Adaptive Learning with Surrogate Assisted Training Models using Limited Labeled Acoustic Sample Sequences

Guilherme Zucatelli Nossa
Rosângela Fernandes Coelho

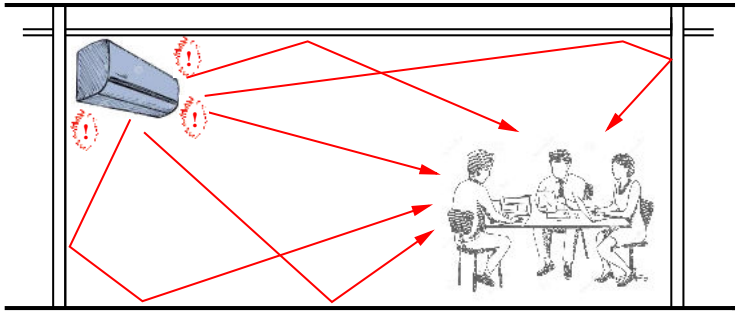
Laboratory of Acoustic Signal Processing
Military Institute of Engineering

2021 IEEE Statistical Signal Processing (SSP) Workshop
11-14 July 2021

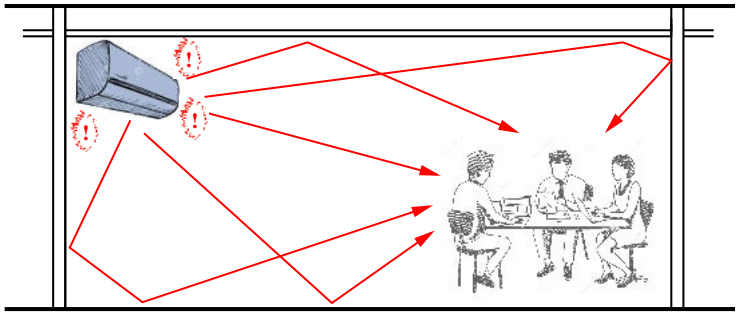
Plan

- Motivation and Challenges
- Objectives
- The Reverberation Effect
- Proposed Adaptive Method: the ALSSmod
- Experimental Results
  - Adaptive Learning: Improving with Surrogates
  - Feature Fusion: pH+MFCC
  - Acoustic Source Classification (MFCC+GMM)
  - ROC and AUC Analysis
  - Separability and Sparse Coding: Bhattacharrya and K-SVD
- Conclusion

Motivation

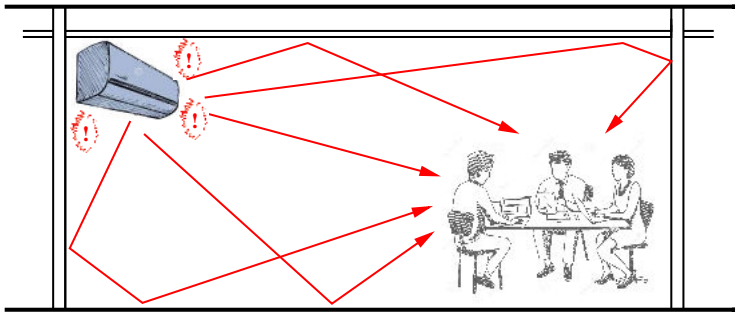


Motivation



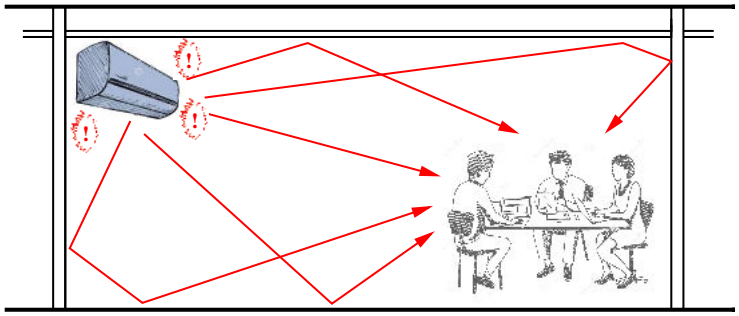
- The Reverberation Effect on Limited Labeled Samples for Acoustic Sources Classification:
  1. Change temporal and spectral characteristics
  2. Modify the nonstationary behavior
  3. May decrease the accuracy of acoustic source classification systems

Motivation



- Applications:
  1. Hearing aid devices
  2. Smart Homes
  3. Robot Navigation
  4. Surveillance Systems

Motivation



- Applications:
  1. Hearing aid devices
  2. Smart Homes
  3. Robot Navigation
  4. Surveillance Systems

- **Challenges:** diversity of sources and environments, multiple temporal and spectral characteristics e non-stationarity of acoustic signals.

Objectives

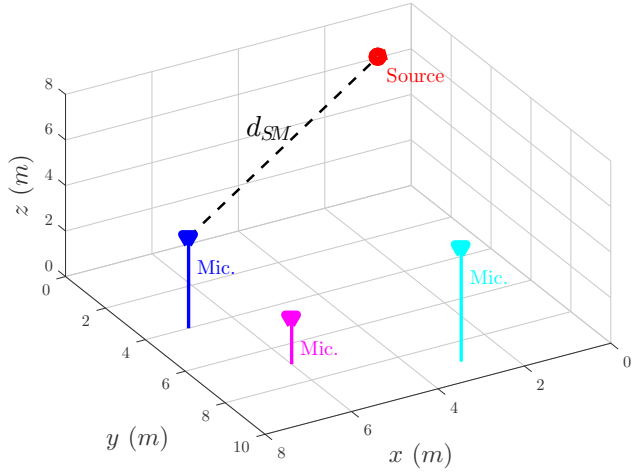
- Increase acoustic source classification accuracy under multiple reverberant environments.
- Attain good representative and discriminative models.
- Improve the limited labeled acoustic samples by the usage of adaptive learning.

