MULTISTREAM DIARIZATION FUSION USING THE MINIMUM VARIANCE BAYESIAN INFORMATION CRITERION

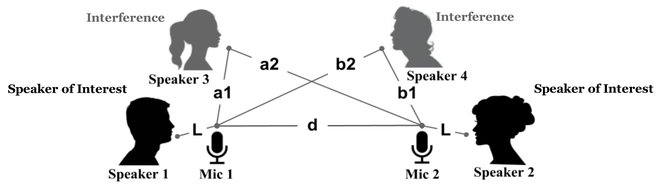
Tae Jin Park, Panayiotis Georgiou
taejinpa@usc.edu, georgiou@sipi.usc.edu

Introduction

- ***Motivation:***
  - Different features for diarization are reliable under different acoustic conditions.
- ***Goal:***
  - Improve the performance of diarization system by fusing two diarization streams in the clustering phase
- ***Approach:***
  - Propose Minimum Variance Bayesian Information Criterion (MVBIC) to estimate the proper ratio to fuse two diarization streams

Open Source USCDiarLibri

- ***USCDiarLibri2,4 Dataset:***



  - The speech data comes from the LibriSpeech ASR corpus
  - Assumes 2 speakers of interest among 4 speakers
  - Models reverberation, overlap, and interfering sources:

$$x_{\text{mic}}[t] = \frac{1}{r}\,\bar{H}[t] * x_{\text{source}}\left[t - r\frac{f_s}{v_s}\right]$$

- **Overlap :** Randomly chosen
- **Interfering sources:** Active at random intervals with overlap
- **Time delay:** Calculated according to the position of each speaker
- **Reverberation:** Impulse Response (IR) is simulated according to the distance to model acoustic degradation in real life

Single Stream Speaker Diarization

- ***Features:***
  - Diarization stream 1: Root Mean Square Energy (RMSE)
  - Diarization stream 2: 13 Mel Frequency Cepstral Coefficients (MFCC)
- ***Segmentation:***
  - KL-distance based speaker change detection
- ***Clustering:***
  - Distance Measure: Bayesian Information Criterion (BIC)
  - Clustering algorithm: Agglomerative hierarchical clustering
- ***Evalutation:***
  - Diarization Error Rate (DER): speaker error, false alarm speech, missed speech and overlap speaker

Proposed Diarization Fusion Framework

- ***Two BIC streams:***
  - N_S : Number of segments in a session
  - $N_S\mathbf{C}2$ (Choose 2 segments out of N_S segments) BIC distances from both RMSE and MFCC
  - An $N_S\mathbf{C}2$ by 2 matrix is used to calculate weight based on MVBIC
- ***MVBIC:***
  - Optimizes the weighted sum of BIC distances in the minimum variance sense
  - Efficiently weights BIC distances according to their reliability
- ***Noisy BIC model:*** Correct BIC stream through a noisy channel
  - The hidden, correct BIC streams: b
  - Two observed, noisy BIC streams: $\tilde{b}_i$

$$\tilde{b}_i = b + n_i \tag{1}$$

- ***Estimated BIC value:*** Obtain the optimal fusion weights

$$\hat{b} = \sum_{i=1}^{M} \omega_i\, b_i = \mathbf{w}^T\mathbf{b}$$
$$Var\left[\hat{b}\right] = \mathbf{w}^T\Sigma_b\mathbf{w} \tag{2}$$

- ***Assumption:*** Mean zero and uncorrelated noise random variables

$$\sigma_{b,i}^2 = \sigma^2 + \sigma_{n,i}^2$$
$$\sigma_{b,ij} = \sigma_{b,ji} = \sigma^2$$
$$\text{where } i \neq j \text{ and } i, j \in [1, M] \tag{3}$$

