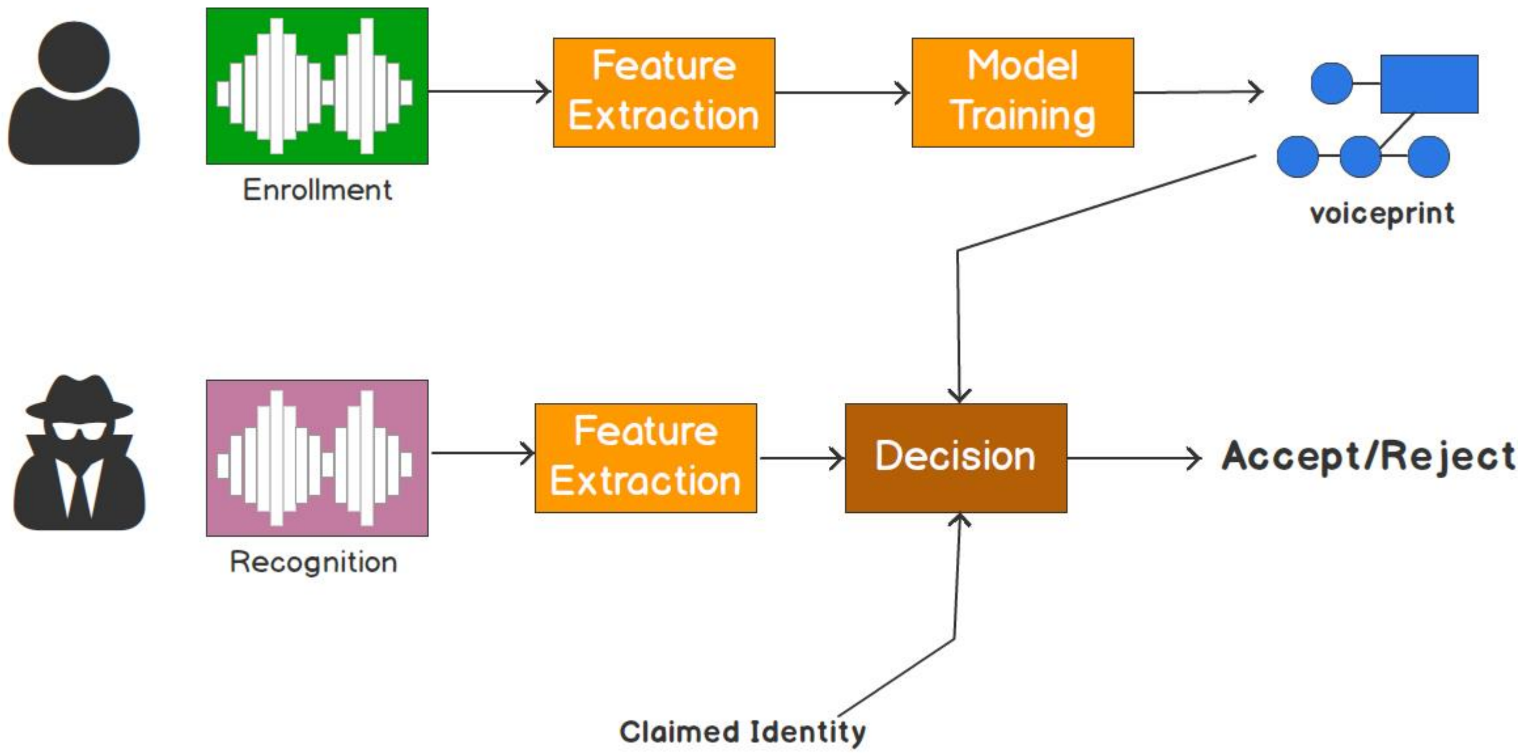
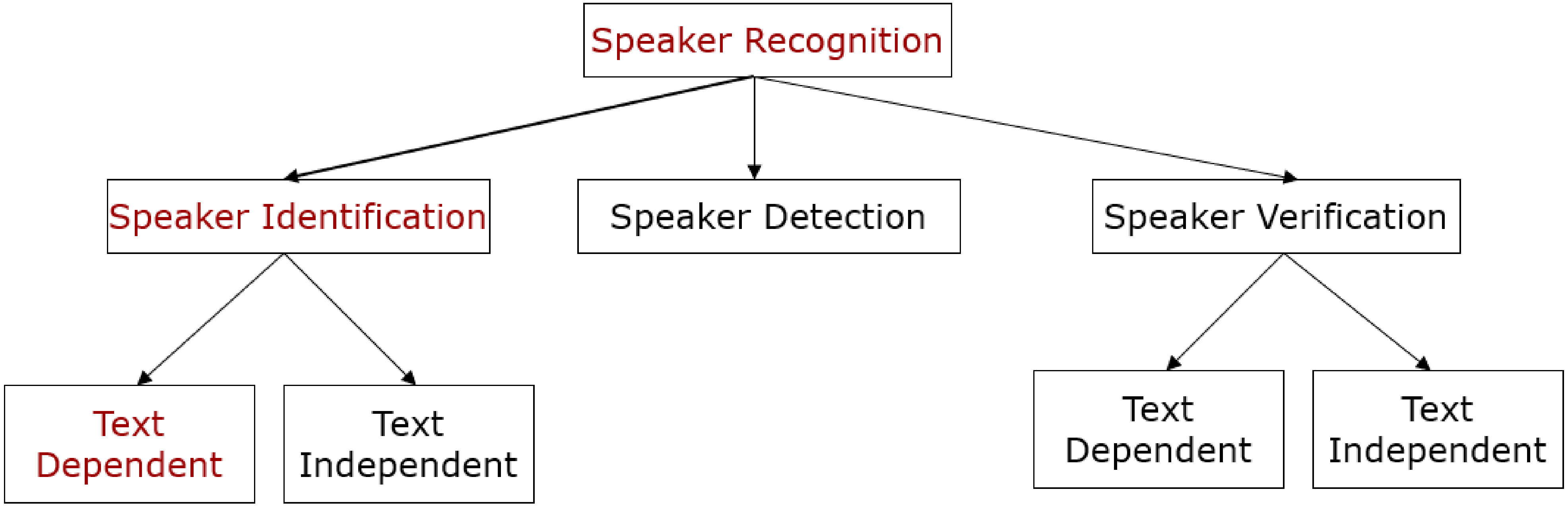
x-vector论文阅读和声纹项目进展

