AP16-OL7: A Multilingual Database for Oriental Languages and A Language Recognition Baseline

Dong Wang, Lantian Li, Difei Tang and Qing Chen

CSLT/RIIT, Tsinghua University & SpeechOcean, Inc.

http://cslt.riit.tsinghua.edu.cn

IEEE APSIPA ASC, Dec. 13-16, 2016, *Jeju, Korea*

OUTLINE

> > > > >

1 AP16-OLR Challenge

2 AP16-OL7 Database

Oriental Languages

- Language families
  - Austroasiatic languages (e.g., Vietnamese, Cambodia)
  - TaiKadai languages (e.g, Thai, Lao)
  - Hmong-Mien languages (e.g., dialects in south China)
  - Altaic languages (e.g., Korea, Japanese)
  - Indo-European languages (e.g., Russian)
  - …
- Characteristics
  - Complex acoustic and linguistic patterns
  - International interaction
  - Cultural integration

Multilingual research for OL

- Interesting areas for OL
  - Comparative phonetics
  - Evolutionary linguistics
  - Second language acquisition
  - Social linguistics
  - Mixlingual and multilingual phenomena
- OL Multilingual speech and language processing
  - Thanks to SpeechOcean and APSIPA ASC 2016!

AP16-OLR Challenge

- Oriental language recognition (OLR) challenge
  - Given a segment of speech and a language hypothesis, the task is to decide whether that target language was in fact spoken in the given segment (yes or no).
  - Tried to following the evaluation metric of LRE15.
- **Collaboration between research, commercial companies and data providers!**
- Participants
  - Academic plus industrial
- Resources
  - AP16-OL7 database (thanks to SpeechOcean)
    - Development set
    - Evaluation set

AP16-OLR Challenge

- Evaluation metric
  - Cavg

$$C(L_t, L_n) = P_{Target} P_{Miss}(L_t) + (1 - P_{Target}) P_{FA}(L_t, L_n)$$

$$C_{avg} = \frac{1}{N} \sum_{L_t} \left\{ \begin{array}{l} P_{Target} \cdot P_{Miss}(L_t) \\ + \sum\limits_{L_n} P_{Non-Target} \cdot P_{FA}(L_t, L_n) \end{array} \right\}$$

  - EER and *min*DCF

$$C_{Det} = C_{\text{miss}} \times P_{\text{miss}} \times P_{\text{Target}} + C_{\text{FalseAlarm}} \times P_{\text{FlaseAlarm}} \times \left(1 - P_{\text{Target}}\right)$$

  - IDR (Identification rate)

$$IDR = \frac{T_c}{T_c + T_i}$$

Baseline system

- GMM i-vector
  - 19-dim MFCCs + Δ + ΔΔ
  - 2,048 Gaussian components
  - 400-dim i-vector
  - 6-dim of LDA projection space
- Decision methods
  - Cosine distance scoring
    - The test i-vector and averaged language i-vector
  - SVM-based scoring
    - Kernel functions (Linear, Poly, RBF)

Baseline system

- Visualization with T-SNE

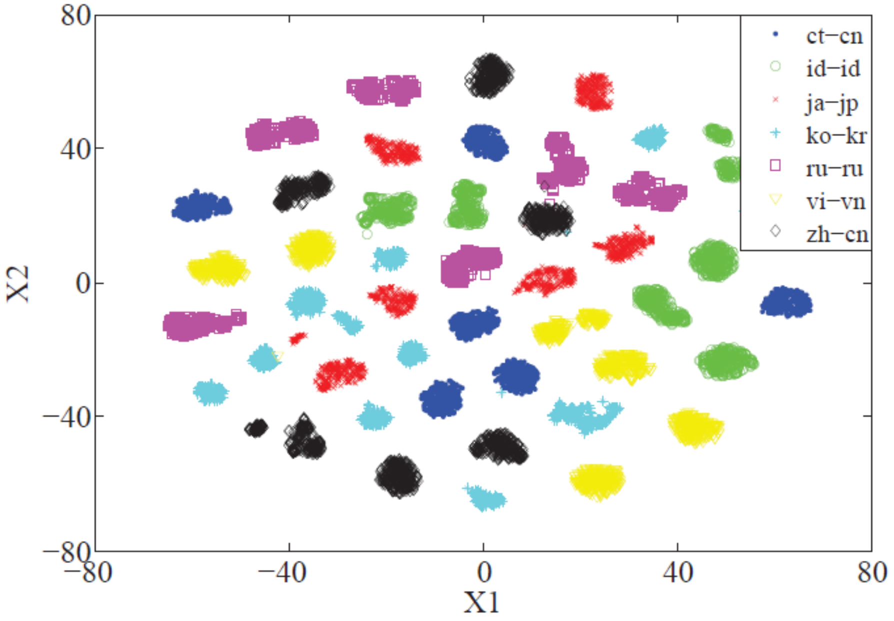


Fig. 1. Original i-vectors plotted by t-SNE. Each color/shape represents a particular language.

