IBM T. J. Watson Research Center

The IBM RT07 Speaker Diarization Evaluation Systems

Jing Huang, Etienne Marcheret,

Karthik Visweswariah,

Gerasimos Potamianos

Outline sad spkr.

- Data for training (SAD) and development
- Speech activity detection systems
  - **Changes from RT06**
- Speaker diarization systems:
  - **IBM1: Used for speaker diarization task.**
  - **IBM2: Used for SASTT task.**
  - **Speaker model refinements.**
  - **SAD Impact on Speaker error rate.**
  - **Our brittle cluster merge threshold.**
- Results on the development data (lect.)
- Results on the eval07 data (lect.)
- Post game analysis.
- Conclusions

Data for Training/Development

- **Training data:**
  - Relevant to SAD step only.
  - ICSI meeting (70 hours)
  - NIST meeting pilot (15 hours)
  - RT04 dev/eval (2.5 hours)
  - RT05 dev (6 hours)
  - AMI seminars (16 hours)
  - CHIL03/04 data (4 hours)
  - CHIL06/07 dev (6 hours)
  - 11 five-minute segments from CHIL RT06s eval

- **Development data:**
  - SAD tuning: 17 five-minute segments from CHIL RT06s eval
  - SPKR tuning: 27 five-minute segments from CHIL RT06s eval (-UPC coffee break)

SAD Systems

- **RT06:**
  - Based on low latency telephony system:
    - Objective:
      - Minimize FA (non-stationary noise, leakage from echo cancel.)
      - Little impact on WER in clean.
    - System
      - 3 class
        - SIL, AMN, BRN, VN (silence,background noise, breathe, vocalized noise)
        - K, S, SH, TS (unvoiced fricatives, plosives) (sp/sil = f(adjacent class))
        - AA, AE, AH,……. (voiced)
      - Model likelihoods fused with energy contour.
      - 60 msec block average for frame level score.
      - Heuristic smoothing to deal with eating into words (FA/FR for RT tasks).
- **RT07:**
  - Objective:
    - Min FA+FR wrt reference alignment.
    - 2 class models: (SIL, AMN, BRN, VN), (all speech phones)
    - Sp/Sil: 5 state HMM, bottom up clustering of SI AM, MAP adapt to CHIL data.

SAD performance

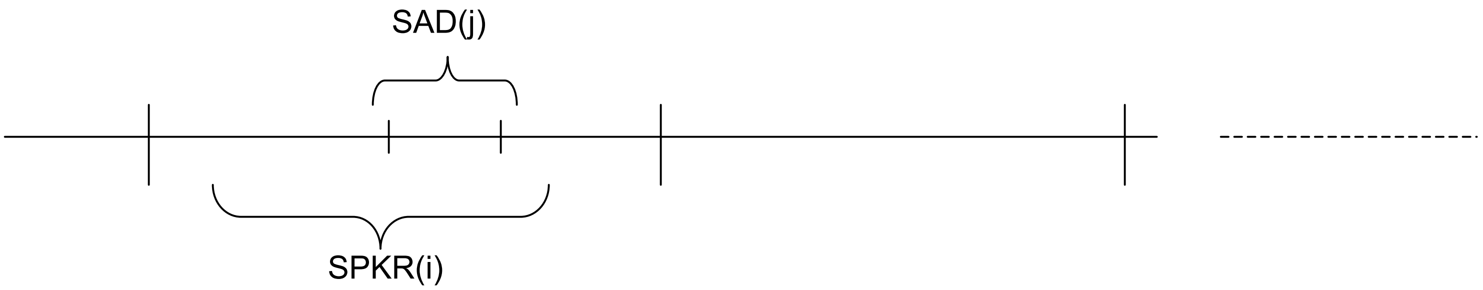
- **Tune FA/FR = f(num speech gauss, num sil gauss)**
  - sad.<num sp gauss>.<num sil gauss>
    - 27 segment eval06:

#Sp G	#Sil G	FR	FA	DER
100	48	1.6	2.2	3.8
100	**32**	**1.0**	**3.3**	**4.3**
100	16	0.4	6.2	6.6
256	32	0.3	7.3	7.6
256	16	0.2	12.2	12.4

- 17 segment eval06
  - RT07 sys: sad.100.32 SDM = 3.0% DER
  - RT06 sys: SDM = 7.5% DER, MDM (Rover) = 5.2% DER

Speaker Diarization System (1)

- **Procedure (IBM1, Diarization task) (19 dim MFCC, no c0)**
  - (1) Initial uniform blocking based on minimum number of frames/spk (4k).
  - (2) Diagonal covariance single Gaussian model for each "speaker" block, "SAD" block.
  - (3) Iterative refinement of speaker models:
    - Mahalanobis measure: assign SAD block, re-estimate speaker model.
  - (4) Cluster Merging: Mahalanobis measure
    - Merge stopped on development test set tuned threshold.
  - (5) To the remaining speaker clusters make final SAD block assignment (Mah.).

