# KU-ISPL speaker recognition system for SRE 2016

Suwon Shon, Hanseok Ko

*School of Electrical Engineering, Korea University, Seoul, South Korea*


## Abstract

- Main issue of SRE 16 is language mismatch compensation
- KU-ISPL system uses 4 different i-vectors ( GMM, DNN, Sup-GMM, BNF)
- Conventional domain mismatch compensation techniques were used ( IDVC, Interpolated PLDA)
- Speaker clustering on unlabeled minor/major dataset were don for Interpolated PLDA and Calibration using AHC
- Gender Classification and Language Clustering were done
- Proposed several language mismatch compensation techqniques (ILVC, GL-norm, AEDA)

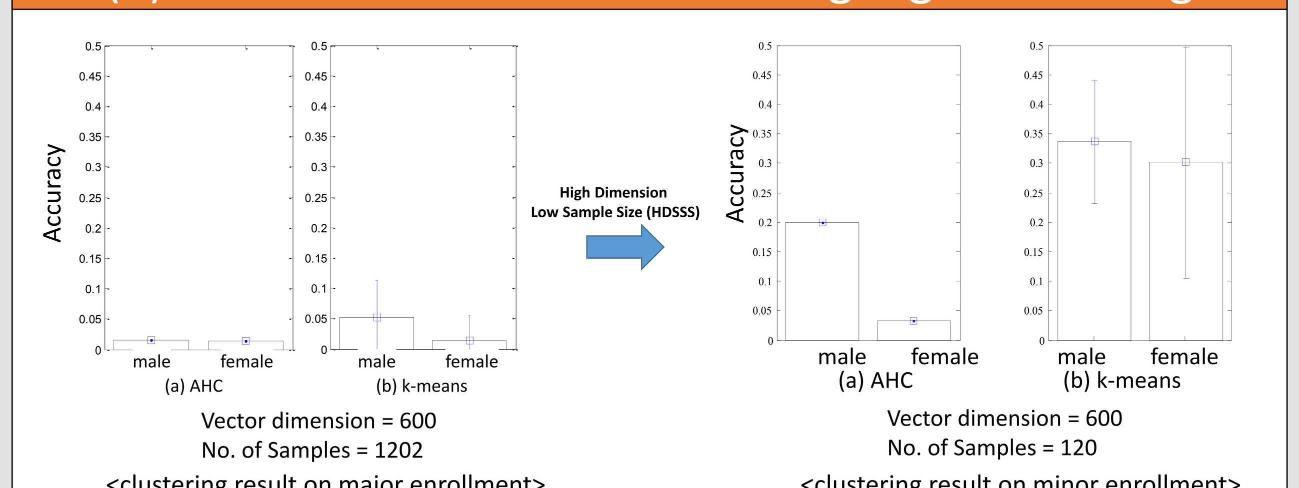

## Callmynet Dataset Statistics

| Dataset | Category | Language | Labels availability (before deadline) | Numbers of Utt. | Numbers of Spk. | Numbers of Calls |
|---|---|---|---|---|---|---|
| Dev. | Enrollment | Minor | O | 120 | 20 | 60 |
| | Test | Minor | O | 1207 | 20 | 140 |
| | Unlabeled | Minor | X | 200 | 20* | 200* |
| | Unlabeled | Major | X | 2272 | X | X |
| Eval | Enrollment | Major | X | 1202 | 802 | 602 |
| | Test | Major | X | 9294 | X | 1408 |

* means information from the SRE16 plan documents.

<Statistics of development and evaluation dataset>

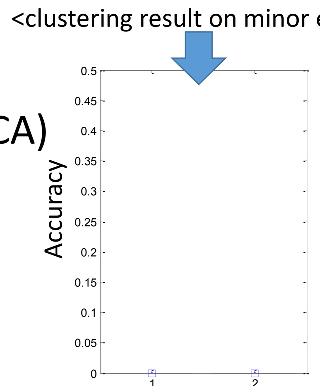
## (1) Gender Classification and Language Clustering




- 2-step approach for language clustering
  1. Initializing on IDVC subspace (alike PCA)
  2. K-means

