

NIST SRE-06: Speaker Recognition Evaluation



Sanjay Patil, Vinod Prakash, Moosa Hassan,
Wooil Kim, Pongtep Angkititrakul, John H.L. Hansen

{sap061000, vxp052000, syed.hassan, wikim, angkitit, john.hansen}@utdallas.edu

Center for Robust Speech Systems (CRSS)
<http://crss.utdallas.edu>

Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.



01	System Configuration	04	TEO BACKGROUND	07	GMM-based Feature Compensation	10	comparative analysis of our systems	13	Results – 10sec4w-10sec4w with Feature Compensation
UTD	UTD	UTD	UTD	UTD	UTD	UTD	UTD	UTD	
02	System Description (10sec4w-10sec4w)	05	Frame Selection using TEO	08	GMM-based Feature Compensation	11	Results – 1conv4w-1conv4w	14	Conclusions
UTD	UTD	UTD	UTD	UTD	UTD	UTD	UTD	UTD	
03	System Description (1conv4w-1conv4w)	06	Average TEO Energy and Corresponding Thresholds	09	Employing Multiple Environmental Models	12	Results – 10sec4w-10sec4w	15	References:
UTD	UTD	UTD	UTD	UTD	UTD	UTD	UTD	UTD	

01 System Configuration

- Features:
 - 19 MFCC (HTK Tool) : 20 ms window, 10 ms skip rate
 - 300 – 3,400 Hz band limited
 - Frame selection:
 - (i) Normalized energy (ii) Adaptive TEO energy
 - CMS
- Models:
 - GMM-UBM with MAP adaptation
 - Gender-dependent UBM – trained from NIST 2004, all the train files
 - Mixture components: 256 (1conv4w), 64 (10sec4w)
 - Means-only MAP-adaptation
 - Score Normalization – Z-norm

02 System Description (10sec4w-10sec4w)

- Three systems submitted
 - Energy-based frame selection + Znorm
 - TEO-based frame selection + Znorm
 - Feature-compensated system
- Znorm:
 - 200 files from NIST 2004 test files
- Threshold:
 - $P_{\text{false alarm}} = 0.2$ based on NIST 2004 test sets

03 System Description (1conv4w-1conv4w)

- One system submitted
 - Energy-based frame selection + Znorm
- Znorm:
 - 100 files from NIST 2004 test files
- Threshold:
 - $P_{\text{false alarm}} = 0.2$ based on NIST 2004 test sets

04 TEO BACKGROUND

- Teager Energy Operator (TEO)

$$\Psi_t[x(t)] = \dot{x}(t)^2 - x(t)\ddot{x}(t) \quad \Psi_d[x(n)] = x^2(n) - x(n+1)x(n-1)$$
 - When applied to single frequency signal,
$$x(t) = A \cos(\omega t + \phi), \quad \Psi[x(t)] = A^2 \omega^2$$

$$x(n) = A \cos(\Omega n + \phi), \quad \Psi[x(n)] = A^2 \Omega^2$$
- Used for speech resonance in practice
 - Gabor filter: $h(t) = \exp(-\alpha^2 t^2) \cos(\omega_0 t)$
 - $H(\omega) = \frac{\sqrt{\pi}}{2} \left[\exp\left(-\frac{(\omega-\omega_0)^2}{4\alpha^2}\right) + \exp\left(-\frac{(\omega+\omega_0)^2}{4\alpha^2}\right) \right]$
 - Output of Gabor filter considered as AM-FM signal

05 Frame Selection using TEO

- TEO (Teager Energy Operator)
 - If $T_{\text{sp}} > T_{\text{noise}}$, \rightarrow current window is a speech candidate.
 - If $T_{\text{sp}} < T_{\text{noise}}$, \rightarrow current window is a noise candidate.
 - Speech energy threshold: T_{noise} : the noise energy threshold
- Thresholds updating
 - when the current analysis window is a speech candidate
$$\frac{T_{\text{sp}}}{T_{\text{noise}}} = \alpha \times (T_{\text{sp}}) + (1-\alpha) \times \bar{T}_{\text{sp}}$$

$$\bar{T}_{\text{noise}} = \bar{T}_{\text{noise}} \times \frac{T_{\text{noise}}}{T_{\text{sp}}}$$
- when the current analysis window is a noise candidate

$$\frac{T_{\text{sp}}}{T_{\text{noise}}} = \beta \times (\frac{T_{\text{sp}}}{T_{\text{noise}}}) + (1-\beta) \times \bar{T}_{\text{sp}}$$

$$\bar{T}_{\text{noise}} = \bar{T}_{\text{noise}} \times \frac{T_{\text{noise}}}{T_{\text{sp}}}$$

06 Average TEO Energy and Corresponding Thresholds

Figure showing Average TEO Energy versus corresponding thresholds. It includes a waveform plot of 'Digital Noise Source' and a plot of 'Averaged TEO Energy' versus 'noise threshold'.

