# Automatic Detection of Expanding H I Shells Using Artificial Neural Networks

ANIK DAIGLE AND GILLES JONCAS

Département de Physique, Université Laval, Québec, QC G1K 7P4, Canada; anik.daigle@phy.ulaval.ca, joncas@phy.ulaval.ca

MARC PARIZEAU

Département de Génie Électrique et de Génie Informatique, Université Laval, Québec, QC G1K 7P4, Canada; parizeau@phy.ulaval.ca

AND

MARC-ANTOINE MIVILLE-DESCHÊNES

Canadian Institute for Theoretical Astrophysics, 60 St. George Street, Toronto, ON M5S 3H8, Canada; mamd@cita.utoronto.ca

**ABSTRACT.** The identification of expanding H I shells is difficult because of their variable morphological characteristics. The detection of H I bubbles on a global scale has therefore never been attempted. In this paper, an automatic detector for expanding H I shells is presented. The detection is based on the more stable *dynamical* characteristics of expanding shells and is performed in two stages. The first one is the recognition of the dynamical signature of an expanding bubble in the velocity spectra, based on the classification of an artificial neural network. The pixels associated with these recognized spectra are identified on each velocity channel. The second stage consists of looking for concentrations of those pixels that were first pointed out and deciding if they are potential detections by morphological and 21 cm emission variation considerations. Two test bubbles are correctly detected, and a potentially new case of a shell that is visually very convincing is discovered. About 0.6% of the surveyed pixels are identified as part of a bubble. These may be false detections but still constitute regions of space with high probability of finding an expanding shell. The subsequent search field is thus significantly reduced. In the near future, we intend to conduct a large-scale H I shell detection over the Perseus arm using our detector.

## 1. INTRODUCTION

The study of neutral hydrogen bubbles is intimately related to our knowledge of the interstellar medium (ISM) and its different phases. The evolution of bubbles is indeed a function of parameters such as the filling factors of the cold and hot phases, the number and distribution of supernovae (SN) and Wolf-Rayet (WR) stars in the Galaxy, etc. In order to develop a significant knowledge of the parameters related to the physics of bubble expansion, one needs to analyze many observational cases. There are very few known expanding H I structures; small bubbles associated with individual stars are particularly unheard of because of the insufficient resolution of the H I surveys previous to the Canadian Galactic Plane Survey (CGPS; see § 2). However, the CGPS data have sufficient resolution for the study of those smaller objects. How common are those objects in reality? If they reveal themselves as rare, could we have to conclude, for example, the paucity of WR stars in the Galaxy? In a more general context, we believe that it would be interesting to trace and study small bubbles because they are isolated and therefore involve fewer energy injection processes. The study of such cases could thus lead to more secure conclusions and better constraints than the study of larger scale structures such as superbubbles, which are caused by multiple contributions and several stars over several hypothetical generations.

Since the morphological characteristics of H I bubbles are very different from one bubble to another, their detection on a morphological basis is difficult. On the other hand, their dynamical characteristics are much more stable; the expansion velocity of bubbles associated to individual stars (and of radii of a few parsecs) is generally $\sim$10 km s$^{-1}$ (Gervais & St-Louis 1999; Cappa & Herbstmeier 2000). Hence, the two opposite sides of the shell have a typical maximum relative velocity of 20 km s$^{-1}$. In the CGPS H I cubes, this results, because of the Doppler shift, in a characteristic signature detectable in the *velocity spectra* (see Fig. 1): two distinct peaks can be seen, one blueshifted and one redshifted, and separated by approximately 20 km s$^{-1}$. Our strategy is as follows: we seek first to detect expanding bubbles using their dynamical characteristics, hence by detecting the characteristic two-peaked profile in the velocity spectra associated with bubbles. Because this is a pixel-by-pixel approach, the bubble morphology is no more of prime importance, and we want to use it only afterward, as a further confirmation to the detection. The presence of two peaks separated by 20 km s$^{-1}$ in a velocity spectrum is indeed not sufficiently restrictive to make a reliable bubble detection; much structural noise—for example, velocity crowding on a line of
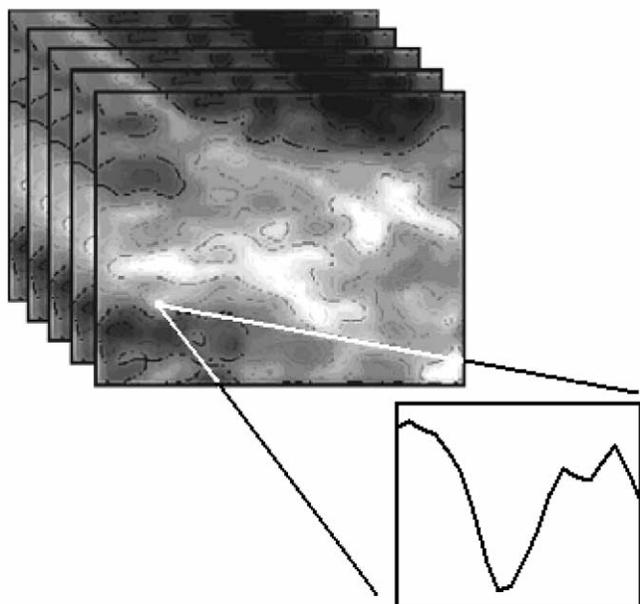
FIG. 1.—Velocity spectra extracted from a 21 cm data cube. Each velocity channel is an image of the H I gas at a given velocity.

TABLE 1
PARAMETERS OF THE CGPS 21 cm SURVEY
(FROM HIGGS 1999)

| Parameter | Value |
|---|---|
| Survey | |
| Number of synthesis telescope fields ...... | 190 |
| Field dimensions (deg) .................... | 2.5 |
| Spacing of field centers (deg) .............. | 1.86 |
| Spatial resolution ......................... | $1\rlap{.}'0 \times 1\rlap{.}'0 \csc \delta$ |
| Number of velocity channels .............. | 272 |
| Spectral resolution (km s$^{-1}$) .............. | 1.319 |
| Channel separation (km s$^{-1}$) .............. | 0.82446 |
| rms noise (field center) (K) ................ | 2.9 |
| Mosaics | |
| Number of mosaics ........................ | 36 |
| Size of mosaics (deg) ..................... | $5.12 \times 5.12$ |
| Overlap of mosaics (deg) ................. | 1.1 |
| Area covered (deg$^2$) ...................... | 666 |
| Galactic longitude coverage (deg) ........ | 74.2 to 147.3 |
| Galactic latitude coverage (deg) .......... | −3.6 to 5.6 |

sight or the presence of converging clouds—is likely to cause false shell detections. As will be seen, the morphology and other indicators are therefore used for confirmation. Also, we intend to first confine our survey to the Perseus arm ($l = 100°$–$140°$) in order to avoid the confusion in the velocity-distance relation existing for $l < 90°$ and $l > 270°$. This way, we believe that a detector pointing to regions of space where dynamical characteristics of expanding shells are found performs a valuable cleaning of the data and significantly reduces the subsequent search field.

## 1.1. Expanding Structures and the Models of the ISM

The so-called bubbles, shells, or cavities, referring to the ISM gas, are technically local regions that lack gas with boundaries that are neutral or ionized and of variable density. Because of their expansion and their often circular morphology, it is generally acknowledged that these objects result from a local momentum injection in the ISM. One of two important questions that can be asked refers to the nature of this energy source. The second question inquires about the effect those objects have on the structure of the ISM. Answering these questions is related more or less directly to the particular characteristics of the shells, which are deduced from observational data in a variety of spectral domains. Those characteristics are mainly the mass of the shell, its expansion velocity, its dimensions, and the density inside and outside of the bubble. Deriving those quantities from observational data allows one to compare them with hydrodynamical models of expanding bubbles in a given

medium, which can be elaborated in relation to any plausible energy injection hypothesis and models of the ISM.

