Draft version March 23, 2015
Preprint typeset using LaTeX style emulateapj v. 12/16/11

[IDENTIFICATION OF SPECTRAL LINES USING SPARSE CODING]

Andrés Riveros, Karim Pichara, Pavlos Protopapas, Diego Mardones, Paulina Troncoso and Mauricio Araya

Draft version March 23, 2015

ABSTRACT

Astronomy is facing new challenges on how to analyze big data and therefore, how to search or predict events/patterns of interest. The automatic detection and identification of spectral lines is an astronomical problem that has not been solved yet, so currently the identification is limited to the manual analysis of spectra by radio-astronomers. The use of spectroscopy allows to describe the chemical composition of astronomical objects through the emissions from the interaction between the radiation and the matter, thus causing emission lines. New observations in previously unexplored wavelength regions that will be available thanks to project as the Atacama Large Millimeter Array (ALMA), with which it is intended to use machine learning techniques to identify automatically these emission lines. Using simulated data based on the observations that has been obtained from the radio-telescope ALMA, it is proposed an algorithm with which to identify lines in order to determine the molecules that compose the observed galaxies. For this will be used the technique of Sparse Coding, that will evaluate the molecules that can origin the observed spectra, and it is expected to be found the best combination of molecules to recreate that spectra. From this algorithm, the astronomers may obtain a probability associated to the possible combinations of molecules composing astronomical objects.

Subject headings: spectral lines: emission lines; spectroscophy techniques; method: data mining

1. INTRODUCTION

This project is developed as part of a collaborative project between several Chilean universities for the initiative of the creation of the chilean virtual observatory Observatorio Virtual Chileno (ChiVO). ChiVO is an online platform that will make available to astronomers the measures of the radio-telescope Atacama Large Millimeter Array (ALMA). Also, ChiVO will provide several tools in order to process the data of ALMA measurements and to get specialized information of those measurements.

The data from the radio-telescope ALMA are of data cubes. The three dimensions correspond to two spatial dimension, and the third one of wave frequency. This means that for each spatial point can be obtained a spectrogram. The Pontifical Catholic University of Chile (PUC) will participate with the development of an algorithm to identify spectral lines using data mining techniques. This tool takes as input a observed spectra and returns a list with the best prediction of molecules that by the theoretical behavior their spectral lines describes the observed spectra.

Currently, there is not enough available ALMA data to train a model, so this project has been developed using synthetic data. The Astronomical SYnthetic Data Observatory (ASYDO) project, a parallel project of ChiVO, will be the used in order to generate synthetic data to test the algorithm.

The objective of this investigation is to develop an algorithm that allows to identify spectral lines automatically in a simulated observed spectra. For this purpose, the objectives are the next:

The spectral lines must be codified into a convenient numerical representation in order to apply sparse coding techniques on them.

Must be determined the best possible collection of base spectra from the previous representation, in order to elaborate a dictionary. The dictionary must be able to generate the most part of the simulated spectra data of ALMA.

Each combination of dictionary elements should give possible models of molecules that origin a spectra, so the combination of words of the dictionary represents a combinations of molecules that generates the spectra.

Finally, the combination that minimizes the error between the observed and generated spectra will be the combination of molecules predicted.

With this, the algorithm will be able to give as output the probabilities associated to the different possible combination of molecules that originated the observed spectra studied.

2. RELATED WORK

The detection of spectral lines following the traditional method is limited to the manual analysis of the data in order to found the frequencies associated with the highest peaks in the spectral. Those peaks and a map with theoretical frequencies, and the experience of the astronomer, allow to classify those observations to certain molecules and its energies states.

The lack of scalability of this not automatized method, and the impractical mechanical process is not suitable for a big amount of data (P. Schilke and Phillips 2001). The difficulty of predicting new maps between the observed frequencies and the theoretical frequencies is given too by the blending of the lines. For those reasons, it would be desirable to automate this task.

El problema de mezclas de lneas y superposiciones son producto de tanto ruido como la falta de sensibilidad para distinguir entre dos lneas en frecuencias cercanas. Lo anterior tambin puede producir peaks dobles en ciertas lneas (Cernicharo et al. 2013).

