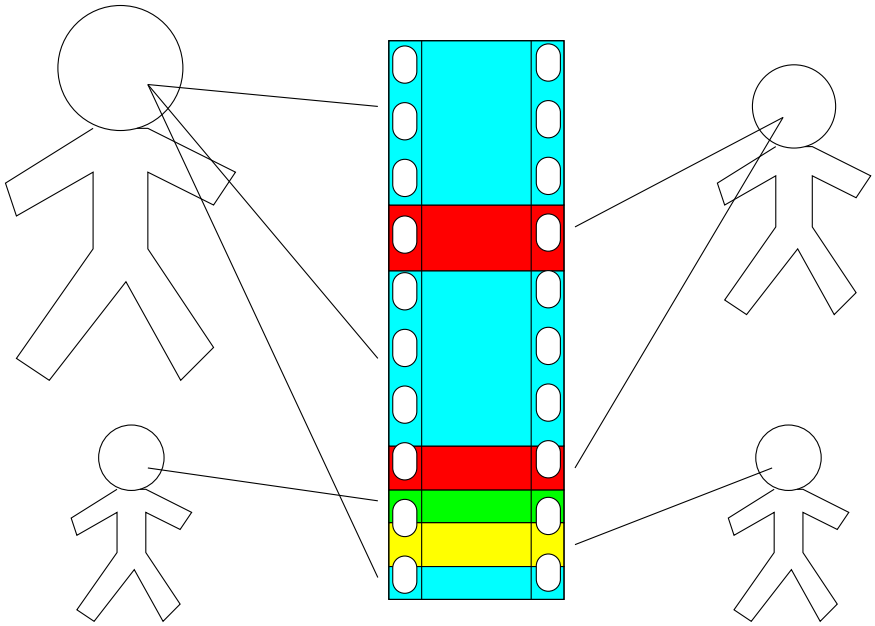
Eurospeech '99

Who Spoke When ?

Automatic Segmentation and Clustering for Determining Speaker Turns



Sue Johnson

Email : sej28@eng.cam.ac.uk

http://svr-www.eng.cam.ac.uk/Research/Projects/mdr/

Cambridge University Engineering Department

Aim

- Identify who spoke when

Applications

- Speaker Identification (forensic, security)
- Speaker Adaptation (speech recognition)
- Speaker Retrieval (searching databases)
- Dialogue Display (understanding transcriptions)
- Speaker Tracking (targeting speakers)
- Text Summarisation (track news announcers)

