Echo-aware signal processing for audio scene analysis

Diego Di Carlo

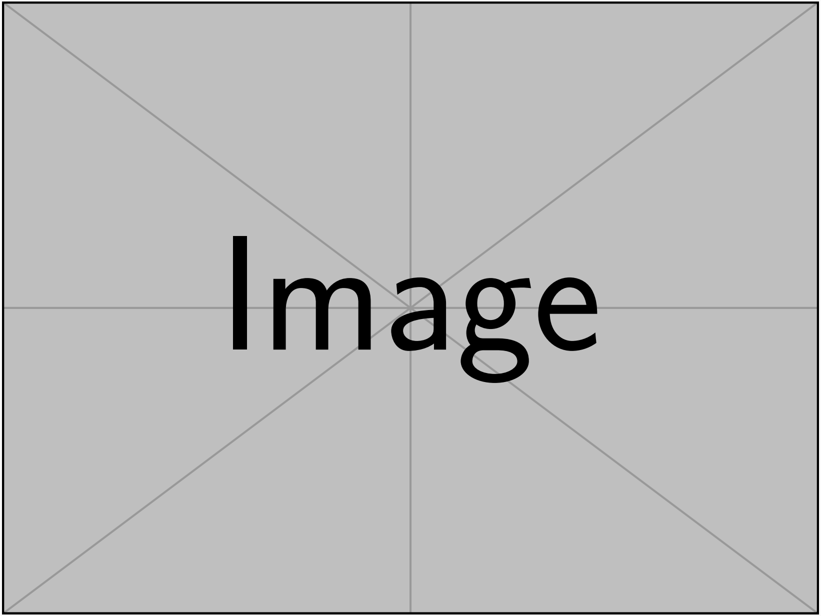
November 17, 2020

suprevisors: Antione DELEFORGE, Nancy BERTIN
collaborators: Clément Elvira, Robin Scheibler, Ivan Dokmanić, Sharon Gannot, Pini A

INRIA IRISA

Introduction

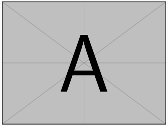
Scenario

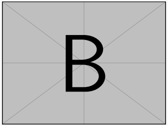


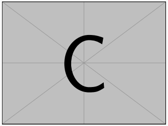
Echo-aware signal processing for audio scene analysis

Sound recorded by microphones carries information:

- Semantic information about source nature and semantic content
- Spatial information about due to *sound propagation*
- Temporal information about event

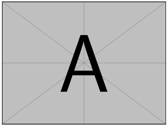


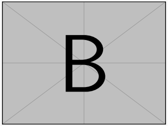


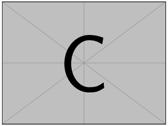


Audio Scene Analysis

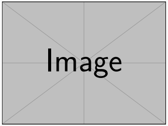
is the extraction and organization of all the information in the sound







→



Typical problems

- What?
  - *Sound Source Separation*
  - *Speech Enhancement* (denoising, dereverberation)
  - *Automatic Speech Recognition*
  - ...
- Where?
  - *Sound Source Localization* (DOA estimation, Mic calibration)
  - *Room Geometry Estimation*
- When?
  - *Speaker Diarization*
  - *Text/Lyrics alignment*
- How?
  - *Acoustic Channel Estimation*
  - *Acoustic Measurements*

Also known as auditory scene analysis or computer auditory scene analysis.

Inverse and Forward problems

Blind and Informed problems

Everything is connected

Signal Processing

Offer mathematical models, frameworks and tools to tackle such ASA problems

General Pipeline

- (Mathematical Models)
- Signal representation (STFT, Features)
- Enhancement (denoising, dereverberation)
- Parameter Estimation (DOA, Localization)
- Adaptive Processing (Filtering)

Acoustic Echoes

- Product of the sound propagation
- Sound repetition
  - “same” content: can be integrated
  - “different” sounds: carry info about the reflection
  - different direction of arrival: spatial information

Echo-aware processing
between anechoic processing and reverberant processing

Turning echoes into friends
Typically reverberation is considered as ”foe” for the processing.