ADSPLAB
张皓然

2019/11/21


Enrollment
Feature Extraction
Model Training
voiceprint
Recognition
Feature Extraction
Decision
Accept/Reject
Claimed Identity

Speaker recognition is the process of automatically recognizing who is speaking by using the speaker-specific information included in speech waves to verify identities being claimed by people accessing systems;



Deep Neural Network Embeddings for Text-Independent Speaker Verification

Definition：Speaker verification (SV) is the task of authenticating the claimed identity of a speaker

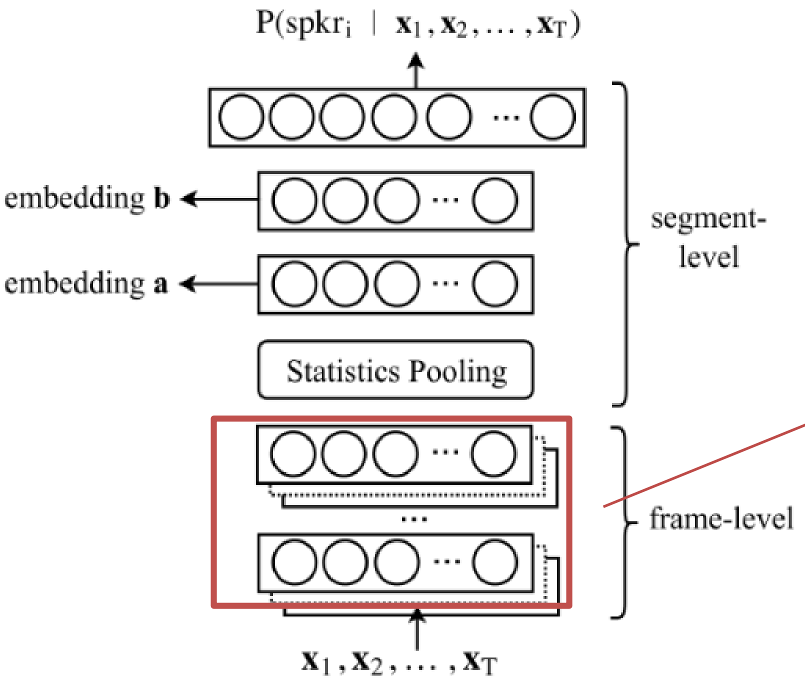
Process：utterances are mapped directly to fixed-dimensional speaker embeddings
and pairs of embeddings are scored using a PLDA-based backend

x-vector，它已经成为了几乎所有的Challenges和papers的新baseline，SRE18上主办方默认使用xvector
作为基线（如果说话人识别任务的训练集并不是很充足时，x-vector一类的embedding很可能会过拟合）

