阐述

研究的是小数据&多信道环境下的语种识别。具体来说是在小数据和多信道环境下如何提高语种识别的性能。

首先，做对比实验，希望通过依次增加训练数据量实验，对比实验结果，确定小数据和多信道对语种识别性能的影响。

然后，研究在小数据和多信道环境下使用哪些方法可以明显改善语种识别性能。目前所使用的方法是数据增强和加入共享信息两种。使用的语言信息借助 chain model。

所使用的基线模型有 ivector, xvector 。

repo

Target

Investigation on how to improve the performance of LID system under low-resource & multi-domain condition.

Related papers

[1] xvector: https://www.isca-speech.org/archive/Odyssey_2018/pdfs/38.pdf

[2] multi-Domain：https://www.isca-speech.org/archive/Odyssey_2018/pdfs/35.pdf

[3] classfier bancked: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7312905

[4] the parameter settings of ivector baseline & training data: https://arxiv.org/pdf/1806.00616.pdf AP18-OLR Challenge: Three Tasks and Their Baselines

[5] civ=400 : Language Recognition via Ivectors and Dimensionality Reduction

[6] Phonetic Temporal Neural Model for Language Identification

[7] dnn-ivector : [6] & [7] I-vector representation based on bottleneck features for language identification https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6680440

[8]Phonetic:the mitll nist lre 2009 language recognition system

I-vector

[9]Front-End Factor Analysis for Speaker Verification

[10]Discriminative and generative approaches for long and shortterm speaker characteristic modeling application to speaker verification

[11]i-vector representation based on bottleneck features for language identification

[12]Analyzing the Effect of Channel Mismatch on the SRI Language Recognition Evaluation 2015 System

[13] speaker verification using end-to-end adversarial language adaptation

[14] Unsupervised Domain Adaptation via Domain Adversarial Training for Speaker Recognition

[15] Unsupervised Domain Adaptation by Backpropagation

[16] HOW TO IMPROVE YOUR SPEAKER EMBEDDINGS EXTRACTOR IN GENERIC TOOLKITS

[17] UNSUPERVISED ADVERSARIAL DOMAIN ADAPTATION FOR ACOUSTIC SCENE CLASSIFICATION

[18] AP18-OLR Challenge: Three Tasks and Their Baselines

Information of the Dataset

Training data (106h) from AP18-OLR.
In-domain test data from AP18-OLR, More details refer to this paper[18]
(包括汉语普通话、粤语、日语、韩语、俄语、越南语、印尼语、藏语、维吾尔语、哈萨克语。16kHz-16bit，channel=mobile)
粤语，韩语和普通话属于 Confusing-language，所以应避免在少量种类语种的情况下同时有这三种。

out-of-domain test data
(包括汉语普通话、日语、俄语、越南语、藏语、维吾尔语,6 种。*kHz-*bit，channel=video) 提取语音段特征之前，首先规整了语音的参数：16kHZ 采样率，16bit 量化级，采用 wav 格式保存。

训练集和集内外测试集的划分如下面 Table 0 中所示，其中训练数据总时长 106.58h，集内测试数据总时长为 34.05h，集外测试集总时长 15.71h。

Table 0 训练与测试集分布

语种标识	具体描述	训练数据		集内测试数据		集外测试数据	
		channel：mobile		channel：mobile		channel：video	
		时长/h	句子总数	时长/h	句子总数	时长/h	句子总数
ka-cn	哈萨克语	9.52	4200	-	1800	-	-
ct-cn	广东话	10.21	7559	-	1800	-	-
id-id	印尼语	10.68	7671	-	1800	-	-
ko-kr	韩语	7.92	7196	-	1800	-	-
ti-cn	西藏语	11.93	11100	-	1800	3.07	3698
uy-id	维吾尔语	13.61	5800	-	1800	2.57	3096
zh-cn	普通话	10.32	7198	-	1800	2.35	2823
ja-jp	日语	8.00	7662	-	1800	2.63	3155
ru-ru	俄语	12.95	7190	-	1800	2.19	2635
vi-vn	越南语	11.43	7200	-	1800	2.9	3503

Basic ideas

1. confirm the problem of low-resource & multi-domain for LID task.

