Audio Segmentation and its Applications in Speaker Characterization

Paula López Otero

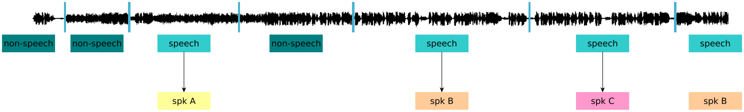
Multimedia Technology Group (GTM)

Multimedia Technology Group (GTM)

- Division of a signal into homogeneous segments
    - Speaker segmentation

- Division of a signal into homogeneous segments
  - Speaker segmentation

- Speaker diarization ("who spoke when")
- Indexing of multimedia information
  - classification of music (song title, genre...)
- Automatic speech recognition (ASR)
  - Removal of non-speech segments
  - When we know "who spoke when" ⇒ Speaker adaptation

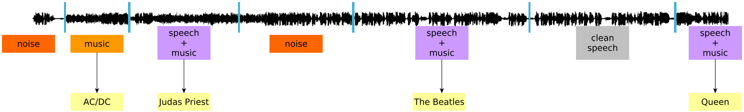


- Speaker diarization ("who spoke when")
- Indexing of multimedia information
  - classification of music (song title, genre...)
- Automatic speech recognition (ASR)
  - Removal of non-speech segments
  - When we know "who spoke when" ⇒ Speaker adaptation

Multimedia Technology Group (GTM)

- Speaker diarization ("who spoke when")
- Indexing of multimedia information
  - classification of music (song title, genre...)
- Automatic speech recognition (ASR)
  - Removal of non-speech segments
  - When we know "who spoke when" ⇒ Speaker adaptation

Multimedia Technology Group (GTM)



- Speaker diarization ("who spoke when")
- Indexing of multimedia information
  - classification of music (song title, genre...)
- Automatic speech recognition (ASR)
  - Removal of non-speech segments
  - When we know "who spoke when" ⇒ Speaker adaptation

Multimedia Technology Group (GTM)

- Speaker diarization (“who spoke when”)
- Indexing of multimedia information
  - classification of music (song title, genre...)
- Automatic speech recognition (ASR)
  - Removal of non-speech segments
  - When we know “who spoke when” ⇒ Speaker adaptation

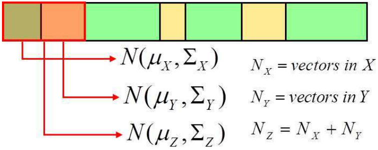
Multimedia Technology Group (GTM)

- Speaker diarization ("who spoke when")
- Indexing of multimedia information
  - classification of music (song title, genre...)
- Automatic speech recognition (ASR)
  - Removal of non-speech segments
  - When we know "who spoke when" ⇒ Speaker adaptation

Multimedia Technology Group (GTM)

Multimedia Technology Group (GTM)

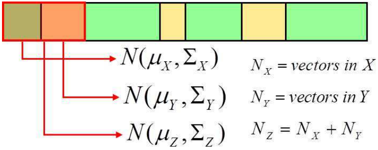
The BIC algorithm



- A window of data (Z) is taken and divided into two sub-windows (X, Y) at frame i
- X, Y and Z are modelled with a multivariate Gaussian
- A hypothesis test is applied for acoustic change detection
  - H_0: No acoustic change in window Z
  - H_1: The window contains an acoustic change at point i

G. Schwarz, "Estimating the dimension of a model"

The BIC algorithm



- BIC: The maximum likelihood ratio between H_0 and H_1

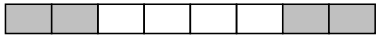
$$R(i) = L_X + L_Y - L_Z = N_Z \log|\Sigma_Z| - N_X \log|\Sigma_X| - N_Y \log|\Sigma_Y|$$

$$\Delta BIC(i) = BIC(H_1) - BIC(H_0) = R(i) - \lambda \frac{1}{2}(d + \frac{1}{2}d(d+1))log(N_Z)$$

