Utterance Partitioning with Acoustic Vector Resampling for I-Vector Based Speaker Verification

Wei Rao and Man-Wai Mak

Odyssey 2012

Department of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong SAR, China

The Hong Kong Polytechnic University
香港理工大學

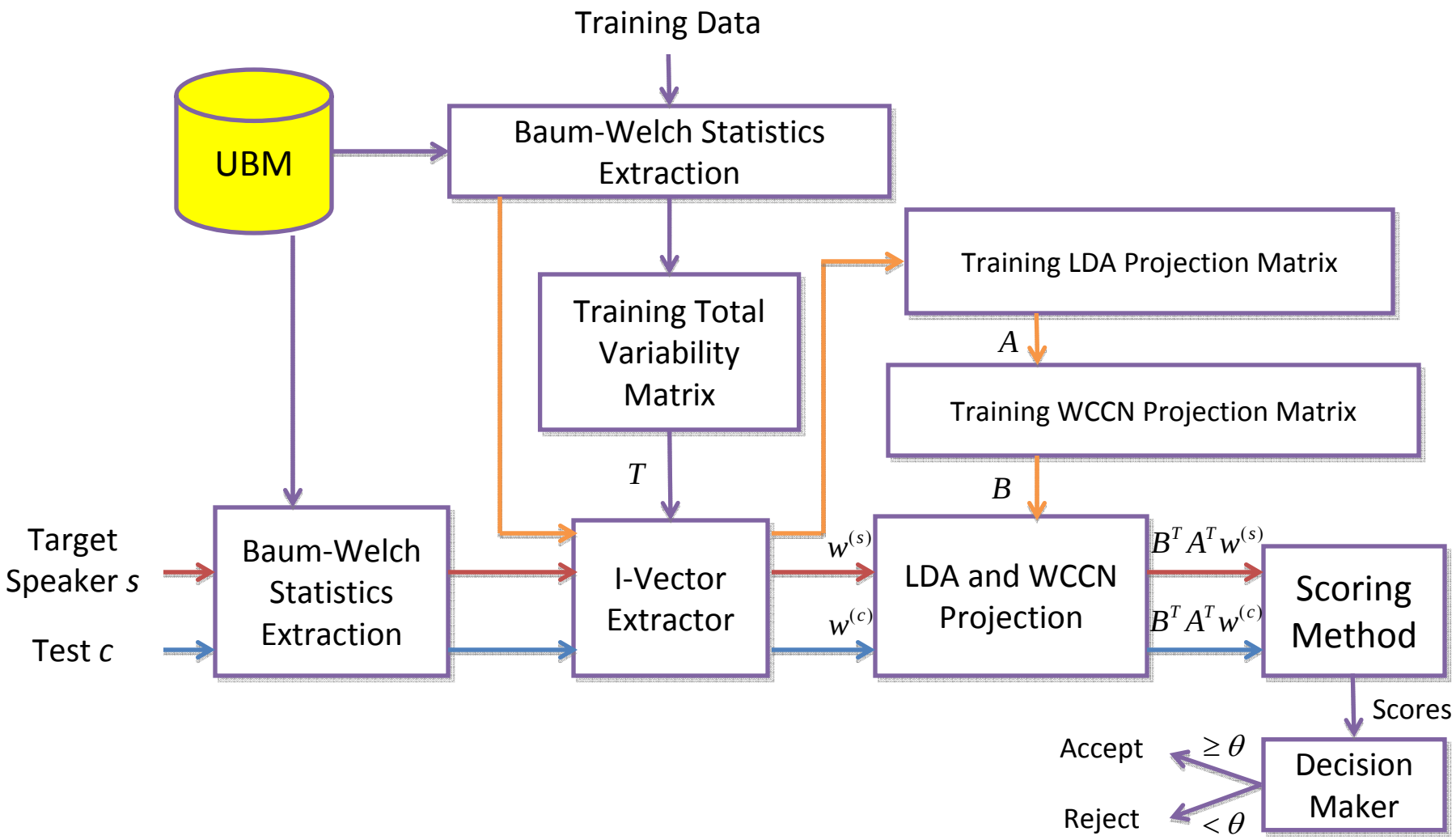
Summary

- We investigate the effect of varying the **conversation length** and the number of **recordings (sessions) per speakers** on LDA+WCCN projection matrices in i-vectors based speaker verification.
- We demonstrate that the amount of **speaker-dependent information** that an i-vector can capture will become **saturated** when the utterance length exceeds a certain threshold.
- We propose to maximize the capability of i-vectors by partitioning a long conversation into a number of sub-utterances in order to produce more i-vectors per conversation.

Key Results

- Using more i-vectors per conversation can enhance the capability of LDA and WCCN when the number of conversations per training speaker is limited
- Increasing the number of i-vectors per target speaker alleviates the data imbalance problem in SVM scoring, leading to 22% relative improvement when compared with cosine distance scoring.

