Robust one-shot spectroscopic multi-component gas mixture detection via randomized smoothing

Mohamed Sy[1], Emad Al Ibrahim[1] and Aamir Farooq†[1]

[1]Physical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia
[1]{mohamed.sy, emad.ibrahim, aamir.farooq}@kaust.edu.sa

Abstract

Spectroscopic methods are well-established and widely used tools in analytical chemistry. They leverage the interaction between light and matter to extract information about chemical species and their abundances. Application of spectroscopic methods is hindered by the need for large datasets and the presence of unknown interference. These problems present significant challenges in developing reliable machine learning models for spectroscopic gas sensing. In many real-world applications, data are scarce, and absorbance signals are often corrupted by noise or overlapping spectral features from interfering species, making accurate detection and classification difficult. To address these challenges, we apply a set of targeted augmentation strategies aimed at improving model robustness and selectivity in gas sensing tasks. Specifically, we propose a one-shot learning approach with Voigt profile augmentation to handle pressure-induced spectral variations. Additionally, we use fictitious augmentations to mitigate the impact of unknown interfering species. Furthermore, we apply randomized smoothing to enhance resilience to unseen perturbations and domain shifts, promoting consistent performance in noisy, real-world conditions. Our models significantly outperform undefended baselines, offering a reliable, data-efficient solution for gas detection. Research in this area holds significant societal impact, with potential applications in occupational safety (detecting hazardous or toxic gas exposure), healthcare (identifying biomarkers in exhaled breath), and environmental protection (monitoring air pollutants and greenhouse gases). Code and models are available at .

1 Introduction

Spectroscopic gas sensing has the potential to impact many lives through applications in safety, health, and the environment [1, 2]. Traditional chemometric techniques like Partial Least Squares (PLS) [3] and Independent Component Analysis (ICA) [4] are often ineffective or sub-optimal in extracting useful information from spectroscopic data. This led to a plethora of machine learning-based solutions to spectroscopic problems[5], ranging from dealing with noise[6], unknown interference[7, 8], and unknown reference spectra[9], to addressing low sensitivity and selectivity [10–12].

However, significant challenges remain in applying machine learning-based spectroscopic solutions to real-world scenarios. Models trained on clean laboratory or simulated data often struggle when confronted with real-world target data. This study addresses two critical domain shifts that contribute to these challenges: variations in pressure and interference, particularly in data-scarce regimes.

Efforts in literature have attempted to tackle pressure dependence by simulating large datasets for multi-class and multi-label classification problems [13, 14]. This is effective but only a viable route for a limited number of chemical species that could be accurately simulated. Voigt convolutions

NeurIPS 2024, Machine Learning and the Physical Sciences.

have been proposed as a cheap method to account for pressure dependence with limited data [15]. This approach leads to certified robustness guarantees when coupled with randomized smoothing for multi-class classification [16–18]. Lastly, fictitious augmentations were proposed to mitigate the effect of unknown interfering species for regression tasks [7]. This study advances the field of multi-label spectroscopic mixture classification by tackling three pivotal questions: (1) How can models be trained effectively with minimal or incomplete data? (2) Which augmentation techniques are most effective in managing spectral variability and interference? (3) How can model robustness be ensured against unforeseen perturbations?

2 Data

The study leverages diverse data sources to simulate real-world constraints in a controlled manner. The source data represents scenarios with limited availability, a common challenge in spectroscopic studies. Through augmentations and mixing operations, this data is transformed into a synthetic dataset suitable for model training. Target data is derived from high-quality spectral simulations and is restricted to a small number of chemical species. Lastly, experimental data, though scarce, is utilized for demonstration purposes. These different types of data are shown in Figure 1 and further explained in this section.

Source data: Using the HAPI software [19] and spectroscopic parameters from HITRAN [20] and JPL [21], we generate reference spectra r_i for the seven target species at P=0.05 Torr with a wavenumber resolution of 0.016 cm^{-1} (reference spectra are labeled 1-7 in Figure 1). This limited amount of information is used as the basis for training many of the models presented to emulate data-scarce applications.