美国国家标准技术署（NIST）主办的说话人识别技术评测
（Speaker Recognition Evaluation， SRE）

Deep Neural Network Embeddings for Text-Independent Speaker Verification

训练思路：阶段一（对神经网络的训练）



Layer	Layer context	Total context	Input x output
frame1	$[t-2, t+2]$	5	120x512
frame2	$\{t-2, t, t+2\}$	9	1536x512
frame3	$\{t-3, t, t+3\}$	15	1536x512
frame4	$\{t\}$	15	512x512
frame5	$\{t\}$	15	512x1500
stats pooling	$[0, T)$	T	1500Tx3000
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512
softmax	$\{0\}$	T	512xN

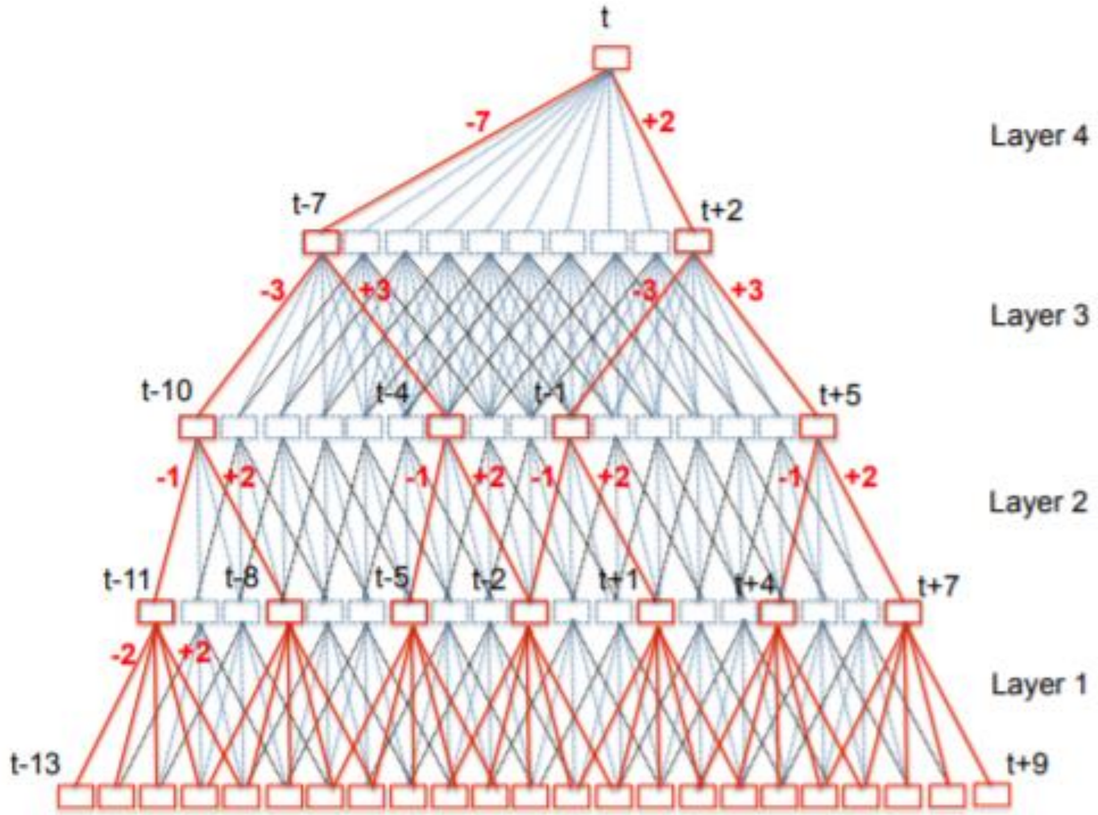
multiclass cross entropy

$$E = -\sum_{n=1}^{N}\sum_{k=1}^{K} d_{nk} ln(P(spkr_k \mid \mathbf{x}_{1:T}^{(n)}))$$

pooling层之前的结构是TDNN，pooling层之后接着两层全向连接层最后加一个softmax层为输出。输出的神经元个数和我们训练集中说话人个数保持一致。可以看到图中所写，输出是一个后验概率。

