MASTER THESIS

# Comparison of Observation Confidence Estimators in Modified Adaptive GMM for Robust Speaker Verification

Ma Xinjie

Supervisor: Professor KIM Jin Young

# Content

INTRODUCTION

Proposed Approach

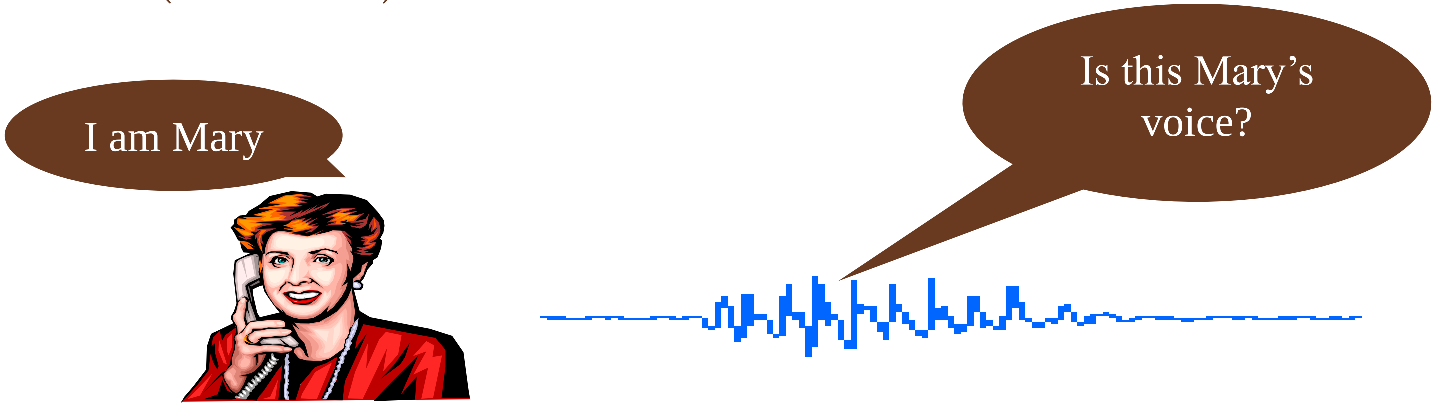
- Modified Adaptive GMM
- Observation Confidence Estimation
  - MMSE log-STSA
  - Low Rank Matrix Recovery
  - Multiple Low Rank Representation
  - Adaptive Multiple Low Rank Representation

EXPERIMENTAL RESULTS

CONCLUSION

# Introduction

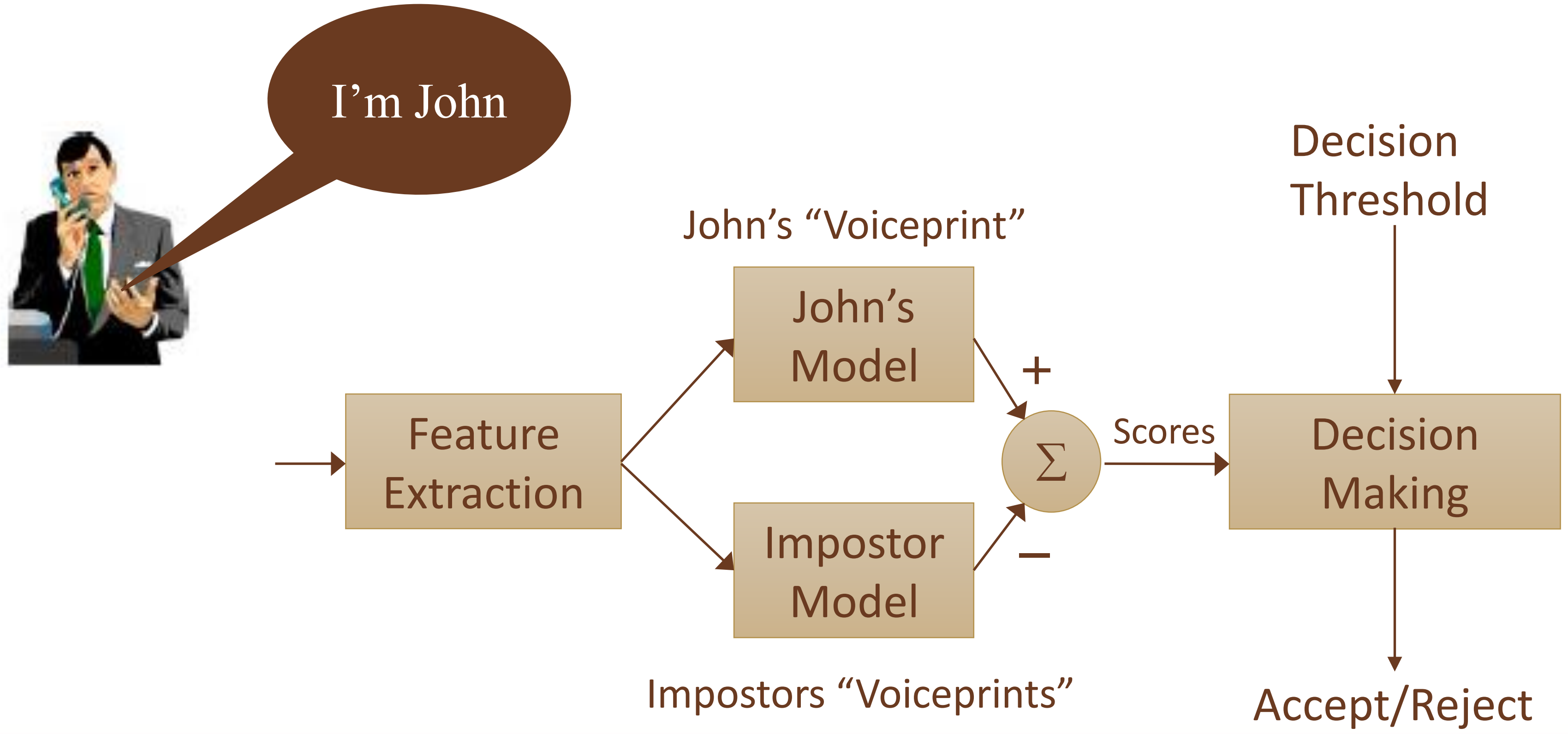
- Speaker Verification
  - To verify the identify of a claimant based on his/her own voices.
  - Binary classification (detection) task




