Beamformit and NMF for Speech Enhancement

V. Rajbabu
rajbabu

Department of Electrical Engineering
Indian Institute of Technology Bombay

01 Feb 2016

Beamformit - Open-source Beamforming Tool

Origin - beamforming for speaker diarization of meetings

- Xavier Anguera, Chuck Wooters and Javier Hernando, *Acoustic beamforming for speaker diarization of meetings*, IEEE Trans. Audio, Speech, and Lang. Proc., Sep. 2007, vol. 15, no. 7, pp. 2011-2023.
- Xavier Anguera, *Robust Speaker Diarization for Meetings*, PhD Thesis, UPC Barcelona, 2006

Objective

Acoustic beamforming and front-end processing for speaker diarization in meeting room domain using multiple distant microphones (MDM)

Tool can be obtained from `github`:
`https://github.com/xanguera/BeamformIt`

Beamformit - Block Diagram

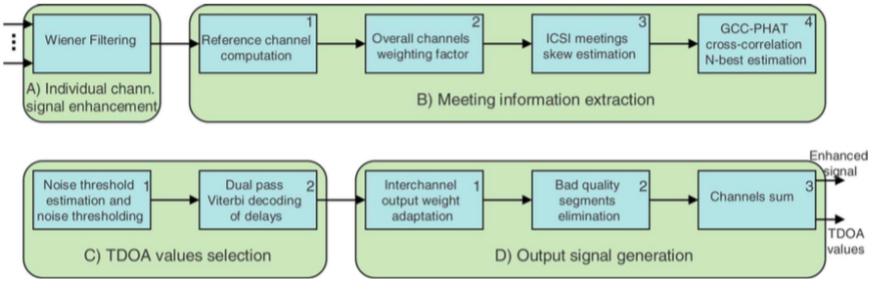

Image Taken from [Xaviera 2007]

1

1

- Xaviera 2007 Xavier Anguera, Chuck Wooters and Javier Hernando, *Acoustic beamforming for speaker diarization of meetings*, IEEE Trans. Audio, Speech, and Lang. Proc., Sep. 2007, vol. 15, no. 7, pp. 2011-2023.

System Setup

- Multiple microphones
  - unknown number, location
  - nonuniform settings
- Multiple speakers
  - unknown number, location
- Interference (natural room noise)
- Data: Rich Transcription evaluations (RT06, RT05, etc.)

System Implementation

Essential component is weighted delay-and-sum, where output $y[n]$ is obtained as

$$y[n] = \sum_{m=1}^{M} W_m[n] x_m \left[n - TDOA^{(m,ref)}[n] \right] \quad (1)$$

$W_m[n]$ - relative weight for mic m, with $\sum_{m=1}^{M} W_m[n] = 1$

$x_m[n]$ - signal for each channel

$TDOA^{(m,ref)}[n]$ - time-delay of arrival between channel m and reference channel *ref*

Individual Channel Signal Enhancement

- Additive noise removal using Wiener filtering
- Performed on individual channels
- Speech or non-speech and noise power estimation

- Does not use multichannel information - but could use

Meeting Information Extraction

Performed in three steps

- Reference channel computation/estimation
- Channel weighting factor computation
- N-best TDOA estimation using GCC-PHAT

Reference Channel Estimation

- Typical reference channel refers to the 'best quality' channel
- Can be obtained if the room layout and microphone array types/configuration are known

- Uses time-average of cross-correlations to obtain the reference channel

$$\overline{xcorr_i} = \frac{1}{K(M-1)} \sum_{k=1}^{K} \sum_{j=1, j \neq i}^{M} xcorr[i, j; k]$$

M - number of channels, $K = 200$ number of 1sec. blocks

- Channel i with highest $\overline{xcorr_i}$ is the reference channel

Channel Weighting Factor

- Uses a channel weighting factor to normalize the input signals
- Weighting factor obtained using a windowed maximum averaging
  - every window has some speech signal

N-best Delays (TDOA) Estimation

- TDOA estimated using generalized cross-correlation phase transform (GCC-PHAT)
- Classical correlation (GCC) between two signal $x_i[n]$ and $x_{ref}[n]$

$$R_{xcorr}^{i,ref}(d) = \sum_{n=0}^{N} x_i[n] x_{ref}[n+d]$$

- GCC-PHAT, robust to noise and reverberation, using $X_i(f)$ and $X_{ref}(f)$ in the frequency domain

$$R_{PHAT}^{i,ref}(d) = \mathcal{F}^{-1}\left(\frac{X_i(f)X_{ref}^*(f)}{|X_i(f)X_{ref}^*(f)|}\right)$$