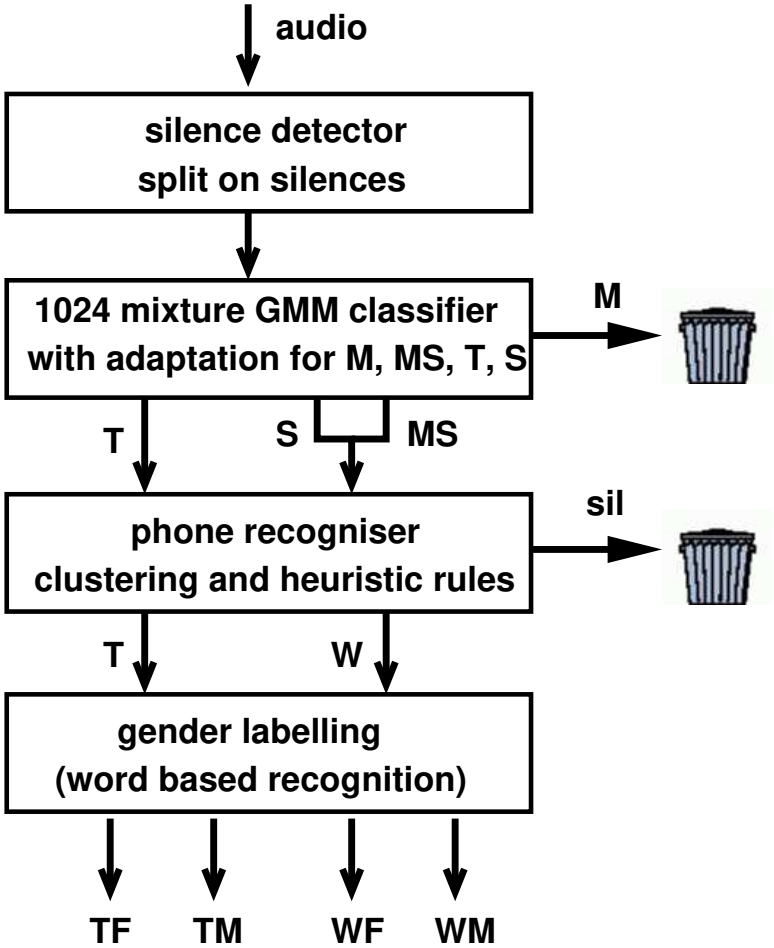
Problems

- All speakers are unknown
- Number of speakers is unknown

Automatic Segmentation

Aim

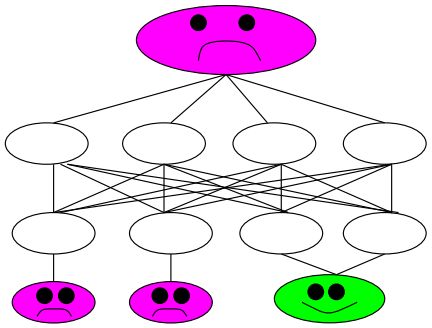
- Produce acoustically homogeneous segments
- Ideally 1 speaker, 1 noise/channel condition
- Segment length between 1 and 30 seconds



Automatic Clustering

Aim : Group similar segments of audio together

Algorithm



Top-down approach

Correlation matrix with AHS

Local recombination allowed

Speaker-Adaptation Scheme

- Designed for subsequent MLLR
- Minimum occupancy requirement (30 s)
- Speaker split not really important

Speaker-Identification Scheme

- Needs 1 cluster per speaker
- No occupancy constraints
- Extra recombination required
  (e.g. gain from split, allowable overlap)
- Special case for singletons

Evaluating Cluster Performance (1)

Aim:

- Reward large pure (single-speaker) clusters
- Penalise multi-speaker and incomplete clusters

Definitions:

N_c = number of clusters

n_{ij} = # utterances in cluster i from speaker j

$n_j = \sum_{\mathrm{i}} n_{ij}$ = # utterances said by speaker j

$n_i = \sum_{\mathrm{j}} n_{ij}$ = # utterances in cluster i

$p_i = \sum_{\mathrm{j}} \frac{n_{ij}^2}{n_i^2}$ = purity of cluster i

Rand Index (smaller number better)

$$I_{RAND} = \frac{1}{2}\left(\sum_{\mathrm{i}} n_i^2 + \sum_{\mathrm{j}} n_j^2\right) - \sum_{\mathrm{i}}\sum_{\mathrm{j}} n_{ij}^2$$

= same spkr/diff cluster + diff spkr/same cluster

BBN-metric (larger number better)

$$I_{BBN}(C) = \sum_{\mathrm{i}} n_i p_i - QN_c$$

Evaluating Cluster Performance (2)

Cluster Efficiency

$$\eta(C) = \frac{I_{BBN}(C) - I_{BBN}(S)}{I_{BBN}(P) - I_{BBN}(S)}$$

I(S) = singleton case = Nu (1 - Q) $\eta(S) = 0$

I(P) = perfect case = Nu - Q Ns $\eta(P) = 1$

I(1) = one-cluster case $\eta(1)$ **often negative**

Choosing Q

Make I(1) = I(S) (agglomerative/divisive even)

$$I(1) = \frac{1}{N_u}\left(\sum_j n_j^2\right) - Q$$

Hence $$Q_{crit} = \frac{N_u^2 - \sum_j n_j^2}{N_u(N_u - 1)}$$

Note, $0 \leq Q_{crit} \leq 1$ since:

$$N_u = \sum_j n_j \leq \sum_j n_j^2 \leq \left(\sum_j n_j\right)^2 = N_u^2$$

Experimental Procedure

Data Used

- 1996 Hub-4 development data
- ~ 100 minutes of U.S. broadcast news
- 488 manually labelled speaker turns
- 553 automatically generated segments

Complete System

- Automatic segmentation run on audio
- Gender and bandwidth dependent clustering
- Cluster lists concatenated to give final output

Cases Considered

- One-cluster (no top-down clustering)
- Singletons (no bottom-up clustering)
- Adaptation-based (occupancy > 30s)
- Speaker-based (2 parameter sets)
- Perfect speaker clusters (manually labelled)

Results (1) : Utterance based

Perfect Segmentation (1 speaker/segment)