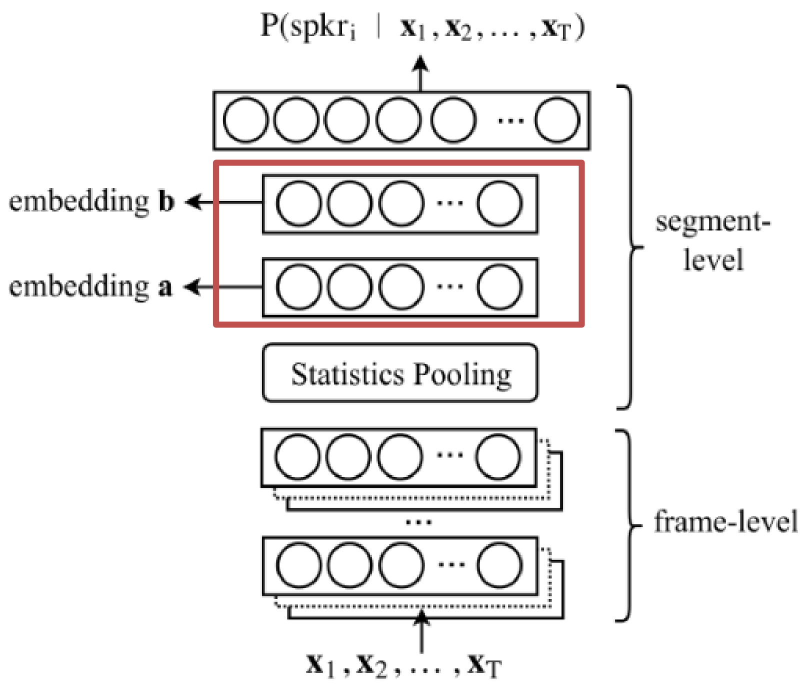
Deep Neural Network Embeddings for Text-Independent Speaker Verification

Layer	Layer context	Total context	Input x output
frame1	$[t-2, t+2]$	5	120x512
frame2	$\{t-2, t, t+2\}$	9	1536x512
frame3	$\{t-3, t, t+3\}$	15	1536x512
frame4	$\{t\}$	15	512x512
frame5	$\{t\}$	15	512x1500
stats pooling	$[0, T)$	T	1500Tx3000
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512
softmax	$\{0\}$	T	512xN



Deep Neural Network Embeddings for Text-Independent Speaker Verification

训练思路：阶段二（PLDA）



去掉已经训练好的神经网络的softmax层
We do not consider the presoftmax affine layer because of its large size and dependence on the number of speakers
利用这些embeddings训练PLDA模型

Deep Neural Network Embeddings for Text-Independent Speaker Verification

实验：

Table 1: *EER(%) on NIST SRE10*

	10s-10s	5s	10s	20s	60s	full
ivector	11.0	9.1	6.0	3.9	2.3	1.9
embedding a	11.0	9.5	5.7	3.9	3.0	2.6
embedding b	9.2	8.8	6.6	5.5	4.4	3.9
embeddings	7.9	7.6	5.0	3.8	2.9	2.6
fusion	8.1	6.8	4.3	2.9	2.1	1.8

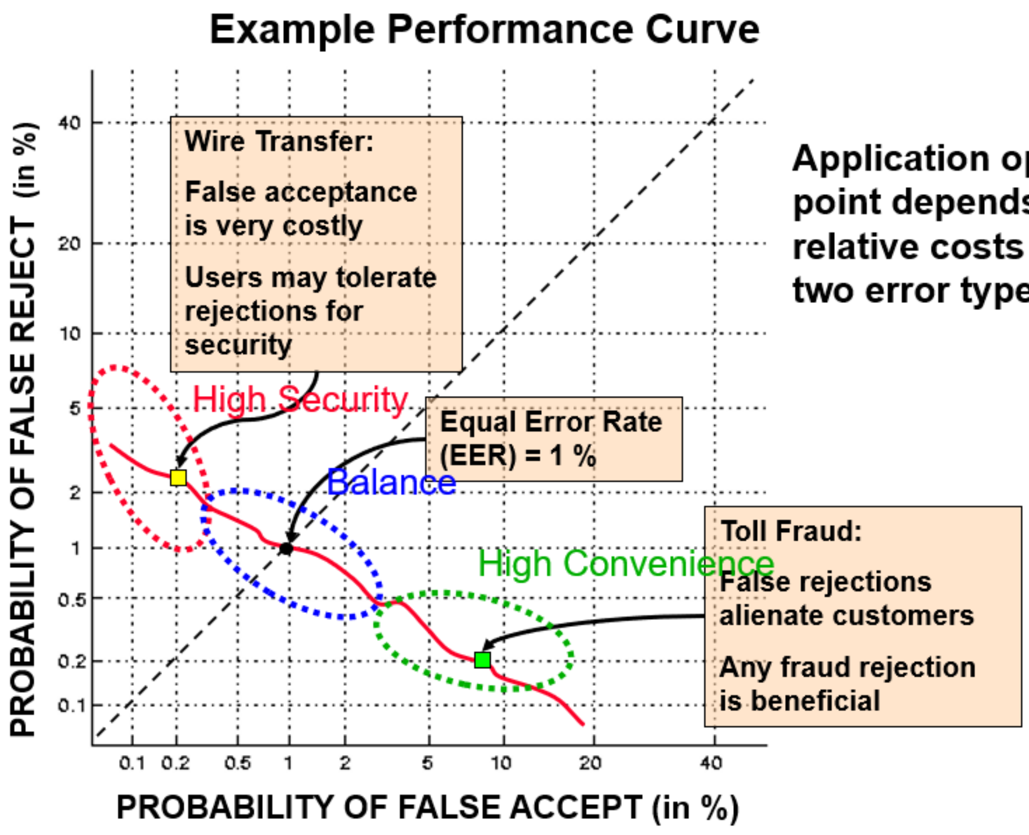
Table 3: *EER(%) on NIST SRE16*

	Cantonese	Tagalog	pool
ivector	8.3	17.6	13.6
embedding a	7.7	17.6	13.1
embedding b	7.8	17.4	13.1
embeddings	6.5	16.3	11.9
fusion	6.3	15.4	11.3