Neutral hydrogen is a major component of the Galactic ISM ($M_{H_I} \approx 2 \times 10^9 \ M_\odot$). The pioneering work of Field, Goldsmith, & Habing (1969) opened the way to multiphase models of the ISM. However, their initial model did not take into account violent events such as SN explosions. Those were considered by McKee & Ostriker (1977), who introduced the hot and ionized phase of the ISM (hot ionized medium, or HIM). This three-phase model efficiently predicts the pressure, electron density, and velocity dispersion of the ISM clouds, as well as the intensity of the soft X-ray emission of the ISM. The values of the filling factors, however, are less certain. McKee (1995) and Cox (1995; Cox & Smith 1974) do not agree about the place taken by the HIM in the ISM; McKee believes that the HIM makes up the intercloud medium, while Cox believes that it has a small filling factor and is confined to dispersed bubbles. It is the determination of the HIM filling factor that will settle between the different ISM models, bringing the important link between the study of expanding shells and large-scale modeling of the ISM. The HIM filling factor can be determined only from observational data.

Four categories of hypotheses can be proposed to explain the existence of expanding structures: (1) the sweeping of the ISM by the winds of massive stars (Lozinskaya 1992; Oey, Clarke, & Massey 2001), (2) the progression of an ionization front followed by a recombination (McCray & Kafatos 1987), (3) one or several SN explosions (Tenorio-Tagle & Palous 1987), and (4) the collision of high-velocity clouds of neutral gas with the disk of the Galaxy. The choice of the right hypotheses is difficult but determinant, since the astrophysical implications regarding the different ISM models derived from the existence of expanding objects are completely different in
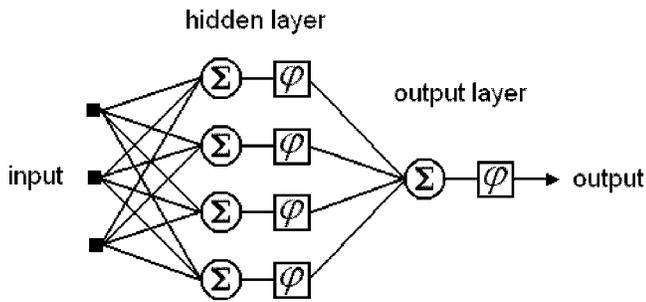
FIG. 2.—Schematic representation of a feed-forward neural network. To each connection is associated a synaptic weight (the free parameters). A neuron first computes the weighted sum of its synaptic connections before applying a usually nonlinear transfer function $\varphi$ to produce an output. The dimensions of the input and output spaces are at the user's discretion. In this paper, the input vector (the input data) is a 12 component segment of a velocity spectrum, and the output is a one-dimensional degree of membership to the class of expanding bubbles.

each of the four cases. There exists no complete and general theory explaining all the aspects of the stellar winds and SN interactions with the ISM. Nevertheless, there are some simplified models that are able to explain the observations (Elmegreen & Lada 1977; McCray & Kafatos 1987; Lozinskaya 1992).

What makes the choice of a model difficult is (1) that they are all incomplete and (2) that the observational data are too limited to supply enough typical cases to make a significant validation. This last point in itself clearly justifies our purpose: even if our automatic bubble detector could find only a few very regularly shaped objects, this would be an appreciable contribution to the discrimination of existing models of bubble formation and evolution. The high resolution of the CGPS makes it possible to visualize bubbles of radius down to a few parsecs in the Perseus arm. Those are especially ill represented in the current database and are thus not well studied even if they are less complex than large-scale structures. Considering the extent of the CGPS field (666 deg$^2$), the need for an automatic detector for expanding shells is obvious.

### 1.2. Proposed Strategy

Few papers have been published on the automatic detection of expanding shells (Mashchenko, Thilker, & Braun 1999; Thilker, Braun, & Walterbos 1998; Mashchenko & St-Louis 2002), and their aim was mostly on the comparison of spherical structures observed in the velocity channels of H I data cube, with structures predicted by hydrodynamical simulations of expanding bubbles in the ISM. They are therefore limited to morphological recognition.

Our approach is to recognize the *dynamical behavior* of an expanding bubble, characterized by its velocity spectra. The 21 cm data collected by the CGPS contain H I abundance *and* kinematical information; as mentioned earlier, a characteristic
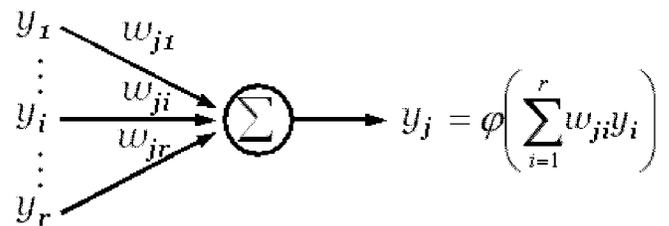


FIG. 3.—Neuron $j$ in the output layer sums the weighted outputs of the neurons $i = [1, r]$ of the preceding layer. The transfer function $\varphi$ is then performed to produce the neuron $j$ output $y_j$.

dynamical profile can in principle be found in all velocity spectra extracted from every pixel of the bubble. This detection technique avoids the difficulties linked to the often distorted morphology of the bubbles. Bubbles are indeed hardly ever spherical and are most often incomplete, which implies the subjectivity of a human observer and the difficulty of a purely morphological detection. Working with the velocity spectra allows a pixel-by-pixel evaluation of the likelihood of finding a bubble at a particular location in the data cube. Subsequent steps are to look for high concentrations of selected pixels in the data cube and finally to consider the morphology of the detection, which is no longer the principal visual characteristic of an expanding H I shell.

The presence of two peaks separated by 20 km s$^{-1}$ in a velocity spectrum is unfortunately not sufficiently restrictive to make a reliable bubble detection. As will be discussed in §§ 3 and 4, it is partly for this reason that we used an artificial neural network (ANN) for the first detection step. This tool is now often encountered in astrophysics, mostly as a classification device. We therefore present an overview of our ANN classifier in § 3 after the description of the CGPS data in § 2. In § 4, we expose the pixel-by-pixel detection method for the velocity spectra, while § 5 deals with the morphological validation of hypothesized bubbles. In § 6, the detector is performed on a simulated purely turbulent H I gas data cube. We present some preliminary results in § 7 before concluding in § 8.

## 2. THE DATA

Many 21 cm surveys of the Galaxy were accomplished in order to perform large-scale study of the ISM: Heiles & Habing (1974), Weaver & Williams (1973), Colomb, Pöppel, & Heiles (1980), and in particular the CGPS (Normandeau 1996; Taylor 1999; Higgs 1999). The CGPS aimed at systematically observing the Galactic plane, in the radio continuum and in the 21 cm H I line, between longitudes 75° and 145° and latitudes −3° and 5°. This project is the result of contributions from Canadian university researchers and is carried out at the Dominion Radio Astrophysical Observatory (DRAO).

The synthesis telescope is a seven-antenna interferometer oriented east-west and with a maximum length of 604 m. This allows an ~1′.0 resolution at 1420 MHz, while the previous
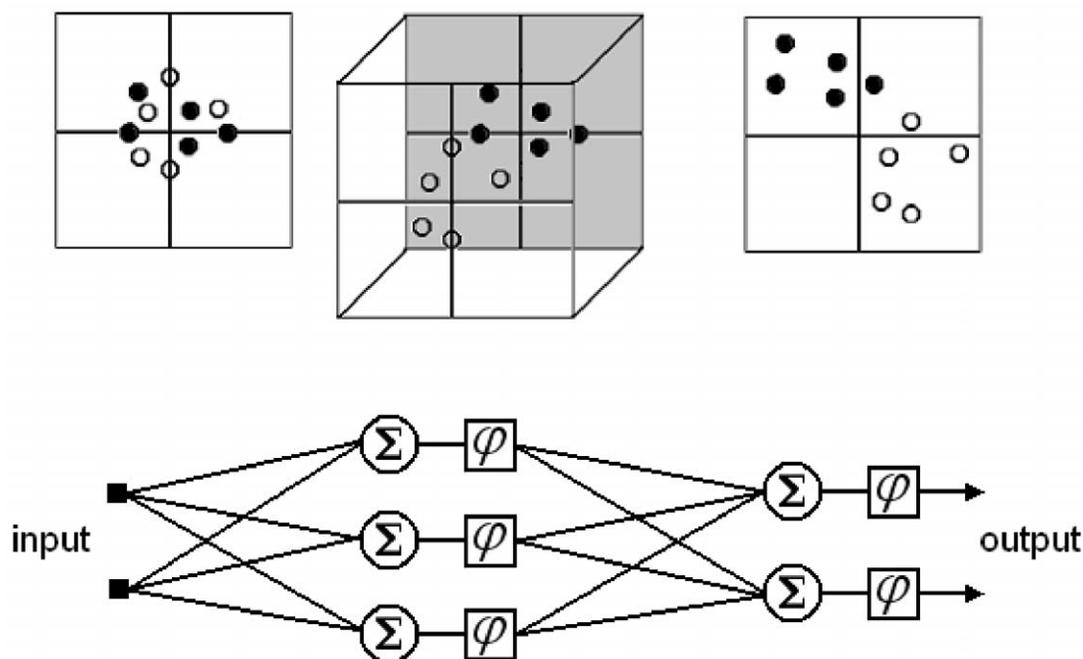
FIG. 4.—Mapping of data from an input space into an output space, in order to achieve a classification. The ANN three-neuron *hidden layer* acts as a three-dimensional intermediate space allowing the separation of intricate data.

