Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2477525 C 2012/01/24

(11)(21) 2 477 525

(12) BREVET CANADIEN
CANADIAN PATENT

(13) C

(86) Date de dépôt PCT/PCT Filing Date: 2003/02/12
(87) Date publication PCT/PCT Publication Date: 2003/09/04
(45) Date de délivrance/Issue Date: 2012/01/24
(85) Entrée phase nationale/National Entry: 2004/08/26
(86) N° demande PCT/PCT Application No.: EP 2003/001361
(87) N° publication PCT/PCT Publication No.: 2003/073416
(30) Priorité/Priority: 2002/02/28 (ITTO2002A000170)

(51) Cl.Int./Int.Cl. *G10L 15/16* (2006.01)

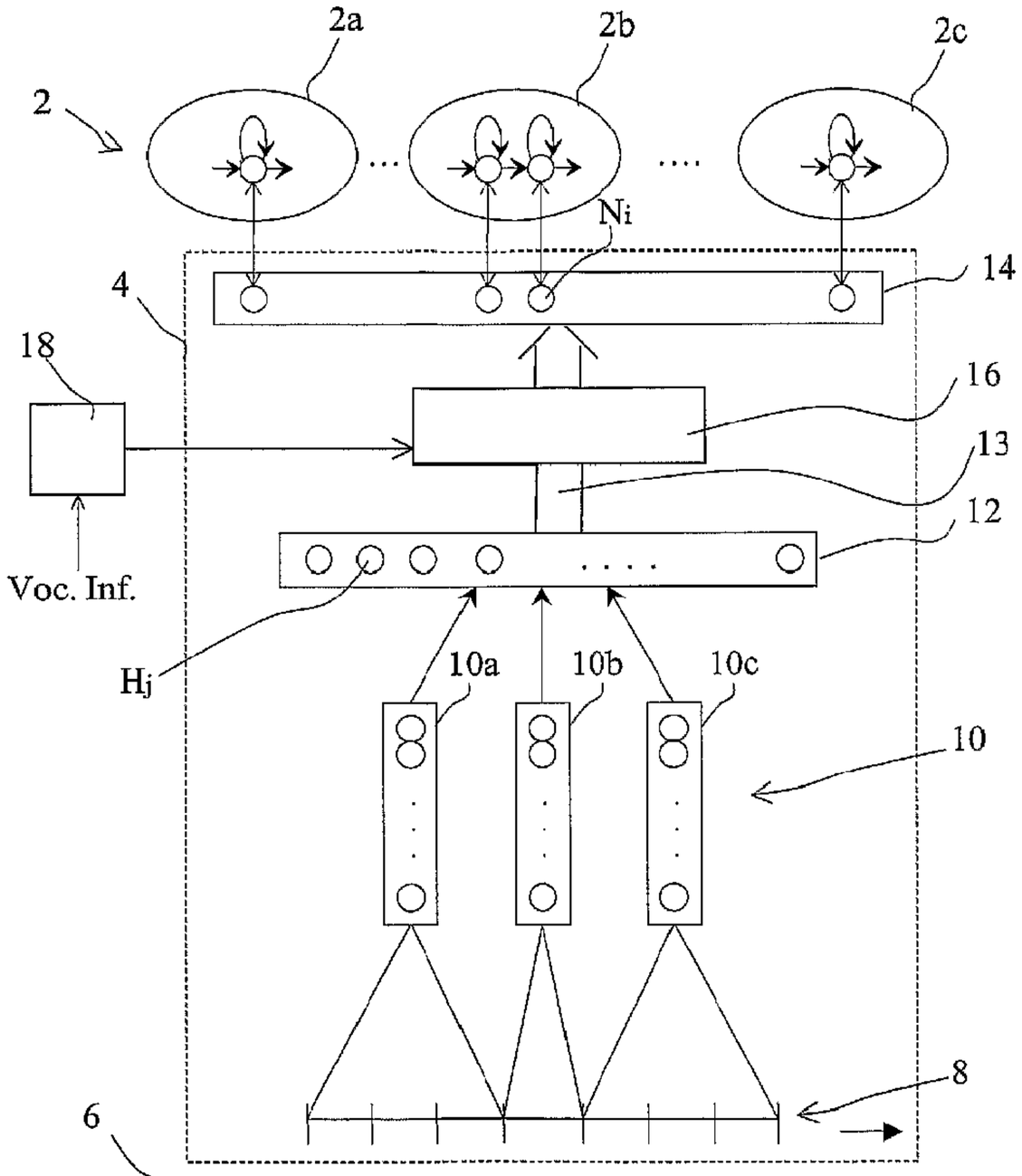
(72) Inventeurs/Inventors:
ALBESANO, DARIO, IT;
GEMELLO, ROBERTO, IT

(73) Propriétaire/Owner:
LOQUENDO S.P.A., IT

(74) Agent: RIDOUT & MAYBEE LLP

(54) Titre : PROCEDE PERMETTANT D'ACCELERER L'EXECUTION DE RESEAUX NEURONAUX DE RECONNAISSANCE VOCALE ET DISPOSITIF DE RECONNAISSANCE VOCALE ASSOCIE

(54) Title: METHOD FOR ACCELERATING THE EXECUTION OF SPEECH RECOGNITION NEURAL NETWORKS AND THE RELATED SPEECH RECOGNITION DEVICE




