基于多模态生成对抗网络和三元组损失的说话人识别

陈 莹* 陈湟康

(江南大学轻工过程先进控制教育部重点实验室 无锡 214122)

摘 要：为了挖掘说话人识别领域中人脸和语音的相关性，该文设计多模态生成对抗网络(GAN)，将人脸特征和语音特征映射到联系更加紧密的公共空间，随后利用3元组损失对两个模态的联系进一步约束，拉近相同个体跨模态样本的特征距离，拉远不同个体跨模态样本的特征距离。最后通过计算公共空间特征的跨模态余弦距离判断人脸和语音是否匹配，并使用Softmax识别说话人身份。实验结果表明，该方法能有效地提升说话人识别准确率。

关键词：说话人识别；跨模态；生成对抗网络；3元组损失

中图分类号：TN912.3, TP391 **文献标识码**：A **文章编号**：1009-5896(2020)02-0379-07

DOI: 10.11999/JEIT190154

Speaker Recognition Based on Multimodal Generative Adversarial Nets with Triplet-loss

CHEN Ying CHEN Huangkang

(*Key Laboratory of Advanced Process Control for Light Industry* (*Ministry of Education*), *Jiangnan University*, *Wuxi* 214122, *China*)

Abstract: In order to explore the correlation between face and audio in the field of speaker recognition, a novel multimodal Generative Adversarial Network (GAN) is designed to map face features and audio features to a more closely connected common space. Then the Triplet-loss is used to constrain further the relationship between the two modals, with which the intra-class distance of the two modals is narrowed, and the inter-class distance of the two modals is extended. Finally, the cosine distance of the common space features of the two modals is calculated to judge whether the face and the voice are matched, and Softmax is used to recognize the speaker identity. Experimental results show that this method can effectively improve the accuracy of speaker recognition.

Key words: Speaker recognition; Cross-modal; Generative Adversarial Network (GAN); Triplet-loss

1 引言

说话人识别是生物识别领域中的一个活跃的研究问题，并且在商业应用和公共安全领域都有一定的应用需求。但是如何将面部和语音这两类异质模态信息适当融合，仍然是当前一项非常重要的任务。

多年来，研究者们在特征层融合和决策层融合都进行了探索。在早期，研究人员主要使用共同的特征转换方法，例如Bredin等人[1]利用典型相关分析(Canonical Correlation Analysis, CCA)来融合说话人识别的视听特征，而Haghighat等人[2]提出了判别相关分析(Discriminant Correlation Analysis, DCA)来融合视听特征以进行识别，不过这些转换方法会导致信息丢失。决策层融合方面，Cheng等人[3]利用所提出的(Incremental Kernel Fisher's Discriminant, IKFD)方法来获得人脸识别分数，并使用(Gaussian Mixed Model, GMM)来产生语音识别分数，然后融合分数；后来Soltane等人[4]提出了一种自适应贝叶斯方法来融合面部和语音模式的分数。这些融合方法虽然不会导致信息丢失，但却无法充分利用这些信息。

随着深度学习的发展及其在多模态学习领域取得成功，Hu等人[5]和Geng等人[6]使用CNN融合人脸特征和音频功能，取得了良好的效果；Ren等人[7]提出了一种用于说话人识别的多模态长短期记忆网络(Multimodal LSTM)学习模态之间的联系并进行了决策层分数融合，在一定程度上提升了识别性能。但他们并没有实现强力的模态之间相关性的约束；Liu等人[8]则是使用CNN提取特征，采用双向

收稿日期：2019-03-15；改回日期：2019-09-09；网络出版：2019-09-19
*通信作者： 陈莹 chenying@jiangnan.edu.cn
基金项目：国家自然科学基金(61573168)
Foundation Item: The National Natural Science Foundation of China (61573168))

LSTM分类。受跨模态检索领域的生成对抗网络(Cross-Modal Generative Adversarial Nets, CMGAN)启发，本文提出一种声视频多模态生成对抗及3元组损失网络 (Audio-Visual Generative Adversarial & Triplet-loss Network, AVGATN)用于说话人识别。本网络首先通过生成对抗的形式拉近声视频模态在所生成的公共空间上的距离，再利用3元组损失拉远公共空间上不同类别之间的距离，从而得到良好的识别性能。

2 跨模态生成对抗网络(CMGAN)

