Novel Clustering Selection Criterion for Fast Binary Key Speaker Diarization

Héctor Delgado\textsuperscript{1}, Xavier Anguera\textsuperscript{2}, Corinne Fredouille\textsuperscript{3}, Javier Serrano\textsuperscript{1}

\textsuperscript{1}CAIAC, Universitat Autònoma de Barcelona, Barcelona, Spain
\textsuperscript{2}Sinkronigo S.L., Barcelona, Spain
\textsuperscript{3}CERI/LIA, University of Avignon, France

September 10, 2015
1 Introduction

2 Binary Key Speaker Diarization

3 Speeding-up binary key speaker diarization

4 Proposed final clustering selection criterion for binary key speaker diarization

5 Experiments and results

6 Conclusions
1. Introduction

2. Binary Key Speaker Diarization

3. Speeding-up binary key speaker diarization

4. Proposed final clustering selection criterion for binary key speaker diarization

5. Experiments and results

6. Conclusions
1. Introduction

Speaker diarization

Who spoke when?

No prior information about

- Number of speakers
- Speaker identities
1. Introduction

Speaker diarization

Who spoke when?

No prior information about
- Number of speakers
- Speaker identities
1. Introduction

Speaker diarization

Who spoke when?

No prior information about

- Number of speakers
- Speaker identities
1. Introduction

Generic speaker diarization scheme

- **Iterative scheme**
- **Intensive computation**: Speaker models re-training or adaptation, Viterbi re-assignment...
- **Long processing times**
A fast diarization system based on binary keys speaker modeling was presented\textsuperscript{1}

Based on a speaker modeling based on binary keys\textsuperscript{2}

Promising results and important speed gain on the NIST meeting audio databases

- Average DER: 25.06 %
- Speed $\approx 0.1 \times$RT

\textsuperscript{1}Anguera, X.; Bonastre, J.-F. “Fast speaker diarization based on binary keys,” in Proc. Acoustics, Speech and Signal Processing (ICASSP), 2011

1. Introduction

Binary Key speaker diarization evolution (2)

- Applied to broadcast TV data (REPERE dataset) with similar results ($\sim 23\%$ DER)$^3$
- Further improved by using Cumulative Vectors and Cosine Distance: ($\sim 19\%$ DER)$^4$

---


Main goal: Fast and accurate speaker diarization system which does not require external training data

But...

- Clustering selection still does not return near-optimum solution
- A suitable final clustering selection criterion is required
- Can execution time be further reduced? (< 0.1xRT)

---

2. Binary Key Speaker Diarization

Outline

1. Introduction
2. Binary Key Speaker Diarization
3. Speeding-up binary key speaker diarization
4. Proposed final clustering selection criterion for binary key speaker diarization
5. Experiments and results
6. Conclusions
2. Binary Key Speaker Diarization

Binary Key Speaker Diarization System

Input audio
2. Binary Key Speaker Diarization

Binary Key Speaker Diarization System

Input audio

MFCCs
2. Binary Key Speaker Diarization

Binary Key Speaker Diarization System

Input audio

MFCCs

KBM
2. Binary Key Speaker Diarization

Binary Key Speaker Diarization System

- Input audio
- MFCCs
- Binary Keys
- KBM

Input audio

MFCCs

Binary Keys

KBM
2. Binary Key Speaker Diarization

Binary Key Speaker Diarization System

Input audio → MFCCs → Binary Keys

KBM → \( C_1 \), \( C_2 \), ..., \( C_N \) \( N \) Initial clusters
2. Binary Key Speaker Diarization

Binary Key Speaker Diarization System

Input audio
MFCCs
Binary Keys
KBM
...
...
N Initial clusters
C_1 C_2 \ldots C_N
...
Cluster Binary Keys
2. Binary Key Speaker Diarization

Binary Key Speaker Diarization System

Input audio

MFCCs

Binary Keys

KBM

N Initial clusters

Cluster

Binary Keys

Data assignment

\[ S(v_{f_1}, v_{f_2}) = \frac{\sum_{i=1}^{N} v_{f_1} \land v_{f_2}}{\sum_{i=1}^{N} v_{f_1} \lor v_{f_2}} \]
2. Binary Key Speaker Diarization

