R64: NIST 2005 Speaker Recognition Evaluation

Walt Andrews, R64
Jaime Hernández-Cordero, R64

NIST 2005 Speaker Recognition Workshop
June 7-8 2005
Montreal, Canada

Outline

- System Description
  - Evaluation Submission
- System Block Diagram
  - Phonetic Speaker Recognition
  - Parallel Phonetic Speaker Recognition
- Conclusion

System Description

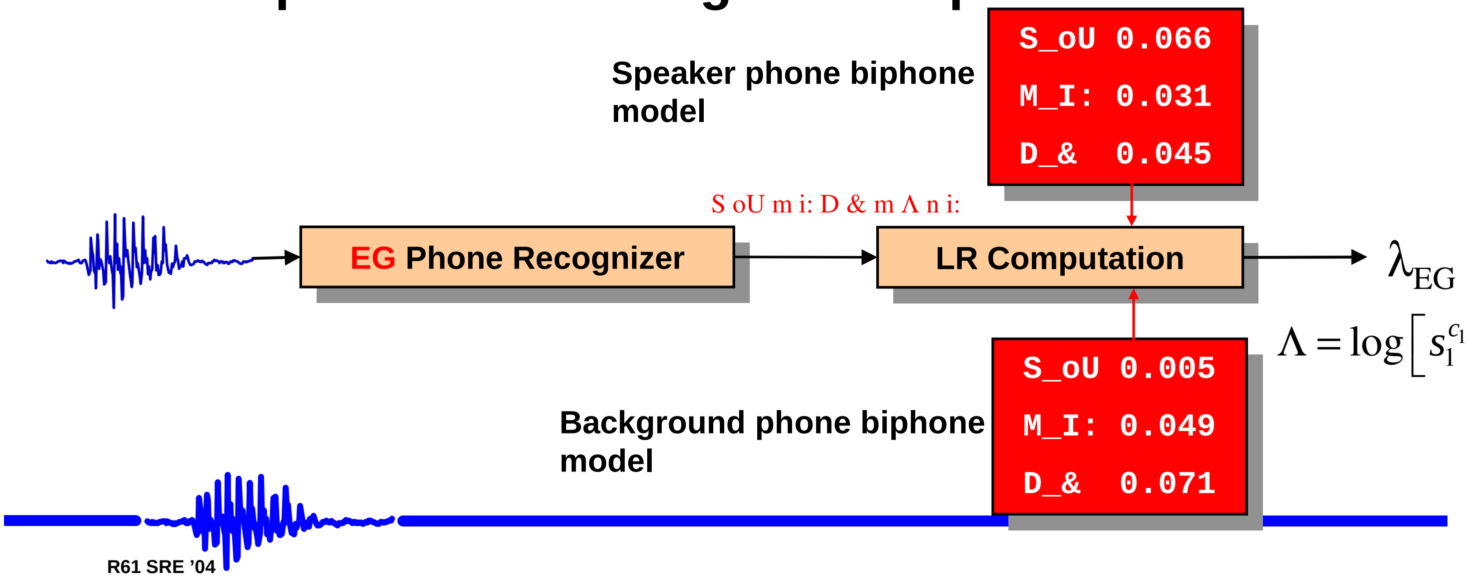
- System submitted same as SRE03
  - Using background data from SWB I, SWB II, SWB Cell and some SRE04 dev data
  - Modified for a LINUX Cluster
    - SUN GRID with ~ 290 CPUs
    - Tokenization SLOW, not using the GRID!
    - Model training and testing VERY FAST! (hrs to min)
- Tokenizer-level linear fusion
  - Available gender-dependent phone recognizers EG, GE, JA, MA, SP
  - Equal weighting 5-way fusion

System Description
Eval Submission

- **R64-1:** Triphones
  - SAD gated phones (xtalk)
  - Cmin applied to the Background during testing
    - Empirically determined by minimizing the EER on the the development data, SRE04
  - **Detection Threshold**: selected the threshold that minimized the DCF on SRE04 dev data

Phonetic Speaker Recognition

- Use cumulative pronunciation details to recognize speakers
  - Individual realization of phonetic information
  - No ASR needed
- PPRLM Tokenizer (HTK 1.4)
  - No language model
- Training: Create models from n-phone counts of speaker and background speech
- Recognition: Compute likelihood score between Test, Speaker and Background n-phone models