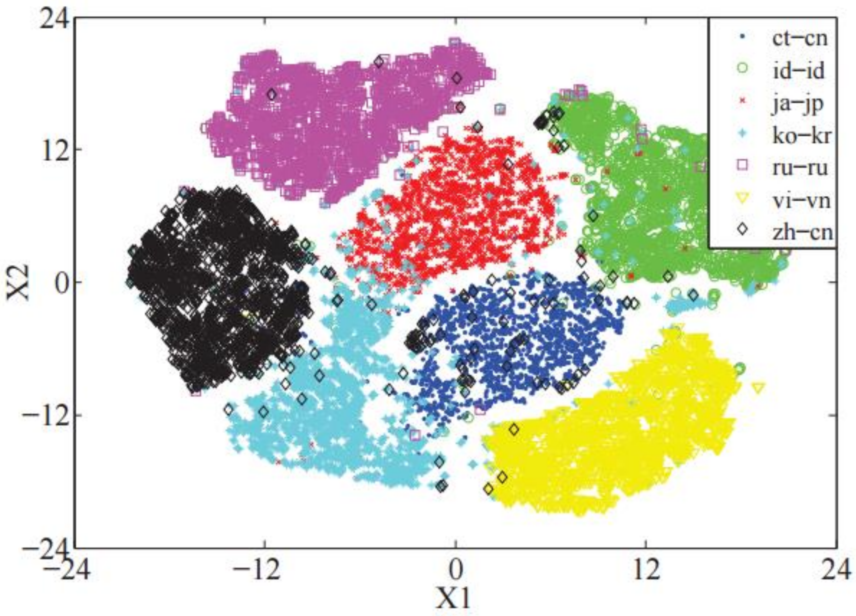


Fig. 2. LDA-transformed i-vectors plotted by t-SNE. Each color/shape represents a particular language.

Baseline results

System	*Cavg* *100	EER (%)	*min*DCF	IDR (%)
i-vector	5.63	6.65	0.0659	89.16
L-vector	4.15	4.76	0.0472	90.19
i-vector-SVM (Linear)	5.68	5.62	0.0558	87.07
i-vector-SVM(Poly)	**3.06**	**3.06**	**0.0303**	**92.73**
i-vector-SVM(RBF)	3.86	3.83	0.0381	90.80
L-vector-SVM(Linear)	3.52	3.49	0.0344	91.82
L-vector-SVM(Poly)	3.37	3.37	0.0334	91.99
L-vector-SVM(RBF)	3.40	3.36	0.0333	92.04

Challenge procedure

- June 06 AP16-OL7 training/dev data release
- July 30 AP16-OL7 test data release
- 12:00 pm, August 1, prior submission deadline
- 12:00 pm, Oct. 2, full submission deadline
- 12:00 pm, Dec, 10, extended submission deadline
- APSIPA 2016: challenge result release

Submissions

- More than 10 downloads, 8 submissions
- Only 1 prior submission (USTC), others are extended submissions
- Some extended submissions downloaded the data in late Nov., so the time usage was even less than the prior submission (e.g., NUS and I2R)