本节将简要介绍一下生成对抗网络(Generative Adversarial Nets, GAN)和CMGAN，并论述其在说话人识别领域的局限性。

2.1 生成对抗网络

生成对抗网络自2014年由Goodfellow等人[9]提出来以后，近年来在机器视觉领域中得到了广泛的关注和研究如图像识别[10]、多视图学习[11]、背景消减[12]等。GAN通常由生成器(generator)与判别器(discriminator)两部分组成，通过一组编码器和解码器对原始输入进行特征提取和数据重构，得到一个生成样本，记为$G(z)$。随后，该生成样本作为判别器的负样本输入(即假样本)，与真实数据x，一同作为判别器的输入训练判别器，最大化正确判别分数，使它学会正确区分真假数据。之后固定D，通过最小化正确判别分数，使得G生成的假数据尽可能地混淆D的判别。如此循环往复直至网络收敛稳定，最终就能使输入z通过G生成逼近真实数据分布的数据。具体目标函数如式(1)所示

$$\min_G \max_D V(G,D) = E_{x\sim p_{\text{data}}(x)}[\lg D(x)] + E_{x\sim p_z(z)}[\lg(1-D(G(z)))] \tag{1}$$

2.2 跨模态生成对抗网络

许多GAN相关的研究工作[13–16]具有有限的灵活性，因为它们仅通过单向路径网络结构解决从一种模态到另一种模式的生成问题。针对跨模态检索任务，Peng等人[17]提出了CMGAN，他们利用GAN对不同模态数据的联合分布进行建模，以学习共同特征，旨在进一步构建各种大规模异构数据的联系。

该网络分别对图片和文本各构建一个生成器，在生成器的编码器末端，两个模态都增加了两层全连接层，并共享了最后一层参数以提取共同特征s_p^t, s_p^i。其中，生成器中提取的样本原始特征h_p^t，h_p^i作为真实数据，生成器中解码器的重构输出分别为r_p^t, r_p^i。在判别器部分，该网络为模态间判别器D_{Ci}, D_{Ct}和模态内判别器D_I, D_T。因此，该网络的目标函数也变为如式(2)所示

$$\min_{G_I,G_T}\max_{D_I,D_T,D_{Ci},D_{Ct}} \mathcal{L}_{\text{GAN}_1}(G_I,G_T,D_I,D_T) \mathcal{L}_{\text{GAN}_2}(G_I,G_T,D_{Ci},D_{Ct}) \tag{2}$$

其中，$\mathcal{L}_{\text{GAN}_1}$与$\mathcal{L}_{\text{GAN}_2}$分别是多模态网络中模态内部的GAN目标函数与模态间的GAN目标函数，具体公式为

$$\mathcal{L}_{\text{GAN}_1} = E_{i\sim p_i}[D_I(i) - D_I(G_{\text{Idec}}(i))] + E_{t\sim p_t}[D_T(t) - D_T(G_{\text{Tdec}}(t))] \tag{3}$$

$$\mathcal{L}_{\text{GAN}_2} = E_{i,t\sim p_{i,t}}[D_{Ci}(G_{\text{Ienc}}(i)) - \frac{1}{2}D_{Ci}(G_{\text{Tenc}}(t)) - \frac{1}{2}D_{Ci}(G_{\text{Ienc}}(\hat{i})) + D_{Ct}(G_{\text{Tenc}}(t)) - \frac{1}{2}D_{Ct}(G_{\text{Ienc}}(i)) - \frac{1}{2}D_{Ct}(G_{\text{Tenc}}(\hat{t}))] \tag{4}$$

其中，i,t分别为原始的图片样本和文本样本，$\hat{i},\hat{t}$表示分属于不同类别的不匹配样本，G_{Idec}, G_{Tdec}为重构特征，G_{Ienc}, G_{Tenc}为生成器特征。

然而，该网络在测试时通过计算多个样本之间的余弦距离并投票，这种策略更适应于跨模态检索任务，而无法直接给出单独两个样本之间是否匹配以及从属于哪个类别的结果，因此本文在此基础上对模型进行修改，并将其应用于说话人识别领域。

3 声视频多模态生成对抗及3元组损失网络

受CMGAN体现出来的增强不同模态间联系的能力的启发，本文针对说话人识别问题，提出一种声视频多模态生成对抗及3元组损失网络(AVGATN)，通过不同模态样本间的对抗学习，得到多模态数据的公共特征空间，并基于此增加基于3元组的身份识别及特征匹配判断网络，获得良好的说话人识别性能。

