Digit-dependent Local I-Vector for Text-Prompted Speaker Verification with Random Digit Sequences

Peixin Chen, Wu Guo, Guoping Hu
presented by Peixin Chen
University of Science and Technology of China

ISCSLP 2016 ·Tianjin, China

Outline

- **Background and Motivation**
  - Speaker verification categories
  - Shortcomings of the traditional i-vector
  - Related works of local vectors
- **Digit-dependent local i-vector**
  - Local I-Vector
  - Digit-dependent PLDA
- **Experiments**
  - Database and trials
  - System flowchart
  - Configurations
  - Performance
- **Conclusion**

Background and Motivation

- Speaker verification categories:
  - text-independent:

    speaker can freely say any content

  - text-dependent:(higher accuracy, shorter utterances)

    a specific set of words or a fixed passphrase must be used

  - text-prompted:(prevent playback attack)

    word sequences are randomly produced

Background and Motivation

- Shortcomings of the traditional i-vector:
  - not take advantage of lexical information
  - not perform well with short utterances

- Related works of local vectors:

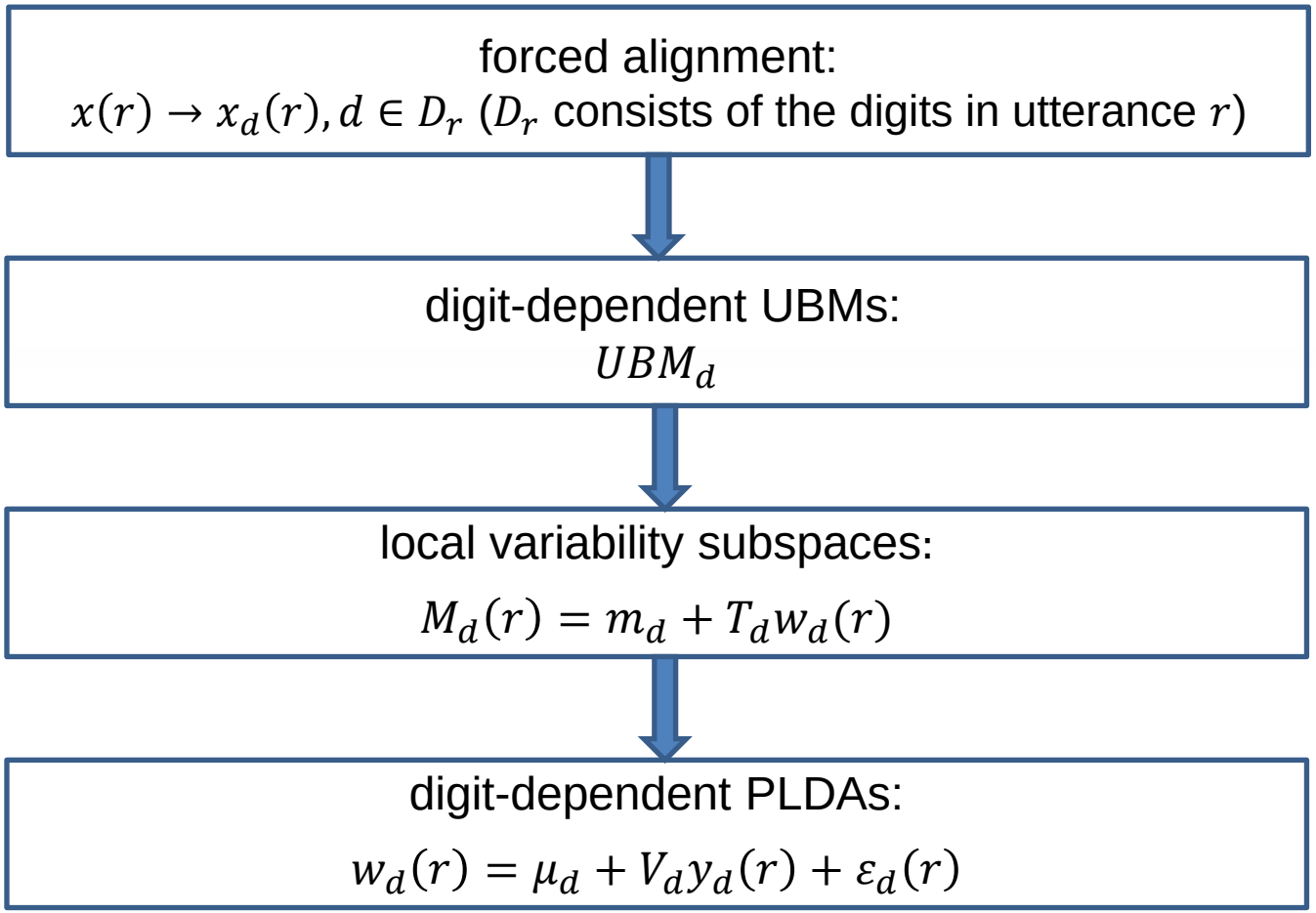
- Global and local vectors:
  T.Stafylakis. JFA for speaker recognition with random digit strings
  - global vector models the whole utterance
  - local vector models segmentation such as a word or a phoneme in an utterance

- Phone-centric local vector:
  L.Chen. Phone-centric local variability vector for text-constrained speaker verification
  - 22 phone-centirc local vectors

the lexicons are limited to 10 digits ⟹ local vectors based on digit acoustic units?