使用方法：Incremental learning:
25h 训练数据 -> 50h 训练数据 -> 75h 训练数据 -> 106h 训练数据
将训练数据集分别抽取 25h、50h、75h，抽取规则为随机抽取、语种类别数据平衡。得到 train_25h train_50h train_50 train_106h 四种不同数据量的训练集。
得出实验结果如 table 1：

RESULT:

【Table 1】

		in-domain				Out-domain			
		pos	cos	lr	lda-plda	pos	cos	lr	lda-plda
i-vector	25h	-	69.65	71.43	73.67	-	24.26	31.94	29.31
	50h	-	83.05	84.00	86.30	-	31.79	37.28	36.34
	75h	-	88.56	88.05	90.06	-	**35.26**	**40.15**	**38.43**
	106h	-	**89.68**	**90.62**	**91.05**	-	34.23	37.51	37.59
x-vector	25h	58.00	58.21	61.70	65.92	27.98	21.84	31.05	36.11
	50h	72.77	66.08	76.76	82.34	30.07	22.45	35.56	37.44
	75h	78.92	72.11	82.87	87.15	28.60	**24.36**	**37.76**	**39.05**
	106h	**81.66**	**72.42**	**86.34**	**89.83**	**28.73**	23.44	35.48	38.43

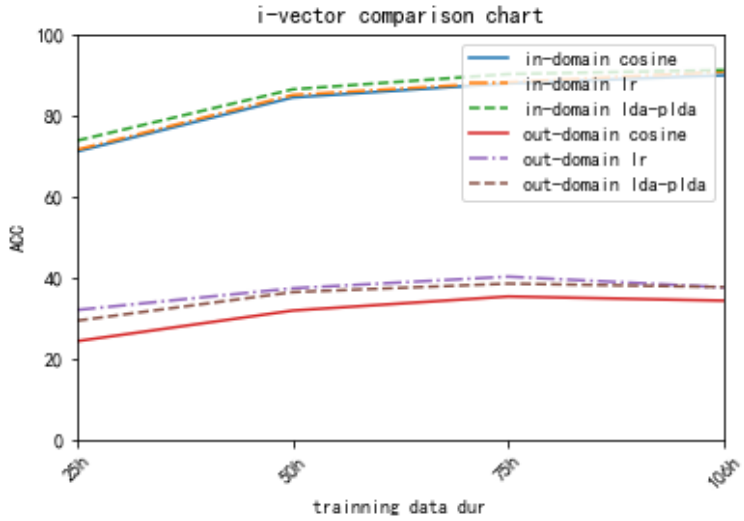


Fig.1

结论：

（1）从图 1 中可以很直观的看出：在 in-domain 时，数据量越小识别率越低。在数据量相同的时候，out-of-domain 的识别率比 in-domain 的识别率低很多。从本实验结果来看，25h&out-of-domain 的识别率比 106h&in-domain 的识别率最高可低 59.11。