3.1 网络结构

考虑到说话人识别任务中，模态内部的特征联系是不被关注的，因此本文去除了CMGAN中的重构网络和模态内部的判别器部分，并增加了身份识别网络和特征匹配判断网络。网络具体结构图如图1所示。

从图1中可以看出，网络的最初始输入分别为人脸图片和语音信号，在输入GAN之前，图1分别对它们进行预处理。语音部分采用传统语音识别中常用的梅尔倒谱系数(Mel Frequency Cepstrum Coefficients, MFCC)特征，考虑到说话人识别任务中的样本是一个时域窗口截取的一段音视频，人脸部分采用普通的卷积神经网络的全连接层特征，将该样本内多张人脸特征并联，则得到一个代表该样本的人脸特征矩阵。

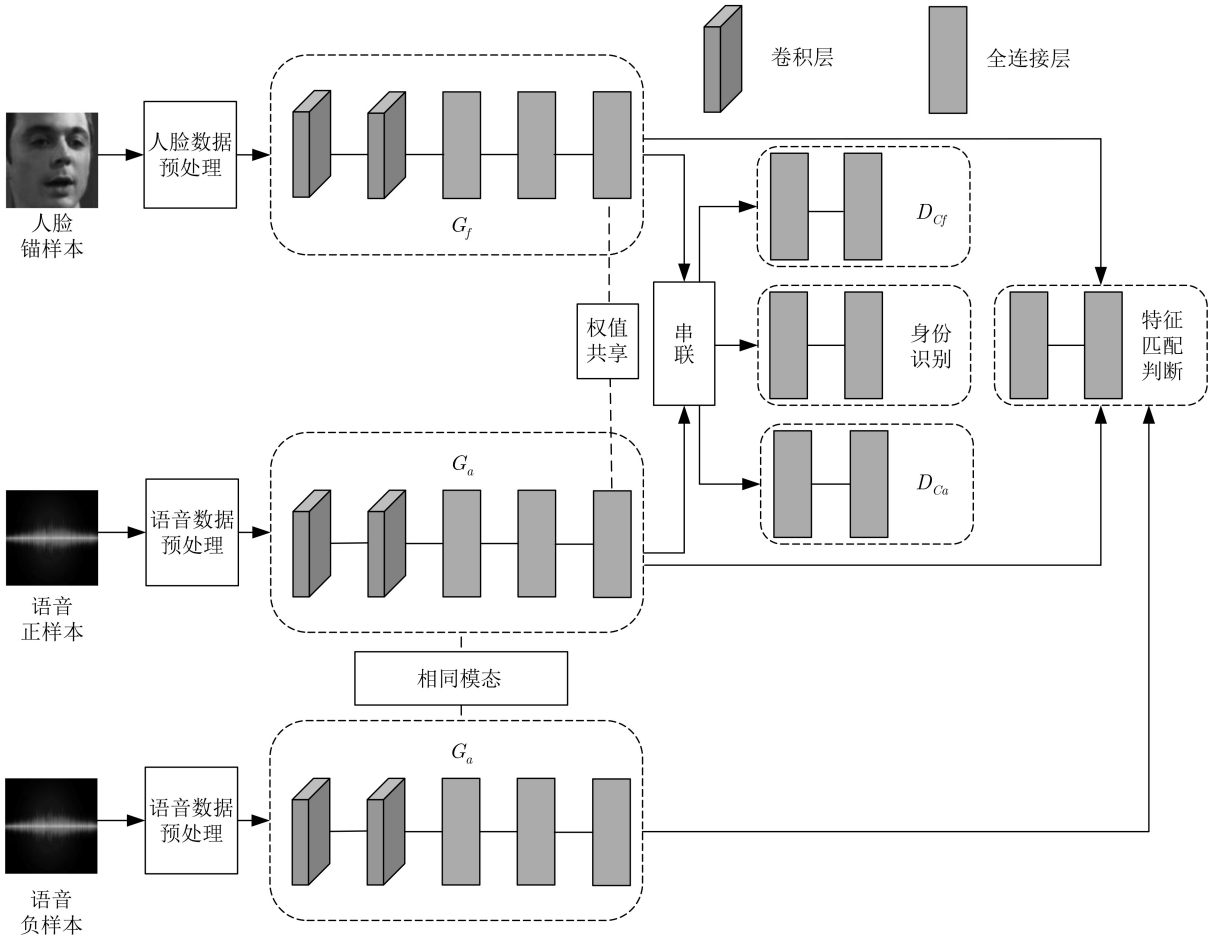


图 1 本文所提网络结构图

人脸模态和语音模态各有一个生成器G_f, G_a，由2层卷积层，2层全连接层组成。为了提取公共空间的特征，在2个生成器末端共享了一层全连接层，记人脸和语音在公共层之前得到的特征分别为h_f和h_a，经过公共层之后得到的特征分别为s_f和s_a。在公共层之后，网络分为了3部分。

一部分为判别器部分，包含2个跨模态判别器D_{Cf}, D_{Ca}。对于D_{Cf}来说，h_f与s_f的串联特征为真实数据，h_f与s_a的串联特征为虚假数据，对于D_{Ca}来说，h_a与s_a的串联特征为真实数据，h_a与s_f的串联特征为虚假数据。

另一部分为身份识别网络，该网络的输入为人脸公共特征s_{f}和语音公共特征s_a的串联，网络由2层全连接层组成，最后一层的激活函数为softmax，用于身份(IDentification, ID)分类。

最后一部分为特征匹配判断网络，该部分网络由2层全连接层组成，输入为人脸公共特征和语音公共特征的3元组，是2个模态共享的网络，用于判断样本的2个模态特征是否来自同一个ID。

3.2 目标函数及网络训练

本文所提网络模型采用的是分步骤训练策略，首先对GAN部分进行预训练，随后固定GAN，训练身份识别网络以及特征匹配判断网络，下文将逐步介绍。