large-scale H I surveys have resolution poorer than 4′. Each antenna is about 9 m in diameter, which gives a synthesized field of view of 2°.5 at 21 cm. Because of the 13 m minimum length of the interferometer, the minimum sampling resolution is ∼1° at 21 cm. A 26 m antenna was used to observe the larger scales up to the survey dimensions. The total field covers 666 deg² in 190 overlapped individual fields composing 36 5°.12 mosaics. As a result of the primary beam of each antenna, the attenuation of the signal increases with the distance to the field center. This and the partial overlap of the individual fields cause the noise to vary in the final images. For this reason, a noise-level map is supplied for each mosaic. Table 1 summarizes the parameters of the DRAO 21 cm H I data.

## 3. ARTIFICIAL NEURAL NETWORKS (ANNs)

An ANN is a network of many simple parallel operating elements (see Fig. 2). This configuration is inspired by the way biological neurons effectively communicate: neurons located in successive layers are massively interconnected by synaptic connections allowing electric signal transmissions. One neuron thus collects signals from many neighbors, and every signal is weighted by the efficiency of the synaptic connection that is involved. Learning can be defined technically as the process of reinforcement and inhibition of those connections, in order to create a neural network appropriate to the task to be learned. In the case of ANNs, a free parameter called the *synaptic weight* stands for the connection efficiency.

A schematic representation of an example of a feed-forward

neural network is shown in Figure 2. This network has a three-dimensional input space, one hidden layer containing four neurons, and its output layer contains one neuron—the output space is one-dimensional. The neurons are represented as summation signs because they compute the weighted sum of their synaptic connections. To each connection is associated a synaptic weight, which is an adjustable free parameter. The dimensions of the input and output spaces, as well as the hidden layer's dimensions, must be determined by the user.

### 3.1. The Back-Propagation Algorithm

ANN learning consists in the adjustment of all the synaptic weights of all the connections until a satisfactory mapping is achieved. The so-called supervised learning of an ANN consists in feeding the network with a training set of input-output mappings $[x(n), d(n)]$, where $d(n)$ is the desired output for the $n$th input vector $x(n)$. After each presentation, the ANN error is computed and is used to adjust the free parameters, so that the error will be minimized if the same situation is encountered. The learning process consists of adjusting all the weights, so that at each iteration the network heads for the direction of the more negative gradient on the error surface. The usual learning algorithm used to achieve this iterative minimization is called the back-propagation algorithm, because the error at the network output is propagated backward through the layers in order to calculate the proper adjustment for each weight in each layer.

The back-propagation algorithm updates the weights for each neuron, beginning with the output layer. Let $e_j(n)$ be the ob-
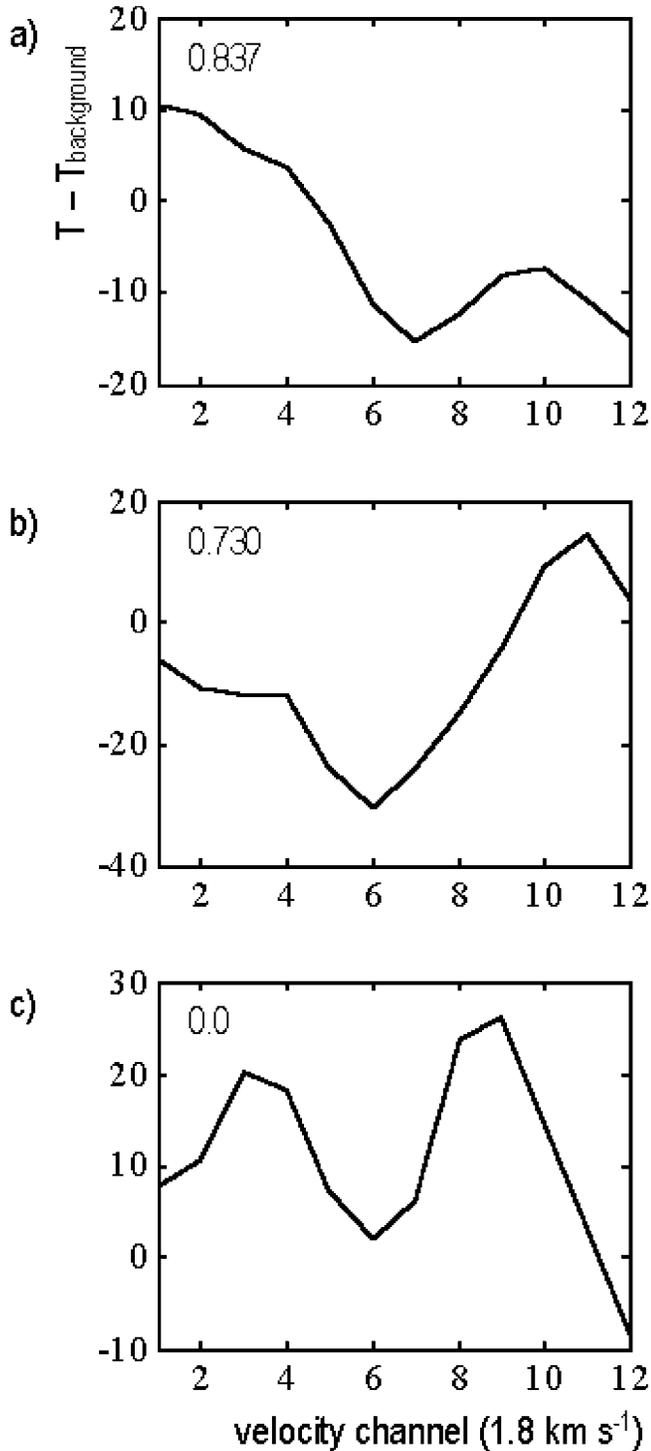
served error at the output neuron $j$ for the training datum $n$:

$$e_j(n) = d_j(n) - y_j(n), \tag{1}$$

where $d_j(n)$ is the desired output for neuron $j$ and $y_j(n)$ is the observed output. The goal is to minimize $E(n)$, the sum of the mean square errors (MSE) observed on the set of $C$ output neurons for the training datum $n$:

$$E(n) = \frac{1}{2} \sum_{j \epsilon C} e_j^2(n). \tag{2}$$

The output of neuron $j$ is defined by

$$y_j(n) = \varphi[v_j(n)] = \varphi\left[\sum_{i=1}^{r} w_{ji}(n)y_i(n)\right], \tag{3}$$

where $\varphi[\ldots]$ is the neuron transfer function (a usually nonlinear function in order to produce a nonlinear mapping), $v_j(n)$ is the weighted sum of the neuron $j$ inputs, $w_{ji}(n)$ is the connection weight between the neuron $i$ in the preceding layer (containing $r$ neurons) and the neuron $j$ in the output layer, and $y_i(n)$ is the neuron $i$ output (see Fig. 3).

To minimize $E(n)$, the weight $w_{ji}(n)$ must be updated in the direction in which the error gradient $\partial E(n)/\partial w_{ji}(n)$ is decreasing:

$$\Delta w_{ji}(n) = -\eta \frac{\partial E(n)}{\partial w_{ji}(n)}, \tag{4}$$

where $0 \leq \eta \leq 1$ is the learning rate. The chain rule gives

$$\Delta w_{ji}(n) = -\eta \frac{\partial E(n)}{\partial w_{ji}(n)} = \eta \delta_j(n)y_i(n), \tag{5}$$

$$\delta_j(n) = e_j(n)\phi_j', \tag{6}$$

where $\delta_j(n)$ is the local gradient of the error surface.