X-vector are better on the short duration conditions
X-vector may be more robust to this domain mismatch

SRE16 presented the challenge of language mismatch between the predominantly English training data and the Cantonese and Tagalog evaluation.

EER(equal error rate) :调整阈值，使得误拒绝率（False Rejection Rate，FRR）等于误接受率（False Acceptance Rate，FAR），此时的FAR与FRR的值称为等错误率。

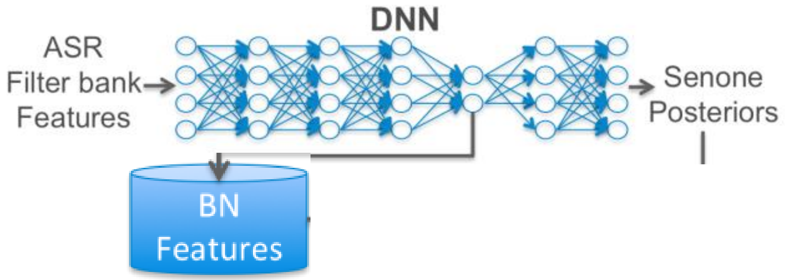


Application operating point depends on relative costs of the two error types

X-VECTORS: ROBUST DNN EMBEDDINGS FOR SPEAKER RECOGNITION

Daniel Povey

			SITW Core			SRE16 Cantonese		
			EER(%)	DCF10^{-2}	DCF10^{-3}	EER(%)	DCF10^{-2}	DCF10^{-3}
4.1	Original systems	i-vector (acoustic)	9.29	0.621	0.785	9.23	0.568	0.741
		i-vector (BNF)	**9.10**	**0.558**	**0.719**	9.68	0.574	0.765
		x-vector	9.40	0.632	0.790	**8.00**	**0.491**	**0.697**
4.2	PLDA aug.	i-vector (acoustic)	8.64	0.588	0.755	8.92	0.544	0.717
		i-vector (BNF)	8.00	**0.514**	**0.689**	8.82	0.532	0.726
		x-vector	**7.56**	0.586	0.746	**7.45**	**0.463**	**0.669**
4.3	Extractor aug.	i-vector (acoustic)	8.89	0.626	0.790	9.20	0.575	0.748
		i-vector (BNF)	7.27	**0.533**	0.730	8.89	0.569	0.777
		x-vector	**7.19**	0.535	**0.719**	**6.29**	**0.428**	**0.626**
4.4	PLDA and extractor aug.	i-vector (acoustic)	8.04	0.578	0.752	8.95	0.555	0.720
		i-vector (BNF)	6.49	0.492	0.690	8.29	0.534	0.749
		x-vector	**6.00**	**0.488**	**0.677**	**5.86**	**0.410**	**0.593**
4.5	Incl. VoxCeleb	i-vector (acoustic)	7.45	0.552	0.723	9.23	0.557	0.742
		i-vector (BNF)	6.09	0.472	0.660	8.12	0.523	0.751
		x-vector	**4.16**	**0.393**	**0.606**	**5.71**	**0.399**	**0.569**



Detection cost function (DCF) =

- P(FR) C(FR) P(target) + P(FA) C(FA) (1-P(target))