Thesis objective

1. provide new methodologies and data to process and estimate acoustic echoes
2. extend previous classical methods for audio scene analysis

Echo-aware signal processing
for audio scene analysis

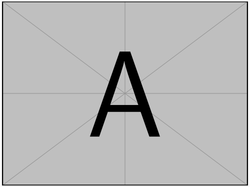
Modeling

Echoes and Room Acoustics

Sound propagates and interacts with space

- it travels with a certain speed and it is attenuated;
- it is absorbed and reflected by surfaces;
- and it is scattered, diffracted, etc.

This is describe by the so called RIRs



Elements of reverberation

- Direct path
- Early Echoes
- Reverberation tails

RIRs can be modeled with the Image Methods

- specular reflection only
- “playing billiard in a concert hall”
- for shoebox room it is is the solution for physics
- in frequency domain it writes as

RIRs accounts for
the geometry of the room

- Room shape and size
- Mic and Source position
- presence of objects

the acoustic properties of the audio scene

- surface materials
- objects materials

examples

Room Impulse Response

$$\tilde{x}_i = (\tilde{h}_i * \tilde{s})(t) \longrightarrow \tilde{X}_i(f) = \tilde{H}_{ij}(f)\tilde{S}(f)$$

the linear filtering effect due to the propagation of sound from a source to a microphone in a indoor space

Observation

Our vision is limited both in time (finite and discrete) and in frequency (finite and discrete)

$$x_i[n] = ... \tag{1}$$

Signal model in the frequency domain

$$x_i = (h_i * s)(t) \ \longrightarrow \ X(f) = H_i(f)S(f)$$

Approximations

- Narrowband Approximation
- DTFT echo model in the DFT

Interim Conclusion I

Approximations

- Echoes are well described by specular reflection
- Echoes are off-grid by nature
- Sampling and quantization make them hard
- Processing in the discrete frequency domain, but with continuous time echo model

Acoustic Echo Estimation

Acoustic Echo Retrieval

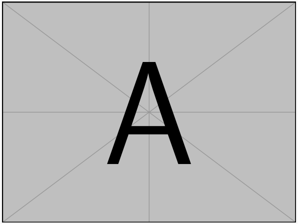
Given the echo model

$$H_{ij}(f) = \sum_{r=0}^{R} \alpha e^{2\pi},$$

The acoustic echoes retrieval (AER) problem

Estimating early (strong) acoustic reflections:

- their time of arrivals $\rightarrow$ TOAs Estimation
- their amplitude
  $\hookrightarrow$ closed-from knowing τ [?]



Note that an order of r

Taxonomy of Acoustic Echo Estimation

- based on the emitted signal knowledge:

Active approaches

- Signal is emitted and known
- Intrusive
- Single channel
- Methods: Least-Square estimation, Inverse Filtering (Equalization)
- Application:
  measurements, calibration, sonars, slam

Passive approaches

- Emitted signal is not known
- Not intrusive (for passive listening)
- Multichannel
- Methods:
  blind deconvolution problem
  ill-posed and *ill*-conditioned
  $\hookrightarrow$ statistics, *sparsity* etc
- Application:
  Robot hearing (Table Top Scenario),
  Pre-processing step

Taxonomy of Acoustic Echo Estimation

- based on the estimated filter:

RIR-based approaches

1. RIRs are first estimated as SIMO BCE problem
2. Echoes extracted from first part of the RIRs with peak picking and disambiguation

Pros

- SIMO BCE is well studied (elegant framework)
- It works well is some scenarios and in practice
  $\hookrightarrow$ if not limitation

Cons

- Full RIR
- dependent of manually tuned peak picking
- Pathological issue (sampling and body-guard
- Complexity

RIRs-agnostic approaches

1. Estimation directly in the echoes parameters space $\{\tau, \alpha\}$
   and direction of arrivals can be used instead

Performed with

- Cross-correlation
  on-grid, eg. EM, Acoustic Cameras
- Cross-relation with super-resolution
  off-grid, [?, ?]