  Or

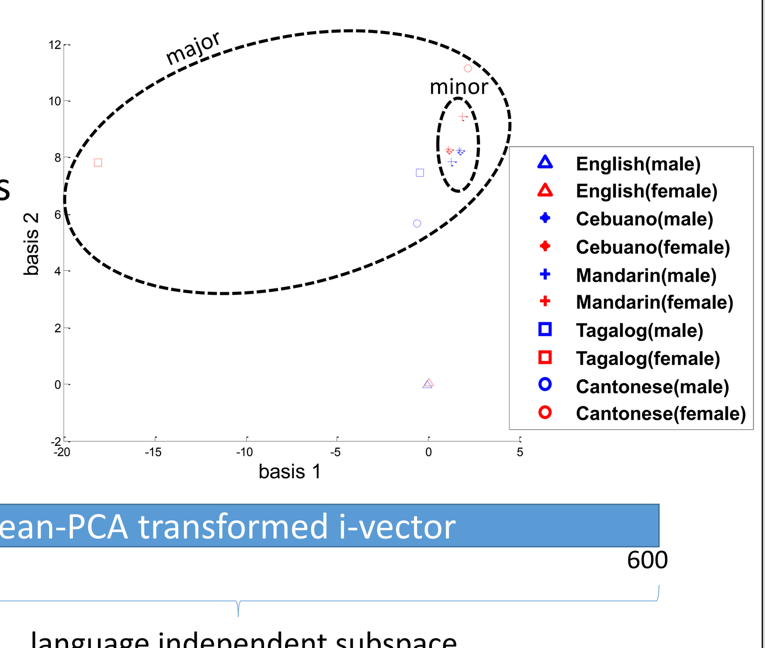
  1. Initialing using AHC
  2. K-means



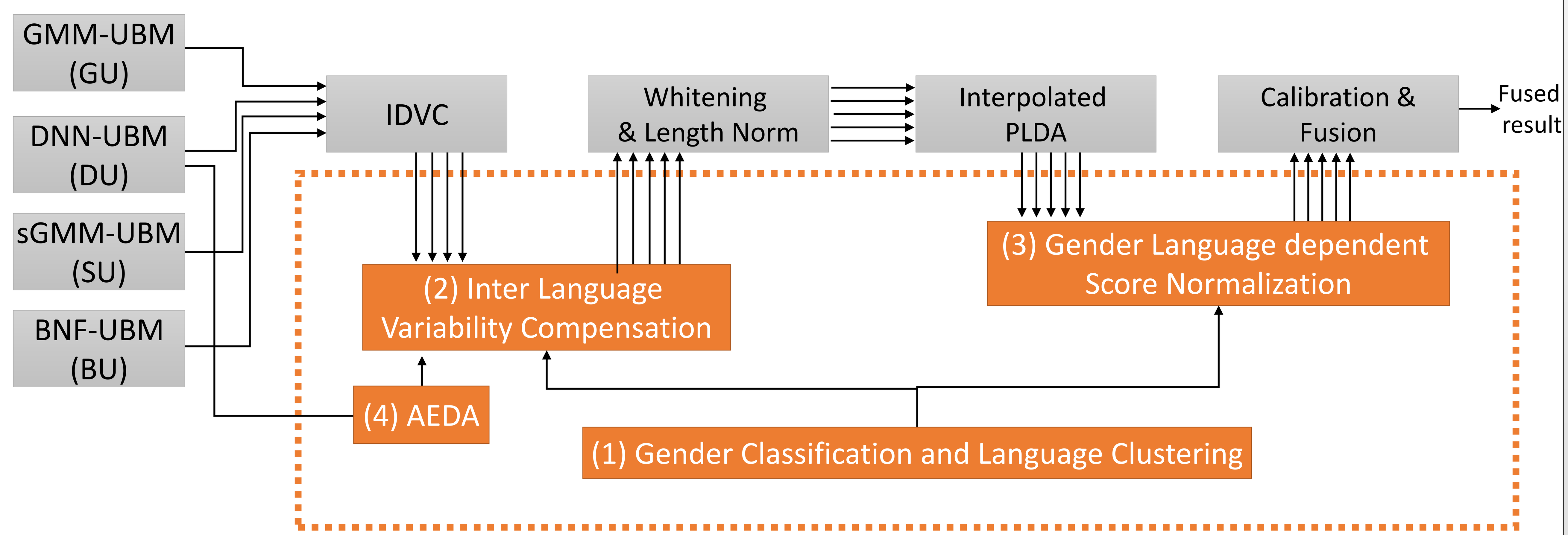

<2-step clustering result on minor enrollment>

## (2) Inter Language Variability Compensation

- Using estimated gender and language labels
  - Mean of each categories
- Removal of language dependent subspace




