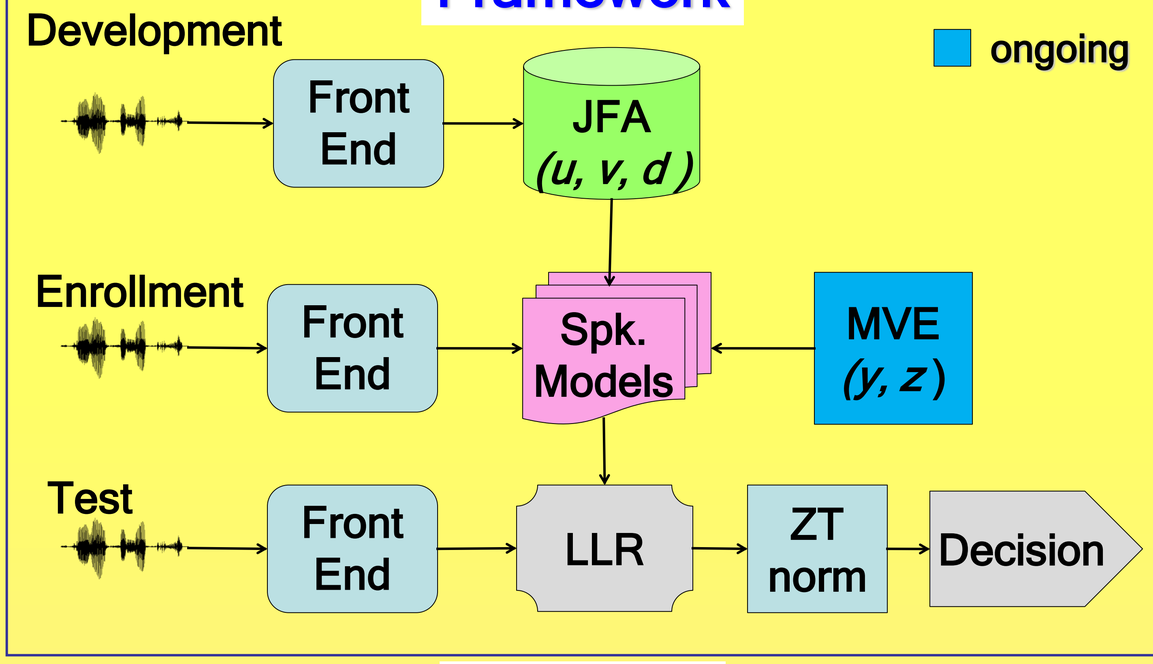
NTUT Submission for the NIST SRE 2010

Wei-Hsiung Ting and Yuan-Fu Liao
Department of Electronic Engineering and Graduate Institute of Computer and Communication Engineering,
National Taipei University of Technology, Taipei, Taiwan.
huaan.vgs@msa.hinet.net, yfliao@ntut.edu.tw

Overview

- **Task**
  - Core - Core
- **System Description – 3-level Normalization**
  - Frontend – histogram equalization (HEQ)
  - Model – Joint Factor Analysis (JFA)
  - Score – ZT Normalization
- **Ongoing development**
  - Minimum verification error (MVE)-based JFA training

Framework



Front End

- **Feature Extraction**
  - MFCC39 : 13 cepstral coefficients with C0, and their **Δ, and Δ² terms**
- **Target Speech Extraction**
  - Silence/Unvoiced Removal : voice detection using snack
  - Interviewee/Interviewer Separation: ASR labeling
- **Histogram Equalization (HEQ)**
  - CDF: Gaussian distribution

Joint Factor Analysis

- **UBM**
  - Gender-independent
  - 1024 Mixtures
- **Eigen-Space**
  - Eigen-voice *v*: 300
  - Eigen-channel 1 *u1*:100 (tel)
  - Eigen-channel 2 *u2*: 50 (mic)
  - Eigen-channel 3 *u3*: 50 (int)
  - Common basis *d*
- **JFA Algorithm**
  - ML then MD
- **Development Data**

UBM	SRE04-1, 3, 8, 16 sides, SRE05-mic, and SRE8-followup
v	SRE04-8, 16sides, SRE05-8conv4w, SRE06-8conv4w, SRE08-8conv4w, SRE05-mic, and SRE08-followup
u1	SRE04-8, 16 sides, and SRE08-8conv4w
u2	SRE05-mic
u3	SRE08-followup
d	SRE04-8, 16sides, SRE05-8conv4w, SRE06-8conv4w, SRE08-8conv4w, SRE05-mic, and SRE08-followup