- Applications: security system, access control, telephone banking, …
- Challenges: noise problem, channel effect, text-independent recognition, …

# Introduction

- Speaker verification process



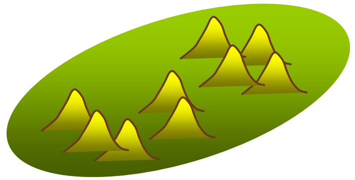
# Motivation and Contribution

- We develop the text-independent speaker verification system under uncontrolled noisy environments.
- Contribution:
  - Propose to use the LRR for observation confidence calculation in the MAGMM framework.
  - Develop an adaptive MLRR method that is more robust and effective for speaker-verification system.
  - Propose a fusion method that employ both MMSE log-STSA and LRR methods to calculate observation confidence value.

# Modified Adaptive GMM

- Gaussian Mixture Model (GMM) is used to represent each speaker by a finite mixture of multivariate Gaussians

$$p(x|\lambda) = \sum_{i=1}^{M} \omega_i f(x|\mu_i, \Sigma_i)$$



- The acoustic vectors of a general population is modeled by another GMM called the Universal Background Model (UBM)

$$\lambda^{(ubm)} = \{\omega_j^{(ubm)}, \mu_j^{(ubm)}, \omega_j^{(ubm)}\}_{j=1}^{M}$$

- The objective function (GMM likelihood) can described as:

$$L(X|\lambda) = log\big(p(X|\lambda)\big) = \sum_{n=1}^{N} \log\big(p(\boldsymbol{x}_n|\lambda)\big)$$

# Modified Adaptive GMM

- In baseline GMM training, observation vectors are considered as clean or free from noise.
- However, speech signals are often **affected and corrupted by various types of noise**.
- Each observation vector may have a different weighted factor and should be treated differently.
- We define $\boldsymbol{\rho_n}$ **as the confidence value** of the *n-th* observation ranging from 0 to 1.
- With observation pairs of $\{\boldsymbol{x_n}, \boldsymbol{\rho_n}\}$, the objective function can be modified by considering the confidence measure as:

$$L(X|\lambda) = log\big(p(X|\lambda)\big) = \sum_{n=1}^{N} \rho_n \log\big(p(\boldsymbol{x}_n|\lambda)\big)$$

# Modified Adaptive GMM

| Observation confidence computation |
|---|
| **Input**: input speech signal $s(t)$<br><br>**Output**: observation-confidence value |
| 1. Find **enhanced (reference) speech** $r(t)$<br><br>2. Apply the **amplitude normalization** to the enhanced speech<br><br>3. Compute the **frame SNR** values between the input speech and the enhanced speech.<br><br>4. Convert the **frame SNR values into the observation-confidence** values by using the simple **sigmoid function**. |

# Minimum Mean Square Error Logarithm Short-time Spectral Amplitude (MMSE log-STSA)

- Let $x(t)$ and $d(t)$ denote the clean speech and noise part, respectively. The noisy observation $y(t)$ is given by

$$y(t) = x(t) + d(t), \qquad 0 \leq t \leq T$$

- Let $X_k = A_k e^{j\alpha_k}$, $D_k$ and $Y_k = R_k e^{j\vartheta_k}$ denote the *k-th* Fourier expansion coefficient of the clean speech $x(t)$, the noise part $d(t)$ and the observation signal $y(t)$, respectively.