首先对于GAN网络，优化目的依然是使生成器生成的特征混淆判别器的判别，从而使得s_f和s_a更接近。因此，目标函数如式(5)所示

$$\min_{G_f,G_a}\max_{D_{Cf},D_{Ca}}\mathcal{L}_{\mathrm{GAN}}(G_f,G_a,D_{Cf},D_{Ca}) \tag{5}$$

在训练过程中，判别器率先被更新，为了使整个网络都能使用梯度下降法优化，对D_{Cf}, D_{Ca}的优化将通过最小化如式(6)和式(7)所示的损失函数实现

$$\mathcal{L}_{D_{Cf}}=\frac{1}{N}\sum_{p}^{N}[\lg D_{Cf}(h_f^p,s_a^p)+\lg(1-D_{Cf}(h_f^p,s_f^p))] \tag{6}$$

$$\mathcal{L}_{D_{Ca}}=\frac{1}{N}\sum_{p}^{N}[\lg D_{Ca}(h_a^p,s_f^p)+\lg(1-D_{Ca}(h_a^p,s_a^p))] \tag{7}$$

其中，N为批次大小，(h,s)表示将两种特征串联。该损失函数的目的是为了使每个判别器区分出输入的串联特征中，公共空间特征是否来自于本模态。

每更新完一次判别器，则进行若干次生成器的更新，G_f, G_a的优化也将通过最小化如式(8)和式(9)所示的损失函数实现

$$\mathcal{L}_{G_f} = \frac{1}{N}\sum_{p}^{N}\lg(1 - D_{Cf}(h_f^p, s_a^p)) \tag{8}$$

$$\mathcal{L}_{G_a} = \frac{1}{N}\sum_{p}^{N}\lg(1 - D_{Ca}(h_a^p, s_f^p)) \tag{9}$$

该损失函数的目的是为了使生成器生成的公共空间特征，能够被每个模态的判别器识别成本模态内部生成的特征。通过交替反复训练，最终同一样本的语音特征和人脸特征，由GAN生成的2个模态的公共空间特征互相靠近，在余弦距离度量上距离趋近于1。