			SITW Core			SRE16 Cantonese		
			EER(%)	$DCF10^{-2}$	$DCF10^{-3}$	EER(%)	$DCF10^{-2}$	$DCF10^{-3}$
4.1	Original systems	i-vector (acoustic)	9.29	0.621	0.785	9.23	0.568	0.741
		i-vector (BNF)	**9.10**	**0.558**	**0.719**	9.68	0.574	0.765
		x-vector	9.40	0.632	0.790	**8.00**	**0.491**	**0.697**
4.2	PLDA aug.	i-vector (acoustic)	8.64	0.588	0.755	8.92	0.544	0.717
		i-vector (BNF)	8.00	**0.514**	**0.689**	8.82	0.532	0.726
		x-vector	**7.56**	0.586	0.746	**7.45**	**0.463**	**0.669**
4.3	Extractor aug.	i-vector (acoustic)	8.89	0.626	0.790	9.20	0.575	0.748
		i-vector (BNF)	7.27	**0.533**	0.730	8.89	0.569	0.777
		x-vector	**7.19**	0.535	**0.719**	**6.29**	**0.428**	**0.626**
4.4	PLDA and extractor aug.	i-vector (acoustic)	8.04	0.578	0.752	8.95	0.555	0.720
		i-vector (BNF)	6.49	0.492	0.690	8.29	0.534	0.749
		x-vector	**6.00**	**0.488**	**0.677**	**5.86**	**0.410**	**0.593**
4.5	Incl. VoxCeleb	i-vector (acoustic)	7.45	0.552	0.723	9.23	0.557	0.742
		i-vector (BNF)	6.09	0.472	0.660	8.12	0.523	0.751
		x-vector	**4.16**	**0.393**	**0.606**	**5.71**	**0.399**	**0.569**

4.1. Original systems

This echoes recent studies that have
found that the large gains achieved by BNFs in English speech
are not necessarily transferable to non-English data

			SITW Core			SRE16 Cantonese		
			EER(%)	DCF10^{-2}	DCF10^{-3}	EER(%)	DCF10^{-2}	DCF10^{-3}
4.1	Original systems	i-vector (acoustic)	9.29	0.621	0.785	9.23	0.568	0.741
		i-vector (BNF)	**9.10**	**0.558**	**0.719**	9.68	0.574	0.765
		x-vector	9.40	0.632	0.790	**8.00**	**0.491**	**0.697**
4.2	PLDA aug.	i-vector (acoustic)	8.64	0.588	0.755	8.92	0.544	0.717
		i-vector (BNF)	8.00	**0.514**	**0.689**	8.82	0.532	0.726
		x-vector	**7.56**	0.586	0.746	**7.45**	**0.463**	**0.669**
4.3	Extractor aug.	i-vector (acoustic)	8.89	0.626	0.790	9.20	0.575	0.748
		i-vector (BNF)	7.27	**0.533**	0.730	8.89	0.569	0.777
		x-vector	**7.19**	0.535	**0.719**	**6.29**	**0.428**	**0.626**
4.4	PLDA and extractor aug.	i-vector (acoustic)	8.04	0.578	0.752	8.95	0.555	0.720
		i-vector (BNF)	6.49	0.492	0.690	8.29	0.534	0.749
		x-vector	**6.00**	**0.488**	**0.677**	**5.86**	**0.410**	**0.593**
4.5	Incl. VoxCeleb	i-vector (acoustic)	7.45	0.552	0.723	9.23	0.557	0.742
		i-vector (BNF)	6.09	0.472	0.660	8.12	0.523	0.751
		x-vector	**4.16**	**0.393**	**0.606**	**5.71**	**0.399**	**0.569**

4.2. PLDA augmentation

4.3. Extractor augmentation

We use a 3-fold augmentation that combines the original "clean" training list with two augmented copies. To augment a recording, we choose between one of the following randomly:

- **babble**: Three to seven speakers are randomly picked from MUSAN speech, summed together, then added to the original signal (13-20dB SNR).
- **music**: A single music file is randomly selected from MUSAN, trimmed or repeated as necessary to match duration, and added to the original signal (5-15dB SNR).
- **noise**: MUSAN noises are added at one second intervals throughout the recording (0-15dB SNR).
- **reverb**: The training recording is artificially reverberated via convolution with simulated RIRs.

			SITW Core			SRE16 Cantonese		
			EER(%)	DCF10^{-2}	DCF10^{-3}	EER(%)	DCF10^{-2}	DCF10^{-3}
4.1	Original systems	i-vector (acoustic)	9.29	0.621	0.785	9.23	0.568	0.741
		i-vector (BNF)	**9.10**	**0.558**	**0.719**	9.68	0.574	0.765
		x-vector	9.40	0.632	0.790	**8.00**	**0.491**	**0.697**
4.2	PLDA aug.	i-vector (acoustic)	8.64	0.588	0.755	8.92	0.544	0.717
		i-vector (BNF)	8.00	**0.514**	**0.689**	8.82	0.532	0.726
		x-vector	**7.56**	0.586	0.746	**7.45**	**0.463**	**0.669**
4.3	Extractor aug.	i-vector (acoustic)	8.89	0.626	0.790	9.20	0.575	0.748
		i-vector (BNF)	7.27	**0.533**	0.730	8.89	0.569	0.777
		x-vector	**7.19**	0.535	**0.719**	**6.29**	**0.428**	**0.626**
4.4	PLDA and extractor aug.	i-vector (acoustic)	8.04	0.578	0.752	8.95	0.555	0.720
		i-vector (BNF)	6.49	0.492	0.690	8.29	0.534	0.749
		x-vector	**6.00**	**0.488**	**0.677**	**5.86**	**0.410**	**0.593**
4.5	Incl. VoxCeleb	i-vector (acoustic)	7.45	0.552	0.723	9.23	0.557	0.742
		i-vector (BNF)	6.09	0.472	0.660	8.12	0.523	0.751
		x-vector	**4.16**	**0.393**	**0.606**	**5.71**	**0.399**	**0.569**