(57) Abrégé/Abstract:
A method for accelerating neural network execution (4) in a speech recognition system, specifically for recognition of words contained in one or more subsets of a general vocabulary, involves the following steps. - at the recognition system initialisation

CA 2477525 C 2012/01/24

(11)(21) 2 477 525

(13) C

(57) Abrégé(suite)/Abstract(continued):
phase, calculating the union of vocabulary subsets and determining the acoustic-phonetic units required for recognising the words contained in that union; re-compacting the neural network eliminating all the weighted connections afferent to computation output units corresponding to unnecessary acoustic-phonetic units; - executing only the re-compacted network at each instant of time.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
4 September 2003 (04.09.2003)

PCT

(10) International Publication Number
WO 03/073416 A1

(51) International Patent Classification[7]: G10L 15/16

(21) International Application Number: PCT/EP03/01361

(22) International Filing Date: 12 February 2003 (12.02.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
TO2002A000170 28 February 2002 (28.02.2002) IT

(71) Applicant *(for all designated States except US)*: **LOQUENDO S.p.A.** [IT/IT]; Via Nole, 55, I-10149 Torino (IT).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **ALBESANO, Dario** [IT/IT]; c/o Loquendo S.p.A., Via Nole, 55, I-10149 Torino (IT). **GEMELLO, Roberto** [IT/IT]; c/o Loquendo S.p.A., Via Nole, 55, I-10149 Torino (IT).

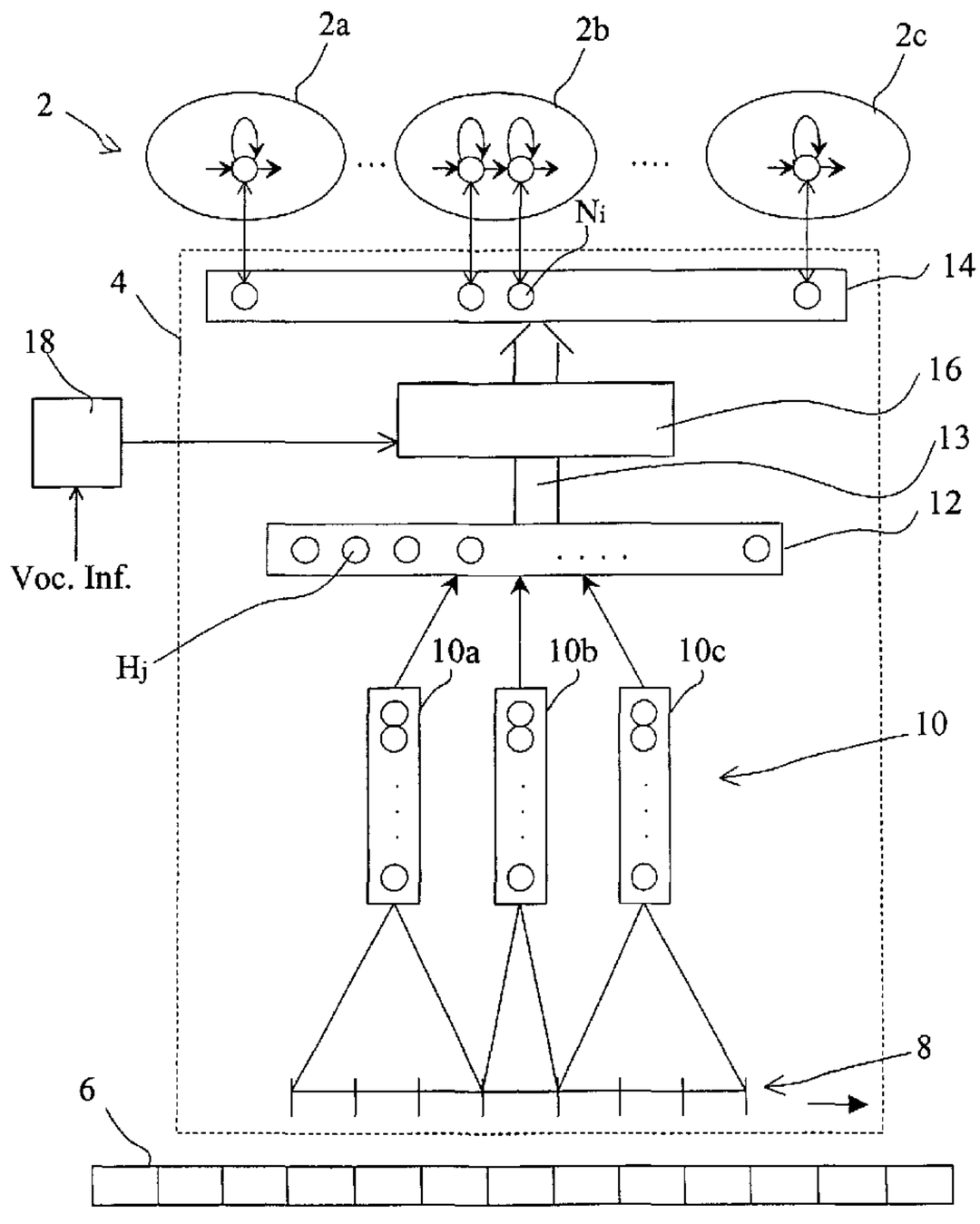
(74) Agents: **GIANNESI, Pier, Giovanni** et al.; Pirelli S.p.A., Viale Sarca, 222, I-20126 Milano (IT).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE,

*[Continued on next page]*

(54) Title: METHOD FOR ACCELERATING THE EXECUTION OF SPEECH RECOGNITION NEURAL NETWORKS AND THE RELATED SPEECH RECOGNITION DEVICE




(57) Abstract: A method for accelerating neural network execution (4) in a speech recognition system, specifically for recognition of words contained in one or more subsets of a general vocabulary, involves the following steps. - at the recognition system initialisation phase, calculating the union of vocabulary subsets and determining the acoustic-phonetic units required for recognising the words contained in that union; re-compacting the neural network eliminating all the weighted connections afferent to computation output units corresponding to unnecessary acoustic-phonetic units; - executing only the re-compacted network at each instant of time.