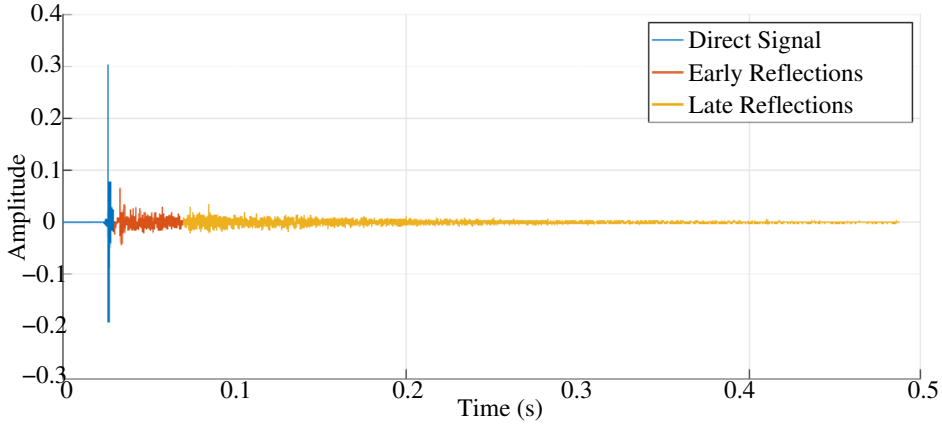
The Reverberation Effect

- Definitions:
  - RIR: Room Impulse Response, $h(t)$ (signals derived by convolution)

$$s(t) = h(t) \circledast x(t),$$

    where $s(t)$ is the reverberated signal and $x(t)$ is the original signal.
  - T_{60}: Time needed for a 60 dB reduction on reverberated signal.
  - DRR: *Direct-to-Reverberant Ratio*
  - d_{SM}: Source-Microphone Distance





T_{60} and Databases

- AIR Database*:

Name	d_{SM} [m]	T_{60} [s]	DRR [dB]	Reverberation
Meeting	1.5	0.36	2.7	Low
Stairway	3.0	1.00	-3.4	Moderate

- LASP_RIR Database**:

Name	d_{SM} [m]	T_{60} [s]	DRR [dB]	Reverberation
LASP_1	1.2	0.65	-3.1	Moderate
LASP_2	1.6	0.79	-4.3	Moderate

Low: $T_{60} < 0.4$s

Moderate: 0.4s $< T_{60} < 2.0$s

* [Jeub, Marco, Magnus Schafer, and Peter Vary. "A binaural room impulse response database for the evaluation of dereverberation algorithms." 2009 16th International Conference on Digital Signal Processing. IEEE, 2009.]

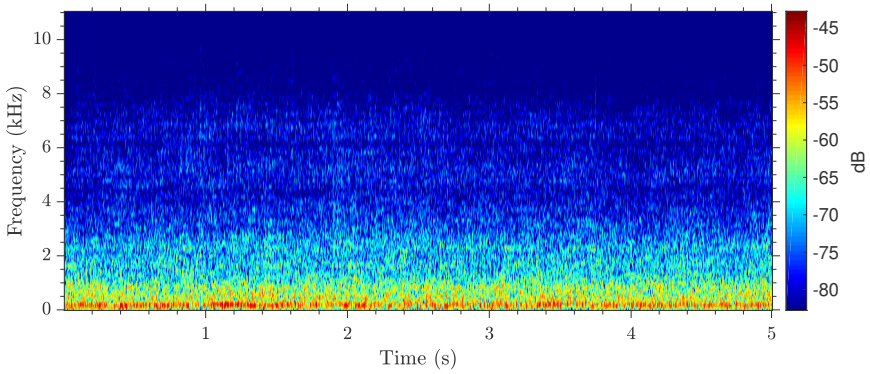
** [Aveilable at lasp.ime.eb.br.]

Reverberation Effect

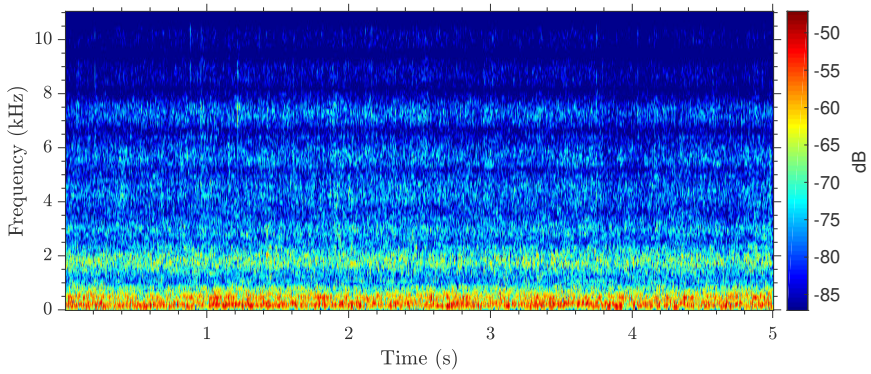
- Waterfall Source and Meeting Reverberation ($T_{60} = 0.36$ s).
- Impact on Acoustic Source Classification (Baseline* 12-MFCC + 4GMM)

Reverberation Effect

- Waterfall Source and Meeting Reverberation ($T_{60} = 0.36$ s).
- Impact on Acoustic Source Classification (Baseline* 12-MFCC + 4GMM)



No Reverb (77.2% Acc.)



Meeting Reverb (5.0% Acc.)

* [Mesaros, Annamaria, et al. "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge." IEEE/ACM Transactions on Audio, Speech, and Language Processing 26.2 (2017): 379-393.]