I-Vector Based Speaker Verification



Issues in I-Vector Based SV

- **Limited** number of recordings (sessions) per speaker for estimating the LDA and WCCN projection matrices.
- **Data imbalance** in training the target-speaker support vector machines when SVM scoring is used, i.e., each SVM is trained with 1 target-speaker's i-vector and many background speakers' i-vectors.

Estimating LDA and WCCN projection matrices

- Linear Discriminant Analysis (LDA)
  - find a set of orthogonal axes for minimizing the within-class variation and maximizing the between-class separation, i.e. maximizing

$$J(A) = \mathrm{tr}\left\{ \left(A^{\mathsf{T}} S_w A\right)^{-1} \left(A^{\mathsf{T}} S_b A\right) \right\}$$

For fixed S, small M_i could cause low rank in S_w

$$S_w = \sum_{i=1}^{S} \frac{1}{M_i} \sum_{j=1}^{M_i} (w_j^i - \mu^i)(w_j^i - \mu^i)^{\mathsf{T}}$$

$$S_b = \sum_{i=1}^{S} (\mu^i - \mu)(\mu^i - \mu)^{\mathsf{T}}$$

$$\mu^i = \frac{1}{M_i} \sum_{j=1}^{M_i} w_j^i$$

Estimating LDA and WCCN projection matrices

- Within Class Covariance Normalization (WCCN)
  - to normalize the within-speaker variation

For fixed S, small M_i could cause low rank in W

$$W = \sum_{i=1}^{S} \frac{1}{M_i} \sum_{j=1}^{M_i} (A^{\mathsf{T}} w_j^i - \widetilde{\mu}^i)(A^{\mathsf{T}} w_j^i - \widetilde{\mu}^i)^{\mathsf{T}}$$

$$\widetilde{\mu}^i = \frac{1}{M_i} \sum_{j=1}^{M_i} A^{\mathsf{T}} w_j^i \qquad W^{-1} = BB^{\mathsf{T}}$$

Cosine distance scoring:

$$S_{\cos}\left(w^{(c)}, w^{(s)}\right) = \frac{\left\langle B^{\mathsf{T}} A^{\mathsf{T}} w^{(c)}, B^{\mathsf{T}} A^{\mathsf{T}} w^{(s)} \right\rangle}{\|B^{\mathsf{T}} A^{\mathsf{T}} w^{(c)}\| \|B^{\mathsf{T}} A^{\mathsf{T}} w^{(s)}\|}$$

Problems in Estimating LDA and WCCN Projection Matrices

- Both LDA and WCCN require the computation of a within class covariance matrix, which requires a number of recordings (sessions) per speaker
- However, the number of recordings per speaker is limited
- The lack of multiple recordings per speaker could lead to inaccurate within-speaker-scatter matrix.

Data-imbalance Problem in SVM Scoring

- SVM scoring is not preferred because of the data-imbalance problem, i.e., one target-speaker's i-vector vs. many background i-vectors.

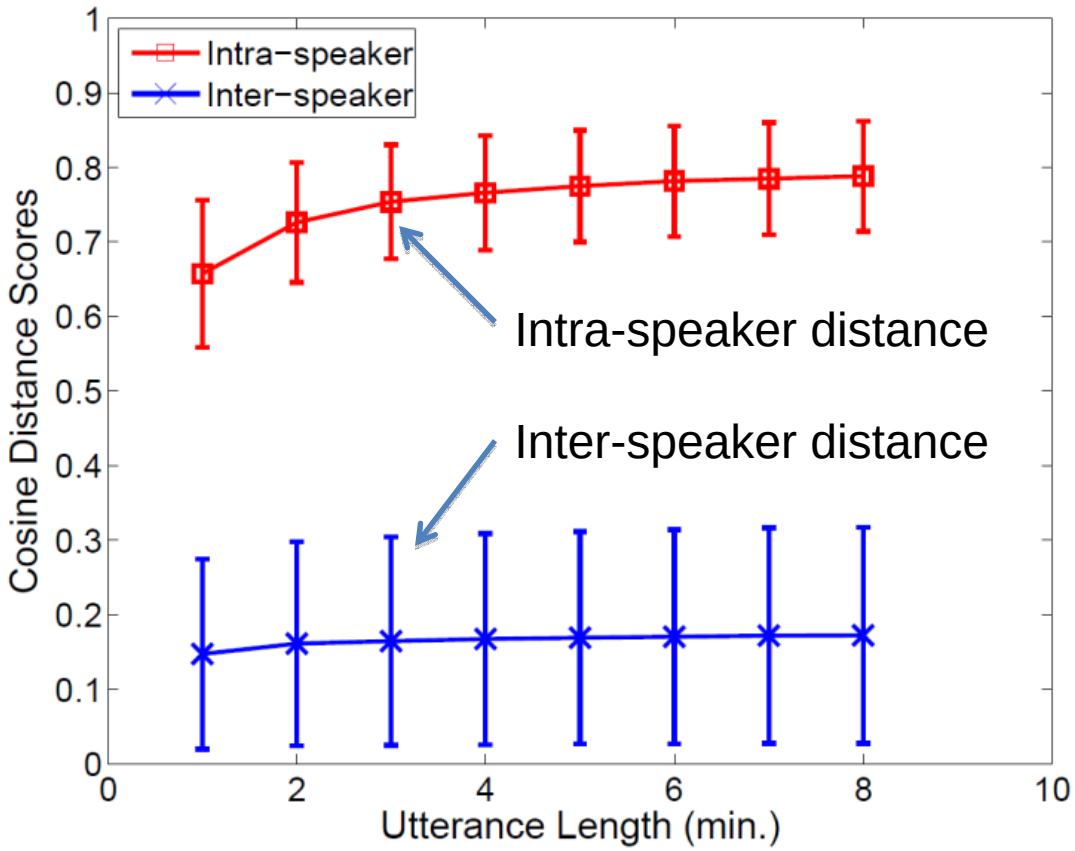
LDA+WCCN Projection: $w \leftarrow B^T A^T w$