PLDA and extractor augmentation

On SITW the x-vectors are now 10-25% better than i-vector (acoustic) and are slightly better than i-vector (BNF) at all operating points. On SRE16 Cantonese,the x-vectors continue to maintain the large lead over the i-vector systems established in Section

			SITW Core			SRE16 Cantonese		
			EER(%)	DCF10^{-2}	DCF10^{-3}	EER(%)	DCF10^{-2}	DCF10^{-3}
4.1	Original systems	i-vector (acoustic)	9.29	0.621	0.785	9.23	0.568	0.741
		i-vector (BNF)	**9.10**	**0.558**	**0.719**	9.68	0.574	0.765
		x-vector	9.40	0.632	0.790	**8.00**	**0.491**	**0.697**
4.2	PLDA aug.	i-vector (acoustic)	8.64	0.588	0.755	8.92	0.544	0.717
		i-vector (BNF)	8.00	**0.514**	**0.689**	8.82	0.532	0.726
		x-vector	**7.56**	0.586	0.746	**7.45**	**0.463**	**0.669**
4.3	Extractor aug.	i-vector (acoustic)	8.89	0.626	0.790	9.20	0.575	0.748
		i-vector (BNF)	7.27	**0.533**	0.730	8.89	0.569	0.777
		x-vector	**7.19**	0.535	**0.719**	**6.29**	**0.428**	**0.626**
4.4	PLDA and extractor aug.	i-vector (acoustic)	8.04	0.578	0.752	8.95	0.555	0.720
		i-vector (BNF)	6.49	0.492	0.690	8.29	0.534	0.749
		x-vector	**6.00**	**0.488**	**0.677**	**5.86**	**0.410**	**0.593**
4.5	Incl. VoxCeleb	i-vector (acoustic)	7.45	0.552	0.723	9.23	0.557	0.742
		i-vector (BNF)	6.09	0.472	0.660	8.12	0.523	0.751
		x-vector	**4.16**	**0.393**	**0.606**	**5.71**	**0.399**	**0.569**

Including VoxCeleb
the x-vector exploits the large increase in the amount of in-domain data better than the i-vector systems.

conclusions

- We found that data augmentation is an easily implemented and effective strategy for improving their performance.

- More generally, it appears that x-vectors are now a strong contender for next-generation representations for speaker recognition.

- Our software framework has been made available in the Kaldi toolkit. An example recipe is in the main branch of Kaldi at https: //github.com/kaldi-asr/kaldi/tree/master/egs/ sre16/v2 and a pretrained x-vector system can be downloaded from http://kaldi-asr.org/models.html.

AOTO_Speaker_Verification 验证（BiGRU_DNN)

注册语音(zhr) -> deep_feature1 (1x512)

```
deep_feature_1 = {ndarray} [ -4.63145995  16.28163576  16.49755669   7.97190905   3.70570898\n  10.
```

注册语音：时长>3s，采样率16000
测试语音：时长>3s，采样率16000

测试语音(非zhr) -> deep_ feature2 (1x512)

```
deep_feature_2 = {ndarray} [  8.89405799  11.43006706 -11.15161753  -3.22878671  -4.29865265\n
```

```
Final Score is -0.030330123949945036 and threshold is 0.1959
Different Person
```

测试语音(zhr) -> deep_ feature2 (1x512)

```
deep_feature_2 = {ndarray} [ -5.69877028   8.70336246  17.89974022  10.87302303  -9.84257126\n  -6.
```

```
Final Score is 0.5571417215548106 and threshold is 0.1959
Same Person
```

输入测试语音到输出结果耗时：

```
1.2009568214416504 s
```

SPKID(xvector)

Aoto speaker 数据描述： 145个说话人，每人6句语音(有缺失)，文本相关（ ‘帮我开个门’ ）， 2s~3s　(50.6M)