The case of the hidden layers is similar, but the local gradient of the neuron $j$ is now function of all the local gradients of the neurons $k$ in the layer following the one whose neuron $j$ is being updated:

$$\delta_j(n) = \phi_j' \sum_{k \epsilon C} \delta_k(n)w_{kj}(n). \tag{7}$$



Fig. 5.—Three velocity spectra. The MLP output is indicated on each graph. (*a*) Spectrum effectively positive and detected as such (belongs to GSH 138−01−094); (*b*) negative spectrum, insufficient MLP output; (*c*) negative spectrum, insufficient MLP output.
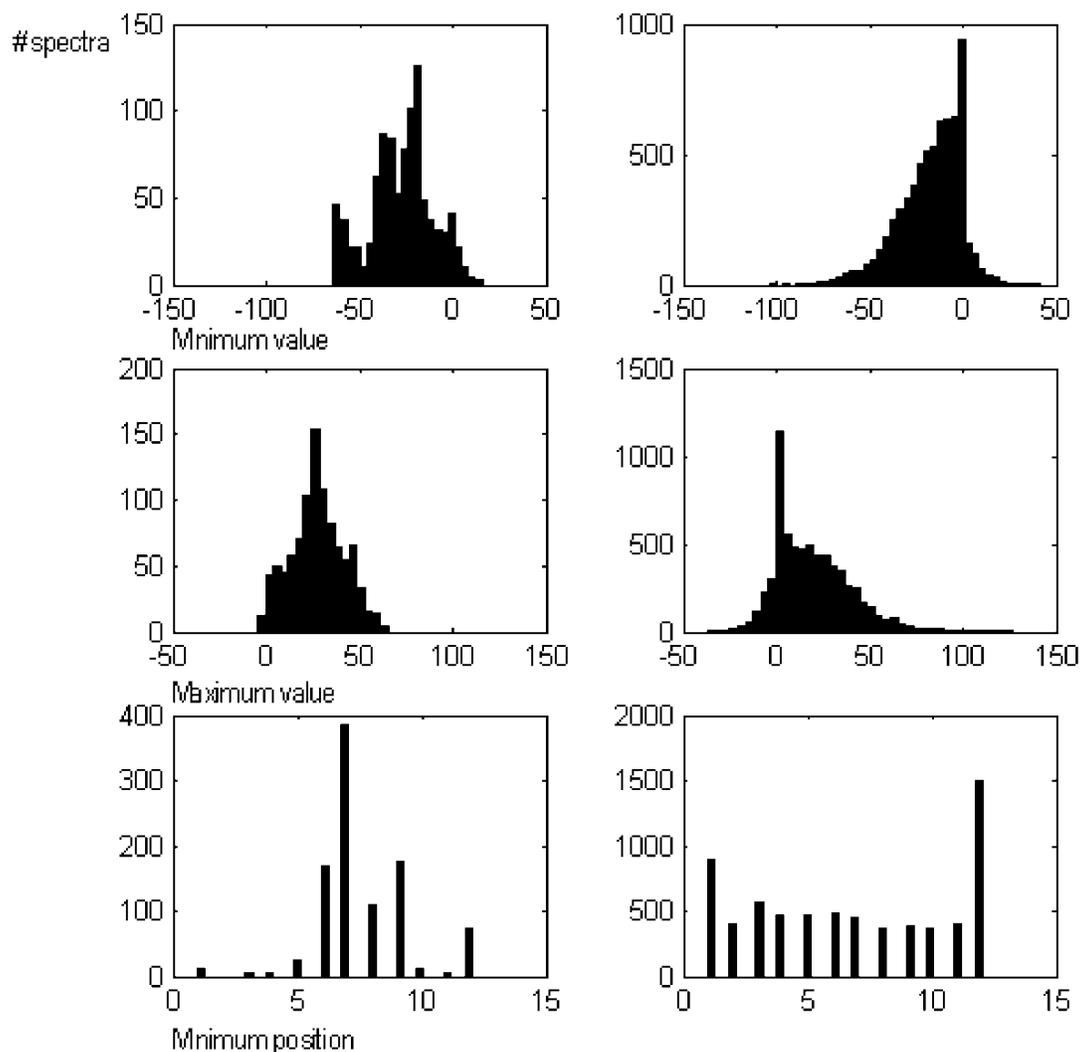
Fig. 6.—Histograms illustrating the statistics of three visual features in positive (*left*) and negative (*right*) spectra sets. *Top:* Value of the minimum; *middle:* value of the maximum; *bottom:* position of the minimum.

## 3.2. Classification with ANNs

An ANN may be described as an algorithm that maps data from an input space into a user-chosen output space. Among other possibilities, mapping a data set into an output space having meaningful categories is equivalent to classification. Figure 4 illustrates the classification of data by mapping the input set into a higher dimensional intermediate space, permitting the separation of two categories. The data are afterward mapped into the output space as separated sets or classes. It is this particular application of the ANNs that we want to take advantage of for this pattern-recognition task in the velocity spectra.

One interesting thing is the generalization capability of the ANN, which can interpolate between new and unknown input data. This generalization ability, as well as the speed of the convergence and the final ANN global performance, crucially depends on the training set. For all the eventual input data to be properly represented in the output space, the training set must cover most of the space. It must include varied examples and cover equally the different cases that may be encountered while using the trained ANN. A very common case in the training set will be very well delimited in the output space, while a rare one could be associated to some other class or be bluntly treated as noise.

The main distinction between ANNs and more classical classifiers, such as cross-correlation, is thus the determination of the classification criteria. A classical classifier needs precise criteria that are sufficient to make a distinction between classes. Those criteria are initially known and chosen by the user, hence they must have a known signification or at least be empirically
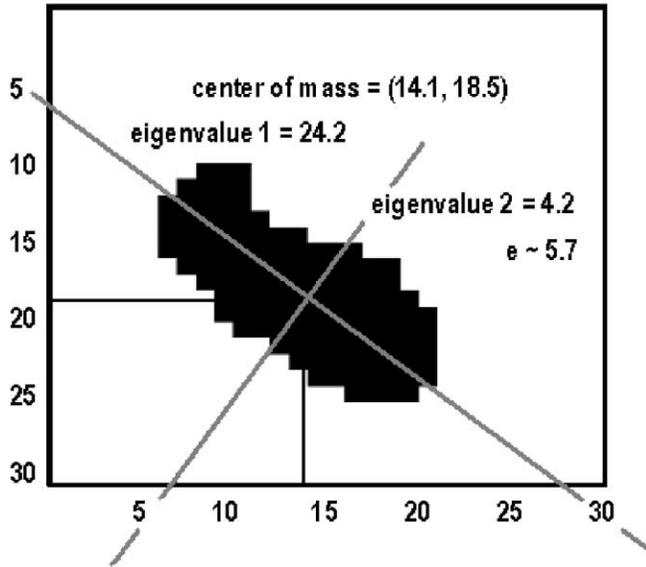
Fig. 7.—Characterization of a blob by its association with an ellipse. The ellipse eccentricity is estimated by the ratio of the eigenvalues of its two principal axes.

identified and accessible. To use an ANN classification, one requires only a good and representative training set. The ANN will, in a way, choose its classification criteria by itself, by adjusting its synaptic weights. The only known conditions are the ANN configuration (number of layers, number of neurons, transfer functions) and the initial values of the free parameters. The only a priori "criterion" is the error minimization algorithm. What is obtained is therefore a totally empirical classification, with no a priori statistical model.

This last point may cause some uneasiness: the fact that the ANN chooses the distinction criteria may be seen as a loss of control on the part of the user. An ANN could choose criteria that are entirely irrelevant to the user's purpose, and consequently it would be unable to process other data than the ones of the training set. On the other hand, when intuitive criteria are not available or are not sufficiently restrictive to achieve a good classification, an ANN may be able to isolate hidden characteristics and use them to make relevant decisions. Moreover, its generalization ability allows the recognition of distorted and noisy data, which a classical classifier might not be able to process.