Reverberation Effect

- Waterfall Source and Meeting Reverberation ($T_{60} = 0.36$ s).
- Impact on Acoustic Source Classification (Baseline* 12-MFCC + 4GMM)

Reverb Free

Actual Source	Classified Source							
	Chainsaw	Dogs	Fan	Rain	Shower	Siren	Subway	Waterfall
Chainsaw	**7.2**	0.0	60.4	0.0	28.4	4.0	0.0	0.0
Dogs	16.8	**68.6**	5.2	0.0	0.4	8.8	0.2	0.0
Fan	1.6	0.0	**28.8**	1.0	0.6	6.2	61.8	0.0
Rain	0.0	0.0	0.6	**81.2**	0.8	0.0	3.6	13.8
Shower	0.0	0.0	0.0	0.0	**99.8**	0.0	0.0	0.2
Siren	0.0	0.0	0.4	0.0	37.8	**61.8**	0.0	0.0
Subway	1.0	0.0	10.2	4.6	0.2	0.0	**84.0**	0.0
Waterfall	0.0	0.0	7.2	8.6	6.6	0.4	0.0	**77.2**

$\Rightarrow$

Meeting room ($T_{60} = 0.36$ s)

Actual Source	Classified Source							
	Chainsaw	Dogs	Fan	Rain	Shower	Siren	Subway	Waterfall
Chainsaw	**26.6**	0.0	47.8	0.0	6.8	18.8	0.0	0.0
Dogs	13.4	**55.6**	10.2	0.0	0.2	20.2	0.4	0.0
Fan	0.0	0.0	**94.8**	0.0	0.0	4.0	1.2	0.0
Rain	0.0	0.0	44.2	**50.2**	0.4	0.0	3.8	1.4
Shower	0.0	0.0	0.0	0.0	**94.6**	4.6	0.0	0.8
Siren	2.0	0.6	27.8	0.0	15.6	**54.0**	0.0	0.0
Subway	1.4	0.0	37.2	0.0	0.0	0.2	**61.2**	0.0
Waterfall	0.0	0.0	79.6	12.6	2.2	0.6	0.0	**5.0**

Waterfall Accuracy: $77.2\% \rightarrow 5.0\%$

Overall System Accuracy: $63.6\% \rightarrow 55.2\%$

* [Mesaros, Annamaria, et al. "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge." IEEE/ACM Transactions on Audio, Speech, and Language Processing 26.2 (2017): 379-393.]

Proposed ALSSmod* **

- ALSSmod - *Modified Adaptive Learning with Surrogate Assistance*
  - Adaptive learning for limited labeled nonstationary acoustic sources.
  - Focused on improving classification accuracy.
  - Original acoustic models replaced by selected Surrogates.
  - Robust to reverberation effect.

Proposed ALSSmod* **

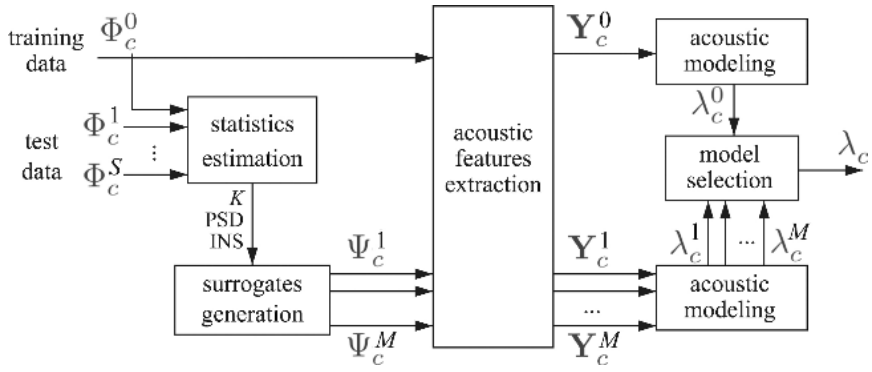
- ALSSmod - *Modified Adaptive Learning with Surrogate Assistance*
  - Adaptive learning for limited labeled nonstationary acoustic sources.
  - Focused on improving classification accuracy.
  - Original acoustic models replaced by selected Surrogates.
  - Robust to reverberation effect.

- Surrogates Generation given a target acoustic signal $x(t)$:
  1. Uncorrelated samples sequence with target Kurtosis K_x.
  2. FIR filtering to obtain PSD decay behavior.
  3. Make short-time adjustment for nonstationary (INS - Index of Nonstationary).

* [G. Zucatelli, R. Coelho and L. Zão, "Adaptive Learning With Surrogate Assisted Training Models for Acoustic Source Classification," in IEEE Sensors Letters, vol. 3, no. 6, pp. 1-4, June 2019]
** [G. Zucatelli and R. Coelho, "Adaptive Learning with Surrogate Assisted Training Models using Limited Labeled Acoustic Sample Sequences", 2021 IEEE Statistical Signal Processing Workshop.]

Proposed ALSSmod* **

- Scheme:



C: Number Classes
M: Number Surrogates
Φ_c: Labeled Data
Ψ_c: Surrogates
Y_c: Feature Matrices
λ_c: Models

- ALSSmod Model Selection $(\Gamma + 1)$:
  Occurs if model λ_c^m increases the average classification rate $(R^{\Gamma+1} > R^{\Gamma})$ and the source accuracy rate $(R_c^m > R_c)$

$$\lambda_c \leftarrow \lambda_c^{\hat{m}}, \quad \text{where } \hat{m} = \max_{1 \leq m \leq M} R^{\Gamma+1}.$$

* [G. Zucatelli, R. Coelho and L. Zão, "Adaptive Learning With Surrogate Assisted Training Models for Acoustic Source Classification," in IEEE Sensors Letters, vol. 3, no. 6, pp. 1-4, June 2019]
** [G. Zucatelli and R. Coelho, "Adaptive Learning with Surrogate Assisted Training Models using Limited Labeled Acoustic Sample Sequences", 2021 IEEE Statistical Signal Processing Workshop.]