WO 03/073416 A1

WO 03/073416 A1

ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

**Published:**

— *with international search report*

# METHOD FOR ACCELERATING THE EXECUTION OF SPEECH RECOGNITION NEURAL NETWORKS AND THE RELATED SPEECH RECOGNITION DEVICE

## Technical Field

This invention refers to automatic signal recognition systems, and specifically regards a method for accelerating neural network execution and a speech recognition device using that method.

## Background Art

An automatic speech recognition process can be schematically described as a number of modules placed in series between a speech signal as input and a sequence of recognised words as output:

- a first signal processing module, which acquires the input speech signal, transforming it from analogue to digital and suitably sampling it;

- a second feature extraction module, which computes a set of parameters that well describe the features of the speech signal in terms of its recognition. This module uses, for example, spectral analysis (DFT) followed by grouping in Mel bands and a discrete transformed cosine (Mel based Cepstral Coefficients);

- a third module that uses temporal alignment algorithms and acoustic pattern matching; for example a Viterbi algorithm is used for temporal alignment, that is to say it manages the time distortion introduced by the different rates of speech, while for pattern matching it is possible to use prototype distances, the likelihood of Markovian states or *a posteriori* probability generated by neural networks;

- a fourth linguistic analysis module for extracting the best word sequence (present only for recognition of

continual speech); for example, it is possible to use models with bigrams or trigrams of words or regular grammar.

In the above model the neural networks enter into the third module as regards the acoustic pattern matching aspect, and are used for estimating the probability that a portion of speech signal belongs to a phonetic class in a set given *a priori*, or constitutes a whole word in a set of prefixed words.

Neural networks have an architecture that has certain similarities to the structure of the cerebral cortex, hence the name neural. A neural network is made up of many simple parallel computation units, called neurones, densely connected by a network of weighted connections, called synapses, that constitute a distributed computation model. Individual unit activity is simple, summing the weighted input from the interconnections transformed by a non-linear function, and the power of the model lies in the configuration of the connections, in particular their topology and intensity.

Starting from the input units, which are provided with data on the problem to solve, the computation propagates in parallel in the network up to the output units that provide the result. A neural network is not programmed to execute a given activity, but is trained using an automatic learning algorithm, by means of a series of examples of the reality to be modelled.

The MLP or Multi-Layer Perceptron model currently covers a good percentage of neural network applications to speech. The MLP model neurone sums the input weighting it with the intensity of the connections, passes this value to a non-linear function (logistic) and delivers the output. The neurones are organised in levels: an input

level, one or more internal levels and an output level. The connection between neurones of different levels is usually complete, whereas neurones of the same level are not interconnected.

With specific regard to speech recognition neural networks, one recognition model in current use is illustrated in document EP 0 623 914. This document substantially describes a neural network incorporated in an automaton model of the patterns to be recognised. Each class is described in terms of left-right automatons with cycles on states, and the classes may be whole words, phonemes or other acoustic units. A Multi-Layer Perceptron neural network computes automaton state emission probability.

It is known however that neural network execution is very heavy in terms of the required computing power. In particular, a neural network utilises a speech recognition system like the one described in the aforementioned document has efficiency problems in its sequential execution on a digital computer due to the high number of connections to compute (for each one there is an input product for the weight of the connection), which can be estimated as around 5 million products and accumulations for each second of speech.

An attempt at solving this problem, at least in part, was made in document EP 0 733 982, which illustrates a method for accelerating neural network execution, for processing correlated signals. This method is based on the principle that, since the input signal is sequential and evolves slowly over time in a continuous manner, it is not necessary to re-compute all the activation values of all the neurones for each input, but it suffices to propagate the differences with respect to the previous input in the

network. In other words, the operation is not based on absolute values of neurone activation at time t, but on the difference with respect to the activation at time t-1. Therefore at each point of the network, if a neurone has, at time t, activation sufficiently similar to that of time t-1, it does not propagate any signal forward, limiting the activity exclusively to those neurones with an appreciable change in activation level.

However, the problem remains, especially in the case of small vocabularies that use only a small number of phonetic units. Indeed, in known systems each execution of the neural network envisages computing of all output units, with an evident computation load for the system.

**Summary of the invention**

This invention proposes to solve the problem of how to speed up neural network execution, in particular in cases in which activation of all the output units is not necessary, leaving the functional characteristics of the overall system in which the neural network is used unchanged.

The advantage of the invention is that it does not compute all the neural network unit output activations exhaustively, but in a targeted manner instead according to the real needs of the decoding algorithm.

In this way not all the neural network output units are necessary, allowing for reduced mode network processing, reducing the processing load on the overall system.