Condition	Nc	I_{RAND}	$\eta_{BBN}^{Q=0.5}$	$\eta_{BBN}^{Q=0.95}$
one-cluster	1	112807	-1.065	0.000
singleton	488	6021	0.000	0.000
adaptation	81	5376	0.336	0.646
speaker (1)	92	4286	0.476	0.707
speaker (2)	165	4937	0.464	0.616
perfect	77	0	1.000	1.000

- Adaptation scheme splits speakers well (65%)

Automatic Segmentation (dominant speaker)

Condition	Nc	I_{RAND}	$\eta_{BBN}^{Q=0.5}$	$\eta_{BBN}^{Q=0.96}$
one-cluster	1	145850	-1.037	0.000
singleton	553	6778	0.000	0.000
adaptation	106	6309	0.380	0.638
speaker (1)	119	4999	0.506	0.691
speaker (2)	151	5144	0.485	0.649
perfect	68	0	1.000	1.000

- Speaker split (70%) indepedent of segmentation

Frame-Based Scoring

Better Scoring

- Compare automatic and manual systems directly
- Segmentation can also be evaluated
- Effects of segmentation & clustering quantified

More Realistic Evaluation

- mistakes with longer utterances more critical
- More accurately reflects the task of tracking a (potentially unknown) anchor news-reader

N_f = total number of FRAMES

n_{ij} = # FRAMES in cluster i from speaker j

Perfect case : $I^F(P) = N_f - QN_s$

Singleton case: $I^F(S) = N_f(1-Q)$

$$Q^F_{crit} = \frac{N_f^2 - \sum_j n_j^2}{N_f(N_f-1)}$$

Results (2) : Frame Based

Perfect Segmentation (1 speaker/segment)

Condition	Nc	I^*_{RAND}	$\eta^{Q=0.5}_{BBN}$	$\eta^{Q=0.95}_{BBN}$
one-cluster	1	166400	-0.902	0.000
singleton	591639	8611	0.000	0.000
adaptation	81	6382	0.585	0.782
speaker (1)	92	4355	0.673	0.828
speaker (2)	165	4648	0.811	0.900
perfect	77	0	1.000	1.000

* x 10^6

Automatic Segmentation (dominant speaker)

Condition	Nc	I^*_{RAND}	$\eta^{Q=0.5}_{BBN}$	$\eta^{Q=0.95}_{BBN}$
one-cluster	1	179500	-0.902	0.000
singleton	614510	9288	0.000	0.000
adaptation	106	7962	0.672	0.827
speaker (1)	119	5785	0.762	0.875
speaker (2)	151	5943	0.789	0.889
perfect	68	0	1.000	1.000

- Less recombination gives better results (90%)
- Similar conclusions to utterance-based case

Results (3) - Automatic vs Manual

Automatically generated segments compared to manual labelling of frames of speech

Condition	Nc	I^{*}_{RAND}	$\eta^{Q=0.5 \text{ or } 0.95}_{BBN}$
one-cluster	1	164100	-0.902
singleton	614510	8551	0.000
adaptation	106	7394	0.620
speaker (1)	119	5525	0.699
speaker (2)	151	5649	0.723
perfect	68	586	0.892

* x 10^{6}

- 11% error introduced by segmentation
- 17% error introduced by clustering

Note, since the segmenter may discard frames as non-speech, or accept frames which have no transcription, there is a slight frame mismatch hence $I_{RAND}(P)$ and $\eta^{Qcrit}_{BBN}(1)$ are non-zero

Results (4) : Summary

η Q=0.5 BBN		Segmentation	
		Manual	Automatic
Clustering	Manual	1.000	0.892
	Automatic	0.811	0.723

Segmentation (39%)

- Speech erroneously discarded as music
- Non-speech erroneously retained
- Classification errors (WB/NB, M/F)
- Boundary errors

Clustering (61%)

- Combining different speaker segments
- Not combining same speaker segments

Drop for clustering indepedent of segmentation

Conclusions

- Automatic segmentation/clustering of audio is possible
- Cluster efficiency provides a measure of speaker split
- Can treat agglomerative and divisive clustering evenly
- Adaptation scheme gives ~65%, speaker scheme ~70%
- Frame efficiency more accurately reflects some tasks
- Adaptation scheme gives ~80%, speaker scheme ~90%
- Relative influence of segmentation and clustering seen
- Segmentation causes 11% drop, clustering 18%
- Relative performance depends on scoring and thus task