- Decision:
  - $\Delta BIC(i) > 0$
  - λ must be tuned for each dataset

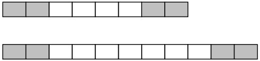
G. Schwarz, "Estimating the dimension of a model"

Segmentation Strategy



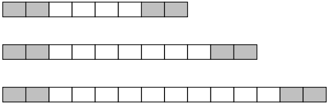
M. Cettolo, M. Vescovi, "Efficient Audio Segmentation Algorithms based on the BIC"

Segmentation Strategy



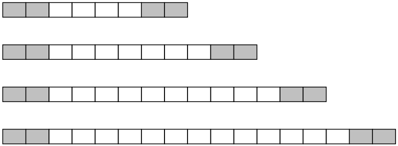
M. Cettolo, M. Vescovi, "Efficient Audio Segmentation Algorithms based on the BIC"

Segmentation Strategy



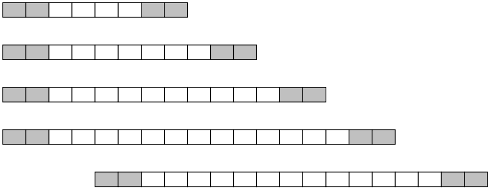
M. Cettolo, M. Vescovi, “Efficient Audio Segmentation Algorithms based on the BIC”

Segmentation Strategy



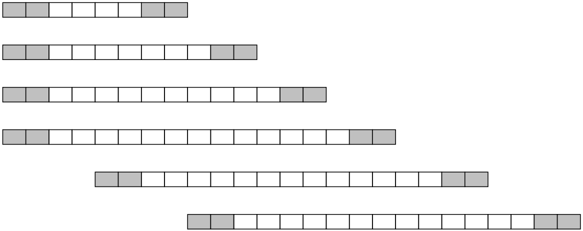
M. Cettolo, M. Vescovi, "Efficient Audio Segmentation Algorithms based on the BIC"

Segmentation Strategy



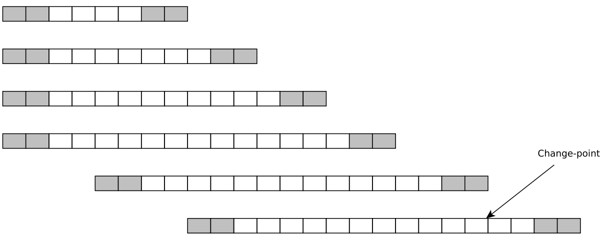
M. Cettolo, M. Vescovi, "Efficient Audio Segmentation Algorithms based on the BIC"

Segmentation Strategy



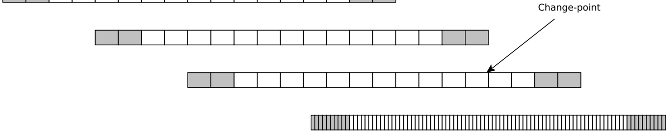
M. Cettolo, M. Vescovi, "Efficient Audio Segmentation Algorithms based on the BIC"

Segmentation Strategy



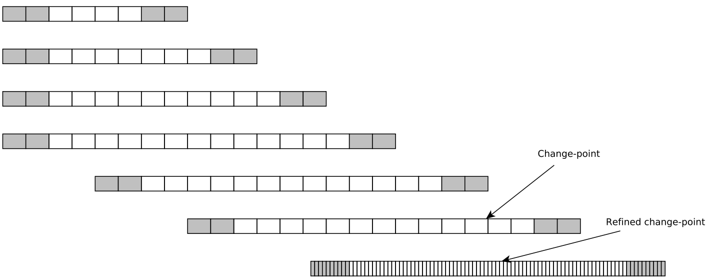
M. Cettolo, M. Vescovi, "Efficient Audio Segmentation Algorithms based on the BIC"