$$S_{\text{SVM}}\left(w^{(c)}, w^{(s)}\right) = \underbrace{\alpha_0^{(s)} K\left(w^{(c)}, w^{(s)}\right)}_{\text{From target-speaker}} - \underbrace{\sum_{i \in S^{(b)}} \alpha_i^{(s)} K\left(w^{(c)}, w^{(b_i)}\right)}_{\text{From background speakers}} + d^{(s)}$$

- We show that it is possible to generate a number of target-speaker's i-vectors for SVM training although each target speaker only have one enrollment utterance.

Effect of Utterance Length on I-Vectors

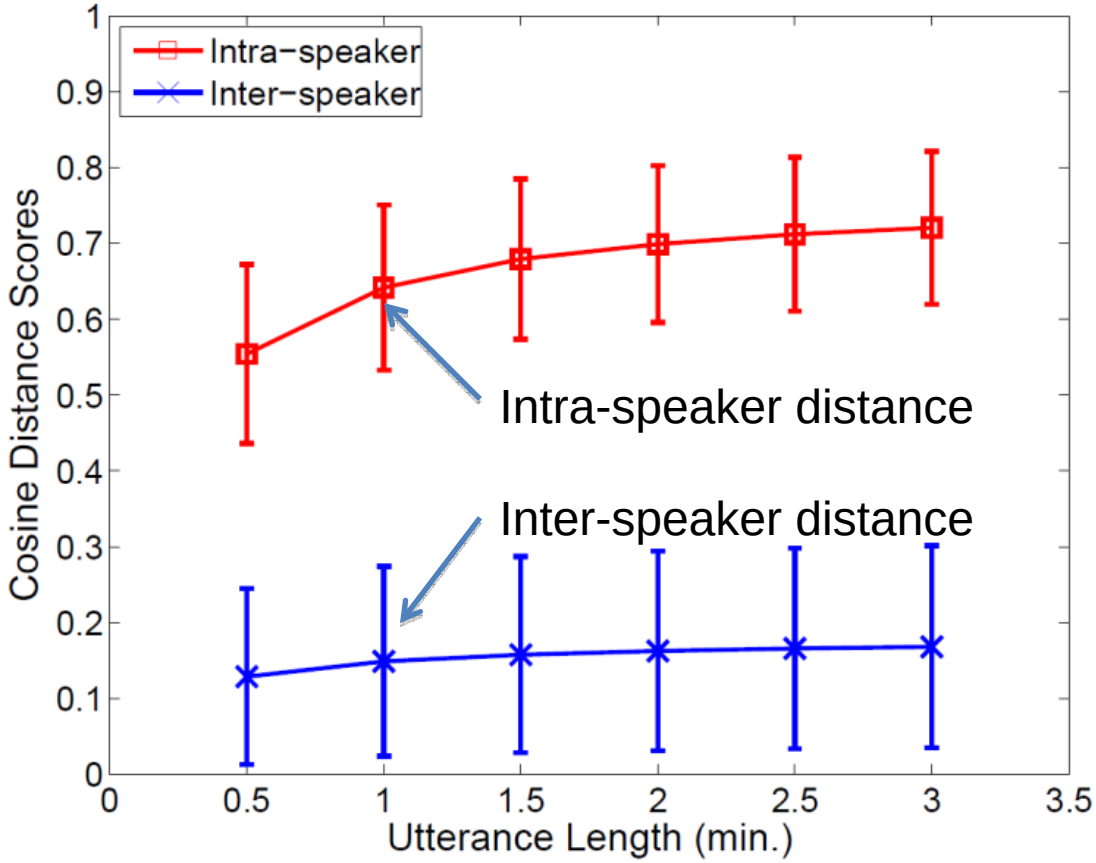
29 male speakers, each providing 4 interview conversations