Un problema importante a la hora de identificar fre-

cuencias subyace en lneas pticamente delgadas, que tienden a dar resultados incorrectos. Usualmente, el uso de lneas de isotopos para su correccin resulta en un proceso costoso en tiempo y por lo mismo no es apto para datos masivos (P. Schilke and Phillips 2001).

Nummelin et al. (Nummelin A, Bergman P, Hjalmarson A, Friberg P, Irvine W M, Millar T J and S 1998) propone el uso de un ajuste manual de las lneas a una forma arbitraria dada por una gausiana, obteniendo por cada lnea su frecuencia observada, el peak en el brillo de temperatura y el ancho de la velocidad (ancho total a media altura), para as proceder con la identificacin de la lnea al asociarla con una molcula en cierto estado de energa.

Para la identificacin de lneas considerando las relaciones entre brillo de temperatura en un mismo espectro, es necesario asumir temperatura y origen homogneo, dado que la diferencia de temperatura cambia la relacin en serie de intensidades de lneas hiper-finas (Nummelin A, Bergman P, Hjalmarson A, Friberg P, Irvine W M, Millar T J and S 2000).

Esto es importante a la hora de utilizar datos simulados con el fin de representar fielmente las caractersticas fsicas de las estructuras a utilizar para entrenar, de modo que el modelo sea posteriormente aplicable sin mayores variaciones al utilizar datos reales de ALMA.

Es posible detectar patrones en las lneas que corresponden a la misma molcula e ispotropo a partir de intensidad relativa considerando que existe una razn entre diferencias de velocidad que es constante para un conjunto de lneas de emisin. Esto permite buscar patrones no tan solo de manera individual, sino que a travs del anlisis manual de series de lneas que se asocian a una misma molcula o tomo en sus diferentes estados energticos.

Los esfuerzos para desarrollar una herramienta automtica de deteccin de lneas actualmente apuntan a herramientas semi-automticas que utilizan como base complejos modelos fsicos y qumicos para la clasificacin de lneas.

XCLASS [1], CASSIS [2] y WEEDS [3] son herramientas que apuntan a modelar la composicin de los espectros de tal forma que las simulaciones se asemejen a lo observado, existiendo grandes esfuerzos en realizar dichos modelamientos para solo la identificacin de lneas. (Schilke et al. 2011).

Estas herramientas hacen uso de catlogos que contienen informacin sobre lneas espectroscpicas de molculas y sus frecuencias tericas de laboratorio, las que estn disponibles pblicamente en catlogos como (JPL [4], CDMS [5], Toyama [6]) (Schilke et al. 2011), los que han sido compilados en Splatalogue [7] (Remijan and Markwick-Kemper 2008; Remijan 2010).

Las tcnicas anteriormente descritas no son escalables al no ser procesos automatizados y depender de anlisis o ajustes manuales que con la inminente llegada de enormes cantidades de datos provenientes de instrumentos como ALMA, dejan de ser aplicables. Por esto es necesario buscar algoritmos de clasificacin que deleguen la tarea de identificar y clasificar lneas espectrales.

[1] https://www.astro.uni-koeln.de/projects/schilke/XCLASS
[2] http://cassis.cesr.fr
[3] https://www.iram.fr/IRAMFR/GILDAS
[4] http://spec.jpl.nasa.gov
[5] http://www.astro.uni-koeln.de/cdms
[6] http://www.sci.u-toyama.ac.jp/phys/4ken/atla
[7] http://www.splatalogue.net

3. BACKGROUND

3.1. *Sparse Coding*

3.2. *Spectroscopy*

Product of the radiation emitted by stellar objects, and its interaction with matter, emission lines are generated. This emission lines are distinctive for certain energy levels of molecules that make astronomical objects. The detection of emission lines and subsequent association with the molecules that cause them, about the structure of stellar objects.

The combination of these emission lines for each object allows a fingerprint of this unique, given its internal characteristics and different factors such as the temperature of the object, the speed with which it travels through space, etc.

With this information, it would have a large number of spectral lines, so arises the idea of detecting these spectral lines as an astronomical problem of interest to apply data mining techniques.