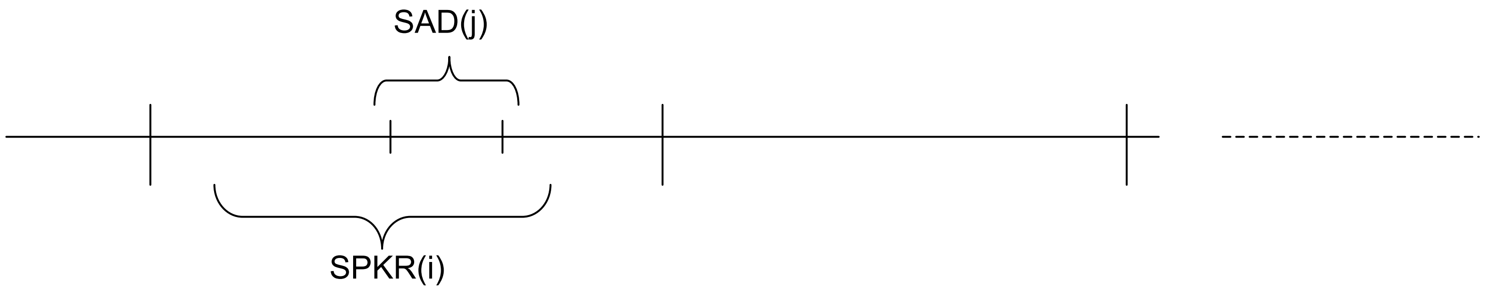


$$dist(i,j) = \sum_{d=1}^{D} \frac{\left(\mu_i(d) - \mu_j(d)\right)^2}{\left(\sigma_i(d) + \sigma_j(d)\right)}$$

← Assignment and cluster merge

Speaker Diarization System (2)

- **Procedure (IBM2, SASTT) (19 dim MFCC, no c0)**
  - (1) Initial uniform blocking based on minimum number of frames/spk (4k).
  - (2) FC single Gaussian model for each “speaker” block, “SAD” block.
  - (3) Iterative refinement: Maximum Log Likelihood measure.
  - (4) Cluster Merging: Likelihood loss.
    - Merge stopped on development test set tuned threshold.
  - (5) To the remaining speaker clusters make final SAD block assignment (Likelihood).



$$LL1(i,j) = \sum_j + (\mu_i - \mu_j)(\mu_i - \mu_j)^T$$

$$LL(i,j) = -0.5\,\mathrm{tr}\left(\Sigma_i^{-1} * LL1(i,j)\right) - 0.5\log\left(\left|\Sigma_i\right|\right)$$

← SAD(j) Assignment to speaker(i)

$$dist(i,j) = N\left(\log\left(\left|\Sigma\right|\right) - \frac{n_i}{N}\log\left(\left|\Sigma_i\right|\right) - \frac{n_j}{N}\log\left(\left|\Sigma_j\right|\right)\right)$$

← Merge on loss of likelihood.