（2）表 2 的实验结果充分表明上下文对语种识别很重要。

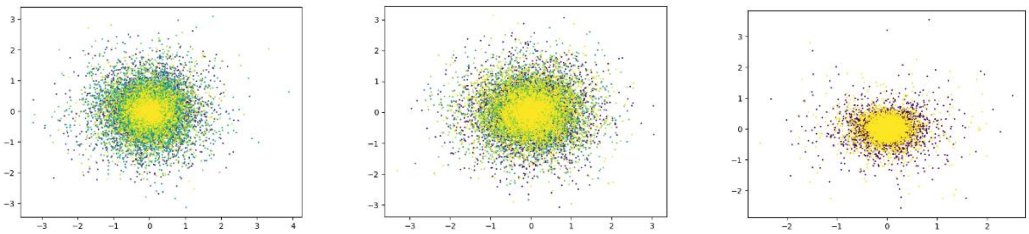
（3）lda-plda 后端参数设置中，从图 3 中可以看出：无论在 i-vector 系统还是 x-vector 系统，lda 设置为 9 最佳。在 i-vector 系统中，集内测试 lda 维度设置为 9 最佳，而对于集外则在 10 之后处于下降趋势。在 x-vector 系统中，集内测试同样在设置为 9 的时候最佳，而集外则在处于 10 的时候存在最低点。此结论在 aug 测试中同样存在。

（4）无论在 i-vector 还是 x-vector 系统，集外测试集在 75h 的识别率比 106h 的识别率还要好，应该是过拟合？

Analysis:

(1) **t-SNE**

For further analysis, first I used t-SNE to plot the distribution of the in-domain i-vectors and out-of-domain i-vectors. then, I used it to plot the distribution of the in-domain i-vectors and out-of-domain i-vectors



in-domian　　out-domain　　in & out-domain

Fig.2

The left figure is the distribution of in-domain i-vectors, the middle figure is the distribution of out-of-domain i-vectors, and the right one is the distribution of in & out-of-domain i-vectors.

(2) **confusion matrix**

Second, I used confusion matrix to analyze the confusion between languages

106h-ivector

Confusion matrix:

```
  kazak     tibet     uyghu     ct-cn     id-id     ja-jp     ko-kr     ru-ru     vi-vn     zh-cn
[[1.714e+03 1.300e+01 1.600e+01 0.000e+00 1.200e+01 1.100e+01 2.000e+00 2.500e+01 3.000e+00 4.000e+00]
 [1.000e+00 1.798e+03 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 1.000e+00]
 [2.900e+01 1.000e+00 1.759e+03 0.000e+00 0.000e+00 4.000e+00 2.000e+00 4.000e+00 0.000e+00 1.000e+00]
 [8.400e+01 1.100e+02 8.000e+00 1.785e+03 1.700e+01 5.400e+01 2.820e+02 1.100e+01 5.900e+01 1.460e+02]
 [2.600e+01 0.000e+00 1.000e+00 8.000e+00 2.477e+03 1.400e+01 1.900e+01 6.000e+00 4.000e+00 2.000e+00]
 [1.700e+01 2.000e+00 2.000e+00 1.000e+00 1.200e+01 2.506e+03 6.000e+00 2.000e+00 0.000e+00 0.000e+00]
 [7.100e+01 1.700e+01 4.800e+01 7.000e+01 3.900e+01 1.630e+02 1.795e+03 3.800e+01 8.900e+01 6.800e+01]
 [2.180e+02 0.000e+00 3.000e+00 0.000e+00 9.000e+00 4.000e+00 1.500e+01 1.543e+03 0.000e+00 4.000e+00]
 [3.700e+01 0.000e+00 0.000e+00 8.000e+00 4.200e+01 0.000e+00 1.400e+01 1.000e+00 2.288e+03 6.000e+00]
 [3.700e+01 0.000e+00 7.000e+00 6.400e+01 1.700e+01 9.000e+00 7.400e+01 5.800e+01 3.900e+01 2.095e+03]]
MAP: 0.8968239597083757
ACC: 0.8961044850573671
```

Conclusion：

In（1）, it can be observed that for the in-domain & out-domain set, the distribution is very messy, and i-vectors from different languages are mixed together. and the right one demonstrated that is domain-dependent.

in-domain 与 out-of-domain 数据的边缘分布不同，及数据整体不相似。

In（2）, it can be observed that ru-ru is easily misjudged as kazak and ko-kr is easily misjudged as ja-jp

Process:

1.1 i-vector system

According to the paper[5], we set civ to 400.and other parameters refer to [18], then we will do the experiment and select the appropriate cnum=1024.