(a) 8 mins, interview_mic

Effect of Utterance Length on I-Vectors

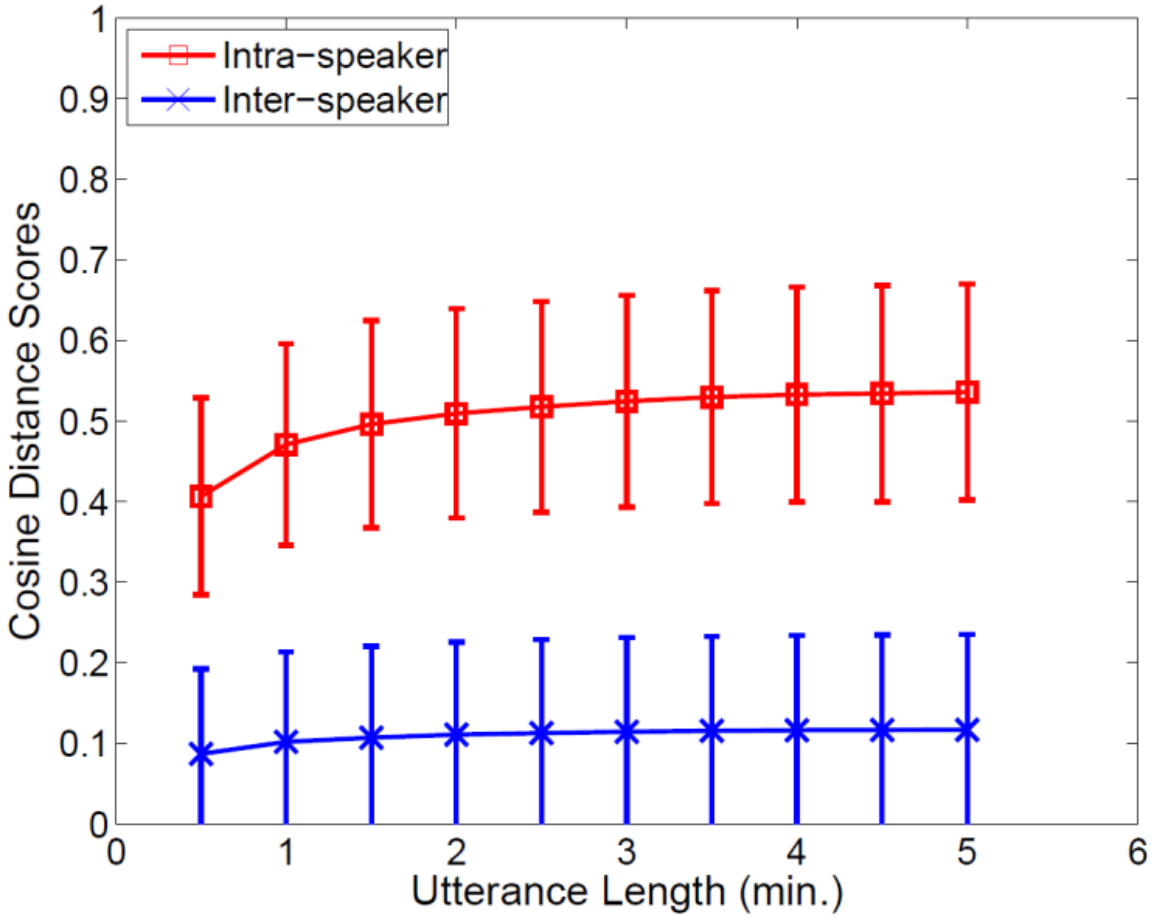
196 male speakers, each providing 4 interview conversations