## Performance comparison on development trials




## Performance comparison on development trials (un-equalized)

| System Name (i-vector and applied techniques) | S-norm EER | S-norm $_{min}C_{primary}$ | S-norm $_{act}C_{primary}$ | GL-norm EER | GL-norm $_{min}C_{primary}$ | GL-norm $_{act}C_{primary}$ |
|---|---|---|---|---|---|---|
| GU-IDVC-WTLN-IPLDA | **18.3927** | 0.7017 | 0.7153 | 18.3720 | **0.7110** | **0.7239** |
| DU-IDVC-WTLN-IPLDA | 18.8587 | **0.6935** | **0.7057** | **18.3513** | 0.7114 | 0.7314 |
| SU-IDVC-WTLN-IPLDA | 19.9720 | 0.7109 | 0.7281 | 19.5112 | 0.7140 | 0.7336 |
| BU-IDVC-WTLN-IPLDA | 21.0128 | 0.7404 | 0.7718 | 20.5727 | 0.7418 | 0.7804 |
| DU-AEDA-WTLN-IPLDA | 19.7494 | 0.7272 | 0.7408 | 19.3662 | 0.7254 | 0.7502 |
| Fusion of 5 sub-systems | 16.7357 | 0.6253 | 0.6347 | 16.4095 | 0.6345 | 0.6396 |
| GU-IDVC-ILVC-WTLN-IPLDA | **16.4043** | 0.6849 | 0.7024 | **16.4872** | 0.6790 | 0.6881 |
| DU-IDVC-ILVC-WTLN-IPLDA | 17.0568 | **0.6454** | **0.6702** | 16.9221 | **0.6346** | **0.6515** |
| SU-IDVC-ILVC-WTLN-IPLDA | 17.6471 | 0.7075 | 0.7113 | 17.4814 | 0.6837 | 0.6930 |
| BU-IDVC-ILVC-WTLN-IPLDA | 18.3927 | 0.7197 | 0.7431 | 18.2425 | 0.7074 | 0.7336 |
| DU-AEDA-ILVC-WTLN-IPLDA | 18.0768 | 0.7040 | 0.7112 | 17.8749 | 0.6807 | 0.7053 |
| Fusion of 5 sub-systems | 13.8567 | 0.5800 | 0.5839 | 13.53 ↑19% | 0.5651 ↑9% | 0.5742 ↑9% |

## (3) Gender and Language dependent Score Normalization

- GL-norm

Enrollment – imposter score normalization

$$\text{score}_{GL}(\omega'_s, t'_i \mid G, L) = \frac{\text{score}(\omega'_s, t'_i) - \mu_{\omega_s|G,L}}{\sigma_{\omega_s|G,L}} + \frac{\text{score}(\omega'_s, t'_i) - \mu_{t_i|G,L}}{\sigma_{t_i|G,L}}$$

Imposter – test session score normalization

$G = \{\text{male}, \text{female}\}$

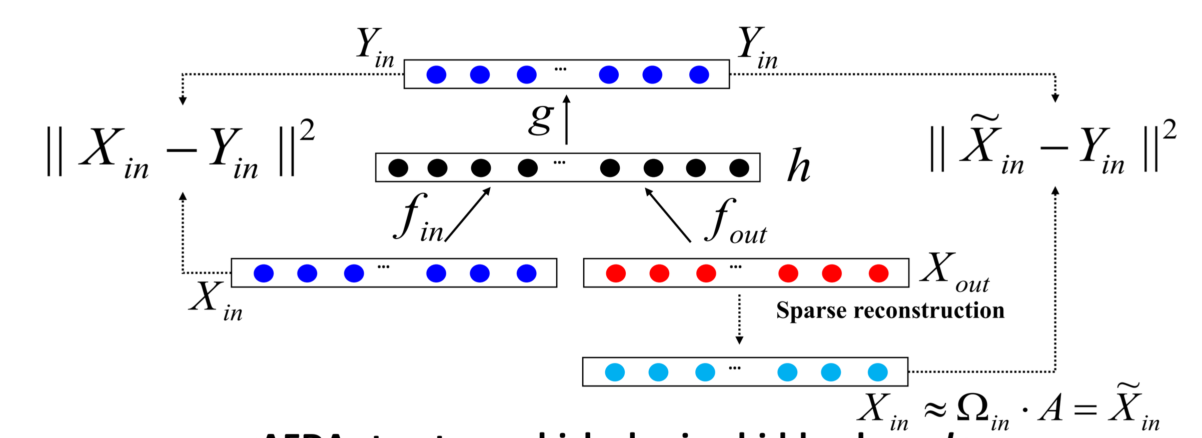
$L = \{\text{Cebuano}, \text{Mandarin}, \text{Tagalog}, \text{Cantonese}\}$

$$\mu_{\omega_s|G,L} = \frac{1}{S} \sum_{\omega_s \in G \cap L} score(\omega_s, \lambda_{imp})$$

