MULTI-STAGE SPEAKER DIARIZATION FOR CONFERENCE AND LECTURE MEETINGS

presented by Xuan Zhu

RT-07S workshop
Baltimore, Maryland
May 11, 2007

INTRODUCTION

Task

- speaker diarization (SPKR): who spoke when

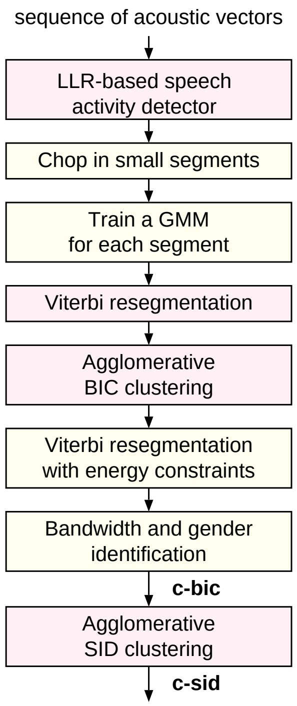
Sub-types of meeting data

- conference room meetings
- lecture room meetings
- coffee break (no LIMSI participation this year)

Challenges of meeting data

- spontaneous speech with overlaps
- variability in audio SNR configurations derived from the use of different types of microphones in recording room
- different styles of participant interaction across sub-domains

DIARIZATION SYSTEM (1)


sequence of acoustic vectors

LLR-based speech activity detector

Chop in small segments

Train a GMM for each segment

Viterbi resegmentation

Agglomerative BIC clustering

Viterbi resegmentation with energy constraints

Bandwidth and gender identification

c-bic

Agglomerative SID clustering

c-sid

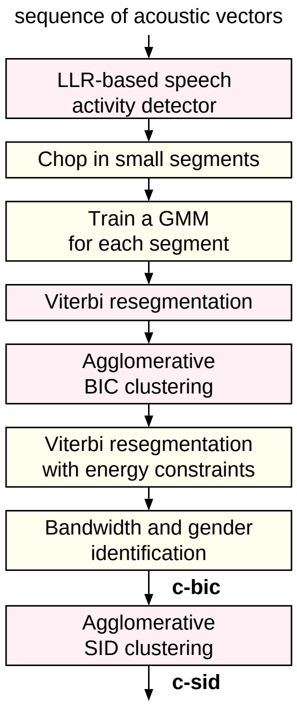
Front-end

- 38 features: 12 MFCC + 12 Δ + 12 $\Delta\Delta$ + Δ logE + $\Delta\Delta$ logE

LLR-based speech activity detector (SAD)

- GMMs for speech and non-speech models
- log-likelihood (LLR) ratio between 2 models computed for each frame
- different prior probabilities for each SAD acoustic model
- transition points detected at the maxima of the mean of LLR over smoothing window

DIARIZATION SYSTEM (2)


sequence of acoustic vectors

LLR-based speech activity detector

Chop in small segments

Train a GMM for each segment

Viterbi resegmentation

Agglomerative BIC clustering

Viterbi resegmentation with energy constraints

Bandwidth and gender identification

c-bic

Agglomerative SID clustering

c-sid

Chop into small segments

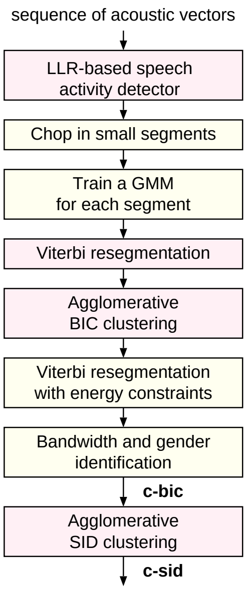
- 2 sliding windows of 5 sec, local divergence measure

$$G(w_1, w_2) = (\mu_2 - \mu_1)^T \Sigma_1^{-1/2} \Sigma_2^{-1/2} (\mu_2 - \mu_1)$$

