 39×39 行列、b は 1×39 行列であり、SVM の特徴量の次元は 39×39 + 1×39 (=

* Age estimation using MLLR transform-based features. by Toshiya Wada, Takahiro Shinozaki, Sadaoki Furui (Tokyo Institute of Technology)

1560)となるが、行列 A をブロック対角行列としてブロックサイズを変化させることによって、次元数を調節することができる。例えばブロックサイズを 13 とすると、A には 13×13 のブロックが対角線上に3ブロック並ぶので、A の要素数は 13×13×3 となり、特徴量の次元は 13×13×3 + 1×39(= 546)となる。

3　実験条件

学習および評価は日本語話し言葉コーパス(CSJ)のデータを用いた。男女計 300 人を 60 人ずつの集合に 5 分割し、交差確認法で実験を行った。各話者が(1)18~29 歳、(2)30~44 歳、(3)45~69 歳のどの年齢層に属するか推定を行う。訓練話者(240 人)は 1 人当たり平均 3 秒の発話データを 100 個(計 300 秒)、試験話者(60 人)は 10 個(計 30 秒)利用しており、それらの訓練・試験データから 1 人につき MLLR 変換行列を 1 つ推定している。MLLR 変換行列のブロックサイズは予備実験により 13 と定めた。

また比較実験として、年齢層毎に GMM を学習し推定する実験を行った。この GMM は、年齢層毎にクラス分けされていることを除いて SAT に用いた初期 GMM と同様に学習した。1 人当たり 10 発話あるので、それぞれの発話で年齢層を推定し、合計時間の長い年齢層をその人の年齢層とした。

4　実験結果

Fig.1 は従来手法として GMM 判別器を用いた場合、提案手法で MLLR 初期モデルとして GMM を用いた場合、および HMM を用いた場合の年齢層認識率である。

GMM 判別器を用いた実験では、GMM の混合数を 4, 8, 16, … , 4096 と変化させたところ、混合数 2048 の場合に最大の認識率 55.3%が得られた。Fig.1 において MFCC(GMM)として示したのはこの値である。

GMM を初期モデルとして MLLR 変換行列を推定する実験では、混合数 128 で SAT を 2 回行った場合に最大の認識率 59.7%が得られた。この値は Fig.1 において GMM-MLLR (SVM) として示した。

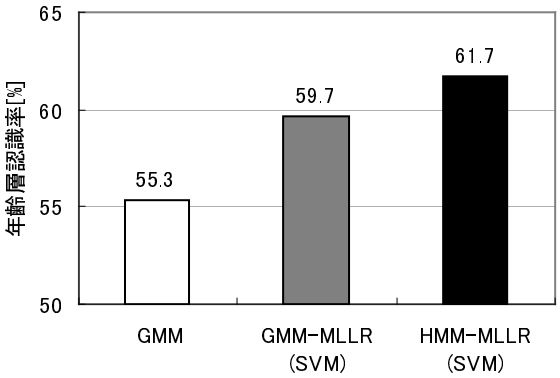

Fig.1 GMM 判別器および MLLR を特徴量として SVM 判別器を用いた年齢層認識率

HMM を初期モデルとして MLLR 変換行列を推定する実験では、混合数 32 で SAT を 1 回行った場合に最大の認識率 61.7%が得られた。Fig.1 における HMM-MLLR (SVM)の値はこの値である。初期モデルとして HMM を用いた方が、GMM を用いた場合より高い精度が得られた。また GMM 判別器を用いた従来手法と比べると認識率が 6.4%増加した。

5　まとめ

話者認識の分野で提案された、MLLR 変換行列の要素を特徴量として SVM をモデル化し推定する手法を、年齢推定に応用した手法を提案した。CSJ を用いた評価実験によって、提案手法により従来のケプストラムを特徴量として GMM 判別器を用いるシステムよりも、高い認識性能が得られることを示した。

参考文献

[1] 西村竜一 他, “安心ウェブの実現に向けた 大人・子ども発話のネット収集実験”, IPSJ SIG Technical Report, 2009.

[2] Andreas Stolcke et al., “MLLR Transforms as Features in Speaker Recognition”, INTERSPEECH 2005, 2425-2428.

[3] Taros Anastasakos et al., “A Compact Model for Speaker-Adaptive Training”, ICSLP 96, 1137-1140 vol.2.

[4] Steve Young et al., The HTK Book, Cambridge University Engineering Department, 2006.