Model Refinement Steps

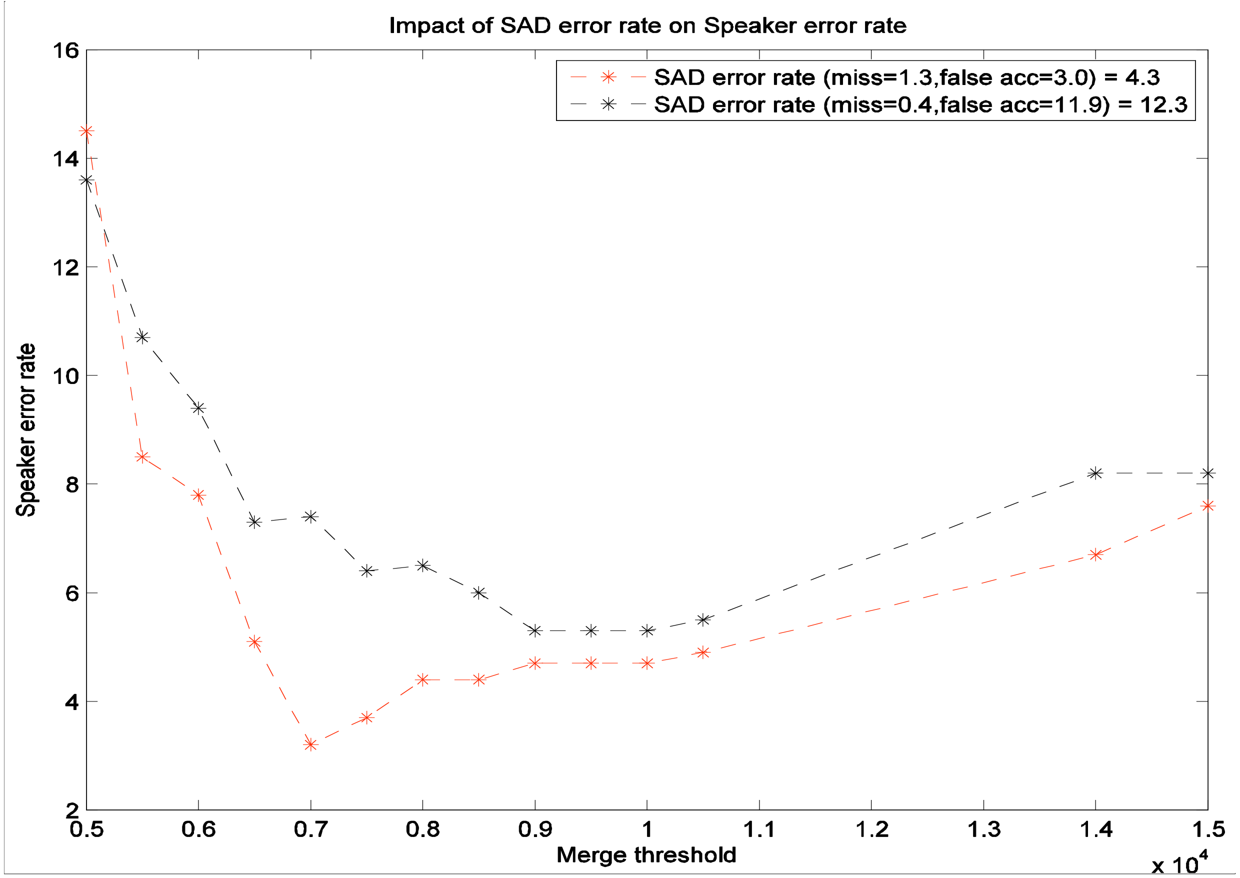
- **Word level alignment (on the IBM 1 system)**
  - feedback to trim SAD segments (squeeze down on SAD FA).
    - (IBM 1 + align)

- **GMM Speaker Models (on the IBM 2 system)**
  - Iterative refinement from output of cluster merge step.
    - EM to build GMM (10 mix,. diag, cov.).
    - Frame level re-labeling
      - Score for each speaker model = 150 msec smoothing window (+-75msec).
    - (IBM 2 + refine)

  - Allows us to generate speaker boundaries within the SAD segments.
  - Replace with SIV system?

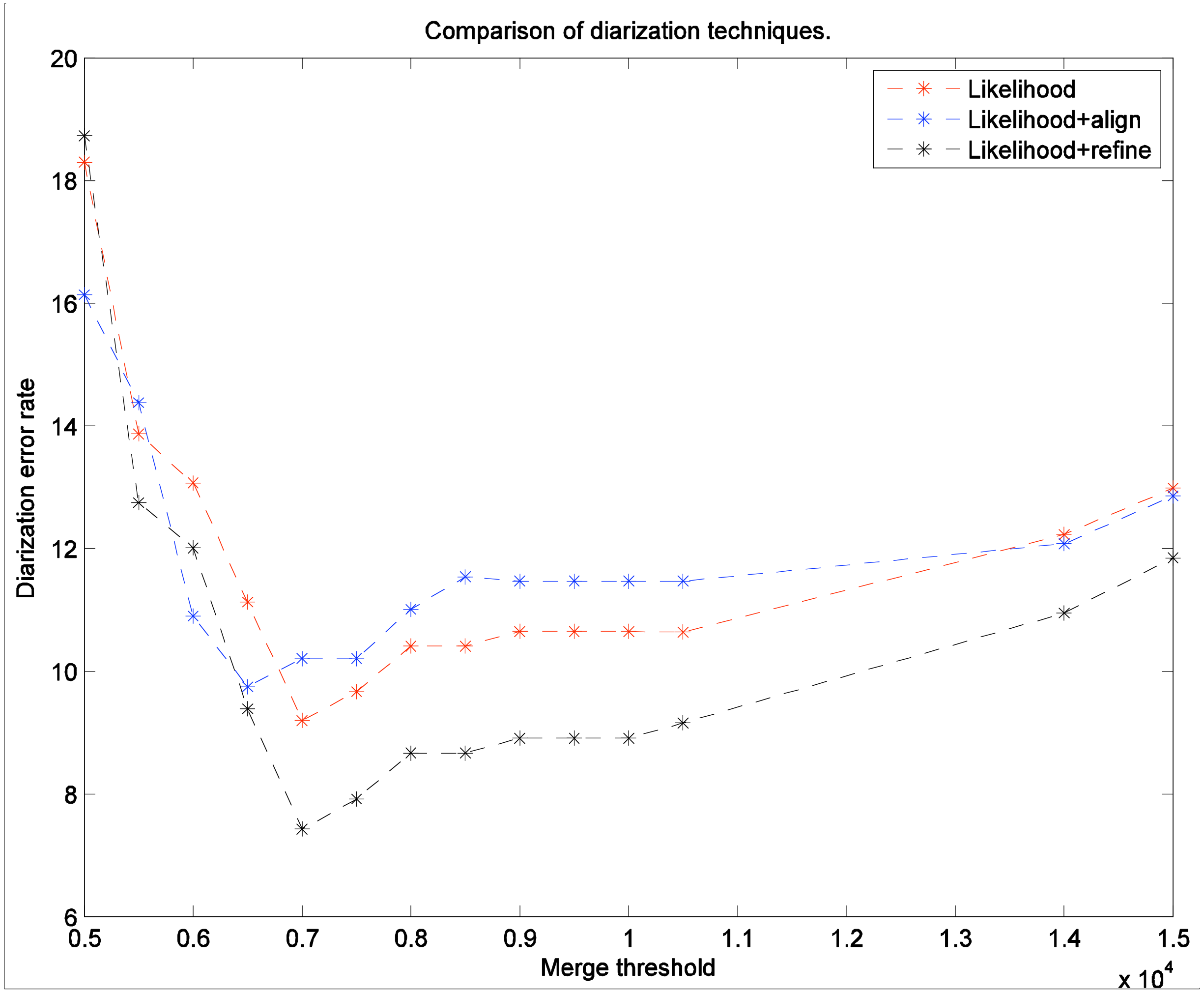
SAD Impact on Speaker Error Rate (IBM2 system)