These last two points justify our use of an ANN for the detection of the kinematic signature of expanding H I bubbles. Another classifier could have been used, but based on uncertain classification criteria: for example, cross-correlation would require a spectral template that we do not possess. Indeed, the search of a profile with peaks separated by 20 km s$^{-1}$ is not sufficiently restrictive. Our detector is thus required to find other criteria to recognize the bubbles' velocity spectra. Furthermore, the data are noisy; in addition to the detection noise,

we have to deal with structural noise: the expanding shells evolve in an H I environment made of clouds of various sizes, filaments, and other various structures that disturb the spectra. We believe that an ANN classifier may be more likely than any other classifier to integrate a sufficiently general and restrictive model of an expanding bubble velocity profile, in order to recognize it amongst all other structures, and despite the noise.

There are several possible ANN configurations. We chose a multilayer perceptron (MLP), which is a feed-forward neural network such as the one described by Figure 2, because it proved its value in many practical classification problems and also because it does not make any a priori assumption about data distribution, contrary to other statistical classifiers. The main shortcoming of the MLP in this particular context is that it may take decisions about regions of space that were not covered by the training set. The MLP will not reject a case pleading incompetence. Independent rejection criteria allowing one to discard some of the MLP decisions will be described in §§ 4 and 5.

The main difficulty of our proposal is about the training set. The small number of known certified observational bubbles makes it difficult to assemble a set of diversified and representative examples. This certainly requires special attention, as will be discussed later.

For more details about the theory and implantation of ANNs, the reader can consult Haykin (1998), Demuth & Beale (2000), and Principe, Euliano, & Lefebvre (2000).

## 4. DYNAMICAL DETECTION

The detection of the dynamical characteristics of a bubble in the velocity spectra is a pixel-by-pixel one; the goal is to assign a degree of confidence to each pixel, expressing how much the corresponding velocity spectrum matches the dynamical profile that is sought. This first step is taken care of by an MLP with two hidden layers (the layers between the input and output layers) containing 17 and two neurons, respectively. The error minimization is performed by a Levenberg-Marquardt back-propagation algorithm, which is a faster variation of the standard back-propagation algorithm explained in § 3 (Hagan & Menhaj 1994). The learning rate is 0.17. Those conditions allow to reach a good MLP performance (MSE ≤ 0.014) in a relatively short period of time. The MATLAB Neural Network Toolbox has been used.

Too long a training can compromise the MLP generalization capability: the worst limit of the MLP is that it will learn to recognize each individual training spectrum and be unable to recognize any other similar spectrum that is unknown. A standard cross-validation technique that averts overtraining is the use of a validation set, which is a data set composed as the training set, but that is not used for the training itself. The MLP performance is computed for each iteration over the validation set, and the training is stopped when the errors over the val-

TABLE 2
Evolution of the False Detections through the Detection Stages

| | GSH 138−01−094 | | G73.4+1.55 | |
|---|---|---|---|---|
| STAGES OF THE DETECTION | False Detections (pixels) (%) | Number of Preserved Blobs | False Detections (pixels) (%) | Number of Preserved Blobs |
| Threshold of 0.8 ............... | 2.3 | 375 | 3.0 | 475 |
| After filtering the spectra ...... | 1.9 | 305 | 2.7 | 456 |
| After filtering the blobs ........ | 0.5 | 29 | 0.7 | 57 |

idation set begin to rise. The training set contains ∼1300 spectra, and the validation set contains ∼200. This ratio is standard.

The vectors presented at the input layer of the MLP are 12 channel segments of the velocity spectra. This number of channels spans ∼20 km s$^{-1}$, a velocity interval sufficient to contain the dynamical profile of the bubble. The background (an average spectrum over the whole data cube) is subtracted beforehand, in order to filter the irrelevant global characteristics such as the local diffuse material associated with the spiral arm. Only the local H I enhancements and dynamical characteristics (such as an expanding bubble) are preserved. The MLP output is a number in the range [0, 1] expressing how well the 12 channel segment matches the dynamical profile of an expanding bubble. Figure 5 shows examples of spectra classified by the MLP. Even if the three examples show a two-peaked profile, the MLP output is different for each. Some information that is not directly accessible or intuitive, but that allows some discrimination, is therefore extracted by the MLP.

The results presented in this paper were obtained with a committee of four MLPs. We simply averaged the results coming from each of the four networks. The MLPs were trained over velocity spectra taken from six different data cubes, each of which contained one known bubble. Two of those bubbles are artificial bubbles; spheres of relatively high brightness temperatures have been added in real and otherwise unaltered H I data cubes. This is a way to enrich and improve the training set by making it more representative; even though those two bubbles are not real, inserting them into the training set introduces representative dynamical characteristics inside various H I environments and thus contributes to enrich knowledge of the MLP. The four real bubbles are G132.6−0.7−25.3, the bubble related to WR 144 (Galactic coordinates 80.04+0.93−26.2), the bubble related to WR 139 (76.6+1.43−69.07), and the bubble related to WR 149 (89.53+0.65−66). The MLP performance will naturally get better as the training set will be enriched with new bubble instances.

As mentioned in § 3.2, an MLP is not well suited for rejecting data; it will output a decision for every input datum, even if it is unable to make a proper classification given the experience acquired during its training. The MLP taken alone may thus make an important number of false detections: generally, 2%–10% of the pixels are incorrectly identified as "bubble" in the whole data cube. A filtering is therefore performed by verifying a number of conditions in each selected spectrum. By comparing our known examples of positive spectra (spectra belonging to known bubbles), we have been able to isolate some discriminant features most of them have in common. However, the negative spectra may have every possible shape and occasionally present some of these features. For this reason, we rather identify *rejection criteria*, allowing us to exclude a detection made by the MLP. These conditions are supplied by statistics compiled over all the positive cases we have. Those characteristics are very pragmatic, such as the values and positions of the spectrum extrema. Applying those rejection criteria on the MLP detections allows us to exclude a significant part of the false detections but keeps most of the correct ones.

Figure 5*a* illustrates a typical positive spectrum that shows some positive visual characteristics:

1. The minimum position is between channels 6 and 9.
2. The minimum value is less than −12 (negative values are due to the background subtraction).
3. The maximum value is greater than zero.

The approximately central position of the minimum is a quite intuitive feature that depends directly on the dynamics of expanding bubbles. However, the values of the extrema (which are brightness temperatures relative to the average background) depend on physical conditions that are less stable, such as the ratio between the shell and ambient medium densities. In fact, the rejection conditions were chosen on a purely empirical basis: we noticed only that they could make some distinction between a positive set and a negative set. Figure 6 shows the statistics of the three characteristics stated above, to visually compare positive and negative sets. One can notice some trend for the positive sets. If some permitted intervals are established for each condition, an important part of the false detections can be rejected. In the examples presented in § 7, we simply established lower and upper limits such that 88% of the spectra of the positive histograms are preserved.

## 5. MORPHOLOGICAL DETECTION

We believe that pixel-by-pixel dynamical detection is helpful to begin the search because of the difficulties arising from a morphological detection in several successive velocity chan-
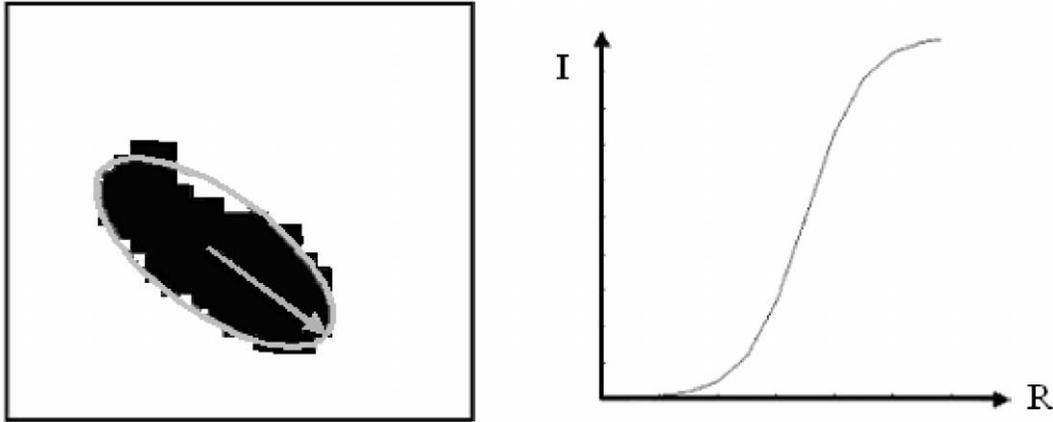
Fig. 8.—Increasing intensity criterion: the H I abundance must increase with the bubble radius. What is effectively verified is that the pixels' intensity on the ellipse principal axis increases as we progress away from the blob center. This "center" is not the center of mass but the minimum of the intensity neighboring the center of mass.

nels. This first step taken, we are left with a cube in which patches of positively identified pixels are found. In a single velocity channel, this results in connected objects, or "blobs," of various dimensions and shapes.