Binary Key Speaker Diarization System

Input audio
MFCCs
Binary Keys
KBM

N Initial clusters
C_1 C_2 ... C_N

Cluster Binary Keys

Data assignment
\[ S(v_{f1}, v_{f2}) = \frac{\sum_{i=1}^{N} v_{f1} \land v_{f2}}{\sum_{i=1}^{N} v_{f1} \lor v_{f2}} \]

Closest cluster pair merging
C_1 ... C_{N-1}

N-1 clusters
2. Binary Key Speaker Diarization

Binary Key Speaker Diarization System

Input audio
MFCCs
Binary Keys
KBM
...
...
N Initial clusters
C_1 C_2 C_N
...
Cluster
Binary Keys
...C_1 C_2 ... n
0
0
1
1
.
.
.
0
0
0
1
0
.
.
.
1
1
0
0
0
.
.
.
1
0
0
1
1
.
.
.
0
0
0
1
1
.
.
.
0
Héctor Delgado
Novel Clustering Selection Criterion for Fast Binary Key Speaker Diarization
2. Binary Key Speaker Diarization

KBM training

Select 1\textsuperscript{st} Gaussian
Initialize KL2 distances

\[ \arg\max_{i} KL2(x_i, \theta_i) \]
\[ v_{KL2}[i] = KL2(\theta_i, \theta_{1st}) \]

Update KL2 distances

Select Gaussian with highest KL2 distance

\[ v_{KL2}[i] = \min(v_{KL2}[i], KL2(\theta', \theta_i)) \]
2. Binary Key Speaker Diarization

Binary key computation

\[ x[n] \]
2. Binary Key Speaker Diarization

Binary key computation

$\lambda_1 \quad x[1] \quad \lambda_2 \quad x[2] \quad \lambda_3 \quad x[M]$

$\lambda_8 \quad \ldots \quad \lambda_8 \quad \ldots \quad \lambda_8 \quad \ldots$

$\lambda_N \quad \ldots \quad \lambda_N \quad \ldots \quad \lambda_N \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$

$\ldots \quad \ldots \quad \ldots \quad \ldots$
Binary key computation

\[ L(x[i] \mid \lambda_1) \]
\[ L(x[i] \mid \lambda_2) \]
\[ L(x[i] \mid \lambda_N) \]

\[ x[1] \quad x[2] \quad \ldots \quad x[M] \]

KBM
N Gauss
2. Binary Key Speaker Diarization

Binary key computation

\[ x[n] \]

\[ \lambda_1 \quad \lambda_2 \quad \lambda_3 \]

\[ x[1] \quad x[2] \quad \ldots \quad x[M] \]

\[ L(x[i] | \lambda_1) \quad L(x[i] | \lambda_2) \]

\[ \lambda_N \]

\[ L(x[i] | \lambda_N) \]
2. Binary Key Speaker Diarization

Binary key computation

\[
x[n] \
\lambda_1 \quad \lambda_2 \quad \lambda_3 \quad \cdots \quad \lambda_N \\]

\[
x[1] \quad x[2] \quad \cdots \quad x[M] \\]

\[
L(x[i]|\lambda_1) \quad L(x[i]|\lambda_2) \quad \cdots \quad L(x[i]|\lambda_N) \\]

Cummulative Vector

\[
0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad \cdots \quad 0 \\]
2. Binary Key Speaker Diarization

Binary key computation
2. Binary Key Speaker Diarization

Binary key computation

\[ \begin{align*}
\lambda_1 & \xrightarrow{L(x[i]|\lambda_1)} x[1] \\
\lambda_2 & \xrightarrow{L(x[i]|\lambda_2)} x[2] \\
\lambda_3 & \xrightarrow{\cdots} \lambda_M \\
\lambda_N & \xrightarrow{L(x[i]|\lambda_N)} \text{KBM} \\
\end{align*} \]

Cummulative Vector

0 2 0 1 0 1 0 1  \ldots  1
2. Binary Key Speaker Diarization

Binary key computation

\[ KBM \]
\[ N \text{ Gauss} \]
\[ x[n] \]
\[ L(x[i]|\lambda_1) \]
\[ L(x[i]|\lambda_2) \]
\[ L(x[i]|\lambda_N) \]
\[ x[1] \]
\[ x[2] \]
\[ x[M] \]
\[ \lambda_1 \]
\[ \lambda_2 \]
\[ \lambda_3 \]
\[ \lambda_8 \]
\[ \lambda_N \]