- TDOA between i-th microphone and *ref* microphone is

$$TDOA^i = \underset{d}{\operatorname{argmax}}\left(R_{PHAT}^{i,ref}(d)\right)$$

N-best Delays (TDOA) Estimation

- TDOA for the current segment need not point at the correct speaker
  - Spurious noises or events while a speaker is active
  - Multiple speakers speaking simultaneously
  - Nonspeech acoustic data or noise
- For each analysis segment, a N-length TDOA vector is maintained along with the GCC-PHAT values

$$\left\{ TDOA_n^i, \quad GCC - PHAT_n^i \right\}$$

for mic i with $m = 1, \ldots, M$, $i \neq ref$, and $n = 1, \ldots, N$.

- N-best TDOAs for all channels, for the entire meeting is obtained at this point

TDOA Selection and Post-processing

- Noisy TDOA thresholding using a continuity filter on the TDOA values for segment *c* using the GCC-PHAT values

$$TDOA_n^i[c] = \begin{cases} TDOA_n^i[c-1], & \text{if } GCC - PHAT_1^i[c] < \theta_{noise} \\ TDOA_n^i[c], & \text{if } GCC - PHAT_1^i[c] \geq \theta_{noise} \end{cases}$$

- Dual-step Viterbi post-processing
  - to maximize speaker continuity by using appropriate TDOA
  - First step: to choose the two-best delays for a single channel
  - Second step: consider all two-best delays across all-channels, to select consistent (best) TDOA value

Output Signal Generation

- Weighted delay-and-sum to obtain the enhanced signal
- Weight for channel m at segment c is computed as

$$\mathcal{W}_m[c] = \begin{cases} \frac{1}{M}, & c = 0 \\ (1-\alpha)\mathcal{W}_m[c-1] + \alpha\overline{xcorr}_m[c], & \text{otherwise.} \end{cases}$$

- Windowing to smooth discontinuities in the signal at segment boundaries

- This enhanced signal (along with TDOA values) can be used in speech processing systems or speaker diarization system

Experiments

- Database: NIST RT evaluations (2004 - 2006)
  - Meeting excerpts (multiple participants interact around a meeting table)
  - Various number of channels, layouts, types of microphones
- Evaluation metrics: NIST diarization error rate (DER) and word error rate (WER)
- Typical acoustic segments were 250 ms duration (i.e., TDOAs every 250 ms were estimated)

Results: Speaker Diarization

Speaker diarization using enhanced acoustic signal (without using TDOAs)

System	**DER** (%)
RT06 Baseline	17.15
Hand-picked ref. ch	17.09
No noise thresh.	18.31
No TDOA-Viterbi	17.16
No TDOA post-proc.	18.72
No adaptive wts.	17.48
No ch. elim.	17.14

- Baseline: uses the complete proposed beamforming system
- Minor differences between development and evaluation datasets

Results: Speaker Diarization

Speaker diarization using enhanced acoustic signal and TDOAs

DER for development set

System	**DER** (%)
SDM	24.88
MDM + TDOA	14.64
MDM (no TDOA)	19.04

DER for evaluation set

System	**DER** (%)
SDM	19.80
MDM + TDOA	14.76

Advantage in using TDOA

- information about speaker position is useful
- depends on the acoustic channels used

Results: Speech Recognition

Dataset: RT05s, RT06s datasets

WER for RT05s

System	WER (%)
SDM	47.7
MDM	45.8

WER for RT06s

System	WER (%)
SDM	57.3
MDM	55.5

Using Beamformit Tool

- Have attempted using Beamformit
- Works fine as is
- Strength: No knowledge of arrays, channels, number of speakers
- Results from Beamformit match results from our implementation of beamforming algorithms (atleast interms of DOA, source localization)

- No DERs or WERs computed on results from these algorithms

[These results in a separate set of slides]

Summary

Beamformit can be used as an initial tool to get started with speaker diarization

- Handles source separation implicitly
- Can be used as baseline system

Moving forward

- Need to integrate Beamformit with a ASR/Diarization system
- TDOA outputs from Beamformit to be used in the tools used for DER
- Look at exploiting prior knowledge (room, microphone config.)
- Other beamforming algorithms (both noise and reverberation)
- Source separation algorithms will help in overlapping regions