- Our purpose is to find the estimator $\hat{A}_k$ in order to minimize the distortion measure under noisy observation $y(t)$, which is given as follows:

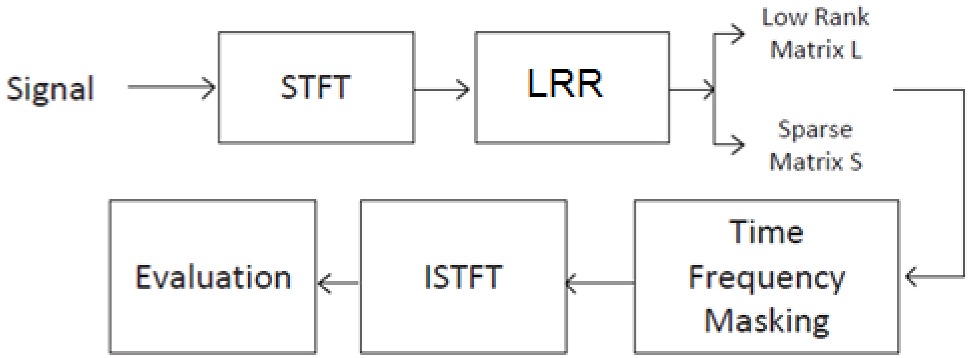
$$\hat{A}_k = e^{\{E[ln \ A_k|Y_k]\}}$$

$$\hat{A}_k = \frac{\xi_k}{1+\xi_k} exp\left\{\frac{1}{2}\int_{\upsilon_k}^{\infty} \frac{e^{-1}}{t} dt\right\} R_k \qquad \upsilon_k = \frac{\xi_k}{1+\xi_k}\gamma_k;\ \xi_k = \frac{\lambda_x(k)}{\lambda_d(k)};\ \gamma_k = \frac{R_k^2}{\lambda_d(k)}\ .$$

where $\xi_k$ and $\gamma_k$ are described as the a priori and a posteriori signal-to-noise ratio (SNR), respectively.

Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 33, No. 2, pp. 443-445, 1985.

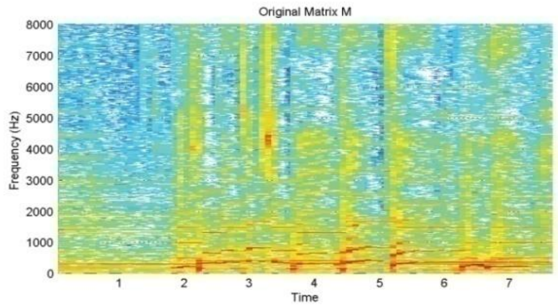
# Low Rank Matrix Recovery (LRR)



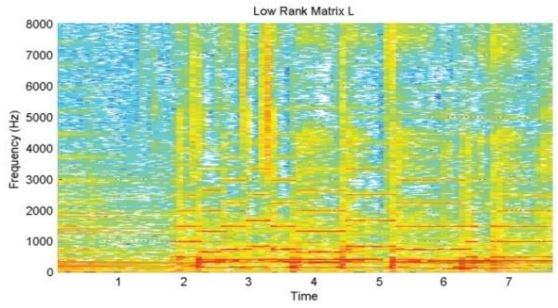

- Time frequency masking $M_b$ as follows:

$$M_b(m,n) = \begin{cases} 1 & |S(m,n)| > \text{gain} * |L(m,n)| \\ 0 & \text{otherwise} \end{cases}$$

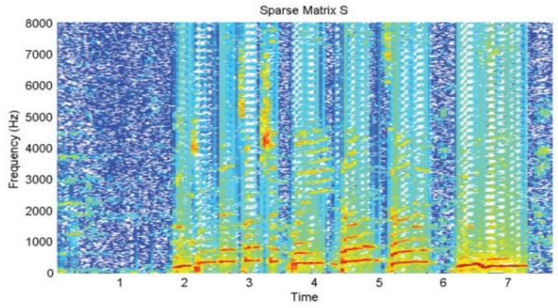
$$\begin{cases} X_{\text{singing}}(m,n) &= M_b(m,n)M(m,n) \\ X_{\text{music}}(m,n) &= (1 - M_b(m,n))M(m,n) \end{cases}$$




(a) Original Matrix $M$




(b) Low-Rank Matrix $L$




(c) Sparse Matrix $M$

P. S. Huang, S. D. Chen, P. Smaragdis and H. M. Johnson, "Singing-Voice Separation from Monaural Recordings Using Robust Principal Component Analysis," Proc. IEEE ICASSP, pp. 57–60, 2012.