Cumulative Vector

1 15 8 26 3 19 3 2 17
2. Binary Key Speaker Diarization

Binary key computation

\[
L(x[i]|\lambda_1) \quad L(x[i]|\lambda_2) \quad \ldots \quad L(x[i]|\lambda_N)
\]

Cummulative Vector:

\[
1 \quad 15 \quad 8 \quad 26 \quad 3 \quad 19 \quad 3 \quad 2 \quad \ldots \quad 17
\]

Binary Key:

\[
0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0
\]
Binary Key Speaker Diarization

Binary key computation

\[ x[n] \]

\[
\begin{align*}
\lambda_1 & \quad x[1] \\
\lambda_2 & \quad x[2] \\
\lambda_3 & \\
\lambda_8 & \\
\lambda_N & \\
L(x[i]|\lambda_1) & \\
L(x[i]|\lambda_2) & \\
L(x[i]|\lambda_N) & \\
\end{align*}
\]

Cummulative Vector

<table>
<thead>
<tr>
<th>1</th>
<th>15</th>
<th>8</th>
<th>26</th>
<th>3</th>
<th>19</th>
<th>3</th>
<th>2</th>
</tr>
</thead>
</table>

Binary Key

| 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 |

KBM N Gauss
2. Binary Key Speaker Diarization

Binary key computation


Final clustering selection

$$T_s = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Select the optimum clustering by using the T-test metric$^8$
- $m_1, \sigma_1, n_1, m_2, \sigma_2$ and $n_2$ are the mean, standard deviation and size of intra-cluster and inter-cluster distance distributions, respectively

Not very accurate. Main system’s drawback!

---

Outline

1. Introduction
2. Binary Key Speaker Diarization
3. Speeding-up binary key speaker diarization
4. Proposed final clustering selection criterion for binary key speaker diarization
5. Experiments and results
6. Conclusions
Current Gaussian selection process based on the **Symmetrized Kullback-Leibler** (KL2) distance

\[
D_{KL2} = D_{KL}(P\|Q) + D_{KL}(Q\|P)
\]

KL divergence for normal multivariate distributions:

\[
KL(P\|Q) = \frac{1}{2} \left( \text{tr}(\Sigma_Q^{-1}\Sigma_P) + (\mu_Q - \mu_P)^t\Sigma_P^{-1}(\mu_Q - \mu_P) - k - \ln\left(\frac{\det\Sigma_P}{\det\Sigma_Q}\right) \right)
\]

Involves matrix operations like traces, inversions and determinants
3. Speeding-up binary key speaker diarization

**Speeding up KBM training: Cosine distance**

- How to lighten Gaussian comparison?
- Can we do *without the Gaussian covariance matrices* and use the *means* only?

Cosine distance:

\[ D_{\text{cos}}(a, b) = 1 - S_{\text{cos}}(a, b) \]

where

\[ S_{\text{cos}}(a, b) = \frac{a \cdot b}{\|a\| \|b\|} \]

- Does not take into account the vector’s magnitude, but the direction defined
- Measures the cosine of the angle between the two vectors being compared
4. Proposed final clustering selection criterion for binary key speaker diarization

Outline

1. Introduction

2. Binary Key Speaker Diarization

3. Speeding-up binary key speaker diarization

4. Proposed final clustering selection criterion for binary key speaker diarization

5. Experiments and results

6. Conclusions
Final clustering selection

- Given a clustering solution $C_k$ of $k$ clusters $c_1, c_2, ..., c_k$, each cluster containing CVs representing speech segments

- **Within-Cluster Sum of Squares (WCSS):**

  $$W(C_k) = \sum_{i=1}^{k} \sum_{x \in c_i} \| x - \mu_i \|^2$$

  where $\mu_i$ is the mean of the points of cluster $c_i$ (centroid)

- Set of clustering solutions $C = (C_1, ..., C_{N_{init}})$ with a decreasing number of clusters (from a single cluster to $N_{init}$ clusters)

- Calculate WCSS for all $C_i$ in $C$ using the cosine distance
Final clustering selection: Elbow criterion

The graph plots the within-class sum-of-squares against the number of clusters. The elbow criterion is used to determine the optimal number of clusters, which is the point where the decrease in the within-class sum-of-squares begins to level off.
4. Proposed final clustering selection criterion for binary key speaker diarization