然而，因为在GAN中没有引入人脸与语音不匹配的训练样本，因此在处理说话人识别任务时，无论样本的人脸与语音是否匹配，几乎所有样本的公共空间特征的余弦距离都会趋近于1。若是在GAN训练中引入不匹配样本，则GAN无法训练成功。因此本文将前述训练好的GAN部分参数固定，在公共层后面增加了特征匹配判断网络，记人脸特征和语音特征在该部分网络最后一层的输出分别为t_f和t_a，则其损失函数定义为

$$\mathcal{L}_{\text{triplet}} = \frac{1}{N}\sum_{p}^{N}\max(d(t_{\boldsymbol{f}}^p, \tilde{t}_{\boldsymbol{a}}^p) - d(t_{\boldsymbol{f}}^p, t_{\boldsymbol{a}}^p) + \text{margin}, 0) \tag{10}$$

$$\left.\begin{aligned} d(t_{\boldsymbol{f}}^p, \tilde{t}_{\text{a}}^p) &= \text{abs}(\cos(\text{t}_{\text{f}}^p, \text{t}_{\text{a}}^p)) \\ d(t_{\boldsymbol{f}}^p, t_{\text{a}}^p) &= \text{abs}(\cos(\text{t}_{\text{f}}^p, \text{t}_{\text{a}}^p)) \end{aligned}\right\} \tag{11}$$

其中，$t_{\boldsymbol{f}}^p$, $\tilde{t}_{\boldsymbol{a}}^p$与$t_{\boldsymbol{a}}^p$为一个3元组，$t_{\boldsymbol{f}}^p$是锚样本，$t_{\boldsymbol{a}}^p$是正样本，$\tilde{t}_{\boldsymbol{a}}^p$是负样本。余弦距离计算后取绝对值是为了方便测试时使用阈值法进行判别，设置margin的目的是用于困难样本挖掘，使得模型只对区分错误或区分度不高的样本进行训练。

在说话人识别任务中，除了判断单个样本的人脸与语音是否匹配之外，还需要判断匹配样本属于哪个ID。因此在公共空间的特征层之后，增加了一个简易的识别网络，该网络的分类由softmax层实现，其损失函数定义为

$$\mathcal{L}_{\text{id}} = \frac{1}{N}\sum_{p}^{N} - \lg\left(\frac{e_p^{\boldsymbol{f}_{y_p}}}{\sum_j e_p^{\boldsymbol{f}_j}}\right) \tag{12}$$

其中，$e_p^{\boldsymbol{f}_j}$表示第p个样本的class score向量$\boldsymbol{f}$中第j个元素，y_p表示第p个样本的真实ID标签在向量 中的序号。该部分的网络输入为人脸模态和语音模态的公共特征的串联。

4 实验及结果分析

4.1 网络设置

本节将介绍本文的实验参数设置以及实验结果，并对实验结果作简要分析，验证本文所提模型的有效性。实验选取的数据集是由Ren等人[7]提供的开源数据集，其中的人脸与语音数据都采集于美剧《生活大爆炸》，分别属于剧中5个主要角色Sheldon, Leonard, Howard, Raj和Penny。训练样本和测试样本都是时域上的0.5 s窗口截取的音视频数据，同时为了便于对比实验结果，本文采用Ren等人[7]的数据预处理方法，即对人脸样本提取53×49的CNN特征[18]，将时域窗口中同一人的人脸特征并联，对语音样本提取25×49的MFCC特征。另外，除了生成人脸和语音同属于一个ID的匹配样本之外，本文按照锚样本与正样本、负样本的3元组的方法，增加了人脸和语音分属于不同ID的不匹配样本。最终实验的训练和测试过程中各使用了25000个匹配样本和25000个不匹配样本，其中训练集在迭代过程中划分为40000个训练样本和10000个验证样本。

4.2 参数分析

实验中，共设置70000次迭代，批次大小为100，前30000次用于训练GAN，后40000次用于训练身份识别网络和度量学习网络。其中训练GAN过程中，每更新一次判别器之后，要更新5次生成器。此外，网络所有层后面都添加了BN层。测试时，样本经过网络前向的传播之后，计算两个模态在度量学习最后一层的全连接层特征的余弦距离，如果距离大于某个阈值，则认为该样本的人脸和语音属于同一个人，反之则不属于。对于判断为属于同一个人的样本，取出其身份识别网络的输出结果作为该样本的预测ID。对于训练过程中margin值的选择，本文通过在验证集上多次实验进行筛选，模态匹配及识别ROC曲线如图2所示。