07 GMM-based Feature Compensation

- Estimate statistical transformation of clean speech distribution under noisy condition
- Compensate input speech using the difference

Diagram illustrating the GMM-based Feature Compensation process:

```

    graph LR
      CS[clean speech GMH] --> NSM[NOISY SPEECH MODEL ESTIMATION]
      NSM --> NS[noisy speech]
      NS --> FC[FEATURE COMPENSATION (MMSE)]
      FC --> RO[RESTORED OUTPUT]
      NI[NOISY INPUT] --> FC
  
```

08 GMM-based Feature Compensation

- Estimate pdf of clean speech
- Estimate pdf of noise-corrupted speech
- Reconstruct based on MMSE

Figure showing the results of 1conv4w-1conv4w system:

- Each file conversation length: 5 Min. Train & 5 Min. Test
- EER (%): Overall: 18.94, Male: 16.34, Combined: 19.48
- English (test & train): 19.26, Non-English (test and train): 20.02

09 Employing Multiple Environmental Models

Diagram illustrating the Employing Multiple Environmental Models process:

```

    graph LR
      subgraph Pre[Pre-obtained]
        direction TB
        C1[clean speech GMH database 1] --> N1[noisy speech GMH 1]
        C2[clean speech GMH database 2] --> N2[noisy speech GMH 2]
        C3[clean speech GMH database E] --> N3[noisy speech GMH E]
      end
      N1 --> NSM[NOISY SPEECH MODEL ESTIMATION]
      N2 --> NSM
      N3 --> NSM
      NSM --> FC[FEATURE COMPENSATION (MMSE)]
      FC --> RO[RESTORED OUTPUT]
      NI[NOISY INPUT] --> FC
  
```

10 comparative analysis of our systems

Figure showing comparative analysis of our systems:

- EER: Combined: 19.48, Female: 18.94, Male: 16.34, English: 19.26, Non-English: 20.02
- Min-DCF: Combined: 0.019, Female: 0.018, Male: 0.017, English: 0.018, Non-English: 0.020

11 Results – 1conv4w-1conv4w

Figure showing the results of 1conv4w-1conv4w system:

- Each file conversation length: 5 Min. Train & 5 Min. Test
- EER (%): Overall: 18.94, Male: 16.34, Combined: 19.48
- English (test & train): 19.26, Non-English (test and train): 20.02

12 Results – 10sec4w-10sec4w

Figure showing the results of 10sec4w-10sec4w system:

- 10 Sec. Train & 10 Sec. Test
- EER (%): Overall: Female: 32.80, Male: 26.70, Combined: 30.72
- English (test & train): 31.26, Non-English (test and train): 29.76

13 Results – 10sec4w-10sec4w with Feature Compensation

Figure showing the results of 10sec4w-10sec4w with Feature Compensation system:

- 10 Sec. Train & 10 Sec. Test
- EER (%): Overall: Female: 21.22, Male: 20.86, Combined: 20.98

14 Conclusions

- Conclusion:
 - Achieved our aim of reaching within the top five as compared to the previous year's performance
 - a learning experience
 - helped us to understand our strengths and weakness
 - Future work:
 - implement a fusion system

15 References:

- Moreno, P.J., "Speech Recognition in Noisy Environments." PhD thesis, Carnegie Mellon University, 1996.
- Reynolds, D.A., Quater, T.F., and Dunn, R.B., "Speaker verification using adapted Gaussian Mixture Models," *Digital Signal Processing*, 10(1-3):19-41, Jan/April/Jul 2000.
- Kuerten, J., "Speaker Verification using Feature Normalization for text-independent Speaker Verification systems," *Digital Signal processing*, 10(1), 2000.
- Teager H.M. and Teager S.M., "Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract," *Speech Production and Speech Modeling*, Hardcastle W.J. and Marschall A. (eds.), Kluwer Academic Publishers, Boston, USA, 1990.