

Albayzin-2010 Audio Segmentation Evaluation: Evaluation Setup and Results

Taras Butko, Climent Nadeu and Henrik Schulz

TALP Research Center
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya, Barcelona, Spain



10.11.2010, Vigo



Outline

- Audio segmentation task in Albayzin-2010 evaluation
- Database and evaluation setup
- Results from participants
- Conclusions

Motivation

Audio segmentation task is proposed for the first time

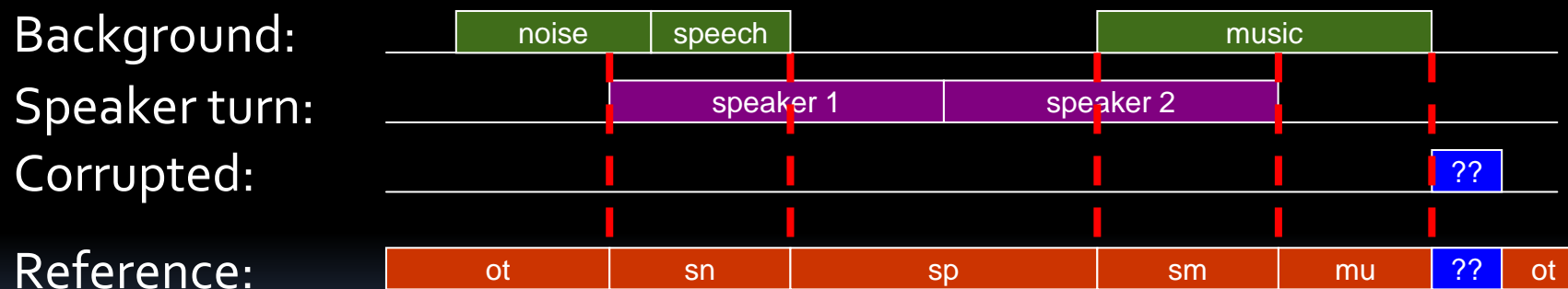
- Automatic indexing, subtitling, content analysis and information retrieval
- Improving speech technologies: ASR, SAD, speaker diarization
- Other applications: surveillance systems, sports highlights generation, violence detection, advertising detection etc

Practical motivation:

- A large and freely available annotated database is recently recorded in 2009 (Technoparla project)

Database

- 3/24 TV channel, broadcast news domain
- 87 hours of manually annotated audio (24 sessions, approximately 4 hours long each)
- Annotation layers:



- 5 acoustic classes are defined for evaluation:

Class	Music [mu]	Speech [sp]	Speech over music [sm]	Speech over noise [sn]	Other* [ot]
Proportion	5 %	37 %	15 %	40 %	3 %

*not evaluated

Metric

- We proposed a specific metric for evaluation:

$$Error = average_i \left(\frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)} \right)$$

Compare with NIST speaker diarization metric:

$$Error = \frac{dur(miss) + dur(fa) + dur(subst)}{dur(ref)}$$

This way we give more weight to minor classes (with lower value of $dur(ref_i)$).
We stimulate the participants to detect music and speech over music class