其中，假正率表示不匹配样本中被判断成匹配样本的比例，与通常的ROC曲线不同，本文图中

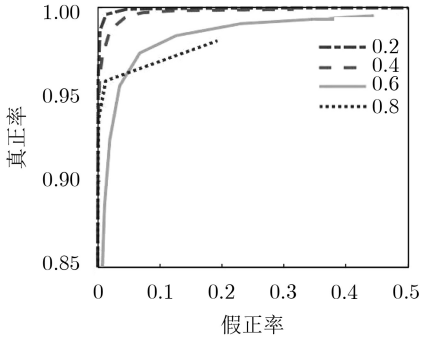


图 2 不同margin值的ROC

真正率表示的是匹配样本中被正确判断，并且被正确识别出ID的比例。可以从图2看到，margin取0.2时，对应的ROC曲线的AUC面积最大，因此本文最终采用0.2的margin值。

而对于阈值的选取，本文在验证集上对比了采用不同阈值时的多个指标的识别结果，如图3所示。观察到当阈值选取0.5时，网络性能最好，因此本文选取0.5的阈值进行测试。值得一提的是，测试其他margin值训练的网络时，网络的综合性能也在0.5附近取得了峰值。

4.3 模块必要性分析与特征选择

4.3.1 公共层必要性分析

为了验证在多模态网络中，建立声视频双模态公共空间的必要性，本文对比了将公共层分别嵌入两个模态前后实验结果，并将其绘制成ROC曲线。从图4可以看出，去除公共层之后，网络的识别判断性能大幅度下降。这是因为没有公共层时，两个模态的特征分布相差甚远，网络无法通过3元组的距离约束来建立模态之间的联系。

4.3.2 特征匹配判断网络必要性分析

由于预训练GAN已经使得两个模态的特征在公共空间距离靠近，为了验证使用3元组损失训练特征匹配判断网络的必要性，本文对比了选用公共层特征进行识别与选用特征匹配判断网络特征进行识别的实验结果。其中直接选用公共层特征的实验结果如图5所示。

图5中所示匹配准确率表示匹配样本和不匹配样本分别被正确判断的数目总和占全部测试样本的比例，ID识别准确率即上述的真正率，总准确率表示不匹配样本被正确判断的数目与匹配样本被正确判断并识别出ID的数目之和，占全部测试样本的比例。从图5(a)中可以看到，无特征匹配判断网络的实验结果在全阈值范围内都不如有特征匹配判断网络的结果，并且都维持在50%的精度左右。这表明不使用特征匹配判断网络时，匹配样本与不匹配样本在余弦距离空间没有区分度，互相混淆分布。

从图5(b)中可以看到，无特征匹配判断网络的ID识别结果与有特征匹配判断网络的识别结果相持平甚至优于，且均在90%的精度以上。由于ID识别只针对匹配样本，而ID识别网络与特征匹配判断网络互相独立，因此该实验结果的比较只与匹配样本的判断有关，当阈值为0.9时，无特征匹配判断网络的识别精度依然高达90%以上，表明通过预训练GAN得到的不同模态的公共层特征，在余弦距离空间靠近，验证了GAN预训练的有效性。