Digit-dependent local i-vector

- Overview of digit-dependent local i-vector:



Digit-dependent local i-vector

➢ Tradition i-vector(global):

- Total variability space:

$$M(r) = m + Tw\ (r)$$

- $w(r)$ represents speaker identity information

➢ Digit-dependent local i-vector:

- Ten separate local variability space:

$$M_d(r) = m_d + T_d w_d(r)$$

- $w_d(r)$ represents speaker-digit combination information

Digit-dependent local i-vector

➢ Digit-dependent PLDA:

- PLDA model:

$$w_d(r) = \mu_d + V_d y_d(r) + \varepsilon_d(r)$$

- PLDA scores:

$$s_d(r_1, r_2) = \log \frac{P(w_d(r_1), w_d(r_2) | H_s)}{P(w_d(r_1) | H_d) P(w_d(r_2) | H_d)}$$

H_s: $w_d(r_1)$ and $w_d(r_2)$ come from the same speaker-digit combination

H_d: $w_d(r_1)$ and $w_d(r_2)$ come from different speaker-digit combination

- Fusion of PLDA scores:

$$s(r_{test}, r_{enroll}) = \frac{1}{|D_r|} \sum_{d \in D_r} s_d(r_{test}, r_{enroll})$$

Experiments

➢ Database and trials:

- Database: RSR2015 Part III
- Types of trials:

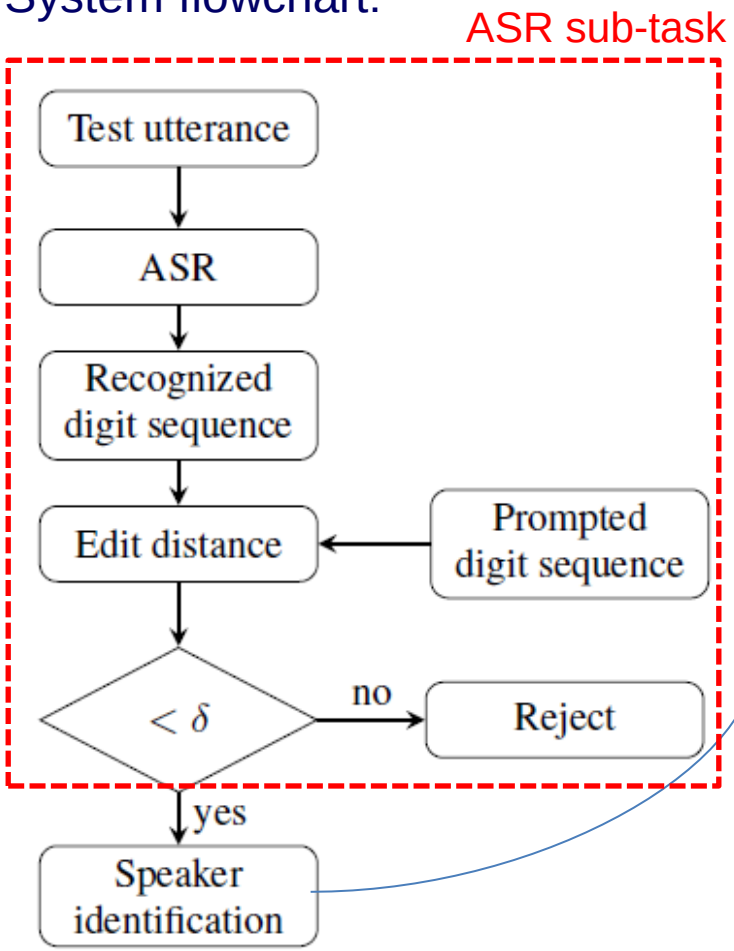
	Correct lexicons	Wrong lexicons
Target	TAR-correct	TAR-wrong
Impostor	IMP-correct	IMP-wrong

- Number of trials:

	Male	Female
TAR-correct	1046	918
TAR-wrong	25659	22878
IMP-correct	55213	42285
IMP-wrong	40648	34987