MVE-based JFA training (ongoing)

- **MVE Criterion**
  - Class Conditional Likelihood Function

$$g_i(\chi;\Lambda) = (vy + dz)^* \Sigma^{-1}(F - N_m - N_c)$$

  - Class Misclassification Measure

$$d_i(\chi) = -g_i(\chi;\Lambda) + \max_{j, j\neq i} g_j(\chi;\Lambda)$$

  - Smoothed Zero-one function

$$l(d) = \frac{1}{1+\exp(-\gamma d + \theta)}$$

- **GPD training Algorithm**
  - General rule

$$\Lambda_{t+1} = \Lambda_t - \varepsilon_t \nabla l(\chi;\Lambda)|_{\Lambda=\Lambda_t}$$

  - Parameters adjustment (y & z)
    - Factor *y*

$$y_l^{(i)}(n+1) = y_l^{(i)}(n) - \varepsilon \frac{\partial l_i(\chi;\Lambda)}{\partial y_l^{(i)}} | \Lambda = \Lambda_n,\ l = 1,2,...,R_s$$

    - Factor *z*

$$z_l^{(i)}(n+1) = z_l^{(i)}(n) - \varepsilon \frac{\partial l_i(\chi;\Lambda)}{\partial z_l^{(i)}} | \Lambda = \Lambda_n, l = 1,2,...,CF$$

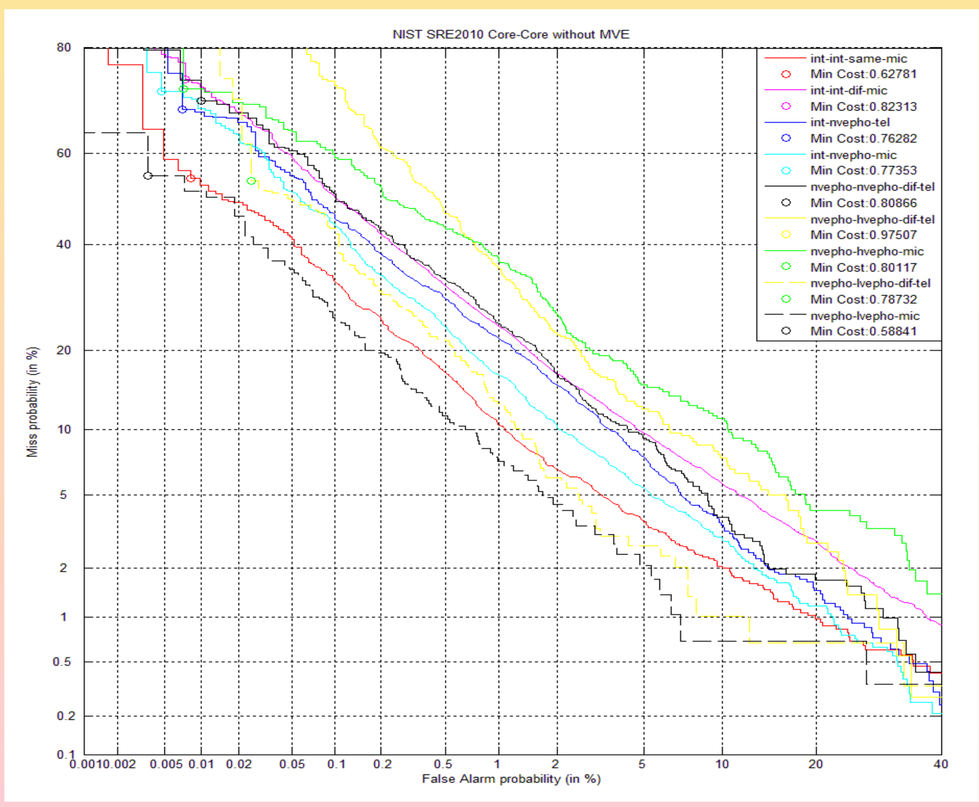
Scoring

- **LLR**
  - Linear Scoring – *LPT* assumption

$$LLR = (vy + dz)^* \Sigma^{-1}(F - N_m - N_c)$$

- **ZT normalization**
  - T-norm and then Z-norm (1000) from SRE05-mic & SRE06-1conv4w

Experimental Results



Conclusions

- **Our first try is working but not good enough yet**
- **Weakness**
  - Too simple front-end and JFA model
  - Didn't consider high/low vocal effect
- **Ongoing tasks**
  - MVE-based JFA training
  - Prosodic modeling