从图5(c)中可以看到，在总准确率指标上，有特征匹配判断网络的识别结果明显优于无特征匹配判断网络的识别结果。从图5(a)，图5(b)，图5(c)，

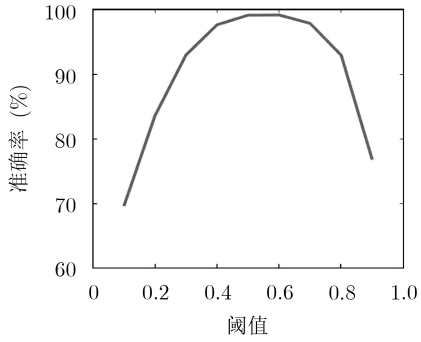


图 3 不同阈值的识别结果

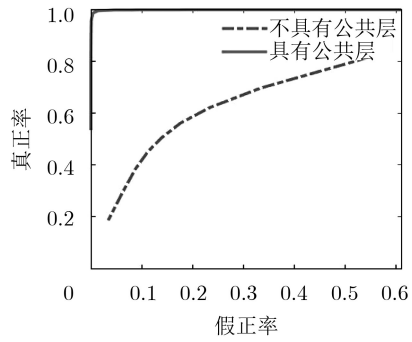
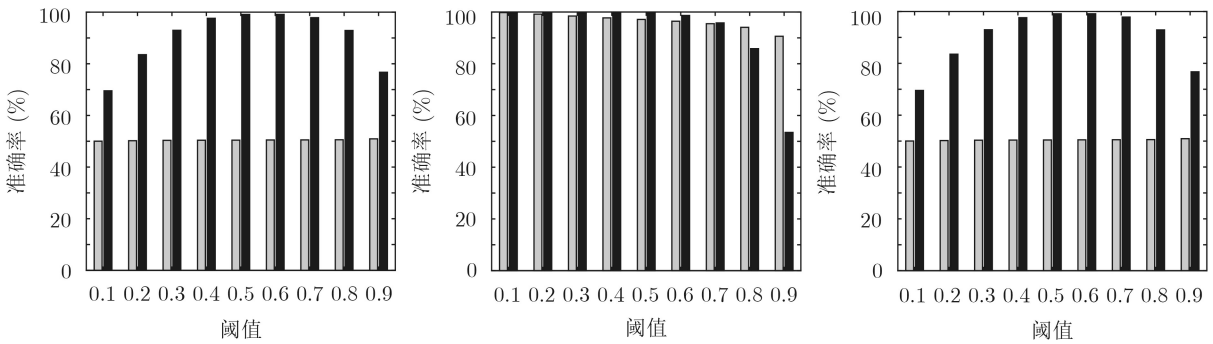


图 4 是否具有公共层的ROC曲线对比



(a) 匹配准确率对比结果　(b) ID识别准确率对比结果　(c) 总准确率对比结果

无特征匹配判断　有特征匹配判断

图 5 有无特征匹配判断网络识别结果对比

可以发现不使用特征匹配判断网络时，不论样本的两个模态是否匹配，特征的余弦距离都趋近于1。因此该实验验证了特征匹配判断网络能有效地拉远不匹配样本不同模态间的距离，拉近匹配样本不同模态间的距离，明显地增加了类别之间的区分度。

4.3.3 识别层输入模式分析

在网络的身份识别模块，本文试验了3种特征，分别是语音的公共特征、人脸的公共特征以及二者的串联特征，评价指标采用ID识别准确率，结果如表1所示。

表 1 不同特征的身份识别准确率(%)

特征	ID识别准确率
语音公共特征	95.57
人脸公共特征	99.41
串联特征	99.59

由表1可知，使用串联特征时，网络的身份识别性能最好，略高于人脸公共特征，这是因为特征中增加了语音模态的信息，使得特征的判别力增强。值得一提的是，单语音特征的识别性能也远远高于其原始MFCC特征，这是因为由于多模态GAN，两个模态的特征在公共空间不断靠近，人脸的分类信息也在训练过程中注入了语音特征中。

4.4 与现有方法的比较

本文选取了近几年说话人识别领域的方法，与本文所提模型进行了对比，结果如表2所示。

表 2 说话人身份识别准确率(%)

方法	ID识别准确率	匹配准确率
Multimodal Correlated NN[6]	83.26	–
Multimodal CNN[5]	86.12	–
Multimodal LSTM[7]	90.15	94.35
Deep Heterogeneous Feature Fusion.[8]	97.80	–
本文AVGATN	99.41	99.02

从表2中可以看出，本文所提模型在判断样本是否匹配的任务上有一定的性能提升，这表明通过GAN将两个模态的特征映射到同一分布空间，再利用3元组损失训练拉近类内距离，拉远类间距离，可以有效地增加类别之间的区分度。在说话人身份识别方面，本文的精度相比较之前的研究者都有显著提升，与文献[8]的模型相比也有所提升。这是因为人脸的CNN特征本身具有良好分类性能，并且通过GAN之后，映射到公共空间的人脸特征分布也学到了一些语音特征中的分类信息，使得特征的分类性能在原有基础上得到进一步提升。

5 结论

本文通过对说话人识别任务中的人脸和语音分别建模，搭建并训练多任务说话人识别网络，通过共享全连接层，将两个模态的特征映射到公共空间且使其靠近，在此基础上，对公共空间特征构建多任务网络，其中对人脸模态的公共空间特征进行单独身份识别网络训练，同时对公共空间的特征进行特征匹配判断学习，使得同属于一个ID的不同模态的特征在余弦空间接近，属于不同ID的疏远，最终使用阈值法判断样本的两个模态特征是否匹配。实验结果表明，本文所提方法能有效提高说话人识别精度。

参 考 文 献

[1] BREDIN H and CHOLLET G. Audio-visual speech synchrony measure for talking-face identity verification[C]. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, USA, 2007: Ⅱ-233–Ⅱ-236.