Synthetic data: Reference spectra from the source data are then used to create synthetic data. First they are passed into a mixing operator that assumes ideal Beer-Lambert blending behavior (i.e. , $A_{mixture} = -ln[I/I_0] = \Sigma_i c_i r_i L$ where mixture absorbance can be summed linearly by the molar contributions of the reference absorbances for a given path length L, mole fraction c_i and species r_i). Random mixtures are generated such that each species contributes a minimum of 1% to the total absorbance, approximating the experimental detection limit. To improve generalization, we use fictitious augmentations (flip, mirror, dilate) and Voigt profile convolutions [15, 7]. Fictitious augmentations help manage unknown interference by modifying known spectra, while Voigt convolutions simulate spectral variations due to pressure changes.

Target data: Once again, we leverage the HAPI software [19] and spectroscopic parameters from HITRAN [20] and JPL [21], to simulate a clean training dataset. Generated data is of the same wavenumber resolution (0.016 cm^{-1}) but now spans a large pressure range P=0.001 - 16 Torr. Additionally, interfering species beyond the target species are treated as unknowns. (only appearing in the test data - see species 8-12 in Figure 1).

Experimental data: A limited set of 30 experimental spectra from a THz microelectronics spectrometer were used to test the proposed models [13, 14]. The data consist of 5 mixture spectra (with CH_3CN as an unknown interferent), and the remaining 25 are pure spectra at varying pressure conditions. This dataset serves as a demonstration; however, a larger dataset is required to draw robust conclusions.

3 Models

For the sake of fair comparison, an architecture similar to that presented in TSMC-net [14] is used as a base classifier for all experiments. Only the last layer was altered to convert the problem from a multi-class classification (via label powerset conversion) to a multi-label classification and the rest of the 1D CNN was kept the same. Figure 2 shows how the base classifier f_θ can be converted to a robust classifier g by randomized smoothing based on perturbations sampled from $p(\epsilon)$ (i.e. , $g(x) = argmax_{k \in C}\, \mathbb{E}_{\epsilon \sim p(\epsilon)}[f_\theta(x+\epsilon)_k]$). It is important to note that in this work, the augmentations used for training the base classifier are consistently aligned with the perturbations applied during randomized smoothing.

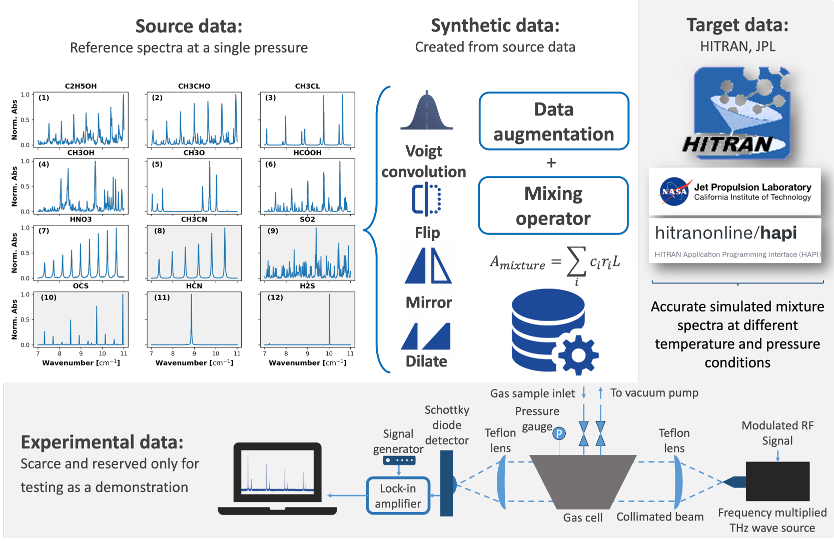

Figure 1: Data sources used in this study. Source data emulates real life conditions of data scarcity at a single pressure. The normalized spectra of species considered are shown where (1-7) represent target species and (8-12) represent interfering species that are presumed unknown at training. Synthetic data is then generated from the source data by augmentations and a mixing operator. On the other hand, target data comes from the HITRAN and JPL databases which rely on experimentally fitted parameters to simulate clean mixture spectra at a given temperature and pressure. Finally, the experimental apparatus used to generate THz spectra for demonstration is shown (reproduced from [13]).