On the side of data mining, there will be enough data to techniques and develop an algorithm for automatic identification. By astronomers, it may have a tool that automates the identification of emission lines in spectrograms belonging to astronomical objects.

Note that the objective of the algorithm is to support the rigor of a classification made by an expert, to serve as an initial preprocessing for this sort your lines, so it does not seek to completely replace the work of the expert, and is expected margin mistake for certain identifiable cases.

After defining the problem, need to establish a data set with which to begin developing the algorithm. Given the amount of spectrum needed to develop a predictive model, as currently do not have enough measurements of radio telescope, we chose to use a simulated spectra.

4. METHODOLOGY

The methodology can be divided in several steps in order to find a solution for the problem:

Identification of the problem: Identify the problem to solve: The scope of the problem and the type of data that will be used, in this case, the type of simulated ALMA data. Must be defined the parameters and subset of molecules to use.

Simulation of data: Especificaciones: Se deben definir con los parmetros de la simulacin, as como un subconjunto de molculas, en conjunto con astrnomos. Para esto deben considerarse las limitaciones de la simulacin.

Related work: Se estudiar el estado del arte en la clasificacin de lneas espectrales. Se buscarn situaciones similares con el fin de aplicar tcnicas adecuadas a este caso. For the codification of the spectra in order to represent the presence of molecules in a observed spectra. To archive this, the previous related work like Raman spectroscopy (Howle, ett al. 2005) (O' Connell et all) can be useful. (Howley et al. 2005).

Development of the solution: se realizar una iteracin de soluciones al problema de codificacin sparse, analizando el efecto de los parmetros en los modelos a utilizar y determinando la cantidad de parmetros necesarios para abarcar mayor complejidad en la identificacin de los espectros. Optimize the solution of the sparse coding problem. Analize the effect of each parameter of the model.

4.1. *Splatalogue*

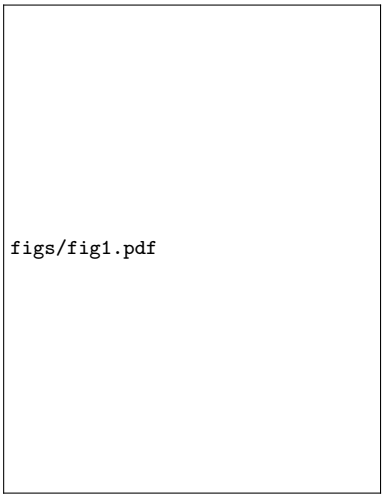


FIG. 1.— ...

4.2. *Synthetic Data*

To develop the algorithm for identifying spectral lines web service that provides simulated data goat is used. This project, called Astronomical Observatory Data synthetic (ASYDO) runs parallel to this project as part of the tools you can use to astronomers as web service. This project is available at [8]https://github.com/ChileanVirtualObservatory/ASYDO.

The simulation will generate a set of training to develop the identification algorithm proposed in this project. As this service simulation an important input algorithm development, there have been a series of meetings to work together and get a set of simulated appropriate data.

With the help of astronomers involved in the goat project has been given the necessary parameters to perform the simulation. Defined the complexity and importance of replicating certain features that help you get curves that approach spectra enough data to be obtained from observations of ALMA.

For the algorithm developed is correctly simulated data, it should be included in the meta-data cubes emission lines present in the spectra. This will be possible to evaluate the predictive model and determine metrics to validate the predictions.

The characteristics of ALMA measurements have yielded the following specifications when simulating cubes ALMA data type:

Ancho de banda espectral	4000 MHz
Resolusin espectral	1 MHz

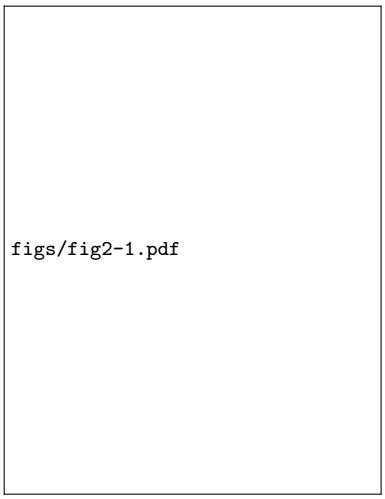
TABLE 1
COMMON SPECIFICATIONS FOR ALL SIMULATED DATA CUBES.

