**Improved Binary Key Speaker Diarization System**

Héctor Delgado¹, Xavier Anguera², Corinne Fredouille³, Javier Serrano³

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

**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
- Speaker 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

**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

**Similarity measures for BK/CV**

- Similarity measures
- Geometric distances for BKs
- Cosine distances for CVs

**Similarity measures**

- 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

**Discussion**

- **Significant performance improvements** for the system output
- **Subtle improvement** for the optimum output
- Estimation of $\lambda$ 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

**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 (FP2012-31024). 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.