### 5.1. The Dimensions of the Object

The object dimension in pixels depends on its intrinsic dimension and on its distance. As a velocity channel is associated to a Galactocentric distance, it is possible to establish a lower limit, a function of the velocity channel, to the number of pixels corresponding to a blob of a few parsecs in diameter, and thus necessary for a blob to be a real bubble. The higher limit for the bubble's size is indirectly determined by the dynamically based detection, since the maximum expansion velocity for bubbles of a few parsecs is typically 10 km s$^{-1}$.

### 5.2. The Object Shape

One should keep in mind that the blobs with which we now have to work are in fact *concentrations of positive detections*; that is to say, their shapes do not necessarily match the visual shape of the bubble. Despite this fact, it is clear that a concentration of detections having a roughly circular and condensed shape is more likely to be a bubble than some very elongated and sharp filament. It is thus relevant, at this point, to take morphology into account.

To begin with, the size and orientation of the object are characterized using a principal component analysis: the directions of the two principal axis of the blob are calculated, and the eccentricity of an associated ellipse is estimated by the ratio of the eigenvalues of the axis (see Fig. 7). A new rejection criterion is then defined on the basis of a superior limit to the eccentricity.

### 5.3. Increasing Intensity Criterion

Because an H I bubble corresponds to a local lack of gas, another rejection criterion for a given velocity channel, directly linkable to the physics of the phenomenon and independent of the dynamics or morphology, is the fact that the intensity of the 21 cm emission should increase in a radial fashion around its center (see Fig. 8).

## 6. TURBULENCE: SIMULATION OF SELF-SIMILAR H I IN THE PERSEUS ARM

As mentioned earlier, the structural noise due to various structures in H I may cause false detections. Apart from the rejection criteria and the morphological confirmation just exposed, there is not much to be done against it. It is however of interest to verify whether our detector could be tricked by turbulence alone. We performed our detector on a simulated purely turbulent H I gas data cube such as described below. The subsequent rejection criteria on the velocity spectra were not used, because the brightness temperatures relative to the average background could not be consistent in real and simulated cubes. The committee of MLPs does retain some spectra as positive, but those detections do not pass the morphological test; in fact, no blob at all can be isolated, the maximum number of connected pixels in a concentration of positive detections being 18 ($\sim\frac{1}{4}$ of minimum number accepted by the detector).

The rest of this section describes how we have simulated the spectro-imagery observation of the observed part of the Perseus arm on the hypothesis that the H I density structure and kinematics are due only to turbulence and Galactic rotation. In this section, we describe how we modeled the H I three-dimensional density and velocity fields and how we constructed the spectro-imagery observation from these.
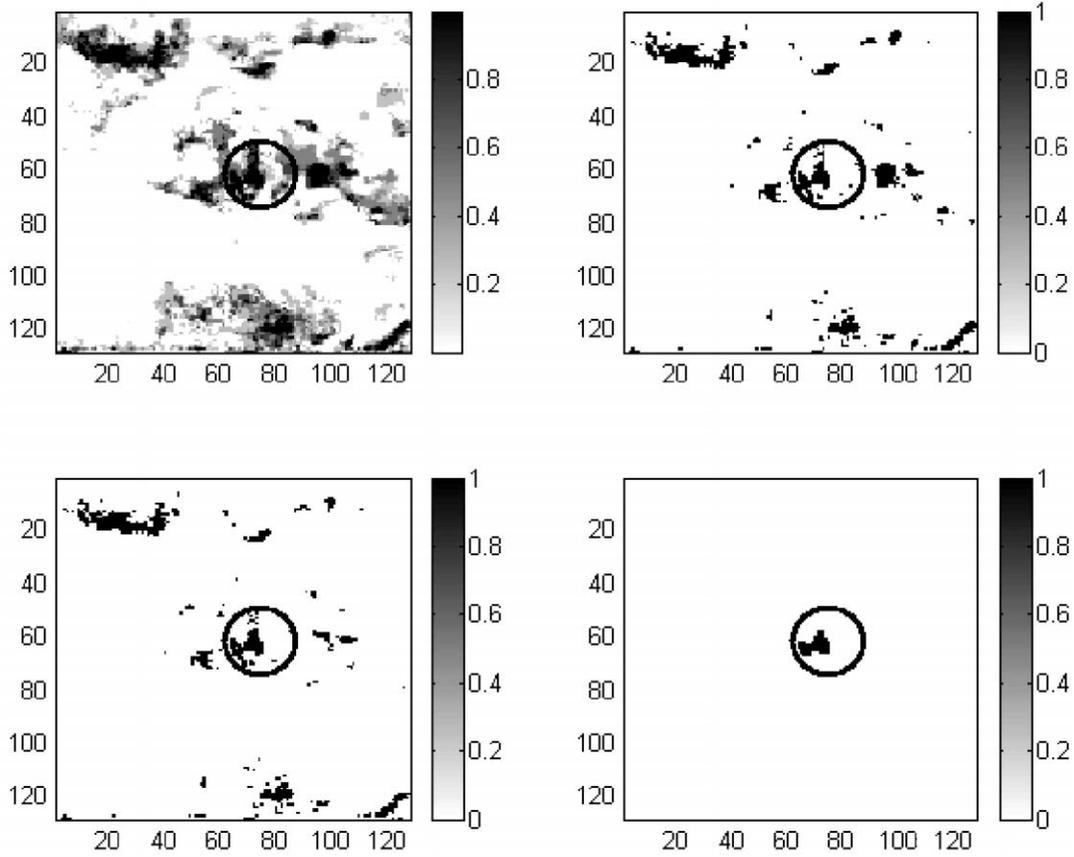
Fig. 9.—Results of the detection of GSH 138−01−094. The ellipse indicates the bubble location and scale at its maximum expansion. *Top left:* MLP detection; *top right:* after a 0.82 threshold; *bottom left:* after application of the rejection criteria on the spectra; *bottom right:* after application of the rejection criteria on the blobs.

We consider that the center of the Perseus arm is at a distance of 2 kpc from the Sun and that its depth is 0.3 kpc. As the field observed at DRAO is 2°.6 × 2°.6, this translates into a surface of ~0.1 × 0.1 kpc at the Perseus arm distance. To simulate the Perseus arm, we have produced a 128 × 128 × 384 pixel box that represents the observed portion of the Perseus arm (0.1 × 0.1 × 0.3 kpc). Each cell in that box has a linear size of 0.75 pc.

## 6.1. The Three-dimensional Velocity Field

The simulated three-dimensional velocity field [$v(x, y, z)$] is the sum of the Galactic rotation component [$V(x, y, z)$] and a turbulent component [$\delta v(x, y, z)$]. To lighten the text, we drop the ($x, y, z$) in the following.

### 6.1.1. The Galactic Rotation Velocity Component

To simulate the Galactic rotation velocity component, we proceeded as follows. Each point in the simulated cube has been attributed a longitude $l$, latitude $b$ (corresponding to our

DRAO observation), and a distance from the Sun (the center of the box being at 2 kpc). The Galactic radius $R$ of each cell element is then given by

$$R = \frac{D \cos (l) - R_{\odot}}{\cos (\theta)},\qquad(8)$$

where

$$\theta = \tan^{-1}\left[\frac{D \sin (l)}{D \cos (l) - R_{\odot}}\right],\qquad(9)$$

with $R_{\odot} = 8.5$ kpc.