**Brief Description of Drawings**

These and other features of the invention are

clarified in the following description of a preferred form of embodiment, given by way of example and by no means limiting, and in the annexed drawings in which:

Figure 1 is a schematic representation of an MLP or Multi-Layer Perceptron type neural network;

Figure 2 is a schematic illustration of a speech recognition device, incorporating a neural network produced according to this invention, and;

Figure 3 is a schematic illustration of an application of the method described in the invention.

**Detailed description of the preferred embodiments**

In reference to Figure 1, an MLP or Multi-Layer Perception type neural network includes a generally very high number of computing units or neurones, organised in levels. These levels comprise an input level 8, an output level 14 and two intermediate levels, generally defined as hidden levels 10 and 12. Neurones of adjacent levels are connected together by weighted connections, for example the connection $W_{ij}$ that unites neurone $H_j$ of hidden level 12 with neurone $N_i$ of output level 14, and each connection envisages computing of the neurone input value for the weight associated with the same connection. The great number of connections to compute, which can be estimated as around 5 million products and accumulations for each second of speech, highlights an efficiency problem in the sequential execution of the neural network on a digital computer.

For applications in the field of speech recognition, in terms of computing units or neurones, a neural network of the type illustrated in figure 1 may have the following structure:

| | |
|---|---|
| Input level: | 273 units (neurones) |
| First hidden level: | 315 units |

Second hidden level: 300 units
Output level: 686 units
(Total units: 1,574)

and the following values in terms of the number of weighted connections:

between input level and first hidden level:
7,665 connections
between first hidden level and second hidden level:
94,500 connections
between second hidden level and output level:
205,800 connections
(Total connections: 307,965)

It is evident that about 2/3 of the connections are located between the second hidden level and the output level. In effect, this is the connection level that has the greatest influence on the total computing power required by the overall system.

The proposed solution envisages not computing all the neural network output unit activations exhaustively, but in a targeted way according to the real needs of the decoding algorithm used, as will be explained in more detail later on.

Effectively, in many cases, at least in the field of speech recognition, it would suffice to compute only a subset of output units instead of the totality of output units as is commonly the case in traditional type systems.

To better understand the basic principle of this method and the related device, let us now analyse a model of a recognition system produced according to this invention, with reference to Figure 2.

The recognition model represented in Figure 2 substantially comprises a neural network 4 incorporated in an automaton model 2 of the patterns to recognise. Each

class is described in terms of left-right automaton 2a, 2b, 2c (with cycles on states), and the classes can be whole words, phonemes or acoustic units. A Multi-Layer Perceptron neural network 4 computes automaton state emission probability.

The input window, or input level 8, is 7 frames wide (each 10 ms long), and each frame contains 39 parameters (Energy, 12 Cepstral Coefficients and their primary and secondary derivatives). The total number of computing units or neurones in the input layer is indeed equal to 7 x 39 = 273. The input window "runs" on a sequence of speech signal samples 6, sampled on input from the preceding modules (signal processing, etc.).

The first hidden level 10 is divided into three feature extraction blocks, the first block 10b for the central frame, and the other two 10a and 10c for the left and right contexts. Each block is in turn divided into sub-blocks dedicated to considering the various types of parameter (E, Cep, $\Delta$E, $\Delta$Cep).

The second hidden level 12 transforms the space between the features extracted from the first hidden level 10 and a set of self-organised features (for example silence, stationary sounds, transitions, specific phonemes).

The output level 14 integrates this information estimating the probability of emission of the states of the words or acoustic-phonetic units used. The output is virtually divided into various parts, each corresponding to an automaton 2a, 2b, 2c, which in turn corresponds to an acoustic-phonetic class. The acoustic-phonetic classes used are so-called "stationary-transition" units, which consist of all the stationary parts of the phonemes plus all the admissible transitions between the phonemes

themselves. The stationary units 2a, 2c are modelled with one state, and each correspond to a network output unit, whereas the transition units are modelled with two states, for example unit 2b in Figure 2, and correspond to two network output units.

The advantage of the invention is that the weighted connections 13, which connect the second hidden level 12 computing units with those of the output level 14, can be re-defined on each occasion, working through a specific selection module 16 that permits selection of those connections thereby reducing their number. In practical terms the module 16 permits elimination of all connections afferent to output units not used at a given moment, thereby reducing the overall computing power necessary for network execution.

The selection module 16 is in turn controlled by a processing module 18 which, receiving speech information (Voc. Inf) input from the recognition system related to subsets of words that in that given moment the system has to recognise, translates that information into specific commands for the selection module 16, according to the method described.

According to the invention, it is possible to apply the method when the words to be recognised are contained in a subset of the general vocabulary of words that the speech recognition system is capable of recognising. The smaller the subset of words used, the greater the advantages will be in terms of reducing the required computing power. In effect, when using speech recognition on small vocabularies, for example the 10 numbers, the active vocabulary does not contain all the phonemes; thus not all the outputs of the network are necessary for their decoding and recognition.

The method for accelerating execution of the neural network 4 envisages the execution of the following steps in the order given:

- determining a subset of acoustic-phonetic units necessary for recognising all the words contained in the word subset of the general vocabulary;

- eliminating from the neural network 4, by means of the selection module 16, all the weighted connections ($W_{ij}$) afferent to computing units ($N_i$) of hidden level 14 corresponding to acoustic-phonetic units not contained in the previously determined subset of acoustic-phonetic units, thereby obtaining a compacted neural network 4', optimised for recognition of the words contained in that vocabulary subset;

- exclusively executing, at each instant in time, the compacted neural network (4').