# Multiple Low Rank Representation (MLRR)

- We are able to obtain the **low-rank representations of *X*** with respect to **multiple dictionaries** ($A_1$, $A_2$ : one dictionary for speech and the other for the noisy components).

$$\min_{Z_1, Z_2} \alpha \|Z_1\|_* + \beta \|Z_2\|_* + \lambda \|X - A_1 Z_1 - A_2 Z_2\|_1$$

- Equivalent equation

$$\min_{Z_1, Z_2, J_1, J_2, E} \alpha \|J_1\|_* + \beta \|J_2\|_* + \lambda \|E\|_1$$

$$subjectto \quad \begin{matrix} X = A_1 Z_1 + A_2 Z_2 + E \\ Z_1 = J_1, Z_2 = J_2 \end{matrix}$$

- The **Augmented Lagrange Multiplier (ALM)** can be used to find the solution for this problem. For example, $J_1$ can be updated as follows

$$J_1 = argmin \alpha \|J_1\|_* + \frac{\mu}{2} \left\| J_1 - \left( Z_1 + \frac{Y_1}{\mu} \right) \right\|_F^2$$

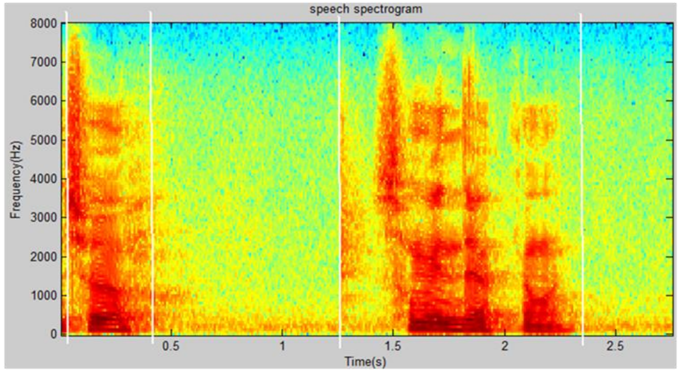
Y. H. Yang, "Low-Rank Representation of Both Singing Voice and Music Accompaniment via Learned Dictionaries," Proc. Int. Society of Music Information Retrieval, pp. 427-432, 2013.

# Adaptive Multiple Low Rank Representation (AMLRR)

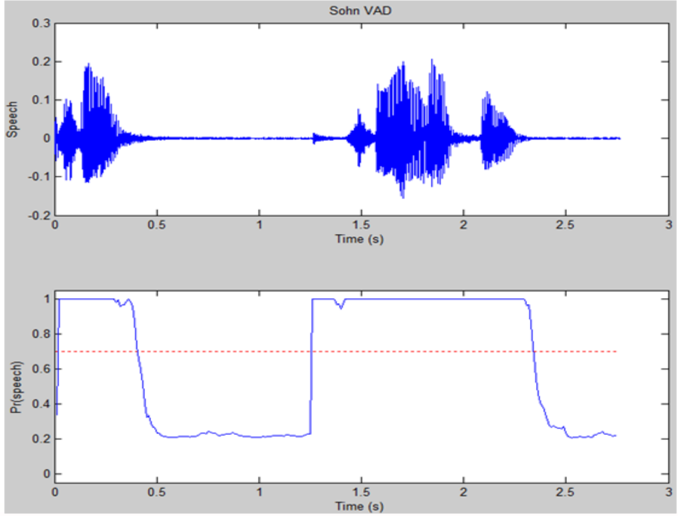
- We introduce to **impale the penalty factor based on the voice-activity side information** of the input signal to improve the quality of the enhanced speech
- To apply the **adaptive MLRR**, we first have to **divide the magnitude spectrogram** $X$ **into column-block** $[X^1, X^2, \dots, X^N]$, and **compute different penalty factors** $\alpha^l, l = 1, \dots, N$ **for each block**.
- Voice detection method (based Jongseo Sohn's method) is used to divide $X$ into blocks.
- An adaptive version for the updating of $J_1 = [J_1^1, \dots, J_1^l, \dots, J_1^N]$ can be described as follows

$$J_1^l = argmin\ \alpha^l \|J_1^l\|_* + \frac{\mu}{2} \left\| J_1^l - \left( Z_1^l + \frac{Y_1^l}{\mu} \right) \right\|_F^2$$

# Adaptive Multiple Low Rank Representation (AMLRR)




a) Segmentation of the magnitude spectrogram into 4 consecutive blocks of speech and non-speech segments.




b) Original signal and probability of speech and non-speech parts.

We set $\alpha^{l}$ as the **mean value of the speech probability** in the corresponding block.

# Observation Confidence based on AMLRR

- Calculate frame SNR values

$$SNR_n = 10\, log \frac{\sum_{i \in frame_n} r_n^2[i]}{\sum_{i \in frame_n} (s_n[i] - r_n[i])^2}$$

where $SNR_n$ denotes the SNR value of the *n-th* frame, and $s_n[i]$ and $r_n[i]$ are the noisy and the reference (enhanced) speech samples, respectively, in the *n-th* frame of the analyzed signal.