Proposed ALSSmod: Surrogate Generation

- Uncorrelated samples with Kurtosis K_x:* **
  - Start sequence of independent random numbers $\{W_m\}$, $0 < W_m \leq 1$.
  - Perform transformation:

$$Y_m = \left[\log \frac{1}{W_{2m-1}}\right]^n \sin(2\pi W_{2m})$$

  - Resulting Kurtosis:

$$K_x = \frac{3}{2}\frac{\Gamma(4n+1)}{[\Gamma(2n+1)]^2},$$

    where Γ is the gamma function and $n \in [0, \infty)$.

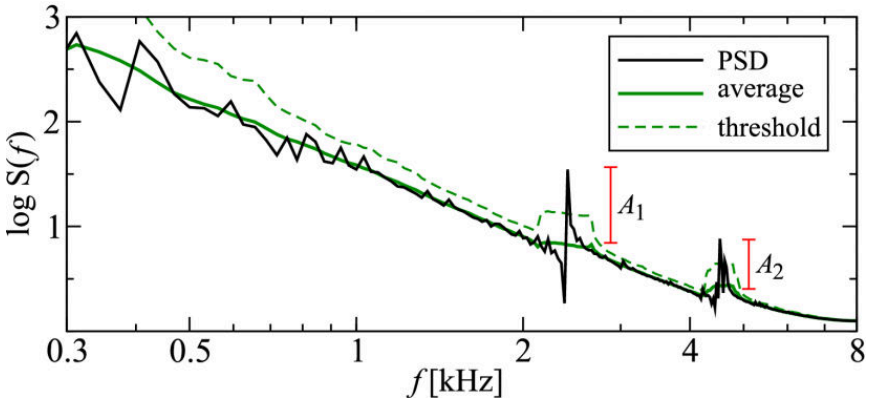
* [R. J. Webster, "A random number generator for ocean noise statistics," in IEEE Journal of Oceanic Engineering, vol. 19, no. 1, pp. 134-137, Jan. 1994]
** [L. Zão and R. Coelho, "Generation of coloured acoustic noise samples with non-Gaussian distributions." IET signal processing 6.7 (2012): 684-688]

Proposed ALSSmod: Surrogate Generation

- Target PSD decay of $\beta/2$ obtained by Al-Alaoui filter rule*:

$$H(z) = \left[\frac{7T}{8}\frac{(1+z^{-1}/7)}{(1-z^{-1})}\right], \text{ where } T \text{ is the sample rate}$$

- PSD peak detection**:



Incorporated as:
$h'(t) = h(t) + \sum_{p=1}^{P} A_p sin(2\pi f_p t)$,
for P peaks at frequency bins
$f1, \ldots, f_P$.

* [Al-Alaoui, Mohamad Adnan. "Novel digital integrator and differentiator." Electronics letters 29.4 (1993): 376-378.]
** [G. Zucatelli, R. Coelho and L. Zão, "Adaptive Learning With Surrogate Assisted Training Models for Acoustic Source Classification," in IEEE Sensors Letters, vol. 3, no. 6, pp. 1-4, June 2019]

Proposed ALSSmod: Surrogate Generation

- Nonstationarity of Surrogate Signals:
  $\Rightarrow$ INS* - Index of Non-Stationarity
  - Objective measure based on temporal-frequency analysis
  - The Kullback-Leibler divergency (KL) determine the distance between the short-time spectrum (T_h) and the global spectrum (T)
  - The INS is the ratio of KLs from the original signal and the corresponding values of stationary references

$$\text{INS} \begin{cases} \leq \gamma: & \text{stationary,} \\ > \gamma: & \text{non-stationary.} \end{cases}$$

- Short-time Amplitude Adjustment:

$$A_p \leftarrow r^2 A_p, \text{ where } r = INS_{\text{Target}}/INS$$

* [Testing Stationarity With Surrogates: A Time-Frequency Approach, P. Borgnat, P. Flandrin, P. Honeine, C. Richard and J. Xiao, *IEEE Transactions on Signal Processing*, vol. 58, no. 7, Jul 2010.]