The unnecessary connection elimination phase is schematically represented in figure 3, which shows a highly simplified example of a neural network with just four neurones $H_1..H_4$ in hidden level 12 and eight neurones $N_1..N_8$ in the output level 14. In a traditional type neural network each neurone in hidden level 12 is connected to all the neurones in the output level (14), and at each execution of the neural network, computes all the corresponding connections ($W_{ij}$).

Supposing in this case that the vocabulary subset only requires the presence of the acoustic-phonetic units corresponding to output neurones $N_3$ and $N_5$, highlighted in Figure 3, the method described in the invention permits temporary elimination of all the weighted connections afferent to unused acoustic-phonetic units, and the reduction of neural network execution to a limited number of connections (connections $W_{31}$, $W_{32}$, $W_{51}$, $W_{52}$, $W_{33}$, $W_{53}$, $W_{34}$,

$W_{54}$ highlighted in Figure 3).

With regard to the choice of the acoustic-phonetic unit subset, it generally suffices to include only the stationary units and the transition units of the phonemes that make up the words in the subset of the general vocabulary.

However, since there are a lot fewer stationary units than transition units, for example in the Italian language there are 27 to 391, it is preferable to leave all the stationary units in any case and eliminate all the transition units not present in the active vocabulary.

Moreover, since stationary units 2a and 2c are modelled with a single state, and correspond to a single network output unit, whereas the transition units are modelled with two states (and two network output units), the reduction in terms of network output units is in all cases very high.

The general vocabulary subset used in a given moment by the speech recognition system can be the union of several subsets of active vocabularies. For example, in the case of an automatic speech recognition system for telephony applications, a first word subset could consist simply of the words "yes" and "no", whereas a second subset could consist of the ten numbers from zero to nine, while the subset used at a given moment could be the union of both subsets. The choice of a vocabulary subset obviously depends on the complex of words that the system, at each moment, should be able to recognise.

Referring to the previously described neural network structure, with 1,574 computing units or neurones and 307,965 weighted connections and considering a subset containing the ten numbers (0 to 9), applying the method described in the invention would reduce the weighted

output connections to 26,100, and the total weighted connections to 128,265 (42% of the initial number). This would in theory save 58% of the required computing power.

In the case of the ten numbers, the selected acoustic-phonetic units would be 27 stationary plus 36 transition (of which 24 with 2 states), that is to say, just 87 states out of the 683 total.

Although this example refers to the Italian language, it does not depend on language and the method is indeed entirely general.

In terms of the algorithm for implementing the method in a recognition system, the method is structured into the following steps:

1) at the recognition initialisation phase computing the union of active vocabularies required for recognition;

2) re-compacting the last level of the neural network always leaving all the stationary units and only the transition units contained in the active vocabulary;

3) executing only the re-compacted network at each instant in time.

The experimental results obtained by implementing the illustrated method in a speech recognition system, with the task of recognising the ten numbers, gave an average reduction of 41% in required computing power. Vice-versa, as expected, there was no appreciable difference on large scale vocabularies (2000 words).

# CLAIMS

1. Method for accelerating neural network execution in a speech recognition system, for recognising words contained in a subset of a general vocabulary of words that the speech recognition system is capable of recognising, said neural network comprising a number of computing units organised in levels, among which at least one hidden level and one output level, the computing units of said hidden level being connected to the computing units of said output level via weighted connections, said computing units of said output level corresponding to acoustic-phonetic units of said general vocabulary, the method comprising:

- determining a subset of acoustic-phonetic units necessary for recognising all the words contained in said general vocabulary subset;
- eliminating from the neural network all the weighted connections afferent to computing units of said output level that correspond to acoustic-phonetic units not contained in said previously determined subset of acoustic-phonetic units, thus obtaining a compacted neural network optimised for recognition of the words contained in said general vocabulary subset;
- executing, at each moment in time, only said compacted neural network.

2. Method according to claim 1, in which said acoustic-phonetic units comprise stationary units and transition units, and said step of determining the subset of acoustic-phonetic units consists of determining the stationary units and transition units present in said general vocabulary subset.

3. Method according to claim 1, in which said acoustic phonetic units comprise stationary units and transition units, and said step of determining the subset of acoustic-phonetic units consists of selecting all the stationary units and determining the transition units present in said general vocabulary subset.

4. Method according to claim 2 or 3, in which said general vocabulary subset is the union of several subsets of vocabularies active at any given moment.

5. Speech recognition system, comprising a neural network with a number of computing units organised in levels, including at least one hidden level and one output level, the computing units of said hidden level being connected to the computing units of said output level via weighted connections, said computing units of said output level corresponding to acoustic-phonetic units of a general vocabulary of words to be recognized, the speech recognition system comprising means for accelerating neural network execution, for recognising words contained in a subset of said general vocabulary, said means comprising:

- a first module for determining the subset of acoustic-phonetic units necessary for recognising all the words contained in said general vocabulary subset;
- a second module for selecting, from among the weighted connections connecting the computing units of hidden level with those of output level, the weighted connections afferent to computing units corresponding to acoustic-phonetic units contained in said subset of acoustic-phonetic units determined by said first module, thus

obtaining a compacted neural network optimised for recognition of the words contained in said general vocabulary subset.