Segmentation Strategy



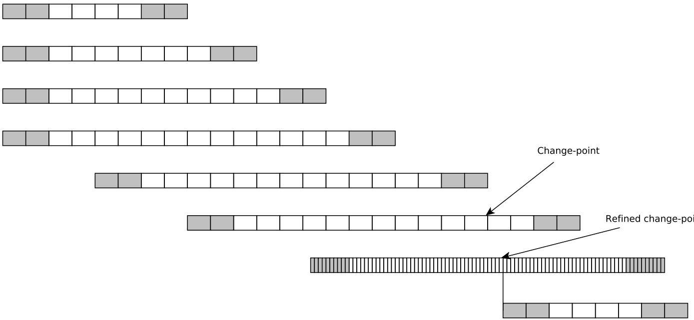
M. Cettolo, M. Vescovi, “Efficient Audio Segmentation Algorithms based on the BIC”

Segmentation Strategy



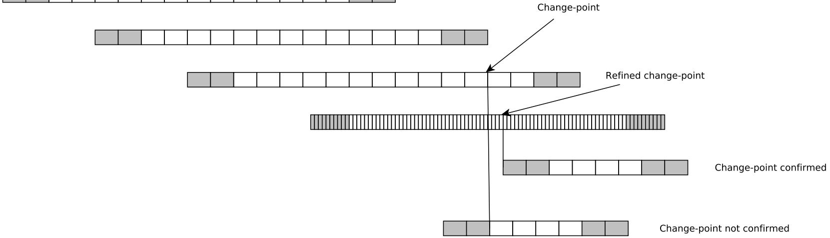
M. Cettolo, M. Vescovi, "Efficient Audio Segmentation Algorithms based on the BIC"

Segmentation Strategy



M. Cettolo, M. Vescovi, "Efficient Audio Segmentation Algorithms based on the BIC"

Segmentation Strategy



M. Cettolo, M. Vescovi, "Efficient Audio Segmentation Algorithms based on the BIC"

Issue: sensitivity

Too sensitive ⇒ false alarms

Too little sensitive ⇒ deletions

Issue: sensitivity

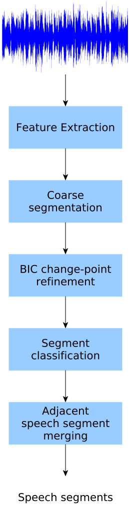
Too sensitive ⇒ false alarms

Too little sensitive ⇒ deletions

Multimedia Technology Group (GTM)

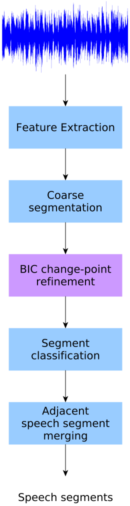
Reducing the false alarm rate

Baseline System

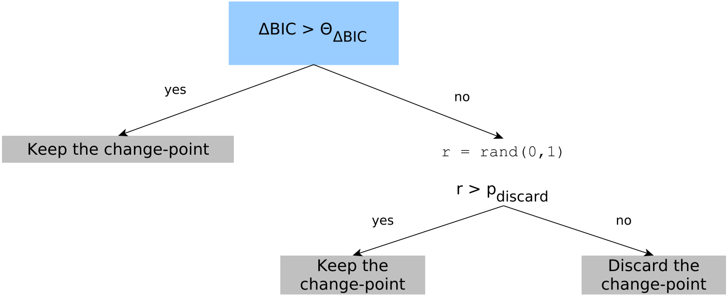


Reducing the false alarm rate

Baseline System



Reducing the false alarm rate



Adaptive Strategy

- $p_{discard}$ and Θ_{BIC} increase when the system is accepting too many false alarms
- $p_{discard}$ and Θ_{BIC} decrease when the system is not accepting false alarms
- Initially: $p_{discard} = 0$, $\Theta_{BIC} = 0$

Poisson-based Strategy

- $p_{discard}$ follows a Poisson cumulative density function
- $p_{discard}$ depends on the expected number of change-points (mean of the distribution)
- Initially: $p_{discard} = 0$

Uniform-based Strategy

- $p_{discard}$ is constant
  - $p_{discard} \downarrow$: high tolerance to false alarms
  - $p_{discard} \uparrow$: low tolerance to false alarms