1.2 Xvector-system

网络结构：dim[input-512-512-512-512-1500-pooling-512-512-output]

参数配置：

num_repeats =5*10^4

frames_per_iter =5*10^10

并且根据以下结果设置 apply_cmvn_sliding=true。

【Table 2】

	Xvector【106h】		
	pos	cosine	lr
cmvn_sliding=false	72.88	67.36	73.35
cmvn_sliding=true	81.66	72.42	86.34

Based on the results, the selection parameter is set to apply_cmvn_sliding=true, and it fullly prove that the context is very important for Language Recognition.

1.3 Classifer-bankend1---lr 参数设置【控制变量】：

Normalizer=*（根据不同训练数据进行调整）

Max-steps=30

Power=0.15

mix-up=40

1.4 classifer-bankend2---lda-plda【lda 维度设置】

in fig 3，我们可以看出，无论在 i-vector 系统还是 x-vector 系统，lda 设置为 9 最佳。

在 i-vector 系统中，集内测试 lda 维度设置为 9 最佳，而对于集外则在 10 之后处于下降趋势。在 x-vector 系统中，集内测试同样在设置为 9 的时候最佳，而集外则在处于 10 的时候存在最低点。此结论在 aug 测试中同样存在。

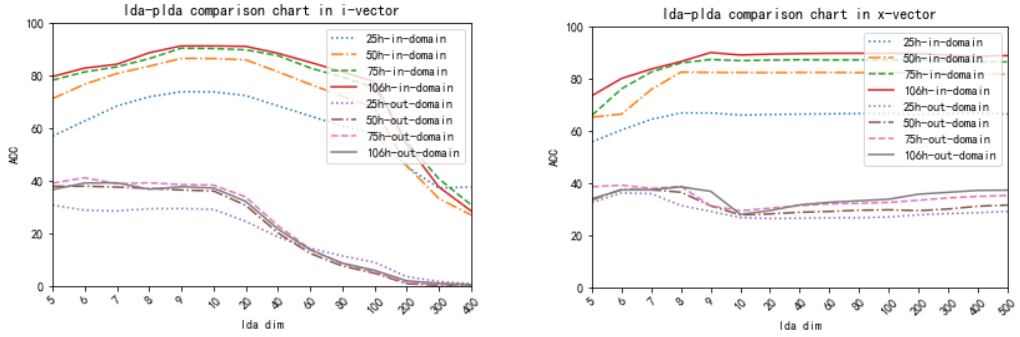


fig 3

For the investigation on how to improve the performance of LID system under low-resource & multi-domain condition, do the following experiments.

2. 解决方法 1---data augmentation: 通过数据增强

在原始“干净”数据集上增加增强数据。
25h+ * 训练数据 -> 50h+ * 训练数据 -> 75h+ * 训练数据 -> 106h+ * 训练数据

Information of the train-aug set:
Including: speed perturbation, vol, reverb, musan(music, noise, babble)

---

aug1: 2fold=clear+(sp+vol+reverb+musan)
aug2: 5fold=clear+sp+vol+reverb+musan

We use augmentation to increase the amount and diversity of the language system training data.aug1 为“干净”语音和各种噪音的叠加组成，aug2 为“干净”语音和分别一份的噪音组成。实验结果如下：

RESULT:

i-vector

	in-domain				Out-domain			
	pos	cos	lr	lda-plda	pos	cos	lr	lda-plda
25h	-	69.65	71.43	**73.67**	-	24.26	**31.94**	29.31
25h_aug1	-	68.42	69.89	65.70	-	20.05	25.23	24.51
25h_aug2	-	72.21	72.52	**81.31**	-	34.21	**43.31**	38.77
50h	-	83.05	84.00	**86.30**	-	31.79	**37.28**	36.34
50h_aug1	-	76.09	76.46	77.85	-	26.85	31.27	32.05
50h_aug2	-	86.17	86.54	**88.57**	-	37.64	**44.61**	41.69
75h	-	88.56	88.05	**90.06**	-	35.26	**40.15**	38.43
75h_aug1	-	87.54	87.83	89.25	-	40.56	**44.94**	44.39
75h_aug2	-	89.96	90.09	**91.86**	-	38.61	**44.86**	42.98
106h	-	89.68	90.62	**91.05**	-	34.23	37.51	**37.59**
106h_aug1	-	88.74	89.25	90.31	-	41.62	**46.74**	46.54
106h_aug2	-	91.52	91.83	**92.26**	-	39.59	**45.36**	45.02
25h	58.00	58.21	61.70	65.92	27.98	21.84	31.05	36.11

x-vector	25h_aug2	60.01	61.37	62.05	**73.06**	27.27	24.56	33.57	**36.98**
	50h	72.77	66.08	76.76	82.34	30.07	22.45	35.56	37.44
	50h_aug2	75.51	67.32	76.89	**85.56**	30.00	24.58	36.43	**38.86**
	75h	78.92	72.11	82.87	87.12	28.60	24.36	37.76	39.05
	75h_aug2	82.40	73.66	83.07	**89.06**	27.97	24.74	37.97	**39.75**
	106h	81.66	72.42	86.34	89.83	28.73	23.44	35.48	38.43
	106h_aug2	71.70	53.09	89.89	**94.18**	17.24	31.03	**43.73**	39.44

Conclusion：

（1）从集内数据的测试结果来看，aug1 情况下的识别率普遍比原始数据下的识别率低，由此可猜测噪音破坏了 aug1 中的原始数据。

（2）从总体实验结果看，通过 aug2 的数据增强之后的识别率普遍比原始数据下的识别率高，说明数据增强有利于提高小数据&跨信道的识别率。并且在 25h-ivector 系统中，集内识别率提高了 7.64，集外识别率提高了 11.37。以下表格是 aug2 相对于原始数据的识别率提升情况。

	dur	in-domain	Out-domain
i-vector	25h	**7.64**	**11.37**
	50h	2.27	**7.33**
	75h	1.80	**4.79**
	106h	1.21	**9.15**
x-vector	25h	7.14	0.87
	50h	**3.22**	1.42
	75h	**1.94**	0.70
	106h	**4.35**	5.3

集内时，25h 情况下 aug2 方法对 i-vector 性能提升较为明显。其它时长时对 x-vector 性能提升明显。但在集外时，aug2 对 i-vector 性能提高较大。

3. **解决方法 2---phonetic knowledge**:加入共享语言信息 with the help of ASR system to feed phonetic information.

ASR model: chain model
Bnf layer: （1）out-linear / （2）out-xent-linear[output-dim=256]

---

bn1: output.config = output-xent.linear

bn2: output.config = output.linear

本实验中的输出层选取了 2 层做对比,第一种是 out-linear 层,第二种 out-xent-linear 层，其输出都是 256 维，且都为最后一层的线性层。

实验流程：fbank[40]→Asr model[256]→ → xvector

将 40 维 fbank 特征输入值 ASR model 中提取 BN 特征，再将提取出的 BN 特征按常规方式输入至 x-vector 模型中进行语种识别。

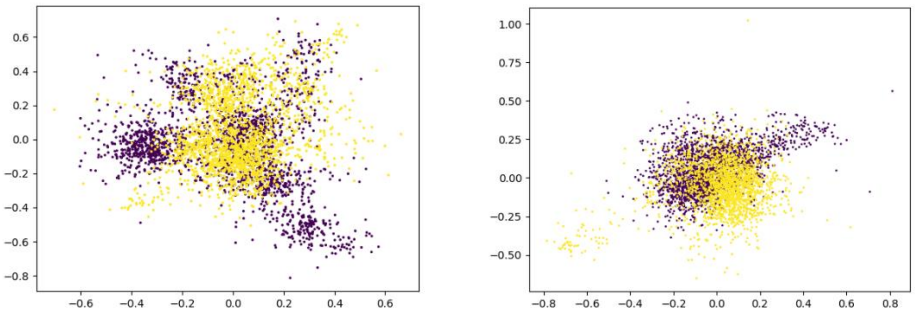
RESULT:

