Improved Binary Key Speaker Diarization System

Héctor Delgado[1], Xavier Anguera[2], Corinne Fredouille[3], Javier Serrano[1]

[1]CAIAC, Autonomous University of Barcelona, Cerdanyola del Vallès, Spain / [2]Sinkronigo S.L., Barcelona, Spain / [3]University of Avignon, CERI/LIA, France
hecdelflo@gmail.com, xanguera@gmail.com, corinne.fredouille@univ-avignon.fr, javier.serrano@uab.cat

EUSIPCO, 31th August – 4th September 2015
Nice, France

Introduction

Speaker diarization is the task of segmenting an audio document into speaker-homogeneous segments.
- Who spoke when?
- Speaker identities are unknown
- Number of speakers is unknown

Applications:
- Enable speaker adaptation in ASR systems
- Enable speaker recognition in multi-speaker data
- Spoken document indexing and retrieval
- Spoken document rich transcription

Binary Key speaker diarization: Fast speaker diarization system based on the binary key speaker modeling. Fast alternative with up to 0.037 xRT (real-time factor, see Interspeech'15 paper).

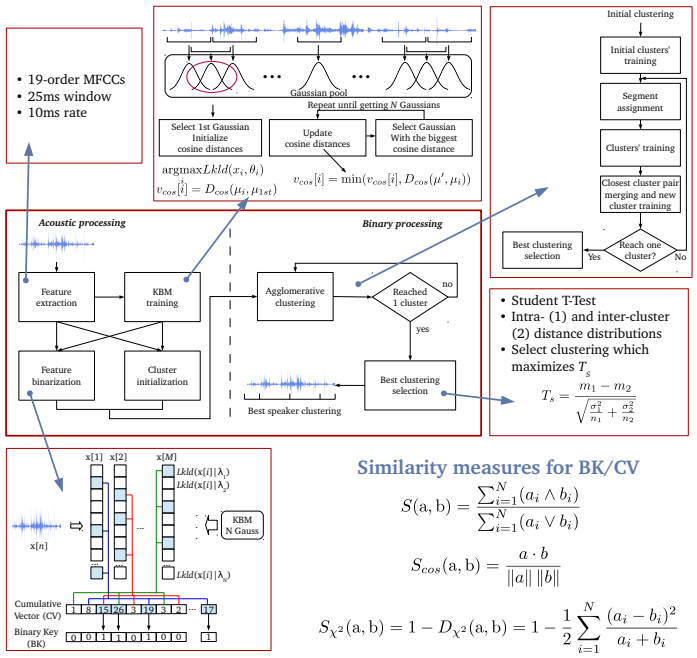
Challenge 1: Binary key speaker modeling
- Speed ups achieved at the cost of a degradation of diarization performance
- It is thought that the binarization step discards speaker related information useful for segregating speakers
- Improve speaker modeling to get closer to state-of-the-art

Challenge 2: Intra-session and intra-speaker variability (ISISV)
- Highly varying background conditions in TV and radio audio data, even within an audio session (background noise, background music, clean environment, etc.)
- Such variability may lead systems to model a given speaker into several clusters
- Compensating ISISV on the binary key domain

Goals

- **Use the cumulative vectors (CV) as speaker models in place of binary keys**
- **Propose suitable similarity measures for CVs**
- **Perform intra-session intra-speaker variability compensation through the Nuisance Attribute Projection (NAP) on the binary key domain**

Binary key speaker diarization system



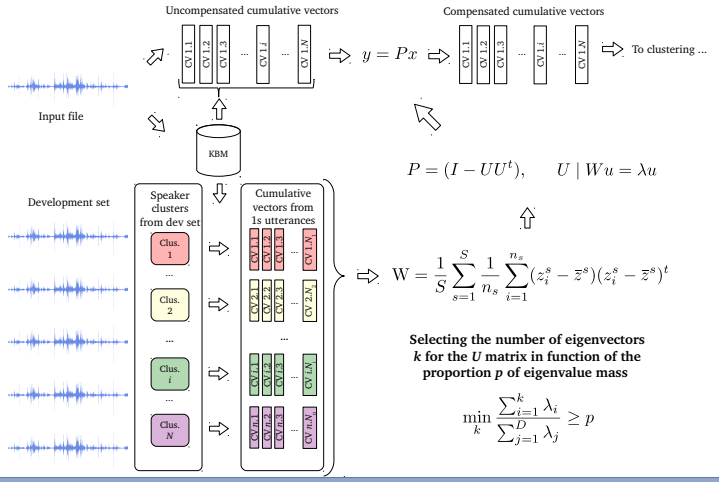
Similarity measures for BK/CV

$$S(a,b) = \frac{\sum_{i=1}^{N}(a_i \wedge b_i)}{\sum_{i=1}^{N}(a_i \vee b_i)}$$

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

$$S_{\chi^2}(a,b) = 1 - D_{\chi^2}(a,b) = 1 - \frac{1}{2}\sum_{i=1}^{N}\frac{(a_i - b_i)^2}{a_i + b_i}$$

Session variability compensation



$$y = Px$$

$$P = (I - UU^t), \qquad U \mid Wu = \lambda u$$

$$W = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(z_i^s - \bar{z}^s)(z_i^s - \bar{z}^s)^t$$

Selecting the number of eigenvectors k for the U matrix in function of the proportion p of eigenvalue mass

$$\min_k \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{j=1}^{D}\lambda_j} \geq p$$

Experimental results

System set-up

KBM training:
- 2s window
- Rate set to obtain around 2000 Gaussians.

