基于 Feature Warping 和 ICA 的发音方式鲁棒的说话人确认

陈胜，徐明星

（普适计算教育部重点实验室，清华信息科学与技术国家实验室（筹）

清华大学计算机科学与技术系，北京 100084）

摘　要：在文本无关的说话人确认中，如果训练语音和测试语音的发音方式不匹配，会导致说话人确认系统的性能下降。针对此问题，本文在特征域上，将 Feature Warping 和 ICA 相结合，对声学特征进行变换补偿，提高了系统对发音方式变化的鲁棒性。首先通过 ICA 将训练语音和测试语音的特征向量映射到各维独立的向量空间，进而通过 Feature Warping，将各维的特征分布映射成为高斯分布，从而消除因发音方式不同而导致的失配影响，最后将得到的各维独立的服从标准高斯分布的特征矢量作为 GMM-UBM 系统的输入，完成发音方式变化的说话人确认。实验结果表明，该方法与使用传统方式的基线系统相比，系统等错误率(EER)下降约 25.48%。

关键词：Feature Warping；ICA；说话人确认；发音方式鲁棒

中图分类号：　TN912.34

Feature Warping and ICA based Speaking-style independent Speaker Verification

CHEN Sheng, XU Mingxing

(Key Laboratory of Pervasive Computing, Ministry of Education

Tsinghua National Laboratory for Information Science and Technology（TNList）

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: Performance of an actual speaker recognition system degrades considerably due to mismatches between training and testing data in speaking style. For better decision-making in a text-independent speaker verification system, a method by combining ICA and Feature Warping in feature compensation field is proposed to eliminate the impact of mismatching speaking style. Firstly, the feature vectors are linearly transformed by ICA to ensure that the components of the feature vectors are independent . Then, the transformed features are nonlinearly mapped into the standard normal distribution by Feature Warping. Finally, the compensated feature vectors are used in the GMM-UBM system to complete the speaker verification process. Compared to the baseline system of the traditional GMM-UBM system, around 25.48% relative improvement in equal error rate (EER) is obtained by the combination of ICA and Feature Warping in the speaking style variation database.

Key words: Feature Warping; ICA; speaker verification; speaking-style independent

说话人确认是说话人识别的一个重要分支，通过语音中所包含的说话人特定信息，来证实说话人的身份，给出接受或者拒绝的结果。

说话人确认系统的性能主要受到说话人无关的因素和说话人相关的两类因素的影响，说话人无关的因素包括训练语音和测试语音在信道上的不匹配、环境噪音的干扰等。说话人相关的因素包括说话人的表达方式、语速、音量、情感状态、身体状况等，这些因素都可以统一为发音方式的变化。

说话人确认的鲁棒性研究主要集中于信道不匹配和噪音干扰的影响，针对鲁棒的说话人确认这一问题，研究人员在特征域方面来提升系统性能。Pelecanos 提出的 Feature Warping[1]方法，通过将特征向量的的分布映射到一个统一个概率分布中，来解决训练语音和测试语音因信道不匹配所带来的系统性能下降的问题。Bing Xiang 提出的 Short-time Gaussianization[2]方法，通过训练 CGM 模型，将特征矢量分离为各维独立的分量，再通过一个非线性变化使得特征向量的各维服从标准正态分布，使得说话人系统更为鲁棒。Kenny 提出的 Joint Factor Analysis[3]则同时对说话人相关因素和说话人无关因素进行建模，来提高说话人确认系统的鲁棒性。

上述方法的提出主要针对的是说话人确认中的训练语音和测试语音的信道不匹配所带来的问题，这与本文所要研究的训练语音和测试语音的发音方式不匹配的问题有相同的本质属性。本文在这些方法的基础上，针对训练语音和测试语音在发音方式的不匹配会导致说话人确认系统性能的下降这一问题，提出将针对跨信道问题的 Feature Warping 和在盲信号分离领域中的 ICA 结合，首先用 ICA 实现原始特征各维的统计独立分离，然后经过 Feature Warping 进行特征分布映射，通过在特征域上对特征向量进行补偿，来提高说话人确认系统的性能。

本文篇章组织如下：首先介绍 Feature Warping 和 ICA，然后提出基于 Feature Warping 和 ICA 的 GMM-UBM 说话人识别，并通过多组实验进行对比分析，最后是结论和下一步工作展望。

1 Feature Warping

Feature Warping 是一种针对训练语音和测试语音的信道不匹配的问题而提出的解决方法。通过将训练语音和测试语音的特征向量的分布映射到一个统一的概率分布中，来减少因训练语音和测试语音信道的不匹配而导致的说话人确认系统系统性能下降的影响。