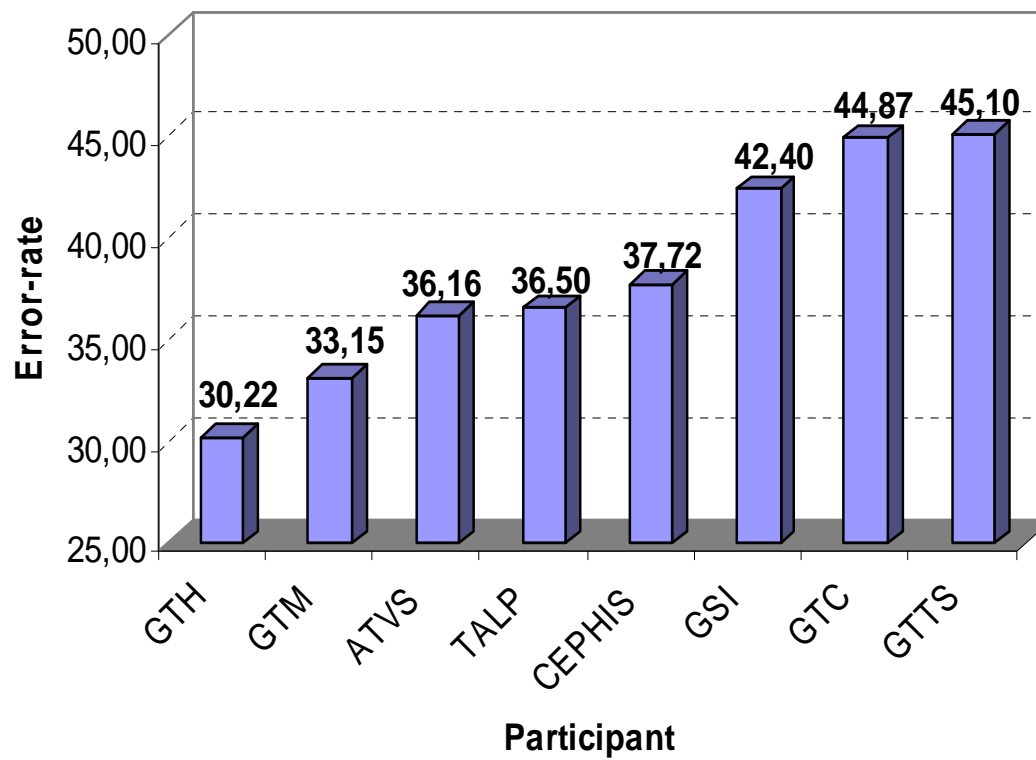
Participants

- 10 groups registered, 8 submitted results:
 - GTTS (Universidad del País Basco)
 - GTC-VIVOLAB (Universidad de Zaragoza)
 - GSI (Universidade de Coimbra, Portugal)
 - TALP (Universitat Politècnica de Catalunya)
 - CEPHIS (Universitat Autònoma de Barcelona)
 - ATVS(Universidad Autónoma de Madrid)
 - GTM (Universidad de Vigo)
 - GTH (Universidad Carlos III de Madrid / Universidad Politécnica de Madrid)

Evaluation conditions

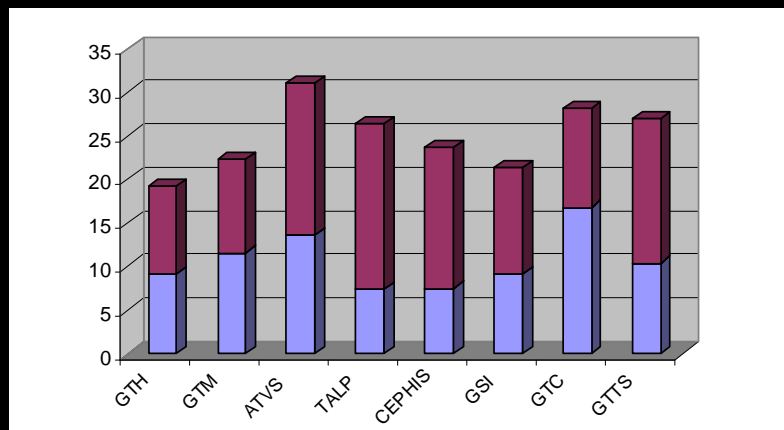
- 2/3 of the database for training/development, 1/3 for testing
- 3 months were given to participants to design their own segmentation system
- 2 weeks were given to perform testing
- Any publicly available data could be used to train the model in addition to the provided
- Listening to test data was not allowed

Results (I)