[2] HAGHIGHAT M, ABDEL-MOTTALEB M, and ALHALABI W. Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition[J]. *IEEE Transactions on Information Forensics and Security*, 2016, 11(9): 1984–1996. doi: 10.1109/TIFS.2016.2569061.

[3] CHENG H T, CHAO Y H, YEH S L, *et al*. An efficient approach to multimodal person identity verification by fusing face and voice information[C]. 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, Netherlands, 2005: 542–545.

[4] SOLTANE M, DOGHMANE N, and GUERSI N. Face and speech based multi-modal biometric authentication[J]. *International Journal of Advanced Science and Technology*, 2010, 21(6): 41–56.

[5] HU Yongtao, REN J S J, DAI Jingwen, *et al*. Deep multimodal speaker naming[C]. The 23rd ACM International Conference on Multimedia, Brisbane, Australia, 2015: 1107–1110.

[6] GENG Jiajia, LIU Xin, and CHEUNG Y M. Audio-visual speaker recognition via multi-modal correlated neural networks[C]. 2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops, Omaha, USA, 2016: 123–128.

[7] REN J, HU Yongtao, TAI Y W, *et al*. Look, listen and learn-a multimodal LSTM for speaker identification[C]. The 30th AAAI Conference on Artificial Intelligence, Phoenix, USA, 2016: 3581–3587.

[8] LIU Yuhang, LIU Xin, FAN Wentao, *et al*. Efficient audio-visual speaker recognition via deep heterogeneous feature fusion[C]. The 12th Chinese Conference on Biometric

Recognition, Shenzhen, China, 2017: 575–583.

[9] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, *et al*. Generative adversarial nets[C]. The 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014: 2672–2680.

[10] 唐贤伦, 杜一铭, 刘雨微, 等. 基于条件深度卷积生成对抗网络的图像识别方法[J]. 自动化学报, 2018, 44(5): 855–864.
TANG Xianlun, DU Yiming, LIU Yuwei, *et al*. Image recognition with conditional deep convolutional generative adversarial networks[J]. *Acta Automatica Sinica*, 2018, 44(5): 855–864.

[11] 孙亮, 韩毓璇, 康文婧, 等. 基于生成对抗网络的多视图学习与重构算法[J]. 自动化学报, 2018, 44(5): 819–828.
SUN Liang, HAN Yuxuan, KANG Wenjing, *et al*. Multi-view learning and reconstruction algorithms via generative adversarial networks[J]. *Acta Automatica Sinica*, 2018, 44(5): 819–828.

[12] 郑文博, 王坤峰, 王飞跃. 基于贝叶斯生成对抗网络的背景消减算法[J]. 自动化学报, 2018, 44(5): 878–890.
ZHENG Wenbo, WANG Kunfeng, and WANG Feiyue. Background subtraction algorithm with Bayesian generative adversarial networks[J]. *Acta Automatica Sinica*, 2018, 44(5): 878–890.

[13] RADFORD A, METZ L, and CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. *arXiv*: *1511*.06434, 2015.

[14] DENTON E, CHINTALA S, SZLAM A, *et al*. Deep generative image models using a laplacian pyramid of adversarial networks[C]. The 28th International Conference on Neural Information Processing Systems, Montreal, Canada, 2015: 1486–1494.

[15] LEDIG C, THEIS L, HUSZÁR F, *et al*. Photo-realistic single image super-resolution using a generative adversarial network[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 105–114.

[16] WANG Xiaolong and GUPTA A. Generative image modeling using style and structure adversarial networks[C]. The 14th European Conference on Computer Vision, Amsterdam, Netherlands, 2016: 318–335.

[17] PENG Yuxin and QI Jinwei. CM-GANs: Cross-modal generative adversarial networks for common representation learning[J]. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2019, 15(1): 98–121.

[18] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, *et al*. Improving neural networks by preventing co-adaptation of feature detectors[J]. *Computer Science*, 2012, 3(4): 212–223.

陈　莹：女，1976年生，教授，博士，研究方向为信息融合、模式识别.

陈湟康：男，1994年生，硕士生，研究方向为说话人识别.