EURECOM submission to the Albayzin 2016 Speaker Diarization Evaluation

Jose Patino[1], Héctor Delgado[1], Nicholas Evans[1], Xavier Anguera[2]
[1]EURECOM, Sophia-Antipolis, France
[2]ELSA Corp., Portugal
{patino,delgado,evans}@eurecom.fr, xavier@elsanow.io

Introduction

Speaker diarization is the task of segmenting an audio document into speaker-homogeneous segments and clustering those segments according to speaker identities

Applications:

- Enabling speaker adaptation in ASR systems
- Enabling speaker recognition in multi-speaker data
- Spoken document indexing and retrieval

Albayzin Evaluation: Segmenting broadcast audio documents according to different speakers and attributing those segments to the speaker who uttered them, without any prior information about the speaker identities nor their number

EURECOM submission:

- Infinite impulse response – constant Q, Mel-frequency cepstral coefficients (ICMC) [1]
- Speaker diarization system based on the binary key modelling [2], an efficient and compact speech and speaker representation
- External training data is not required: the test data itself is used for training
- System does not perform overlapping speech detection

Database

Database: 100 hours split into two parts:

- 87 hours for **training** from the **Catalan broadcast news database**
- 4 and 16 hours for **development** and **test** respectively from the **Corporación Aragonesa de Radio y Televisión (CARTV)**

Acoustic classes annotations provided

- Speech
- Music
- Background noise

Evaluation metric

Diarization error time for each segment *n*

$$E(n) = T(n)[\max(N_{ref}(n), N_{sys}(n)) - N_{correct}(n)]$$

$T(n)$:	Duration of segment *n*
$N_{ref}(n)$:	Number of speakers that are present in segment *n*
$N_{sys}(n)$:	Number of system speakers present in segment *n*
$N_{correct}(n)$:	Number of reference speakers in segment *n* which are correctly assigned by the diarization system

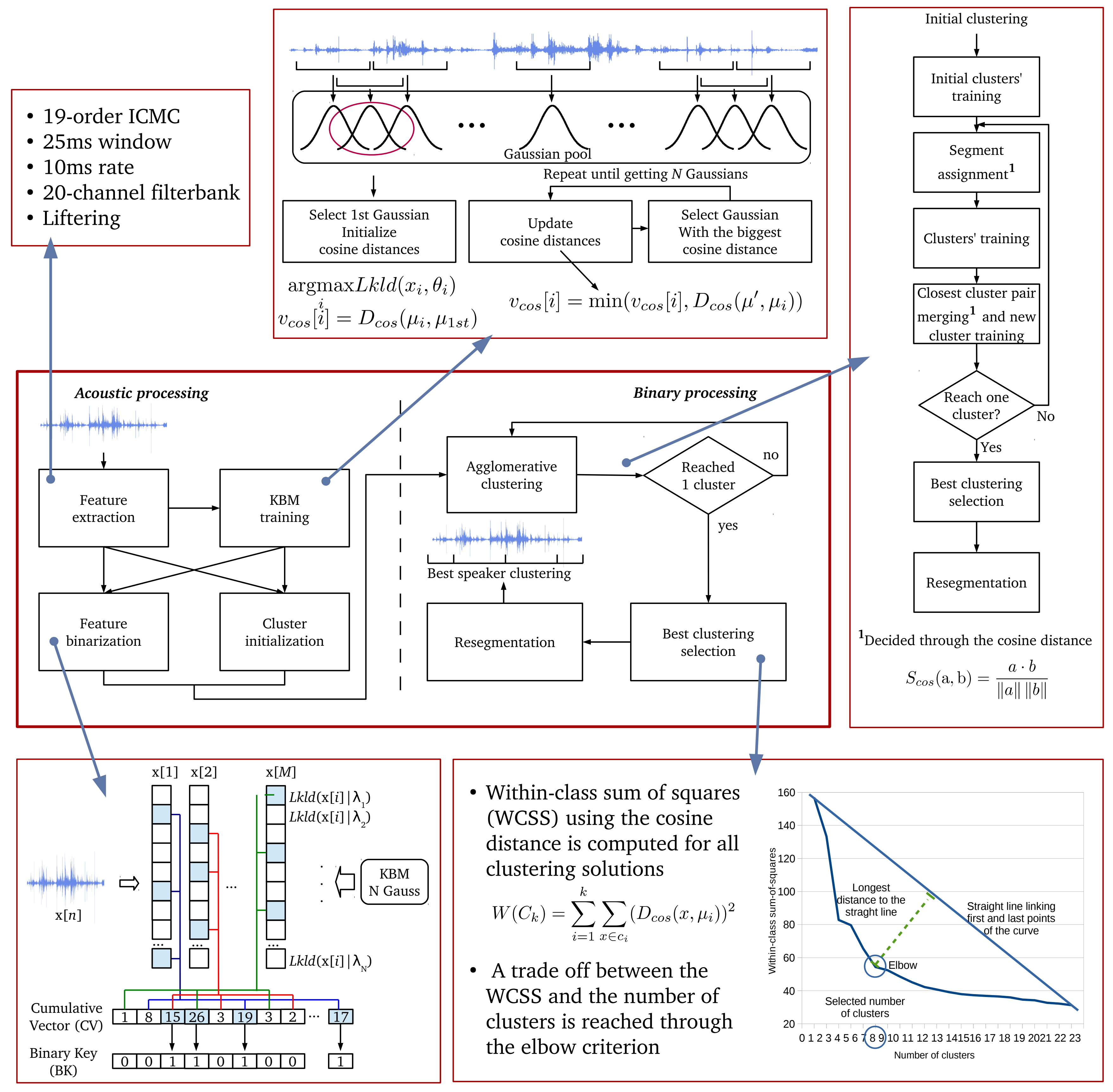
Diarization Error Rate (DER)