6. System according to claim 5, in which the acoustic-phonetic units comprise stationary units and transition units, and said first module comprises the means for determining both the stationary units and transition units present in said general vocabulary subset.

7. System according to claim 5, in which said acoustic-phonetic units comprise stationary units and transition units, and said first module comprises means for selecting all the stationary units and determining the transition units present in said general vocabulary subset.

8. System according to claim 6 or 7, in which said general vocabulary subset is the union of several subsets of vocabularies active at any given moment.

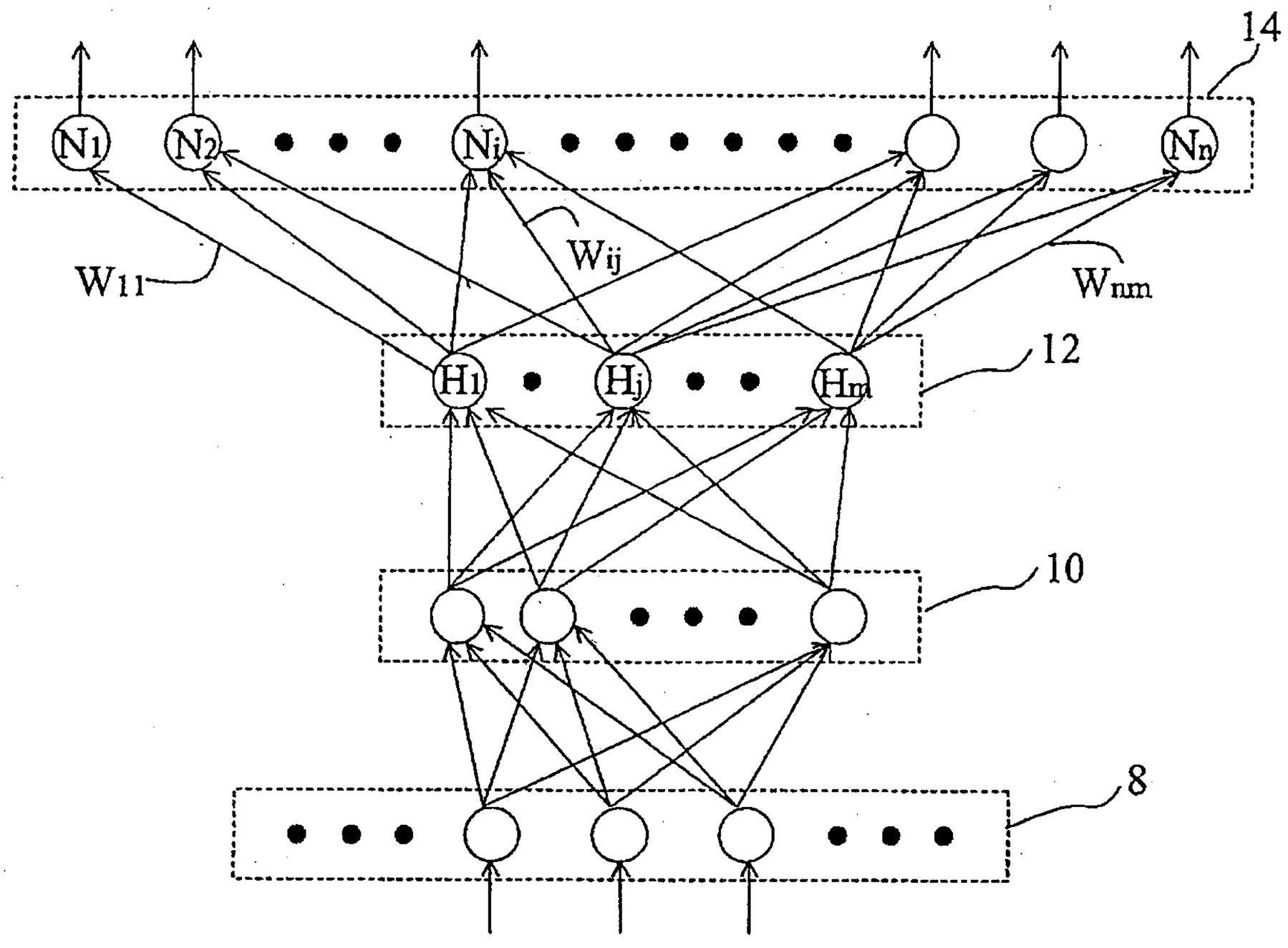



Fig. 1 (PRIOR ART)