And as input parameters, the user will need to provide:

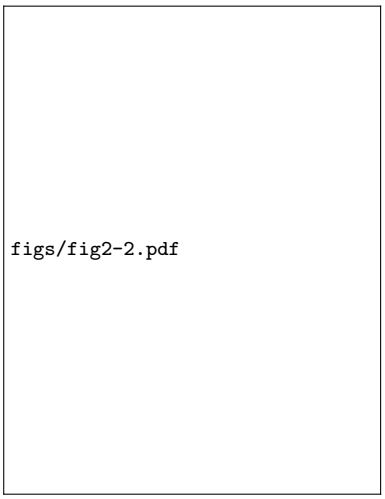
Frecuencia central	MHz
Resolusin espectral	1 MHz
Asencin	Grados
Declinacin	Grados
Ancho de las lineas	(fwhm)

TABLE 2
PARAMETERS PROVIDED BY THE USER FOR DATA CUBES SIMULATED.

To represent measurements manner close to reality, was chosen with the help of a set of molecules astronomers and their isotopes. The criterion was to select molecules with a structure that is not in excess complicated, so that not involve unnecessary complexity for the algorithm. In addition, this set of molecules should be representative of the molecules that users expect of astronomical objects. These are shown in the following table.



(a) Title 2.1.



(b) Title 2.2.

FIG. 2.— In figure 2(b) is shown that ...

Nombre	Frmula	Istopos
Carbon Monoxide	'CO'	'COv=0', 'COv=1','13COv=0', 'C18O', 'C17O','13C17O', '13C18O'
Diazenylium	'N2H'	'N2H+v=0', 'N2D+', '15NNH+', 'N15NH+'
Cyanide Radical	'CN'	'CNv=0', '13CN', 'C15N'
Hydrogen Cyanide	'HCN'	'HCNv=0', 'HCNv2=1', 'HCNv2=2','HCNv3=1' 'HC15Nv=0', 'H13CNv2=1' 'H13CNv=0' 'HCNv1=1', 'HCNv3=1', 'DCNv=0' 'DCNv2=1', 'HCNv2=4', 'HCNv2=1 1-v2=4 0'
Carbon Monosulfide	'CS'	'CSv=0', '13C34Sv=0', 'C36Sv=0' 'C34Sv=0', 'CSv=1-0', '13CSv=0' 'C33Sv=0', 'CSv=1', 'C34Sv=1'
Thioxoethenylidene	'CCS'	'CCS', 'C13CS', '13CCS', 'CC34S'
Hydrogen sulfide	'H2S'	'H2S', 'H2S', 'H234S', 'D2S'
Thioformaldehyde	'H2CS'	'H2CS', 'H213CS', 'H2C34S'
Sulfur Dioxide	'SO2'	'SO2v=0', '33SO2', '34SO2v=0', 'SO2v2=1'
Sulfur Dioxide	'OSO'	'OS18O', 'OS17O'
Formaldehyde	'H2CO'	'H2CO', 'H2C18O', 'H213CO'
Formylium	'HCO'	'HCO+v=0', 'HC18O+', 'HC17O+', 'H13CO+'
Cyanobutadiyne	'HC5N'	'HC5Nv=0', 'HC5Nv11=1', 'HCC13CCCN', 'HCCCC13CN' 'HCCC13CCN', 'H13CCCCCN' 'HC13CCCCN'
Methanol	'CH3OH'	'CH3OHvt=0', '13CH3OHvt=0 ', 'CH318OH' 'CH3OHvt=1 ', '13CH3OHvt=1'

TABLE 3
SET OF MOLECULES AND ISOTROPIC WITH WHICH THE SIMULATIONS WERE PERFORMED.

4.3. *Codification*

The proposed algorithm aims at identifying the molecular components that are part of different simulated astronomical objects. For this, we analyze the spectral lines from the spectra of these astronomical objects, in order to find patterns and predict its composition.