L. Docio-Fernandez, P. Lopez-Otero, C. Garcia-Mateo, "An Adaptive Threshold Computation for Unsupervised Speaker Segmentation"

Adaptive Strategy

- $p_{discard}$ and Θ_{BIC} increase when the system is accepting too many false alarms
- $p_{discard}$ and Θ_{BIC} decrease when the system is not accepting false alarms
- Initially: $p_{discard} = 0$, $\Theta_{BIC} = 0$

Poisson-based Strategy

- $p_{discard}$ follows a Poisson cumulative density function
- $p_{discard}$ depends on the expected number of change-points (mean of the distribution)
- Initially: $p_{discard} = 0$

Uniform-based Strategy

- $p_{discard}$ is constant
  - $p_{discard} \downarrow$: high tolerance to false alarms
  - $p_{discard} \uparrow$: low tolerance to false alarms

P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, "Novel Strategies for Reducing the False Alarm Rate in a Speaker Segmentation System"

Multimedia Technology Group (GTM)

Adaptive Strategy

- $p_{discard}$ and Θ_{BIC} increase when the system is accepting too many false alarms
- $p_{discard}$ and Θ_{BIC} decrease when the system is not accepting false alarms
- Initially: $p_{discard} = 0$, $\Theta_{BIC} = 0$

Poisson-based Strategy

- $p_{discard}$ follows a Poisson cumulative density function
- $p_{discard}$ depends on the expected number of change-points (mean of the distribution)
- Initially: $p_{discard} = 0$

Uniform-based Strategy

- $p_{discard}$ is constant
  - $p_{discard} \downarrow$: high tolerance to false alarms
  - $p_{discard} \uparrow$: low tolerance to false alarms

P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, "Novel Strategies for Reducing the False Alarm Rate in a Speaker Segmentation System"

Multimedia Technology Group (GTM)

Database

- TC-STAR 2006 ASR evaluation campaign
- Spanish parliament sessions

Metrics

$$\text{Precision: } P = \frac{c}{c+i} \times 100$$

$$\text{Recall: } R = \frac{c}{c+d} \times 100$$

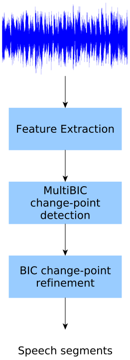
$$\text{F-score: } F = \frac{2PR}{P+R}$$

Results

System	P	R	F
Baseline	57.46	85.59	65.99
Adaptive	66.98	85.59	73.17
Uniform	80.97	81.75	81.22
Poisson	76.21	83.09	79.05

Multimedia Technology Group (GTM)

MultiBIC Strategy



BIC

$$\Delta BIC(i) = L_i - \lambda P$$

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

$$L_i = \frac{L}{2}log|\Sigma| - \frac{L_1}{2}log|\Sigma_1| - \frac{L_2}{2}log|\Sigma_2|$$

MultiBIC

$$\Delta MultiBIC(i,j) = L_{ij} - \lambda P$$

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

$$L_{ij} = \frac{L}{2}log|\Sigma| - \frac{L_1}{2}log|\Sigma_1| - \frac{L_2}{2}log|\Sigma_2| - \frac{L_3}{2}log|\Sigma_3|$$

P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, "MultiBIC: an Improved Speaker Segmentation Technique for TV Shows"

Database

- TV shows
  - *sit-coms*: canned laughter, jingles
  - drama series: background music

Metrics

$$\text{Precision: } P = \frac{c}{c+i} \times 100$$

$$\text{Recall: } R = \frac{c}{c+d} \times 100$$

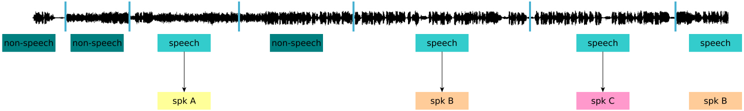
$$\text{F-score: } F = \frac{2PR}{P+R}$$