Final clustering selection: Elbow criterion

![Graph showing the Elbow criterion for selecting the number of clusters. The x-axis represents the number of clusters, and the y-axis represents the within-class sum-of-squares. The graph shows a sharp decrease in the sum-of-squares as the number of clusters increases from 0 to 3, followed by a gradual decrease as the number of clusters increases further.]
Final clustering selection: Elbow criterion

![Graph showing the Elbow criterion for clustering selection. The x-axis represents the number of clusters, and the y-axis represents the within-class sum-of-squares. The graph illustrates a straight line linking the first and last points of the curve, indicating the optimal number of clusters.](image-url)
Final clustering selection: Elbow criterion
4. Proposed final clustering selection criterion for binary key speaker diarization

**Final clustering selection: Elbow criterion**

![Graph showing Elbow criterion](image)

- **Within-class sum-of-squares** vs. **Number of clusters**
- **Elbow** point on the graph indicates the optimal number of clusters.
- **Longest distance to the straight line**
- **Straight line linking first and last points of the curve**

Héctor Delgado

**Novel Clustering Selection Criterion for Fast Binary Key Speaker Diarization**
4. Proposed final clustering selection criterion for binary key speaker diarization

Final clustering selection: Elbow criterion

![Graph showing the elbow criterion for clustering selection.](image)

- **Elbow criterion**
  - **Selected number of clusters**
  - **Longest distance to the straight line**
  - **Straight line linking first and last points of the curve**

Number of clusters: 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, 52, 56, 60, 64, 68, 72, 76, 80, 84, 88, 92, 96, 100, 104, 108, 112, 116, 120, 124, 128, 132, 136, 140, 144, 148, 152, 156, 160.
5. Experiments and results

Outline

1. Introduction
2. Binary Key Speaker Diarization
3. Speeding-up binary key speaker diarization
4. Proposed final clustering selection criterion for binary key speaker diarization
5. Experiments and results
6. Conclusions
Evaluation setup

Database: **REPERE Phase 1 test set** (broadcast TV shows)

- Standard 19-order MFCCs
- Ground-truth SAD labels
- KBM settings
  - 2s window
  - Shift adjusted to get a pool of around 2000 Gaussians
  - KBM size: free parameter
- Cumulative Vectors: 5 top Gaussians at frame level
- Clustering initialization: 25 uniform-initialized clusters
- Data mapping: 1s segments (adding 1s after and before = 3s)

Evaluation metrics:

- DER with 0.25s forgiveness collar, overlapping speech is accounted
- Real-time factor ($xRT$): $xRT = \frac{t_{system}}{dur_{speech}}$ (excluding feature extraction)
5. Experiments and results

New KBM training results

**SysOut**: System output

**OptOut**: Result of the best clustering selected manually
5. Experiments and results

New KBM training results

**SysOut**: System output

**OptOut**: Result of the best clustering selected manually
5. Experiments and results

Final clustering selection results

![Graph showing DER vs KBM size for OptOut and FAST KBM]
Final clustering selection results

5. Experiments and results

- OptOut
- FAST KBM
- WCSS Euclidean

DER vs KBM size graph
5. Experiments and results

Final clustering selection results

![Graph showing clustering selection results](image-url)
## Results summary

<table>
<thead>
<tr>
<th>KM Size</th>
<th>DER (%)</th>
<th>Abs. red.</th>
<th>Rel. red.</th>
<th>xRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline system</td>
<td>576</td>
<td>19.12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+ Fast KBM</td>
<td>704</td>
<td>17.5</td>
<td>1.62</td>
<td>8.47</td>
</tr>
<tr>
<td>+ New criterion</td>
<td>320</td>
<td>15.15</td>
<td>3.97</td>
<td>20.76</td>
</tr>
</tbody>
</table>
6. Conclusions

Conclusions and future work

- New KBM training based on cosine distance significantly speeds up the process
- New clustering selection criterion based on WCSS improves DER of output solution
- Obtained performance very near to state-of-the-art in this database, while being very fast

Future work

- The error floor has not been reached yet
- Further refine final clustering selection
- More speed ups? xRT < 0.01? (100 times faster-than-real-time)
Thank you!

hecdelflo@gmail.com

Download the speaker diarization system Matlab code from:

http://hectordelgadome.software