For an astronomical object, observe their spectral lines can help identify the molecules that compose it, since for each energy state of these molecules, being those present in the object, manifested in emission lines along their spectra observed at certain frequencies.

The algorithm takes advantage of this behavior and looks for patterns based on the presence of certain lines. When a molecule is present in the composition of an object, probably a number of lines should be observed along the spectrum. This means the fewer quee observed theoretical lines of said molecule, the lower the probability that the molecule forms part of the composition of the object.

So if there is confusion in identifying a spectroscopic line between two molecules, it is possible to perform a probabilistic prediction of the molecule to which corresponds the line based on the presence of each theoretical line along the spectrum being observed.

4.4. *Diseo del Prototipo*

Como se mencion anteriormente, para el diseo del prototipo se utiliz el proyecto ASYDO. La temperatura en estas simulaciones no posee unidades, ya que las magnitudes de las lneas son relativas a la lnea ms alta de CO, a la cual se le ha asignado un valor arbitrariamente. Esto significa que los valores en s mismos no poseen un significado.

El proceso para simular los cubos de prueba consisti en encontrar todas las molculas y todos sus isotopos dentro de un rango de frecuencia. El rango de frecuencia utilizado para estas pruebas fue desde los 602000 MHz hasta los 606000 GHz, aproximadamente, que corresponde a la banda 9 de ALMA. Se corri un script de lneas combinadas de varios subconjuntos aleatorios de isotopos, subconjunto de tamao variable del total de molculas tericamente existentes en esta ventana de frecuencias utilizada.

El objetivo del algoritmo es entonces recuperar la lista de molculas con la cual se generaron los cubos de datos.

Para realizar esta tarea solo se pueden utilizar los espectrogramas observados. A travs de mtodos estadsticos se pretende predecir la presencia de ciertas molculas y validar dichas predicciones al conocerse las molculas utilizadas para generar las simulaciones.

FIG. 3.— Se desea recuperar la lista de molculas con la que se simul el espectrograma.

4.5. *Implementacin*

El cdigo de la implementacin se encuentra disponible en el repositorio del proyecto de ChiVO en git. [9]

4.5.1. *Etapa de Deteccin*

El proceso de deteccin de lneas utiliza un parmetro de sensibilidad que determina si una medicin es considerada una potencial lnea espectroscpica. Esta sensibilidad depende de la desviacin estndar del ruido en una regin sin lneas visibles. Para obtener este parmetro se forz a que en cada cubo, el pixel (0 , 0) no tuviese lneas espectrales, sino que solo ruido. Esto en la prctica se puede obtener seleccionando una regin vaca del cubo de datos que el usuario previamente seleccione.

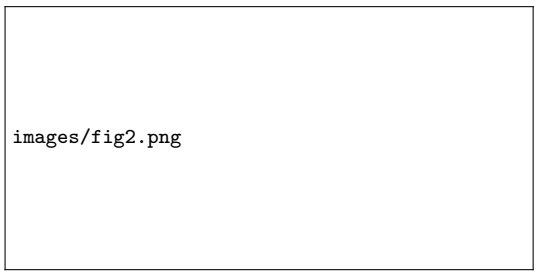


FIG. 4.— Espectro al situarse en un pxel espacial en particular.

A continuacin, para cada punto espacial del cubo de datos, se toma cada espectrograma de manera independiente. Lo primero es reducir el ruido de el espectro utilizando un filtro de Savitzky-Golay (Howley et al. 2005). Este filtro utiliza un parmetro que indica el nmero de mediciones consecutivas a utilizar para suavizar la curva, por lo que el ancho de las curvas simuladas corresponde aun buen parmetro para ser asignado. Con este filtro se puede obtener un espectro con menor variacin y as, identificar con mayor claridad las lneas. Este filtro se aplica al cubo completo, incluido el pxel con solo ruido, para hacer comparables los espectros al calcular la sensibilidad.