Following Burton (1992), the Galactic angular velocity $\Omega$ has been estimated to be

$$\Omega = \Omega_{\odot}[1.0074(R/R_{\odot})^{0.0382} + 0.00698],\qquad(10)$$

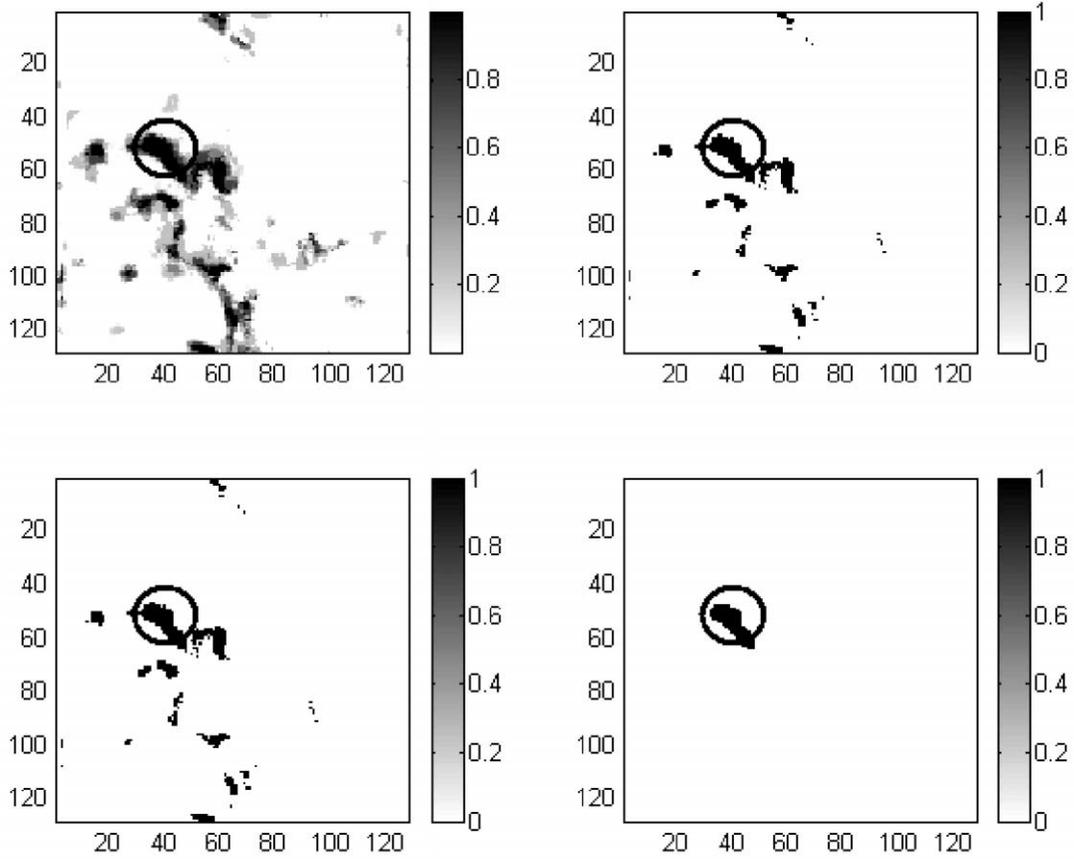with $\Omega_{\odot} = 220$ km s$^{-1}$. Finally, the angular velocity compo-

FIG. 10.—Same as Fig. 9, but for G73.4+1.55.

nent along the line of sight is given by the following expression:

$$V = \left(\frac{R_\odot}{R}\,\Omega - \Omega_\odot\right)\sin\,(l)\,\cos\,(b). \qquad (11)$$

### 6.1.2. The Turbulent Velocity Component

To simulate the self-similar properties of the velocity structure, we have used three-dimensional fractional Brownian motion simulations (Stutzki et al. 1998; Miville-Deschênes, Levrier, & Falgarone 2003). We have simulated a 128 × 128 × 384 pixel cube with spectral index −11/3, to simulate Kolmogorov-type turbulence. The average of the turbulent velocity field was set to zero ($\langle\delta v\rangle_{x,y,z} = 0$), and its dispersion was set to 20 km s$^{-1}$.

### 6.2. The Density Field

The simulated three-dimensional density field [$n(x, y, z)$] was also built using a fractional Brownian motion simulation with a spectral index of −11/3. No correlation between velocity and density was introduced. We have subtracted the minimum value

from the three-dimensional density cube to allow only positive density values. Because of the relatively small size of the observed region, we did not introduce any variation of the average density with the height above the Galactic plane.

### 6.3. Building the Spectro-Imagery Observation

The spectro-imagery observation (also called position-position-velocity [PPV] cube) has been constructed by making the assumption that every gas cell on the line of sight emits a thermally broadened Gaussian [$\sigma = (k_BT/m)^{1/2}$] centered at $v(x, y, z) = V(x, y, z) + \delta v(x, y, z)$ and of amplitude $n(x, y, z)$ (see Miville-Deschênes et al. 2003 for details). We have considered optically thin cold neutral medium gas at $T = 100$ K. We have constructed a 128 × 128 × 128 pixel PPV cube that gives the column density on each line of sight and at each velocity $u$ (velocity bin is 1.65 km s$^{-1}$) using the following equation:

$$N_H(x, y, u) = \alpha \sum_z n(x, y, z) \exp\left\{-\frac{[u - v(x, y, z)]^2}{2\sigma^2}\right\}.$$
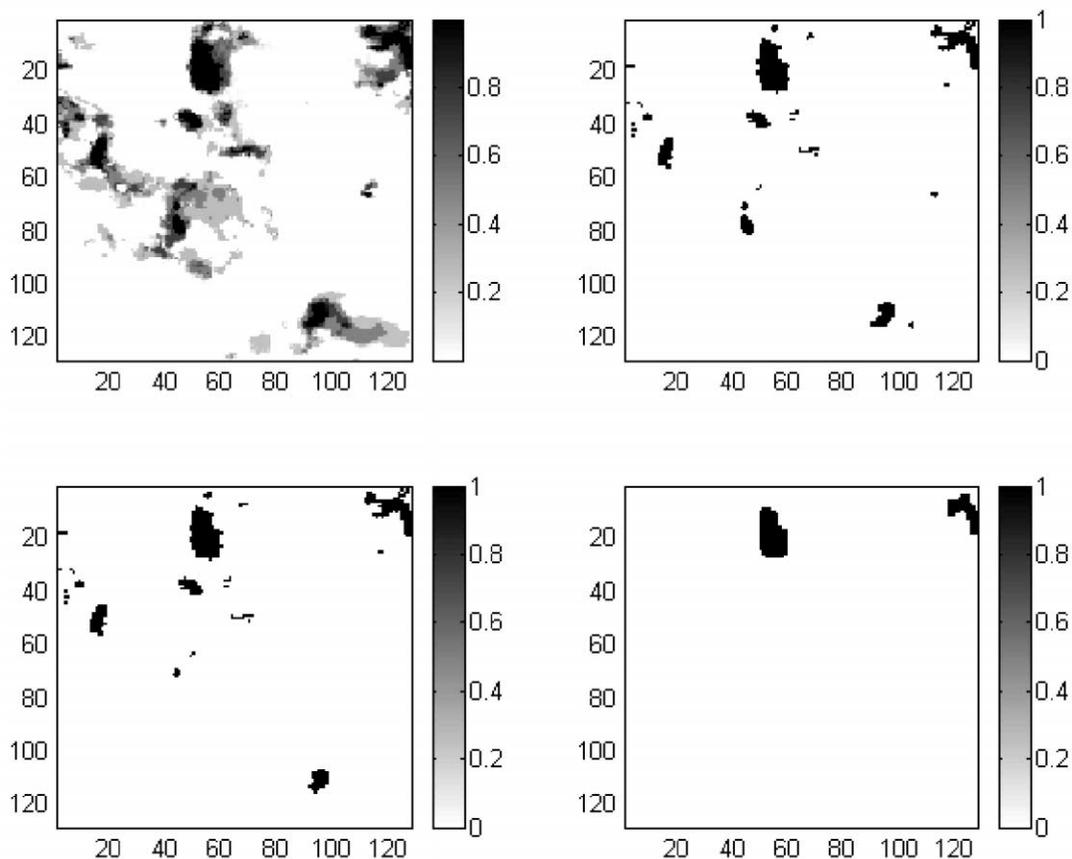
$$(12)$$

FIG. 11.—Detection at coordinates $(l, b, v) = (138°, -0°.5, -7)$. Corresponds to the location of an unknown bubble, at least visually convincing in its morphology as well as in its dynamics.