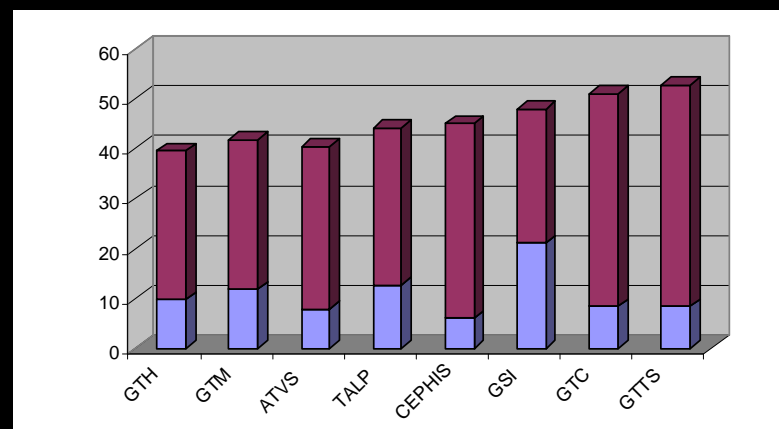


Results (II). Misses and FA

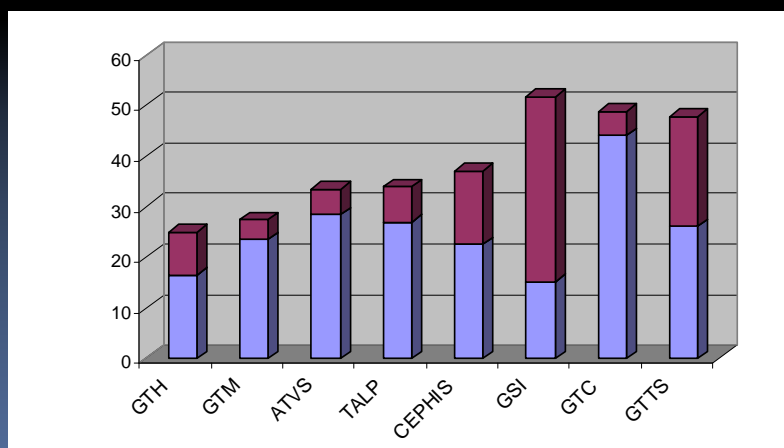
Misses False alarms



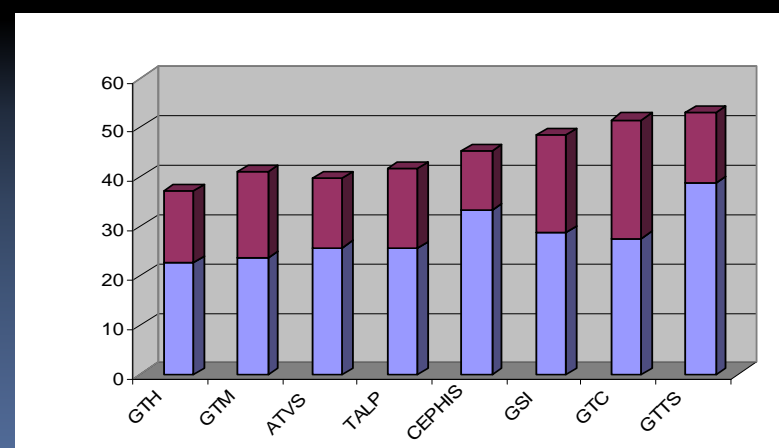
music



speech



speech over music



speech over noise

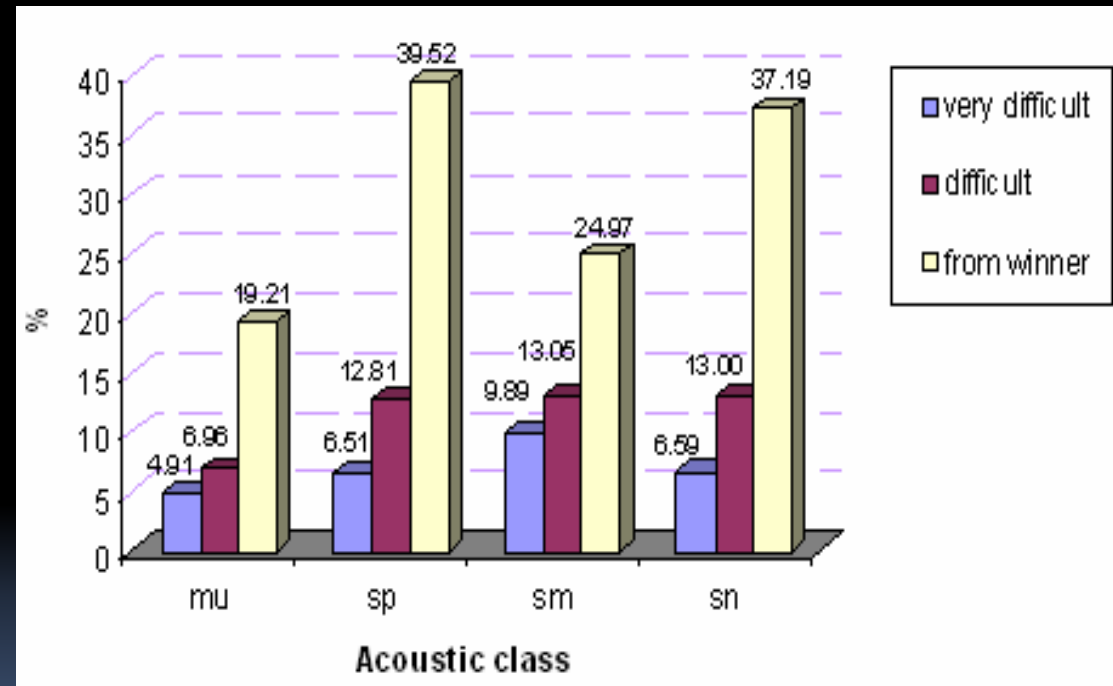
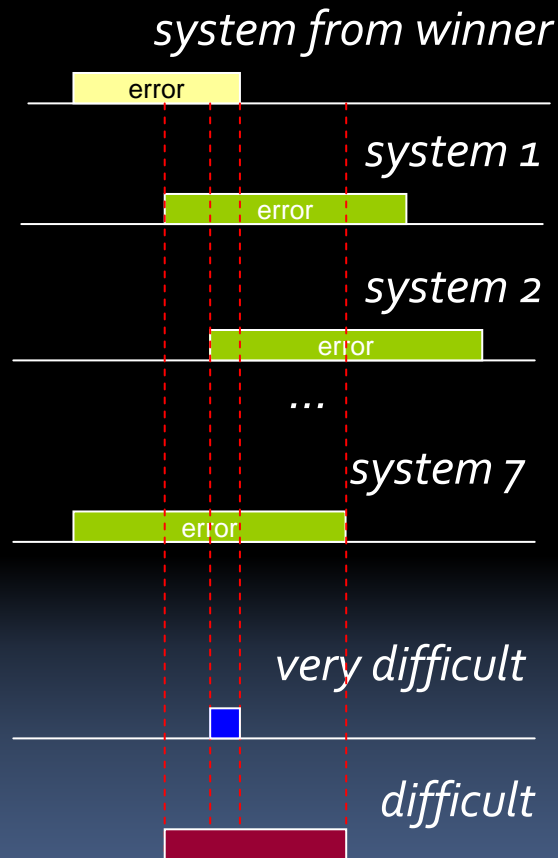
Results (III). Confusion matrix

	mu	sp	sm	sn
mu	89.4	0.1	8.0	2.5
sp	0.0	70.6	2.9	26.5
sm	1.8	1.2	87.0	10.0
sn	0.3	10.2	8.3	81.2

The matrix shows the percentage of hypothesized ACs (rows) that are associated to the reference ACs (columns), so that all the numbers out of the main diagonal correspond to confusions

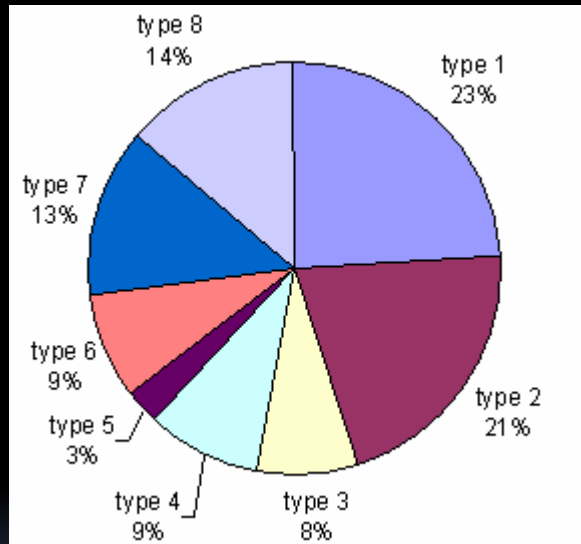
Difficulty of segmentation task

Hypothesis segmentation:




- *Very difficult*: 8/8 participants produced errors
- *Difficult*: 7 or more participants produced errors
- *From winner*: winner of evaluations produced errors

Analysis of “very difficult” segments



<i>Type of error</i>	<i>Description</i>
Type 1	Low level of background sound
Type 2	Speech in background
Type 3	Annotation error
Type 4	The microphone is affected by the wind
Type 5	Singing in background
Type 6	Noise in background is more dominant than music for the [sm] class
Type 7	The quality of music in background is low
Type 8	Other

-  type 2
-  type 4
-  type 6
-  type 7

Conclusions

- The audio segmentation evaluation was organized for the first time. 8 participants submitted their results in time and in correct way
- Main source of segmentation errors: low level of background sound and overlapped speech (but not mistakes of annotators)
- Audio segmentation task is still challenging

Thank you!