$$DER = \frac{\sum_{n\in\Omega} E(n)}{\sum_{n\in\Omega}(T(n)N_{ref}(n))}$$

DER may be decomposed as:

- ***Speaker error:*** time wrongfully assigned to a speaker
- ***Missed speech:*** time where speech is present but is not labelled by the system
- ***False alarm:*** amount of time that has been assigned to speech which is not present

Binary key speaker diarization system



- 19-order ICMC
- 25ms window
- 10ms rate
- 20-channel filterbank
- Liftering

$\underset{i}{\text{argmax}} Lkld(x_i, \theta_i)$

$v_{cos}[i] = D_{cos}(\mu_i, \mu_{1st})$

$v_{cos}[i] = \min(v_{cos}[i], D_{cos}(\mu', \mu_i))$

[1]Decided through the cosine distance

$$S_{cos}(\mathrm{a},\mathrm{b}) = \frac{a \cdot b}{\|a\|\,\|b\|}$$

- Within-class sum of squares (WCSS) using the cosine distance is computed for all clustering solutions

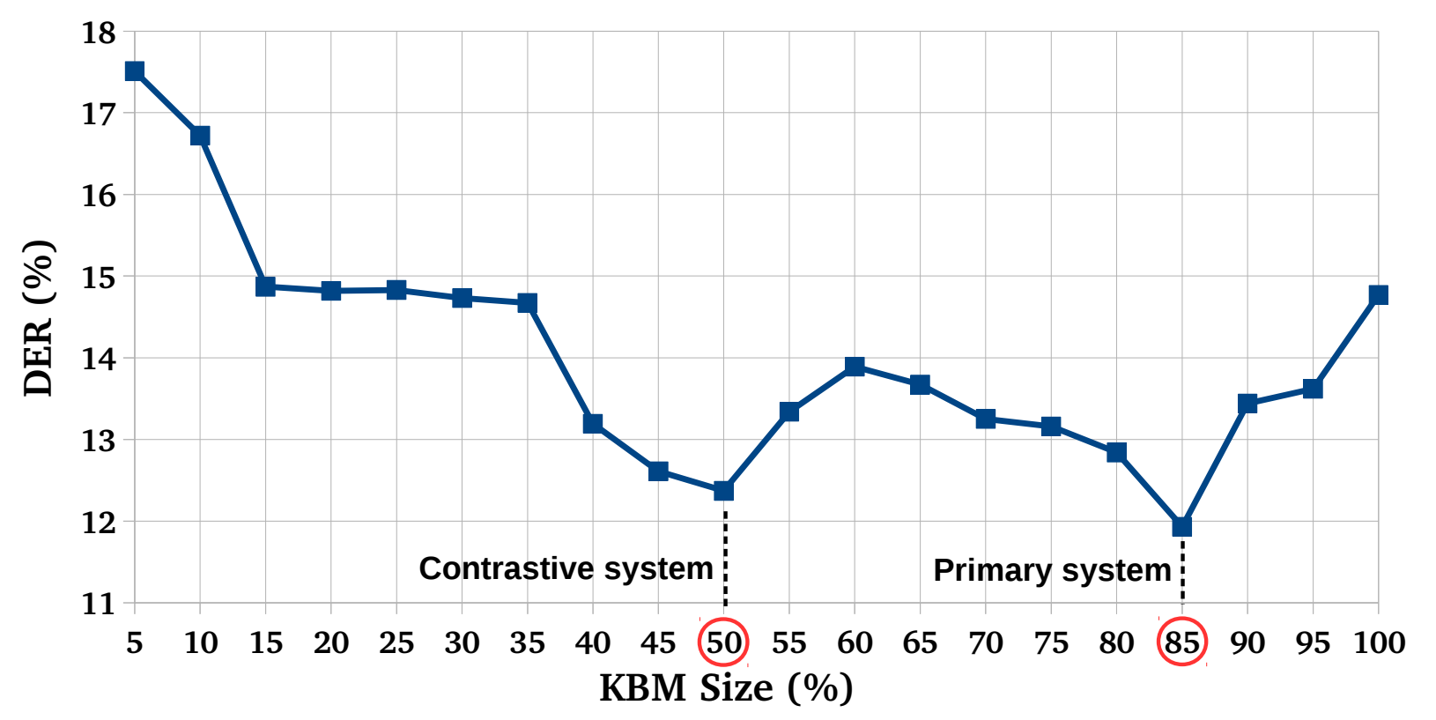
$$W(C_k) = \sum_{i=1}^{k}\sum_{x\in c_i}(D_{cos}(x,\mu_i))^2$$

- A trade off between the WCSS and the number of clusters is reached through the elbow criterion