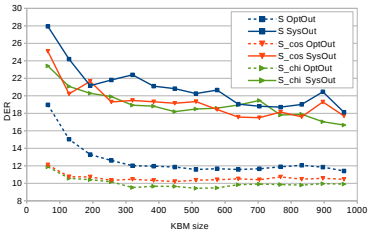
Binary key estimation:
- Top 5 Gaussians at frame level
- Top 20% of components at segment level

Clustering initialization:
- 25 flat-initialized uniform clusters

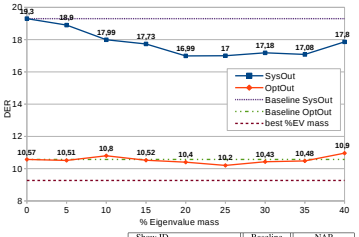
Segment assignment:
- 1s segments, extended 1s before and after (totaling 3s)

Database:
- REPERE phase 1 test set: 3 hours of labeled data distributed into 28 excerpts from French TV channels
- REPERE phase 1 development set: 3 hours of labeled data, used for the estimation of W

Similarity measures for BK/CV



Session variability compensation



Discussion

Similarity measures

- CVs provide more accurate speaker modeling than the BKs
- Chi-squared similarity is the best performing measure
- Cosine similarity also outperforms the baseline similarity metric
- For all similarity metrics, the output clustering selection is far from returning the optimal solution

Session variability compensation

- Significant performance improvements for the system output
- Subtle improvement for the optimum output
- Estimation of k as a function of a proportion p of the eigenvalue mass is effective, but not optimal
- p is still very dependent on the input audio file
- System output is still far from the error floor

Show ID	Baseline DER	NAP p	NAP DER
BFMTV_BFMStory_1	8.3	25	5.93
BFMTV_BFMStory_2	16.28	25	15.74
BFMTV_BFMStory_3	5.75	40	4.6
BFMTV_BFMStory_4	4.35	40	4.21
BFMTV_CulturelEtVous_1	26.54	10	22.55
BFMTV_CulturelEtVous_2	29.61	25	30.65
BFMTV_CulturelEtVous_3	20.7	25	17.97
BFMTV_CulturelEtVous_4	3.88	35	5.74
BFMTV_CulturelEtVous_5	11.34	40	10.01
BFMTV_CulturelEtVous_6	18.65	25	18.37
BFMTV_CulturelEtVous_7	24.8	5	25.26
LCP_CaVousRegarde_1	8.02	25	8.02
LCP_CaVousRegarde_2	9.11	25	9.1
LCP_CaVousRegarde_3	20.79	30	15.04
LCP_EntreLesLignes_1	34.23	15	31.97
LCP_EntreLesLignes_2	11.49	20	9.29
LCP_EntreLesLignes_3	6.92	10	5.49
LCP_LCPInfo13h30_1	8.54	25	6.91
LCP_LCPInfo13h30_2	2.8	20	0.57
LCP_LCPInfo13h30_3	11.98	35	8.96
LCP_PileEtFace_1	16.29	20	15.89
LCP_PileEtFace_2	15.35	10	13.31
LCP_PileEtFace_3	26.51	20	32.88
LCP_PileEtFace_4	14.23	35	10.64
LCP_PileEtFace_5	7.97	10	5.26
LCP_TopQuestions_1	3.83	30	2.87
LCP_TopQuestions_2	1.54	15	0.57
LCP_TopQuestions_3	3.48	30	2.76
Overall	10.57	-	**9.27**

Conclusions

- The **use of cumulative vectors as speakers models**, together with the **proposed similarity measures,** are beneficial for the task of speaker diarization, outperforming the binary key baseline diarization system
- **Nuisance Attribute Projection** on the cumulative vector space provides slight performance gains through the proposed automatic method for estimating k
- However, **k is very dependent on the input audio file** and a better estimate for k would yield better performance
- The dependence on the local KBM introduces a **great negative impact on efficiency** since the development utterances for estimating W must be projected to the local KBM for each input file (baseline 0.07 xRT versus new 0.5 xRT).

Future work:
- Global KBM for processing all the input files will allow to estimate W only once and reuse it for all tests
- Improve output clustering selection (addressed in Interspeech'15 paper)

Acknowledgments

This work is part of the project "Linguistic and sensorial accessibility: technologies for voiceover and audio description", funded by the Spanish Ministerio de Economía y Competitividad (FFI2012-31024). This work was partially done within the French Research program ANR Project PERCOL (ANR 2010-CORD-102). This article is supported by the Catalan Government Grant Agency Ref. 2014SGR027.