

huaan.vgs@msa.hinet.net, yfliao@ntut.edu.tw

- **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

The diagram illustrates the JFA framework, which is divided into three main stages: Development, Enrollment, and Test. A legend indicates that blue boxes represent 'ongoing' processes.

- Development:** An input waveform is processed by a 'Front End' block, which then feeds into a 'JFA (u, v, d)' cylinder. This stage is marked as 'ongoing'.
- Enrollment:** An input waveform is processed by a 'Front End' block, which feeds into 'Spk. Models' (represented by a stack of pink rectangles). A blue box labeled 'MVE (y, z)' also feeds into the 'Spk. Models'. This stage is marked as 'ongoing'.
- Test:** An input waveform is processed by a 'Front End' block, which feeds into an 'LLR' (Log-Likelihood Ratio) block. The 'LLR' block feeds into a 'ZT norm' block, which finally feeds into a 'Decision' block (represented by a chevron shape).

Arrows indicate the flow of data between these components.

- **Feature Extraction**
 - MFCC39 : 13 cepstral coefficients with C0, and their Δ , and Δ^2 terms
- **Target Speech Extraction**
 - Silence/Unvoiced Removal : voice detection using snack
 - Interviewee/Interviewer Separation: ASR labeling
- **Histogram Equalization (HEQ)**
 - CDF: Gaussian distribution

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

- **General rule**

$$\Lambda_{t+1} = \Lambda_t - \varepsilon_t \nabla l(\chi; \Lambda) |_{\Lambda = \Lambda_t}$$
- **Parameters adjustment (y & z)**
 - **Factor y**

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

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

- **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

NIST SRE2010 Core-Core without MVE

Legend:

- int-int-same-mic
- Min Cost 0.62781
- int-int-dif-mic
- Min Cost 0.82313
- int-nvpho-tel
- Min Cost 0.76282
- int-nvpho-mic
- Min Cost 0.77353
- nvpho-nvpho-dif-tel
- Min Cost 0.80866
- nvpho-nvpho-dif-tel
- Min Cost 0.97507
- nvpho-nvpho-mic
- Min Cost 0.80117
- nvpho-nvpho-dif-tel
- Min Cost 0.78732
- nvpho-nvpho-mic
- Min Cost 0.58411

- 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