Pro

- No need for full RIRs
- Sub-sampling accuracy
- Low complexity
- Sparsity and Non-negativity are respected

Cons

- Exploratory

AER as discrete SIMO BCE

Key ingredient – *Cross relation identity*

$$x_i = h_i * s$$

$$h_2 * x_1 = h_2 * h_1 * s = h_1 * h_2 * s = h_1 * x_2$$

Ideas

1. Sampled version of x_1, x_2 are available $(\mathbf{x}_1, \mathbf{x_2})$
2. Assume echoes belong to multiples of the sampling frequency
3. Identify echoes $\rightarrow$ find sparse vectors $\mathbf{h}_1, \mathbf{h}_2$
4. Lasso-like problem

$$\widehat{\mathbf{h}}_1, \widehat{\mathbf{h}}_2 \in \underset{\mathbf{h}_1, \mathbf{h}_2 \in \mathbf{R}^n}{\arg\min} \; \|\mathbf{x}_1 * \mathbf{h}_2 - \mathbf{x}_2 * \mathbf{h}_1\|_2^2 + \lambda \mathrm{Reg}(\mathbf{h}_1, \mathbf{h}_2)$$

$\mathrm{Reg}(\mathbf{h}_1, \mathbf{h}_2) \longrightarrow$ sparse promoting regularizer

5. Pick picking

✓ [?] ✓ [?]

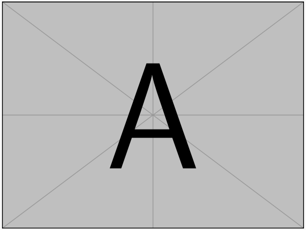
✓ [?] ✓ [?]

Limitations / bottleneck

Limitations

- Echoes are not necessarily “on grid”
- *Body guard* effect [?]

  $\longrightarrow$ low recall $\implies$ low accuracy
  $\longrightarrow$ slow convergence



Increase the sampling frequency, F_s

$\longrightarrow$ Increase Precision

Computational bottleneck

- Bigger vectors and matrices

  $\longrightarrow$ memory usage

- Computational complexity: at best $\mathcal{O}(F_s^2)$ per iteration
- the higher the sampling frequency, the more ill-conditioned

  $\longrightarrow$ slow convergence

Observation 1: the cross relation remains true in the frequency domain

$$\mathcal{F}x_1 \cdot \mathcal{F}h_2(n/F_s) = \mathcal{F}x_2 \cdot \mathcal{F}h_1(n/F_s) \qquad n = 0 \ldots N-1$$

Observation 2: $\mathcal{F}\delta_{\text{echo}}$ is known in closed-form

Observation 3: $\mathcal{F}\mathrm{x_i}$ can be (well) approximated by DFT

$$\mathbf{X}_i = \mathsf{DFT}(\mathbf{x}_i) \simeq \mathcal{F}\mathbf{x}_i(nF_s) \qquad n = 0 \ldots N-1$$

Idea: Recover echoes by matching a finite number of frequencies

$$\underset{h_1, h_2 \in \text{measure space}}{\arg\min} \ \tfrac{1}{2}\|\mathbf{X}_1 \cdot \mathcal{F}h_2(f) - \mathbf{X}_2 \cdot \mathcal{F}h_1(f)\|_2^2 + \lambda \|h_1 + h_2\|_{\text{TV}} \quad \text{s.t.} \begin{cases} h_1(\{0\}) = 1 \\ h_l \geq 0 \end{cases}$$

Instance of a **BLasso** problem [?] (Sliding Frank-Wolfe algorithm)

no Toeplitz matrix

Solutions is a train of Dirac

anchor prevents trivial solution

Experiments

- simulation data with ISM with Pyroomacoustics
- 1 source, 2 microphones, random room geometry
- Full RIRs
- 2 sources: broadband and speech
- 2 datasets: different SNR, different RT60

Methods

- BSN: Blind Sparse and Nonnegative SIMO BCE [?]
- IL1C: Iteratively-weighted ℓ_1 Constraint SIME BCE [?]
- Blaster: Proposed off-grid approach

Metrics

- RMSE
- Precision

`Blaster`- Results

Lantern- data-driven AER

Observation 1: Mapping from observation to echo is extremely difficult
Later echoes are not considered, may help

Observation 2: We have acoustic simulators
Acoustic simulators based on ISM

source position, room $\leftarrow$ reverberation elements $\leftarrow$

annotation for free

Observation 3: (Deep) Learning-based methods successful for localization
Echoes are strongly related to the source position

