Towards a complete Binary Key System for the Speaker Diarization Task

Héctor Delgado¹, Corinne Fredouille², Javier Serrano¹
¹CAIAC, Autonomous University of Barcelona, Barcelona, Spain / ²University of Avignon, CERI/LIA, France
hector.delgado@uab.cat, corinne.fredouille@univ-avignon.fr, javier.serrano@uab.cat

INTERSPEECH, 14-18 September 2014, Singapore

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

State of the art. Very competitive performance:
- 7-10% DER for broadcast news (NIST RT04 evaluation)
- 12-14% DER for meetings (NIST RT09 evaluation)

Challenges. Generally, systems are quite slow
- Computational expensive algorithms run iteratively.
- Real time factors (RTF) above 1
- Too long processing times for real-life, time-critical applications

Alternatives:
- Speed up by algorithm tuning with little speed improvement
- Use of parallel hardware (e.g. GPU). Great speed gains, but there is dependence on complex, nonstandard hardware and programming methodologies

One more alternative. Recently, a speaker diarization method based on the novel Binary Key speaker modeling was presented
- The technique runs around 10 times faster than real time with little increase of DER
- 27% DER with 0.103 XRT with meeting audio data
- Results considered still preliminary

Goals

- To evolve the speaker diarization based on binary keys in order to reach state-of-the-art performance while preserving high speed
- To further investigate the technique towards a complete binary key system for speaker diarization, including also Speech Activity Detection (SAD) within the binary domain
- To check the feasibility of the technique under different audio contexts – TV broadcast audio

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 (FFI2012-31023). This work was partially done within the French Research program ANR Project PERCOL (ANR 2010-CORD-103). This article is supported by the Catalan Government Grant Agency Ref. 2009SGR700.

Binary key speaker diarization system

Experimental results

Binary key SAD

Training data:
- RT04 dataset for meeting audio
- REPERE development dataset for TV audio

Audio classes:
- Speech / nonspeech for meeting audio
- Speech / music / speech+music / telephone for TV audio

Data assignment using 0.3s segments

Conclusions

- The proposed binary key SAD method has shown potential for this audio classification problem
- System performance when processing TV audio data remains quite similar to performance with meeting audio, even with no system adaptation to the new domain
- The final clustering selection algorithm fails when finding the optimum clustering

Future work:
- Revise the final clustering selection algorithm so that the system can be applied to real cases
- Improve precision of segment boundaries in the binary key SAD system
- It is thought that an in-depth analysis of system parameters (KBM training, segment assignment, binary key computation) can result in a further improvement of system performance.