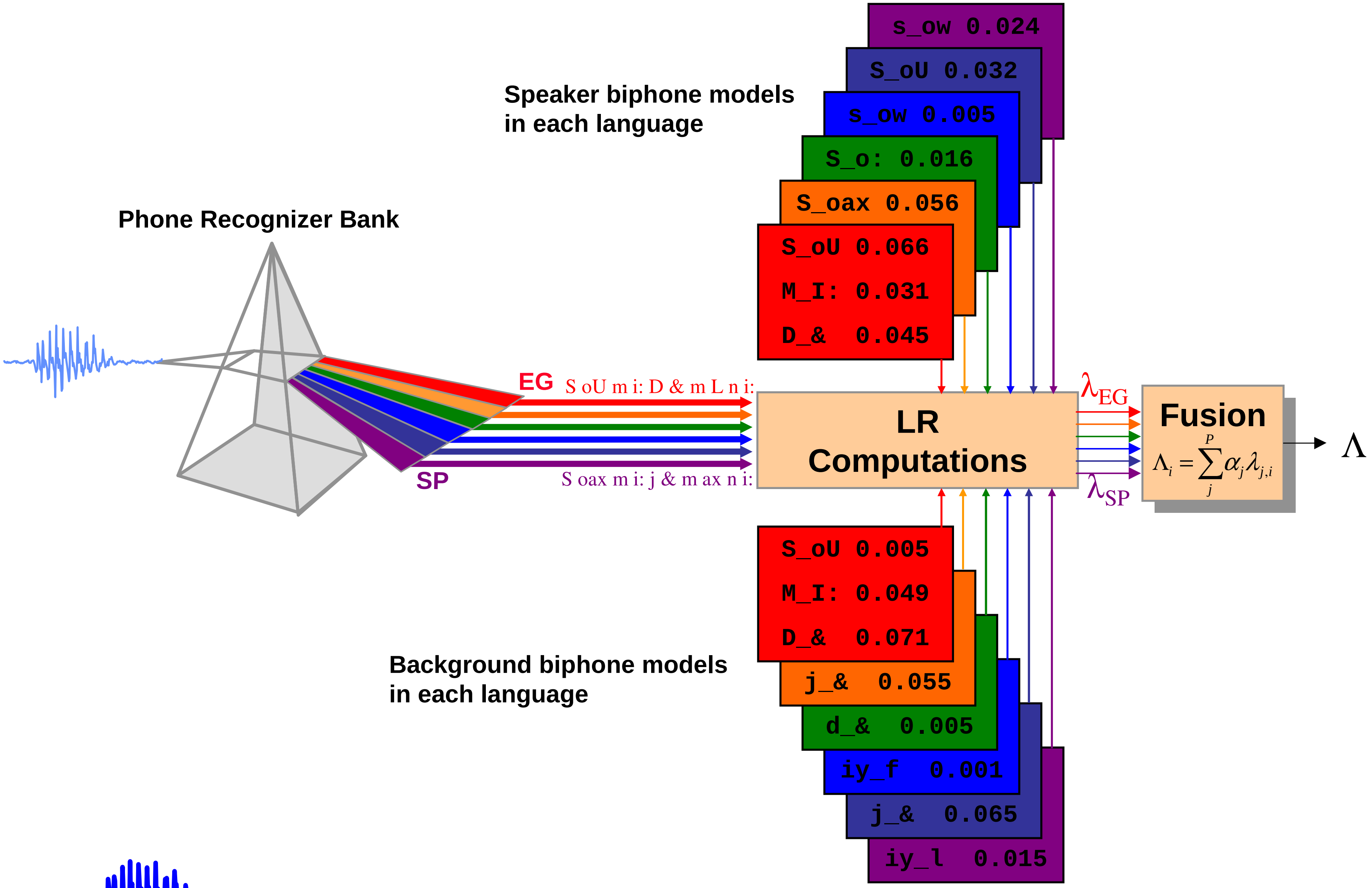


```
S_oU 0.066
M_I: 0.031
D_&  0.045
```

```
S_oU 0.005
M_I: 0.049
D_&  0.071
```

$$\Lambda = \log\left[s_1^{c_1} s_2^{c_2} \ldots s_i^{c_i}\right] - \log\left[b_1^{c_1} b_2^{c_2} \ldots b_i^{c_i}\right]$$

Parallel Phonetic Speaker Recognition



Speaker biphone models in each language

s_ow 0.024

S_oU 0.032

s_ow 0.005

S_o: 0.016

S_oax 0.056

S_oU 0.066

M_I: 0.031

D_& 0.045

Phone Recognizer Bank

EG S oU m i: D & m L n i:

SP S oax m i: j & m ax n i:

LR Computations

λ_{EG}

λ_{SP}

Fusion

$$\Lambda_i = \sum_{j}^{P} \alpha_j \lambda_{j,i}$$

Λ

Background biphone models in each language

S_oU 0.005

M_I: 0.049

D_& 0.071

j_& 0.055

d_& 0.005

iy_f 0.001

j_& 0.065

iy_l 0.015

Conclusions and Future Work

- Performance was not as expected
- Combining several data source for background training produced no improvements.
- Upgrade Tokenizers to latest HTK package
  - Take full advantage of the SUN GRID LINUX Cluster
- Include the Universal Phonetic Recognizer, UPR, to the tokenizers.
  - Fusion
- Fuse with other speaker recognition systems to check for orthogonality

Phone Recognition

- Phone recognizer from PPRLM
  - Phone recognition w/o language model constraints
  - Consistency, not accuracy, is desired
- Features: 12 cepstral, 13 delta-cepstral coeffs
- Frames: 20 ms length, 10 ms update
- HMM (HTK 1.4)
  - Trained on OGI multi-language corpus
    - English, German, Hindi, Japanese, Mandarin, Spanish
  - Fully connected, 3-state, null grammar, Viterbi decoder
- Output: phone symbol, start time, stop time and log-likelihood
- Gender dependent and independent models
  - No gender dependent Hindi models

Scoring

$$\Lambda = \log\left[s_1^{c_1} s_2^{c_2} \ldots s_i^{c_i} \right] - \log\left[b_1^{c_1} b_2^{c_2} \ldots b_i^{c_i} \right]$$

- s is the number of times symbol i occurred in the speaker model divided by the total number of all symbols in the speaker model
- b is the number of times that symbol i occurred in the background model divided by the total number of all symbols in the background model
- c is the number of time that symbol i occurred in the test segment