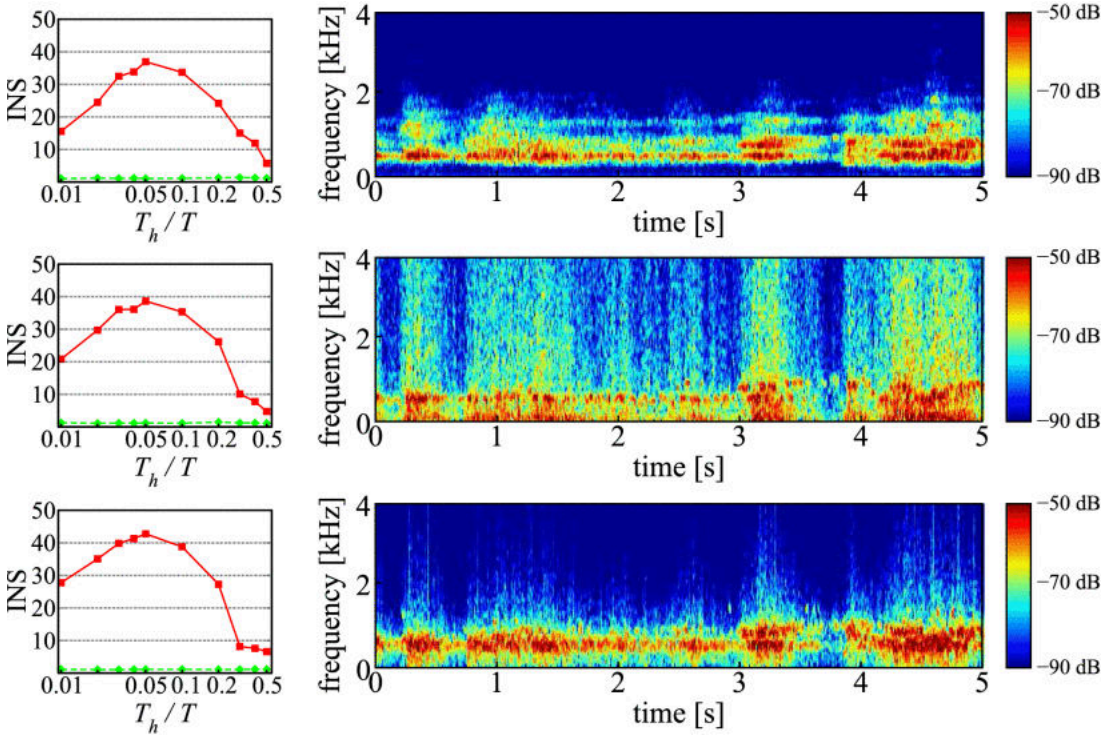
Proposed ALSSmod

- Nonstationarity of Surrogate Signals:

Original Source Dogs

Surrogate Dogs

Selected Surrogate Dogs



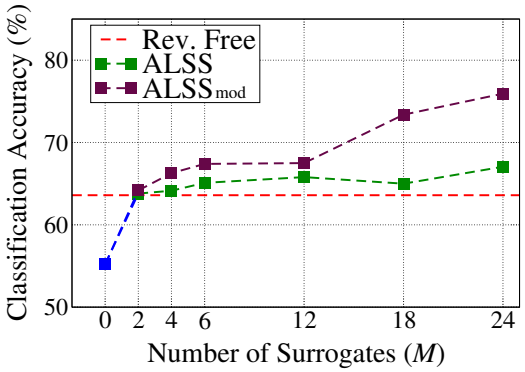
Experimental Setup

- 8 acoustic sources: Chainsaw, Dogs, Fan, Rain, Shower, Siren, Subway and Waterfall.*
- Reverberation database:
  LASP_RIR* $\Rightarrow$ LASP1 ($T_{60} = 0.65$ s) and LASP2 ($T_{60} = 0.79$ s)
  AIR $\Rightarrow$ Meeting ($T_{60} = 0.36$ s) and Stairway ($T_{60} = 1.0$ s)
- Results:
  - Adaptive Learning: Improving with Surrogates
  - Feature Fusion: pH***+MFCC
  - Acoustic Source Classification (12-MFCC + 4-GMM)
  - ROC and AUC Analysis
  - Separability and Sparse Coding: Bhattacharrya and K-SVD**

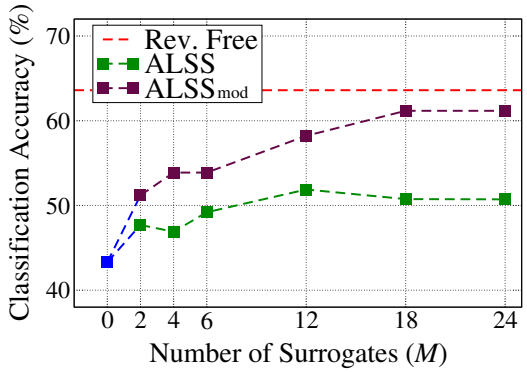
* [Available at lasp.ime.eb.br]
** [Aharon, Michal, Michael Elad, and Alfred Bruckstein. "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation." IEEE Transactions on signal processing 54.11 (2006): 4311-4322.]
*** [Sant'Ana, R., Rosângela Coelho, and Abraham Alcaim. "Text-independent speaker recognition based on the Hurst parameter and the multidimensional fractional Brownian motion model." IEEE Transactions on Audio, Speech, and Language Processing 14.3 (2006): 931-940.]

Adaptive Learning

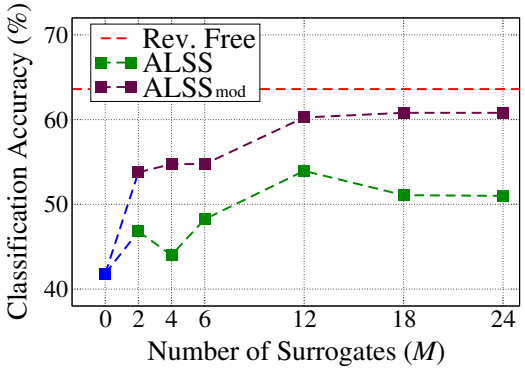
- Improving with Surrogates:



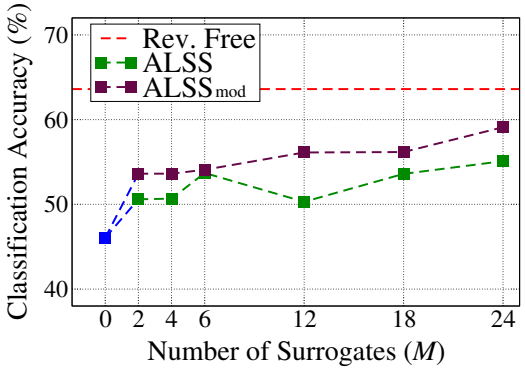
($T_{60} = 0.36$ s)



($T_{60} = 0.65$ s)



($T_{60} = 0.79$ s)