GMM estimation for each segment

- 8-component GMM with diagonal covariance matrix per segment

DIARIZATION SYSTEM (3)



sequence of acoustic vectors

LLR-based speech activity detector

Chop in small segments

Train a GMM for each segment

Viterbi resegmentation

Agglomerative BIC clustering

Viterbi resegmentation with energy constraints

Bandwidth and gender identification

c-bic

Agglomerative SID clustering

c-sid

BIC Agglomerative clustering

- Gaussian with full covariance matrix
- merge criterion

  $$\Delta BIC = (n_i+n_j)log|\Sigma| - n_i log|\Sigma_i| - n_j log|\Sigma_j| - \lambda P$$

  with penalty

  $$P = \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log N$$

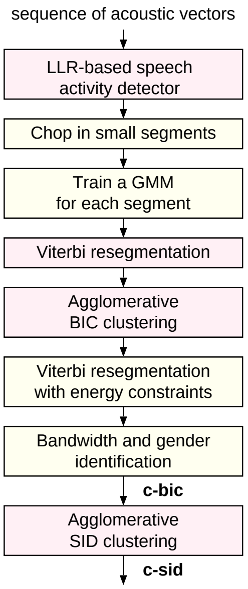
- stop criterion

  $$\Delta BIC >= 0$$

BIC penalty

- local: $N = n_i + n_j$
- global: $N = \Sigma_k n_k$

DIARIZATION SYSTEM (4)



sequence of acoustic vectors

LLR-based speech activity detector

Chop in small segments

Train a GMM for each segment

Viterbi resegmentation

Agglomerative BIC clustering

Viterbi resegmentation with energy constraints

Bandwidth and gender identification

c-bic

Agglomerative SID clustering

c-sid

SID clustering

- 15 MFCC + Δ + Δ logE, feature warping (Gaussian normalization)
- Universal Background Models (UBM) with 128 Gaussians
- MAP adaptation of matching UBM
- cross log-likelihood ratio between clusters c_i and c_j

  $$clr(c_i, c_j) = \frac{1}{n_i} log \frac{f(x_i|M_j)}{f(x_i|UBM)} + \frac{1}{n_j} log \frac{f(x_j|M_i)}{f(x_j|UBM)}$$

  with x_i the data from cluster c_i, M_i the model for cluster c_i, n_i the size of segment x_i

- threshold δ

ADAPTATION TO MEETINGS

System structure

- removing bandwidth detection module from the RT06 system (assumption of no telephone speech in meetings)

Audio input condition

- using beamformed signals generated from ICSI delay&sum signal enhancement system for the Multiple Distant Microphone (MDM) condition

ACOUSTIC MODEL TRAINING

RT06 SAD models and UBMs

- speech and non-speech models trained on far-field data:
  7 ISL lectures recorded in 2003
- 4 UBMs (male/female, studio/telephone) trained on a subset of 1996/1997 English Broadcast News data (same as BN system)

New SAD models and UBM

- using forced alignment segmentations to train speech and non-speech models and UBM independent of the gender and bandwidth
- new training data used to estimate SAD models and UBM:
  8 RT-04S development conferences + 8 RT-04S evaluation conferences + 10 RT-05S evaluation conferences
- different types of acoustic features along with various feature normalization techniques investigated for model training
- same SAD models and UBM for conference and lecture test data

DEVELOPMENT CORPUS DESCRIPTION

Conference development dataset (conf dev07s)

- 9 conference meetings from RT-06S evaluation data
- collected by 5 laboratories: CMU, EDI, NIST, TNO and VT
- a duration of about 15 minutes per excerpt
- forced alignment references available for scoring

Lecture development dataset (lect dev07s)

- 28 lecture meetings from RT-06S evaluation dataset
- recoded by 5 CHIL partner sites: AIT, IBM, ITC, UKA and UPC
- audio lengths ranging from 23 to 44 minutes
- forced alignment references available for scoring

LLR-BASED SAD USING VARIOUS TYPES OF FEATURES

Configuration for LLR-based SAD

- 256 Gaussians in each SAD acoustic model
- prior probability for speech and non-speech models being 0.8:0.2
- smoothing window with a duration of 50 frames