可以把 Feature Warping 视为一种非线性变换 T，将原始特征向量 X 变换为弯折后的向量 X'

$$X'=T(X)$$

Feature Warping 算法如下：

1) 对输入语音提取声学特征，K 维特征矢量数据按维分为 K 条数据流，各数据流按照以下步骤同时处理，直至各维数据流全部处理完成。
2) 在各数据流的开始处加上大小为 N 的滑动窗口，窗中含有 N 个相邻帧的同一维数据。
3) 将滑动窗口内的数据从小到大排序，确定各数据的序号值。
4) 假设滑动窗口中间位置的数据数值为 x，其排序后的名次为 r，则窗口中间位置处的原始值 x 根据以下公式变换为 x'

$$(r-1/2)/N=\int_{-\infty}^{x'} f(z)dz$$

5) 若数据流未结束，则滑动窗口向右移动一帧，进入第 4 步，否则，待各维数据流都处理完后，算法结束。

在第 4 步中，一般取 $f(z)=\frac{1}{2\pi}\exp(-\frac{z^2}{2})$，即标准正态分布密度函数。

信道不匹配的训练语音和测试语音通过以上 Feature Warping 处理，变换后特征矢量的每一维都服从标准正态分布，这样就减少了因信道不匹配所带来的消极影响。

2 独立成分分析（ICA）

ICA(Independent Component Analysis) [4]是从盲信号分离（Blind Source Separation,BSS)发展起来的一种信号处理方法，其目的是将混合信号分解为相互独立的成分。

标准 ICA 的数学模型为

$$X=A*S$$

其中 A 为 N*N 的混合矩阵，$X=[x_1,x_2,\cdots,x_N]$ 是 N 维观察向量，$S=[s_1,s_2,\cdots,s_N]$ 是相互统计独立的 N 维原始信号。

ICA 的目的就是以统计独立为目标，寻找分离矩阵 $W=[w_1,w_2,\cdots,w_N]$，通过如下计算

$$U=W*X$$

使得 $U=[u_1,u_2,\cdots,u_N]$ 的各分量之间是统计独立的，可以通过基于负熵最大化估计的判别准则和固定点快速迭代算法[5]计算分离矩阵 W。

3 基于 Feature Warping 和 ICA 的 GMM-UBM 说话人确认

3.1 GMM-UBM 结构的说话人确认

高斯混合模型-通用背景模型（GMM-UBM[6]）是说话人确认系统中最为常用的一种模型。GMM（高斯混合模型）是一种通用的概率模型，通过多个高斯分布的概率密度函数的组合来刻画特征矢量在概率空间中的分布。该模型可以表示为

$$p(\vec{x}\mid\lambda)=\sum_{i=1}^{M} w_i N_i(\vec{x})$$

其中 w_i 为混合权重，且满足 $\sum_{i=1}^{M} w_i=1$，N_i 是构成混合高斯密度函数的高斯分量，M 是高斯混合模型的混合数。

UBM（全局背景模型）本质上是一个高斯混合模型 GMM,是用大量语音训练得到的被所有说话人共享的背景说话人模型。目标说话人的建立有两种方式，一种是通过目标说话人的语音数据训练得到，但是通常需要较多的目标说话人语音数据。另一种就是自适应的方式，先训练得到 UBM，在 UBM 的基础上用少量的目标说话人语音，通过 MAP 自适应得到目标说话人模型。

GMM-UBM 模型打分时，采用似然比评分的方式，即用目标说话人模型（GMM）的评分减去 UBM 的评分作为最终分数。

$$S(U)=\log p(U\mid\lambda)-\log p(U\mid\bar{\lambda})\begin{cases}>\theta 接受\\ \le\theta 拒绝\end{cases}$$

其中 θ 是判决阈值，λ 是目标说话人模型，$\bar{\lambda}$ 是全局背景模型。

3.2 基于 Feature Warping 和 ICA 的 GMM-UBM 说话人确认系统

Feature Warping 和 ICA 的结合是一种在特征域上对特征向量进行补偿的方法，其核心思想是将发音方式不匹配的训练语音和测试语音的特征向量的各维进行统计独立分离，然后统一映射为标准正态分布，从而减少因训练语音和测试语音的发音方式不匹配而带来的性能急剧下降的影响。