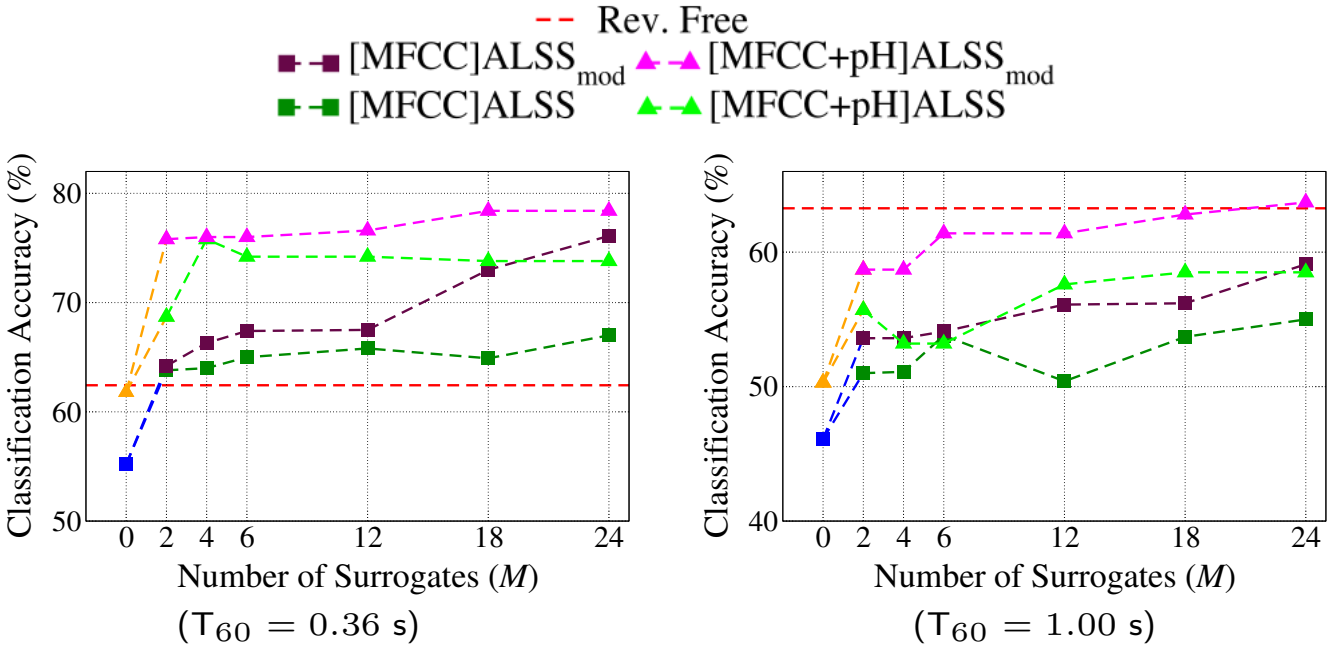


($T_{60} = 1.00$ s)

Best Result: 76.8% accuracy for $T_{60} = 0.36$ s $\rightarrow$ Gain of 12.5 p.p.

Feature Fusion pH+MFCC

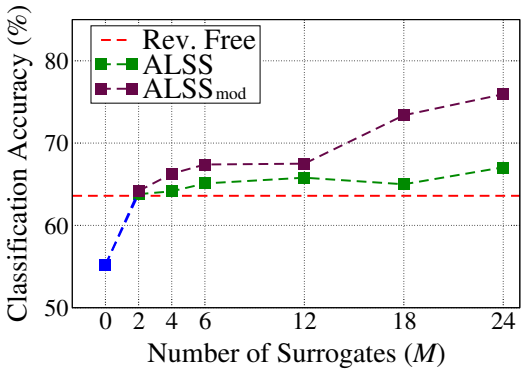
- Adoption of pH feature vectors for Acoustic Source Classification
- Rooms Meeting ($T_{60} = 0.36$ s) and Stairway ($T_{60} = 1.00$ s).
- Comparison of ALSS and ALSSmod for 7-pH+12-MFCC and 4GMM.



Highest Accuracy: pH+MFCC $\rightarrow$ 78.4% (a) and 63.7% (b)

Acoustic Source Classification

- Baseline 12-MFCC+4GMM on Meeting Reverberation ($T_{60} = 0.36$):



Actual Source	Classified Source							
	Chainsaw	Dogs	Fan	Rain	Shower	Siren	Subway	Waterfall
Chainsaw	**26.6**	0.0	47.8	0.0	6.8	18.8	0.0	0.0
Dogs	13.4	**55.6**	10.2	0.0	0.2	20.2	0.4	0.0
Fan	0.0	0.0	**94.8**	0.0	0.0	4.0	1.2	0.0
Rain	0.0	0.0	44.2	**50.2**	0.4	0.0	3.8	1.4
Shower	0.0	0.0	0.0	0.0	**94.6**	4.6	0.0	0.8
Siren	2.0	0.6	27.8	0.0	15.6	**54.0**	0.0	0.0
Subway	1.4	0.0	37.2	0.0	0.0	0.2	**61.2**	0.0
Waterfall	0.0	0.0	79.6	12.6	2.2	0.6	0.0	**5.0**
Average Accuracy: **55.2**								

(Without Learning)

Actual Source	Classified Source							
	Chainsaw	Dogs	Fan	Rain	Shower	Siren	Subway	Waterfall
Chainsaw	**2.2**	0.0	0.0	0.0	0.0	97.8	0.0	0.0
Dogs	1.4	**59.8**	12.6	0.0	0.0	26.2	0.0	0.0
Fan	0.0	0.0	**97.6**	0.0	0.0	0.8	1.4	0.2
Rain	0.0	0.0	14.0	**63.4**	0.0	0.4	10.6	11.6
Shower	0.0	0.0	0.0	0.0	**98.8**	1.0	0.0	0.2
Siren	0.0	0.0	0.6	0.0	0.0	**99.4**	0.0	0.0
Subway	0.4	0.0	38.2	0.2	0.0	2.6	**58.6**	0.0
Waterfall	0.0	0.0	0.2	21.0	10.0	16.4	0.4	**52.0**
Average Accuracy: **66.5**								