(b) 3 mins, interview_mic

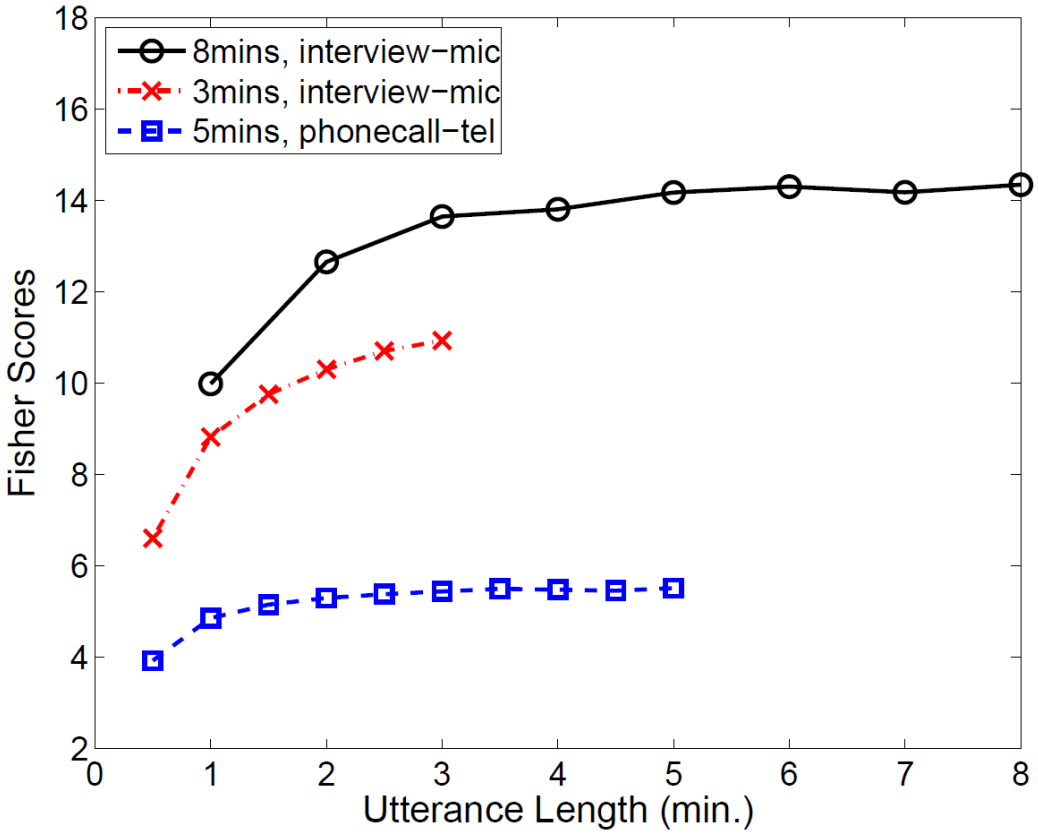
Effect of Utterance Length on I-Vectors

47 male speakers, each providing 4 telephone conversations



(c) 5 mins, phonecall_tel

Effect of Utterance Length on I-Vector



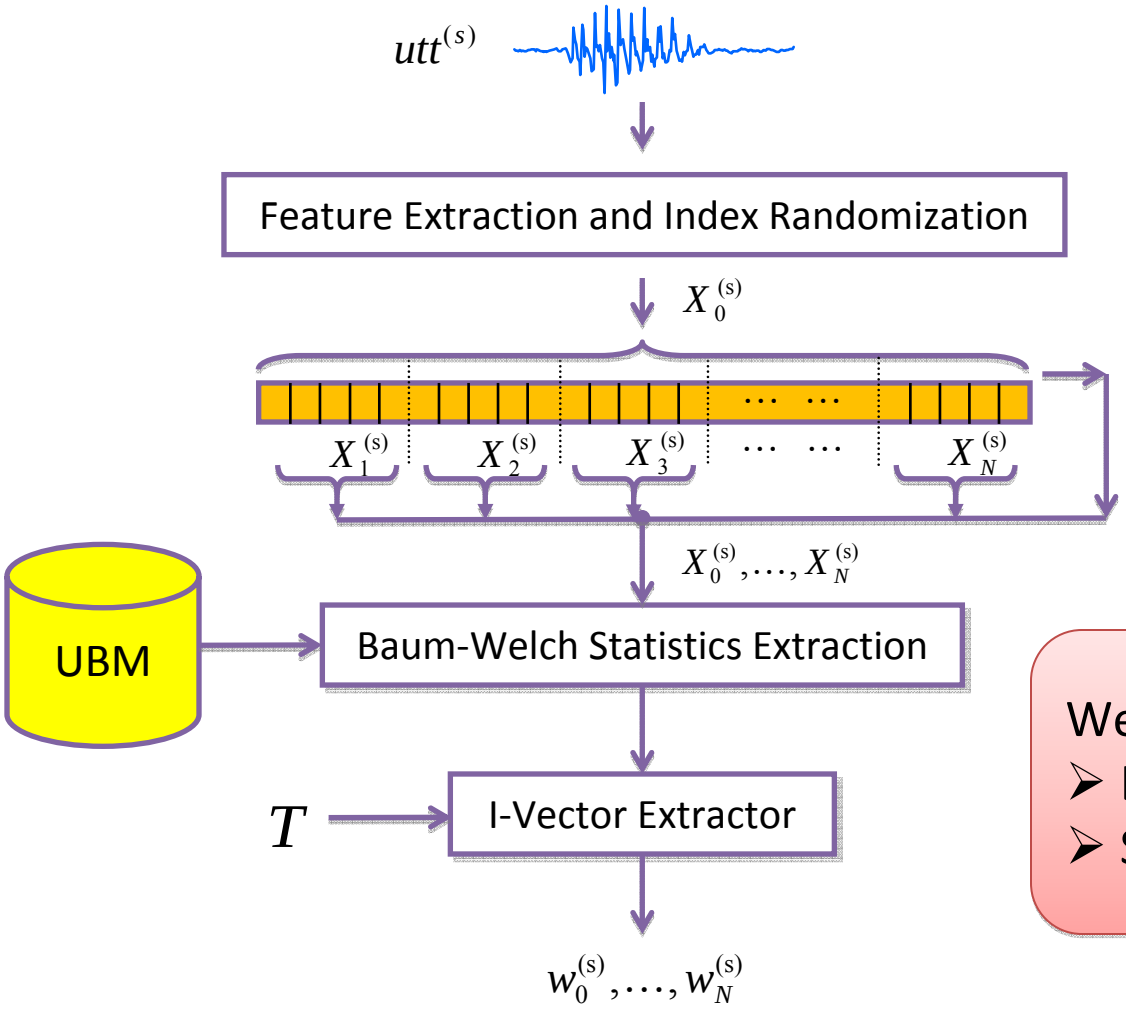
The discriminative power of i-vectors becomes saturated for segment length exceeding 2-3 minutes.