- Perform Min-Max normalization

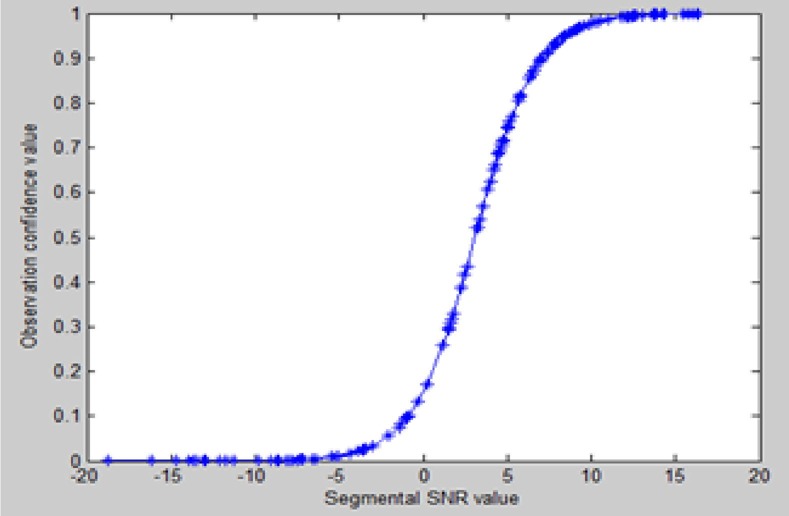
$$X_{norm} = \frac{(X - X_{\min}) * (Input_{max} - Input_{min})}{X_{max} - X_{min}} + Input_{min}$$

# Observation Confidence based on AMLRR

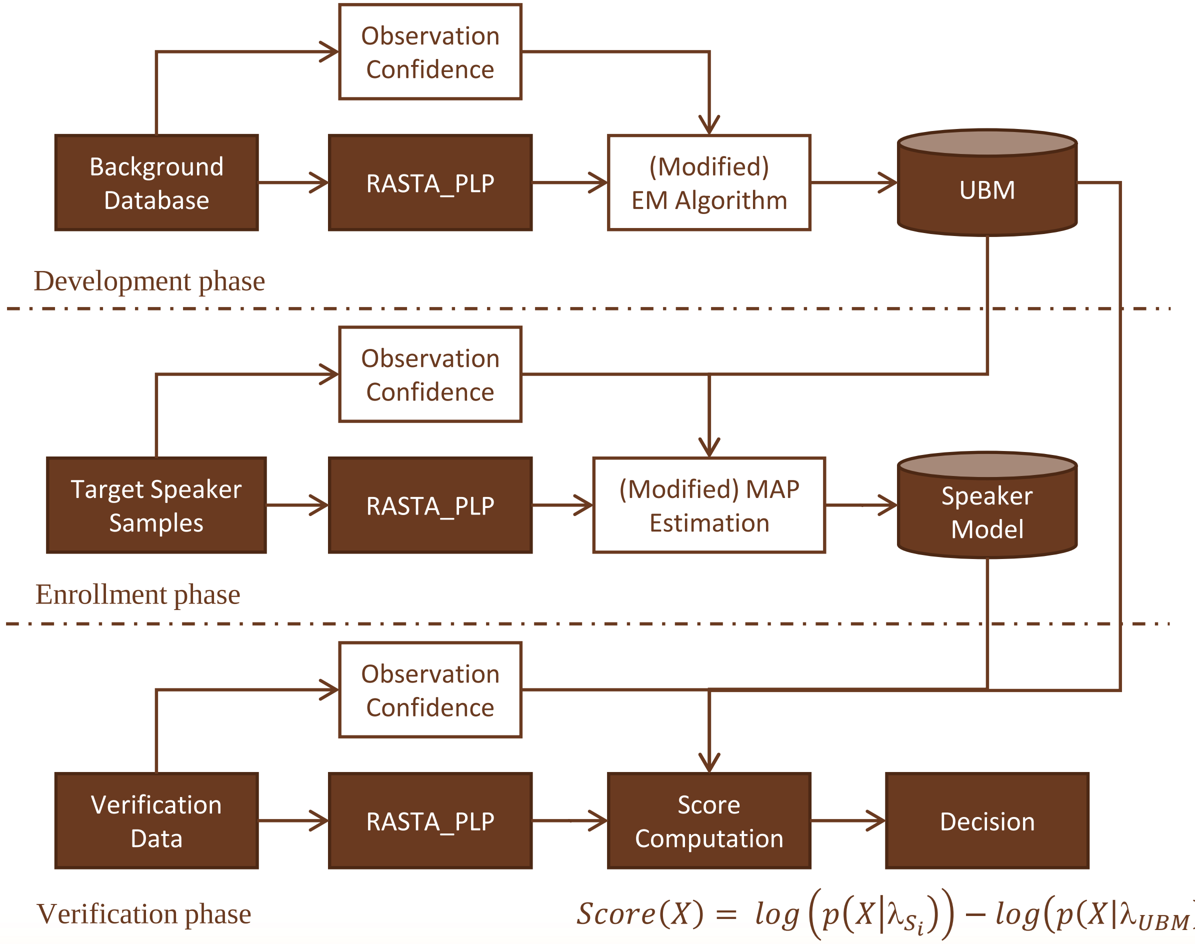
- Convert frame SNR value to observation confident value using sigmoid function