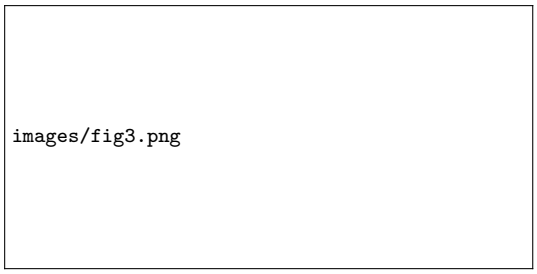


FIG. 5.— Espectro al reducir el ruido con un filtro de Savitzky-Golay.

Se comienza con la determinacin de los puntos mximos de la curva observada. Es posible asignar un parmetro que determina la distancia mxima que debe existir entre el brillo de dos frecuencias consecutivas para ser considerados un mximo. Sin este parmetro, la curva tendra una serie de falsos mximos y mnimos producto del ruido existente en las mediciones. As, cada mximo local detectado que est por sobre el parmetro de sensibilidad es un candidato a lnea de emisin.

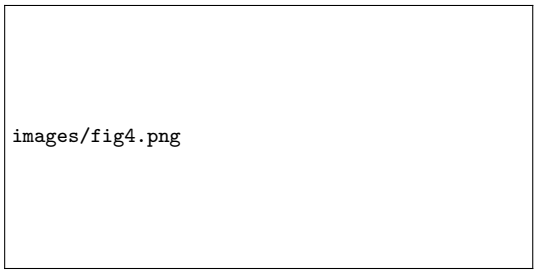


FIG. 6.— Puntos mximos locales por sobre el parmetro de sensibilidad.

Al final del proceso, se crea un vector del tamao del ancho de banda observado, donde se cada elemento del vector representa 1 Mhz del espectro. En cada frecuencia donde se detect una lnea, se asigna el valor 1, y cero en caso contrario.

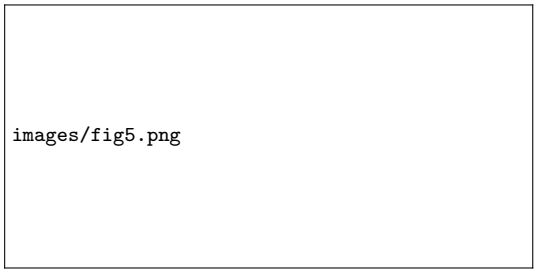


FIG. 7.— Grfico del vector observado con valores no nulos en los puntos detectados.

[9] https://github.com/ChileanVirtualObservatory/DISPLAY

0	0	...	1	...	1	...	1	...	0	0

TABLE 4
Forma del vector del espectro observado para el caso analizado.

4.5.2. *Etapa de Prediccin*

A continuacin, se utiliza el catalogo de lneas espectroscpicas tericas Splatalogue para obtener la lista de todas las frecuencias tericas en el rango observado.

Para cada isotopo con lneas tericas, se crea un vector del tamao de la ventana en Mhz, similar al de la etapa de deteccin, donde el valor de cada posicin es 1 si existe una lnea terica en dicha frecuencia para dicho isotopo, y cero en otro caso. Cada uno de estos vectores corresponde a una palabra de un diccionario de molculas.

Posteriormente, se procesan las palabras de modo que solo tengan valores distintos de cero en las frecuencias donde el espectro observado es distinto de cero. Para esto, se asigna en dichas frecuencias la diferencia entre 1 y distancia exponencial entre la frecuencia observada y la frecuencia terica ms cercana, utilizando como parmetro sigma igual al ancho de las lneas espectrales.

Con esto se espera que las palabras que tienen frecuencias tericas ms cercanas a las observadas, tendrn valores mayores, con lo que el algoritmo de sparse coding tenga preferencia a elegir dichas palabras.

La implementacin se asegura de que cada frecuencia terica cambie el valor de una y solo una frecuencia observada, y en caso de que haya otra frecuencia terica que tambin tenga como frecuencia observada ms cercana a la misma frecuencia, se asigna la menor distancia a dicha frecuencia observada. Con esto, cada palabra del diccionario queda con la misma o con menor cantidad de frecuencias distintas de cero.