Results

Test		c	d	i	P	R	F
1	BIC	466	158	6	98.06	66.13	78.98
	MultiBIC		96	2	99.46	79.28	88.21
2	BIC	497	169	5	98.66	66.36	79.34
	MultiBIC		82	4	99.17	83.50	90.67
3	BIC	402	147	0	100	63.43	77.63
	MultiBIC		61	0	100	84.83	91.79

Multimedia Technology Group (GTM)

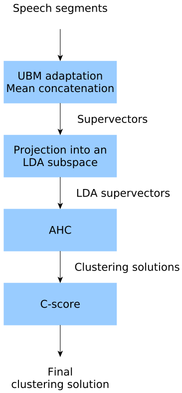
Speaker Diarization



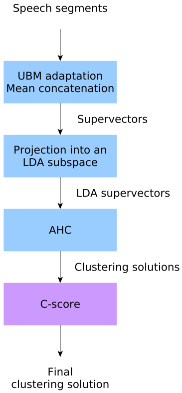
- Audio and speaker segmentation
  - Non-speech segments are removed
  - Speech segments of different speakers are divided
- Clustering
  - Agglomerative hierarchical clustering
  - How many clusters?

Applications: Speaker Diarization

The C-score Technique



Multimedia Technology Group (GTM)


Speech segments
UBM adaptation
Mean concatenation
Supervectors
Projection into an
LDA subspace
LDA supervectors
AHC
Clustering solutions
C-score
Final
clustering solution

- A number of clusters n^* is selected such that:
  - Intra-cluster similarity (I_n) is minimized
  - Extra-cluster similarity (E_n) is maximized

$$C\text{-}score_n = \frac{2I_n(1-E_n)}{I_n + (1-E_n)}$$

$$n^* = \underset{i=n_{min},\ldots,n_{max}}{\arg\max}\; C\text{-}score_i$$

Database

- Albayzin 2010 speaker diarization evaluation database
- Broadcast news programmes

Metrics

- Speaker error rate (SPKE)
- False alarm rate (FAS)
- Missed speech (MISS)

Results

Segmentation	Method	FAS	MS	SPKE (%)
Manual	C-score	0 %	0 %	16.1 ± 0.9
	BIC			29.0 ± 1.1
Automatic	C-score	2.2 %	7.3 %	15.0 ± 0.7
	BIC			19.4 ± 0.8

Multimedia Technology Group (GTM)

Database

- Transcrigal-DB
- Broadcast news programmes in Galician language

Metrics

$$WER = \frac{S + D + I}{W}$$

Results

	Speakers	Substitutions	Deletions	Insertions	WER
Manual segmentation	All	14.9 %	4.9 %	4.4 %	24.2 %
	Habitual	12.1 %	3.4 %	4.6 %	18.3 %
	Others	18.2 %	5.9 %	4.3 %	28.4 %
Automatic segmentation	All	14.9 %	6.5 %	3.6 %	25.0 %
	Habitual	10.6 %	4.5 %	3.6 %	18.7 %
	Others	17.9 %	7.8 %	3.7 %	29.4 %
Automatic segmentation and clustering	All	13.7 %	6.2 %	3.3 %	23.1 %
	Habitual	9.9 %	4.3 %	3.3 %	17.6 %
	Others	16.4 %	7.5 %	3.3 %	27.2 %

P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, A. Cardenal-Lopez, "On the Influence of Automatic Segmentation and Clustering in Automatic Speech Recognition"

Multimedia Technology Group (GTM)

Ongoing work: Emotion Recognition

- Automatic estimation of speaker's affect and depression level
- Application of speaker verification techniques to emotion recognition
    - Feature dimensionality reduction
    - Emotional characterization of speech segments

E. Sanchez-Lozano, P. Lopez-Otero, L. Docio-Fernandez, E. Argones-Rua, J.L. Alba-Castro, "Audiovisual Three-Level Fusion for Continuous Estimation of Russell's Emotion Circumplex"

Thank you for your attention

plopez@gts.uvigo.es