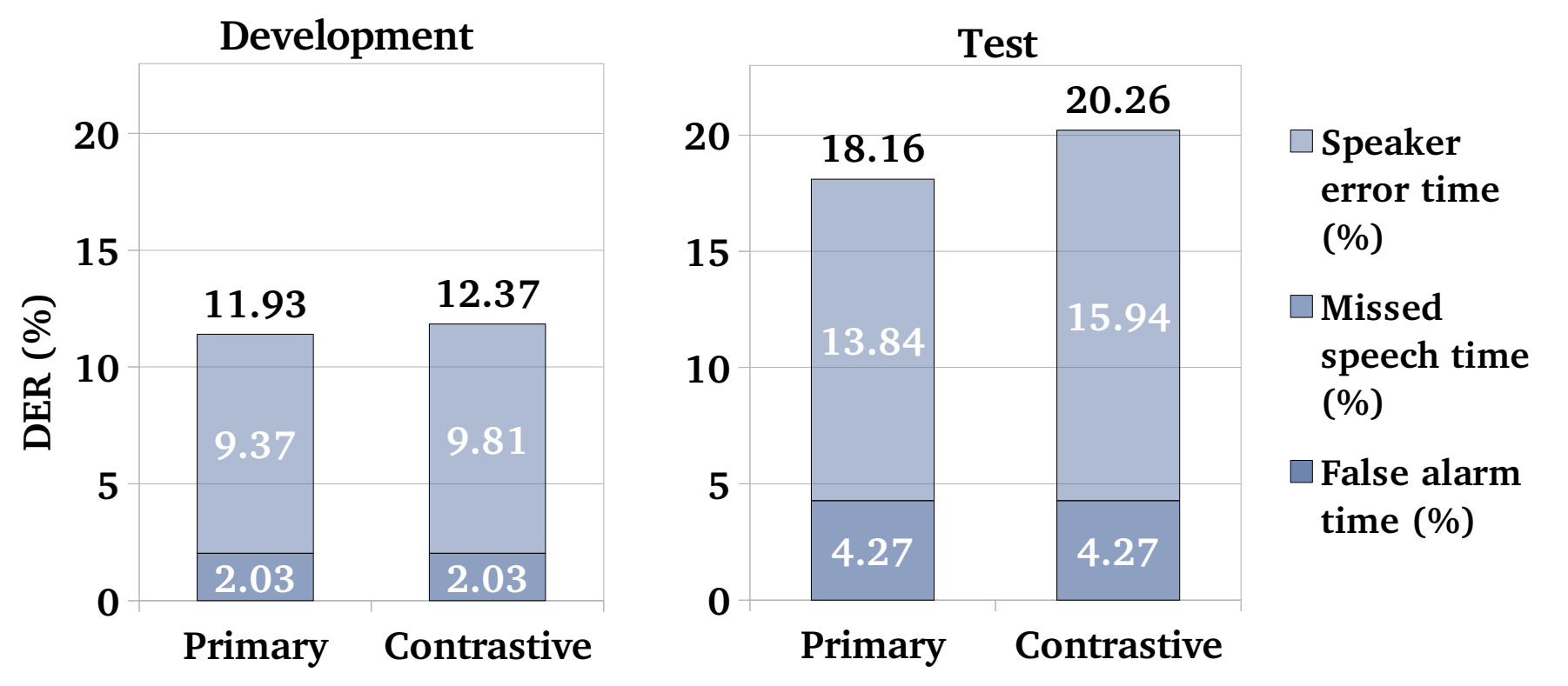
160
140
120
100
80
60
40
20
Within-class sum-of-squares
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Number of clusters
Longest distance to the straight line
Straight line linking first and last points of the curve
Elbow
Selected number of clusters

Results

System results



DER trend for different KBM sizes on the development set



DER of submitted systems on the development and test data (with 0.5% and 0.05% false alarm rates, respectively)

- KBM size as a percentage of the initial number of Gaussian components, which is related to the length of the speech content
- Exploration of relative KBM size led to two working points, submitted as primary and contrastive system

- Primary system, with a bigger relative KBM, outperforms contrastive system on development and test
- DER on test data is an approx. absolute 8% worse than on the development set for both primary and contrastive systems