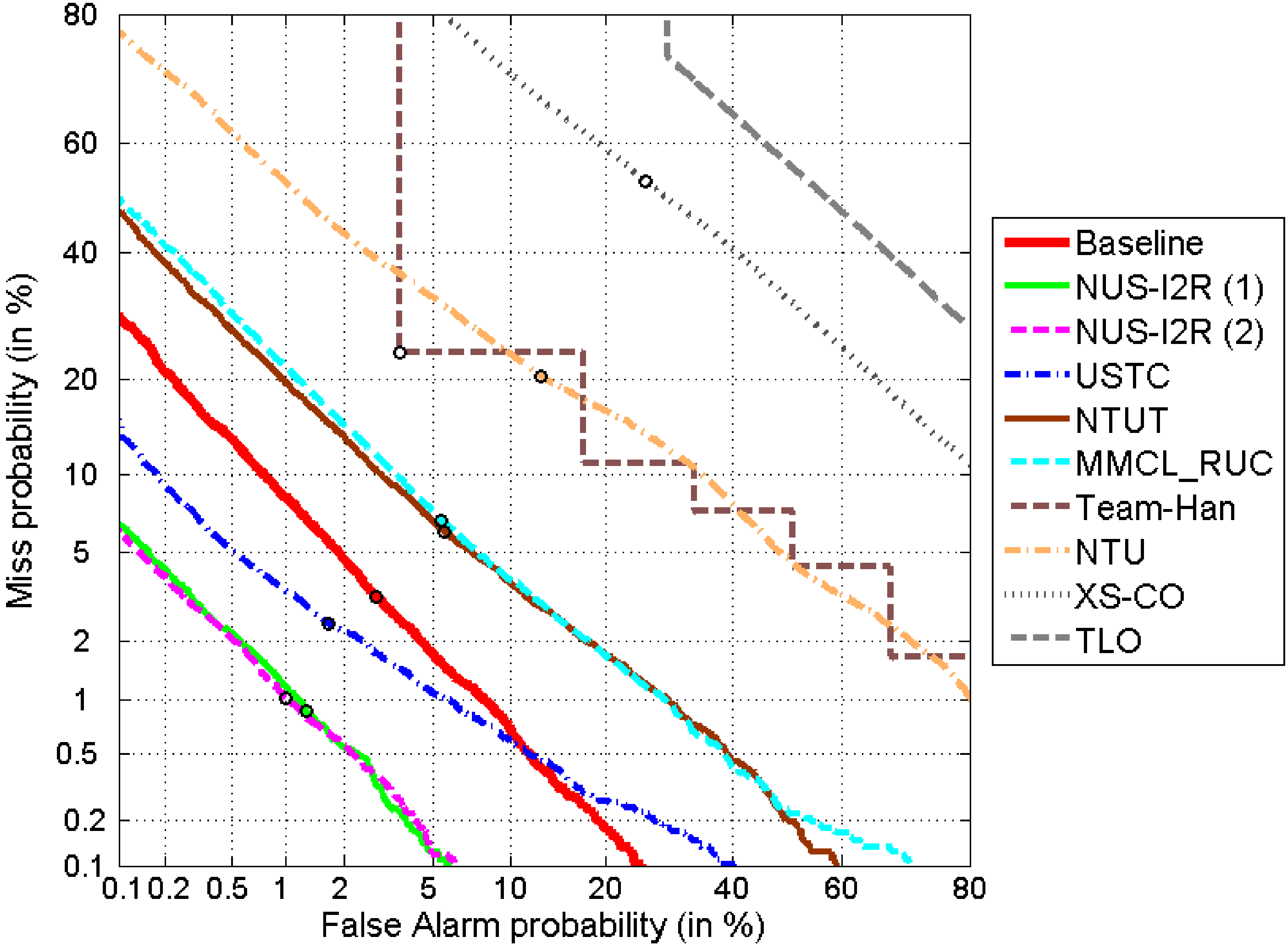
Submissions

Team	Main researchers
NUS and I2R, Singapore	Haizhou Li, Hanwu Sun, Kong Aik Lee, Nguyen Trung Hieu, Bin Ma
USTC, China	Wu Guo
NTUT, Taiwan, China	Yuanfu Liao , Sing-Yue Wang
MMCL_RUC, China	Haibing Cao, Qin Jin
PJ-Han, Germany	Anonymous
NTU, Singapore	Haihua Xu
XS-CO, China	Anonymous
TLO, China	Anonymous

Results

Rank	Team	*Cavg* *100	EER (%)	*min*DCF	IDR (%)
1	NUS and I2R (1), Singapore	1.13	1.09	0.0108	97.56
2	NUS and I2R (2), Singapore	1.70	1.02	0.0101	97.60
3	USTC, China	1.79	2.17	0.0205	96.94
4	NTUT, Taiwan, China	5.86	5.88	0.0586	87.02
5	MMCL_RUC, China	6.06	6.16	0.0610	86.21
6	PJ-Han, Germany	14.00	17.34	0.1365	77.65
7	NTU, Singapore	14.72	17.44	0.1657	71.44
8	XS-CO, China	36.99	40.26	0.3924	31.91
9	TLO, China	50.00	53.34	0.4999	12.37

- ***Red submissions are better than the baseline***
- ***USTC is the only prior submission***
- ***NUS and I2R systems used 40 days***


80
60
40
20
10
5
2
1
0.5
0.2
0.1
Miss probability (in %)
0.1
0.2
0.5
1
2
5
10
20
40
60
80
False Alarm probability (in %)
Baseline
NUS-I2R (1)
NUS-I2R (2)
USTC
NTUT
MMCL_RUC
Team-Han
NTU
XS-CO
TLO

Some findings

- i-vector plus SVM can perform pretty well on AP16-OL7
- The challenge seems not very ‘challenging’, as the test utterances are relative long
- More complicated tasks will be designed in OLR17.
- Thanks to all the participants, and congratulations to the rank winners!