Proposed energy normalization based on voicing factor

- voicing factor $\boldsymbol{v}$ computed as maximum peak of the autocorrelation function (excluding lag zero)
- harmonic energy defined as $\boldsymbol{E_h = v.E_0}$
- energy normalized relative to 10% highest harmonic energy

SAD RESULTS ON CONFERENCE MDM DEV DATA

SAD acoustic features	*missed speech error (%)*	*false alarm speech error (%)*	*overlap SAD error (%)*
baseline	1.3	4.3	5.6
baseline+e	1.1	4.0	5.1
baseline+env	1.1	3.3	4.3
baseline+e+mvn	0.8	3.0	3.9

Different kinds of acoustic features used in LLR-based SAD

- baseline: 12 MFCC + 12 Δ + 12 $\Delta\Delta$ + Δ logE + $\Delta\Delta$ logE
- baseline+e: adding raw energy to baseline features
- baseline+env: baseline features plus normalized energy relying on voicing factor
- baseline+e+mvn: performing standard variance normalization on both the baseline features and raw energy

SAD RESULTS ON LECTURE MDM DEV DATA

SAD acoustic features	*missed speech error (%)*	*false alarm speech error (%)*	*overlap SAD error (%)*
baseline	2.4	5.3	7.8
baseline+e	0.5	11.2	11.8
baseline+env	0.9	4.7	5.7
baseline+e+mvn	1.0	5.6	6.6

- use of raw energy degrades largely SAD performance on lectures
- mismatch between conference training and lecture test leads to a higher SAD error

SPKR RESULTS ON CONFERENCE MDM DEV DATA

UBM acoustic features	*speaker match error (%)*	*overlap DER (%)*
15plp+Δ+ΔlogE+w	28.4	36.2
15plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+w	23.3	31.1
12plp+Δ+ΔlogE+w	22.9	30.6
12plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+w	27.9	35.7
12plp+Δ+ΔlogE+mvn	33.8	41.6
12plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+mvn	32.0	39.8

SID clustering with UBMs trained on different types of features

- “w” being feature warping, “mvn” being variance normalization
- each UBM with 128 Gaussian component
- with SAD acoustic models trained on “baseline+e+mvn”
- BIC penalty weight $\lambda = 3.5$ and SID threshold $\delta = 0.5$

SPKR RESULTS ON LECTURE MDM DEV DATA

UBM acoustic features	*speaker match error (%)*	*overlap DER (%)*
15plp+Δ+ΔlogE+w	10.0	17.5
15plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+w	10.2	17.7
12plp+Δ+ΔlogE+w	10.3	17.8
12plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+w	10.2	17.7
12plp+Δ+ΔlogE+mvn	10.5	18.0
12plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+mvn	10.2	17.7

SID clustering with UBMs trained on different kinds of features

- no significant changes in diarization performances for lectures
- 128 Gaussian per UBM
- with SAD acoustic models trained on “baseline+e+mvn”
- BIC penalty weight $\lambda = 3.5$ and SID threshold $\delta = 0.5$

EVALUATION RESULTS

data type *& condition*	*SPKR as SAD* *error (%)*	*non-overlap* *DER (%)*	*overlap* *DER (%)*
conference MDM	3.2	23.0	26.1
conference SDM	3.5	26.6	29.5
lecture MDM	10.1	24.5	25.8
lecture SDM	10.0	24.3	25.6

Same SAD models and UBM for conference and lecture data

- SAD acoustic models trained on “baseline+e+mvn” feature set
- UBM trained on “12plp+Δ+ΔlogE+w” feature set

Configurations of diarization

- BIC penalty weight $\lambda = 3.5$ for both conference and lecture
- SID threshold δ set to 0.6 for conference and 0.5 for lecture

CONCLUSIONS

Speaker diarization system for meeting data

- diarization results obtained on the conference evaluation data similar to ones on the development data
- higher DER rate on the lecture evaluation data than the development data can be attributed to the higher participant interaction in this year’s lecture data
- beamformed MDM signals effective for conference but not for lecture