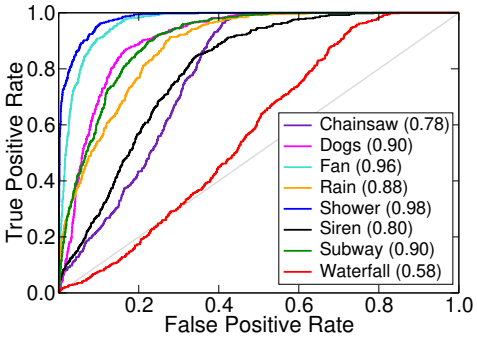
(ALSS)

Actual Source	Classified Source							
	Chainsaw	Dogs	Fan	Rain	Shower	Siren	Subway	Waterfall
Chainsaw	**64.2**	0.0	20.0	0.0	0.0	15.8	0.0	0.0
Dogs	1.4	**58.2**	4.0	0.0	0.0	36.0	0.4	0.0
Fan	0.0	0.0	**99.0**	0.2	0.0	0.2	0.4	0.2
Rain	0.0	0.0	13.4	**57.6**	0.0	0.0	11.2	17.8
Shower	0.8	0.0	0.0	0.0	**98.8**	0.2	0.0	0.2
Siren	4.4	0.0	0.0	0.0	0.0	**95.6**	0.0	0.0
Subway	0.0	10.2	24.4	0.0	0.0	2.0	**63.2**	0.2
Waterfall	5.6	0.0	2.8	3.2	5.6	4.2	0.6	**78.0**
Average Accuracy: **76.8**								

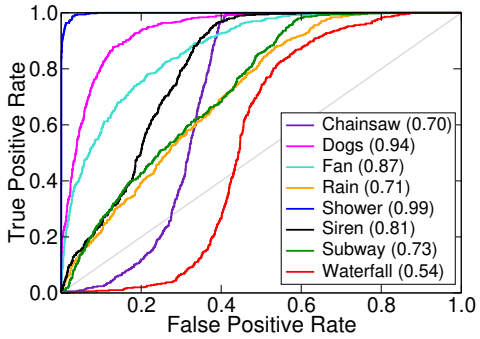
(ALSSmod)

ROC and AUC Analysis

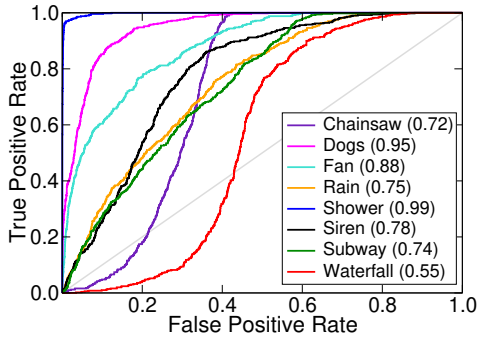
- Without Learning:



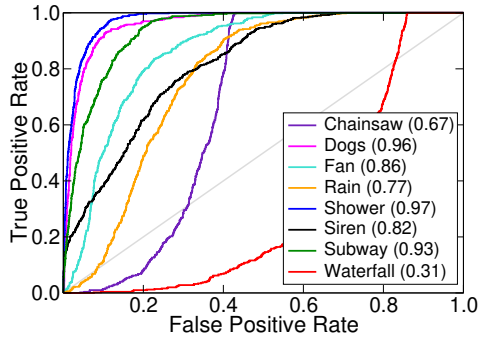
($T_{60} = 0.36$ s)



($T_{60} = 0.65$ s)

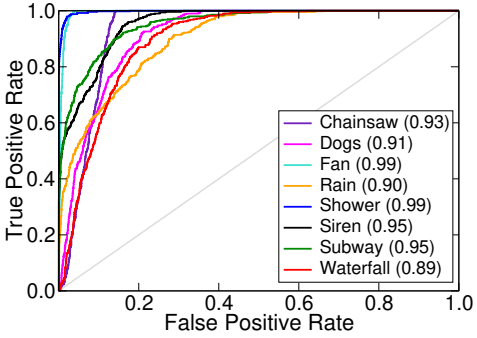


($T_{60} = 0.79$ s)

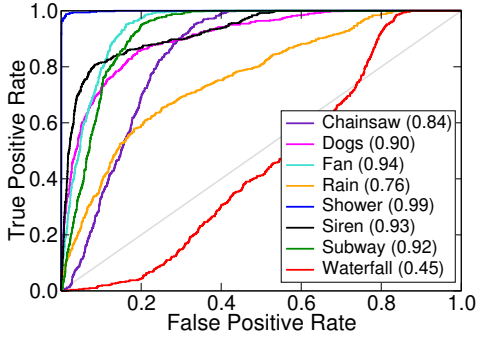


($T_{60} = 1.00$ s)

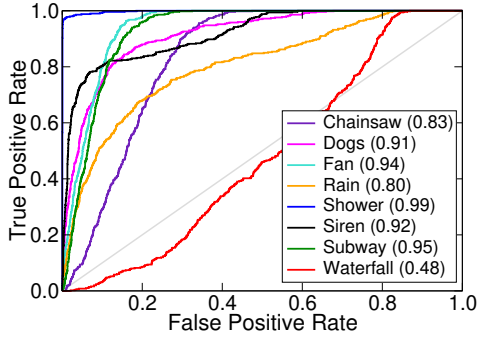
- ALSSmod:



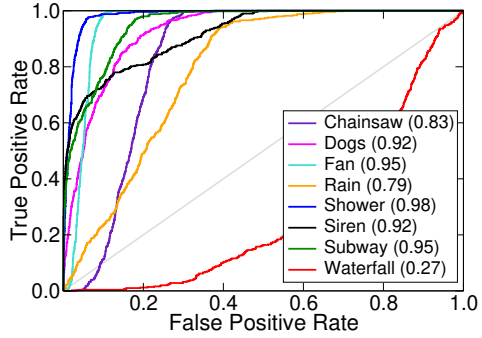
($T_{60} = 0.36$ s)



($T_{60} = 0.65$ s)