It is not necessary to record very long utterances for the i-vectors to achieve good performance.

Dividing the long utterance into a number of sub-utterances to produce more i-vectors per conversation.

$$S_{\text{Fisher}} = \frac{\left(\mu_{\text{intra}} - \mu_{\text{inter}}\right)^2}{\sigma^2_{\text{intra}} + \sigma^2_{\text{inter}}}$$

Utterance Partitioning with Acoustic Vector Resampling (UP-AVR)



If the process is repeated R times, we can obtain *RN* + 1 i-vectors from one conversation

We can apply UP-AVR in:

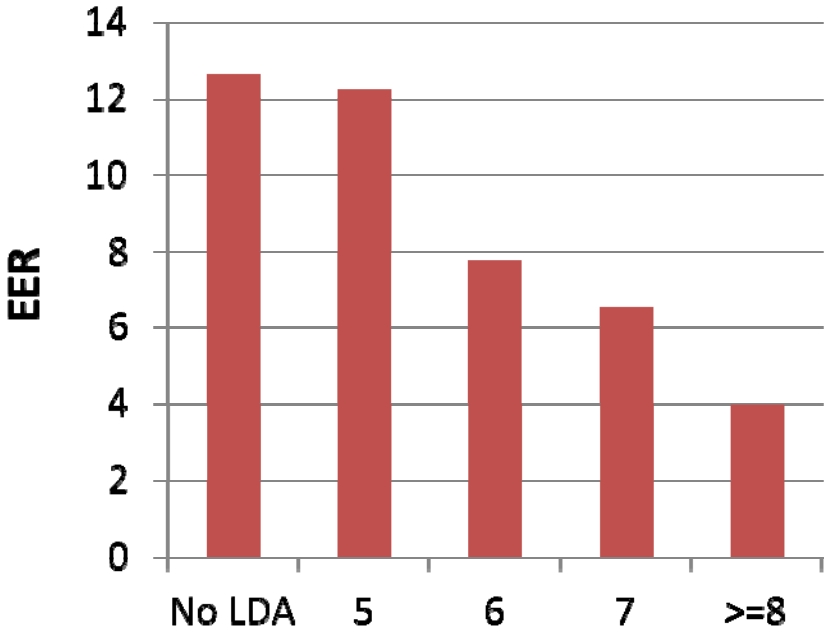
- LDA and WCCN estimation
- SVM scoring

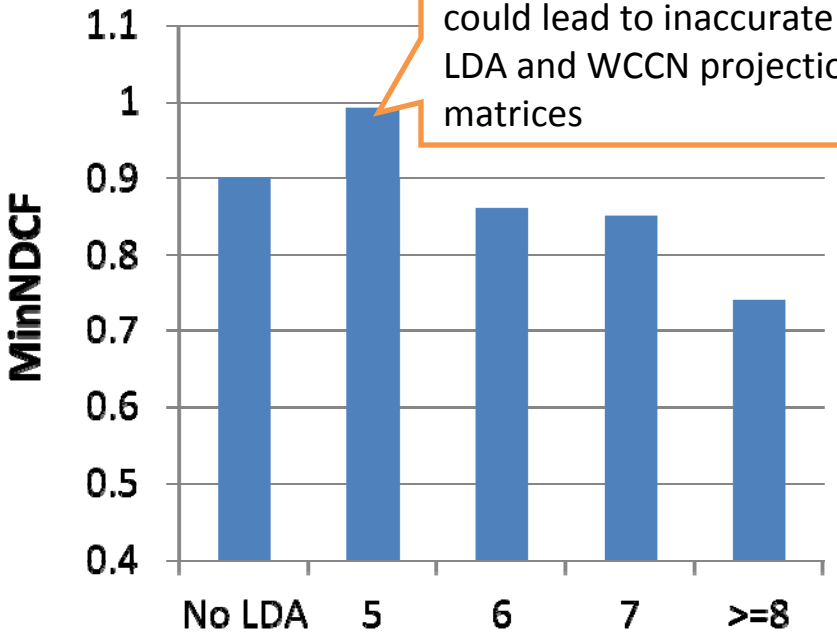
Experiment Setup

- **Evaluation dataset:** cc1, cc2, cc4, cc7 and cc9 of NIST SRE 2010 male extended core set (i.e., interview and mic data)
- **Parameterization:** 19 MFCCs + △ + △ → 60-Dim
- **UBM:** gender-dependent, 1024 mixtures
  - Training data were selected from NIST 2005, 2006 and 2008 SRE.
- **Total Variability Matrix:** gender-dependent, 400 total factors
  - Training data were selected from NIST 2005, 2006 and 2008 SRE.
- **LDA + WCCN Projection:** 400 → 150 dim
  - UP-AVR for LDA and WCCN: 111 male speakers were selected from NIST 2008 SRE. Each speaker provided at least 8 utterances.
  - UP-AVR for SVM scoring: Same training data as for training the total variability matrix.
- **Score Norm:** ZT-norm