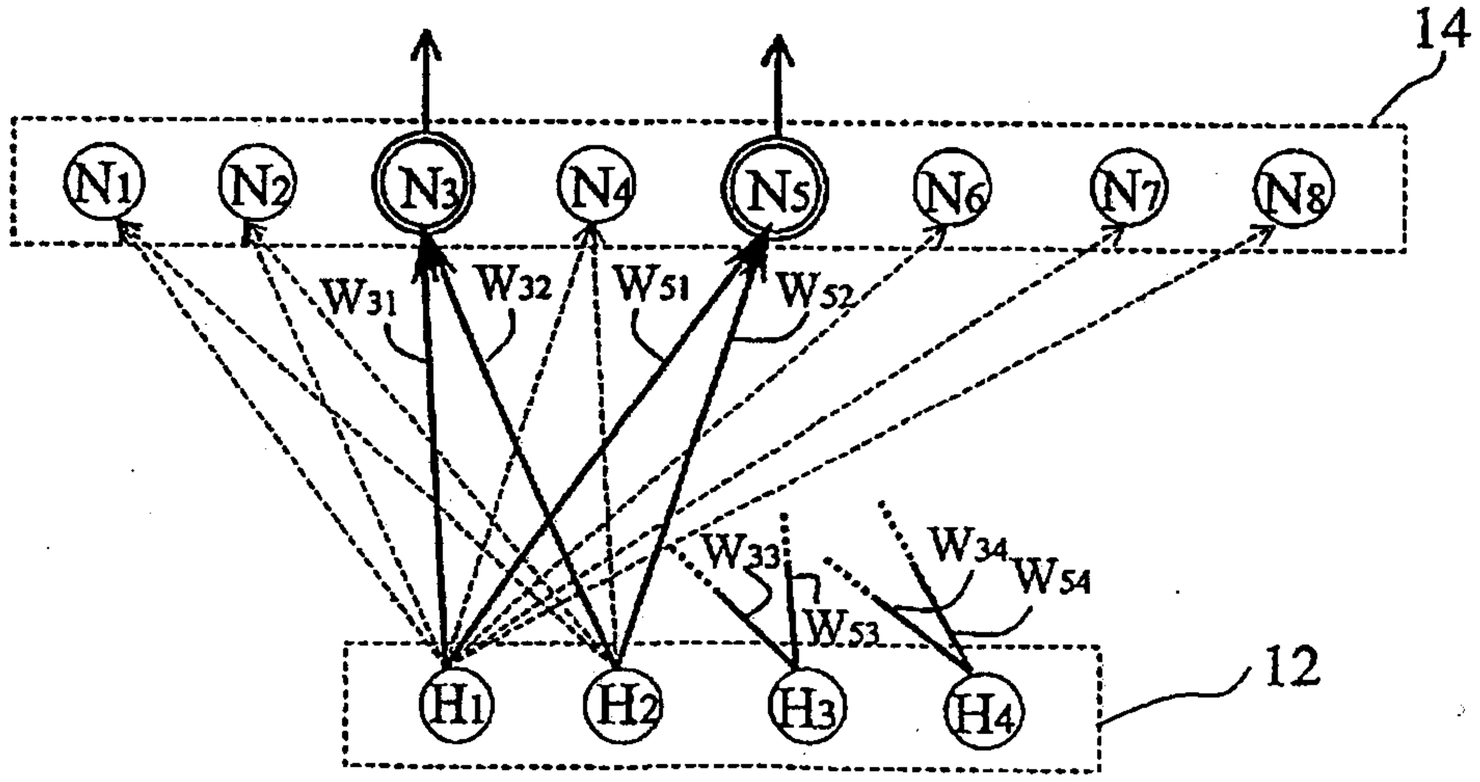



Fig. 3

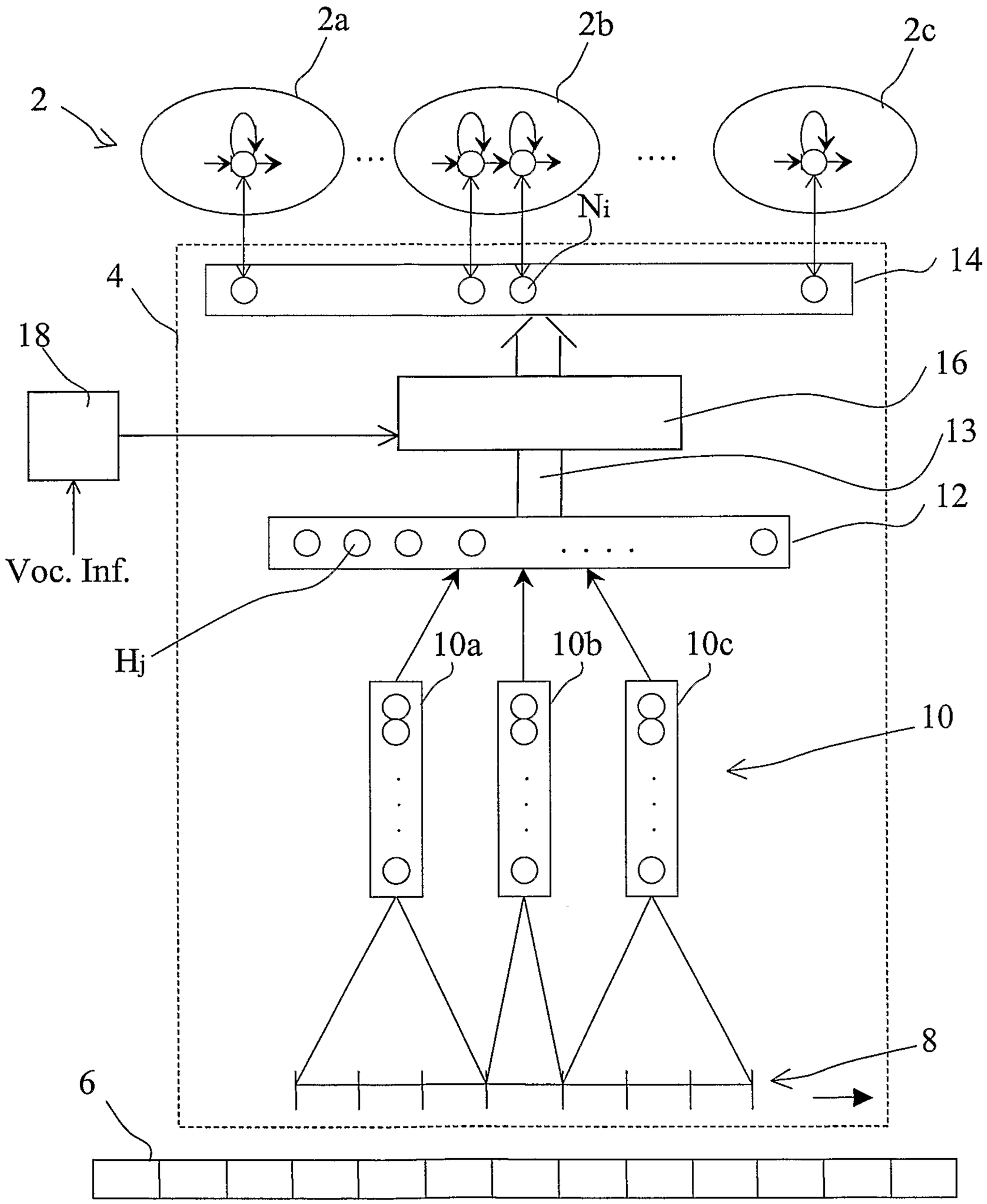