- **Reduce SAD error:**
  - 12.3% (0.5 Miss, 11.8 FA) → 4.3% (1.3 Miss, 3.0 FA)
  - Average reduction in speaker error rate of 20.3% (56.7% at opt. merge thresh.)
  - Expected? Refinement is not robust enough? Threshold is moving!



Development test set

- **Merge threshold**
  - Impact of word level alignment and GMM speaker level refinement (on IBM 2).



Development test set (27 seg.)

System	opt. thresh.	SAD (%) (1.3FR,3.0FA)	Speaker error(%)	DER(%)
IBM 1	*0.6*	*4.3*	*6.6*	*10.9*
IBM 2	*7000*	*4.3*	*5.0*	*9.3*
IBM 1+align	*0.6*	*4.3*	*5.6*	*9.9*
IBM 1+align **(Sub. Diar)**	*Site Specific*	*4.3*	*3.9*	*8.2*
IBM 2 + refine **(Sub. SASTT)**	*7000*	*4.3*	*3.2*	*7.5*

Evaluation test set

System	opt. thresh. (devset)	SAD(%) (2.4FR, 3.9FA)	Speaker error(%)	DER(%)
IBM 1	*0.6*	*6.3*	*21.9*	*28.2*
IBM 2	*7000*	*6.3*	*24.8*	*31.1*
IBM 1+align	*0.6*	*6.3*	*21.0*	*27.3*
IBM 1+align (Sub. Diar)	*Site Specific*	*6.3*	*23.7*	*30.0*
IBM 2 + refine (Sub. SASTT)	*7000*	*6.3*	*21.4*	*27.7*

Align step gives 4% relative.
Refine step gives 10.9% relative

Our threshold is not stable.

Evaluation set, post game analysis

System	thresh.	SAD(%) (2.4FR, 3.9FA)	Speaker error(%)	DER(%)
IBM 1	*0.6*	*6.3*	*21.9*	*28.2*
IBM 2	*7000*	*6.3*	*23.7*	*30.0*
IBM 1	*0.9 (opt)*	*6.3*	*18.5*	*24.8*
IBM 2	*15000(opt)*	*6.3*	*17.6*	*23.9*
IBM 1+align	*0.9(opt)*	*6.3*	*18.7*	*25.0*
IBM 2 + refine	*15000(opt)*	*6.3*	*16.5*	*22.8*

26% - 1 spkr.

Conclusions

- **Likelihood based FC model more sensitive to tuned threshold**
  - Approx. 1% absolute reduction in speaker error rate.
- **Iterative GMM refinement**
  - Approx. 1% absolute reduction in speaker error rate.

- **Cluster merge thresholds not generalizing:**
  - Modified BIC?

- **Replace Iterative GMM refinement step with SIV system.**

- **Multiple channels (Beamforming).**
- **Can’t we use camera info?**