Results

- Effect of varying the number of recording sessions per speaker on the effectiveness of LDA and WCCN projection

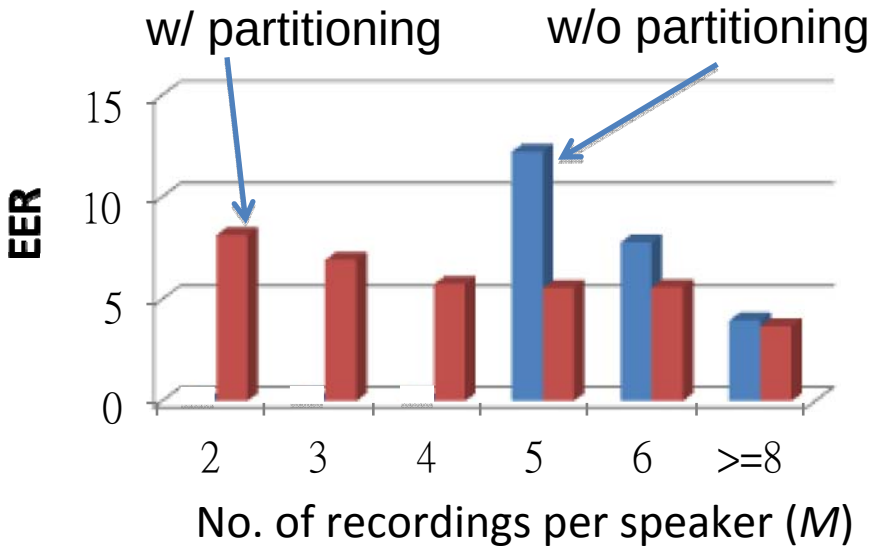


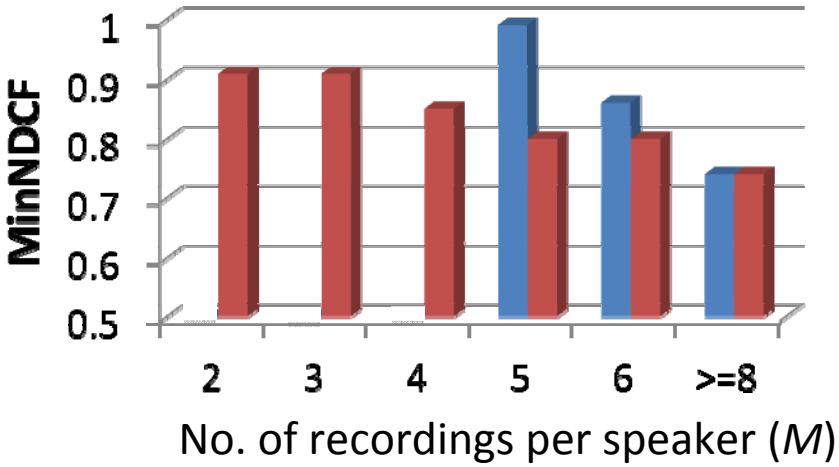


Insufficient number of recordings per speaker could lead to inaccurate LDA and WCCN projection matrices

By increasing the number of recordings per speaker, the effectiveness of LDA and WCCN projection improves significantly

Results: UP-AVR for LDA and WCCN





- When $M \leq 4$, numerical difficulty occurs, and utterance partitioning can effectively overcome such difficulty
- When $M = 5$, utterance partitioning is the most effective
- When $M \geq 8$, the effectiveness of utterance partitioning diminishes.

Without UP-AVR

UP-AVR

UP-AVR for SVM Scoring

Scoring Methods	Common Condition					
	CC1	CC2	CC4	CC7	CC9	Mic
Cosine Distance	0.38	0.52	0.53	0.99	0.44	0.63
SVM	0.30	0.47	0.45	0.99	0.29	0.53
SVM+UP-AVR (N=4, R=1)	0.28	0.45	0.41	0.99	**0.23**	0.49
SVM+UP-AVR (N=4, R=4)	**0.26**	**0.44**	**0.41**	0.99	0.24	**0.49**

(a) MinNDCF

Scoring Methods	Common Condition					
	CC1	CC2	CC4	CC7	CC9	Mic
Cosine Distance	1.72	2.88	2.81	9.23	**1.71**	2.98
SVM	1.87	3.30	3.07	8.94	3.31	3.35
SVM+UP-AVR (N=4, R=1)	1.57	3.04	2.97	8.37	3.04	3.07
SVM+UP-AVR (N=4, R=4)	**1.46**	**2.76**	**2.70**	**8.30**	3.18	**2.73**

(b) EER(%)

Conclusions

- Utterance partitioning can enhance the effectiveness of LDA and WCCN projections under insufficient speech resources.
- After performing utterance partitioning, SVM scoring outperforms cosine distance scoring by 22% and 9% in terms of minimum DCF and EER, respectively.