Vector Obervado	0	0	0	1	0	1	0	1	0	0	0
Palabra Terica	0	0	1	0	0	1	0	0	0	0	1
Palabra Recalculada	0	0	0	0.88	0	1	0	0.6	0	0	0

TABLE 5
Ejemplo de reclculo de una palabra utilizando para la distancia exponencial sigma = 2.

Finalmente, se tiene un problema de optimizacin con una formulacin de sparse coding, donde se intenta acercar al mximo la combinacin lineal entre palabras de tal forma que se construya el vector del espectro observado. El paramero de sparsity permite restringir la cantidad de palabras que se desea que el modelo utilice para formar la observacin. En este caso se asign un valor elevado para que el modelo pudiese acercarse lo mximo posible a la observacin sin lmite de palabras. La funcin a minimizar en esta formulacin es la norma l-2 y corresponde al error cuadrtico medio.

5. RESULTS

$$min_\alpha ||x - D\alpha||_2^2 \; s.a. \; ||\alpha||_1 \leq \lambda$$

x : vector observado, D : diccionario recalculado, λ : parmetro de spartity, $||()||_2$: Norma l-2

La solucin ptima debera utilizar solo las palabras asociadas a los isotopos presentes en el espectro. En la siguiente tabla se puede ver la simulacin realizada y los isotopos rescatados:

Nombre	Frmula	Istopos
Hydrogen Cyanide	'HCN'	'HC15Nv=0', 'H13CNv2=1', 'H13CNv=0'
Thioformaldehyde	'H2CS'	'H213CS'
Sulfur Dioxide	'SO2'	'SO2v=0', 'SO2v2=1'
Sulfur Dioxide	'OSO'	'OS18O', 'OS17O'
Formaldehyde	'H2CO'	'H2C18O', 'H213CO'

TABLE 6
Conjunto de istopos con los que se realizaron simul la implementacin.

Nombre	Frmula	Istopos	Alpha
Hydrogen Cyanide	'HCN'	'HC15Nv=0'	0.0000
		'H13CNv2=1'	0.0000
		'H13CNv=0'	0.0000
Thioformaldehyde	'H2CS'	'H213CS'	0.0000
Sulfur Dioxide	'SO2'	'SO2v=0'	0.0000
		'SO2v2=1'	0.2694
Sulfur Dioxide	'OSO'	'OS18O'	0.9612
		'OS17O'	0.5589
Formaldehyde	'H2CO'	'H2C18O'	0.0000
		'H213CO'	0.0000

TABLE 7
Set of isotopes in the simulation and is alphas.

Cabe destacar que en este caso solo hubo un falso positivo, el isotopo '33SO2', con un valor de alpha de 0.6592. Esto significa que el algoritmo tiende a elegir los isotopos de molculas con mayor cantidad de lneas en el ancho de banda observado, y es un aspecto a mejorar.

Adems, ciertas molculas no son detectadas, y esto se puede deber a dos factores: en primer lugar, el algoritmo no logr detectarlas en la etapa anterior al no ser suficiente el filtro de ruido aplicado para que fuera identificada como mximo local o, en segundo lugar, no super la sensibilidad dado que la variabilidad de la simulacin hizo que no fuera observable.