Here $\alpha$ is a normalization factor that takes into account the spectral resolution and the linear size of a cell along the line of sight.

## 7. PRELIMINARY RESULTS

The two test cubes have Galactic coordinates $(\Delta l; \Delta b; \Delta v)$ = $(136°.75, 139°.30; -0°.10, -2°.65; +12.6, -136.7)$ and $(\Delta l; \Delta b; \Delta v) = (73°.1, 75°.7; +0°.25, +2°.9; +60.7, -102.9)$ and contain, respectively, bubbles GSH $138-01-094$ and G73.4+1.55. Those two bubbles were not used for the MLP training, therefore the detector possesses no preliminary knowledge of those particular cases. It is believed that GSH $138-01-094$ could have been caused by an SN (Stil & Irwin 2001), while G73.4+1.55 is linked to WR 134 (Gervais & St-Louis 1999).

It took a few hours to train the MLPs, and their MSE on the validation set is $\leq 0.014$. The detection itself on a cube of $128 \times 128 \times 79$ 12 channel spectra takes ~5 minutes with a 1.9 GHz Pentium 4 and 512 Mbytes RAM.

Figure 9 shows the velocity channel corresponding to $v \approx -94$ km s$^{-1}$, where GSH $138-01-094$ should be detected, through the stages of the detection. Figure 10 shows the same

stages for G73.4+1.55, at $v \approx -11.4$ km s$^{-1}$. Table 2 traces the evolution of the amount of false detections for each step, in pixel percentages and in number of preserved blobs.

Figure 11 shows a detection at $(l, b, v) \sim (138°, -0°.5, -7)$ in the same data cube where GSH $138-01-094$ is found. The presence of a bubble is visually very convincing, as much in its morphology through successive velocity channels as in its dynamical signature in the velocity spectra. It would certainly be relevant to check the eventual presence of a WR star or a supernova remnant at this location.

### 7.1. Discussion

Of the detections, 0.5%–0.7%, or fewer than 30–60 blobs, do not correspond to any known bubbles. It is, however, tempting to verify those false detections in the hope of finding out that some of them were justified. What we want to point out is that, even if our detector gives several false alarms, those very limited and confined regions of the sky constitute a search field considerably reduced compared to the whole sky. The purpose of our detector is not to make a final decision about the validity of a detected bubble, but to raise flags in order to

guide the researchers toward regions of the sky with a relatively high detection probability.

## 8. CONCLUSIONS

The next step to this work is to use the detector on a large section of the CGPS H I data. The eventual new cases of bubbles found using the detector will be integrated into the MLP training set in order to improve its specialization. For the first time, an evaluation of the amount and distribution of expanding bubbles of a few parsecs radius will be possible.

Such an evaluation is critical to evaluate the relative importance of the HIM in the ISM, which is traditionally addressed in terms of the porosity parameter $Q$. This parameter is defined as the ratio between the volume occupied by the bubbles in the Galaxy and the total volume of the Galaxy; $Q$ is thus directly related to the HIM filling factor, if we suppose that the HIM fills the bubbles (McKee & Ostriker 1977). For example, Oey & Clarke (1997) deduced $Q$ from an analytical expression of the distribution and size of bubbles and superbubbles. If their prediction for a maximum in the H I cavities' radii distribution was verified, this would constitute a new constraint for the SN remnant's evolution, for the ISM conditions, and for the typical energy injected by SNe in galaxies. However, an observational verification of this prediction in the small-radii regime is not possible because of the lack of known cases. Our results could help to clarify this issue.

In this paper, it has been shown that a detector using the dynamical features in the velocity spectra of H I data cubes can point out zones where there is a relatively high probability of finding an expanding bubble. This detector is based on the classification of an MLP and on some independent dynamical and morphological rejection criteria. The test bubbles have been correctly detected with a false-detection percentage of ∼0.6%, which significantly reduces the subsequent search field. This preliminary cleaning could be followed by a human inspection or other more complex automatic analysis processes, which could be applied on a few objects rather than a vast amount of rough data.

## REFERENCES

Burton, W. B. 1992, in The Galactic Interstellar Medium, ed. P. Pfenniger & P. Bartholdi (Heidelberg: Springer), 1
Cappa, C. E., & Herbstmeier, U. 2000, AJ, 120, 1963
Colomb, F. R., Pöppel, W. G. L., & Heiles, C. 1980, A&AS, 40, 47
Cox, D. P. 1995, in ASP Conf. Ser. 80, Problems with the Diffuse Interstellar Medium, ed. A. Ferrara, C. F. McKee, C. Heiles, & P. R. Shapiro (San Francisco: ASP), 317
Cox, D. P., & Smith, B. W. 1974, ApJ, 189, 105
Demuth, H., & Beale, M. 2000, Neural Network Toolbox, for Use with MATLAB, User's Guide, Version 3.0 (Natick: Mathworks)
Elmegreen, B. G., & Lada, C. J. 1977, ApJ, 214, 725
Field, G. B., Goldsmith, D. W., & Habing, H. J. 1969, ApJ, 155, 149
Gervais, S., & St-Louis, N. 1999, AJ, 118, 2394
Hagan, M. T., & Menhaj, M. 1994, IEEE Trans. Neural Networks, 5, 989
Haykin, S. S. 1998, Neural Networks, a Comprehensive Foundation (2d ed.; Upper Saddle River: Prentice Hall)
Heiles, C., & Habing, H. J. 1974, A&AS, 14, 1
Higgs, L. A. 1999, in ASP Conf. Ser. 168, New Perspectives on the Interstellar Medium, ed. A. R. Taylor, T. L. Landecker, & G. Joncas (San Francisco: ASP), 15
Lozinskaya, T. A. 1992, SNe and Stellar Wind in the Interstellar Medium (New York: AIP)
Mashchenko, S., & St-Louis, N. 2002, in ASP Conf. Ser. 260, Automatic Shell Detection in CGPS Data, ed. A. F. J. Moffat & N. St-Louis (San Francisco: ASP), 65

Mashchenko, S., Thilker, D. A., & Braun, R. 1999, A&A, 343, 352
McCray, R., & Kafatos, M. 1987, ApJ, 317, 190
McKee, C. F. 1995, in ASP Conf. Ser. 80, The Multiphase Interstellar Medium, ed. A. Ferrara, C. F. McKee, C. Heiles, & P. R. Shapiro (San Francisco: ASP), 292
McKee, C. F., & Ostriker, J. P. 1977, ApJ, 218, 148
Miville-Deschênes, M. A., Levrier, F., & Falgarone, E. 2003, ApJ, in press
Normandeau, M. 1996, Ph.D. thesis, Univ. Calgary
Oey, M. S., & Clarke, C. J. 1997, MNRAS, 289, 570
Oey, M. S., Clarke, C. J., & Massey, P. 2001, in Proc. 40th Conf. on Dwarf Galaxies and Their Environment, ed. K. S. De Boer, R.-J. Dettmar, & U. Klein (Berlin: Shaker), 181
Principe, J. C., Euliano, N. R., & Lefebvre, W. C. 2000, Neural and Adaptive Systems: Fundamentals through Simulation (New York: Wiley)
Stil, J. M., & Irwin, J. A. 2001, ApJ, 563, 816
Stutzki, J., Bensch, F., Heithausen, A., Ossenkopf, V., & Zielinsky, M. 1998, A&A, 336, 697
Taylor, A. R. 1999, in ASP Conf. Ser. 168, Radio Continuum Results from the Canadian Galactic Plane Survey, ed. A. R. Taylor, T. L. Landecker, & G. Joncas (San Francisco: ASP), 3
Tenorio-Tagle, G., & Palous, J. 1987, A&A, 186, 287
Thilker, D. A., Braun, R., & Walterbos, R. M. 1998, A&A, 332, 429
Weaver, H. F., & Williams, D. R. 1973, A&AS, 8, 1