Thank You

I-Vector Extraction

$$m_s = m + Tw$$

m_s is the speaker- and channel-dependent GMM-supervector;
m is the GMM-supervector of the UBM;
T is a low-rank total variability matrix;
w is a low-dimension vector called the i-vector.

Support Vector Machine Scoring

Given the SVM of target speaker s, the verification score of claimant c is given by:

$$S_{\text{SVM}}\left(w^{(c)}, w^{(s)}\right) = \alpha_0^{(s)} K\left(w^{(c)}, w^{(s)}\right) - \sum_{i \in S^{(b)}} \alpha_i^{(s)} K\left(w^{(c)}, w^{(b_i)}\right) + d^{(s)}$$

The kernel function $K(\cdot,\cdot)$ can be of many forms. The cosine kernel is shown as below:

$$K\left(w^{(c)}, w^{(s)}\right) = \frac{\left\langle B^{\mathrm{T}} A^{\mathrm{T}} w^{(c)}, B^{\mathrm{T}} A^{\mathrm{T}} w^{(s)} \right\rangle}{\left\| B^{\mathrm{T}} A^{\mathrm{T}} w^{(c)} \right\| \left\| B^{\mathrm{T}} A^{\mathrm{T}} w^{(s)} \right\|}$$

Results

- Effect of varying No. of utterances per speaker on effectiveness of LDA and WCCN projection

No. of utts. Per spk (M)	MinNDCF					
	CC1	CC2	CC4	CC7	CC9	Mic
M=0	0.62	0.84	0.82	0.98	0.63	0.90
M=5	0.97	0.99	0.97	0.99	0.98	0.99
M=6	0.76	0.86	0.79	0.97	0.56	0.86
M=7	0.71	0.81	0.82	**0.96**	0.56	0.85
M>=8	**0.54**	**0.64**	**0.72**	0.98	**0.42**	**0.74**

No. of utts. Per spk (M)	EER (%)					
	CC1	CC2	CC4	CC7	CC9	Mic
M=0	5.10	10.77	9.65	15.08	5.96	12.60
M=5	8.52	12.75	11.72	18.43	10.25	12.21
M=6	4.50	8.52	6.10	12.84	5.13	7.76
M=7	3.83	6.84	5.19	12.28	4.27	6.53
M>=8	**2.43**	**3.79**	**3.85**	**10.87**	**4.09**	**3.96**

UP-AVR for LDA and WCCN

Systems	No. of utts. per speaker (M)					
	2	3	4	5	6	>=8
Without UP-AVR	---	---	---	0.99	0.86	0.74
UP-AVR(2)	0.97	0.94	0.85	**0.80**	0.81	0.74
UP-AVR(4)	0.94	0.91	0.85	0.84	**0.80**	**0.74**
UP-AVR(8)	**0.91**	**0.91**	**0.85**	0.82	0.83	0.75

1(a) MinNDCF

Systems	No. of utts. per speaker (M)					
	2	3	4	5	6	>=8
Without UP-AVR	---	---	---	12.21	7.76	3.96
UP-AVR(2)	15.24	8.97	6.71	6.21	6.18	3.83
UP-AVR(4)	9.51	7.28	**5.74**	5.68	5.61	3.74
UP-AVR(8)	**8.13**	**6.93**	6.34	**5.55**	**5.58**	**3.67**

1(b) EER(%)

UP-AVR for SVM Scoring

Scoring Methods		Common Condition					
		CC1	CC2	CC4	CC7	CC9	Mic
CDS		0.38	0.52	0.53	0.99	0.44	0.63
SVM	C = 1	0.33	0.49	0.50	0.91	0.28	0.52
	C = 0.01	0.30	0.47	0.45	0.99	0.29	0.53
SVM+UP-AVR(4)	C = 1	0.29	0.47	0.46	0.92	0.29	0.49
	C = 0.01	0.28	0.45	0.41	0.99	**0.23**	0.49
SVM+UP-AVR(16)	C = 1	0.30	0.46	0.47	**0.89**	0.24	0.49
	C = 0.01	**0.26**	**0.44**	**0.41**	0.99	0.24	**0.49**

2(a) MinNDCF

UP-AVR for SVM Scoring

Scoring Methods		Common Condition					
		CC1	CC2	CC4	CC7	CC9	Mic
CDS		1.72	2.88	2.81	9.23	**1.71**	2.98
SVM	C = 1	1.87	3.12	3.06	10.05	2.56	3.26
	C = 0.01	1.87	3.30	3.07	8.94	3.31	3.35
SVM+UP-AVR(4)	C = 1	1.71	3.00	2.86	10.05	2.56	3.10
	C = 0.01	1.57	3.04	2.97	8.37	3.04	3.07
SVM+UP-AVR(16)	C = 1	1.64	3.03	2.79	9.31	2.56	3.12
	C = 0.01	**1.46**	**2.76**	**2.70**	**8.30**	3.18	**2.73**

2(b) EER(%)