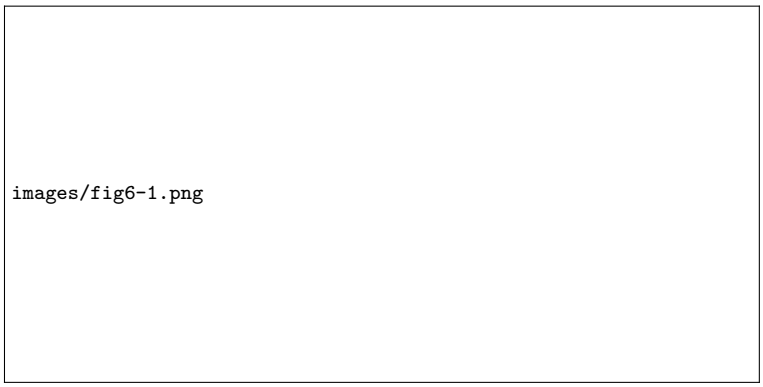
5.1. *Presence of Isotopes*

6. CONCLUSIONS

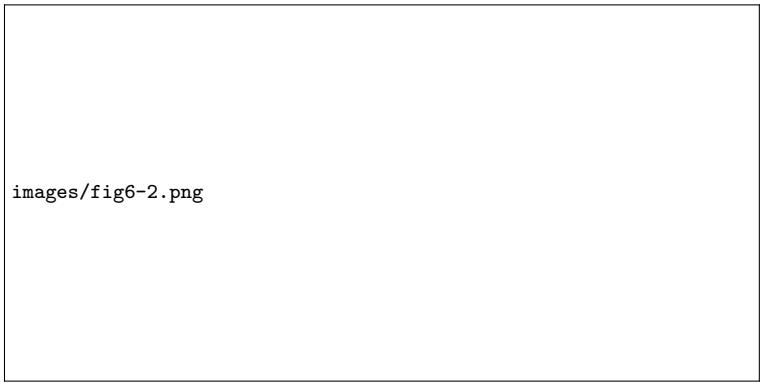
The sparse coding technique allows to identify much of the molecules, and even when it fails, the mismatch correspond to an isotope of the predicted molecule. With this, it is possible to give a basic notion of whick is the molecular composition of the astronomical objects simulated.

The data used to train the model did not included the high of the peaks observed in the spectra, given the limitation of the simulation. Those limitations reside in the inability to comply the theoretical relative heights of lines of the same isotope in a given spectra, assign the heights randomly.

This limitation was solved not using the height of the observed lines, but only the observed frequencies. So the solution consisted in the use of only the detected frequencies and assign for each isotope a value wich depends on



(a) Figura 6.1.



(b) Figura 6.2.

FIG. 8.— Rn la figura 8(a) se grafica una palabra terica y en la 8(b) su equivalente recalculado

the distance of the nearest theoretical frequency. Thus, the algorithm can relax the match between observed an theoretical frequencies.

The main difficulty of the current approach of this solution lies in that some molecules in some frequencies intervals are present with a large number of lines. As the words depends only on the observed frequency. Without consider the height, the importance of some observations is lost, besides it is a higher or a lower line, if it exist of it is just noise.

Therefore, an natural extension of this algorithm will be the incorporation of the height of the observed lines. With that information from future real data, will be possible to assign weights to avoid making this generalization of the importance of the observations. Also, one ratio between the relative heights of the lines of the same isotope help to have more representative words of reality.

REFERENCES

J. Cernicharo, B. Tercero, A. Fuente, J. L. Domenech, M. Cueto, E. Carrasco, V. J. Herrero, I. Tanarro, N. Marcelino, E. Roueff, M. Gerin, and J. Pearson. Detection of the Ammonium Ion in Space. *The Astrophysical Journal*, 2013.

Tom Howley, Michael G. Madden, Marie-Louise O Connell, and Alan G. Ryder. The Effect of Principal Component Analysis on Machine Learning Accuracy with High Dimensional Spectral Data. 2005.

Ohishi M Nummelin A, Bergman P, Hjalmarson A, Friberg P, Irvine W M, Millar T J and Saito S. A Three-Position Spectral Line Survey of Sagittarius B2 between 218 and 263 GHz. I. The Observational Data. *The Astrophysical Journal Supplement Series*, 1998.

Ohishi M Nummelin A, Bergman P, Hjalmarson A, Friberg P, Irvine W M, Millar T J and Saito S. A Three-Position Spectral Line Survey of Sagittarius B2 between 218 and 263 GHz. II. Data Analysis. *The Astrophysical Journal Supplement Series*, 2000.

T. R. Hunter D. C. Lis P. Schilke, J. Beneford and T. G. Phillips. A line survey of orion-kl from 607 to 725 ghz p. *The Astrophysycal Journal Supplement Series*, 2001.

A. J. Remijan. Splatalogue - Motivation, Current Status, Future Plans. 2010.

A. J. Remijan and A Markwick-Kemper. Splatalogue: Database for Astronomical Spectroscopy. 2008.

Peter Schilke, Rainer Rolffs, and Claudia Comito. Analysis tools for spectral surveys. *Proceedings of the International Astronomical Union*, 2011.