		in-domain				Out-domain			
		pos	cos	lr	lda-plda	pos	cos	lr	lda-plda
x-vector	25h	58.00	58.21	61.70	65.92	27.98	21.84	31.05	36.11
	25h_bn1	**96.89**	96.72	96.81	96.19	61.79	58.06	**64.51**	59.75
	25h_bn2	96.40	96.08	96.51	96.11	52.29	51.00	53.87	53.37
	50h	72.77	66.08	76.76	82.34	30.07	22.45	35.56	37.44
	50h_bn1	98.31	98.15	98.53	**98.72**	60.11	60.99	**64.53**	63.30
	50h_bn2	98.44	98.30	98.63	98.50	49.68	52.88	57.29	56.85
	75h	78.92	72.11	82.87	87.12	28.60	24.36	37.76	39.05
	75h_bn1	98.76	98.55	98.65	98.49	64.31	63.18	**66.40**	65.67
	75h_bn2	98.98	98.75	**99.10**	98.85	52.53	53.90	56.63	53.51
	106h	81.66	72.42	86.34	89.83	28.73	23.44	35.48	38.43
	106h_bn1	98.86	98.83	98.91	**99.07**	**70.71**	67.99	68.99	65.19
	106h_bn2	98.97	98.91	99.07	98.79	60.04	56.45	56.90	53.20

结论:

（1） 从实验结果看：在集内测试集中，25h 的 BN 特征比 fbank 特征识别率高了近 40%，且与 106h 的识别率只相差 2%左右。似乎 BN 特征已经近似解决了小数据问题。在集外测试集中，同时长的训练集，BN 特征比 fbank 特征的识别率最高提升了 32.28%。由此可知，加入共享语言信息极大解决了小数据&跨信道问题。

（2） 实验结果可看出，bn1 比 bn2 解决小数据跨信道的能力更好。理由是？
分析：



x-vectors　　bnf-xvectors

fig 4

左边是加入共享语言信息之前的结果，右边是之后的。

从图中可以看出原本 in-domain 与 out-of-domain 数据的边缘分布不同，及数据整体不相似。但是通过 DNN-ASR 结构之后，out-of-domain 与 in-domain 的距离靠近，可视为近似投影到了一个子空间中。所以想要提升跨信道情况下的性能，要将两个不同的域投影到一个公共子空间中，以消除域不匹配。

4. 2+3 融合方法 1、方法 2

		in-domain				Out-domain			
		pos	cos	lr	lda-plda	pos	cos	lr	lda-plda
x-vector	25h	58.00	58.21	61.70	65.92	27.98	21.84	31.05	36.11
	25h_bn1	96.89	96.72	96.81	96.19	61.79	58.06	64.51	59.75
	25h_aug2_bn1	96.25	**97.11**	96.88	96.47	**62.85**	60.30	62.66	61.17
	50h	72.77	66.08	76.76	82.34	30.07	22.45	35.56	37.44
	50h_bn1	98.31	98.15	98.53	98.72	60.11	60.99	64.53	63.30
	50h_aug2_bn1								
	75h	78.92	72.11	82.87	87.12	28.60	24.36	37.76	39.05
	75h_bn1	98.76	98.55	98.65	98.49	64.31	63.18	66.40	65.67
	75h_aug2_bn1								
	106h	81.66	72.42	86.34	89.83	28.73	23.44	35.48	38.43
	106h_bn1	98.86	98.83	98.91	99.07	70.71	67.99	68.99	65.19
	106h_aug2_bn1								

5. ... thinking ...

（1）将训练集加入集外数据

（2）迁移学习

（3）生成式模型