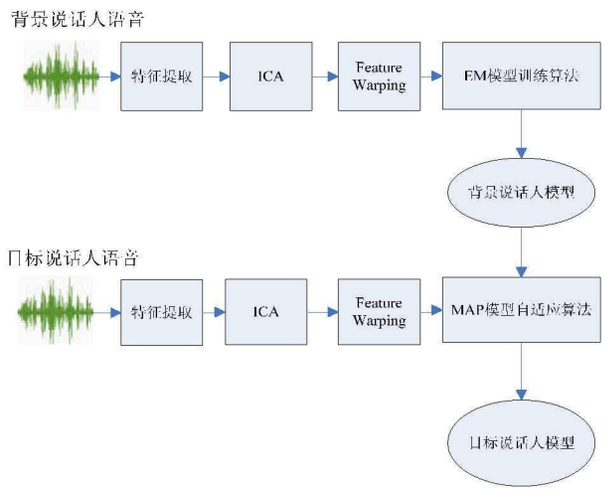

图 1 目标说话人模型的训练流程

说话人确认分为两个阶段，目标说话人模型的训练阶段和测试语音的确认阶段。目标说话人模型的训练流程如图 1 所示，首先分别提取目标说话人和背景说话人语音的特征矢量，然后进行 ICA 和 Feature Warping 处理，分别获得最终的目标说话人和背景说话人的语音特征。在背景说话人语音特征上，通过 EM 模型训练算法得到背景说话人模型，加上目标说话人语音特征，通过 MAP 模型自适应算法，最终训练得到目标说话人模型。

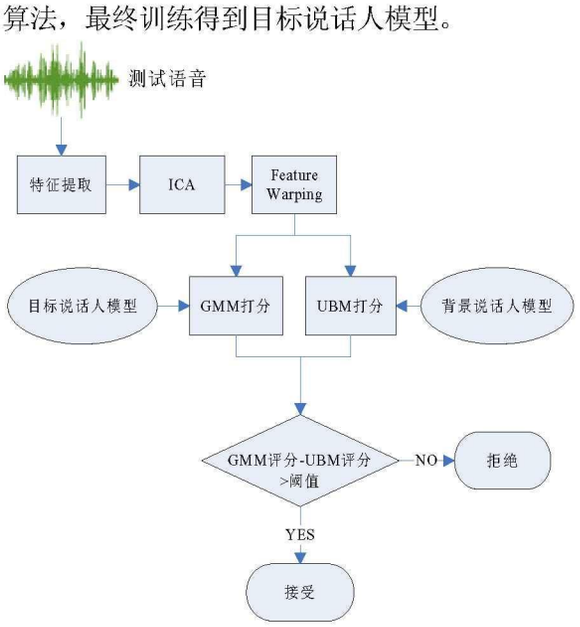

图 2 基于 Feature Warping 和 ICA 的 GMM-UBM 说话人确认系统

基于 Feature Warping 和 ICA 的 GMM-UBM 说话人确认系统如图 2 所示，首先提取测试语音的特征矢量，然后通过 ICA，将其映射到各维独立的向量空间，接下来通过 Feature Warping，获得最终各维独立的服从标准正态分布的特征矢量，进而分别进行目标说话人模型和背景说话人模型评分，通过分数规整，将最终得分与阈值进行比较，完成说话人的确认或拒绝。

4 实验结果与分析

4.1 多发音方式语音数据库

与[7]一样，本文考虑了发音方式变化的 6 个方面，包括表达方式、语速、音量、情感状态、身体状况和语种，故一种发音方式可以用以下 6 元组<表达方式、语速、音量、情感状态、身体状况、语种>来表示。

数据库共有 12 种不同的发音方式。若将发音方式<自然发音、中速、音量适中、中性、正常状态、汉语>作为基准发音方式，则通过每次改变基准发音方式中的一个维度，可以获得其他 11 种发音方式。

语音数据是在无明显噪声的办公室中录制的，均为同一电话信道，采样率为 8KHz。实验使用了数据库中 50 个说话人的数据，每人 12 种发音方式，每种发音方式的总时长为 3 分钟。

4.2 评估标准

本文采用的说话人确认的评估标准是在本研究领域被广泛采用的 DET（Detection Error Trade-off）[8]曲线和等错误率（Equal Error Ratio, EER），DET 曲线的横坐标和纵坐标分别对应 FA（错误接受率）和 FR(错误拒绝率)。

4.3 对比实验与结果分析

对比实验涉及到三个不同的识别系统，，分别是基线系统，基于 Feature Warping 的说话人确认系统和 Feature Warping 和 ICA 结合的说话人确认系统。

基线系统采用的特征矢量为 39 维的 MFCC，说话人模型为 GMM-UBM。在包含 12 种发音方式的 30 人的时长为 18 小时的数据基础上，训练得到包含 1024 个高斯混合的 UBM 模型。对时长为 22s 的目标说话人语音上通过 MAP 自适应得到目标说话人模型，语音的发音方式是在 12 种发音方式中随机选择的。在包含 12 种发音方式的 1900 段语音上做测试，每段测试语音的时长为 22s。由于目标模型训练语音的发音方式是随机选择的，所以最终系统性能(EER)是经过 10 轮“训练－测试”后的性能平均值。