训练集：每人5条语音
测试集：每人1条语音

```
epoch 0, loss_tr=12.058552 err_tr=1.000000 loss_te=4.946517 err_te=0.950021 err_te_snt=0.946154
epoch 1, loss_tr=9.243673 err_tr=0.997500 loss_te=4.926078 err_te=0.875500 err_te_snt=0.869231
epoch 2, loss_tr=6.998516 err_tr=0.972187 loss_te=4.903053 err_te=0.786025 err_te_snt=0.761538
epoch 3, loss_tr=4.971113 err_tr=0.863437 loss_te=4.887064 err_te=0.780013 err_te_snt=0.769231
epoch 4, loss_tr=3.565307 err_tr=0.712187 loss_te=4.869194 err_te=0.721533 err_te_snt=0.676923
epoch 5, loss_tr=2.406699 err_tr=0.513750 loss_te=4.860543 err_te=0.710446 err_te_snt=0.692308
epoch 6, loss_tr=1.651469 err_tr=0.331562 loss_te=4.851987 err_te=0.685518 err_te_snt=0.669231
epoch 7, loss_tr=1.216467 err_tr=0.229375 loss_te=4.836324 err_te=0.635470 err_te_snt=0.630769
epoch 8, loss_tr=0.804816 err_tr=0.126562 loss_te=4.844227 err_te=0.666138 err_te_snt=0.661538
epoch 9, loss_tr=0.685921 err_tr=0.099687 loss_te=4.840653 err_te=0.636263 err_te_snt=0.615385
epoch 10, loss_tr=0.496485 err_tr=0.063437 loss_te=4.825949 err_te=0.632245 err_te_snt=0.615385
epoch 11, loss_tr=0.403659 err_tr=0.051562 loss_te=4.838764 err_te=0.673383 err_te_snt=0.653846
epoch 12, loss_tr=0.384862 err_tr=0.045937 loss_te=4.824498 err_te=0.640017 err_te_snt=0.630769
epoch 13, loss_tr=0.327981 err_tr=0.037187 loss_te=4.822965 err_te=0.610876 err_te_snt=0.600000
epoch 14, loss_tr=0.309444 err_tr=0.034063 loss_te=4.822235 err_te=0.633240 err_te_snt=0.623077
epoch 15, loss_tr=0.295286 err_tr=0.033750 loss_te=4.818516 err_te=0.641127 err_te_snt=0.607692
epoch 16, loss_tr=0.216414 err_tr=0.020625 loss_te=4.806367 err_te=0.634496 err_te_snt=0.615385
epoch 17, loss_tr=0.187519 err_tr=0.021250 loss_te=4.799305 err_te=0.616913 err_te_snt=0.600000
epoch 18, loss_tr=0.282146 err_tr=0.039375 loss_te=4.819955 err_te=0.627702 err_te_snt=0.607692
epoch 19, loss_tr=0.197313 err_tr=0.020312 loss_te=4.810517 err_te=0.605578 err_te_snt=0.592308
epoch 20, loss_tr=0.189448 err_tr=0.022187 loss_te=4.806847 err_te=0.618051 err_te_snt=0.607692
```

预想计划：使用aishell数据集进行与训练，再用aoto数据集fine-tune

VGGVox speaker identification

VoxCeleb1(40G) VoxCeleb2(80G)

Layer	Support	Filt dim.	# filts.	Stride	Data size
conv1	7×7	1	96	2×2	254×148
mpool1	3×3	-	-	2×2	126×73
conv2	5×5	96	256	2×2	62×36
mpool2	3×3	-	-	2×2	30×17
conv3	3×3	256	384	1×1	30×17
conv4	3×3	384	256	1×1	30×17
conv5	3×3	256	256	1×1	30×17
mpool5	5×3	-	-	3×2	9×8
fc6	9×1	256	4096	1×1	1×8
apool6	1×n	-	-	1×1	1×1
fc7	1×1	4096	1024	1×1	1×1
fc8	1×1	1024	1251	1×1	1×1

Dataset	VoxCeleb1	VoxCeleb2
# of POIs	1,251	6,112
# of male POIs	690	3,761
# of videos	22,496	150,480
# of hours	352	2,442
# of utterances	153,516	1,128,246
Avg # of videos per POI	18	25
Avg # of utterances per POI	116	185
Avg length of utterances (s)	8.2	7.8

test_file	test_speaker	1	2	3	result	correct
data/wav/test/1-3.wav	1	0.782583	0.891363	0.805392	1	1
data/wav/test/2-3.wav	2	0.963036	0.459733	0.713374	2	1
data/wav/test/3-3.wav	3	0.553141	0.590193	0.496234	3	1

7,000 +

speakers

VoxCeleb contains speech from speakers spanning a wide range of different ethnicities, accents, professions and ages.

1 million +

utterances

All speaking face-tracks are captured "in the wild", with background chatter, laughter, overlapping speech, pose variation and different lighting conditions.

2,000 +

hours

VoxCeleb consists of both audio and video. Each segment is at least 3 seconds long.

工作预想：使用aoto数据微调VGGVox模型参数 使用kaldi

Thanks for your attention!