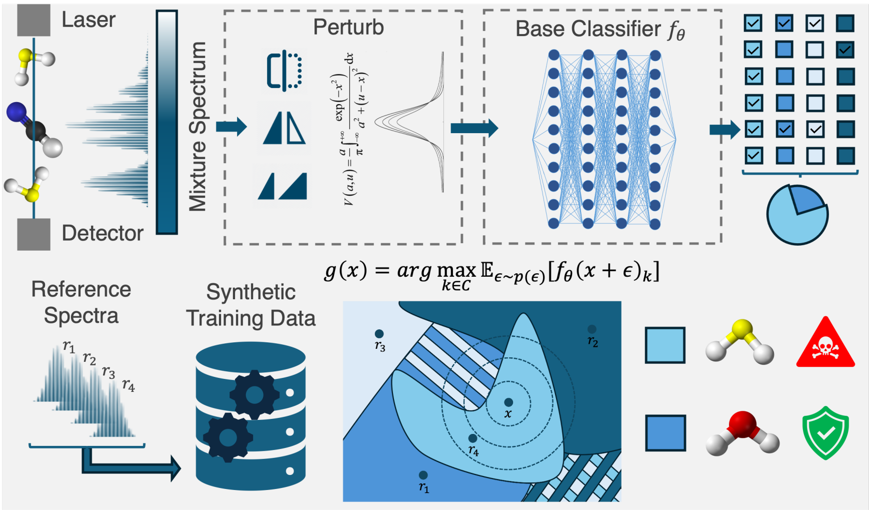
Laser
Detector
Mixture Spectrum
Perturb
$V(a,u) = \frac{a}{\pi}\int_{-\infty}^{\infty} \frac{\exp(-x^2)}{a^2+(u-x)^2} dx$
Base Classifier f_θ
$g(x) = arg \max_{k \in C} \mathbb{E}_{\epsilon \sim p(\epsilon)}[f_\theta(x+\epsilon)_k]$
Reference Spectra
r_1
r_2
r_3
r_4
Synthetic Training Data
x

Figure 2: Workflow diagram of this study. Reference spectra denoted as r_i are used to create synthetic data to train the base classifier f_θ. At test time, a query mixture spectrum x is obtained from an experiment or a high-quality simulation software. The mixture spectrum is perturbed N number of times by fictitious and Voigt augmentations and then passed to the base classifier. A majority vote is then taken to smoothen the decision boundary and give a robust prediction of dangerous toxic gasses.

4 Results

Models are evaluated using metrics such as accuracy, F1-score, and precision with a test dataset of 1,640 mixture observations containing up to 12 VOCs, across 164 pressure conditions. In interference-free conditions, the **Baseline** model, trained on both source and target data, achieved 99% accuracy. If trained on source data only, the **Baseline** model achieved 77% accuracy. The **Baseline + Voigt** model, which used Voigt convolutions, reached 87% accuracy, while the **Baseline + Voigt (V) + Randomized Smoothing (RS)** model, incorporating additional augmentations and randomized smoothing, improved accuracy to 92%.

For classification under interference, the **Baseline** model achieved 76% accuracy. The **Baseline + FA** model, using fictitious augmentations, improved to 92%. If trained using source data only, the **Baseline** model showed 69% accuracy, and the **Baseline + Voigt** model reached 70%. The **Baseline + Voigt + Fictitious Augmentations (FA)** model achieved 83% accuracy, with the **Baseline + Voigt (V) + Fictitious Augmentations (FA) + Randomized Smoothing (RS)** model improving to 88%. All results are summarized in Table 1.