基于Feature Warping的说话人确认系统则是在将最初提取得到的MFCC基础上，通过Feature Warping得到补偿后的特征矢量，再进行模型训练和测试。参照文中[2]所作的滑动窗口大小选择实验对比结论，当选取滑动窗口大小为300时，特征矢量映射为特定分布的效果最佳，故本文在实验中也将滑动窗口大小设为N=300。特征映射后的分布选取标准正态分布。

Feature Warping和ICA结合的说话人确认系统则是在对特征向量进行Feature Warping处理之前，先进行ICA分析，得到了各维独立的特征矢量之后，再进行Feature Warping处理。其中，ICA分析中的分离矩阵W是在一个包含8名说话人、12种发音方式的4.8小时的开发集上迭代训练出来的；Feature Warping的参数设置与前一组实验一样。

表1　三个说话人确认系统系统的EER

实验	EER
Baseline	32.02%
Feature Warping	30.19%
Feature Warping + ICA	23.86%

三组实验的等错误率如表1所示，在引进Feature Warping之后说话人确认系统，系统系能有所提升，其EER相对于基准系统下降了5.69%，而Feature Warping和ICA的结合则显著提高了系统的性能，EER相对于基准系统下降了25.48%。三组实验的DET曲线如图3所示，可以看到，Feature Warping对基线系统有一定的改善，但效果不是很明显，而Feature Warping和ICA相结合明显提升了基线系统的性能。

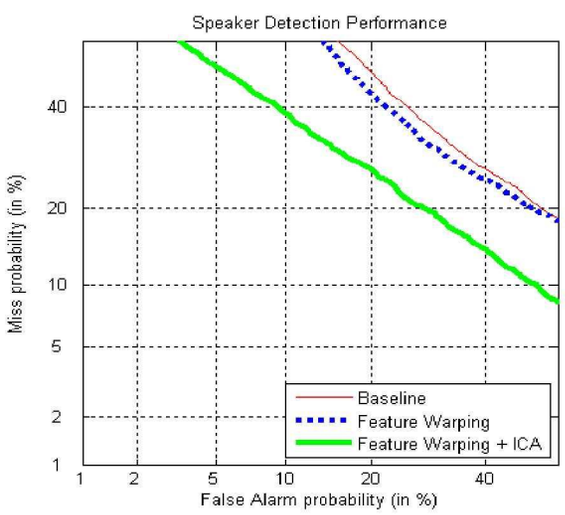

图3 三组实验的DET曲线

5　结论

本文针对因训练语音与测试语音的发音方式不匹配而导致的说话人确认系统性能下降的问题，提出了基于Feature Warping和ICA结合的在特征域上进行特征补偿方法，并采取GMM-UBM模型作为说话人模型，在一个包含多种发音方式的数据库上进行了相关实验测试。

实验结果表明：通过在特征域上Feature Warping和ICA相结合的特征补偿，基于Feature Warping和ICA的说话人确认系统的EER相对于基准系统降低了约25.48%，这说明本方法对面向多种发音方式的说话人确认系统的性能有了明显的提升。

本文专注于在特征域上针对发音方式变化的说话人确认的问题进行研究，今后还将尝试其他特征变换方法，并将在模型域和分数域上开展进一步的研究。

参 考 文 献

[1] Pelecanos and S. Sridharan. Feature Warping for robust speaker verification. Proc. ISCA Workshop on Speaker Recognition , 2001

[2] B. Xiang, U. Chaudhari, J. NavrAtil, G. Ramaswamy, and R. Gopinath. Short-time gaussianization for robust speaker verification. Proc. ICASSP, 2002

[3] P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel. Speaker and Session Variability in GMM-Based Speaker Verification. Audio, Speech, and Language Processing. IEEE Transactions,2007

[4] A. Hyvrinen, J. Karhunen and E. Oja. Independent Component Analysis. John Wiley & Sons,2001

[5] E. Bingham and A. Hyvrinen. A fast fixed-point algorithm for independent component analysis of complex-valued signals. Int. J. of Neural Systems, 2000

[6] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing, vol. 10, pp. 19-41, 2000

[7] Xu Mingxing, ZHANG Lipeng, WANG Linlin. Database Collection for study on speech variation robust speaker recognition. Proc of O-COCOSDA2008. Kyoto, 2008

[8] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. Pmc. Eurospeech, 1997