Idea: Use Deep Learning for AER

- Extend previous work on source localization for Echo Estimation
- Estimate the first echo TOA
  $\hookrightarrow$ simple case, but with important application in SSL

Data

- train:
  - ↪ artificially generated RIR
  - ↪ white noise + noise
  - ↪ instantaneous RTF
- test:
  - ↪ artificially generated RIR
  - ↪ white noise, speech + noise
  - ↪ instantaneous RTF

Architecture

- models: MLP, CNN
- loss: Multi-class regression problem
  - ↪ RMSE
  - ↪ Gaussian regression + uncertainty
  - ↪ Student Regression + uncertainty

Experiments

1. MLP
2. CNN
3. CNN + Noise
4. CNN + Gaussian
5. CNN + Student

Results

1. MLP
2. CNN
3. CNN + Noise
4. CNN + Gaussian
5. CNN + Student

Interim conclusion (2/4)

on Acoustic Echo Retrieval:

- Most of the literature is on Passive and RIR-based, with on-grid approaches
- On-grid approaches suffers by the off-grid nature of the echoes (complexity, sampling)

on `Blaster`:

✓ off-grid parameter-free which exploit dirac closed-form model (non negativity and sparsity)

✓ smaller RMSE due to super-resolution, better for small # of echoes

✗ source dependent and on number of echoes

✗ validate only on synthetic data

→ Multichannel and RTF-based extention

on `Lantern`:

✓ promising results for first echo estimation

✓ direct application for table top application

Echo-aware Application

Sound propagation is [?]

$$x_i(t) = (h * s)(t)$$
$$h(t) = h^d(t) + h^e(t) + h^r(t)$$
$$H(f) = \sum_{r=0}^{R} \alpha_i^{(r)}(f) \mathrm{e}^{-\mathrm{i}2\pi\tau_i^{(r)} f_k}$$

- completely ignored
  $\hookrightarrow h(t) = 1$
- assumed direct path (*anechoic* case)
  $\hookrightarrow h(t) = h^d(t) + \varepsilon(t)$
- fully modeled (*reverberant* case)
  $\hookrightarrow h(t) = h^d(t) + h^e(t) + h^l(t) + \varepsilon(t)$
- early echoes (*multipath* case)
  $\hookrightarrow h(t) = h^d(t) + h^e(t) + \varepsilon(t)$

$\Leftarrow$ strong early reflection and strong reverberation level

- detrimentally affect typical Audio Scene Analysis algorithm

Echo-aware Application

What: echoes as sound repetition

- Sound Source Separation
- Speech Enhancement
  ↪ Dereverberation, Denoising, Room Equalization
- Speaker Verification

Where: echoes as new sound direction

- Sound Source Localization
- Microphone Calibration
- Room Geometry Reconstruction

How: echoes as element of sound propagation

- Blind Acoustic Channel Estimation
  as initialization for other methods
- Acoustic Measurements

What: echoes as sound repetition

- Sound Source Separation
- Speech Enhancement
  $\hookrightarrow$ Dereverberation, Denoising, Room Equalization
- Speaker Verification

Where: echoes as new sound direction

- Sound Source Localization
- Microphone Calibration
- Room Geometry Reconstruction

How: echoes as element of sound propagation

- Blind Acoustic Channel Estimation
  as initialization for other methods
- Acoustic Measurements

The Picnic Scenario:

- Microphone close to a surface (table-top scenario)
- Clear definition of the echo
- One source

Mirage Array
How to access the *image* microphone

Each pair is augmented with echoes

1D SSL

- Estimate the TDOA between two microphones signals with GCC
- Map the TDOA to angles knowing the array geometry

2D SSL

- For each pair:
  1D-SSL
- Compute a global angular spectrum by “fusing” together the estimation of each pairs

Baseline:
GCC-PHAT on true microphones

Proposed Approach:
Using DNN-based TDOA estimation
problem: real value not estimation

Echo-aware Dataset

Acoustic Echo Estimation

Interim conclusion (3/4)

Annotation

Usage

Conclusion

2D Outline

Thesis outline with projects