where $\lambda_{imp} \in \Lambda_{imp}$, imposter session and target models

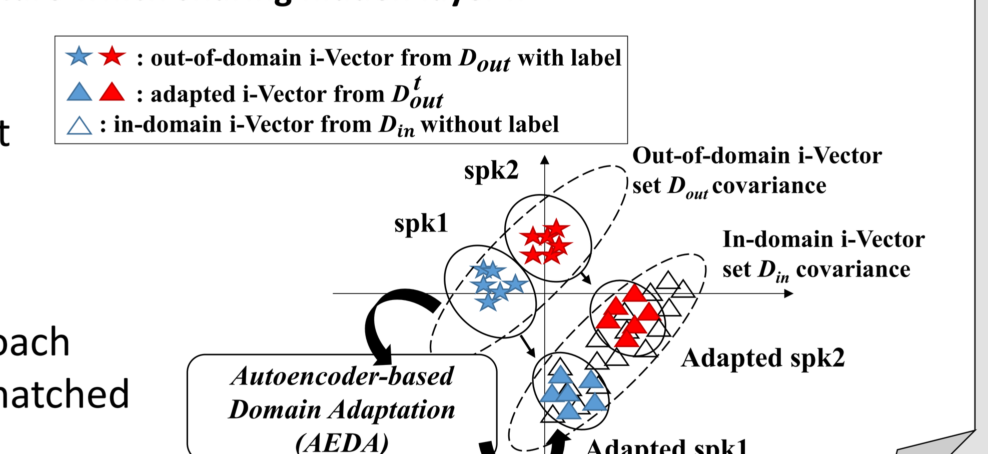
$$\sigma_{\omega_s|G,L} = \sqrt{\frac{1}{S} \sum_{\omega_s \in G \cap L} \left(score(\omega_s, \lambda_{imp}) - \mu_{\omega_s|G,L}\right)^2}$$

## (4) Autoencoder based Domain Adaptation (AEDA)




<AEDA structure which sharing hidden layer h>

- Combining Autoencoder and Denoising Autoencoder to adapt rich out-of-domain dataset to in-domain subspace
- Use sparse reconstruction approach to find out-of-domain i-vector matched in-domain i-vector




## Evaluation Results (equalized)

- Performance comparison on Evaluation trials

| | EER | $_{min}C_{primary}$ | $_{act}C_{primary}$ |
|---|---|---|---|
| Primary | 11.54 | 0.6899 | 12.0722 |
| Cont. 1 | 9.73 | 0.6416 | 1.006238 |
| Cont. 2 | **9.01** | **0.6068** | **3.4574** |

<(**Official**) Un-calibrated version>

| | EER | $_{min}C_{primary}$ | $_{act}C_{primary}$ |
|---|---|---|---|
| Primary | 10.44 | 0.6413 | 0.6886 |
| Cont. 1 | 8.93 | 0.6177 | 0.6297 |
| Cont. 2 | **8.89** | **0.5960** | **0.6277** |

<Calibrated version>

- Difference between submissions
  - ✓Prim. : use **SRE,SWB, minor, major dataset** for score norm.
    + use real gender, language label of minor enroll/test
  - ✓Cont.1 : use only **minor, major dataset** for score norm.
  - ✓Cont.2 : use only **minor, major dataset** for score norm.
    + use real gender, language label of minor enroll/test
- Performance comparison by gender and languages

| | Mandarin Male | Mandarin Female | Cebuano Male | Cebuano Female |
|---|---|---|---|---|
| EER | 5.12 | 9.33 | 12.78 | 18.72 |
| $_{min}C_{primary}$ | 0.1870 | 0.5088 | 0.7203 | 0.8548 |
| $_{act}C_{primary}$ | 0.1905 | 0.5810 | 0.7951 | 0.8892 |

<Development (minor) trials>

| | Cantonese Male | Cantonese Female | Tagalog Male | Tagalog Female |
|---|---|---|---|---|
| EER | 4.28 | 4.66 | 12.08 | 13.09 |
| $_{min}C_{primary}$ | 0.3955 | 0.4618 | 0.7528 | 0.7616 |
| $_{act}C_{primary}$ | 0.4364 | 0.5223 | 0.7752 | 0.7771 |

<Evaluation (major) trials>

## Performance improvement on Eval. trials

| System Name | Raw i-vector w. PLDA EER | Raw i-vector w. PLDA $_{min}C_{primary}$ | Proposed method EER | Proposed method $_{min}C_{primary}$ |
|---|---|---|---|---|
| GMM-UBM (GU) | 14.1897 | 0.8113 | 12.0528 | 0.7342 |
| DNN-UBM (DU) | 14.0957 | 0.7971 | **11.8396** ↑16% | **0.7221** ↑9% |
| Supervised GMM-UBM (SU) | 15.6090 | 0.8458 | 12.5789 | 0.7563 |
| BNF GMM-UBM (BU) | 15.0019 | 0.8468 | 13.9000 | 0.8220 |

## Conclusion

- Male speaker is more distinguishable
- Amount of unlabeled dataset has correlation with performance
- Insufficient knowledge of minor/major language is actively explored to discover rich labels
- Utilizing gender and language labels, language discrepancy is compensated

swshon@ispl.korea.ac.kr