$$\rho_n = \frac{1}{1 + e^{-s(SNR_n - c)}}$$

where $c$ is mean of SNR of a input signal, $s$ is chosen of 0.55



# Modified Adaptive GMM




$$Score(X) = log\left(p\left(X|\lambda_{S_i}\right)\right) - log\left(p\left(X|\lambda_{UBM}\right)\right)$$

# Experimental Results

| | |
|---|---|
| **Database** | Korean drama ( “You come from the star” – 12 first episodes) |
| **Number of speakers** | **7** |
| **UBM model (Impostor model)** | • UBM is trained using random 3000 samples (other speakers, events, …) |
| **Speaker Enrollment** | • 7 target speakers using Clean speeches (from episodes 1 and 6) |
| **Verification Process** | • Clean Test (clean speeches from episodes 7 and 12 + 1200 random samples from other events)<br>• Noisy Test (noisy speeches from episodes 7 and 12 + 1200 random samples from other events) |
| **Modified Adaptive GMM Training** | Modified EM algorithm<br>• Diagonal covariance matrix<br>Modified MAP estimation<br>• Relevance factor: $r = 16$<br>• *128* mixtures |

# Experimental Results

*Information of training and test*

| Speakers | Number of Samples | | |
|---|---|---|---|
| | Training | Testing | |
| | | Clean Test | Noisy Test |
| Speaker 1 | 170 | 208 | 611 |
| Speaker 2 | 348 | 513 | 1167 |
| Speaker 3 | 44 | 114 | 179 |
| Speaker 4 | 162 | 137 | 326 |
| Speaker 5 | 53 | 22 | 216 |
| Speaker 6 | 57 | 61 | 33 |
| Speaker 7 | 11 | 87 | 111 |

# Experimental Results

- ***Feature extraction*** (Relative Spectral Transform - Perceptual Linear Prediction)
  - RASTA_PLP: 42 mels (13 cepstral coefficients + delta + double-delta).

- ***Evaluation***
  - Equal error rate (EER)
  - False reject (miss detection) probability = false accept (false alarm) probability

# Experimental Results

- To enhance performance of the system, we proposed to combine two frame SNR values using MMSE log-STSA and LRR.
- A simple linear combination method is used to estimate observation confidence values.

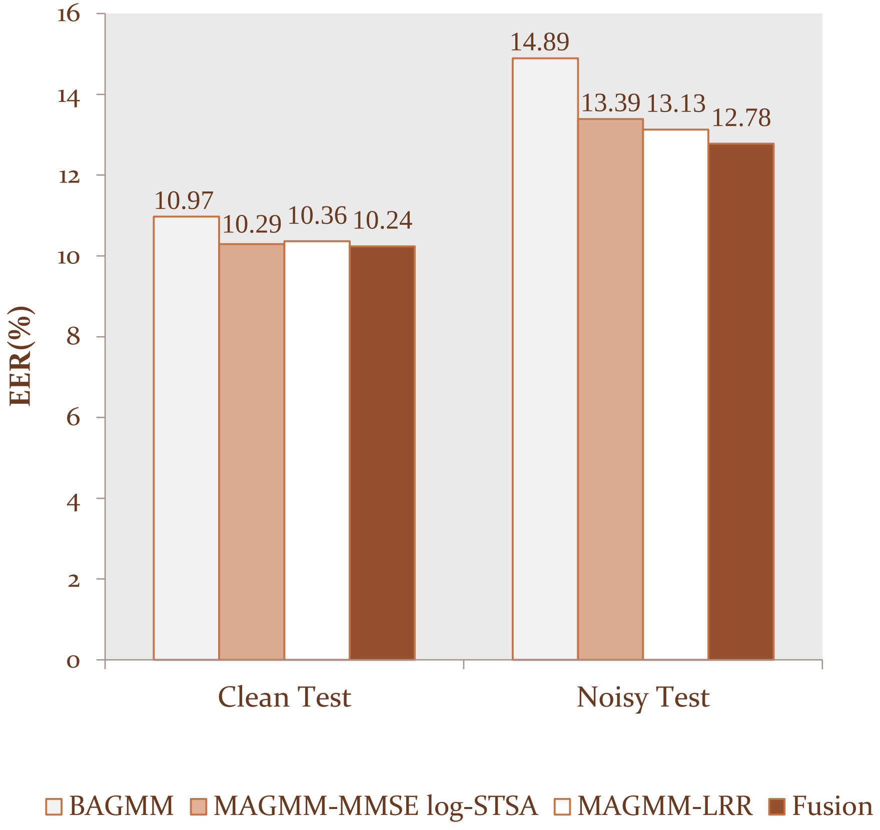
$$FSNR_{SL} = a_S FSNR_S + a_L FSNR_L$$

- We randomly generate weights $a_S$ and $a_L$ and evaluate the performance of system.
- We do experiments 30 times and compare the performances.