Table 1: Summary of test results on target (simulated) data. V refers to Voigt convolutions, FA refers to fictitious augmentations, and RS refers to randomized smoothing.

Models	Trained on	Interference	Accuracy	F1-score	Precision
Baseline	target	No	0.99	1.00	0.99
Baseline	source	No	0.77	0.78	0.83
Baseline + V	source	No	0.87	0.91	0.83
Baseline + V + RS	source	No	0.92	0.98	0.93
Baseline	target	Yes	0.76	0.79	0.79
Baseline + FA	target	Yes	0.92	0.91	0.94
Baseline	source	Yes	0.69	0.75	0.73
Baseline + V	source	Yes	0.70	0.75	0.75
Baseline + V + FA	source	Yes	0.83	0.83	0.86
Baseline + V + FA + RS	source	Yes	0.88	0.89	0.90

The classification models were evaluated using experimental data that included 25 pure components and 5 mixtures. The **Baseline** model, trained on both sources and targets, achieved 83% accuracy in interference-free conditions. When trained on source data only, the **Baseline** model had 64% accuracy, while the **Baseline + Voigt** model reached 72%. The **Baseline + V + RS** model improved to 77% accuracy. In classification under interference, the **Baseline** model achieved 83% accuracy. The **Baseline + FA** model improved to 90%. When trained on source data only, the **Baseline** model showed 63% accuracy, with the **Baseline + V** model reaching 83%. The **Baseline + FA** model achieved 90%, and the **Baseline + V + FA + RS** model improved to 93%. Results are summarized in Table 2. All experiments were ran on a single GPU in <30 minutes.

Table 2: Summary of test results on experimental data combining 25 pure components and 5 mixtures. V refers to Voigt convolutions, FA refers to fictitious augmentations, and RS refers to randomized smoothing.

Models	Trained on	Interference	Accuracy	F1-score	Precision
Baseline	target	No	0.83	0.91	0.86
Baseline	source	No	0.64	0.72	0.67
Baseline + V	source	No	0.72	0.78	0.76
Baseline + V + RS	source	No	0.77	0.83	0.80
Baseline	target	Yes	0.83	0.83	0.83
Baseline + FA	target	Yes	0.90	0.88	0.92
Baseline	source	Yes	0.63	0.65	0.64
Baseline + V	source	Yes	0.83	0.87	0.87
Baseline + V + FA	source	Yes	0.90	0.91	0.92
Baseline + V + FA + RS	source	Yes	0.93	0.95	0.95

5 Conclusion

This study demonstrated effective spectroscopic multi-label classification training techniques under limited data availability. We showed that given a single reference spectrum per target species, one could train a model via Voigt augmentations with accuracy approaching that of a model trained on extensive pressure-dependant data. Given that many applications could encounter complex gas mixtures, we emphasized the importance of defending against unknown interfering species. Since many gaps remain in spectroscopic databases, techniques like the ones presented here could help accelerate adoption of machine learning based spectroscopic solutions in the real world.

References

[1] Christopher S Goldenstein, R Mitchell Spearrin, Jay B Jeffries, and Ronald K Hanson. Infrared laser-absorption sensing for combustion gases. *Progress in Energy and Combustion Science*, 60:132–176, 2017.

[2] Aamir Farooq, Awad BS Alquaity, Mohsin Raza, Ehson F Nasir, Shunchun Yao, and Wei Ren. Laser sensors for energy systems and process industries: Perspectives and directions. *Progress in Energy and Combustion Science*, 91:100997, 2022.

[3] Ali Elkhazraji, Mohamed Sy, Mhanna Mhanna, Joury Aldhawyan, Mohammad Khaled Shakfa, and Aamir Farooq. Selective btex detection using laser absorption spectroscopy in the ch bending mode region. *Experimental Thermal and Fluid Science*, 151:111090, 2024.