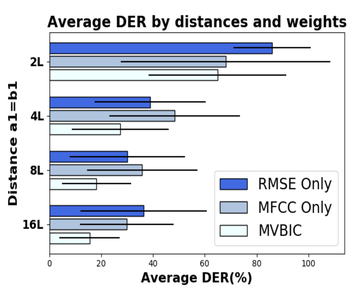
- ***Minimize Noise:*** Set the sum of weights to 1 to keep the signal variance intact and only minimize the variance of noise

$$Var[\hat{b}] \;=\; \left(\sum_{i=1}^{M}\omega_i\right)^2\sigma^2 + \sum_{i=1}^{M}\omega_i^2\,\sigma_{n,i}^2 \;=\; \sigma^2 + \sum_{i=1}^{M}\omega_i^2\,\sigma_{n,i}^2 \tag{4}$$

- ***Minimization problem:***

$$\begin{aligned}&\text{Minimize:} && Var\left[\hat{b}\right] = \mathbf{w}^T\Sigma_b\mathbf{w} \qquad \widehat{\mathbf{w}} = \frac{\Sigma_b^{-1}\mathbf{1}}{\mathbf{1}^T\Sigma_b^{-1}\mathbf{1}}\\ &\text{Subject to:} && \mathbf{w}^T\mathbf{1} = 1\end{aligned} \tag{5}$$

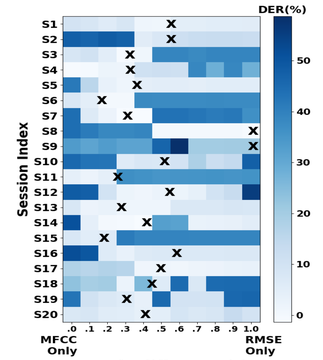
Experimental Results: USCDiarLibri2,4



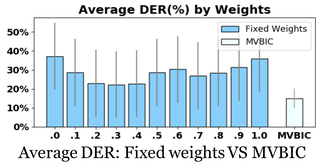
Average DER by distances of interfering speakers

- ***Objective of the Experiment:***
  - Test the effect of the distances from interfering speakers to microphones
- ***Experiment Setup:***
  - Fix distance between primary speakers to 5L
  - Vary distances a1 and b1, keeping a1 = b1
- ***Results:***
  - MVBIC gives lower DER than both of single stream diarization systems.
  - Both features perform worse when interfering speakers are near the primary speakers.

Experimental Results: USCDiarLibri2,4



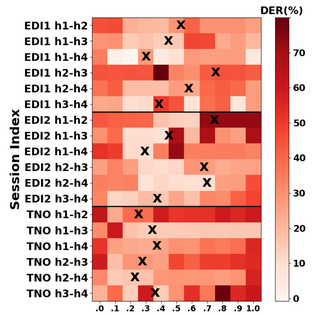
DER for different weights
X Marks MVBIC optimal weight



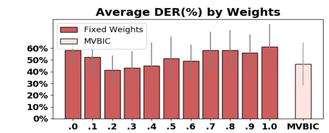
Average DER: Fixed weights VS MVBIC

- ***Objective of the Experiment:***
  - MVBIC vs. fixed BIC weights
  - Evaluated on USCDiarLibri2,4
- ***Experiment Setup:***
  - Random distances [2L, 20L]
- ***Improvement with MVBIC method:***
  - lower DER for each session
  - lower overall DER over fixed BIC weight

Experimental Results: RT06



Estimated weights (X) and DER for RT06S.



Average DER by fixed weights and MVBIC for subset of RT06S dataset.

- ***Objective of the Experiment:***
  - Same test on real-life data (RT06S)
- ***Experiment Setup:***
  - Speech recorded with head-set microphones
- ***Improvement with MVBIC method:***
  - Real-life data shows more irregularity
  - MVBIC method :46.5% fixed BIC method: 41.5%.

Conclusions

- ***Contributions:***
  - Open source data generation framework that models various acoustic conditions
  - Propose MVBIC method to combine two diarization streams without using dev-set data
- ***Further work:***
  - Multiple streams; including i-vector, DNN embeddings

Acknowledgments

http://scuba.usc.edu

This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs through the Psychological Health and Traumatic Brain Injury Research Program under Award No. W81XWH-15-1-0632.