Experiments

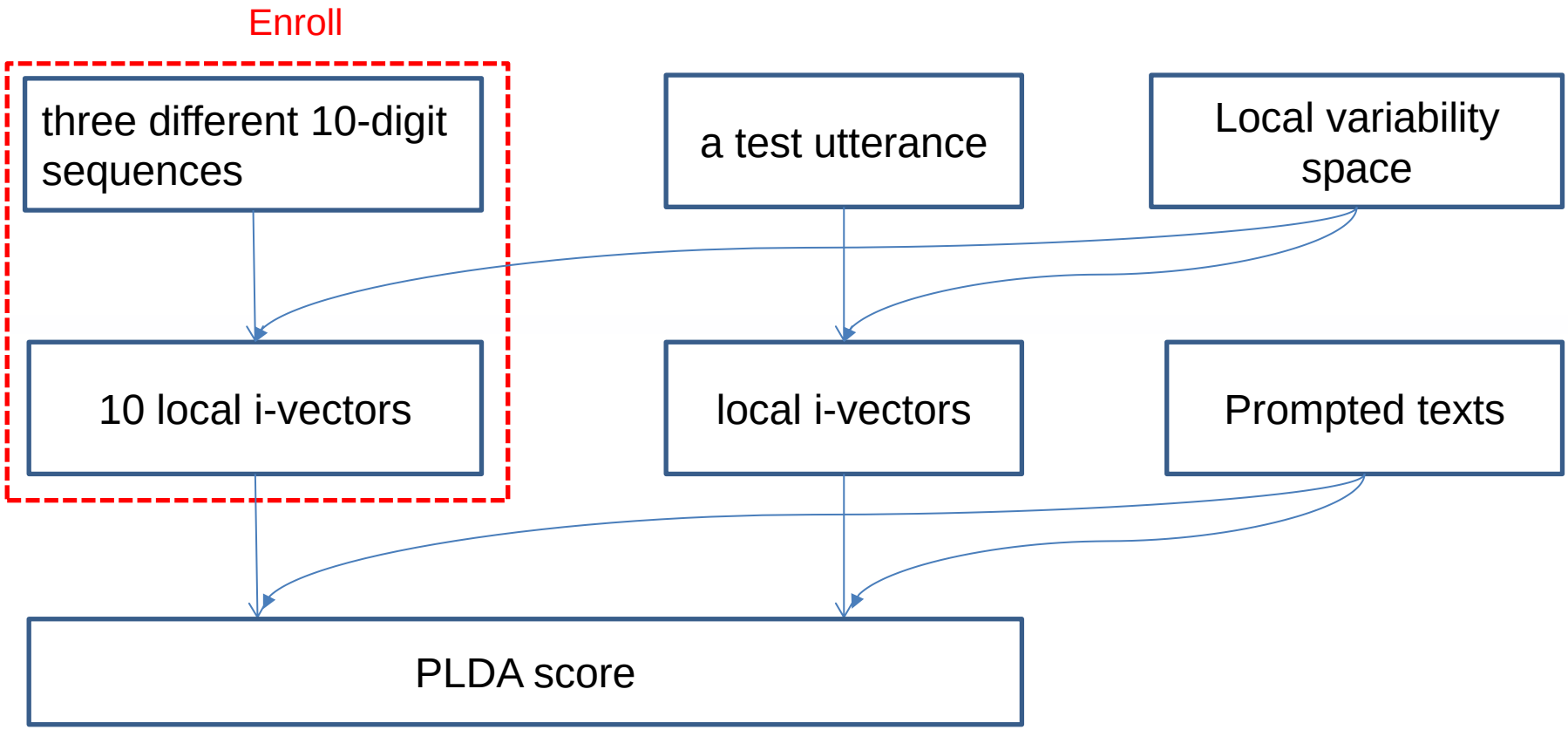
- System flowchart:



- Four different i-vector systems:
- GMM i-vector
- DNN i-vector
- phone-dependent local i-vector
- digit-dependent local i-vector

Experiments

- Speaker identification in local i-vector sub-system:



Experiments

- Configurations of the four i-vector systems：

	Gaussian num	i-vector dim	PLDA speaker factors	features
GMM i-vector	128	400	200	39-dim mfcc
DNN i-vector	80	400	200	39-dim mfcc 120-dim fbank
local i-vector(phone)	16	100	60	39-dim mfcc
local i-vector(digit)	16	100	60	39-dim mfcc

- GMM-HMM systems (for forced alignment)：

GMM-HMM	Phone accuracy	Word accuracy
Phone-unit	72.68%	96.72%
Digit-unit	— —	97.70%

Experiments

- Performance of speaker identification：

	GMM i-vector	DNN i-vector	Phone i-vector	Digit i-vector
male				
EER(%)	2.23	1.85	2.53	**1.55**
DCF08	0.0135	**0.0099**	0.0159	0.0108
female				
EER(%)	3.57	2.35	2.98	**1.47**
DCF08	0.0221	0.0128	0.0161	**0.0113**

Conclusion

- Global GMM i-vector:
  - Gaussians in GMM is ambiguous without clear phonetic definition
- Global DNN i-vector:
  - Does not take advantage of the prompts information
  - Total variability has no phonetic meanings, only represents speaker vocal tract information
- Digit-dependent local i-vector:
  - Represent speaker-digit combination information
  - Flexible, suitable for text-prompted SV with random digit sequences
- Phone-dependent local i-vector:
  - 22 local i-vectors suffer from sparsity of data in short utterance
  - The influence of co-articulation, phone accuracy is lower than word accuracy

Thank You for Listening!

Q&A

Peixin Chen
E-mail: bubble@mail.ustc.edu.cn