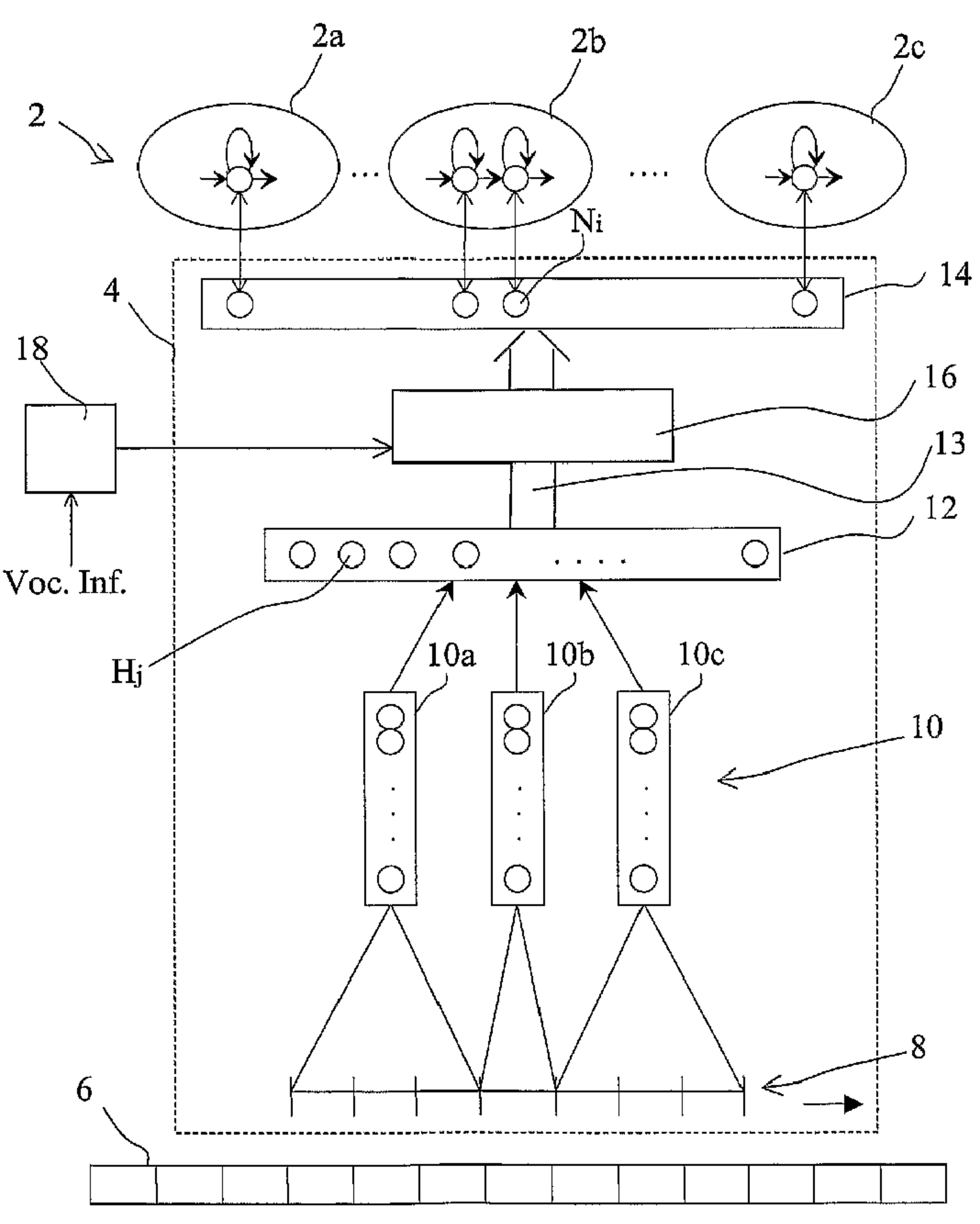



Fig. 2


2a
2b
2c
2
Ni
14
4
18
16
13
12
Voc. Inf.
Hj
10a
10b
10c
10
8
6