# Experimental Results

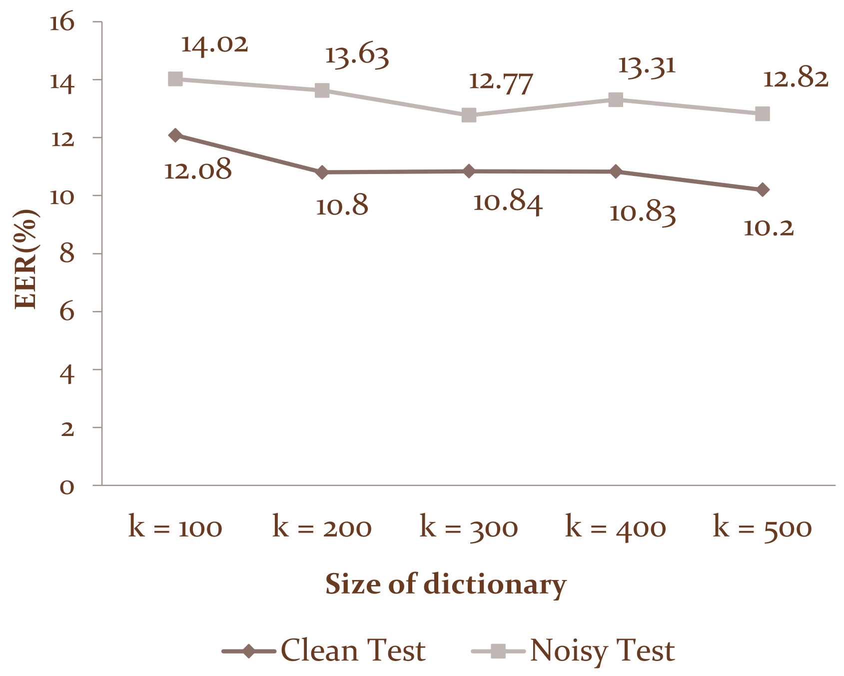
*Performance comparison of the MMSE log-STSA and the LRR.*

| System | EER (%) | |
|---|---|---|
| | Clean Test | Noisy Test |
| **BAGMM** | 10.97 | 14.89 |
| **MAGMM-MMSE log-STSA** | 10.29 | 13.39 |
| **MAGMM-LRR without normalization** | 10.53 | 13.55 |
| **MAGMM-LRR with normalization** | 10.36 | 13.13 |

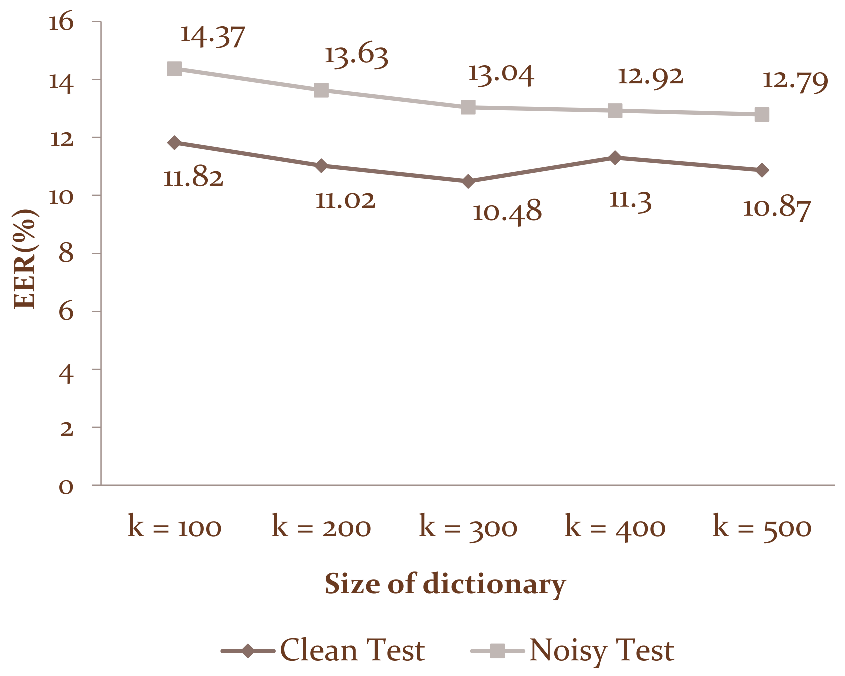



Performance comparison of the fusion of frames SNR values using the MMSE log-STSA and the LRR with other methods where ($a_S$ = 0.2609 and $a_L$ = 0.7391).

