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.

HLTRI

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 UBMs — 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:
  - $P_{(false\ alarm)}$ = 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:
  - $P_{(false\ alarm)}$ = 0.2 based on NIST 2004 test sets

04

TEO BACKGROUND

- Teager Energy Operator (TEO)

$$\Psi_c[x(t)] = [\dot{x}(t)]^2 - x(t)\ddot{x}(t) \quad \Psi_d[x(n)] = x^2(n) - x(n+1)x(n-1)$$

  - When applied to single frequency signal,

$$x(t) = A\cos(\omega t + \phi), \quad \Psi[x(t)] = A^2\omega^2$$
$$x(n) = A\cos(\Omega n + \phi), \quad \Psi[x(n)] = A^2\sin^2\Omega$$

- Used for speech resonance in practice
  - Gabor filter: $h(t) = \exp(-\alpha^2 t^2)\cos(\omega_c t)$

$$H(\omega) = \frac{\sqrt{\pi}}{2\alpha}\left\{\exp\left[-\frac{(\omega-\omega_c)^2}{4\alpha^2}\right] + \exp\left[-\frac{(\omega+\omega_c)^2}{4\alpha^2}\right]\right\}$$

  - Output of Gabor filter considered as AM-FM signal

05

Frame Selection using TEO

- TEO (Teager Energy Operator)
  - If $\bar{T}_{signal} > \bar{T}_{speech}$, → current window is a speech candidate;
  - If $\bar{T}_{signal} < \bar{T}_{speech}$, → current window is a noise candidate.
    - $\bar{T}_{speech}$: speech energy threshold; $\bar{T}_{noise}$: the noise energy threshold
  - Thresholds updating
    - when the current analysis window is a speech candidate

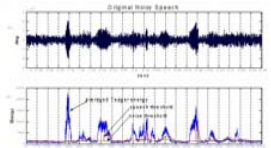
$$\bar{T}_{speech}^{new} = \alpha \times (\bar{T}_{speech}^{old}) + (1-\alpha) \times \bar{T}_{signal}$$
$$\bar{T}_{speech} = \rho_{speech} \times \bar{T}_{speech}^{new}$$

    - when the current analysis window is a noise candidate

$$\bar{T}_{noise}^{new} = \beta \times (\bar{T}_{noise}^{old}) + (1-\beta) \times \bar{T}_{signal}$$
$$\bar{T}_{noise} = \rho_{noise} \times \bar{T}_{noise}^{new}$$

06

Average TEO Energy and Corresponding Thresholds

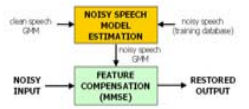


Averaged TEO Energy versus corresponding thresholds
(a) Sample Noisy speech waveform (from CU-Move)
(b) TEO profile and resulting speech and noise thresholds

07

GMM-based Feature Compensation

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



08

GMM-based Feature Compensation

- Estimate pdf of clean speech

$$p(\mathbf{x}) = \sum_{k=1}^{C} \omega_k N(\mathbf{x}; \boldsymbol{\mu}_{x,k}, \boldsymbol{\Sigma}_{x,k})$$

- Estimate pdf of noise-corrupted speech

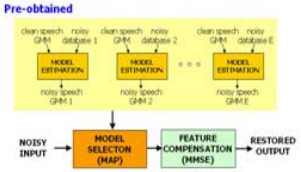
$$p(\mathbf{y}) = \sum_{k=1}^{C} \omega_k N(\mathbf{y}; \boldsymbol{\mu}_{y,k}, \boldsymbol{\Sigma}_{y,k})$$
$$\boldsymbol{\mu}_{y,k} = \boldsymbol{\mu}_{x,k} + \boldsymbol{\mu}_r \ , \ \boldsymbol{\Sigma}_{y,k} = \boldsymbol{\Sigma}_{x,k} + \mathbf{R}_r$$

- Reconstruct based on MMSE

$$\hat{\mathbf{x}}_{MMSE} = E[\mathbf{x}|\mathbf{y}]$$

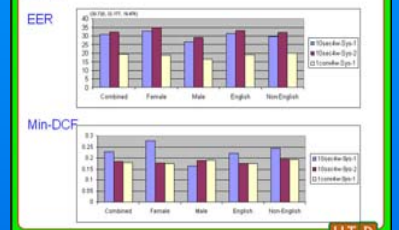
09

Employing Multiple Environmental Models



10

comparative analysis of our systems

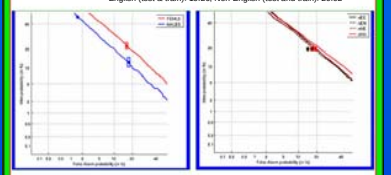


11

Results – 1conv4w-1conv4w

* Each file conversation length: 5 Min. Train & 5 Min. Test

EER (%) Overall:- Female: 18.94, Male: 16.34, Combined: 19.48
English (test & train): 19.26, Non-English (test and train): 20.02

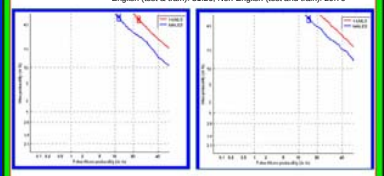


12

Results – 10sec4w-10sec4w

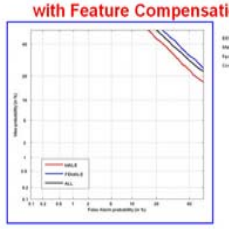
10 Sec. Train & 10 Sec. Test

EER (%) Overall:- Female: 32.89, 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



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:

- Morena, P.J., "Speech Recognition in Noisy Environments," PhD thesis, Carnegie Mellon University, 1996.
- Reynolds, D.A., Quatieri T.F., and Dunn R.B., "Speaker verification using adapted Gaussian Mixture Models," Digital Signal Processing, 10(1-3):19-41. Jan/Apr/Jul 2000.
- Auckenthaler R., Carey M., Lloyd-Thomas H., "Score 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 Marchal A (eds), Kluwer Academic Publishers, Boston, USA, 1990.