[4] Martin Kermit and Oliver Tomic. Independent component analysis applied on gas sensor array measurement data. *IEEE Sensors Journal*, 3(2):218–228, 2003.

[5] Andre Nicolle, Sili Deng, Matthias Ihme, Nursulu Kuzhagaliyeva, Emad Al Ibrahim, and Aamir Farooq. Mixtures recomposition by neural nets: A multidisciplinary overview. *Journal of Chemical Information and Modeling*, 64(3):597–620, 2024.

[6] Mohamed Sy, Mhanna Mhanna, and Aamir Farooq. Multi-speciation using a tunable laser and deep neural networks. In *2023 Conference on Lasers and Electro-Optics (CLEO)*, pages 1–2. IEEE, 2023.

[7] Emad Al Ibrahim and Aamir Farooq. Augmentations for selective multi-species quantification from infrared spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*, 240:104913, 2023.

[8] Mhanna Mhanna, Mohamed Sy, and Aamir Farooq. A selective laser-based sensor for fugitive methane emissions. *Scientific Reports*, 13(1):1573, 2023.

[9] Mohamed Sy, Emad Al Ibrahim, Ali Elkhazraji, and Aamir Farooq. Unsupervised source separation technique for multi-speciation using a single laser: Quantifying hydrocarbons without their reference spectra. In *2024 Conference on Lasers and Electro-Optics (CLEO)*, pages 1–2. IEEE, 2024.

[10] Mhanna Mhanna, Mohamed Sy, Ayman Arfaj, Jose Llamas, and Aamir Farooq. Laser-based selective btex sensing using deep neural networks. *Optics Letters*, 47(13):3247–3250, 2022.

[11] Mhanna Mhanna, Mohamed Sy, Ali Elkhazraji, and Aamir Farooq. Deep neural networks for simultaneous btex sensing at high temperatures. *Optics Express*, 30(21):38550–38563, 2022.

[12] Mhanna Mhanna, Mohamed Sy, Ali Elkhazraji, and Aamir Farooq. Multi-speciation in shock tube kinetics using deep neural networks and cavity-enhanced absorption spectroscopy. *Proceedings of the Combustion Institute*, 40(1-4):105733, 2024.

[13] M Arshad Zahangir Chowdhury, Timothy E Rice, and Matthew A Oehlschlaeger. Voc-net: A deep learning model for the automated classification of rotational thz spectra of volatile organic compounds. *Applied Sciences*, 12(17):8447, 2022.

[14] M Arshad Zahangir Chowdhury, Timothy E Rice, and Matthew A Oehlschlaeger. Tsmc-net: Deep-learning multigas classification using thz absorption spectra. *ACS sensors*, 8(3):1230–1240, 2023.

[15] Mohamed Sy, Emad Al Ibrahim, and Aamir Farooq. Voc-certifire: Certifiably robust one-shot spectroscopic classification via randomized smoothing. *ACS Omega*, 2024.

[16] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

[17] Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. *Advances in Neural Information Processing Systems*, 33:21945–21957, 2020.

[18] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[19] Roman V Kochanov, IE Gordon, LS Rothman, P Wcisło, C Hill, and JS Wilzewski. Hitran application programming interface (hapi): A comprehensive approach to working with spectroscopic data. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 177:15–30, 2016.

[20] Iouli E Gordon, Laurence S Rothman, Christian Hill, Roman V Kochanov, Y Tan, Peter F Bernath, Manfred Birk, V Boudon, Alain Campargue, KV Chance, et al. The hitran2016 molecular spectroscopic database. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 203:3–69, 2017.

[21] HM Pickett, RL Poynter, EA Cohen, ML Delitsky, JC Pearson, and HSP Müller. Submillimeter, millimeter, and microwave spectral line catalog. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 60(5):883–890, 1998.