# Experimental Results




Performance of the MAGMM-MLRR system.



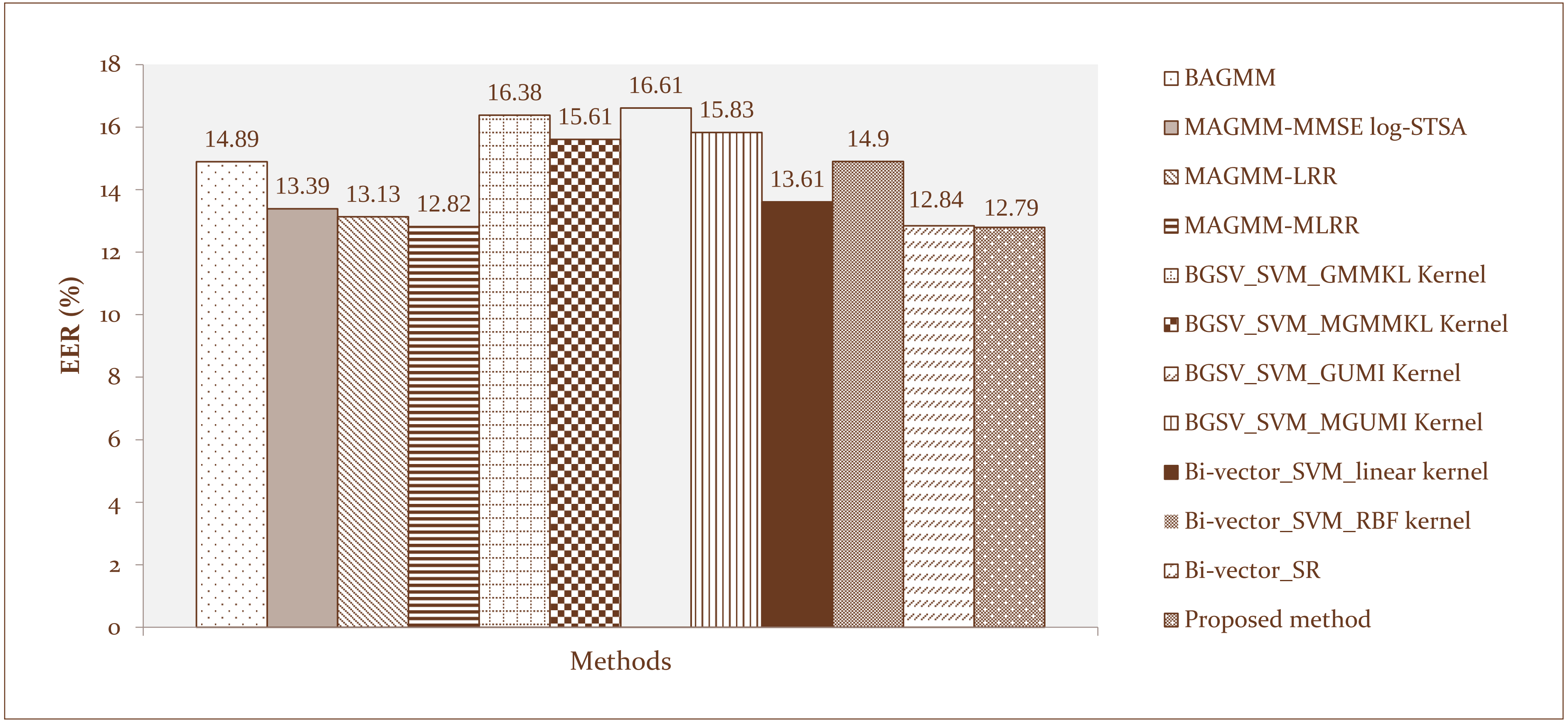

Performance of the MAGMM-Adaptive MLRR system.

# Experimental Results

*Performance comparison of various systems based on the AGMM framework.*

| System | EER(%) | |
|---|---|---|
| | Clean Test | Noisy Test |
| **BAGMM** | 10.97 | 14.89 |
| **MAGMM-MMSE log-STSA** | 10.29 | 13.39 |
| **MAGMM-LRR** | 10.36 | 13.13 |
| **MAGMM-MLRR** | 10.20 | 12.82 |
| **MAGMM-Adaptive MLRR (proposed method)** | 10.87 | 12.79 |

# Experimental Results




Comparison of various speaker-verification systems under noisy conditions.

# CONCLUSION

- The LRR and AMLRR are proposed to find the observation confidence that is incorporated into the MAGMM model.
- A comparison of various techniques for calculating observation confidence is discussed.
- A fusion of the observation confidence estimation methods is used to enhance the performance of verification system.
- Future works:
  - different techniques for feature extraction .
  - channel/session compensation.
  - extend database

# Thank You

# 감사합니다