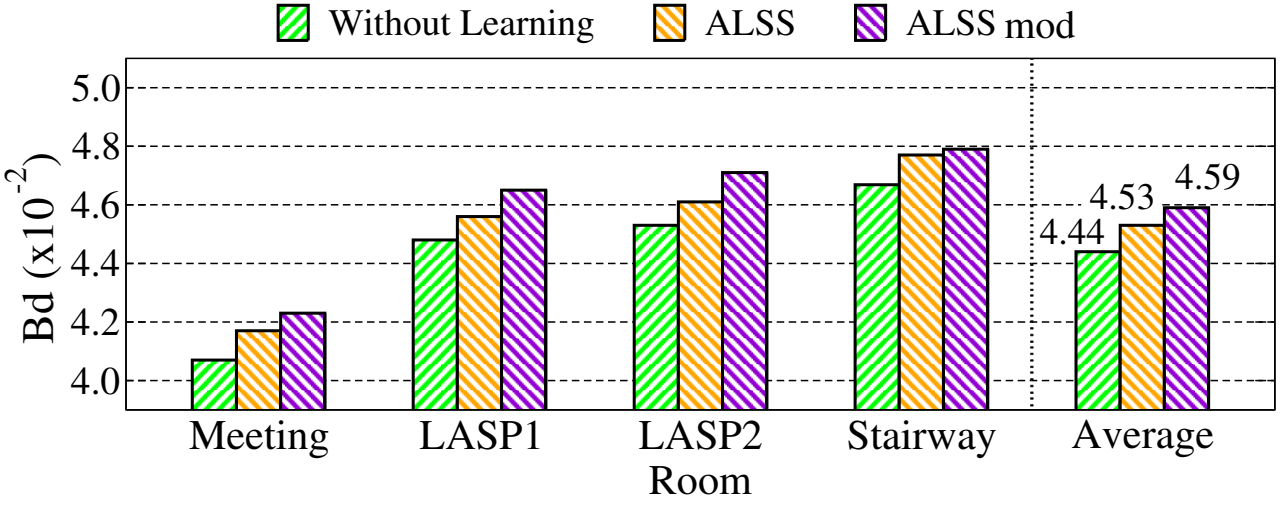
($T_{60} = 0.79$ s)



($T_{60} = 1.00$ s)

Bhattacharrya distance (Bd)

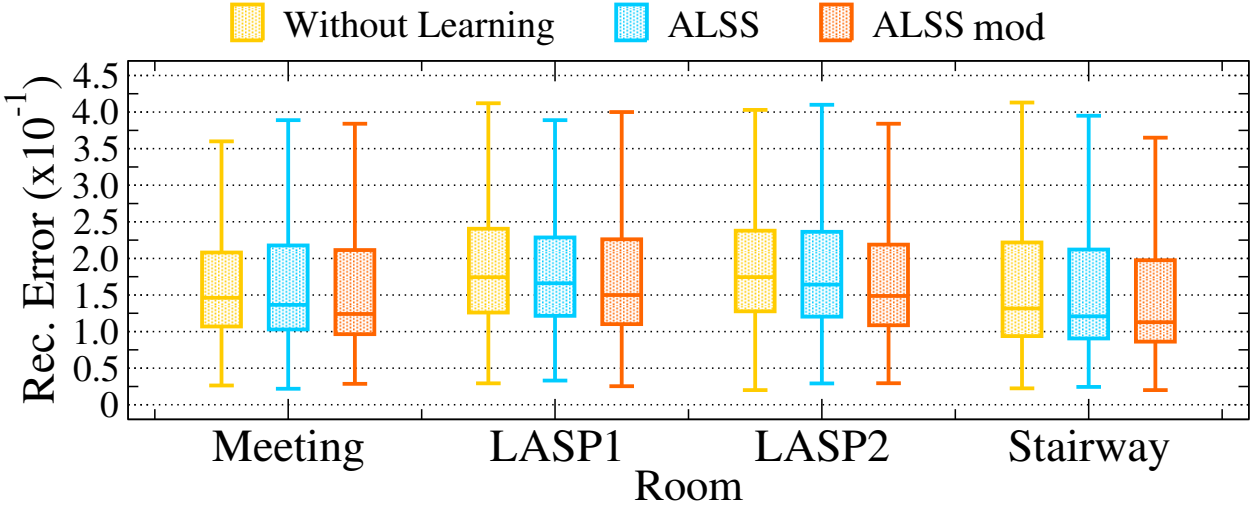
- Objective measure to access model separability on the MFCC domain
- Computed pairwise: $Bd = \frac{1}{2} \ln \frac{\frac{|\Sigma_1+\Sigma_2|}{2}}{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}} + \frac{1}{8}(\mu_1-\mu_2)^T \left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1} (\mu_1-\mu_2)$
- More discriminative models $\Rightarrow$ Higher Bd values



Highest Gain: LASP2 room $T_{60} = 0.79$ s $\rightarrow \Delta Bd = 0.18$

Sparse Coding K-SVD

- Evaluate K-SVD MFCC reconstruction error for reverberated acoustic sources.
- 80 iterations to generate 12x12 dictionaries per class.
- More informative models ⇒ Lower reconstruction errors



Maximal Decline: Meeting room $T_{60} = 0.36$ s $\rightarrow$ 15% error reduction

Conclusion

- The ALSSmod achieved the highest acoustic source classification accuracy for the MFCC-GMM under several scenarious.
- The ALSSmod attained greater AUC values, specially for the most non-stationary sources Chainsaw and Siren.
- Regarding separability and derived sparse coding the ALSSmod acquired the best results overall, which corroborates its capacity to select discriminative, separated and informative models.
- Experiments with pH feature vector demonstrated consistent gains on acoustic source classification, which indicates a good practice for future surrogate learning approaches.

Acknowledgements

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

Conselho Nacional de Desenvolvimento Científico e Tecnológico

Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro

Thank You!