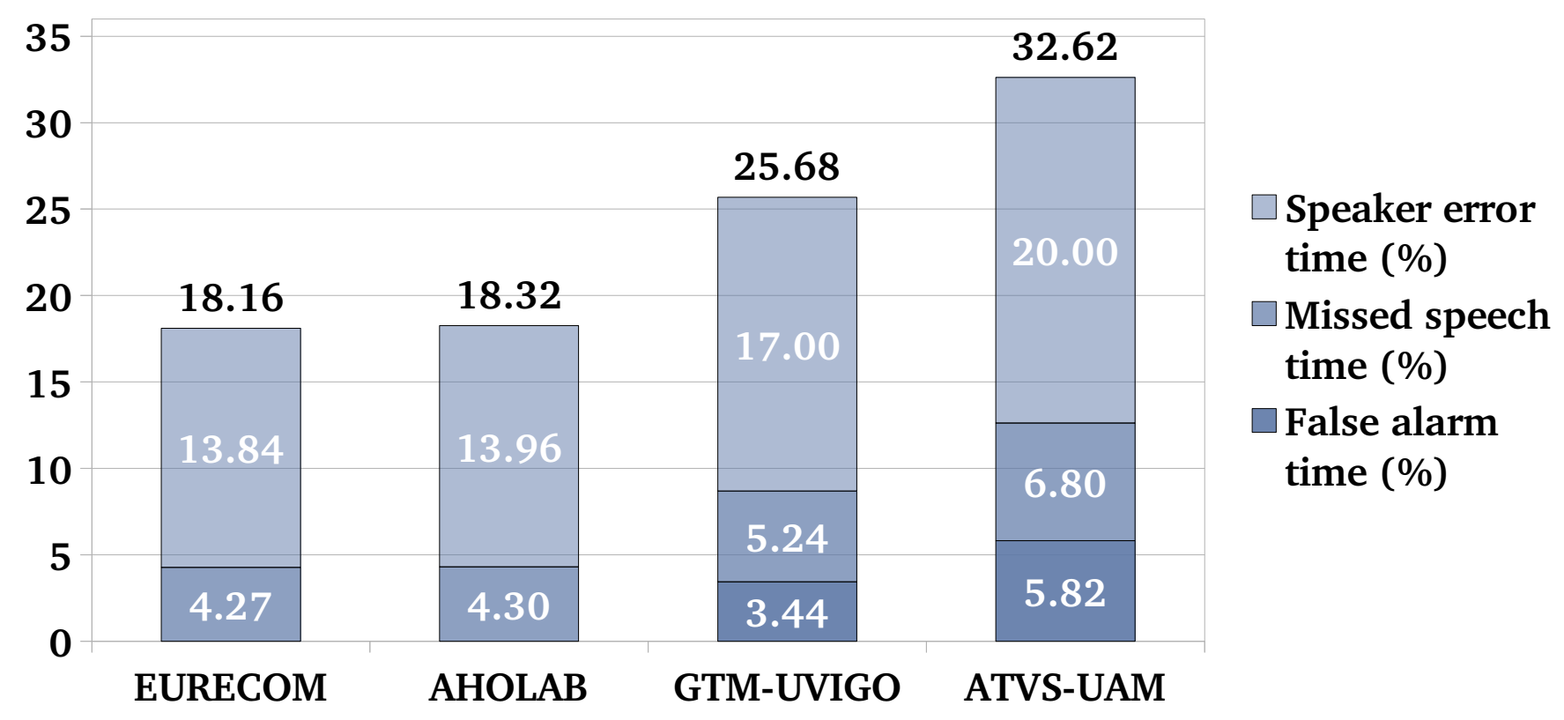
Execution time

	Primary system		Contrastive system	
Task	Time	xRT	Time	xRT
Feature extraction	00:49:11	0.046	00:49:11	0.046
Speaker diarization	00:39:59	0.044	00:32:01	0.035
Overall	01:29:10	0.045	01:21:12	0.0405

Execution time taken by primary and contrastive systems when processing the test data (around 16 hours of audio). Real time factor (xRT) is also provided

- Very low execution times, suitable for processing big amounts of data and for online operation with low latencies
- Contrastive system provides a faster alternative at the cost of decreased performance

All participants results



All participants primary systems results (with 0.05% and 0.06% false alarm rates for systems 1 and 2)

- The top two systems performed very similarly with only an absolute 0.16% difference
- Systems 3 and 4 exhibited higher missed speech and false alarm error rates, possibly influencing speaker error rates

Conclusions

- Speaker diarization system based on the ICMC features and binary key modelling
- Data from the training set was not employed. System tuned on the development set only
- Two different working points were chosen in order to compose the KBM in the primary and contrastive systems
- Best result of experiments on development data is 11.93% DER
- Official result on the Albayzin Speaker Diarization Evaluation is 18.16% DER
- System was the 1st ranked among the submissions, with a slight difference over the 2nd system and an approximate absolute 7% over the 3rd ranked

Acknowledgments

This work was partially supported with funding from the French Agence Nationale de la Recherche Project ODESSA (ANR-15-CE39-0010)

References

[1] Delgado, H., Todisco, M., Sahidullah, M., Sarkar, A. K., Evans, N., Kinnunen, T., & Tan, Z. H. (2016). Further optimisations of constant Q cepstral processing for integrated utterance and text-dependent speaker verification. In IEEE Spoken Language Technology (SLT) Workshop

[2] Delgado, H., Anguera, X., Fredouille, C., & Serrano, J. (2015). Fast single-and cross-show speaker diarization using binary key speaker modeling. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 23(12), 2286-2297