OUTLINE

> > > > >

1 AP16-OLR Challenge

2 AP16-OL7 Database

What is AP16-OL7?

- It’s a speech database build up by Speechocean
- It contents 7 Oriental Languages
- About 71 hours
- All manually transcribed by native speakers
- Pronunciation Lexicon is available for each language
- It’s the first multilingual speech database designed for oriental languages

Parameters -1

Languages	• Mandarin in China • Cantonese in China Mainland & HK • Japanese in Japan • Korean in Korea • Russian in Russia • Indonesian in Indonesia • Vietnamese in Vietnam
Parameter	16 KHz, 16 Bit, Mono Channel
Script Design	Reading Style Dialog, SMS, SNS, Newspaper….

Parameters -2

Data Sets	Training Set	Testing Set	Total
No. of Speakers	18 / Language	6/ Language	168
Utterances	5K-7K/ Language	1.7K-2K/ Language	51779
Recording Hours	7-11Hrs./Language	2-3 Hrs. /Language	71.32

Recording Platform

Platform	Mobile Phone Model	Per. (%)
iOS	iPhone 3GS, iPhone 4	29.0%
Android	HTC Legend (G6), HTC Aria (G9), Samsung i909,	49.5%
	Samsung Nexus-S9020, HTC G18, MOTO XT615	
Windows Mobile	HTC t2222,Samsung i900	21.5%

Gender & Age Distribution

Age Group-Overall	# Speakers (%) - Overall
16 – 30 years	59 %
31 – 45 years	25%
45+ years	16%

Female-Overall	Male-Overall
50%	50%

Other Existing Oriental Language Corpus

Language	Script Type	Existing Hours
Arabic	Sentence/In-Car	565
Chinese	Sentence/In-Car/Conversational/Commanding Words	28000
Cantonese	Sentence	1200
Tagalog	Sentence	500
Hindi	Sentence/Conversational	2600
Indonesian	Sentence/Conversational	3100
Japanese	Sentence/Conversational/Commanding Words	3200
Korean	Sentence/Conversational	1500
Malay	Sentence/Conversational/Commanding Words	1400
North Korean	Sentence/Conversational	700
Russian	Sentence/Conversational/Commanding Words/In-Car	2500
Taiwanese	Sentence/In-Car	1900
Thai	Sentence/In-Car/Conversational	4500
Tibetan	Conversational	300
Urdu	Sentence/Commanding Words	600
Uygur	Sentence/Conversational	500
Ukrainian	Sentence/In-Car	600
Vietnamese	Sentence/In-Car/Conversational	1100

Data Resources Overview

Existing Data Resources Overview

Data Type	Language Coverage	Data Volume
TTS	35 Languages	520 Hours
ASR	65 Languages	85,000+ Hours
Lexicon	48 Languages	5 Million Entries
Text	31 Languages	600 Million Annotated Words

What's Unique?

Diversities

- In-Car Corpus
- Spontaneous Corpus
- Telephony Corpus
- Non-Native Speaker Corpus
- Far-Filed Recording
- Children Speech

….

Uniqueness

- North Korean
- Hebrew
- Catalan
- Urdu
- Ukrainian
- Uygur
- Tibetan

…

Cooperation

Welcome to Approach Us for Cooperation …

- Phonetic & Phonological Analysis
- Speech Recognition
- Speaker Recognition
- Language Recognition
- Language Understanding
- Speech Synthesis

… …

Chenqing@speechocean.com

Cooperation

Mix-lingual Speech Processing Special Session & Speech Recognition Challenge

☞ O-Cocosda2016 – Bali,Oct.2016

Database :- Provided by Speechocean

Chinese-English Mix-lingual Speech
80 Hours Recording Time
Manually Transcripted
Lexicon Available

Baseline:- Provided by Tsinghua University

Chenqing@speechocean.com

Cooperation

It’s Coming……

- O-COCOSDA2017
- APSIPA2017
- Maybe something more…..

wangdong99@mails.tsinghua.edu.cn

Chenqing@speechocean.com

Data With Minimum Cost

KingLine Data Center

Share Exchange Distribution

- Free Membership that never expires;
- 500+ Commercial Corpus, 300+ Academic Corpus;
- Many ways to earn credits & Exchange data with credits;
- Best way to earn credits: share data with us, or distribute data by us

Chenqing@speechocean.com

Data With Minimum Cost

Don’t Miss Any Free Data !

32 Free Corpora Have Been Promoted This Year!

The last free database of this year is coming in 1 week!

Chenqing@speechocean.com

清華大學

Center for Speech and Language Technologies

GROUPING

speechocean

海天瑞声

Thanks a lot.